

Running head: SEMANTIC DIVERSITY

The Role of Semantic Diversity in Lexical Organization

Michael N. Jones, Brendan T. Johns, & Gabriel Recchia

Indiana University

In press, *Canadian Journal of Experimental Psychology*

Correspondence

Michael Jones
Dept. of Psychological and Brain Sciences
1101 E. 10th St.
Indiana University
Bloomington, IN, 47404

Email: jonesmn@indiana.edu
Phone: (812) 856-1490
Fax: (812) 855-4691

Abstract

Recent research has challenged the notion that word frequency is the organizing principle underlying lexical access, pointing instead to the number of contexts that a word occurs in (Adelman, et al., 2006). Counting contexts gives a better quantitative fit to human lexical decision and naming data than counting raw occurrences of words. However, this approach ignores the information redundancy of the contexts in which the word occurs, a factor we refer to as semantic diversity. Using both a corpus-based study and a controlled artificial language experiment, we demonstrate the importance of contextual redundancy in lexical access, suggesting that contextual repetitions in language only increase a word's memory strength if the repetitions are accompanied by a modulation in semantic context. We introduce a cognitive process mechanism to explain the pattern of behavior by encoding the word's context relative to the information redundancy between the current context and the word's current memory representation. The model gives a better account of identification latency data than models based on either raw frequency or document count, and also produces a better-organized space to simulate semantic similarity.

The Role of Semantic Diversity in Lexical Organization

A consistent finding in studies of lexical access is that high-frequency words are identified faster than low-frequency words (Broadbent, 1967; Forster & Chambers, 1973; Kruger, 1975). The efficiency of processing suggests that high-frequency words have a privileged status over low-frequency words in the mental lexicon, as the effect is stable across different response tasks (e.g., word naming and 2AFC choices such as lexical, concreteness, and category decisions). Frequency is core to classic strength accounts of lexical access based on the assumption that each repetition increases memory strength for a word, boosting the efficiency of later access¹. This *principle of repetition* has been influential on formal models of lexical access, leading to the development of rank frequency models of the lexicon (Murray & Forster, 2004), threshold activation accounts (Coltheart et al., 2001), and connectionist models (Seidenberg & McClelland, 1989).

However, recent research is questioning whether humans use frequency information to organize the lexicon. Word frequency may appear to be the organizing factor because it is confounded with a word's contextual diversity (CD)—the number of contexts in which the word has been experienced. But it may be a word's CD and not frequency that humans use to organize lexical priority. Unfortunately, CD is a slippery construct to define and takes on slightly different operational definitions across the handful of experiments that have studied it. Generally, CD is conceptualized as the number of distinct contexts in which a word occurs. If frequency is based on the principle of repetition, CD is based on the *principle of likely need* emphasized by rational models of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991; Dennis & Humphreys, 2001): A word that has been experienced in many contexts during learning is more

¹ Alternatively, repetition may affect the number of instances of a word in memory, increasing availability, as in popular multiple-trace models (e.g., Hintzman, 1986).

likely to be needed in unknown future contexts, hence it is more accessible in the lexicon. The variable is most commonly operationalized as the number of documents in which a word occurs across a text corpus (with no regard for frequency within documents). Adelman and Brown (2008) summarize the theoretical position: “As words tend to cluster in contexts, the likely need of a word in an arbitrary new context relates to the number of contexts the word has been seen in before, not the number of occurrences of the word” (p. 223). This phenomenon is heavily related to the concept of word “burstiness” in information retrieval (Katz, 1996). If humans are sensitive to word frequency information, then repeating a word should be beneficial to later identification of that word. If humans use CD information, however, then repeating a word is of limited use if the repetition is not also accompanied by a modulation in context.

Schwanenflugel and Shoben (1983) originally demonstrated that a variable they termed “context availability” influences word recognition. They define this variable as the ease with which one can think of a particular circumstance in which a word might appear, and have argued that this is the “real” explanatory variable underlying the concreteness effect (cf. Galbraith & Underwood, 1973). In the recognition memory literature, robust evidence has been found that CD benefits item learning and retrieval efficiency. Items with a greater CD at encoding are more likely to be subsequently recognized and tend to be recognized faster (Goldinger & Azuma, 2004; McDonald & Shillcock, 2001; Nelson & Shiffrin, 2006; Pexman et al., 2008). Steyvers and Malmberg (2003) have also demonstrated a role for CD as a strong component of the mirror effect of “frequency” seen in recognition memory. More recently, CD has been found to influence free recall as well (Lohnas, Polyn, & Kahana, 2011). Further, CD has been demonstrated to benefit learning of grammatical classes (Redington, Chater, & Finch, 1998),

speech perception (Apfelbaum & McMurray, 2011), and word-referent mappings (Smith & Yu, 2008).

In a recent study, Adelman, Brown, and Quesada (2006) conducted a corpus analysis of CD, operationalizing it as the number of documents in which a word occurs in large English corpora. They computed diversity and frequency counts for thousands of words and conducted a regression analysis to explore how the two measures predicted lexical decision and naming times for the words taken from Balota et al.'s (2002) English Lexicon Database. Adelman et al. found clear evidence for the superiority of CD over word frequency in predicting word identification latency: CD predicted all variance in latency data that frequency did, and additional unique variance. They argue that previous theories have been constructed based on a false assumption that humans use frequency information to organize the lexicon, and their work suggests that many current models need to be abandoned or revised to adequately explain how the lexicon is organized.

However, it is difficult to determine conclusively if CD is the organizing principle of lexical priority with the methodologies used in existing studies. Firstly, it is questionable whether common operational definitions of CD are valid measures under the principle of likely need. The most common operational definition of CD is simply the number of documents in which a word appears in a text corpus: A word's frequency count is incremented each time it occurs in the corpus, but its diversity count is only incremented each time it occurs in a new document. This operationalization of frequency is fair, but operationalizing diversity as a document count is likely to be an invalid measure of the true contextual diversity of the word. 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 (Schmidt, 1991; Verkoeijen, Rikers, & Schmidt,

2004; Wickens, 1987). It is not directly obvious how counting documents corresponds to classic notions of a change in context, but it seems intuitive that if a document is repeated in the corpus, we should not consider the two repetitions to be different semantic contexts of the word. Further, a frequent discourse topic is likely to have many documents dedicated to it, and so a word that describes a frequent topic is likely to appear in more documents, even though the documents are not truly distinct contextual uses of the word. What is needed is a graded measure of CD by examining the information overlap among a word's linguistic contexts.

Secondly, the superiority of CD over frequency has only been demonstrated with regression analyses: CD accounts for all of the variance in identification latency that frequency does, and additional unique variance when frequency is partialled out (Adelman et al., 2006). However, regression analyses alone do not provide conclusive evidence for the causal role of CD due to confounds with a variety of other variables in addition to frequency, any one of which could plausibly be the causal factor influencing lexical access. For example, access may simply be superior for words that have been experienced more recently; words with a greater CD or frequency are also likely to have a higher recency (but see Balota & Spieler, 1999). Ambiguity, abstractness, imageability, and word length are also confounded with document count and frequency, and are difficult to tease apart. Finally, it has been suggested (Balota; in Adelman et al.) that document count from a text corpus may actually be a better measure of real-world frequency due to the structure of the corpus. Recent corpus-based studies attempt to partial out the confounding variables as covariates. However, the effect of contextual diversity has never before been induced experimentally; to do so would require control over the statistical structure of the language being learned.

The outline of this paper is as follows. Experiment 1 addresses the issue of how to measure CD by introducing a graded *semantic distinctiveness count* and using a corpus analysis and large-scale fits to lexical decision and naming times to demonstrate its superiority over word frequency or document count. Experiment 2 demonstrates a causal effect of diversity using an artificial language paradigm to independently manipulate document count and semantic distinctiveness count. Finally, we introduce a computational model based on co-occurrence learning models and expectation-congruency, which adjusts its encoding strength for a word relative to the information redundancy between the current memorial representation of the word and the current linguistic context in which the word is experienced. Within the same model framework, we are able to compare semantic diversity with nested models considering only raw frequency or document count, and demonstrate the superiority of a semantic diversity learning mechanism in accounting for human word identification latency.

Experiment 1: A Role for Semantic Diversity

Adelman et al. (2006) demonstrated the superiority of document count over raw frequency in fitting lexical decision and naming times. However, operationalizing CD as a document count ignores semantic context—the information overlap between documents. Under the principle of likely need (Anderson & Milson, 1989), repetition of a word in distinct documents would increase its likely need to a greater extent than an equal number of repetitions in redundant documents. For example, if the word *bank* occurs in two very similar documents discussing the government-sponsored buyout of mortgage assets, we would consider the two documents to be very similar contextual uses of *bank* compared to the contextual similarity between one of these financial documents and a document discussing river banks. In addition, the government buyout may be a very frequently discussed discourse topic (having many

documents on the topic), meaning that even though *bank* would receive a large document count, these are not truly distinct contextual uses of the word, hence it is a poor operational definition to be true to the principle of likely need. The example with a homograph like *bank* makes the point clear, but this pattern will be true of all words as a function of slight contextual modulations; a measure needs to consider the graded semantic coherence of the contexts in which words occur to estimate contextual diversity.

Semantic Distinctiveness Count

To introduce a weighted context count, we first quantify the dissimilarity of any pair of documents in which a word occurs as a function of the proportion of overlapping words:

$$\text{Dissimilarity}(doc_i, doc_j) = 1 - \frac{|doc_i \cap doc_j|}{\min\{doc_i, doc_j\}} \quad (1)$$

Document similarity is the intersection of the two sets of words, divided by the size of the smaller document. Function words (e.g., *the*, *is*, *of*, etc.) are filtered out of the calculation using the standard LSA stoplist (Landauer & Dumais, 1997). Document dissimilarity is then 1 – similarity. A dissimilarity value of 1 indicates that there is no content word overlap between the two documents (i.e., high distinctiveness), analogous to if it were presented in two separate lists in a standard memory experiment. A value of 0 indicates that the two documents are identical (i.e., low distinctiveness), similar to if the word were repeated in the same list in a standard memory experiment. A word's *semantic distinctiveness* is defined as the mean dissimilarity over all the documents in which it occurs. A word with a low mean distinctiveness tends to occur in documents that are semantically redundant, whereas a high mean distinctiveness indicates the set of documents that contain the word are semantically unique.

Finally, a word's *semantic distinctiveness count* (SD_Count) is computed to consider the number of documents in which it occurs weighted by the semantic uniqueness of those contexts. For the word set, the distribution of dissimilarity values is standardized to quantiles. In this paper, we use septiles (dividing the cumulative distribution into seven equal bins) as previous research has pointed to seven as the optimal number of divisions to fit identification latencies (Johns & Jones, 2008). Document pairs that are in a higher quantile are more distinct than those in lower quantiles. For a given word, its SD_Count is calculated as the sum of the quantiles in which the set of documents containing it fall:

$$SD_Count = \sum_{i=1}^n \sum_{j=1}^i \text{quantile}(Dissimilarity(doc_i, doc_j)) \quad (2)$$

Words that tend to appear in a greater number of unique documents will have a higher SD_Count. Note that if there were a single quantile, this function would be equal to a document count (incrementing the count by one each time the word appears in a new document with no regard for information overlap). When the distribution is split into more than one quantile, the function produces a greater count for more unique contextual uses of the word. Comparing two words that occur in an equal number of documents, the one that occurs in more redundant documents will have a lower SD_Count than the one that occurs in more distinct documents. If a word were to hypothetically only occur once in a document, and the document was repeated multiple times throughout the corpus, then the frequency count, document count, and SD_Count variables would all be equal. Using SD_Count, we have a variable that is sensitive to the uniqueness of semantic contexts in which words occur—more unique contexts are weighted more heavily than less unique contexts.

Method

We computed word frequency, document count, and SD_Count from three corpora: 1) the Touchstone Applied Science Associates (TASA) corpus (Landauer & Dumais, 1997), 2) a Wikipedia corpus (Recchia & Jones, 2009), and 3) a New York Times (NYT) corpus (Jones & Mewhort, 2004). The TASA corpus was composed of 10,500 documents, with each document having a mean length of 289 words. The Wikipedia corpus was composed of 9,755 documents, with a mean document length of 391 words. The NYT corpus is composed of 9,100 documents with a mean length of 250 words. These are smaller versions of the full corpora, and the reduced size was necessary due to the computational complexity of this measure. Lexical decision times (LDTs) and naming times (NTs) were obtained from the English Lexicon Project (Balota, et al., 2002). Measures were computed for 17,984, 22,673, and 14,609 words for the TASA, WIKI, and NYT corpora, respectively.

Results

Figure 1 illustrates a median split of words with high/low semantic distinctiveness (mean document dissimilarity) by high/low document count for the LDTs and the NYT corpus (the pattern was consistent across all corpora and with NTs as well). A main effect was observed for document count, with words occurring in more documents having faster LDTs than words occurring in fewer, $F(1,4232) = 177.43, p < .001$. Further, a main effect was observed for semantic distinctiveness, with more semantically distinct words having faster LDTs than less distinct words, $F(1,4232) = 143.78, p < .001$. Substituting word frequency for document count produces the same result (frequency and document count are highly correlated). Of particular interest is the finding that semantic distinctiveness and document count interacted, $F(1,4232) = 74.26, p < .001$. As document count increases, words that occur in a greater number of

semantically distinct documents see a greater benefit on their LDTs from the additional contextual occurrences.

However, Figure 1 only demonstrates that words that occur on average in more unique documents have an RT savings in identification. Figure 2 uses the SD_Count variable (a count of the documents in which the word occurs weighted by their semantic distinctiveness) to analyse the proportion of variance explained by SD_Count over word frequency (the zero point) and document count. The top panel shows LDT and the bottom panel NT. As the figure indicates, counting contexts relative their semantic uniqueness gives a much better account of the human data across all corpora for both LDT and NT. Because the variables are strongly correlated with each other, we emulate Adelman et al.'s (2006) original regression analyses and examine variance predicted while systematically partialling out covariates. Table 1 shows the unique variance predicted by each variable for LDT and Table 2 for NT². As the tables indicate, the SD_Count variable gives a significantly better prediction of both LDT and NT than document count, $p < .001$ in all cases, and subsumes the effect of frequency just as effectively as does document count.

Discussion

The corpus analysis suggests that when words with equivalent document counts are considered, those that occur in more semantically distinct contexts see a larger latency savings when compared to those that occur in redundant contexts. Our SD_Count variable follows the principle of likely need and corroborates the findings of Adelman et al. (2006), but clearly demonstrates that the lexical priority of a word depends on both the number and redundancy of the contexts in which it has been experienced. However, our regression analysis shares the same

² Following Adelman et al. (2006), we use log transformations of each predictor in the regression. The ordinal pattern summarized in Tables 1 and 2 is the same with either power or rank transformations.

weaknesses with other correlational studies criticized in the introduction. Several confounding factors may still be the hidden causal variables (e.g., recency, concreteness, etc.). We next rule out these potential confounds by inducing the effect of diversity experimentally.

Experiment 2: Testing Diversity in an Artificial Language

Experiment 2 was designed to test the hypothesis that repetition of contextual occurrences produces greater latency savings for unique contexts than redundant contexts, as well as to compare the effect of contextual diversity on lexical decision times in a controlled paradigm. The CD effects used to support the principle of likely need have never been induced experimentally because in natural languages CD is confounded with many other sources of statistical information (McDonald & Shillcock, 2001). Thus, we used an artificial language paradigm to independently vary frequency/document count and semantic diversity, and to assess the relative contribution of each on identification latency.

Method

Participants

Thirty-two undergraduate students at Indiana University participated in the experiment for partial course credit. All had normal or corrected-to-normal vision.

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 Balota, et al.’s (2002) database, 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.

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 half the participants.

For each participant, a set of 450 training slides was created. Each training slide consisted of a three-word “sentence” in Xaelon 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”; Tarr, 2010). Each object word corresponded to a different geometric shape constructed from geons, 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.

Finally, of the twelve Xaelon words, four were randomly selected and crossed for a factorial combination of two levels of word frequency (hi/low) and two levels of semantic distinctiveness (hi/low). Note that in this experiment, a word’s “document count” and its frequency are equivalent, as there are no repetitions within a context; hence we will just refer to repetitions as the frequency of Xaelon sentences in which the word occurs. Low-frequency words appeared 45 times each in the training slides, while high-frequency words appeared 180 times each. Further, whenever a low-diversity word appeared, it always appeared in the same semantic context (i.e., in the same sentence and with the same image), whereas each high-diversity 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. Participants 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. An example of a training trial is displayed in Figure 3. Following the training trials, participants were confronted with a surprise pseudo-lexical decision task (PLDT) 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 participants during the lexical decision task.

Results

Participants performed quite well at the PLDT task, with a mean accuracy of .88 ($SE = .02$) across all target and foil trials. We set a stringent accuracy criterion of 85%, which trimmed seven 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. Response latencies did not differ as a function of part-of-speech (subject, locative, object), $F(2,23) = 0.11, ns$.

Figure 4 plots mean response latency as a function of frequency and semantic diversity. A repeated-measures ANOVA indicated no significant main effects for either frequency or semantic diversity, 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 diversity at low frequency was non-significant, $t(24) = -1.54, ns$, however, the difference between the levels of diversity at high frequency was statistically reliable, $t(24) = 2.11, p < .05$. Further, the change in PLDT latency across the levels of frequency for low diversity was statistically flat, $t(24) = -1.67, ns$. However, the decrease in PLDT latency over frequency for high diversity was statistically significant, $t(24) = 2.06, p < .05$. These results demonstrate that increasing the repetitions of a pseudoword from 45 to 180 produced no facilitation in PLDT if the contexts in which the word occurred were unchanged. However, processing savings were observed if the increase in frequency was accompanied by a change in contexts across learning.

Discussion

Consistent with the principle of likely need (and Experiment 1), Experiment 2 suggests that lexical access is facilitated for words appearing in a large number of contexts that are high in semantic distinctiveness. However, appearing in a large number of redundant contexts produced equivalent response latencies to a much lower number of repetitions in the redundant context. This finding parallels 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.

A Computational Model of Semantic Diversity

Experiments 1 and 2 demonstrate that semantic diversity is an important source of information used by humans to organize the lexical priority of words. Both experiments point to

an encoding operation by which the word is encoded most strongly if the current episodic context provides novel information about the word not already contained in memory. This conceptual framework is consistent with rational models (Chater & Oaksford, 1997) as well as Rosch's (1978) notion of cognitive economy. It is also consistent with expectancy-congruency effects (e.g., Hirschman, 1988; Ranganath & Rainer, 2003; Schmidt, 1991)—unexpected or distinctive events are more memorable than expected occurrences. All of these well-established effects seem to have overlap with our semantic distinctiveness findings. However, we still lack a mechanistic explanation of how a process might give rise to the structure seen in studies of lexical access.

Contemporary computational models of lexical semantic similarity are also based on frequency (e.g., Landauer & Dumais', 1997, Latent Semantic Analysis). Rather than single-token frequency, however, they depend on the co-occurrence frequency of words across a linguistic corpus. For example, the word *milk* may frequently co-occur in the same contexts as *drink* and *cookie*. As a result, it can be inferred that these words are semantically related. For reviews of the various models, see McRae and Jones (in press) or Riordan and Jones (2011). Typically, these models first weight the word inversely proportional to its document entropy across the corpus. However, the weighting scheme is blind to the semantic content of those documents, which Experiments 1 and 2 suggest is an assumption incompatible with human encoding.

Co-occurrence models of lexical similarity also have the potential to account for lexical access. Typically, co-occurrence models learn from a word-by-document frequency matrix representation of a text corpus, and represent a word's meaning as a vector over semantic components. If the vectors for two words have a correlated pattern over components, they are similar. However, each word's vector also contains information about the word's individual

frequency of occurrence as well (the magnitude of the vector), and this information is often discarded as a nuisance (Durda & Buchanan, 2008; Shaoul & Westbury, 2006). For example, if Murray and Forster's (2004) model of lexical access based on rank frequency is correct, this information is already contained in the vector magnitude of a co-occurrence model—it just happens to be discarded when computing semantic similarity. Any vector contains both phase (direction) and magnitude (length). If a vector representation for a word is thought of abstractly as a “brain state” when the word is processed, then semantic similarity is the similarity of brain state phase patterns between two words, and lexical access is determined by the magnitude (intensity) of the brain state when the word is processed in isolation.

We build our semantic distinctiveness model on a word-by-context matrix, in which each word is represented as a distribution over documents (either by frequency, document count, or weighted by semantic distinctiveness). This architecture allows a single model to be used to simultaneously account for single-word identification latency as well as paired-word semantic similarity data.

The Semantic Distinctiveness Model

The Semantic Distinctiveness Model (SDM) is based on other co-occurrence models of semantic memory, using a word-by-context matrix representation of a text corpus. However the vector representation of a word is “grown” as the word is experienced in contexts. Each time a word is experienced in the corpus, the model compares the prediction of the word's current memory representation to the information in the current context. If the information in the current context is highly consistent with the current contents of memory, the context is encoded at a weaker magnitude. However, if the information in the context is novel compared to the current contents of memory, it is encoded at a much stronger magnitude. This process creates a

representation consistent with our SD_Count variable. The same model framework may be used with a pure frequency or document count, allowing model comparisons from within the same formal framework.

When a word is experienced in a document, each word in the document is retrieved from memory and its environmental context is represented as the sum of the vectors of the other words in the document (cf., in a list memory task, the item's context is the other items occurring in the list with it):

$$Context = \sum_{i=1}^n T_i, \quad (3)$$

where n is the number of words in the document, and T_i is the corresponding memory vector for a given word in the document (again, with function words removed).

Next, the model assesses how similar the current contextual representation of the word is to its current memorial representation. The cosine (normalized dot product) is computed between the target word's contextual representation and its memory representation. If the cosine is relatively high, the current context is redundant with information already stored in memory; hence the current context is encoded at a lower weight. However, if the cosine is relatively low, the current context is more unique from information already stored in the word's memory vector; hence, the current context is encoded at a greater weight. The cosine is transferred through an inverse exponential density function (following Shepard's 1987 universal law of similarity scaling) to reflect the current document's semantic distinctiveness (SD):

$$SD = e^{-\lambda \cos(context, word_i)}, \quad (4)$$

where λ is a fixed parameter with a positive value representing the slope of the similarity gradient. This SD value is then added into the memory matrix for the target word (row) in the

specific context (column) of the matrix. A document count model can be considered to be nested within this model, with a λ fixed at 0.

When a word is first encountered its memory vector will be empty, hence, the SD value will always be 1.0 for the first occurrence of a word, and it will be encoded at maximal strength. The second time a word is experienced, the similarity of this context is compared to the word's current lexical representation (which only contains the first context). If this is a repetition of the first document, the new context will be encoded at minimal strength. If, however, it is a context that is unique from the first, the new context will be encoded at maximal strength. In this fashion the word-by-document matrix has columns added to it each time a new document is learned, with the encoding strength for a document (for a particular word) dependent on the goodness-of-fit between what has been learned and what is being experienced.

Simulation 1: SDM Simulation of Experiment 1

The SDM model was trained on the same corpora used in Experiment 1. The lambda parameter was fixed at 5.5 to encode SD for all following simulations, as preliminary work has suggested that fit to the human identification latency reaches asymptote at this value (similar to the septile in our SD_Count; see Johns & Jones, 2008). The analogue of the regression analysis conducted in Experiment 1 was conducted to assess the variance predicted in LDT and NT comparing a word's representation between versions of the model designed to attend to word frequency, document count, and SD.

For each vector representation, a word's relative access was computed as the sum of its vector elements (magnitude). As in Experiment 1, log transforms of all variables were used, although the pattern reported is consistent across power and rank transforms as well. The other difference from Experiment 1 is that the models were trained on the full versions of the corpora: 37,600 documents from TASA, with an average length of 121 words per document, 40,000

documents from the Wikipedia corpus, with an average document length of 279 and 17,399 documents from The New York Times corpus, with an average document length of 250 words. LDT and NT data were again obtained from Balota et al.'s (2002) database. In the analysis, latencies from 29,799, 35,518, and 20,744 words were used for the TASA, WIKI, and NYT corpora, respectively.

Tables 3 and 4 show the increase in R^2 for each version of the model for LDT and NT, respectively, while controlling for the variance accounted for by the other versions of the model (analogous to Tables 1 and 2). Similar to the corpus counts, the SDM model accounted for significantly more variance in both LDT and NT than models based on either frequency or document count. As was suggested by the corpus analysis, a mechanism that adjusts encoding strength relative to the amount of new information in a context not already contained in memory provides a better account of human single-word identification latency data.

Simulation 2: SDM Simulation of Experiment 2

SDM was next used to simulate the artificial language data from Experiment 2. Xaelon sentences were considered to be distinct contexts, and a word-by-context matrix was constructed using equations 3 and 4. The exact sentences presented to the subjects in Experiment 2 were presented to the model, and predictions for PLDT were computed as the magnitude of a word's memory vector.

The resulting behavior of SDM is plotted in Figure 5.³ Note that neither a frequency nor document count version of the model could simulate this result by design of the stimuli—each would produce parallel effects rather than an interaction. SDM naturally produces a pattern similar to the behavioral data from Experiment 2: Words that were repeated in multiple distinct

³ Note that the scale on the ordinate has been changed to negative magnitude to be consistent with Figure 4. Vector magnitude (word intensity) is negatively correlated with identification latency.

contexts produce a greater representational intensity than words that are repeated in redundant contexts. As with the human subjects, the simulation indicates that contextual repetitions of the word only benefit processing if the repetitions are accompanied by a change in semantic context. While this pattern could not be produced by a model that encoded frequency or contextual occurrences, it is a natural consequence of a mechanism that encodes words relative to their information overlap with what has already been stored.

Simulation 3: A Test of Semantic Similarity

Co-occurrence models of semantic similarity commonly use a word-by-document frequency matrix to determine lexical relatedness. However, the frequency assumption made by these co-occurrence models may also be wrong, and a simple document count or weighted semantic distinctiveness count may be the correct initial matrix representation to mimic human semantic similarities as well as lexical access. While many co-occurrence models apply a dimensional reduction mechanism (e.g., Landuaer & Dumais, 1997), recent work suggests that raw co-occurrence counts from the original matrix may give a better approximations of human semantic similarity if simulated at a sufficiently large scale (Louwerse & Connell, 2011; Bullinaria & Levy, 2007; Recchia & Jones, 2009). Hence, we compare the cosines between specific word vectors learned by the raw frequency, document count, and SD versions of SDM here to explore how they map onto semantic similarities.

For each model, predictions were made for the 45,786 word-pair similarities from Maki, McKinley, and Thompson (2004), computed from WordNet. Maki et al. have presented comparisons to human judgments of semantic similarity suggesting that the WordNet JCN metric is a very close correspondence to human similarity judgments. For each word pair, under each model representation (frequency, document count, SD), the predicted semantic similarity was

simply the cosine of their respective vector representations. The SD version of the model gave a significantly better prediction of the WordNet semantic similarities ($r = .172$) than either frequency or document count ($r = .126$ and $r = .161$, respectively), and also outperformed Landauer and Dumais' (1997) LSA model, which produced $r = .158$.⁴

While semantic similarity among words is not the focus of this paper, this simulation does demonstrate that the incorrect assumption of frequency in lexical access may have also been falsely applied to co-occurrence models of semantic similarity, and semantic diversity may be the correct source of information underlying both. Semantic diversity can be learned by a simple mechanism within a context co-occurrence framework and, as a byproduct, it also seems to produce a better organized semantic space. In this manner, we can represent both lexical access and lexical similarity within the same model, allowing insights into how single- and paired-word tasks are related to the same memorial structure.

General Discussion

Experiments 1 and 2 provide 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 semantic context. Our findings corroborate recent evidence from others (e.g., Adelman et al., 2006; Pexman et al., 2008) suggesting that CD is potentially a more important variable than is frequency in word recognition and memory access. But it also makes a clear case for the importance of semantic context. The interaction of repetition and semantic redundancy found in both experiments is difficult to account for with most existing models of word identification.

⁴ Note that with 45,786 pairs, all numeric differences between correlation coefficients are significant.

Our semantic distinctiveness model implements this pattern by using an expectancy-congruency mechanism to build a word-by-context co-occurrence matrix: The encoding strength for a word in a given context is relative to the information overlap between the context and the current 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 (e.g., Kruschke, 1992). In addition, newer models of lexical access, such as Wagenmakers et al.'s (2004) REM-LD model, are sensitive to contextual variability and seem promising candidates to explain these findings.

Finally, our SDM implementation of a word-by-context matrix not only outperformed matrices based on frequency and document count, but the resulting matrix seemed to produce better semantic organization as well, a free lunch we were not explicitly trying to create. This pattern bolsters the importance of semantic distinctiveness over frequency or document count, and points to an important connection between models of lexical access (based on single-word statistics) and lexical similarity (based on co-occurrence statistics).

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.
- Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, 115, 214-227.
- Anderson, J.R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703-719.
- Anderson, J.R., & Schooler, L.J. (1991) Reflections of the environment in memory. *Psychological Science*, 2, 396-408.
- Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35, 1105-1138.
- Balota, D. A., Cortese, M. J., Hutchinson, K. A., Neely, J. H., Nelson, D., Simpson, G. B., & Treiman, R. (2002). The English Lexicon Project. Retrieved September 30, 2007, from <http://elexicon.wustl.edu/>.
- Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128, 32-55.
- Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39, 510-526.
- Broadbent, D. E. (1967) Word-frequency effect and response bias. *Psychological Review*, 74, 1-15.

- Chater, N., & Oaksford, M. (1997). Ten years of the rational analysis of memory. *Trends in Cognitive Science*, 3, 57-65.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108, 452-478.
- Durda, K., & Buchanan, L. (2008). Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40, 705-712.
- Forster, K.I., & Chambers, S.M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627, 635.
- Galbraith, R. C., & Underwood, B. J. (1973). Perceived frequency of concrete and abstract words. *Memory & Cognition*, 1, 56-60.
- Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, 11, 716-722.
- Hintzman, D.L. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411-428.
- Hirshman, E. (1988) The expectation-violation effect: Paradoxical effects of semantic relatedness. *Journal of Memory and Language*, 27, 40-58.
- 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 Cognitive Science Society Meeting*, 279-284.

- Jones, M. N., & Mewhort, D. J. K. (2004). Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior Research Methods, Instruments, and Computers*, 36, 388-396.
- Katz, S. M. (1996). Distribution of content words and phrases in text and language modeling. *Natural Language Engineering*, 2, 15-59.
- Krueger, L. E. (1975). Familiarity effects in visual information processing. *Psychological Bulletin*, 82, 949-974.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language*, 64, 249-255.
- Louwerse, M. M., & Connell, L. (2011). Symbol interdependency in symbolic and embodied cognition. *Cognitive Science*, 35, 381-398.
- Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, 36, 421-431.
- McRae, K., & Jones, M. N. (in press). Semantic memory. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology*.

- 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.
- Murray, W. S., & Forster, K. I. (2004). Serial Mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111, 721-756.
- Pexman, P., Hargreaves, I., Siakaluk, P., Bodner, G., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, 15, 161-167.
- Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, 4, 193-202.
- Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41, 647-656.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic information. *Cognitive Science*, 22, 425-469.
- Riordan, B., & Jones, M. N. (2011). Redundancy in linguistic and perceptual experience: Comparing distributional and feature-based models of semantic representation. *Topics in Cognitive Science*, 3:2, 303-345.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.
- Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, 19, 523-542.

- 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.
- Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 111, 721-756.
- Shaoul, C. & Westbury, C. (2006). Word frequency effects in high dimensional co-occurrence models: A new approach. *Behavior Research Methods*, 38, 190-195.
- Shepard, R. N. (1987); Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Smith, L.B., & Yu, C. (2008). Infants Rapidly Learn Word-Referent Mappings via cross-situational Statistics. *Cognition*, 106, 333-338.
- 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.
- Tarr, M.J. (2010). Fribble images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University.
<http://www.tarrlab.org/>.
- 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, & 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.

Author Notes

Michael Jones, Brendan Johns, and Gabriel Recchia; Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, 47405. This research was supported by grants from Google Research and NSF BCS-1056744 to MJ. BJ was supported by a post-graduate scholarship from NSERC.

Table 1.
Lexical Decision Time Variance Predicted by SD_Count, Word Frequency, and Document Count

Analysis	Effect (ΔR^2 in %)		
	TASA	WIKI	NYT
SD_Count (After WF)	5.501	6.417	6.282
DC (After WF)	2.341	1.675	0.0 <i>n.s.</i>
SD_Count (After DC)	3.87	6.807	11.557
WF (After DC)	0.0 <i>n.s.</i>	0.382	1.123
DC (After SD)	0.645	2.094	5.025
WF (After SD)	0.0 <i>n.s.</i>	0.0 <i>n.s.</i>	0.0 <i>n.s.</i>
SD_Count(After DC,WF)	4.487	7.731	11.881
WF(After SD, DC)	1.282	1.03	1.485
DC(After SD, WF)	0.641	3.108	5.445

Note. Unless otherwise specified, all values are significant. Raw data were log transformed.

Table 2.

Naming Time Variance Predicted by SD_Count, Word Frequency, and Document Count

Analysis	Effect (ΔR^2 in %)		
	TASA	WIKI	NYT
SD_Count (After WF)	8.49	9.016	7.751
DC (After WF)	3.98	2.654	0.0 <i>n.s.</i>
SD_Count (After DC)	6.511	11.718	13.235
WF (After DC)	0.217	0.0 <i>n.s.</i>	0.847
DC (After SD)	0.471	5.468	6.617
WF (After SD)	1.86	0.819	1.55
SD_Count(After DC,WF)	6.511	12.403	13.868
WF(After SD, DC)	1.86	0.775	1.459
DC(After SD, WF)	0.465	5.833	6.569

Note. Unless otherwise specified, all values are significant. Raw data were log transformed.

Table 3.
Lexical Decision Time Variance Predicted by SDM, Word Frequency, and Document Count Models

	Effect (ΔR^2 in %)		
	TASA	WIKI	NYT
SDM (After WF)	3.048	1.81	5.461
DC (After WF)	1.274	0.786	0.0 <i>n.s.</i>
SDM (After DC)	2.346	0.849	6.901
WF (After DC)	0.03	0.364	1.07
DC (After SDM)	0.511	0.141	0.462
WF (After SDM)	0.0 <i>n.s.</i>	0.704	0.0 <i>n.s.</i>
SDM(After DC, WF)	3.118	1.175	7.348
DC(After SDM, WF)	1.327	0.149	2.001
WF(After SDM, DC)	0.816	0.7	1.549

Note. Unless otherwise specified, all values are significant. Raw data were log transformed.

Table 4.

Naming Time Variance Predicted by SDM, Word Frequency, and Document Count Models

	Effect (ΔR^2 in %)		
	TASA	WIKI	NYT
SDM (After WF)	5.811	3.323	6.568
DC (After WF)	2.904	2.01	0.0‡
SDM (After DC)	4.984	1.213	7.791
WF (After DC)	0.119	0.0 <i>n.s.</i>	0.75
DC (After SDM)	2.062	0.0 <i>n.s.</i>	0.724
WF (After SDM)	0.386	0.132	0.0 <i>n.s.</i>
SDM(After DC, WF)	5.361	1.336	8.243
DC(After SDM, WF)	2.163	0.0 <i>n.s.</i>	1.868
WF(After SDM, DC)	0.485	0.117	1.197

Note. Unless otherwise specified, all values are significant. Raw data were log transformed.

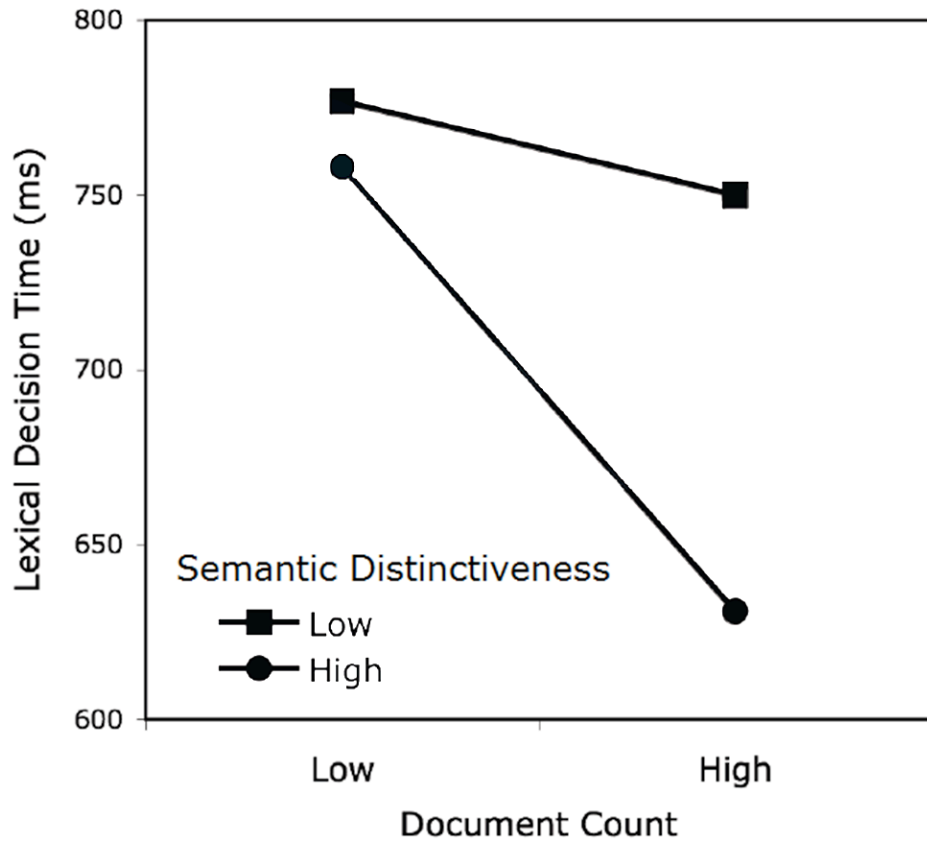


Figure 1. Factorial combination of semantic distinctiveness by document count on lexical decision times. Repetitions of a word in documents produces greater latency savings if the documents are low in semantic redundancy.

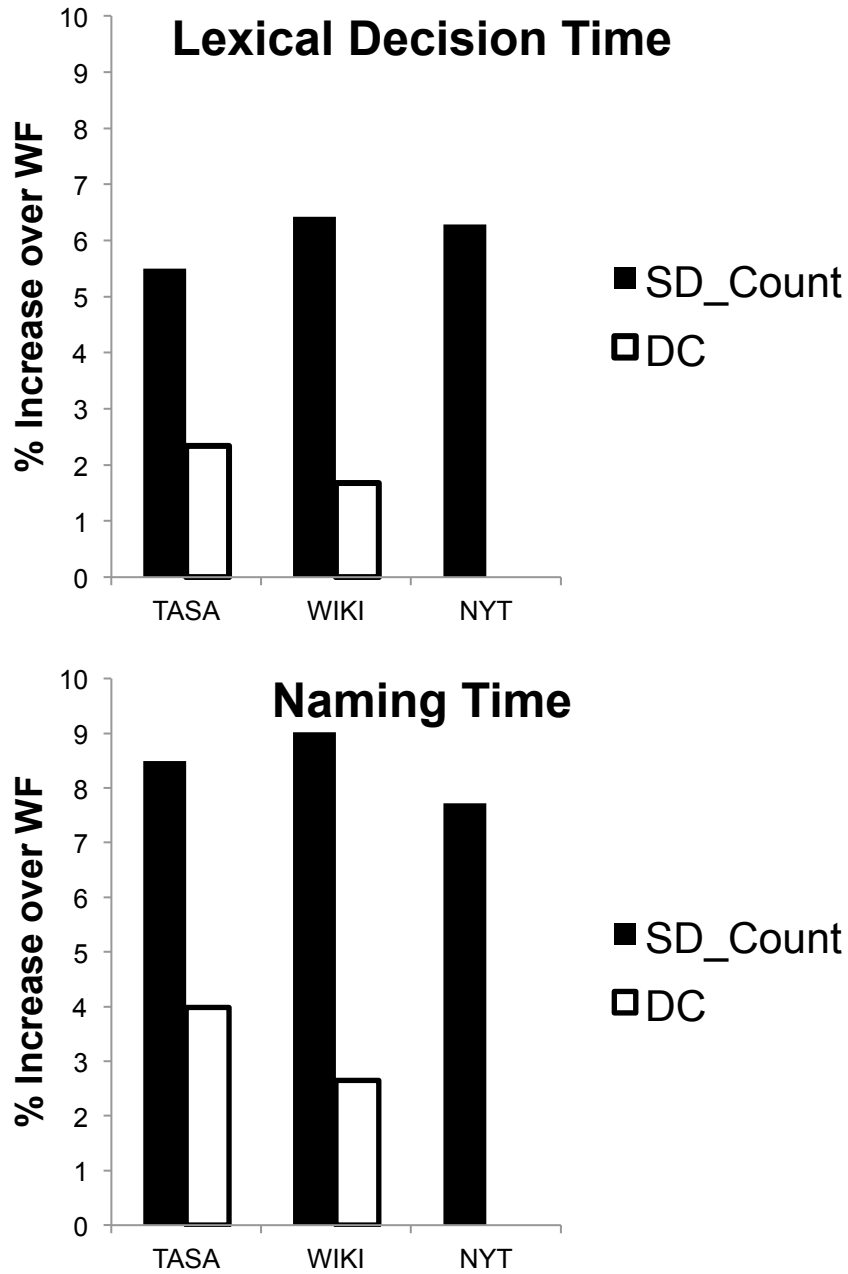


Figure 2. Increase in variance predicted over word frequency for lexical decision times (top panel) and naming times (bottom panel) for document count and SD_Count.

plurt gluds leuts



Figure 3. Example of a training slide seen by participants while learning the “alien” language Xaelon.

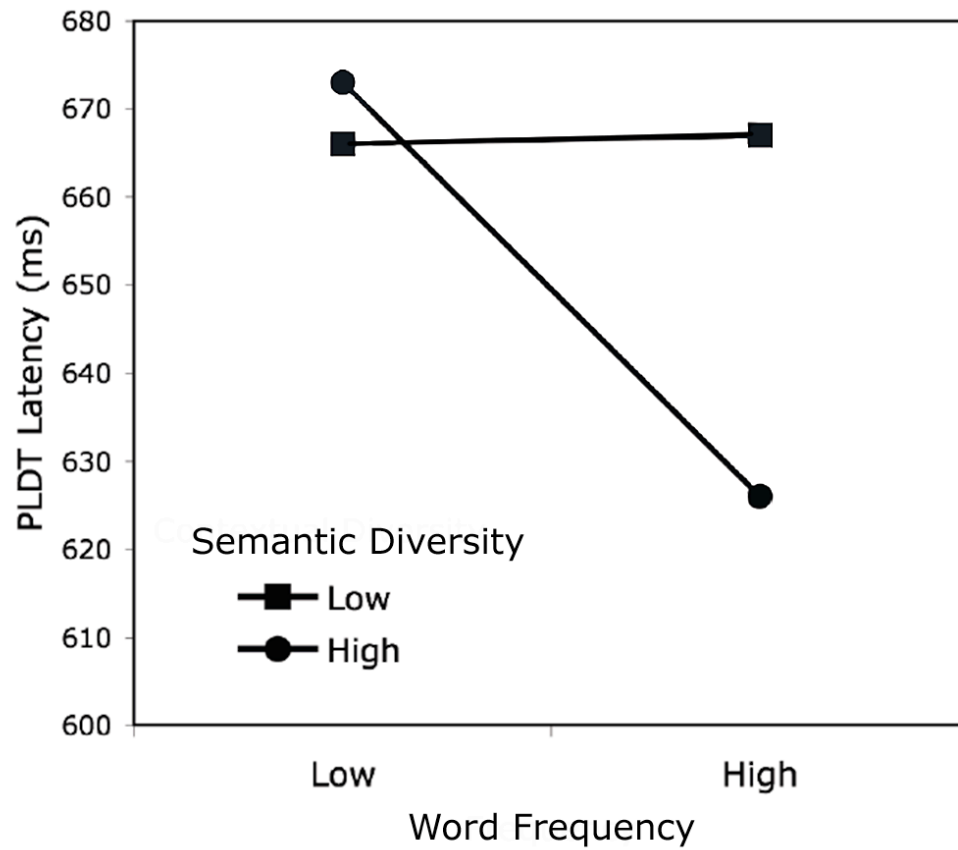


Figure 4. Pseudo-lexical-decision latency in the Xaelon task as a function of token repetition frequency and semantic diversity. Item repetitions produced no detectable latency savings unless the repetitions were accompanied by a change in context.

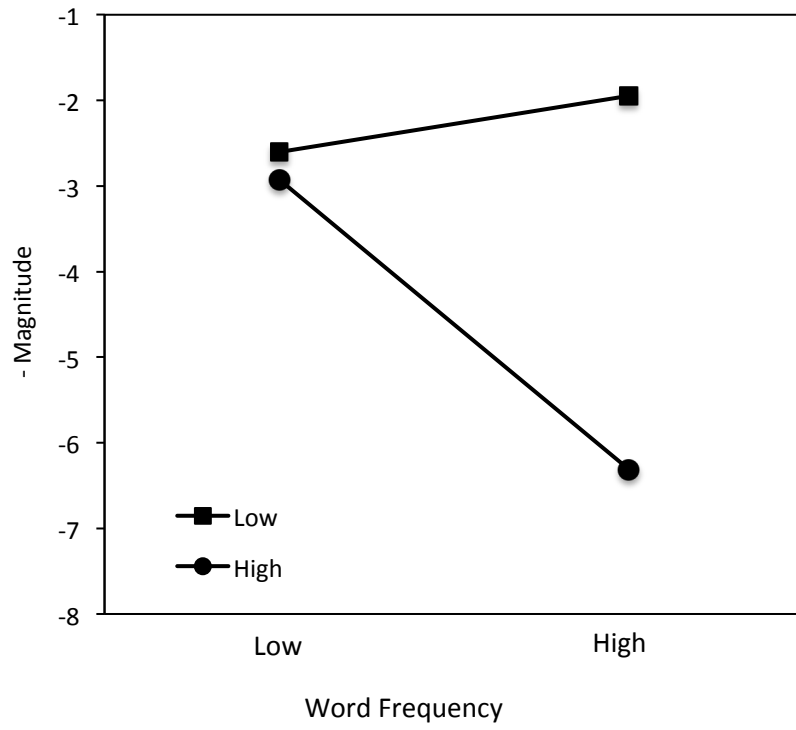


Figure 5. SDM simulation of Experiment 2 (artificial language).