

Context Repetition Benefits are Dependent on Context Redundancy

Gabriel Recchia (grecchia@indiana.edu)

Cognitive Science Program, 1910 E 10th St.
Indiana University, Bloomington, Indiana USA

Brendan T. Johns (johns4@indiana.edu)

Department of Psychology, 1101 E. 10th Street
Indiana University, Bloomington, Indiana USA

Michael N. Jones (jonesmn@indiana.edu)

Department of Psychology, 1101 E. 10th Street
Indiana University, Bloomington, Indiana USA

Abstract

What sources of statistical information do humans leverage to organize the mental lexicon? Recently, accounts emphasizing the importance of word frequency have been challenged by accounts emphasizing contextual diversity (CD). The latter suggest that words will be processed faster if they occur in a greater number of unique semantic contexts. Previous corpus studies have operationalized CD by counting the number of documents that each word appears in. However, there is no guarantee that each document corresponds to a unique semantic context, as documents may address similar topics. We develop a measure of CD that takes into account the semantic similarity of documents, and show that for words appearing in many contexts, appearing in multiple contexts is most beneficial if the contexts are high in semantic distinctiveness. In Experiment 2, we induce a similar effect experimentally using an artificial-language paradigm, demonstrating that repetition of a word does not speed processing unless the repetition is accompanied by a change in context.

Keywords: contextual diversity, word frequency, lexical decision, learning models

Introduction

Classic strength accounts of learning are based on the belief that repetition of an item aids learning, increases memory strength for the item, and increases the probability that the item will be later retrieved. This principle of repetition has also been influential in theories of word identification based on the finding that word frequency is among the most important variables to predict lexical decision times (LDT). A higher-frequency word is likely to be identified as a word more quickly than is a low-frequency word (Forster & Chambers, 1973). This theoretical position entails that on each occurrence, a word's memory trace is strengthened, resulting in the trace being easier to retrieve on subsequent experiences of the word. The classic repetition/frequency principle has led to the development of serial-searched rank frequency models (Murray & Forster, 2004), threshold activation accounts (Coltheart, et al., 2001), and iterative-update connectionist models (Seidenberg & McClelland, 1989).

However, recent evidence has called into question the importance of frequency information in LDT. Adelman, Brown, and Quesada (2006) have demonstrated that word frequency (WF) is confounded with contextual diversity (CD)—the number of different contexts in which a word has

been used. In their analysis, Adelman et al. quantify a word's CD as simply a count of the number of different documents in which the word occurs across a large text corpus, and their results are consistent across many corpora and LDT datasets. Words that appear in more contexts are likely to have a higher frequency, but Adelman et al. argue that it is CD and not WF that is the causal factor of LDT. Similarly, Steyvers and Malmberg (2003) have demonstrated that models based on document count outperform models based on rank frequency in recognition memory tasks.

Contextual diversity is based on the rational approach to memory (Anderson & Milson, 1989; Anderson & Schooler, 1991), specifically the *principle of likely need* (PLN) which states that the more contexts a word occurs in across learning, the more likely the word will be needed in future contexts (Adelman et al., 2006). Psychological notions of context differ greatly, and range from the list in which a word was encoded, to changes in time, to the room in which learning took place (Dennis & Humphreys, 2001; Wickens, 1987). It is not directly obvious how counting documents corresponds to classic ideas of a change in context, but it seems fairly intuitive that if a word is repeated in the same document, we should not consider the semantic context to have changed.

PLN implies that a word will be processed faster if it occurs in a greater number of unique semantic contexts. Hence, operationalizing CD as the number of documents that a word appears in across a corpus may be an invalid measure of the true contextual diversity of the word. A frequent discourse topic is likely to have many documents dedicated to it, and so a word that is used to describe a frequent topic is likely to appear in more documents, even though the documents are not truly distinct contextual uses of the word. Thus, document count risks giving a high CD score to words that actually have very low semantic uniqueness. Ideally, we would like a more nuanced measure. Under PLN, one would expect that repetition of a word in distinct contexts would increase its likely need to a greater extent than an equal number of repetitions in redundant contexts.

At the heart of the debate is the question of whether repetition of a word benefits processing if context does not change. Under strength accounts, any repetition adds another trace to memory and, hence, increases the strength of the trace (or the availability of traces for the word).

Under a PLN account, by contrast, repetition should not benefit processing if context does not change because repeating an item in the same context does not increase its likely need. To contribute to this debate, we first derive a new measure of contextual diversity that is based on the semantic uniqueness of the set of documents in which a word occurs. In Experiment 1, we present a large corpus analysis using our new measure demonstrating that repetition of a word is most beneficial if the repetitions are in more distinct contexts, a finding consistent with Adelman et al.'s (2006) PLN, but not directly answerable with their document count measure of CD.

An additional problem with document count as a measure of CD is that it is confounded with many other variables in addition to WF, any of which could be the causal factor influencing LDT. For example, LDTs could be faster for words that have been experienced more recently; words with a greater document count or WF are also likely to have a higher recency. Ambiguity, abstractness, and imageability are also confounded with document count and WF, and are difficult to tease apart. Finally, it has been suggested (Balota; in Adelman et al. 2006) that document count from a text corpus may actually be a better measure of real-world WF due to the structure of the corpus. Recent corpus-based work attempts to demonstrate the unique effects of CD statistically by partialing out the confounding variables as covariates. However, we take a novel approach here, and in Experiment 2 attempt to induce the effect of CD experimentally with an artificial language learning task followed by a surprise pseudo-lexical decision task. To our knowledge, the effects of CD and WF have never been induced experimentally.

Experiment 1

We conducted a large corpus analysis to assess whether processing of a word is most likely facilitated by simple repetition of the word (WF), number of contexts in which it appears (document count; DC), or whether the benefit of context or frequency was dependent on the uniqueness of the context relative to the word's history of contextual occurrences. The variables computed from the corpus analysis were used to predict LDT on a per-word basis for a large number of observations.

Semantic Distinctiveness

To examine the influence of contextual uniqueness, it is necessary to create a measure of the dissimilarity of documents in which a word has appeared. Though there are many existing models of semantic representation (e.g., HAL or LSA), we did not want to approach the problem from a specific theoretical orientation. Instead the base measure that we use to assess the dissimilarity between two documents is the proportion of words that the documents have in common, or:

$$Dissim(doc_1, doc_2) = 1 - \frac{|doc_1 \cap doc_2|}{\min(|doc_1, doc_2|)} \quad (1)$$

That is, document similarity is the intersection of the two sets of words, divided by the size of the smaller document. This gives the proportion of word overlap between two documents. Function words (e.g. *the, is, of,* etc...) were filtered out of the set of words, so they do not impact the similarity rating. Document dissimilarity is then just 1-similarity. We then define a word's *semantic distinctiveness* (SD) as the mean dissimilarity of the set of documents that contain it:

$$SD_{word} = \frac{\sum_{i=1}^n \sum_{j=1}^i Dissim(doc_i, doc_j)}{n + (n^2 - n)/2}, \quad (2)$$

where n is the number of documents that a word appears in. Equation (2) is the mean dissimilarity between every document in which a word appears, and this SD value signals how distinct the documents that a word occurs in are from each other. A word with a high SD value tends to occur in documents that have a low amount of word overlap (it is more contextually distinct), and a word with a low SD value tends to occur in documents that have a high amount of word overlap (it is less distinct).

Method

To compare the SD measurement of contextual diversity with the straightforward document count (DC) of Adelman et al. (2006), an SD value was computed for each of 18,494 words for which we obtained LDTs from the English Lexicon Project database (Balota, et al., 2000). For each word, SD, DC, and WF values were computed using 9,100 documents from a corpus containing articles from the 1994 New York Times corpus. Documents in this corpus had a mean length of 250 words.

In order to compare SD and DC as measures of contextual diversity, the data were divided into two groups – a high or low SD group, and/or a high or low DC group. A word was placed into the high-SD group if its SD value was in the upper quartile of all the computed SD values, but into the low-SD group if its SD value was in the lower quartile of all the computed SD values. Likewise, a word was placed into the high-DC group if it was in the upper quartile of the DC data, but into the low-DC group if it was in the lower quartile of the DC data.

Results

In order to assess the impact of these two measures on each other, a univariate analysis of variance (ANOVA) was conducted. The means from this analysis are depicted in Figure 1. A main effect was observed for the DC factor, with higher DC words having faster LDTs than lower DC words, $F(1,4232) = 177.43, p < .001$. Further, a main effect was found for SD, with more semantically unique words having faster LDTs than less unique words, $F(1,4232) = 143.78, p < .001$. Substituting WF for DC produces the same result for all analyses reported (because WF and DC are so highly correlated), hence, we only report results from

DC. Of particular interest, however, is the finding that SD and DC interacted, $F(1,4232) = 74.26$, $p < .001$. As the document count increased, words that appeared in a greater number of semantically unique documents saw a greater benefit on their LDTs from the additional contextual occurrences.

It is possible that this interaction is a simple result of the fact that sufficient occurrences are necessary for semantic uniqueness to play its role. That is, appearing in more documents increases chance co-occurrences with less frequent (and hence more unique) words. On a continuous scale, the correlation between log-SD and LDT is $-.433$. When we partial out the effects of DC, the unique partial correlation between log-SD and LDT remains at $-.361$ (and drops to a lower, but still significant, $-.166$ by partialing out log-DC). Hence, even when the number of documents is controlled for, there is a clear relationship in which words that appear in more unique contexts are identified faster, consistent with PLN.

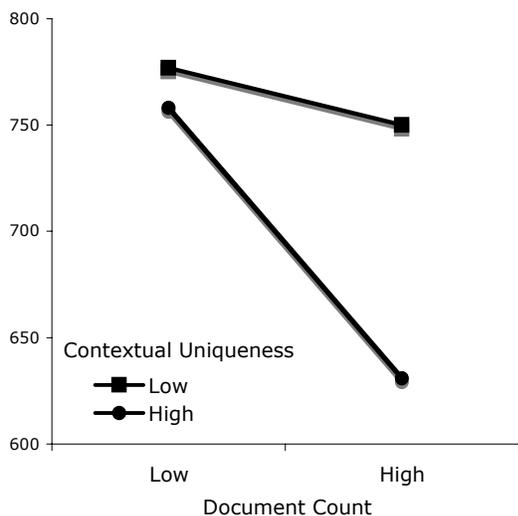


Figure 1. Lexical decision latency as a function of the factorial combination of number of contexts (abscissa) and the semantic uniqueness of those contexts (separate lines). The repetition of a context shows the greatest benefit to processing when the contexts repeated are more unique.

Discussion

This interaction suggests that words appearing in many contexts are processed more quickly if the contexts are highly unique, as would be expected by PLN. However, semantic distinctiveness causes the largest reduction in LDT only at high levels of contextual diversity. Thus, repetition of a word appears to be most beneficial when the context changes (cf. Verkoeijen, Rikers, & Schmidt, 2004). It also suggests that when attempting to measure contextual diversity, it is not enough to simply count the number of different documents that a word appears in. Instead, one

must account for the semantic uniqueness of each context in which the word appears, and this finding has important implications for learning models.

What accounts for the fact that SD has the greatest impact on LDT when a word appears in many documents? One possibility is that because highly frequent words also tend to appear in a large number of documents, the DC measure is somewhat confounded with word frequency, while the new measure of SD is less so. If this is the case, then the results of Experiment 1 suggest that appearing in diverse semantic contexts facilitates processing the most for highly frequent words, and less so for infrequent words. We investigate this hypothesis in Experiment 2 by manipulating CD and WF.

Experiment 2

Experiment 2 was designed to experimentally test the hypothesis that repetition of contextual occurrences produces greater processing gains for unique contexts than for redundant contexts, as well as to compare the effects of contextual diversity and word frequency on lexical decision times in a more controlled setting. The contextual diversity effects used to support PLN over rank-frequency models have never been induced experimentally, perhaps due to the fact that in natural languages, CD is highly confounded with many other sources of statistical information (McDonald & Shillcock, 2001). Thus, we used an artificial language paradigm to independently vary CD and WF, and to assess the relative contribution of each on response latency.

Method

Participants. Thirty-two undergraduate students at Indiana University participated in the experiment for partial course credit. Of these, eight were excluded from the final analysis due to low accuracy on the lexical decision task (85% or lower).

Materials. Participants were trained in an artificial “alien” language referred to as Xaelon. The Xaelon lexicon consisted of a set of twelve one-syllable pronounceable nonwords selected from the Elexicon database (Balota, et al., 2002), equated for number of phonemes, number of letters, and orthographic neighborhood size. A set of twelve foils, to serve as negative examples during the lexical decision task, was selected in the same way. The nonwords comprising the lexicon and the set of foils were selected so as to exhibit no significant differences in bigram count averages, bigram count sums (calculated by position as well as overall), or mean lexical decision latencies from the Elexicon database. To account for potential unforeseen differences in the processability of the lexicon compared with the foils, the set of nonwords that comprised the lexicon was swapped with the set of nonwords comprising the foils for each participant.

For each participant, a set of 450 training slides was created. Each training slide consisted of a three-word “sentence” in Xaelon placed just above an image of a scene

described by that sentence. Of the twelve words in the Xaelon lexicon, four were designated as *subject words*, four as *object words*, and four as *locatives*. Each subject word corresponded to a different unfamiliar image (“Fribbles”) retrieved from Tarr’s (2006) online repository of experimental stimuli, each object word corresponded to a different three-dimensional geometric shape, and each locative corresponded to a different position that the subject could be in relative to the object (above, below, to the left, or to the right of the object). Which words corresponded to which semantic designations were randomized for each participant.

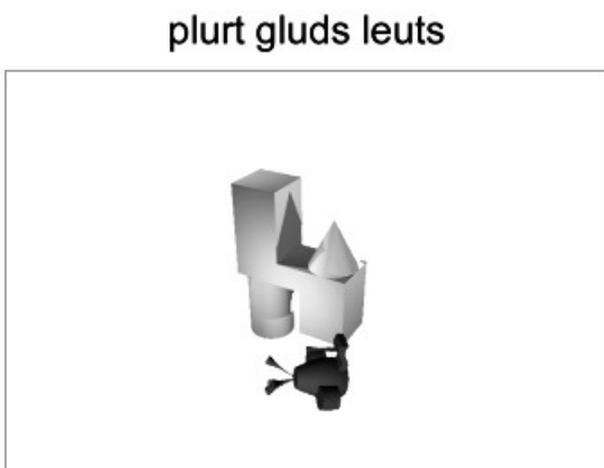


Figure 2. Example of a training slide seen by participants while learning the “alien” language Xaelon.

Finally, of the twelve Xaelon words, four were randomly selected and crossed for a factorial combination of two levels of WF (hi/low) and two levels of CD (hi/low). Low-WF words appeared 45 times each in the training slides, while high-WF words appeared 180 times each. Also, whenever a low-CD word appeared, it always appeared in the same semantic context (i.e., in the same sentence and with the same image), whereas each high-CD word appeared in eight different semantic contexts (i.e., it could appear in any one of eight different sentences, each juxtaposed with its corresponding image).

We selected nonwords and novel images so that participants could not simply translate the artificial language into English words for the subjects or objects; however, we did not attempt to create novel locatives.

Procedure. Participants were asked to imagine that they were explorers charged with the task of learning an alien language called Xaelon. They were also advised that they would see various scenes paired with Xaelon sentences that described them, and that they would later be tested on their knowledge of Xaelon. Participants next viewed 450 training

trials, divided into 10 blocks of 45 images each. Training slides appeared in random order, and each slide was displayed for four seconds, with a one-second intertrial interval.

Following the training trials, participants were confronted with a surprise pseudo-lexical decision task in which they were told that they would be presented with several stimuli, some of which were words from the language that they had just learned, and some which were not. They were asked to press one key if the stimulus was part of the language they had just learned, and another key if it was not. Instructions stressed both speed and accuracy. Participants then completed 288 test trials, divided into 18 blocks of 16 trials each. Each trial consisted of a fixation cross for 500 ms, a blank screen for 200 ms, and finally either a foil or Xaelon word, which remained on the screen until the participant pressed one of the response keys. Exactly 12 examples of each Xaelon word and 12 examples of each foil were presented to each participant during the lexical decision task.

Results and Discussion

Participants performed quite well at the pseudo-lexical decision task (PLDT), with a mean accuracy of .88 ($SE = .02$) across all target and foil trials. We set a stringent accuracy criterion of 85%, which trimmed eight participants. The mean accuracy of the above-threshold participants was .94 ($SE = .01$). Latencies greater than 2.5 standard deviations from a participant’s mean were removed; this resulted in 2.7% of latencies to be trimmed from the analysis. Surprisingly, response latencies did not differ as a function of part-of-speech (subject, locative, object), $F(2,28) = 0.11, ns$.

The reaction time distributions for means were positively skewed; hence, we report results here both from means and medians for correct responses only. Both means and medians show the same overall pattern of results. Figure 3 plots median response latency as a function of WF and CD. The data were analyzed using a repeated-measures ANOVA design. For medians, there was no significant main effects of either WF or CD, but there was a marginally significant WF*CD interaction effect, $F(1,24) = 3.02, p \sim .09$. Analysis on mean response latencies were very similar, revealing no significant main effects for either WF or CD, but a significant frequency-by-diversity interaction $F(1,24) = 4.37, p < .05$.

Post-hoc analyses (Bonferroni correction) revealed that the difference between the levels of CD at low WF was insignificant, $t(24) = -1.54, ns$, however, the difference between the levels of CD at high WF was statistically reliable, $t(24) = 2.11, p < .05$. Further, PLDT latency did not differ significantly between the levels of WF for low CD items, $t(24) = -1.67, ns$. However, the decrease in PLDT latency over WF for high CD was statistically significant, $t(24) = 2.06, p < .05$. Increasing the repetitions of the word from 45 to 180 produced no facilitation in PLDT if the context did not change, but large processing savings were

seen if the increase in frequency was accompanied by a change in context.

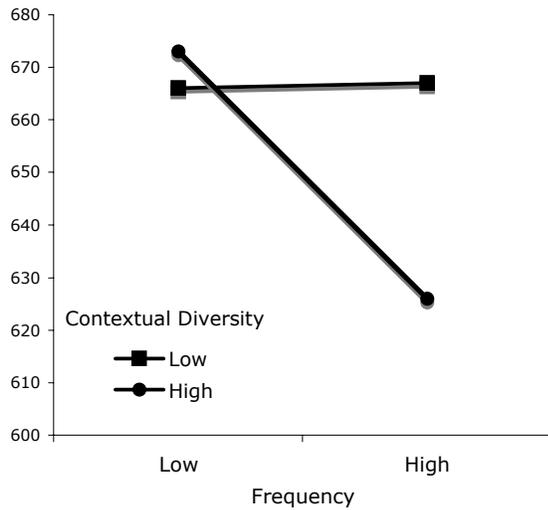


Figure 3. Pseudo-lexical-decision latency (PLDT) in the Xaelon task as a function of token repetition frequency and contextual diversity.

Consistent with PLN (and Experiment 1), we find that processing is facilitated for words appearing in a large number of contexts (High-WF) which are highly semantically distinct (High-CD). However, appearing in a large number of redundant contexts (High-WF, Low-CD) produced equivalent response latencies to a much lower number of repetitions in the redundant context (Low-WF, Low-CD).

This finding is a nice parallel to the results of our corpus analysis in Experiment 1, in which repetition of the word produced greater processing savings if the repetition was in a more semantically distinct context rather than if the repetition occurred in redundant contexts. If each sentence in Experiment 2 is considered to be a different “document,” the WF variable is equivalent to the DC variable in Experiment 1, since no word was ever repeated within a single Xaelon sentence. Likewise, the CD variable of Experiment 2 maps nicely on to the SD value of Experiment 1: a “low” value for each indexes the fact that the documents/sentences in which a word appears have high word overlap (100% overlap in the case of Experiment 2), whereas a “high” value indexes little word overlap across contexts, indicating that the contexts in which a word appears are highly semantically distinct.

From the results of Experiment 2, it appears that contextual variability benefits processing for high-frequency words, but that for low-frequency words, variability of contexts neither facilitates nor inhibits processing. This result is consistent with distributional accounts of child language learning. If children rely on distributional information when acquiring word-classes and word

meanings, one would expect words to be learned most easily when their immediate lexical contexts tend to be highly similar to each other (Mintz, Newport & Bever, 2002). When individuals encounter a word that they have not often encountered before (e.g. a low-frequency word) in a context that is very different from any context that they have previously encountered it in, they may at first have difficulty processing the rare word in this unexpected context. However, for words that are encountered very frequently in a large number of different contexts—that is, words that one could reasonably expect to see in any situation—it is perhaps not surprising that they should be processed the most quickly in the unfamiliar context of participating in a lexical decision experiment.

PLN was originally formulated as an interpretation of evidence that contextual diversity, rather than word frequency, best accounts for lexical processing differences (Adelman et al. 2006). Thus, at first it may seem like the fact that word frequency interacts with contextual diversity should be a problem for this account. However, it actually provides stronger evidence for PLN than was previously available. This is because PLN does not require that contextual diversity trump word frequency in all cases, but states only that words more likely to be needed in memory processes are processed more quickly. Intuitively, neither high-frequency words that appear in redundant contexts *nor* low-frequency words that appear in many contexts are nearly as likely to be needed in memory processes at any given time as are high-frequency words that appear in many contexts. The results of both Experiment 1 and Experiment 2 therefore suggest that lexical processing is optimized for precisely those words that are most likely to be required in any given situation.

General Discussion

The results of Experiments 1 and 2 provide a nice set of converging evidence from both a mega-study (the corpus analysis) and a controlled micro-study (the experimentally induced artificial language). Both the mega and micro seem to point to the same pattern of behavioral data: repetition of a word produces greater processing savings if the repetition is accompanied by a change in context. In the artificial language experiment, in fact, if the increase in repetitions was not accompanied by a change in context, then the increased frequency produced no processing savings whatsoever. Granted that we did not see the same pattern of main effects between the two experiments, but these are likely effects that unfold over a very large scale (in the corpus analysis) and may require much more experience to induce with the artificial language experiment. Nonetheless, the interaction of repetition and contextual redundancy found in both experiments is difficult to account for with most existing models of word learning and semantic organization.

Our findings corroborate evidence from others who have studied PLN (Adelman et al., 2006; McDonald & Shillcock,

2001; Pexman et al., 2008) that context variability is potentially a more important variable than is frequency in word recognition and memory access. Further, these results point towards a role for episodic learning in the semantic representation of words. However, our measure of semantic distinctiveness clearly demonstrates that it is not sufficient to simply define contextual distinctiveness as the number of contexts in which a word appears (as in Adelman, 2006) but, rather, it is important to consider the semantic uniqueness of those contexts when building metrics of word access.

The results of these experiments are difficult to account for with current models based on repetition-based learning, and call for learning models that are sensitive to context and shifts in attentional allocation when encoding an episode dependent on how unique a new context is relative to memory of contexts in which the word has been experienced in the past. Models such as Wagenmakers et al.'s (2004) REM-LD model that are sensitive to contextual variability seem promising candidates based on these findings. In addition, we point to a new co-occurrence based learning model that creates a semantic representation similar to Latent Semantic Analysis (Landauer & Dumais, 1997) which can adapt its representation for a word based on fit to previous context episodes stored in memory (Johns & Jones, 2008).

Acknowledgments

We would like to thank Jason Dawson for assisting with data collection, and Sun Ah Kim for developing the stimuli used for the object words in Experiment 2.

References

- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science*, 17, 814–823.
- Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., et al. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. Available at <http://elexicon.wustl.edu/>, Washington University.
- Dennis, S. & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108(2), 452-477.
- Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627-635.
- Johns, B. T., & Jones, M. N. (2008). Predicting word-naming and lexical decision times from a semantic space model. In V. Sloutsky, B. Love, and K. McRae (Eds.) *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.
- Landauer, T. K. and 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.
- 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.
- Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393-425.
- Murray, W. S. and Forster, K. I. (2004) Serial mechanisms in lexical access: the Rank Hypothesis. *Psychological Review*, 111(3), 721-756.
- Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008) The are many ways to be rich: Effects of three measures of semantic richness on word recognition. *Psychonomic Bulletin & Review*, 15, 161-167.
- Steyvers, M., & Malmberg, K. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29(5), 760-766.
- Tarr, M. J. (2006). Fribbles [Data file]. Available at the Tarr Lab website, <http://titan.cog.brown.edu:8080/TarrLab/stimuli/novel-objects/fribbles.zip/view>
- Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 796-800.
- Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, 48, 332-367.
- Wickens, D. D. (1987). The dual meanings of context: Implications for research, theory, and applications. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and learning: The Ebbinghaus Centennial Conference*. Hillsdale, NJ: Erlbaum.