Naturalistic Word-Concept Pair Learning With Semantic Spaces

Brent Kievit-Kylar[1], George Kachergis[3], and Michael N. Jones[1,2]
{bkievitk, gkacherg, jonesmn}@indiana.edu
[1]Cognitive Science Program
[2]Department of Psychological & Brain Sciences
Bloomington, IN 47405 USA
[3]Psychology Department, Leiden University
Leiden 2311 EZ Netherlands

Abstract

We describe a model designed to learn word-concept pairings using a combination of semantic space models. We compare various semantic space models to each other as well as to extant word-learning models in the literature and find that not only do semantic space models require fewer underlying assumptions, they perform at least on par with existing associative models. We also demonstrate that semantic space models correctly predict different word-concept pairings from existing models and can be combined with existing models to perform better than either model can individually.

Keywords: Childes; natural language processing; semantic space models; associative learning.

Introduction

While the task of word-concept matching may seem trivial to an adult, imagine the task from the perspective of a young child. A child hears a series of vocalizations, which are then parsed into word units, and must perceive instances of objects in their environment through visual inspection. From that information the child must determine a set of objects or concepts that are present. These two tasks are challenging enough, but the child must then find a way to correlate the words that he/she has heard with the objects in the immediate environment. Extracting the correct mappings from the myriad possible ones is complicated by things such as the potential absence of matches between object and word (e.g., an object is mentioned that is not present), and the fact that not all words refer to objects (e.g., verbs and function words).

Several techniques have been proposed to help simplify the word-concept acquisition problem, the majority of which require the child to have prebuilt assumptions (e.g., a novel word must map to a novel object), or to perform complex Bayesian logic calculations (Frank et al. 2007). In this paper, we explore a new class of model that is based on the rapidly growing field of semantic space models. In particular, we generalize Kintsch's (2001) *predication* algorithm to the problem of word-concept learning in semantic spaces.

Kintsch's (2001) algorithm simulates the process of matching based on shared neighbors in a semantic space. The result of our adaptation is a model that learns to map words and objects to semantic clusters, greatly simplifying the problem of word-object mapping. Rather than casting the problem as one of learning associations between independent words and independent objects, a semantic space approach can take advantage of the fact that similar words carry mutually reinforcing information about each other's object referents. In addition, the similarity between a noun and semantically related verb or adjective contains information about the noun's referent. *Bounce* may often be used when a ball is present in the environment, even in absence of the noun *ball.* The semantic similarity between the words *bounce* and *ball* may be used as an indirect cue to the mapping between the noun and object.

We next provide background on the problem, data, and existing models of word-concept learning. Then we turn to a summary of a variety of semantic space models used, and a general purpose technique for creating word-concept learners from semantic models adapted from Kintsch's (2001) algorithm. Finally, we test these models on a labeled fragment of the CHILDES corpus and explore the benefits of combining different semantic models into hybrids.

Child Learning Models

While there are a number of existing word-concept mapping models from the child learning literature, we will focus on two recent models that have both been applied to the object-tagged corpus data that we use (described below).

The data used for training in these simulations are from an annotated version of the Rollins section of the CHILDES corpus (MacWhinney, 2000) used by Frank et al (2007). The entire corpus takes place over approximately ten minutes of talk taken from a caregiver to a child. Each sentence is annotated with the objects that are visible to the child when that sentence was being spoken. Thus the corpus consists of entries in the form $\{W_0, W_1, \ldots, W_n, C_0, C_1, \ldots C_m\}$ where W_i is a word token and C_i is a concept token (each represented by a string). A unique identifier was used to differentiate concepts from words, as both are represented in the dataset by similar strings of characters. In this paper, we will use the angle brackets as delimiters, such that "dog" represents the word dog, and "<dog>" represents the concept or object dog.

Frank et al.'s (2007) Bayesian Framework

Frank et al. (2007) propose a Bayesian model to jointly learn word-concept mappings, as well as which objects a speaker intends to speak about in a situation. Using a model similar to Latent Dirichlet Allocation (LDA) used by Topic models, they assume that words are generated from the lexicon according to what objects are present and are likely

to be talked about (i.e., the intention of the speaker). This model is a computational-level model that does not specify learning mechanisms, but rather specifies how to calculate the likelihood of a particular lexicon, given all of the situations that one has observed. The inferred lexicon is simply a collection of word-object pairings, and tends to be small because the prior favors smaller lexicons. The model handles nonreferential words: if a given word appears with many objects only a few times, these mappings will likely not be added to the lexicon. The inferred lexicon will mostly be comprised of the highest co-occurring word-object pairs; there is no explicit penalization for linking words to multiple objects, nor a word to multiple objects. There is no learning of associations among words, nor among objects. Frank et al. demonstrate impressive performance from this model on subsequent testing of word-object pairings.

The semantic space approach we propose differs theoretically from the Frank et al. (2007) model in at least two ways. Firstly, while the Frank et al. approach attempts to calculate an underlying generator that maps from concepts to words through the lexicon, the semantic space approach is more passive, projecting words and concepts onto points in psychological space. Secondly, the semantic space approach attempts to learn the relations between words, including the relations between concepts. This added structure allows a semantic space model to bootstrap additional partial information from indirect relationships.

Kachergis, Yu, & Shiffrin (2012) Associative Model

Kachergis et al. (2012) introduced an incremental model that learns word-object associations. Competing attentional biases for familiarity (i.e., already-strong associations) and for stimuli with uncertain associates (i.e., high entropy) allow this model to exhibit mutual exclusivity and other word-learning principles, as well as associative learning effects such as blocking and highlighting (Kachergis, 2012).

The model stores knowledge in M, a word-object association matrix that grows during training. Cell $M_{w,o}$ is the strength of association between word w and object o. Before the first trial, M has no information: each cell is set to $1/m$. Association strengths decay, and on each new trial a fixed amount of associative weight, χ, is distributed among the associations between words and objects, and added to the strengths. The rule for distributing χ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., χ) is given to this pair—a prior knowledge bias. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice versa, do not attract much attention. Stimulus uncertainty is captured using entropy (H), a measure that is 0 when the outcome of a variable is certain (e.g., a word appears with only one object), and maximal ($log_2 n$) when all of the n possible object (or word) associations are equally likely (e.g., for a novel stimulus, or one that appears with all stimuli equally). In the model, on each trial the entropy of each word and object is calculated from the normalized row (column) vector of associations for that word (object), $p(M_{w,\cdot})$, as follows:

$$H(M_{w,\cdot}) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}$$

In this equation, α is a parameter governing forgetting, χ is the attention weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly χ associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. This model aims to capture the process of learning simple word-concept associations using basic cues a learner may have.

Semantic Space Models

Semantic space models have seen a great amount of both attention and success in the literature over the past decade. There are a variety of semantic space models currently in the literature, but all are fundamentally based on the assumption that the contexts in which a word occurs may be used to infer its meaning, commonly projected into a high-dimensional psychological space. Words that frequently co-occur in contexts together, or that frequently occur in similar contexts, become more proximal in semantic space. We explore a variety of semantic space model representations of the CHILDES data here, all using the same mapping mechanism adapted from Kintsch's (2001) algorithm. We next very briefly describe each representation model used in our comparison.

BEAGLE

The BEAGLE model (Jones & Mewhort, 2007) uses holographic vector manipulation to represent word similarities. In BEAGLE, each new word encountered is assigned an environmental vector with elements generated independently from a Gaussian distribution, and a lexical vector of the same length but initialized to zeros. When encountering a sentence, the environmental vector of each word is added to the lexical vector of each word it co-occurs with. Similarity is measured using cosine similarities between words' lexical vectors.

ESA

Explicit semantic analysis (Gabrilovich & Markovitch, 2007) was designed for use with the Wikipedia corpus. It uses a centroid-based classifier that correlates given input

text to a weighted list of concepts associated with each target word.

FDTRI

Fixed Duration Temporal Random Indexing, introduced by Jurgens and Stevens (2009), attempts to bypass the computational difficulty inherent in singular value decomposition through the use of random projections onto lower dimensional space. Similar to BEAGLE, each word has an environmental vector, although FDTRI vectors are generated to be sparse. Rather than producing a word-by-meaning matrix, FDTRI incorporates time in a word-by-meaning-by-time tensor. The additional temporal information could be useful in word-concept pairing, if the sequential information given by the caregiver is relevant to object detection. For example, a caregiver may be more likely to start with the label and then continue with a description of the object.

HAL

The Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996) uses a fixed size window that is slid along the corpus. A matrix is built which is an accumulation of pairs of words that co-occur within any given window of text. Order information is partially preserved through the use of "occurring before," and "occurring after" co-occurrence matrices.

LSA

Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) operates by applying singular value decomposition to a word-by-context frequency matrix, reducing the matrix from high dimensionality (documents) to lower dimensional space (latent semantic components). The premise is that the reduction removes irrelevant features of the word usage, yielding a semantic abstraction in the resulting space.

ISA

Baroni, Lenci and Onnis (2007) developed incremental semantic analysis (ISA) to analyze children's speech data. ISA is based on random indexing models with a few variations. First, updating occurrence information includes both the signature (or environmental vector) as well as information on the learned history of the other word. This allows ISA to capture higher order relations. Second, word frequency discounts are updated online as the model learns for information about the distribution of words in the world.

PMI

Pointwise Mutual Information (PMI; Church & Hanks, 1989) is a basic information theoretic metric that looks at the probability of two words occurring together relative to the probability of each word occurring individually. This provides a first order word co-occurrence metric, and has demonstrated remarkable effectiveness at explaining human semantic data without resorting to complex inference mechanisms.

Word-Concept Models

We developed a generalized technique to transform any semantic space model into a word-concept learning model. Word-concept models are divided into a learning phase and a prediction phase. In the learning phase, the model is applied to sentences composed of words. The generalized modification for a word-concept model is to simply concatenate the word tokens with the concept tokens into a single concept/label sensory episode. In the prediction phase, we attempt to assign an object token to each word token.

All semantic space models have the ability to determine similarity between any word pair, thus prediction can be as simple as finding the object with the maximum similarity to the given target word:

$$maxarg_i\left(sim\left(w_{targ}, o_i\right)\right) \tag{1}$$

While this performs reasonably, it is possible to improve on this technique by adding a second step. Building from Kintsch's (2001) *predication* algorithm, we first activate the neighbor set of N most similar words to our target word:

$$\mathrm{NSet}_{targ} = maxarg_i\left(sim\left(w_{targ}, w_i\right)\right) \tag{2}$$

Then for every object, we calculate its activation, Act_i, as the similarity between that object and every one of the top N word matches, weighted by the similarity of that word to the target word:

$$Act_i = \sum_{j \in NSet} sim(w_{targ}, w_j)^p * sim(o_j, w_j) \tag{3}$$

The mechanism provides a match not only to the target word but also to the target's region of semantic space. This is particularly important because there are always more words than concepts. Using Kintsch's (2001) predication allows non-nouns to influence the outcome of the similarity measurement through their similarity to the nouns. Thus, if the word "red" is strongly associated with the word "apple" in the discussion and "red" is also associated with the concept <apple>, then "red" can be used to discover the underlying link between "apple" and <apple>. This mapping can be done implicitly, without knowledge of the part of speech as long as the target words that are to be matched to objects are known.

Experiment

Each of the above models was trained on the CHILDES corpus and the results were compared to the gold standard model in Frank et al. (2007), as well as to a baseline model that simply counts which words and objects co-occur. There are many different ways to evaluate model performance, and

there does not seem to be agreement in the field about the correct measure to use. To remain comparable to Frank et al (2007), we examined the best F-score (the harmonic mean of precision and recall) achieved by each model. We also looked at the overall proportion of pairings matching the golden standard if all 37 words are assigned a concept meaning. We do this in order to determine what mappings the system makes if forced, although we note that this random slice of parent-child interaction may not be enough to disambiguate all of the mappings.

We also explore hybrid models, asking which two models contribute the most complementary (non-redundant) information. This exercise may hint at what combinations of mechanisms are most important for learning in the natural language environment.

Results

One popular way of visualizing the results of word-concept learning is through a confusion matrix as shown in Figure 1. The confusion matrix shows the similarity between word-concept pairings as gradients from black to white (with lighter being a higher association) filling each grid cells. According to the gold standard, each word is associated with exactly one object (except for "bird" which can refer to <duck> or <bird>). The cells outlined in red indicate the correct word-object pairings according to the gold standard.

In Figure 1a and 1b, a winner-takes-all filter has been applied for each word. Thus, the object that has the highest association has been assigned the similarity of 1, and all others have been assigned a similarity of 0. The values returned by the semantic space models cannot be directly interpreted as probabilities for pairing selections. Hence, only relative similarity measures are used here. In Figure 1c, we display the gradient similarity ratings after having been scaled to a power of five (thus exaggerating the differences between predictions).

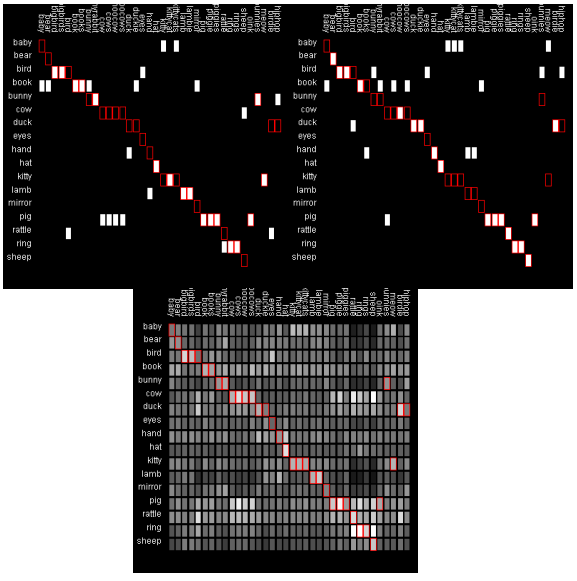

Figure 1: Confusion matrix results.

	BEAGLE	ESA	FDTRI	HAL	ISA	LSA	Kachergis	CoOc	
baby			■						0.07
bear	■	■		■			■	■	0.33
bigbird							■	■	0.13
bigbirds							■	■	0.13
bird									0.00
book							■	■	0.13
books									0.00
bunny	■			■					0.13
bunnyrabbit	■	■		■					0.20
cow	■	■	■	■					0.27
cows	■	■	■	■		■			0.33
moocow	■	■	■	■			■	■	0.47
moocows		■	■		■				0.20
duck	■	■		■		■		■	0.33
duckie		■						■	0.20
eyes		■				■			0.13
hand		■				■		■	0.20
hat		■		■	■	■			0.33
kitty	■				■				0.13
kittycat	■				■				0.13
kittycats	■		■						0.13
lamb	■	■		■	■				0.27
lambie	■	■		■	■				0.27
mirror	■	■					■		0.20
pig	■						■	■	0.20
piggie							■	■	0.13
piggies		■					■	■	0.20
rattle							■	■	0.13
ring							■	■	0.13
rings		■				■		■	0.20
sheep							■	■	0.13
oink		■	■	■			■	■	0.27
bunnies	■		■	■	■	■			0.33
meow	■		■						0.20
birdie	■	■		■			■	■	0.40
bird	■	■			■		■	■	0.33
hiphop	■		■	■		■			0.27
	0.51	0.51	0.27	0.38	0.24	0.27	0.41	0.51	

Figure 2: Word space. For each model by word, a black square indicates if model correctly identified that word.

It is important to understand which word-object pairs each individual model gets correct or incorrect. It is also important to see which word/object pairs are overall more or less likely to be found by the model. Figure 2 shows for each model, and each word, the probability of accurately identifying the target word. Black squares indicate word-by-model pairs in which model correctly identifies the object associated with that word. There is also a calculated average correctness for each word (rightmost column) and for each model (bottom row). Some of the semantic space models are non-deterministic (BEAGLE, FDTRI and ISA). For these models, 100 runs were computed and a correct identification granted when more than half of the runs correctly identified that pairing. No partial credit was given in any form for coming close to correctly matching each word. Parameter values were selected to be those optimized (relative to expected overall matches) for the individual model.

The F-scores are important indicators to help understand how well each model has correctly inferred the word-object pairings. To calculate the F-scores for each model, we first computed a similarity matrix for every word-object pair. For non-deterministic models, the similarities were averaged across 100 similarity runs for the given model. Next, maximum values were determined within each word to select an object. Pairings that were correct were labeled true, and pairings labeled that were incorrect, were labeled false. These similarity measures were then ordered based on strength (both correct and incorrect measures). For each N, precision and recall figures were then calculated for each of the top N word pairings.

This results in the receiver operator curves shown in Figure 3, and Table 1 shows the maximum F score values taken from the ROC curve chart for the top 5 models.

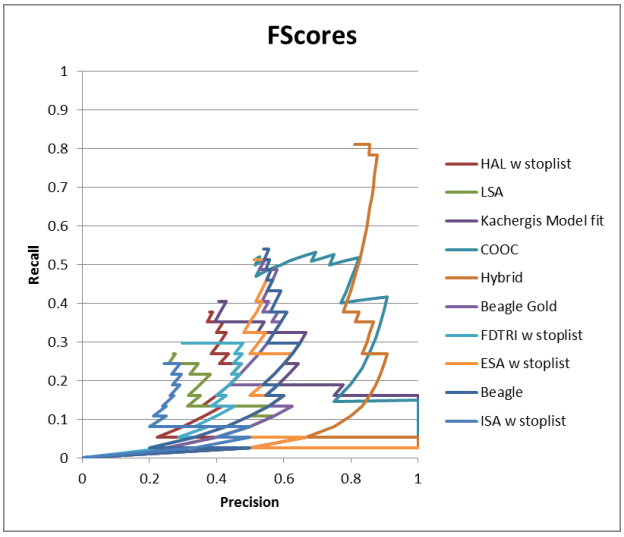

Figure 3: ROC curve for all models.

Table 1. Top F Scores

Model	Hybrid	BEAGLE	Frank	ESA	COOC
Best F	.83	.55	.54	.54	.53

Using the Word 2 Word language visualization tool (Kievit-Kylar & Jones, 2012) we can visualize all of the word/object pairings as a graph. In Figure 4, words are nodes and each similarity measure is an edge. In the visualization below, we see all concepts lined up across the top with each word referring to them shown below. The green connections indicate the strongest similarities observed by the system. Ideally, all lines would link to the word directly above.

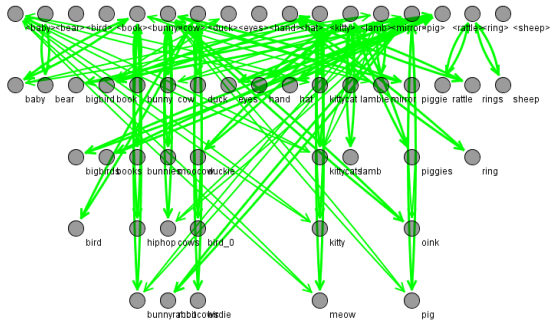

Figure 4: Word network visualization of BEAGLE model solution.

Because different models seem to make different mistakes, we also explored how hybrid models might be able to exploit these differences. Each model M is able to assign some similarity measure to and word-object pair $M_{sim}(W_x,O_y)$. We considered hybrid models of the following form: $M_{A,Bsim}(W_x,O_y) = M_{Asim}(W_x,O_y) * c + M_{Bsim}(W_x,O_y)$. Each pair of models was optimized, relative to the average number of matches with the golden standard, for the constant c.

Figure 5 shows the correct number of matches for each optimum pairing of models. A heat map of colors has been added to indicate highest (red) to lowest (purple) values. The optimum model is a co-occurrence by BEAGLE hybrid with the later having a weight three times greater than the former. This hybrid model results in 30 correct mappings on the gold standard.

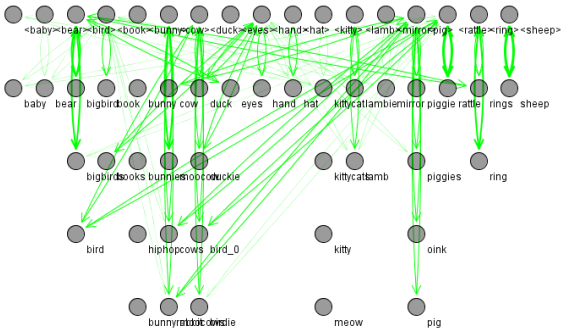

Figure 6: Word network visualization of optimum hybrid.

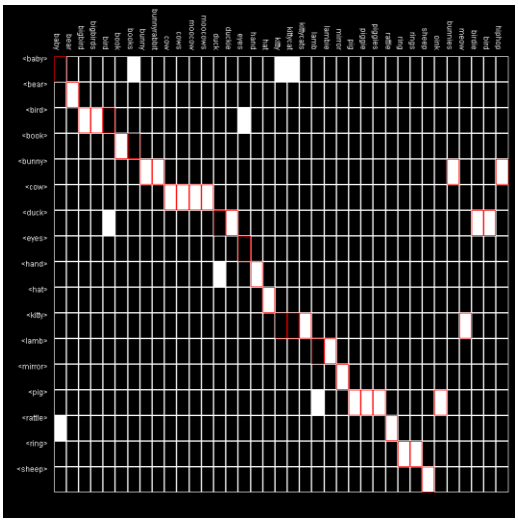

Figure 7: Thresholded confusion matrix for best hybrid.

	beagleE	beagleF	esaE	esaF	fdtriE	fdtriF	halE	halF	isaE	isaF	lsaE	lsaF	Kachergis	cooc
beagleGold	24	27	21	19	12	19	19	15	8	19	10	19	20	30
beagleE		21	18	19	12	11	14	20	8	10	10	21	17	24
beagleF			18	19	12	18	14	23	8	10	10	23	22	22
esaE				19	20	18	20	19	17	18	21	20	23	21
esaF					22	19	22	22	19	19	20	19	24	21
fdtriE						15	15	15	12	16	13	12	19	29
fdtriF							15	19	10	15	15	11	15	24
halE								16	13	15	10	14	16	28
halF									14	15	11	14	15	28
isaE										12	11	8	15	18
isaF											11	11	11	27
lsaE												10	17	23
lsaF													15	22
Kachergis														20

Figure 5: Hybrid Pairings.

Conclusions

The semantic space approach to word-concept learning is a fruitful endeavor with potential to better understand how humans make use of mechanisms and mutually reinforcing information sources across learning. The best pure semantic space models was able to predict the gold standard to a higher degree of accuracy than existing models while still conforming to known semantic and processing constraints. The adaptation of Kintsch's (2001) mechanism for predication allows semantic models to consider not only the semantic similarity between a word and object, but to also consider mutual information from the semantic neighborhoods. This procedure provided a benefit to each of the semantic space models tested.

Hybrid models also provide interesting insight into the word/concept-learning problem. The optimum hybrid model merged the co-occurrence model with BEAGLE. This optimal fusion makes intuitive sense, as the co-occurrence model provides first-order co-occurrence information that can be best supplemented by the higher-order co-occurrence information inherent in the semantic space models. The performance of the hybrid model suggests that infants may be capitalizing on both raw co-occurrence information and an emerging ability for higher-order semantic abstraction. Knowledge of which words are similar to each other from linguistic experience may be used to bootstrap word-object mappings across learning.

Acknowledgments

This research was supported by grants from Google Research and NSF BCS-1056744 to MNJ.

References

Baroni, M., Lenci, A., and Onnis, L. (2007). Isa meets lara: A fully incremental word space model for cognitively plausible simulations of semantic learning. *In Proceedings of the 45th Meeting of the Association for Computational Linguistics.* 49–56.

Church, K.W., Hanks, P. (1989). Word Association Norms, Mutual Information and Lexicography. *In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistics,* 76-83.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational wordlearning. *Advances in neural information processing systems*, 20, 1212–1222.

Gabrilovich, E. and S. Markovitch. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of IJCAI*, 1606-1611.

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

Jurgens D., Stevens K. (2009). Event detection in blogs using temporal random indexing. *In Proceedings of the Workshop on Events in Emerging Text Types*, 9-16.

Kachergis, G. (2012). Learning Nouns with Domain-General Associative Learning Mechanisms. *Proceedings of the 34th Annual Conference of the Cognitive Science Society* 533-538.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An Associative Model of Adaptive Inference for Learning Word-Referent Mappings. *Psychonomic Bulletin & Review*, *19*(2), 317-324.

Kievit-Kylar, B., & Jones, M. N. (2012). Visualizing multiple word similarity measures. *Behavior Research Methods, 44, 656-674.*

Kintsch, W. (2001). Predication. *Cognitive Science, 25,* 173-202.

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.

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

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Lawrence Erlbaum.