UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Comparing Semantic Space Models Using Child-directed Speech

Permalink
https://escholarship.org/uc/item/19m6t658

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 29(29)

ISSN
1069-7977

Authors
Riordan, Brian
Jones, Micheal N.

Publication Date
2007

Peer reviewed

Comparing Semantic Space Models Using Child-directed Speech

Brian Riordan (briordan@indiana.edu)
Department of Linguistics, 1021 E. Third St.
Indiana University, Bloomington, IN 47405 USA

Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. Tenth St.
Indiana University, Bloomington, IN 47405 USA

Abstract

A number of semantic space models from the cognitive science literature were compared by training on a corpus of child-directed speech and evaluating on three increasingly rigorous semantic tasks. The performance of families of models varied with the type of semantic data, and not all models were reasonably successful on each task, suggesting a narrowing of the space of plausible model architectures.

Keywords: semantic space models; child-directed speech; lexical development

Introduction

Semantic space models have proven successful at accounting for a broad range of semantic data, in particular semantic priming (Jones, Kintsch, & Mewhort, 2006; Lowe & McDonald, 2000). Since all the models are successful at accounting for the semantic data in most cases, however, finding tasks where the models make different predictions, and narrowing the space of plausible models, has proven to be quite difficult.

Semantic space models have traditionally been trained on adult language input. Further, the models are trained on very large corpora – in many cases, more data than humans experience. Finally, the models are usually only applied to modeling semantic data after processing the entire training corpus. Each of these steps is problematic.

The corpora semantic space models have been trained on range from Usenet postings (Burgess, Livesay, & Lund, 1998; Rohde, Gonnerman, & Plaut, submitted; Shaoul & Westbury, 2006) to the British National Corpus (Bullinaria & Levy, in press; Lowe & McDonald, 2000) to the TASA corpus (Jones & Mewhort, 2007). These corpora vary widely in their content and representativeness of human experience. However, the rationale for using a particular corpus is rarely supported by an evaluation of its representativeness. For example, Burgess et al. (1998) motivate the use of Usenet by claiming that Usenet represents "everyday speech" and is "conversationally diverse" – without presenting an analysis of the corpus that would justify this claim.

The training corpora for semantic space models are not only diverse, but large. The BNC totals 100 million words, the Usenet corpora used for HAL and HiDEX approach 300 million words, while COALS is trained on more than 1.2 billion words. It has been estimated that at a rate of 150 words per minute (a high estimate), reading 8 hours per day for 365 days of the year, it would take more than four years to read the full 100 million words of the BNC. This would make 12 years to encounter HAL's 300 million words, and 48 years to encounter all of the words COALS is trained on. At the very least, it would seem that these models are trained on the very high end of a scale of possible human input.

For the most part, semantic space modelers have only assessed model predictions after the entire training corpus has been processed (the exceptions being LSA (Landauer & Dumais, 1997) and BEAGLE (Jones & Mewhort, 2007)). What is lacking is a consideration of the rate at which the model learned its representations – information which may be crucial for assessing model plausibility.

In order to remove these potential advantages, in this study we compare a variety of semantic space models from the cognitive science literature using age-stratified child-directed speech (CDS) from the CHILDES database. For several reasons, CDS may offer us the important ability to decide between equally plausible models that perform comparably at a larger learning scale. First, CDS is arguably much more realistic than the adult corpora that semantic space models have been trained on: we know that children learn the meanings of words with this kind of input. Second, since the size of any corpus derived from the CHILDES database will be much smaller than other training corpora, it is more likely to be in the range of input for a human learner. Third, the caregiver speech in the CHILDES database can be divided according to the age of the target child. This allows the construction of training corpora that reflect changes in input over time, similar to what children are actually exposed to.

Two previous studies have explored the behavior of semantic space models when trained on CDS. Li, Burgess, and Lund (2000) trained HAL on the caregiver speech in CHILDES, at the time 3.8 million words. Denhière and Lemaire (2004) derived an LSA space from a 3.2 million word French corpus that included both children's speech and stories, textbooks, and encyclopedia articles written for children. However, it is not clear what is being modeled in these studies, as the training corpora aggregate a great deal of data from the linguistic environments of children of a variety of ages. The modeling target crucially affects the data on which the models should be evaluated.

Experimental Setup

Corpus. Four corpora were constructed from caregiver speech in the American section of the CHILDES database, one for each of four broad age ranges of target child: 12-23 months, 24-35 months, 36-47 months, and 48-60 months. The sizes of the corpora are listed in Table 1[1]. The Age 1 corpus represents all the American caregiver input to 12-24-month-olds in CHILDES; the other corpora were chosen to be of an approximately equal size.

Unlike previous studies that used CDS from the CHILDES database, the age group corpora used in this study were subjected to significant preprocessing. Given the small size of the corpora, the orthographic variation in CHILDES could potentially affect the semantic space models' representations. First, more than 700 word forms were standardized to eliminate orthographic variation. Second, all corpora were stemmed and lemmatized using a version of the Snowball stemmer (Porter & Boulton, 2006) augmented to change irregular verb forms to base forms. Third, most proper names in the corpora were converted to a single generic marker. The reason for this was to avoid proper names appearing in the list of context words for some models (see below).

Models. Semantic space models may be classified into two families based on architecture and representational scheme. One family, exemplified by HAL, computes a word-by-word co-occurrence matrix. In these models, words are more similar when they have appeared with the same neighboring words. Sahlgren (2006) dubs these "paradigmatic" spaces because of their tendency to emphasize paradigmatic relationships between words. Another family, exemplified by LSA, computes a word-by-context matrix, where a context may be a sentence, paragraph, etc. In these models words are similar to the extent that they appear in the same contexts. These spaces emphasize proximal relationships between words ("syntagmatic" spaces)[2].

Models of each family were selected for comparison (Table 2). The paradigmatic models included COALS (Rohde et al., submitted), HAL (Burgess et al., 1998),

Table 1: Sizes of the corpora constructed from the CHILDES database.

	Corpus size (words)	Cumulative corpus size
Age 1	460,384	
Age 2	460,743	921,127
Age 3	458,692	1,379,819
Age 4	450,097	1,829,916

---

[1] See Riordan (2007) for a list of the actual corpora used within each age range.

[2] *Random indexing* models use an alternative representational scheme in which a word's vector is assigned a distributed representation (e.g., Jones & Mewhort, 2007; Sahlgren, 2006) and may approximate either paradigmatic or syntagmatic models.

Table 2: Semantic space algorithms compared in this study.

Space Name	Context Specification	Lexical Association Function	Similarity Measure
"Paradigmatic" models			
COALS	Window (ramped)	Correlation	Correlation
HAL	Window (ramped)	Vector length normalization	Inverse Euclidean distance
HiDEX	Window (ramped)	Word frequency normalization	Inverse Euclidian distance
LLTR	Window	Log-likelihood coefficient	Cosine
Lowe & McDonald	Window	Positive log odds ratio	Cosine
PosPMI	Window	Positive mutual information	Cosine
"Syntagmatic" models			
Full dimensionality	20, 200, 2000 words	Entropy-based (no SVD)	Cosine
LSA	200, 2000 words	Entropy-based and SVD to 300 dimensions	Cosine

HiDEX (Shaoul & Westbury, 2006), a loglikelihood-transformed model (McDonald & Lowe, 1998; Padó & Lapata, 2003), Lowe and McDonald (2000), and a model based on positive pointwise mutual information (Bullinaria & Levy, in press). The syntagmatic models were LSA and a corresponding model without dimensionality reduction. To explore the effect of the size of the context region in syntagmatic models, three versions of the full dimensionality model and two versions of the LSA model were compared. In total, 11 models were compared. All models were reimplemented for this investigation.

For the paradigmatic models, context words were selected using an automatic procedure that approximated a specified number (500) of *content* words, considering words from most to less frequent in the corpora (omitting stop words)[3]. The context window size was constant, and set at 10 empirically (see Riordan (2007)). For consistency with the other models, the Euclidean distance metric used in HAL and HiDEX was converted into a similarity measure.

Target words. Target words for this investigation were selected to be sufficiently frequent and reliable in the full age group corpus. Reliability was determined through a procedure adapted from McDonald and Shillcock (2001). Words with a cumulative frequency of greater than twenty,

---

[3] The implementation of HAL used this context word selection procedure, rather than selecting the context words with the highest variance in the corpus, as in some HAL implementations. Lowe and McDonald's context word selection procedure was maintained.

plus all content words from the MacArthur CDI (Dale & Fenson, 1996) were included, for a total of 1892 targets.

Program of evaluation. The models were subjected to three increasingly rigorous semantic tasks: discriminating related from random word pairs, modeling adult semantic data, and modeling age-of-acquisition data.

Experiment 1: Word pair discrimination

As a first test of the models' abilities to derive adequate semantic representations when trained on CDS, we apply the models to the task of discriminating between, on the one hand, words that are known to be semantically related, and on the other, words that have been paired randomly. We assume that the distributional information available to a semantic space model should be sufficient for the model to locate related words in closer proximity than unrelated words in the high-dimensional space.

The related word pairs for this task were drawn from the University of South Florida Word Association Norms (USF; (Nelson, McEvoy, & Schreiber, 2004). In Nelson et al.'s word association task, subjects were given a *cue,* to which they were asked to respond with the first word that came to mind that was "meaningfully related or strongly associated" (2004: 403). Only one response was produced per cue.

A subset of 49,362 USF pairs that were included in the Maki, McKinley, and Thompson (2004) database formed the pool of candidate related word pairs. Of these, words pairs were constrained to be made up of words that were included in the 1892 word target set[4]. A set of 13,354 word pairs met this criterion. Unrelated word pairs were created by randomly pairing each cue with a response word, with the following constraints: no cue-response pairs from the actual cue-response word pairs could occur; no cue-unrelated response pair could occur more than once; no cue-unrelated response pair could be comprised of the same word repeated.

Similarities in each model were derived for each of the cue-response pairs and each of the cue-unrelated response pairs. These sets of scores were submitted to a oneway ANOVA. Models were deemed to have minimally discriminated between the sets of word pairs if the scores for the cue-response set were statistically greater than the scores for the cue-unrelated response set (indicating tighter clustering in the semantic space). The results for the ANOVAs for each of the spaces are shown in Figure 1. The ANOVAs for all spaces were significant, and in each case the average similarity score for the related word pair set was significantly greater than the unrelated word pair set. Thus, on average, each space located the related words closer in
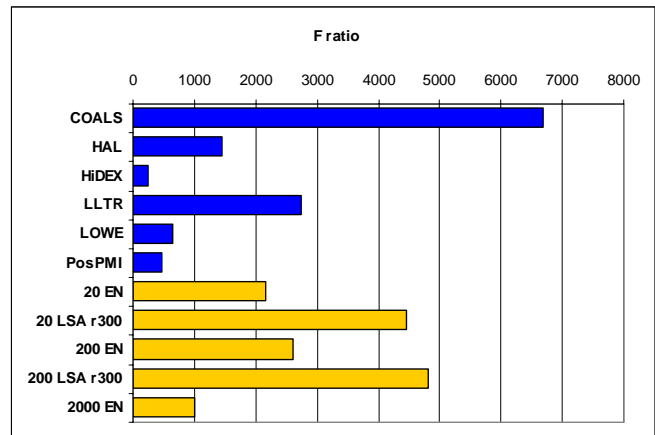

Figure 1: Discriminative abilities of each of the spaces on the full 13,354 word pair set.

semantic space than the unrelated words. At the same time, there was substantial variation in the degree to which related versus unrelated word pairs clustered in the spaces.

It should be noted that we cannot actually conclude from the size of the F-ratio in this task that one model is "better" than another. This is because we don't know what the "real" discrimination of these pairs, either for a child or for an adult, would look like, since experimental data for humans on this task does not exist.

Experiment 2: Modeling adult semantic data

As a more rigorous test of the models' representations and learning rates, we next compare the models on two related tasks where human data exists: modeling word association strengths, and modeling semantic distance in WordNet.

The *forward strength* in the USF word association norms is the probability that a cue will elicit a particular response. Using the 13,354 word pairs from Experiment 1, the models were compared on their abilities to predict these forward strengths from representation similarity. In this experiment the age group corpora are organized cumulatively, so that the models are exposed to greater amounts of age-appropriate speech (see Table 1).

For each of the cumulative corpora, the similarities in each model were derived for each of the 13,354 cue-response pairs. These similarities were entered into a linear regression to predict the forward strengths for the corresponding word pairs. The forward strengths were drawn from the Maki et al. database.

With increasing age-appropriate input, we expect the variance in the adult semantic data that is explained by the models to increase, as the models' semantic representations become more "adult-like". More specifically, after training on each cumulative corpus there should be an increase in the correlations of the word-pair similarity scores derived from the models and the semantic similarity scores from the human data.

Figure 2 plots the change in correlation of each of the models' similarity scores with both types of semantic data

---

[4] Stemming the candidate word pairs further restricted the pool of candidates, since some cue-response pairs became indistinguishable as a result (because of plural words used as cues, etc.).
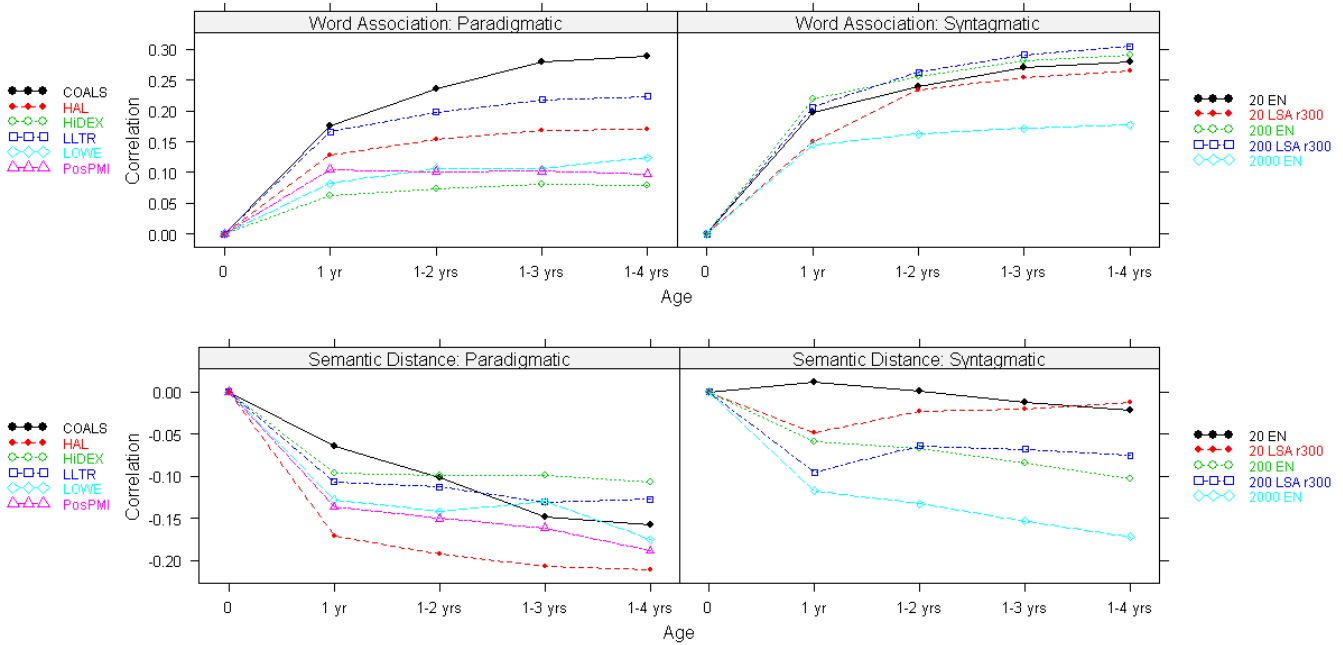
Figure 2: Correlations between the models' similarity scores and the semantic data.

as more age-appropriate caregiver speech is encountered. The syntagmatic models generally explain more of the variance in the word association data than the paradigmatic models. They have higher average correlations with the data after the Age 1 corpus (.184 vs. .120) and after processing each of the corpora (.264 vs. .164). This may be related to the better match of the syntagmatic architecture with the word association data (see Sahlgren (2006)). On the other hand, the syntagmatic models are nearly uniform in their trajectory of improvement over time, while paradigmatic models tend to show more variation.

In this task, even the best models only reached a correlation of about .3 with the word association data. The models' concomitant R^2 was also low, explaining less than 10% of the variance in the data.

Despite the instructions in the word association task to consider "meaningfully related" responses (Nelson et al., 2004), subjects often produce responses that are collocated with the cue but not necessarily semantically related. Although some researchers argue that distinguishing between semantic and associated relationships is futile (Nelson et al., 2004), other data regarding lexical semantic relatedness that focus more on semantic relationships do exist. Maki et al. (2004) derived semantic distances between word pairs in WordNet using the Jiang-Conrath distance measure (JCN). JCN is an information-theoretic measure of semantic distance in the WordNet hierarchy. JCN distances have been shown to correlate highly with human judgments of semantic similarity (Maki et al., 2004).

Word pair similarities for the 13,354 word pair set were used to predict the corresponding JCN distances as reported in Maki et al. (2004). The lower half of Figure 2 plots the correlations for the paradigmatic and syntagmatic models on this task. Note that since the JCN measures are distances,

not similarity scores, the models' scores are negatively correlated with the distances.

On this task, the paradigmatic models explain more variance in the adult data, reflecting the nature of the WordNet resource: WordNet is a hierarchical taxonomy split into noun and verb parts, and the links between words reflect paradigmatic relationships. In addition, the models that perform best on this task are not the same as those that performed the best in accounting for the word association data. For example, while the 200-word context syntagmatic models predicted the most variance in the word association task, here the 2000-word context model was the best. This likely reflects the fact that more paradigmatic information is available in the larger context.

The best models showed monotonically increasing correlations with the semantic data as they were exposed to more input, and relatively high correlations with the data in both tasks. Among the paradigmatic models, COALS and HAL met these criteria, while among the syntagmatic models only the 200-word context full dimensionality space did. HAL performed surprisingly well, especially given that the other paradigmatic models were designed to be improvements on its parameter choices. The LSA models' non-monotonicity and similar performance to the unreduced syntagmatic spaces indicate that dimensionality reduction does not automatically produce spaces that are more highly correlated with human semantic data.

Experiment 3: Modeling age-of-acquisition

Experiment 2 tested the models' overall learning trajectories. Most of the models gradually explained more of the variance in the adult semantic data as they were trained on age-appropriate input. In this experiment, we focus more

closely on models' learning rates by comparing the models' abilities to model age-of-acquisition (AoA) data when trained on the cumulative input of the age group corpora. Models that more closely match AoA data may be said to have learning rates that more closely resemble those of children.

For the purposes of this experiment, as a proxy for acquisition, we consider stabilization in the neighborhoods of words in semantic space. We will define a word's semantic neighborhood as the nearest *n* words in a given space. At a given time, *t,* we can find the semantic neighborhood for a word. At a later time, *t+1,* after the model has been exposed to more input, we can again find the neighborhood for the word, and compare it to the word's neighborhood at time *t*. As we continue this process, we will produce a record of the stabilization of a word's semantic neighborhood over time. We hypothesize that early-acquired words' neighborhoods will stabilize more quickly than those of later acquired words.

For age-of-acquisition (AoA) ratings, the norms of Bird, Franklin, and Howard (2001) were used. After stemming, there were 689 words that overlapped between the Bird et al. norms and the target words.

To compare semantic neighborhoods, we use a modified version of *combinatorial similarity,* originally proposed in Hu, Cai, Graesser, and Ventura (2005):

$$C'_{x,T} = \frac{\left\| S_{1,x,t} \cap S_{2,x,t} \right\|}{\left\| t \right\|}$$

Here, $S_{1,x,t}$ is the top *t* neighbors of a word *x* in space 1. $S_{1,x,t}$ is composed of sets of $s_i(x,y)$, the similarity scores of words *x* and neighbors *y* in space 1. The numerator here is simply the intersection of the top *t* neighbors, ignoring the similarity scores. Instead of dividing by the union of the neighbors in the neighborhoods as Hu et al. propose, we normalize by the number of neighbors in the neighborhoods being compared (e.g. 10).

Sample points of even intervals are established across the age group corpus. At each sample point, the semantic neighborhoods of the target words are computed. The semantic neighborhoods of words at successive sample points are compared using the above measure of neighborhood overlap. Once the neighborhood change history for the target words is established, a change coefficient is calculated for each word. This is computed as the average of the absolute values of the differences in neighborhood overlap from point to point:

$$change = \frac{\sum_{i=0}^{p-1} \left| C'_{i+1} - C'_i \right|}{p}$$

where *p* is the number of sample points and *C'* is computed neighborhood overlap. Early-acquired words should have lower change coefficients, as their neighborhoods stabilize quickly. The neighborhood size was set at 10 and the number of sample points was 8.
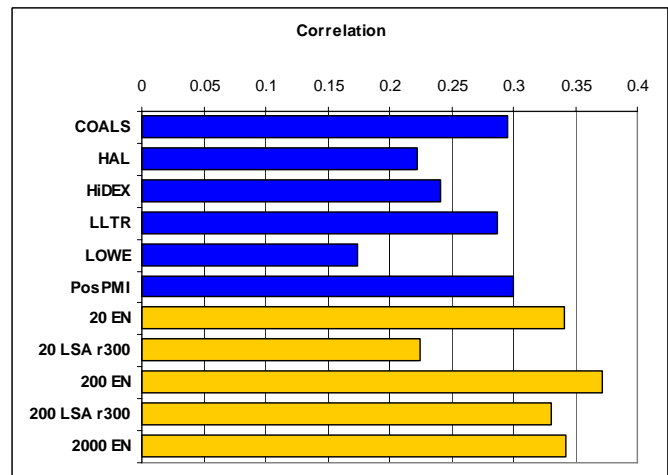

Figure 3: Correlations of the models' average change coefficients with the Bird *et al* AoA norms (Neighborhood size = 10; sample points = 8).

The correlations for the remaining models with the Bird et al. data are presented in Figure 3. Differences in the models' correlations were compared using Williams' ratio for nonindependent correlations (see Steiger, 1980). Among the paradigmatic models, COALS, LLTR, and PosPMI were significantly more correlated with the AoA data than HAL and Lowe and McDonald (e.g. COALS vs. HAL: $t(686) = 2.15$, $p < .05$; COALS vs. Lowe and McDonald: $t(686) = 3.92$, $p < .01$). The syntagmatic models were comparable, with the exception of the 20-word context LSA model, which was significantly less correlated with the AoA data (e.g. 200 LSA r300 vs. 20 LSA r300: $t(686) = 2.83$, $p < .01$). However, the 200-word context full dimensionality model was significantly more highly correlated with the AoA data than PosPMI, the best paradigmatic model: $t(686) = 2.26$, $p < .05$.

In this experiment, evidence of significant correlations between the stabilization patterns in the models and a set of AoA norms were found. While significant, however, the correlations of the models and the data were still rather low (all R^2 values were less than .12). With a few exceptions, the better-performing models on the previous tasks also performed well on this task.

Conclusion

This study represents a first attempt to compare a number of semantic space models on a common corpus with common evaluation tasks. The type of corpus used – CDS – was selected because it is more realistic than previous training corpora in terms of quantity and content.

Using CDS from CHILDES also naturally allowed an examination of the models' learning rates. The learning rate is a crucial yardstick by which to measure models' performance: if models are to be taken as models of both lexical acquisition and representation, as Landauer and Dumais (1997) and others have argued, they must perform

reasonably given a corpus that is an accurate representation of what children learn from.

While all models showed significant discriminative ability between random and related word pairs, over the course of two further tasks, we discovered that some models did not have plausible acquisition rates (given our broad assumptions of what should constitute acquisition in a semantic space). We also found a great deal of variation in the representations that models derived from the same data, which in turn likely affected their learning rates. Because most models differed from each other on a number of parameters, further investigation of the parameters that are the sources of the variation in performance is necessary.

At a wider angle, corroborating Sahlgren (2006), we found that certain families of models are better at certain semantic tasks: "syntagmatic" models better accounted for word association, a task that often emphasizes sequential relationships between words, while "paradigmatic" models better accounted for semantic data in the absence of association. It would appear difficult to maintain the notion that any one semantic space model is an optimal model of human semantic learning and memory.

In general, the models' abilities to explain the variance in the human data were low. There are many possible reasons for this: data sparsity in CHILDES, idiosyncrasy and noise in the semantic data themselves, as well as limitations on learning from co-occurrence data and lack of extra-linguistic information. The relative contributions of these factors to model performance deserve further investigation.

References

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, and Computers, 33*(1), 73-79.

Bullinaria, J. A., & Levy, J. P. (in press). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes, 25*(2/3), 211-257.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavioral Research Methods, Instruments, & Computers, 28*, 125-127.

Denhière, G., & Lemaire, B. (2004). A Computational Model of Children's Semantic Memory. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 297-302). Hillsdale, NJ: LEA.

Hu, X., Cai, Z., Graesser, A. C., & Ventura, M. (2005). Similarity between semantic spaces. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: LEA.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*(4), 534-552.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (Ed.), *Proceedings of the Thirtieth Stanford Child Language Research Forum* (pp. 167-178). Stanford, CA: CSLI.

Lowe, W., & McDonald, S. (2000). The direct route: mediated priming in semantic space. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 806-811). Hillside, NJ: LEA.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, and Computers, 36*(3), 421-431.

McDonald, S., & Lowe, W. (1998). Modelling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society (CogSci '98)* (pp. 675-680). NJ: LEA.

McDonald, S., & Shillcock, R. C. (2001). *Contextual Distinctiveness: A new lexical property computed from large corpora*: University of Edinburgh Informatics Research Report EDI-INF-RR-0042.

Nelson, D. L., McEvoy, C., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers, 36*(3), 402-407.

Padó, S., & Lapata, M. (2003). Constructing Semantic Space Models from Parsed Corpora. In *Proceedings of the 41st Meeting of the Association of Computational Linguistics, Sapporo, Japan* (pp. 128-135).

Porter, M., & Boulton, R. (2006). Snowball stemmer.

Riordan, B. (2007). *Comparing semantic space models using child-directed speech.* Doctoral dissertation, Department of Linguistics and Program in Cognitive Science, Indiana University, Bloomington.

Rohde, D., Gonnerman, L., & Plaut, D. (submitted). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*.

Sahlgren, M. (2006). *The Word-Space Model.* Stockholm University, Doctoral dissertation, Department of Linguistics, Stockholm University.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods, 38*(2), 190-195.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.