The Influence of Contextual Variability on Word Learning

Brendan T. Johns[1], Melody Dye[2], & Michael N. Jones[2]
bj23@queensu.ca, meldye@indiana.edu, jonesmn@indiana.edu
[1]Department of Psychology, Queen's University, Kingston, On.
[2]Department of Psychological and Brain Sciences, Indiana University

Abstract

In a series of analyses over mega datasets, Jones, Johns, & Recchia (2012) and Johns et al. (2012) found that a measure of semantic distinctiveness (SD), which takes into account the semantic variability of a word's contexts, provides a better fit to both visual and spoken word data than traditional measures, such as word frequency or raw context counts. The present study offers strong empirical support for this account's extensibility to natural language. In a self-paced reading experiment, subjects were incidentally exposed to novel words as they rated short selections from articles, books, and newspapers. When novel words were encountered across distinct discourse contexts, subjects were both faster and more accurate at recognizing them than when they are seen in redundant contexts. However, learning across redundant contexts promoted the development of more stable semantic representations. These findings are predicted by a model of SD trained on the same materials as our subjects.

Keywords: Word learning; Contextual Diversity; Semantics

Introduction

Traditionally, many accounts of language processing have proposed that word frequency is the primary type of environmental information used to organize the lexicon, based on consistent findings that high frequency words are easier to process than low frequency words (Broadbent, 1967; Forster & Chambers, 1973). Recently, Adelman, Brown, & Quesada (2006) demonstrated that a measure that builds a word's strength in memory by counting the number of contexts that a word occurs in (operationalized as the number of documents that a word occurs in across a corpus) provides a superior fit to processing times of words.

However, a pure context count as proposed by Adelman, et al. (2006) ignores an important information type: the semantic composition of different contexts. Linguistic contexts are not uniform, but instead are variable in terms of their semantic structure. There have been a number of past results that have suggested that diversity of contexts is important in language processing, such as the importance of contextual availability on word recognition performance (Schwanenflugel & Shoben, 1983), the role of distributional information on a number of lexical variables (McDonald & Shillcock, 2001), age of acquisition effects (Hills, Maouene, Smith, & Riordan, 2010), frequency effects in aphasia (Hoffman, Rogers, & Lambon Ralph, 2011), and the role of context in memory (e.g. Steyvers & Malmberg, 2003; Lohnas, Polyn, & Kahana, 2011; Nelson & Shiffrin, 2013).

To more closely examine the role that contextual variability plays in language learning, Jones, Johns, & Recchia (2012) conducted an artificial language learning experiment that manipulated word frequency and contextual diversity, such that certain words always occurred with different sets of words (high contextual diversity), while others repeatedly occurred with the same set (low contextual diversity). While there was no effect of diversity for low-frequency words, high frequency words were retrieved more quickly when they had been learned across distinct, high diversity contexts.

On the basis of these results, and a corpus analysis, Jones, et al. (2012) proposed a new model of lexical organization, the semantic distinctiveness memory (SDM) model. The SDM builds a word's strength in memory by weighting a new context by how much unique information that context provides about the meaning of the word. In other words, contexts that are redundant with past experience are not encoded as strongly as contexts that provide unique information. Across various corpora, this model was able to account for a larger amount of variance to a mega dataset of lexical decision and naming times (obtained from the English Lexicon Project; Balota, et al., 2002) over word frequency and a context count. Additionally, Johns, Gruenenfelder, Pisoni, & Jones (2012) demonstrated that this advantage for a semantic diversity count extends to spoken word recognition performance, suggesting that contextual variability is a general property of linguistic organization. Similar proposals have been made by others, such as the discriminative learning perspective (Baayen, 2011).

The goal of the current paper is to demonstrate conclusively that contextual variability is an important variable used in word learning. To do this, a unique paradigm that mixes natural language comprehension with an artificial language learning task will be used. Before describing this experiment, a new version of the SDM model will be described. This model will be used to predict the pattern of results that should be seen in the word learning experiment.

The Semantic Distinctiveness Model

In the original semantic distinctiveness model (SDM), a given word's strength in memory is represented in a Word x Document matrix, which can be trained over any large corpus of documents. For each new document that is encountered, a new column is added to the matrix. If a word occurred in that document, then it is assigned a semantic distinctiveness (SD) value in that column, which signals how redundant the new context is, compared to past experience (for specifics about how this is calculated, see Jones, et al., 2012). A word's strength in memory (corresponding to how easily a word is to retrieve from memory) is then just the summed semantic distinctiveness values across its row in the matrix.

While this implementation of the model has proven successful, it is also difficult to scale due to the computational

resources required by an ever-expanding matrix. A more tractable approach is to use a set vector size, similar to that employed in vector accumulation models, such as the BEAGLE model of semantics (Jones & Mewhort, 2007). The core assumption behind BEAGLE is that a word consists of an environmental vector, which is static, and a context vector, which is dynamic, and which changes as a function of the co-occurrence statistics of the linguistic environment. Here, we employ a new vector-based version of the SDM, the vSDM model, which is based on the same mechanisms as the original model, but is less computationally expensive. Since vSDM does not involve an exponentially increasing amount of computation as more documents are processed, it can be run over much larger corpora.

The fundamental operation of the SDM is that it utilizes a expectancy-congruency mechanism to build a word's semantic representation: The encoding strength for a word in a given context is relative to the information overlap between the context and the memorial representation of the word. This mechanism is very similar in principle to models that adjust attention across learning to dimensions that are more diagnostic. The vSDM will have underlying differences in terms of implementation, but this mechanism is the basic assumption about the importance of contextual variability.

Implementation of vSDM

In the vSDM model, the environmental vectors used to represent words are sparse ternary vectors[1]. From these, a context representation of each document can be constructed by summing the environmental vectors of the words that occurred within that document, instead of adding a new column into a matrix, as was done with the original SDM model. The following equation captures this summation process:

$$Context = \sum_{i=1}^{n} env_i \qquad (1)$$

Where n represents the number of words in a document. This context vector allows us to assess the similarity between the current document context and a word's semantic representation. As in the original SDM model, we use the vector cosine (a normalized dot product) as our similarity metric, and subject it to an exponential transformation, so as to properly weight redundant contexts over distinctive ones. Semantic distinctiveness for a given word is thus computed as a function of:

$$SD_i = e^{-\lambda * \cos(sem_i, context)} \qquad (2)$$

Where i is the current word being processed, sem is the semantic representation for that word, and λ is a scaling parameter that determines how much to weight the differences between high and low variability contexts. Note that a key difference arises here between the SDM and vSDM implementations: Whereas in the SDM, a word's semantic distinctiveness value was contained within the word's entry in the Word x Document matrix, in the vSDM, that value is registered with an external counter, which tabulates the accumulated SD values from across the word's document set. After the semantic distinctiveness values for a given document have been calculated, all of the words that occurred in that document then have their semantic representation updated by adding that context vector to their semantic representation. As in the original SDM model, a word's semantic representation will be updated according to how unique the contextual information is. This is accomplished by multiplying the context vector by the word's semantic distinctiveness value, as follows:

$$sem_i = sem_i + (context * SD_i) \quad i = 1, \ldots, n \qquad (3)$$

Where i goes through each word in the document.

Comparison to other measures

To demonstrate that the semantic distinctiveness (SD) measure derived from vSDM provides an advantage over word frequency (WF) and contextual diversity (CD) measures, all three of these values were calculated for each unique word type in a 200,000 document Wikipedia corpus. (By comparison, the largest corpus that the original SDM model could be trained on was a 40,000 document corpus). Table 1 displays the results of a regression for these three variables over lexical decision and naming times obtained from the English lexicon project (Balota, et al., 2002). In line with previous results, the SD variable accounts for the greatest amount of unique variance, while the effects of WF and CD are much reduced. This indicates that the vSDM is a valid measure of semantic diversity, and is also capable of scaling up to more realistically sized corpora.

Table 1.
Unique variance predicted by WF, CD, and SD for LDT and NT

Data	Effect (ΔR^2 in %)		
	SD	CD	WF
LDT	8.33***	1.19***	0.12***
NT	12.34***	1.85***	0.0

Note. All variables were natural log transformed. Analysis done over 36,361 data points. ***p<0.001

However, the central assumption of this model – that contextual variability is an information source that humans attend to in natural language processing – has yet to be empirically demonstrated with natural language materials. The following experiment examined whether the vSDM model would accurately predict how discourse variability influences word learning from context.

Word Learning Experiment

To assess the role contextual variability plays in lexical learning, we asked subjects to read and rate short passages

[1] A vector size of 2,000, with 4 non-zero values sampled equally from {1, -1}, was used.

that contained novel words. Critically, the discourse contexts that each novel word occurred in had been manipulated, such that some words were encountered across highly distinctive contexts, while others were encountered across very similar contexts. Following exposure at reading, subjects completed a pseudolexical decision and semantic similarity judgment task, which offered insight into how the representation of each word had developed as a function of the discourse contexts in which it had been encountered.

Prior research in artificial language tasks suggests that contextual variability confers a distinct processing advantage (e.g., Jones et al. 2012). Our study sought to replicate and extend these findings to natural language, and to test the extensibility of the vector-based semantic distinctiveness (vSDM) model. When trained on the same material as our subjects, vSDM predicted that whereas diverse contexts should strengthen memory for novel words, leading to faster and more accurate recognition judgements, uniform contexts should support the development of a more stable semantic representation.

Method

Participants. Ninety-one undergraduate students at Indiana University participated in the experiment for $10 pay. All were native American English speakers with normal or corrected-to-normal vision. The data from four subjects was discarded: two because they did not complete the experiment, and three because their performance fell below chance on the pseudolexical decision task.

Materials. Our interest was to gauge how participants' representations of novel words developed over reading. To this end, we prepared study materials in which a real target word (e.g., panjandrum) had been replaced with a novel word form (e.g., sattery). Drawing from real words had two advantages. First, it meant that training passages could be taken from natural, real world contexts in which those words occurred. Second, it provided a benchmark semantic representation against which subject judgments could be assessed. However, actually employing real words in training would have made it difficult to separate out learning at study from prior learning. To minimize the effects of pre-experimental exposure, we selected low frequency targets, and replaced each target with a novel word form.

Ten target words were selected for study. All were low frequency (>5 instances per million) and attested in a variety of discourse contexts (e.g., constellation is used in relation to stars, symptoms and freckles). For each target, two distinct sets of reading materials were developed: one set comprising five passages from a single discourse topic (low variability), the other comprising five passages spanning a number of distinct topics (high variability). Passages were short excerpts selected from fiction and non-fiction books, academic journals, and reputable news sources that contained exactly one instance of the target word. To ensure that the length and semantic distinctiveness of each set was kept constant across targets, we measured the length of each passage, and the word overlap between passages. There was no statistical difference in passage length between any paragraph sets, but for all paragraph sets the amount of word overlap was controlled, such that it was significantly larger for high variability paragraphs[2].

During the experiment, each target was randomly replaced with a pronounceable nonword, drawn from a list of twenty nonces. These novel word forms were selected from the English lexicon project (Balota et al., 2002), and matched on number of letters, orthographic neighborhood size, bigram count, and reaction time and accuracy in pseudolexical decision.

Procedure. Participants were told that they were assessing standardized testing materials for clarity and comphrensibility. During the study phase of the experiment, each passage was displayed on screen for a minimum of 10 seconds, after which a rating scale appeared. Subjects were instructed to make a rating on a scale of 1 to 7 assessing how well they understood the passage, with 1 indicating that they understood it perfectly, and 7 indicating that they did not understand it at all. In order to simplify visualization of this data, the scales of this data were flipped in the following graphs, in order to make them more comprehensible. They had as much time as they liked to read and respond. After the subject's rating had been submitted, the passage and scale disappeared, and the program advanced to the next training trial.

The study was designed such that each target word had both a uniform (low variability) and a diverse (high variability) set of passages associated with it, each of which comprised 5 short paragraphs. At the beginning of study, the program randomly assigned half of the targets to the uniform condition, and half to the diverse condition, and loaded the corresponding paragraph sets. Each target was then randomly assigned a pseudoword, which replaced the target across all the passages in which it occurred. Subjects read a total of 50 paragraphs (10 targets x 5 paragraphs), with the order of presentation randomized.

After training, subjects completed a pseudolexical decision task (PLDT). For each word presented at test, subjects were asked to determine whether they had seen that word at study, responding as quickly as accurately as possible. Each word was preceded by a fixation cross that lasted 1 second, after which the subject had been instructed to press '1' if the word had been seen in reading, and '0' if it had not. Both accuracy and reaction time were recorded. Following the design of Jones et al. (2012), each of the 10 studied pseudowords was presented 5 times. The 10 remaining unstudied pseudowords from the original set were also presented 5 times each, as filler items, for a total of 100 trials. Unstudied and studied items were randomly intermixed, and no word was repeated sequentially.

Following the PLDT, participants completed a semantic similarity judgment task. A pair of words was presented on

[2] See http://www.indiana.edu/~clcl/cont_var.pdf for paragraph sets, along with corresponding comparisons.

screen, and subjects were asked to rate how similar the pair was in meaning on a scale from 1 to 7, with 1 being the most similar and 7 being the least similar. Pairs consisted in a studied item and a close associate of the item's target meaning. Each of the 10 studied items was paired with 4 close associates, yielding a total of 40 semantic similarity ratings.

Model Predictions

To establish what the vSDM would predict in these tasks, we trained the model on the same materials as our subjects. For the PLDT, we compared the semantic distinctiveness score for each item following training over uniform passages, against that item's score following distinctive passages. A higher SD value from this model signals a higher strength in memory. As the top panel of Figure 1 illustrates, the model predicts that items learned over diverse contexts should be represented more strongly in memory. Behaviorally, this suggests that subjects should be faster and more accurate at recognizing these items, as compared to those learned across uniform contexts. However, whether or not differences in reaction time will emerge empirically is unclear. In the artificial grammar learning experiment conducted by Jones et al. (2012), response time differences were only observed for "high frequency" words, which had been observed across many contexts. Given that the items in this experiment were only presented 5 times (which is fewer than the low frequency words in Jones, et al.), it is unclear whether a robust speed difference should emerge. However, the use of natural language materials may increase the power of this manipulation.

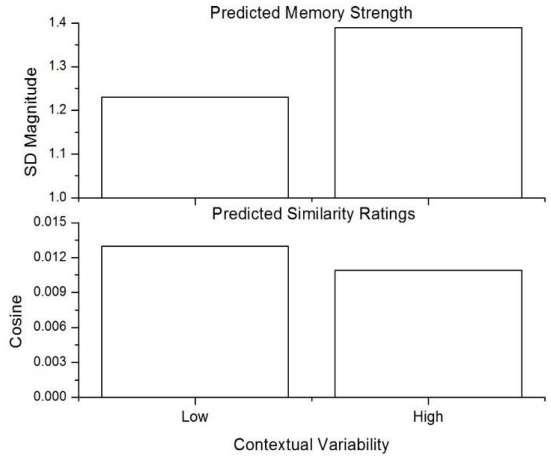

Figure 1. The predictions of the vector-based semantic distinctiveness model (vSDM) after training over the same materials as our subjects.

To determine what the model would predict for the semantic similarity rating task, we calculated the predicted similarity between the pseudowords and its target associates, and compared them across conditions. Representations of the associate words were attained by training the model on the 200k document Wikipedia Corpus, with the λ parameter set to 5. Representations for pseudowords were constructed from either the uniform or diverse paragraphs, using the same environmental vectors as the corpus model. Similarity between the pseudowords and the close associates was computed with a vector cosine. The model's predictions are displayed in the bottom panel of Figure 1. As can be seen, the model predicts that items trained in uniform contexts should actually be more similar to their target associates than items trained in diverse contexts. This is likely because in the low variability condition, the high word overlap among the paragraphs contributes to a more stable semantic representation. A model based on frequency or a context count would predict that these should be equal, at least for memory strength, hence this task is diagnostic in separating different proposed models.

Experiment results

During the study phase of the experiment, subjects supplied comprehension ratings for each of the passages. A 2 (paragraph condition) x 5 (trial number) repeated measures ANOVA revealed a significant effect of paragraph diversity [$F(1,86)=110.26$, $p<0.001$], a small effect of trial number [$F(1,86)=2.377$, $p=0.05$], and a significant interaction [$F(4,344)=4.565$, $p=0.001$]. Figure 2 shows the average comprehension ratings for the low variability and high variability sets, across the five passages. As can be seen, for the first passage, the ratings for the low and high variability passages are equivalent. However, for subsequent passages, the ratings for high variability passages are systematically lower, meaning that they were rated as less comprehensible. Notably, for the within condition the ratings increased, indicating that participant's subjective comprehension ratings to paragraphs within the same discourse topics were increasing, due to the subject receiving more experience with the specific topic under question. However, in the low variability condition, the ratings are relatively stable, suggesting very little overlap in meaning of the different paragraphs across reading. This is an expected result, based on how the different paragraph sets were constructed. How this influences word learning will be examined next.

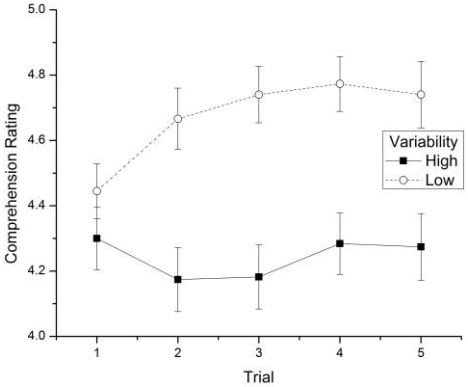

Figure 2. Comprehension ratings awarded to low and high variability passages across trials. Higher ratings indicate that the paragraphs were easier to comprehend.

The vSDM predicted that words that occur in more diverse semantic contexts should have a stronger representation in memory, which means that they should be easier to discriminate and faster to respond to. This prediction was tested with a PLDT task. In the PLDT, average accuracy was 83.9% across conditions (5 subjects were below chance and were excluded from this analysis). The left panel of Figure 3 depicts recognition accuracy for low and high variability paragraphs across all participants. As predicted, subjects were significantly more accurate at recognizing targets seen across highly variable contexts [$t(86)=3.561$, $p<0.001$]. Variability also appears to support more rapid responding. As illustrated in the right panel of Figure 4, subjects were, on average, 25.8 ms faster at identifying words that appeared in high variability paragraphs (RTs were medians), an effect that was significant [$t(86)=2.297$, $p<0.05$]. This effect is not as large as has been previously found, but the trend is in a similar direction to what has been formerly reported. Taken together, these results accord well with the vSDM's prediction that items learned over diverse contexts should be more strongly represented in memory. This also matches the artificial language results and simulation work of Jones, et al. (2012) and Johns, et al. (2012). This demonstrates that subjects are able to more accurately discriminate and also respond faster to words that occurred faster in high variability conditions.

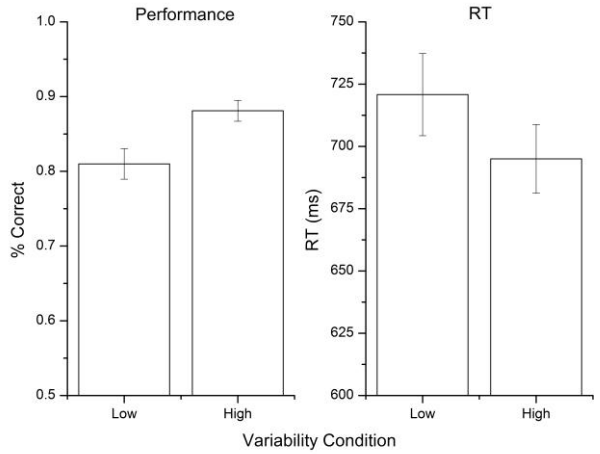

Figure 3. Performance and RT results from the PLDT task.

After completing the PLDT, subjects were asked to rate the semantic similarity between each pseudoword and four close associates of its target meaning. The vSDM predicted that items learned in uniform contexts should be rated as more similar to target associates than items seen across diverse contexts, as redundancy should support the development of a more stable semantic representation. Figure 4 plots the average word pair similarity ratings as a function of training condition, with a lower rating signaling a higher association value. As predicted, subjects rated items trained on the low variability paragraphs as more similar to their target associates, an effect which was significant [$t(86)=3.406$, $p=0.001$]. This demonstrates that unlike the results of the PLDT experiment, subjects are better able to learn the meanings of words that occur in low diversity contexts, as opposed to high diversity contexts.

These results reveal an interesting dissociation between ease of processing and semantic representation. Subjects in our experiment appear to be more efficient at processing items trained over diverse contexts, recognizing those items more quickly and more accurately. At the same time, however, subjects appear to have better discriminated the meanings of items trained in redundant contexts, a finding supported both by their subjective comprehension ratings of the passages, and by their increased similarity ratings in the semantic judgment task. This may be because uniform contexts provide more consistent cues that are necessary to disambiguate the meaning of a word. However, it is unclear whether this would hold across a greater number of presentations. It may be the case, for example, that more exposures are required to form a well-discriminated representation of a contextually 'promiscuous' word. In any case, it suggests that the composition of context can have differing effects on the linguistic system.

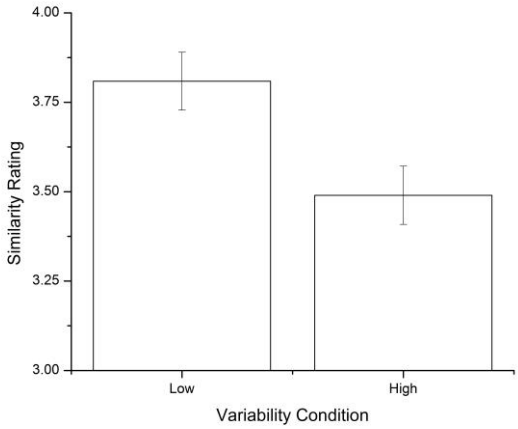

Figure 4. Mean similarity ratings of studied items and target associates by passage training type.

General Discussion

The experiment described here attempted to empirically demonstrate that humans use contextual variability to learn and organize language, based on the suggestions of the SDM model (Jones, et al., 2012; Johns, et al., 2012). To accomplish this, two sets of 5 paragraphs were assembled for 10 low frequency words. One paragraph set was sampled from one discourse topic (low variability) or from five different topics (high variability). The low frequency words were then replaced with pseudowords in order to assess how natural language contexts impact the ability to learn new words.

The vSDM model described in part 1 predicted that subjects should learn words in the high variability condition more efficiently, and perhaps are faster at identifying them (although due to the relatively small number of presentations, whether this effect would emerge was unclear). The model also predicted that semantic similarity ratings to associates of

the replaced words should have an opposite pattern: the low variability paragraphs should lead to slightly higher similarity ratings, due to these paragraphs allowing for a more stable semantic representation to be formed. Both of these predictions were borne out by the experiment. Subjects were significantly better at learning words when they appeared in the high variability condition. Words were also identified faster in this condition. Additionally, it was found that subjects do rate words that were learned in low diversity passages as more similar to semantic associates. These results indicate that contextual variability is an important information source used in language, which can have differential effects depending on the task that is used.

Although the SDM model is capable of efficiently measuring a word's contextual variability, and make corresponding predictions about the effect that this should have on behavior, it is not a conceptual model. However, one clear possibility for the importance of contextual variability in organizing the lexicon comes from the recent understanding of the importance of the use of prediction in language processing (e.g. Altmann & Mirkovic, 2009; Elman, 2009). This type of explanation would propose that people construct expectations about future linguistic input, based on the current context. In terms of contextual variability, words that are low in contextual variability are better able to be predicted with contextual cues, and so should be weighted lower in memory since an expectation-generation process would be able to predict the word's occurrence. However, words that are high in contextual variability are not able to be accurately predicted by contextual cues since they occur in such diverse linguistic environments. These words should be relatively stronger in the lexicon, since the expectation process is not as capable of predicting when these words should occur. Thus, word retrieval should be seen as a more dynamic process, where both past experience with words, and also the current context, combine to drive the retrieval process.

More research needs to be conducted to validate this dynamic view of word retrieval, but the current study, as well as many previous studies, demonstrates that contextual variability is an important aspect in language. At the very least, this research demonstrates that each occurrence of a word is not weighted equally, and hence many models of word retrieval that are dependent on static information sources (e.g. word frequency) to organize lexical retrieval need to be modified in order to account for these effects.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision time. *Psychological Science*, 17, 814-823.

Altmann, G. T. M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33, 583-609.

Baayen, R. H. (2010). Demythologizing the word frequency effect. *The Mental Lexicon*, 5, 436–461.

Balota, D. A., Cortese, M. J., Hutchinson, K. A., Neely, J. H., Nelson, D., Simpson, G. B., & Treiman, R. (2002). The English lexicon project. *Behavior Research Methods*, 339, 445-459.

Broadbent, D. E. (1967) Word-frequency effect and response bias. *Psychological Review*, 74, 1-15.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33, 1-36.

Forster, K.I., & Chambers, S.M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627, 635.

Hills, T., Maouene, J., Riordan, B., & Smith, L. (2010). The associative structure of language and contextual diversity in early language acquisition. *Journal of Memory and Language*, 63, 259-273.

Hoffman, P., Rogers, T. T., Lambon Ralph, M. A. (2011). Semantic diversity accounts for the "missing" word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. *Journal of Cognitive Neuroscience*, 23, 2432-2446.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *Journal of the Acoustical Society of America*, 132, 74-80.

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

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66, 115-124.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language*, 64, 249-255.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44, 295-323.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, 120, 356-394.

Peirce, J.W. (2007) PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8-13.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 82-102.

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 760-766.