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, 2005a). 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

Recent work with distributional analyses has shown that they account for a wide range of phenomena, such as grammatical categorization (Mintz, Newport, & Bever, 2001), semantic priming (Lund, Burgess, & Atchley, 1995), word association strength (Griffiths & Steyvers, 2004), and metaphor comprehension (Kintsch & Bowles, 2002).

Many of these models start with basic distributional statistics (e.g., co-occurrence) and use that information to derive measures of contextual similarity. For example, Hyperspace Analogue to Language (HAL, Burgess & Lund, 2000) computes the 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 contextual usage history for all the words. Each word's row and column can be thought of as a measure of the contextual usage of a given word (in the forward and backward directions). Different rows or columns within the matrix can be compared, giving a measure of the similarity of usage of two words in terms of the other words with which they tend to co-occur. 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.

Another popular model based on word co-occurrence is Latent Semantic Analysis (LSA, Landauer & Dumais, 1997). LSA can be used to compute a co-occurrence matrix that specifies the number of times that a set of words occurs in a set of documents. Comparing two words' vectors in LSA can be thought of as comparing how similar two words are in terms of the set of documents in which they appeared. Words like *stealing* and *criminal* will tend to occur together in the same documents and thus have geometrically close LSA vectors. Singular value decomposition (SVD), a data reduction method similar to principle components analysis, is then typically used in conjunction with LSA 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 the vectors reflects a direct co-occurrence relationship between the words; two words become more similar in LSA the more they tend to appear 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 the similarity (cosine) of their document occurrence vectors is low (0.05) prior to SVD. After SVD, the similarity of their vectors is .634. This happens because, 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.

Newer statistical models have emerged, such as TOPICS (Griffiths, Steyvers, & Tanenbaum, in press) and BEAGLE (Jones & Mewhort, 2007), modeling an impressive range of data beyond HAL and LSA. However, given the emphasis of this paper is to see how far one can go with simple distributional statistics before relying on more complex models, this paper will use only simple statistics like conditional probability and distributional similarity as measured by HAL and LSA.

Measures of Distributional Statistics

Corpus

All distributional measures were obtained from a corpus of text derived from the online Wikipedia (2006) encyclopedia. The corpus contained 1,308,712 actual articles (redirects to other articles, user discussions, and maintenance articles were not used). From this set, we randomly sampled 250,000 articles (19% of 1.3 million articles) from which to compose our corpus. 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. Additionally, all the text in the corpus was lower cased. Punctuations were preserved but specially annotated. The cleaned corpus contains 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 analysis. 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. The final result was a word-by-document matrix used for LSA that was of size 10000x250000 (rows x columns) and a word-by-word matrix used for HAL that was 10000x10000.

Local Co-occurrence

Measures of local co-occurrence assess the proportion of time that two words co-occur in a particular context. The context could be defined as immediate adjacency, or within a window of text of any size. Given a *prime* and a *target* word (see Distributional Comparison 1) the local co-occurrence measure simply computes the number of times the two words appeared together and divides this by the frequency of the occurrence of the prime. This measure is computed for each window and subsequently averaged across all windows. This measure can be thought of to

reflect the direct associate strength between two words within some span of language. For the current study, local co-occurrence was calculated within an 8-word window and within each Wikipedia document.

Distributional Similarity

Distributional similarity is a way to assess indirect, higher-order relationships: how similar two words are in terms of the other words with which they co-occur. Two words may be unlikely to appear in the same sentence but highly likely to co-occur with the same set of other words (like *road* and *street*). Some words are highly unlikely to co-occur within a large sample of language (like *artist* and *sketching*) but likely to occur within the same types of large samples of language. These relationships are, respectively, what are measured by the HAL and LSA models.

Distributional similarity within an 8-word window was calculated using the HAL model. This involved deriving frequency normalized co-occurrence values for all 10,000 x 10,000 tracked word pairs. Co-occurrences involving pairs of the 10,000 words were tracked within an 8 word window, and weighted such that if words occurred immediately adjacently, it counted 8 times more toward the co-occurrence count than if the words co-occurred intervened by 7 other words. Each cell in the 10,000 x 10,000 matrix of co-occurrence values were then normalized for word frequency by dividing that cell's value by the word's total co-occurrence count. Finally, the similarity of the words' rows and columns (reflecting co-occurrences that came after and before the word in the 8-word window) were compared by correlating the vectors.

Distributional similarity within documents was calculated using the LSA model. Frequency counts were tabulated for all 10,000 tracked words within the 250,000 randomly sampled documents. These frequency counts were then normalized using the log entropy frequency normalization method usually used with LSA (Dumais, 1991). The word's similarities (in terms of their vectors of occurrences within documents) were then compared by calculating the cosine of the vectors. This similarity calculation was computed on vectors both before and after the wordxdocument matrix had been reduced by SVD.

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 tested and found significant priming for agents, patients, instruments, and locations. Based on their results and those of Ferretti et al, McRae et al concluded that nouns are not "just" arguments of verbs, and that any lexical item 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 analyses 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 easily through association and therefore that more complicated representations might be unnecessary. Further, the relationships that are missing in the input would imply which representational structures would be necessary, and the biases in learning that would be necessary 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 within an 8-word window; (2) co-occurrence within a document; (3) forward distributional similarity within an 8-word window; (4) backward distributional similarity within an 8-word window; (5) raw distributional similarity within a document; (6) SVD'd distributional similarity within a document. Paired t-tests were then computed for each dependant measure within each thematically related category.

Results

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

distributional similarity differences were statistically significant ($p < .001$) within a document (3) before and (4) after SVD (see Figure 2).

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

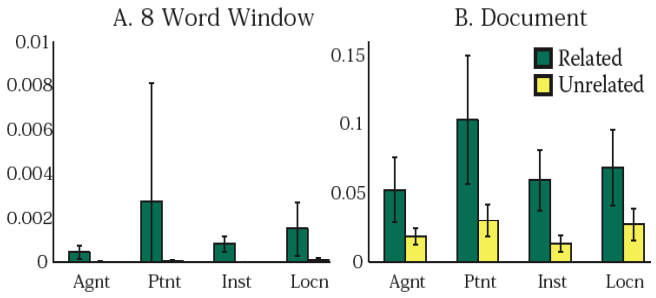
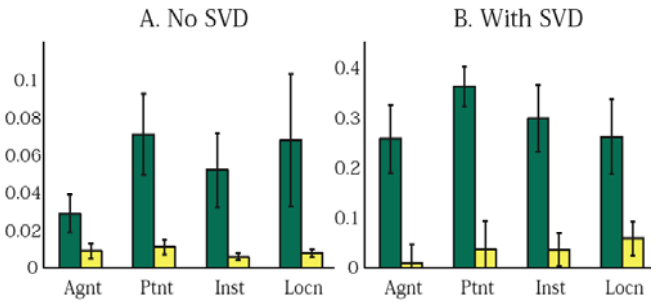
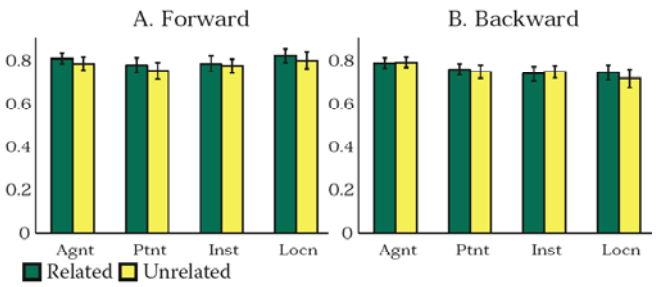


Figure 2. Distributional similarity before and after SVD for thematic role relationships.



The only distributional measure that did not consistently show significantly higher similarity for related pairs relative to unrelated pairs was distributional similarity within an 8-word window (see Figure 3). In one case (verb-agent), forward distributional similarity was significantly more similar for related than unrelated items (for verb-agent), and several other comparisons trended towards significance, but given the number of tests being performed and the power in these analyses due to the number of items, it is best treat these borderline significant effects with caution.

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



Discussion

As we argued in the introduction, the failure of the distributional statistics is often more informative 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 sentences (similarity as typically measured by the HAL model) to be a reliably significant predictor is very informative. The HAL model measures two words similarity in terms of the other words with which they typically co-occur within an 8-word window. Insignificant differences in this case imply that thematically related nouns and verbs are *not* more likely to share other words than unrelated nouns and verbs within the range of a typical sentence's length.

Distributional Comparison 2

McRae et al (2005), and historically many others (typically 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. We have shown in Comparison 1 that this turns out not to be the case. Other researchers less skeptical of distributional statistics might argue that the results from Comparison 1 are not much of a surprise. It is not entirely shocking that words like *knife* and *cut* co-occur more often than *rag* and *cut*. Arguably, a more rigorous test of distributional information's sufficiency would provide some insight into the statistics potential for establishing thematic representations. This requires testing to see if distributional statistics can distinguish not just thematically related from unrelated items, but also distinguish between more graded versions of relatedness.

In one recent study, 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, if not syntactically), or merely a semantic adjunct (allowed to be present, but not required). One such principle is that the noun must be obligatory. For example, the verb *behead* requires an instrument, and typically a pretty specific type of instrument (like a *sword*). In contrast, the verb *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 provide evidence that verbs that are paired with obligatory nouns aided 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

Table 2: 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)	

those nouns which are obligatory those that are not. And as both obligatory and non-obligatory pairs are technically thematically related (but with the obligatory pairs being "more" related), it provides a good test for the ability of distributional information to account for the graded structure of thematic role representations.

Methods

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

Procedure The procedure is the same as in the first distributional comparison.

Results

Means and standard errors for each dependant measure for each of the thematic categories are shown in Table 2. 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 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 within an 8-word window in the backward direction was significantly higher for obligatory items. Distributional similarity in the forward direction was not significantly different. 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.

Discussion

The conclusion from Distributional Comparison 2 is that not only are distributional statistics sufficient to establish associative links between thematically related nouns and verbs, they are also sufficient to establish more graded differences like those between obligatory arguments and non-obligatory adjuncts (as defined by Koenig et al , 2003). 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.

The particular pattern of successes and failures of the different distributional statistics is of considerable interest, and provides implications that will be worth exploring in future studies. The likelihood of co-occurrence within an 8-word window was significantly higher for obligatory noun-verb pairs, but the likelihood of co-occurrence within a document was significant in the reverse direction. One interpretation of this result is that, while obligatory nouns are likely to occur with their verbs, they are only likely to do so in within a very narrow window of the verb. Non-obligatory, optional arguments (or adjuncts) that tend to occur with a verb also show a reliable distributional pattern (i.e. their relatively high likelihood of co-occurring within a document). This might imply a very easily mapped structure for relations: obligatory nouns close to the verb, non-obligatory nouns in the periphery, and both more likely than completely unrelated nouns.

The success of backward but not forward distributional similarity in predicting obligatory pairs is also of interest. The meaning of such a result is that obligatory pairs are more likely to share items that preceded them than non-obligatory items. In a paper of larger scope and length, such a result calls out for a specific analysis of what words it is that these pairs tend to share, whether or not they are function or content words, and whether they come from earlier in the same sentence or from prior sentences. Differences here could have serious implications for how the semantic relatedness of these items interacts with other syntactic information.

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 window results, these results could be followed up to see what the other words are that 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, distributional statistics *are* sufficient to establish 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 a probabilistic constraints model based on direct lexical associations could also be a possible explanation for thematic role relatedness effects. There are two major issues to address in order to assess the fitness of a direct lexical association model relative to a feature-based thematic fit model.

The first question would be how a lexical association model would accommodate the context sensitive activation that Ferretti et al (2001) and other have demonstrated. 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 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 second issue deals with the nature of the graded structure of thematic relationships. Distributional structure clearly provides one type of graded structure that is highly influenced by syntactic use of the nouns and verbs, as well as patterns of co-occurrence with other words. Feature-based fit models likely provide a qualitatively distinct graded structure that focuses on different relationships, like non-lexical conceptual structure. 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.

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.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. DIetrich and A. Markman (Eds.), *Cognitive Dynamics*. New York: Erlbaum.

Dumais, S. T. (1991), Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2), 229-236.

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. (2002). Prediction and semantic association. In: *Advances in Neural Information Processing Systems, 15, 11-18.*

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.

Wikipedia (2006). http://www.wikipedia.org/. Retrieved online, October 14, 2006.

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

Willits, J. A., & Burgess, C. (2005a). Semantic priming: A comparison of feature-based and high-dimensional context models. Talk presented at Psychonomics Society Annual Meeting. Toronto, CA.