CHAPTER EIGHT

Context as an Organizing Principle of the Lexicon

Michael N. Jones*,[1], Melody Dye* and Brendan T. Johns[§]
*Indiana University, Bloomington, IN, United States
[§]University at Buffalo, Buffalo, NY, United States
[1]Corresponding author: E-mail: jonesmn@indiana.edu

Contents

Abstract

Classic accounts of lexical organization posit that humans are sensitive to environmental frequency, suggesting a mechanism for word learning based on repetition. However, a recent spate of evidence has revealed that it is not simply frequency but the diversity and distinctiveness of contexts in which a word occurs that drives lexical organization. This chapter provides an in-depth evaluation of new research on contextual diversity, integrating evidence from word recognition, semantic memory, episodic memory, and information retrieval. The aggregate evidence suggests an expectancy—congruency learning mechanism that updates lexical representations based on the fit between the current contents of memory and the

Psychology of Learning and Motivation, Volume 67
ISSN 0079-7421
http://dx.doi.org/10.1016/bs.plm.2017.03.008

information in the local environmental context. This learning mechanism produces a well-structured lexicon that is adapted to the demands of lexical retrieval and processing.

1. INTRODUCTION

Learning to organize the mental lexicon is one of the most important cognitive functions across development, laying the fundamental structure for future semantic learning and communicative behavior. While there is a considerable amount of variance in linguistic experience across individuals, we must all nonetheless converge on a sufficiently similar lexical organization to successfully communicate and interact with each other. How we arrive at this shared mental organization from our diverse set of experiences has been an active area of research for over 60 years. Part of the similarity across individuals likely owes to the fact that we are all born with the same domain-general learning mechanisms. However, it is also necessary for us to learn from the same sources of statistical information in the environment (Thompson-Schill, Ramscar, & Chrysikou, 2009). Cognitive mechanisms have thus coevolved with environmental structure.

Given our interest in the organization of the lexicon, this chapter focuses on the word as a molar unit. We further delimit our focus to the semantic and contextual components that underpin lexical organization. While orthographic and phonological features are extremely important in both word recognition and lexical organization, they have been studied extensively elsewhere (Adelman et al., 2014; Adelman, Sabatos-DeVito, Marquis, & Estes, 2014; see Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001 for a review). Behavioral research has made clear that the lexicon imposes a minimum of two types of organization: (1) words that are more commonly encountered are more broadly accessible and (2) words that are semantically similar are close to each other (or well connected) in mental space. Although the two types of structure are likely related, the literature on lexical access and lexical similarity tend to stand distinct from one another due to differences in the tasks used to assess them.

Lexical access is most commonly studied using single-word identification latency tasks, such as lexical decision or word naming (Balota et al., 2007). Lexical access is based on characteristics of the word in isolation (e.g., *dog* is identified faster than *hog* given no additional context). In contrast, **lexical similarity** depends on the relation between two or

more words and is more commonly studied using paired tasks, such as semantic priming (Meyer & Schvaneveldt, 1971), free association (Nelson, McEvoy, & Schreiber, 2004), or semantic similarity judgments (Miller & Charles, 1991). For example, *dog* is identified faster when primed with a semantically related word such as *cat* or *leash* than when primed with an unrelated word such as *table*. To study lexical similarity among groups of words, tasks such as category fluency (Troyer, Moscovitch, & Winocur, 1997; e.g., "name all the animals you can in a minute") and subjective organization (Bousfield, 1953) are often used. Formal models exist for both lexical access and lexical similarity, but there is limited cross talk between the literatures. This is an issue because recent research suggests that access and similarity are both based on the same basic statistical structure in the linguistic environment, and there may be important insights from integrating across perspectives.

If you were to build a lexicon that maximized these two types of organization, what information would you use to fill it up? This is essentially the problem faced by a parent teaching words to a child. Do you keep repeating a word until it is obvious that the child comprehends or can produce it? Do you then frequently pair it with a related word until the relation is clearly acquired? Is it important to frequently pair the word with its perceptual referent if it has one? Or would you opt instead to use the word in multiple distinct contexts to resolve ambiguity? Landing on the "right" solution to this problem requires that you know how the learning mechanism operates, and this is one of our field's biggest mysteries. But clues to understanding the mechanism can be inferred from correlations between behavioral patterns and environmental patterns in language use (cf. Anderson & Schooler, 1991). Determining what information the mechanism has evolved to operate over can influence everything from how we teach children in the classroom to how Internet search engines find information for us.

Word frequency is a beloved variable in cognitive psychology. It is an excellent predictor of a word's age of acquisition, speed of identification, and likelihood of being recognized or recalled from a list. It is also a fundamental source of information underlying most models of semantic representation. The effectiveness of word frequency as a predictor of many behavioral variables has made the **principle of repetition** ubiquitous in psychology. Modern strength-based and multiple-trace models are all based on the premise that repetition of a word increases its strength or availability in memory, which produces benefits to both lexical access and

lexical similarity. Repetition is core to most theoretical accounts of learning, and is a fundamental principle to the field of education.

However, a spate of recent findings has called into question the role of frequency information in lexical organization, pointing instead to **contextual diversity** (CD) as the primary source of information humans use to organize their lexicons. CD can be a slippery construct to define, and takes on slightly different operational definitions across the handful of experiments that have studied it. However, CD is generally conceptualized as the number of distinct contexts in which the word occurs. If frequency is based on the principle of repetition, then CD is based on the **principle of likely need** emphasized by rational models of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991): A word that has been experienced in many contexts during learning is more likely to be needed in an unknown future context, making it more broadly accessible in the lexicon.

Importantly, the principle of repetition and the principle of likely need make different empirical predictions. If humans are sensitive to *word frequency*, then repeating a word should be beneficial to later identification of that word and should increase the similarity of related words with each repetition. If, on the other hand, humans rely on *contextual diversity*, then repeating a word should be of limited use if the repetition is not also accompanied by a modulation in context, and repetitions within the same context should be largely irrelevant to lexical organization. While frequency and CD are highly correlated variables, they suggest substantively different learning mechanisms.

Consider the simplified example in Fig. 1. Five distinct contexts are presented, with each letter representing a word within the context. Although the variables of frequency and CD are (and will always be) highly correlated, there are different item-specific predictions under the principles of repetition and likely need. For example, while words A and B both occur in the same number of contexts, the frequency of A is two times that of B due to its many repetitions within contexts. Under the principle of repetition, A should be more available in memory because it has a stronger trace than B. Under the principle of likely need, however, A and B should have equal strengths in memory because the multiple repetitions of A were not accompanied by a change in context—thus, both are equally expected in an unknown future context. The opposite is the case for words B and D: Since they have occurred the same number

Context	Content		WF	CD
1:	A B C	A:	10	5
2:	A A A B	B:	5	5
3:	A A B C	C:	2	2
4:	A A A B	D:	5	1
5:	A B D D D D D			

max(CD) = WF; *min*(WF) = CD
If a word occurs once per context, CD = WF

Under Principle of Repetition A > B; B = D; C < D

Under Principle of Likely Need: A = B; B > D; C > D

Figure 1 A simplified sample of experience to contrast word frequency (WF) and contextual diversity (CD). Five contexts are presented, with each letter representing a word within the context. If each word occurred once within each context, WF = CD, and this would be the maximum value CD could take on. If a word is repeated within contexts, WF > CD. Although the two variables are highly correlated, the principles of repetition and likely need make different item-specific predictions about the memory strength of words in these context arrays (see text).

of times overall, the principle of repetition sees them as equal in strength and availability. But under the principle of likely need, B should be stronger in memory than D, and should also be more available, since B's occurrences were all in distinct contexts, whereas D was repeated all within a single context—thus, B is more likely to be encountered in an unknown future context.

The principle of likely need aligns well with modern distributional models of semantics, which construct a word's meaning from the other words that it cooccurs with over contexts. From a distributional perspective, the multiple presentations of word D are redundant signals to its meaning. By comparison, the presentations of word B, though equal in number, offer more distributional variance to infer its meaning(s) (see Jones, Willits, & Dennis, 2015; McDonald & Shillcock, 2001; Redington, Chater, & Finch, 1998).

The benefit of CD over word frequency in lexical access dates back at least to seminal corpus work by McDonald and Shillcock (2001). But the most influential and well-cited study documenting the CD superiority was conducted by Adelman, Brown, and Quesada (2006). Adelman et al. computed word frequency and CD for thousands of words from two large text corpora, operationalizing CD as the number of documents in which a word occurred. A regression predicting lexical decision time and

word-naming time from Balota et al.'s (2007) English Lexicon Project database found clear evidence of the superiority of CD over frequency: CD predicted all variance in latency data that frequency did, and additional unique variance. The number of different documents that a word occurs in was a better predictor of its recognition latency than was a raw count of the number of times it occurred.

From these data, Adelman et al. argued that previous theories of lexical access had been constructed based on a false assumption that humans use frequency information to organize memory. Their work suggests that many current models either need to be abandoned or revised to adequately explain how the lexicon is organized: Frequency effects are only observed because frequency and diversity are highly correlated. Contextual diversity is the more likely causal factor in lexical organization.

Many of the studies that have found a benefit of CD over frequency have relied on statistical regression to demonstrate that CD can account for a greater amount of variance in human behavioral data. Although statistically significant, in most cases this benefit in fit is rather small, often in the order of a 1% improvement. However, while the predictive benefit of CD over frequency is small, it is also decisive, as it indicates that humans behave consistently with the principle of likely need rather than the principle of repetition. This distinction has wide-reaching consequences for our understanding of lexical learning and organization, as CD signals a radically different learning architecture than has heretofore been assumed that has huge consequences for our understanding of lexical organization and learning.[1] The small improvement of fit for CD over frequency suggests that humans are consistent with the principle of likely need rather than the principle of repetition—these support fundamentally different learning mechanisms. So, the rather modest improvement of fit using one variable over the other is not simply a statistical exercise in using predictive variables: It has very important consequences to how we teach children vocabulary and meaning, which are fundamental building blocks of cognition.

[1] This is similar to the debate on exponential versus power laws of practice and forgetting (e.g., Heathcote et al., 2000). Distinguishing between an exponential and power function can usually only be done in the tail of the distributions, which requires a large amount of data and usually only results in a modest improvement in fit of one model over the other. But this tiny difference in fit is huge on a theoretical scale—the two functions are produced by fundamentally different theories of learning and forgetting (see Farrell & Lewandowsky, 2015).

2. WORD FREQUENCY AND CONTEXTUAL DIVERSITY

In 1885, Ebbinghaus ushered in the modern science of memory with his publication of Über das Gedächtnis ("On Memory"). That seminal work, which would introduce to psychology foundational concepts such as learning and forgetting curves, was based on Ebbinghaus' experimentation with nonsense syllables. In the days since those pioneering studies, much of the research into human cognition has involved verbal stimuli. Today, words are still the default choice for studies of recognition and recall, priming, fluency, and paired associate learning, to name but a few. Accordingly, most major cognitive models have been assessed by their ability to fit data on verbal learning and remembering (see Johns & Jones, 2010). Of course, the goal of cognitive research is to extrapolate beyond words and letters, to general principles of encoding, storage, and retrieval, and to the organization of knowledge, more broadly. Nevertheless, it is uncontroversial that language has traditionally been the primary environmental input to subjects and models, and likewise, to theory construction (Monsell, 1991).

The importance of that fact, and what it implies, has been subject to some debate. Some 70 years after Ebbinghaus, Estes (1955) urged the field to shift "the burden of explanation from hypothesized processes in the organism to statistical properties of environmental events" (p. 145). That injunction was later refined by Simon (1969), who articulated a vision of scientific practice in which both the structure of information in the environment, and the known limits of the human mind, offered constraints on cognitive theorizing. These prescriptions ought to be of particular interest to present-day researchers, as recent innovations in corpus creation and statistical modeling have made it possible to quantify the lexical and distributional properties of words and texts in ways previously unimaginable.

In the study of how words are represented in mind, considerable attention has been paid both to cataloging the manifold dimensions on which words vary, and to characterizing these differences and their measurable behavioral effects (Balota et al., 2007; Rubin, 1980; Whaley, 1978). These lexical and semantic variables include both subjective measures, based on intuition—such as familiarity (Gernsbacher, 1984), concreteness (Paivio, 1971), and age of acquisition (Carroll & White, 1973)—and objective measures derived from large-scale language corpora and dictionaries, such

as word length (Forester & Chambers, 1973) and frequency (Howes & Solomon, 1951).

These two approaches have different strengths, depending on the behavior under examination. While objective measures tend to benefit from greater measurement precision and replicability, there are significant statistical challenges and sampling biases to contend with (Gernsbacher, 1984; Lovelace, 1988). Moreover, objectively derived counts necessarily yield a population-level picture, papering over individual differences (Gardner, Rothkopf, Lapan, & Lafferty, 1987; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). On the other hand, whereas introspective ratings offer a more complete picture of individual experience, they are more difficult to collect at scale, may vary significantly depending on the precise set of instructions, and may be contaminated by other lexical variables (Balota, Pilotti, & Cortese, 2001).

Decades of study on lexical processing has yielded dozens of variables of both stripes, many of which are highly intercorrelated. Of these, **word frequency** is likely the most studied and the most robust in its effects. Frequency assesses a given word's occurrence in the broader linguistic environment, relative to other words. In psychological experimentation, it is often taken as a proxy for a subject's experience with that word. Frequencies are typically estimated on the basis of their occurrence in a corpus, which is a large, curated collection of texts, designed to be representative of the language as a whole (e.g., Baayen, Piepenbrock, & Gulikers, 1995; Davies, 2009). Given variability in corpus size, frequencies are often reported in terms of frequency per million words (van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

That frequency is a principle variable in how words are processed has been recognized since the dawn of psycholinguistics (Howes & Solomon, 1951). Frequency is a well-established predictor of performance in a range of lexical and semantic tasks, with frequency differences detectable across an array of chronometric measures, including response latencies, eye movements, and patterns of neural response and activation. In many perceptual and production tasks, high frequency words are markedly advantaged, posting both faster reaction times and higher accuracy in tests of lexical decision (Scarborough, Cortese, & Scarborough, 1977), naming (Broadbent, 1967; Forster & Chambers, 1973), and perceptual identification (Morton, 1969). In episodic tasks, the story is more complicated, with the standard finding being that high frequency words are better recalled in pure lists (Deese, 1960), but low frequency words are better recalled

in mixed lists (Gregg, 1976) and better recognized (Gorman, 1961; Kinsbourne & George, 1974). When processing words in sentences, higher frequency benefits speed of processing in both first-pass and later eye movements (Inhoff & Rayner, 1986; Rayner & Duffy, 1986; see Plummer, Perea, & Rayner, 2014 for a review).

In line with findings such as these, usage-based theories of language acquisition and processing have focused on the frequency of items and constructions (Goldberg, 2006; Tomasello, 2003), leading to the development of exemplar models based on repetition (Bybee, 2006). Likewise, many computational models of visual word recognition incorporate the principle of repetition, often implicitly. Consistent with Hebbian learning rules, the underlying assumption is that the more times an item has been encountered, the more easily it will subsequently be processed or retrieved. This effect of repetition is typically formalized as a counter in the head, which biases detection of items according to their frequency, either by raising their baseline level of activation (Coltheart et al., 2001) or by lowering their resting state threshold (Morton, 1969). Alternately, in serial search system models, the lexicon is organized by rank frequency, such that words with a higher relative frequency are accessed more quickly and efficiently (Murray & Forster, 2004; Rubenstein, Garfield, & Millikan, 1970; but see Adelman & Brown, 2008).

It seems undeniable that frequency plays a role in the accumulation of knowledge (Gardner et al., 1987; Shiffrin & Nelson, 2013), and its effects are well attested in many cognitive and perceptual domains. What is less clear, however, is the extent of that contribution. There is, for one, the **power law of practice** to consider: While the time it takes to perform a skill—such as recognizing a word, or typing it—tends to diminish in line with the number of practice trials, the rate of improvement is negatively accelerated, decreasing rapidly over repetitions (but see Heathcote, Brown, & Mewhort, 2000). This relationship between practice and performance is ubiquitous (Newell & Rosenbloom, 1981) and suggests that the principle benefits of repetition may be attained within a relatively short span, with modest benefits thereafter (Salasoo, Shiffrin, & Feustel, 1985). In line with this, low frequency words benefit considerably more from additional study in tests of verbal memory than do their high frequency counterparts (Scarborough et al., 1977).

At the same time, a word's frequency is confounded with a number of key environmental variables, including how recently it was experienced (Scarborough et al., 1977) and in how many different contexts (Church &

Gale, 1995a, 1995b) differences that may have important consequences for processing and retrieval. While there is now a considerable body of knowledge about the structure and properties of the lexicon, there is still significant work to be done in disentangling the many intercorrelated environmental variables that may contribute to lexical processing.

Thus, while frequency is a clear precondition to learning and maintaining a behavior, the mechanism by which its effects are achieved is still an open question. In recent years, this question has come to the fore in a debate over the differential contributions of context and repetition to how an item is encoded and remembered (Adelman et al., 2006; McDonald & Shillcock, 2001). This debate gets to the heart of the question of how the mental lexicon is organized, and how the patterns of events that comprise our experience of the world—and the mental processes that operate over them—give rise to this organization.

2.1 Measuring Word Frequency

In tasks involving verbal stimuli, what is typically desired is a model of the "average" speaker's experience with the language. Arriving at such a model is a problem of statistical inference, which concerns how to generalize from a finite sample of text—a corpus—to the hypothetically infinite set of texts that comprise the language as a whole. As with other domains, the object of inquiry is rarely limited to the specifics of the sample under consideration. Rather that sample is taken to be representative of the language more broadly.

For instance, a linguist working with the Brown Corpus might hope to draw conclusions about the use of indirect relative clauses in American English. Similarly, a cognitive psychologist working with a list of word frequencies might wish to investigate the effect of prior knowledge on recognition memory. Neither the linguist nor the psychologist is interested in the Brown Corpus per se. What they are invested in is the model it offers of linguistic experience, which allows predictions to be made about usage or processing.

On the surface, the measurement of word frequency seems to be relatively uncontroversial: Simply count how many times a word occurs in a large sample of language, and this integer provides an estimate of the true parameter. The larger the linguistic sample the count was computed from, the better the estimate. However, even this simple notion of word frequency is more complex under closer inspection (see Baayen, Milin, & Ramscar, 2016). For example, an extremely large text corpus is likely to

yield a count that is closer to the word's "true" relative frequency than is a count from a small corpus. But no single human ever experiences such a large sample of language, so if the aim is to predict human behavioral data, the true environmental parameter is largely irrelevant. Thus, while using the largest text sample possible is, statistically speaking, the best choice for estimating the word's relative frequency in the environment, psychologists are faced with a somewhat different problem—that of determining how large a sample is a reasonable estimate of the experience of the experimental subjects whose data is being modeled.

The type of linguistic experience that frequencies are tabulated from is also very important: The count unit can be specified in different ways—e.g., the n-gram, the word, the stem, or the syllable—and the count yields can vary significantly by the corpus under consideration—different genres and registers can yield very different counts. Recent corpus-based modeling has found, for instance, that frequencies from written text are actually a poorer estimate of visual word recognition than frequencies from television subtitles (van Heuven et al., 2014).

An even thornier issue concerns the statistical assumptions that underpin frequency counts. Traditionally, word frequency distributions in psychology and linguistics have been understood in terms of a random sample model. These belong to the class of "generative" models, which attempt to specify the underlying stochastic process by which some observable data—a document or graph, say—were generated. The random sample model makes the simplifying assumption that a given document (or set of documents) can be treated as a bag of words, with little interesting structure. Statistically, this translates to the idea that the words in the document were generated by a single, fixed parameter distribution, such as the Poisson or Binomial. The assumption, then, is that words are drawn independently, and at random, from some infinite English text, and that a given word's variance over documents is no greater than its mean (Church & Gale, 1995a, 1995b). The theoretical implication is that words and documents are, in a sense, exchangeable, and that context can be ignored.

The problem with these assumptions is that they do not accord well with actual linguistic data (Baayen, 2009). Indeed, many of the conclusions that follow from them are plainly incorrect—for instance, just because "the definite article *the* accounts for roughly one in 17 words in English" does not imply that "the ungrammatical sequence *the the* should occur about once in every 300 words" (Evert, 2006, p. 177). When fit to real language, the random sample model makes a number of

glaring errors. These errors derive from the premise that the likelihood of a word is independent of context, and is, therefore, uniform across documents. On this view, word frequency is simply a fractal property that scales with text size. While word frequency distributions are invariant in precisely this way, the frequency of an individual word is not—and the distinction is critical.

The fractal nature of linguistic distributions is described by Zipf's law, which specifies that in any given text, the relation between word rank and word frequency is governed by an inverse power law (Zipf, 1949). This is a fundamental regularity of linguistic structure, and it is universal, found in texts that span authors, genres, and languages (Wyllys, 1981). However, the shape of the aggregate distribution—which holds stable—should not be identified with the particular words that comprise it—which do not. The systematic patterns of correlation and cooccurrence that exist within texts, are "destroyed when the words in a collection are reshuffled, even while the global word frequencies are preserved" (Serrano, Flammini, & Menczer, 2009, p. 1).

Compare, for instance, a list of the top content words in an anthology of romantic poetry to those of an electrical engineering textbook. There is little overlap. The discrepancy stems from the fact that a word's likelihood is highly context dependent, and varies according to a wide variety of factors—including sentence-level syntactic and semantic constraints, and principles of discourse organization, coherence, and style (Baayen, 2009; Church & Gale, 1995a, 1995b). On examination, only the most frequent 1% of word types are uniformly distributed across texts, in compliance with the random sample model. The occurrence of words of average or lesser frequency—which account for fully 99% of the English lexicon—appears to be conditioned on hidden contextual variables, in violation of the independence assumption (Church & Gale, 1995a, 1995b).

In short, frequency measures are based on raw counts over an entire corpus, and thus collapse the distinction between words that occur uniformly across many texts and words that occur many times in the space of a single text. However, there is reason to suspect that words that vary along this dimension may not only play different roles in discourse, but may also be processed differently. This raises an important question for models of lexical processing: Should a word that occurs n times across many documents, and a word that occurs n times within a single document be classed similarly? Psychologists have devised measures of contextual diversity to assess if and how these different usage patterns affect processing.

2.2 What is a Context?

At first glance, a word's CD also seems fairly simple to compute: Each time a word occurs in a new context, increment its counter by one. For example, if a corpus is split into discrete documents (say, articles), and we assume that each document is a new context (cf., Landauer & Dumais, 1997), then a word's frequency count is incremented each time it occurs in the corpus, but its CD is only incremented each time it occurs in a new document. Once the theorist has defined the contextual unit, CD is just as efficient to tabulate as word frequency. However, CD has significantly more difficulties in operational definition than frequency due to the many and varied theoretical views on what constitutes a context.

Are multiple adjacent paragraphs really distinct contexts? Are two paragraphs from the same article really from the same context? What about speech, time, and the physical environment? How humans dice up their experience into discrete contextual units is of fundamental importance to understanding a mechanism based on CD. While most theorists agree on how to count word frequency, there is much less consensus on how to count contexts. The validity and predictive accuracy of CD depends on the right operational definition of context when computing the variable. We can get insights into what the right definition of context is by looking at prior theory and experimentation in episodic memory. But the clues can also point in the other direction: If CD computed from a particular operational definition of context gives a better account of behavioral data in lexical access and similarity, this reinforces both the mechanistic account and the original theoretical construct that the notion of context was borrowed from.

Psychological notions of context vary widely (e.g., Murnane, Phelps, & Malmberg, 1999; Smith, Glenberg, & Bjork, 1978). In determining the role of context in semantic organization, it is useful to consult classic memory research for clues about what information might be considered bite-sized context units by our cognitive mechanisms. In episodic memory research, for example, it is very common to present lists of randomly selected words to probe the operations of encoding, storage, and retrieval. In this case, context is usually specified as the list in which a word was presented—if two words were on the same list, they were presented in the same context. This is in a similar vein to how Adelman et al. (2006) defined contexts as groups of words clustered by documents—if two words are in the same document, they are in the same context. But even the

classic list learning literature in episodic memory debates what exactly "list context" means (Dennis & Humphreys, 2001; Howard & Kahana, 1999). Is context defined by the other items on the list, the spatial proximity of items at encoding, or the fact that they occurred together as a temporal cluster during learning?

Another group of context effects in memory research has focused on the change in performance when information is studied in one context and retrieved in a different context (e.g., McGeoch, 1932). The classic example is learning in one environment and being tested in another, such as different rooms (Abernethy, 1940; Smith, 1979) or on land versus underwater (Godden & Baddeley, 1975). Studies such as these highlight the association between the information being studied and incidental aspects of the physical environment.

Wickens (1987) differentiated between two types of context, which he called Context Alpha and Context Beta. Context Alpha refers to the environmental surroundings in which an event is experienced, but the context itself does not influence or relate to the event in any meaningful way. Examples would be the cage in which a rat is conditioned, or the classroom in which students learn materials. In contrast, Context Beta is defined by Wickens as "the situation in which one stimulus event combines with another stimulus event to define the correct response or meaning of the event" (p. 146). An example would be the tone to which a rat is conditioned to respond, or the textbook materials that students are encoding.

While useful concepts, Contexts Alpha and Beta are difficult definitions to put to use in word learning. Firstly, Wickens' (1987) principle confounds the **unit** of context and the **content** of context. For models of word learning, the content of a context is usually thought of as the other words with which a word occurs, following from Firth's (1957) famous "You shall know a word by the company it keeps." But the unit across which the company of a target word is computed is still open to interpretation. Wicken's concepts are not completely clear about how far the range of time or space should be considered until a context changes, and to be fair, this is an overextension of what Contexts Alpha and Beta were created for. But most distributional models of word learning also see a word's meaning as an emergent property from statistical redundancies in the environment (see Jones et al., 2015; for a review). Hence, distributional models are formalizations of how incidental context becomes meaningful context after many cooccurrences—the inference process is essentially a continuous shift from a context being Alpha to Beta. The "meaningful"

context was at one point in learning an incidental context. As Nelson and Shiffrin (2013, p. 357) point out, "the way episodic (or event) memories are encoded depends on the knowledge (or semantic memory) of the individual who is encoding them. Conversely, an individual's knowledge must be formed through the episodes they encounter."

In addition to Wickens' (1987) constructs, a broad scan of the memory literature yields three main types of context that have been extensively studied, which can be loosely categorized as time, space, and content. The temporal context in which an item occurs is known to have a profound effect on encoding and retrieval, independent of other factors (e.g., Estes, 1955; Hintzman, Block, & Summers, 1973; Howard & Kahana, 2002). Temporal context is a core principle to many models of semantic learning (e.g., Elman, 1990; Howard, Shankar, & Jagadisan, 2010), thought to be generated by oscillations in hippocampal projections during encoding (Howard, Fotedar, Datey, & Hasselmo, 2005; Polyn & Kahana, 2007). Spatial context (Ekstrom, Arnold, & Iaria, 2014) and scene context (Oliva & Torralba, 2007) are also fundamentally important in encoding and recall, although it may well be the case that the brain encodes spatial context as temporal context (see Miller et al., 2013).

2.3 Measuring Contextual Diversity

Contextual diversity can be measured in a variety of ways depending on the theorist's definition of context. Of these measures, some are subjective, relying on the intuitive assessments of raters—such as context availability, which measures the perceived ease of imagining a context in which a word might appear (Schwanenflugel, Harnishfeger, & Stowe, 1988; Schwanenflugel & Shoben, 1983)—or on the skilled judgment of lexicographers—such as ambiguity/polysemy, which indexes the variability in a word's meaning (Jastrzembski, 1981). Other objective measures are based on analyses of the distributional pattern of words across documents or other verbal contexts.

In text-based analyses, the most straightforward measure of diversity is a raw count of the number of distinct documents in which a given word occurs (Adelman et al., 2006), a measure also known as its dispersion (Baayen, 1996, pp. 17–31; Gries, 2008). Other, more complex measures of diversity have been developed, which compute—for example—the semantic diversity of contexts in which a word occurs (Hoffman, Ralph, & Rogers, 2013; Jones, Johns, & Recchia, 2012), the information a word conveys about its contexts of use (McDonald & Shillcock, 2001), and the

distribution of temporal scales over which a word occurs (Altmann, Pierrehumbert, & Motter, 2009).

Underpinning these disparate measures is the observation that words vary in the extent to which they deviate from the random sample model, and hence, from the presumption of uniformity across contexts. The greater the deviation from what would be expected by chance, the more usage is characterized by "burstiness"—quick bursts of activity, in which a sharp uptick in mention in a particular context is followed by a steep decline to baseline levels (Barabási, 2005). While words differ substantively in this kind of clustering behavior, fully 99% of the lexicon shows some deviation from Poisson (Church & Gale, 1995a, 1995b). What these measures evaluate is degree.

Since a word's meaning is closely related to the contexts in which it appears (Firth, 1957; MacDonald & Ramscar, 2001), contextual diversity can also be tapped indirectly through measures of semantic richness (for a review, see Pexman, Siakaluk, & Yap, 2013). Richness has been indexed in a variety of ways, including number of senses (Jastrzembski, 1981), number of semantic neighbors or associates (Nelson et al., 2004; Paivio, Yuille, & Madigan, 1968), and number of perceptual features (McRae, Cree, Seidenberg, & McNorgan, 2005). Richness can also be extrapolated from linguistic data, and quantified in terms of the semantic diversity of contexts in which a word occurs (Hoffman et al., 2013; Jones et al., 2012), its connectivity to other words in its associative network (Rotaru, Vigliocco, & Frank, 2016; Steyvers & Tenenbaum, 2005), and the proximity of its semantic neighbors (Buchanan, Westbury, & Burgess, 2001).

Variables that index distributional and semantic richness are robust predictors of verbal learning and memory. In semantic tasks, such as naming and lexical decision, words that score higher on measures of CD and semantic richness are responded to more accurately and efficiently (Adelman & Brown, 2008; Jones et al., 2012; McDonald & Shillcock, 2001; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008), a result that has been replicated using different language materials and lexical data (e.g., Brysbaert & New, 2009), and extended to other subject groups, such as young readers (Perea, Soares, & Comesaña, 2013). Similarly, in tests of episodic memory, CD has been found to facilitate recall (Lohnas, Polyn, & Kahana, 2011), while impairing recognition (Anderson, 1974; Steyvers & Malmberg, 2003). CD has also been found to benefit learning of grammatical classes (Mintz, Newport, & Bever, 2002; Redington et al., 1998), speech perception (Pisoni & Lively, 1995), and word-referent mapping (Kachergis,

Yu, & Shiffrin, 2016; Smith & Yu, 2008), and predicts the order of lexical acquisition in early language development (Hills, Maouene, Riordan, & Smith, 2010).

3. FREQUENCY AND DIVERSITY IN LINGUISTIC CONTEXTS

3.1 Context Dispersion in Information Retrieval

While text-based measures of context are relatively new to the psychological literature, they are of longstanding importance in a variety of computing domains, such as information retrieval (IR). There, they have been of considerable interest because of their usefulness in identifying effective keywords for search. Information retrieval is a branch of computer science concerned with the problem of identifying and returning the set of files in some database that best match a user's informational needs (Manning, Raghavan, & Schütze, 2008). Electronic library catalogs, Web search engines, and Desktop file search, are all examples of IR systems, and the retrieval process is broadly similar across platforms. Typically, the system operates over an index database, which contains metadata about the potential files to be pulled. On the basis of a user-generated query, the search algorithm then combs through that index for potential matches, returning a set of hits, rank-ordered by relevance. The system's aim is to balance two potentially contradictory goals: returning all of the most pertinent items (*recall*), and minimizing the number of irrelevant items (*precision*; Salton & Buckley, 1988).

Human memory has a number of clear parallels with information retrieval (Griffiths, Steyvers, & Firl, 2007). In particular, like IR systems, humans are tasked with "managing a huge stockpile of memories" that cannot all be made available at once (Anderson & Schooner, 1991, p. 396). There is thus an analogous optimization problem. While it seems unlikely that minds solve the problem in exactly the same way as machines, the insights yielded from IR are certainly instructive, and well worth considering.

In IR research, a common rule of thumb is that a term's **frequency** f_t is inversely correlated with its usefulness as a keyword, as less common words tend to be more semantically specific (Luhn, 1957). However, frequency is not a perfect guide to lexical richness, and a better metric was soon developed: **inverse document frequency** (idf; Sparck Jones, 1972). If D is

the number of documents, and df_t is the number of documents that contain some term t, then idf can be defined simply as:

$$\mathrm{idf}(t, D) = -\log_2 \frac{df_t}{D} \quad (1)$$

idf is a document-level statistic that measures a word's relative frequency in terms of the number of distinct documents it occurs in, rather than the number of distinct tokens over the entire collection. Underpinning idf is the intuition that words that are spread more evenly across contexts exhibit distinct properties from those that cluster more densely, even when frequency is held constant (Altmann et al., 2009; Church & Gale, 1995a, 1995b). As a metric for selecting effective keywords, it penalizes words that occur promiscuously across a broad range of documents, instead prioritizing bursty words that show up selectively. Terms that score well on this measure have two advantages for retrieval: first, they dramatically reduce the search space, returning a select set of potentially relevant documents; and second, they tend to be highly informative about the documents in which they do occur.

Compare the following keyword sets: (*once*, *somewhat*, *instead*) and (*design*, *boycott*, *intelligence*). While both sets belong to the same frequency class, they have markedly different idf scores: *boycott* belongs to just a few documents, while *instead* is spread across many. Intuitively, a story that makes mention of a *boycott* is likely about boycotts, whereas a story that includes the word *instead* could pertain to almost any topic. *Boycott* is unquestionably the better keyword.

Today, many IR models make use of a composite weighting scheme, **tf-idf**, which incorporates both the term's frequency within a given document, tf(t, d) and its inverse document frequency over the entire collection of documents, idf(t, D). tf-idf can be characterized as a measure of the information in a term, weighted by its probability of occurrence (Aizawa, 2003):

$$\mathrm{tf} - \mathrm{idf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D) \quad (2)$$

tf-idf selects for terms that occur many times in a small number of documents, penalizing terms that occur only a few times in a document, or that occur in many documents. It has the advantage over IDF that it selects for words of "medium" frequency where there is the greatest deviation from Poisson and the random sample model (Church & Gale, 1995a, 1995b).

What these measures reveal is that the best keywords for information retrieval have high variance over documents, and low document entropy,

thus returning a highly select and informative set of documents. This reflects the fact that terms that score highly on these measures tend to be strongly dependent on hidden variables (such as author or discourse topic), helpfully discriminating among documents on the basis of these latent dependencies. A good rule of thumb is that the most useful keywords also tend to be those that deviate the most from what would be expected by chance.

In addition to its applications in search, tf-idf is a common metric used in term weighting in vector space models of semantics, and provides the feature space for models such as latent semantic indexing and Bayesian topic models (Blei, Ng, & Jordan, 2003). Jones et al. (2012) incorporate a principled variant on tf-idf in their computational model of lexical access, 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" (116). That so many models capitalize on these broadly similar measures, suggests that information captures something important about how humans cognize their environment.

In fact, similar observations have been made before. For example, latent semantic analysis (LSA), perhaps the most well-known distributional model of lexical semantics, subjects the raw cooccurrence data between words and documents to a series of transformations in which local term frequency is first log-transformed, and then divided by the term's entropy over documents. Landauer & Dumais (1997) are explicit in noting a connection to learning and information,[2] remarking that this inverse entropy measure, "accomplishes much the same thing as conditioning rules such as Rescorla and Wagner (1972) in that it makes the primary association better represent the informative relation between the entities rather than the mere fact that they occurred together" (216).

3.2 Cognitive Principles of Context Diversity

Information retrieval systems have clearly benefited from measures of word usage that incorporate context. Such context-sensitive measures are also useful for understanding how words pattern across texts and speech.

[2] There are clear parallels between learning and information theoretic principles. In a Shannon sense, the information in a signal can be quantified in terms of the amount of uncertainty it reduces (Shannon, 1948). Likewise, learning is a function of uncertainty reduction, and like any other signal, a cue is informative to the extent that it makes an upcoming outcome more predictable (Gallistel, 2003).

An important question is why contextual diversity matters to how a word is processed and retrieved. This question can be addressed from a number of different theoretical perspectives.

From a rational perspective, a word that has been experienced uniformly across many contexts during learning is more likely to be needed in future contexts, which should render it more broadly accessible in the lexicon (Anderson & Milson, 1989; Anderson & Schooler, 1991). Adelman et al. (2006) interpret the advantage of CD over frequency in lexical access under the principle of likely need: "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 closely related to the concept of word burstiness in information retrieval (Katz, 1996).

This **principle of likely need** derives from the notion that human memory is an adaptive system, tasked with solving a significant data access problem. Central to this problem is the question of how to organize the contents of memory such that desired information can be accessed quickly and reliably. From this vantage, memory systems behave optimally when, like a judiciously sorted index, they make available just those memory traces that are most likely to be needed in the present context. The key idea is that the accessibility of an item in memory should be strongly influenced by its history of use: Items that have been retrieved frequently, or are retrieved regularly, should be easier to access, independent of additional context.

In an influential model of this principle, Anderson and Milson (1989) define the rational objective of memory search as one of maximizing the gain associated with retrieving a target memory, while minimizing the processing costs associated with retrieving an irrelevant memory. Retrieval is thus bounded by a stopping rule, which weighs the expected gains of continuing against its costs, halting search when the costs outweigh the benefits. The specific cost-to-benefit ratio is thought to vary with the nature of the task, and with the available knowledge, time, and attention of the searcher. What is presumed invariant, however, is the ranking algorithm, which sorts traces in memory according to their relevance to a particular query. On this analysis, the challenge is to elucidate the principles that govern the dynamic organization of traces in memory.

In their model, a memory's relevance—or "likely need"—is computed as a joint function of its contextual relevance (associative strength to the

present search query) and its historical relevance (past access history). Expressed formally, the odds of a trace A being called upon is calculated in terms of both its history H_A and the individual cues to memory i, which together comprise the query Q:

$$P(A|H_A \,\&\, Q) = P(A|H_A) * \prod_{i \in Q} \frac{P(i|A)}{P(i)} \tag{3}$$

The first term, the history factor, reflects the past access history of the trace. The second term, the context factor, is equivalent to the associative strength between the query and the trace, and gives the product of the odds ratio of the cues' conditional probabilities. The solution relies on Bayesian estimation techniques.

In estimating the history factor, a principled way to incorporate both current and historical relevance is to assume that the rate at which a given memory trace is accessed varies with its desirability, and that desirability wanes over time. (Both initial desirability and decay rate are assumed to vary over items.) Given such a framing, one possibility is to model a trace's desirability as a Poisson process, with a monotonically decreasing decay function since last access. However, while such a model has the virtue of being simple, it is a poor fit to actual lexical access patterns. Memory retrieval does not conform well to a model in which queries are independent and equally likely at every time point. On the contrary, clustering in memory search is common (Barabási, 2005). Thus, a more useful model would differentiate between items that are accessed routinely and those that are accessed en masse. To this effect, the following assumptions can be added: that hidden variables produce occasional volatility in memory search, registered by rapid upswings in access for a particular memory trace, and that items of the "volatile" type should be assigned faster decay rates than their more consistently accessed counterparts. Incorporating these assumptions produces a model with much better fit to behavioral data.

There are a number of important lessons to be drawn from this work. One is that a rational model of memory search must take stock of not just frequency and recency, but also the intervals over which a trace has been accessed (see also Howard & Kahana, 2002; Nelson & McEvoy, 2000). Another is that memory is finely tuned to the statistical structure of the environment: An item's accessibility in the lexicon depends on the

prevalence and distribution of cues that elicit its retrieval (Anderson & Schooler, 1991).

The principle of likely need is not meant to supplant mechanistic frameworks, but rather to show why the design they implement is rational. Like rational models, mechanistic accounts of learning and memory predict that differences in a word's contextual spread will affect processing and retrieval. In that literature, a wealth of findings suggest that information tends to be encoded to the extent that it is novel and surprising, rather than redundant with past expectations (Jamieson, Crump, & Hannah, 2012; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010). Learning is a means of reducing uncertainty about the world (Rescorla & Wagner, 1972) that selectively highlights reliable temporal and logical contingencies between events, and downplays spurious correlations, building a rich, relational representation of the environment. This is accomplished by weighting the most informative predictors to relevant outcomes, while eliminating redundant or potentially misleading cues (see Rescorla, 1988; Siegel & Allan, 1996).

In considering a document context in these terms, the upper bound on uncertainty over which word might occur next is described by a Poisson random rate process, wherein the unfolding context provides absolutely no information about which word will follow. It is deviations from Poisson that are informative and that will come to be associated with a given context (Gallistel, 2003). Another way to frame the same idea is in terms of the encoding specificity principle, which proposes that the likelihood of retrieving a specific memory depends on the overlap between the retrieval cue and prior traces laid down in memory (Tulving & Thomson, 1973). When a word has been experienced in many different contexts, it will be associated with many more traces, which will have the effect of diminishing its association with any particular one (Anderson, 1974).

4. THE ROLE OF SEMANTIC DIVERSITY

Most operational definitions of CD simply count the number of documents (e.g., articles) in which a word occurs in a large corpus, ignoring repetitions within the document. Operationalizing a context as a document bears similarity to classic notions of list context in episodic memory, or notions of temporal context. But it ignores the *content* of those contexts, which in the case of semantic models **is** the context. For example, if the

same document is repeated many times within a corpus, should the multiple occurrences be considered different semantic contexts of the word? If the term *Clinton* occurs in the context of an FBI email probe over 20 times in a day, are these different contextual uses of *Clinton*, or repetitions of the same context? We experience patterns like this daily, as popular discourse topics in the news dominate our attention, or as the same conversations take place in the same spatial contexts at different times. How should these similar but separate encounters be registered? This is really an empirical question to ask of the psychological mechanism. The principle of likely need is agnostic about this issue; it requires a theorist to first specify what exactly is meant by context.

4.1 Measuring Semantic Diversity

Jones et al. (2012; see also Johns & Jones, 2008) explored the distinction between context as a document and context as the *content* of the document by creating a graded measure of CD referred to as **semantic diversity**. (A very similar metric was independently created and evaluated by Hoffman et al., 2013.) Semantic diversity considers the information overlap in documents when computing a word's CD as a metric of word overlap. With this metric, multiple encounters of the word in the same (or highly similar) documents do not increase its likely need much more than a single encounter, whereas multiple encounters of the word in highly variable contexts significantly increase its likely need.[3] It is the linguistic content of the contexts that matters. The key prediction of semantic diversity is that the repetition of a word in distinct documents will increase its likely need to a greater extent than an equal number of repetitions in redundant documents.

Jones et al. (2012) used the same materials as originally used by Adelman et al. (2006) to demonstrate the superiority of a document count over raw frequency when accounting for human lexical decision and naming times in the English Lexicon Project database (Balota et al., 2007). In addition, they found that the more nuanced semantic diversity measure encapsulated all variance predicted by both frequency and document count, and an impressive amount of additional variance. That is, a measure of CD that considers the information overlap when weighing the number of contexts a word has occurred in yields an improved approximation of what humans

[3] Semantic diversity has a tuning parameter, fit to behavioral data, that controls to extent to which information overlap in the contexts affects a word's CD.

do in speeded word recognition tasks. Consistent with these results, Recchia, Johns, and Jones (2008) found that semantic diversity confers a distinct processing advantage: When document count is held constant, words that occur in documents that have a lower word overlap with each other tend to be identified ~200 ms faster in lexical decision.

There is now an impressive body of work attesting to the superiority of semantic diversity as a predictive measure over document count or raw frequency. This advantage is not exclusive to visual word recognition: When subjects are tasked with identifying spoken words in noise, semantic diversity is the best predictor of accuracy (Johns, Gruenenfelder, Pisoni, & Jones, 2012). In the absence of additional context, words that are higher in semantic diversity tend to be better identified when the auditory signal is degraded. Thus, semantic diversity is a superior predictor of both visual (Jones et al., 2012) and auditory data (Johns et al., 2012). In addition, the predictive variance of semantic diversity is observable even when surface characteristics such as orthography and phonology are factored out (Johns et al., 2012).

Semantic diversity has also proved a useful explanatory tool for making sense of word recognition patterns in patient populations. For example, semantic diversity has been proposed as an explanation for why word frequency effects are seen in patients with semantic dementia, but not in patients with stroke aphasia (Hoffman, Rogers, & Lambon Ralph, 2011). In word recognition tasks, it has also been used to explain patterns of bilingual differences, latency differences in old and young individuals, and interactions between aging and bilingualism (Johns, Sheppard, Jones, & Taler, 2016).

To summarize, early work (e.g., Adelman et al., 2006) demonstrated the superiority of document count over raw frequency in explaining behavioral data, supporting the principle of likely need over the principle of repetition. More recent refinements to computing contextual diversity (e.g., Jones et al., 2012) offer continued support for the principle of likely need as an organizing principle of the lexicon, while clarifying that humans are sensitive to context relative to information overlap rather than some sort of contextual tag. This suggests an expectancy–congruency mechanism (formalized in Johns, Dye, & Jones, 2014) in which words are encoded more strongly if the content of their context is incongruent with what has been previously experienced and stored in memory. We expand on a formal model of this mechanism in Section 5.

4.2 Experimental Evidence

A major limitation to our interpretation of contextual diversity is that the supporting evidence thus far has been purely correlational—we lack conclusive evidence that CD is a causal force in lexical organization. The superiority of CD over frequency has primarily been demonstrated with regression analyses: CD accounts for all of the variance in behavior that frequency does, and additional unique variance when frequency is partialled out. The regression approach has been the standard approach, whether CD is measured as a document count (e.g., Adelman et al., 2006) or as a more continuous semantic diversity metric (e.g., Johns & Jones, 2008).

Yet regression analyses alone do not provide conclusive evidence for the causal role of CD (however it is measured) 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; CD and frequency are highly correlated with recency. Ambiguity, abstractness, and imageability are also confounded with document count, semantic diversity, and frequency and are difficult to tease apart. When words are selected so as to be balanced on a potential confound, a strong effect of CD is still seen (e.g., Hoffman & Woollams, 2015 controlled for the effect of imageability). But it has proven impossible to select a factorial combination of words that allow us to tease apart all confounds from the pure effect of CD.

Recent corpus-based studies have attempted to partial out the confounding variables as covariates. However, to have confidence that likely need is a more plausible organizing principle of the learning mechanism than repetition, what we really need is experimental control of the statistical structure of the language being learned.

To address this concern, Jones et al. (2012; see also Recchia et al., 2008) introduced a new artificial language learning paradigm to examine the effect of CD induced experimentally. In the experiment, subjects were sent to an alien planet to learn the language "Xaelon." On each training trial, the subject was presented with a "sentence" in Xaelon juxtaposed with an image of novel creatures that the sentence described. After having experienced hundreds of these situations, subjects were given a surprise pseudo lexical decision (PLDT) task, in which they were presented with a series of trials with a single word from Xaelon, or a foil that was not from the language, and tasked with making a rapid word/nonword judgment. The PLDT

task was selected because lexical decision was the original focus of Adelman et al.'s (2006) regression study.

To test the hypothesis that repetition of contextual occurrences produces greater latency savings for unique contexts than redundant contexts, the sentences in Xaelon were carefully constructed to have a factorial combination of frequency and CD. Low frequency words occurred 45 times each in the training examples, while high frequency words occurred 180 times. While a low CD word always appeared in the same context (same sentence with same picture), a high diversity word could appear in any one of eight distinct contexts (sentences and pictures).

The results of this experiment demonstrate a clear unique effect of CD on lexical access. Fig. 2 shows an example training exemplar (left panel) and lexical decision latency as a function of word type (right panel). As can be seen, increasing frequency produced no latency savings unless the context changed. These results cannot be due to confounds such as recency or concreteness because these either do not exist in Xaelon, or were randomized across cells.

The Jones et al. (2012) study was the first to provide empirical evidence that manipulating CD independent of frequency caused a latency savings in word recognition. When added to the many regression studies that have shown a better fit of CD over frequency, these results offer important support for the proposal that likely need is a more likely principle of lexical organization than simple repetition. It is important to note, however, that

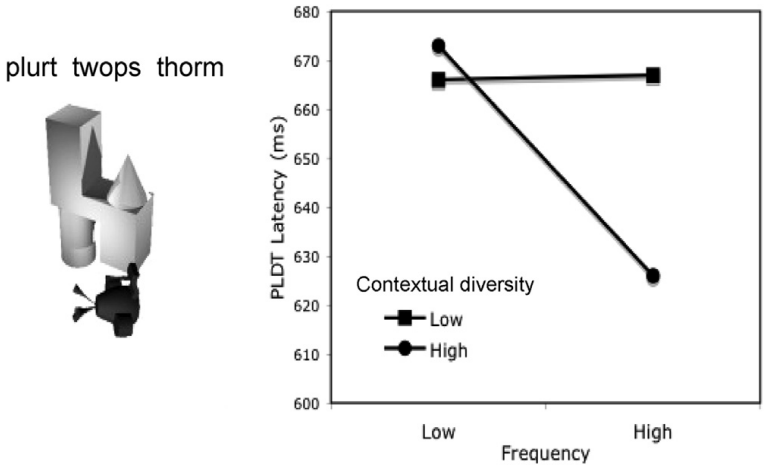

Figure 2 A sample training instance in Jones et al.'s (2012) artificial language learning experiment (left panel), and the pattern of lexical decision latency for pseudowords as a factorial combination of frequency and contextual diversity (right panel).

frequency is not irrelevant. Frequency is a necessary condition for variability to occur—a word needs to have a sufficient frequency in order to have contextual diversity. The Jones et al. results revealed that a critical number of repetitions were required for the item to become sufficiently familiar for contextual diversity to have an effect.

These findings also support semantic diversity as the best operational definition of contextual diversity. If each experience in Xaelon is thought of as a distinct document, then high frequency items that were balanced on the number of "documents" that they appeared in showed a ~60 ms benefit in processing if those documents were also semantically diverse. However, these results do not allow us to compare frequency to document count, because the two variables were purposely confounded in the design of the stimulus materials.

Johns, Dye, and Jones (2016) extended the artificial language findings from Jones et al. (2012) to natural language materials and the learning of new words from contexts of known words. Real words would have made it impossible to separate preexperimental learning from the experimental manipulation. Rather, participants were incidentally exposed to novel words as they rated short passages from books, articles, and newspapers. Unbeknownst to the participants, a core topic word in the text had been consistently replaced with a pronounceable nonword, and the contexts were varied across exposures; all other words were English words with high familiarity. After reading each passage, participants rated how well they comprehended the material.

For example, if the target word in the passage was *course*, then every occurrence of *course* in the passage was replaced by a pseudoword, such as *covella*. No participants reported noticing the manipulation—they presumably assumed the pseudoword was a term outside of their vocabulary. This simple experiment is quite similar to the cooccurrence learning situation many of us face when reading academic material: Often, there are words or acronyms we are not familiar with and we must gradually infer the meaning from multiple exposures of the term juxtaposed within a context of words that we do know. The training passages were drawn from natural real-world contexts in which the target words occurred (encyclopedia and news articles). Each target word naturally occurred in a variety of discourse contexts. For each target, two distinct sets of passages were developed: one set consisting of five passages from a single discourse topic (low variability) and the other comprising five passages spanning a number of distinct topics (high variability).

After reading 50 passages, participants in the Johns, Dye, et al. (2016) study were surprised with a PLDT task followed by a semantic similarity rating task. In the lexical decision task, participants made a speeded decision whether the presented stimulus was a word or not. Target and foil words were all pronounceable nonwords carefully selected from Balota, Burgess, Cortese, and Adams (2002) database to be balanced on orthographic and phonological characteristics. The distinction was that the half of the nonwords experienced in training had been "promoted" to the status of real words when their meanings had been inferred from context. The semantic similarity task simply presented a pair of words on the screen, and participants rated how similar in meaning the two terms were. Presented pairs of words consisted of a studied pseudoword, and a close associate of the pseudoword's target meaning. For example, if *course* was replaced with *covella*, then the participant was asked to rate the similarity in meaning of *covella* to *route*.

The results of Johns, Dye, et al. (2016) were consistent with the artificial language experiment from Jones et al. (2012): pseudowords that occurred in more diverse contexts across training had both a faster response latency and greater accuracy in the lexical decision task than words that occurred in less diverse contexts. It is important to note that no passages were ever repeated in the study. Even the low diversity condition presented distinct contexts, but the contexts were all sampled from the same discourse topic. High diversity contexts were sampled from a number of discourse topics. Hence, the variability was in semantic diversity.

Although high-CD words were identified faster, their passages actually had **lower** comprehension ratings during learning, and they were rated as **less similar** to their target word's associate in the semantic similarity judgments. These results suggest that while words were recognized better and faster when they had been encountered in variable contexts, the same manipulation lead to poorer inferences of semantic similarity and comprehension of the material (cf. Anderson's, 1974 fan effect). This dissociation between lexical access and lexical similarity closely mirrored the findings of Hoffman and Woollams (2015) with strategically sampled real words—CD speeds lexical decision, while slowing semantic relatedness judgments. The pattern is also closely related to Yap, Tan, Pexman, and Hargreaves (2011) finding that words with more senses are identified faster, but are less accurate in semantic classification.

5. INSIGHTS FROM DISTRIBUTIONAL MODELS

Contemporary distributional models of lexical semantic similarity are also based on frequency. Rather than single-token frequency, however, they depend on the cooccurrence frequency of words across a linguistic corpus. Perhaps the best known of these models is LSA (Landauer & Dumais, 1997). The learning mechanisms used by distributional models span a wide range of theoretical constructs, including dimensional reduction, reinforcement learning, Hebbian learning, and probabilistic inference (for an overview, see Jones et al., 2015). Irrespective of learning mechanism, distributional models are all based on the relative cooccurrence frequencies of words (Firth, 1957; Harris, 1970). For example, the word *milk* may frequently cooccur in the same contexts as *drink* and *cookie*. As a result, it can be inferred that these words are semantically related. Furthermore, *milk* will be similar to other words that appear in similar contexts, such as *juice* or *soda*. On the other hand, *milk* will be much less similar to *tree* because both words rarely appear in the same or similar contexts.

Distributional models have seen considerable success in accounting for a variety of semantic phenomena, including synonymy judgments (Landauer & Dumais, 1997), semantic priming (Hare, Jones, Thomson, Kelly, & McRae, 2009; Jones, Kintsch, & Mewhort, 2006), semantic categorization (Jones & Mewhort, 2007; Riordan & Jones, 2010), word association (Griffiths, Steyvers, & Tenenbaum, 2007; Sahlgren, 2006), and developmental trends (Asr, Willits, & Jones, 2016; Riordan & Jones, 2007).

Distributional models of lexical similarity also have the potential to account for lexical access. Typically, distributional 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 similar phase pattern over components, they are similar in meaning. 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 (Bullinaria & Levy, 2007; Durda & Buchanan, 2008; Shaoul & Westbury, 2010).

For example, if Murray and Forster's (2004) model of lexical access based on relative frequency is correct, this information is already contained in the vector magnitude of a cooccurrence model—it just happens to be discarded

when computing semantic similarity (much the same way that computing a correlation coefficient first standardizes each raw variable to z-scores). Any vector contains both phase (direction) and magnitude (length), but magnitude is normalized to unit length when computing a cosine similarity metric. 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.

Very little work has looked at predicting lexical access from the magnitude of a word vector within a cooccurrence model, or how these two types of structure interact in behavioral tasks. Consider the very simple cooccurrence model of Salton and McGill (1983) in which a word's representation is a raw frequency pattern over documents: Each vector element is the word's frequency within a specific document. Two words are semantically similar if their vector patterns are correlated (indicating they have a similar pattern of cooccurrence). However, a word's vector also contains its raw frequency information—simply summing vector elements indicates the marginal intensity of the word in the corpus. Fig. 3 illustrates the two types of information contained in a cooccurrence vector representation using sine waves. The top panel shows three patterns that are in phase (semantically similar) but differ in magnitude (accessibility), and the bottom panel shows three patterns that are out of phase but have the same magnitude. Current cooccurrence models learn the phase patterns but ignore the magnitude of the vector that would allow them to simultaneously explain lexical access.

Models of lexical access and models of lexical similarity are both built on the assumption that frequency is the important raw source of statistical information that humans use to organize the lexicon (the principle of repetition). However, it is equally possible for the cooccurrence vector to be based on contextual diversity rather than frequency. Rather than recording each word's frequency in the document, the matrix a distributional model is trained from could be a sparse binary matrix: if the word occurs in the document it is coded as one, and it is coded as zero otherwise (cf. Kanerva, 2009). We know from all the abovementioned work on semantic diversity that the magnitude of this vector based on CD would be a better model of human lexical access than one built on raw frequency counts. But it is unclear what a CD pattern would do to the semantic similarity structure in a distributional model. Furthermore, the vector could

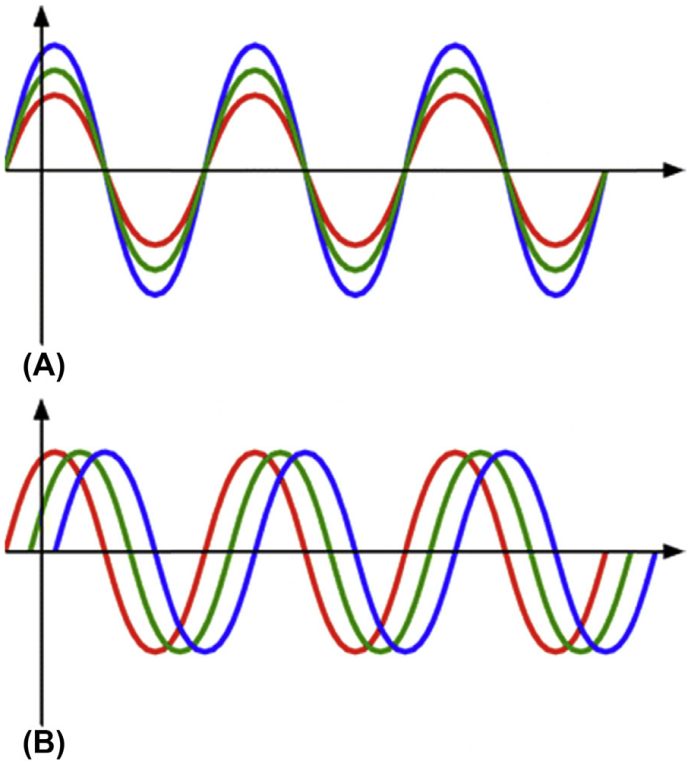

Figure 3 Phase and magnitude in a vector/signal. (A) Three patterns with the same phase but different magnitudes and (B) three patterns with the same magnitude but different phase.

be built dynamically to reflect semantic diversity, and it is possible that such a model would give a better account of both lexical access and lexical similarity from a single representation. Exploration of this idea is what leads to the creation of the semantic distinctiveness model (SDM).

5.1 The Semantic Distinctiveness Model

Johns and Jones (2008; see also Johns et al., 2014) proposed a new distributional model that would in theory be able to explain both lexical access and lexical similarity from the phase and magnitude of a single vector representation. Their SDM is based on other distributional models of semantic memory, using a word-by-context matrix representation of a text corpus. However, cell entries are not simple frequencies—the vector representation for 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. The same model framework may be used with a pure frequency or document count, allowing model comparisons from within the same formal framework.

The fundamental operation of the SDM is the use of this expectancy–congruency mechanism when building a word's semantic representation. Specifically, the encoding strength for a word in a given context is relative to the information overlap between the current environmental context and the representation of a word in memory. When a new document is encountered, a new column is added to the matrix. If a word does not occur in the document, it is assigned a value of 0 for that column. If a word does occur in that document, its expected value for the current context is computed as the sum of the semantic representations of the other words in the document.

The strength with which the word is then encoded into the new column is determined by the similarity of the current context to the word's semantic representation—the higher the similarity, the weaker the word is encoded. That is, if the semantic content of the context is redundant with previously stored information, it does not need to be encoded as strongly, as the memory store already contains this information. Similarity is computed as the vector cosine between the word's existing memory row and the context. The cosine is passed through an exponential transformation such that high similarity is transformed into low distinctiveness, and low similarity of context is transformed into high distinctiveness. The magnitude of the transformation is controlled by the λ parameter, which is a scaling parameter that determines how much to weight the differences between high and low similarity contexts. This transformed value is the semantic distinctiveness, SD:

$$\mathrm{SD} = e^{-\lambda * \cos(\mathbf{context},\ \mathbf{word_i})}$$

The SD value is then encoded into a word's row in the new column in the word-by-document memory matrix. A word's meaning is represented by the pattern of elements in the row corresponding to it (the vector phase), and its individual strength is simply the sum of its vector elements (the vector magnitude). Given equal frequency, words that occur in more semantically unique contexts will have a higher magnitude than words that appear in redundant contexts. Formally, this produces a vector in which

the magnitude of a word is equal to its semantic distinctiveness from Section 4.1, but using an expectancy–congruency mechanism common to reinforcement learning. Note that the model can be manipulated so that a word's vector magnitude is equal to its frequency, document count, or semantic diversity in the same model framework. We already know that a semantic diversity measure gives a better prediction of word recognition data; the open question is how does this affect the quality of semantic representation (phase) in the model?

5.2 Using Phase and Magnitude in the Semantic Distinctiveness Model

The SDM takes advantage of the magnitude of the vector to account for CD in lexical access, and takes advantage of the phase pattern of the vector to account for CD in word meaning and lexical similarity. Hence, the model can predict both word identification latencies and semantic phenomena from the same memory store (Johns et al., 2014; Johns & Jones, 2008; Jones et al., 2012). Fig. 4 shows a two-dimensional scaling solution of SDM vectors representing words from multiple discourse topics after the model has been trained on the TASA corpus. Words that are close in the space are semantically similar. Darker words are higher in individual intensity (based

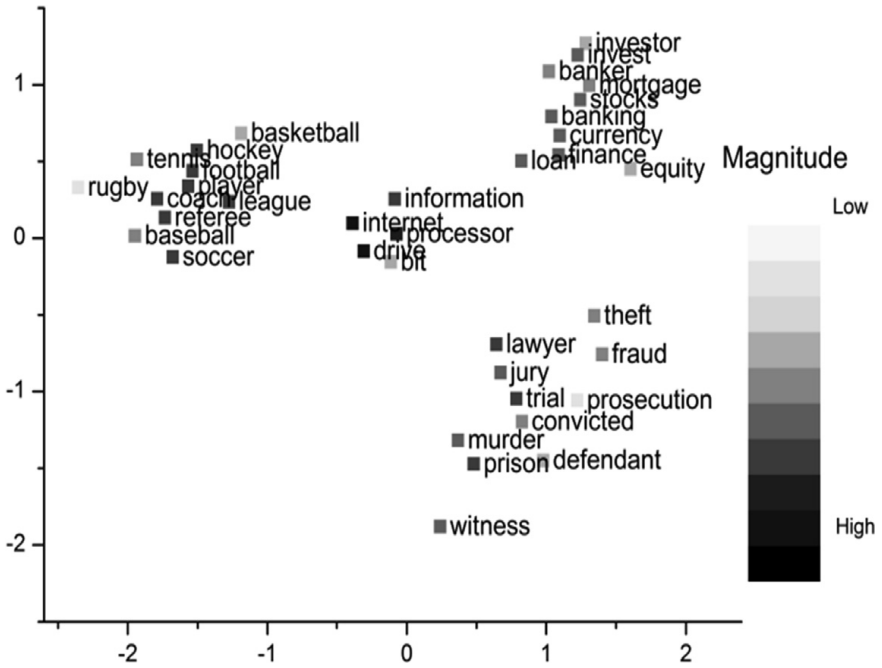

Figure 4 Phase and magnitude in a subset of the lexicon learned by the semantic distinctiveness model.

on semantic diversity). This type of model allows researchers to study the interaction of lexical availability and lexical similarity within a single formal framework. In semantic priming, for example, the identification latency of the target is based partially on its individual intensity (access priority) and partially on its semantic similarity to the prime. These two types of lexical structure are constantly interacting in a variety of behavioral tasks, but also in everyday reading, situational prediction, and online sentence comprehension.

The vector magnitude learned by SDM gives superior predictions of all previously mentioned lexical decision and naming studies than document count or raw frequency. It also accounts for the pattern of experimental data in both the artificial language experiment in Jones et al. (2012), and the natural language experiment of Johns, Dye, et al. (2016). In addition, the model produces better semantic representations using semantic distinctiveness than those learned by paying attention to document count or frequency. When we optimize the right contextual information to make magnitude predictions work best, phase predictions seem to improve for free.

For example, Jones et al. (2012) found that cosine similarities from SDM gave better predictions of word-pair similarity judgments for a sample of ~46,000 words from Maki, McKinley, and Thompson (2004) than did nested versions of the model that used frequency or document count as an information source. Further, SDM explained the dissociation between lexical access and lexical similarity judgments in the experiments of Johns et al. (2014) and Johns, Dye, et al. (2016). Building a distributional model from a mechanism that learns semantic distinctiveness, as opposed to frequency or document count, gives a better representation to explain both types of tasks from within a single formal framework, and points to an important connection between models of lexical access (based on single-word statistics) and lexical similarity (based on cooccurrence statistics).

5.3 Consistency of the Semantic Distinctiveness Model Across Multiple Languages

An original motivation when creating the SDM was to explore how diversity manifests itself across different languages, genres, and registers; e.g., newspapers are very different than children's books,[4] and a measure

[4] And the statistical structure of children's books is very different from the statistical structure of child-directed speech (Montag, Jones, & Smith, 2015).

of frequency or document count is much less sensitive to these content differences. To further explore the stability of SDM's predictions over frequency and document count, we compiled corpora in three different languages and compared predictions of lexical decision latency under each of the three environmental information sources.

English, French, and Dutch corpora were assembled from collections of fiction books. Each corpus was assembled from a set of 2000 books in that language, and restricted to ~300,000 paragraphs. Since it is difficult to identify where paragraph breaks occur in online books, a moving window of 20 sentences was used to estimate context. This methodology yielded differences in the word counts across corpora. The English corpus contains approximately 79 million words, the French corpus 65 million words, and the Dutch corpus 76 million words. Although there is some variability in terms of the overall size of the corpora, these are all very sizeable samples of language, and indeed are the largest that the SDM has been trained on.

Lexical decision times were obtained from the English lexicon project (Balota et al., 2007), the French lexicon project (Ferrand et al., 2010), and the Dutch lexicon project (Keuleers, Diependaele, & Brysbaert, 2010). The data used in the following analyses were z-transformed reaction times. For the English data, 39,351 words were common between the corpus and behavioral data, 37,469 words for the French data, and 14,089 words for the Dutch data. The analysis methods employed here emulate those used by Adelman et al. (2006) and Jones et al. (2012). As in these other studies, all frequency, document count, and SD values were transformed to a log scale. The effect of each variable was assessed in a multiple regression analysis where the amount of unique variance over and above the other variables was measured through percent change in the R^2 value.

The amount of unique variance explained for each variable and language is presented in Table 1. Replicating past results (Johns et al., 2012; Jones et al., 2012), the results of the regression analysis show that the SDM magnitudes explain the lion's share of unique variance, over and above frequency or document count. Additionally, the effects of frequency and document count were minimized across each language, indicating that these variables explain limited unique variance. Interestingly, the amount of unique variance that the SDM explains is actually smallest for the English data, with the model explaining more variance in the French and Dutch data. It is difficult to isolate the reasons for this finding; however, it could

Table 1 Lexical decision time variance predicted by SDM, WF, and DC models across English, French, and Dutch

	Effect (ΔR^2 in %)		
	English	**French**	**Dutch**
SDM (after WF)	1.14***	3.01***	3.53***
DC (after WF)	0.254**	0.62***	1.64***
SDM (after DC)	1.16***	2.7***	2.53***
WF (after DC)	0.16***	0.22***	0.33***
DC (after SDM)	0.0 *n.s.*	0.3***	0.044 *n.s.*
WF (after SDM)	0.08***	0.264***	0.046 *n.s.*
SDM (after DC, WF)	1.59***	2.44***	2.25***
DC (after SDM, WF)	0.22***	0.1**	0.0 *n.s.*
WF (after SDM, DC)	0.38	0.03*	0.0 *n.s.*

*$P < .05$; **$P < .01$; ***$P < .001$.
DC, document count; *SDM*, semantic distinctiveness model; *WF*, word frequency.

be due to differences in syntactic structure across the three languages, or could be unique to fiction writing.

These results demonstrate that the benefit of semantic diversity holds across different languages, an important finding. Additionally, given the advent of large-scale psycholinguistic data sets that are being collected for different languages (e.g. Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2010), it is useful to begin to analyze these data sets to determine the commonalities among them. This type of analysis will allow for empirical insights into the universal properties of language, a key step for quantitative approaches to understanding lexical organization.

6. CONCLUSIONS

> *…an account of frequency effects that is also an account of how new words are* learned *would surely be preferable to one that accounts for frequency effects alone, and then only by means of assumptions made solely for the purpose of explaining frequency effects.*
>
> **Monsell (1991, p. 149)**

The study of the mental lexicon is plagued by an ever-increasing number of highly intercorrelated variables. One approach to this problem has been to establish evaluative methods for appraising relevant variables: judging, for example, the extent to which they are objectively derivable, can predict behavioral data, or can be modeled computationally (Shillcock,

McDonald, Hipwell, & Lowe, 1998). Regression analyses over large-scale data sets have also been a helpful tool in this type of assessment (see e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Kliegl, Olson, & Davidson, 1982).

Such analyses have repeatedly revealed that distributional variables are as good or better at predicting behavioral outcomes as frequency (Adelman et al., 2006; Jones et al., 2012). In lexical decision, local syntactic and morphological cooccurrence probabilities—*not* repeated exposure—are what make frequency such a powerful predictor of response times (Baayen, 2010). Words are processed more rapidly if they occur more often as constituents in other words, cooccur with a larger variety of other words, and are found in a broader array of texts—factors that are tightly coupled with frequency. When these distributional variables are partialled out, frequency's explanatory power is dramatically attenuated. From this perspective, frequency effects may simply be "an epiphenomenon of learning to link form to lexical meaning" (Baayen, 2010, p. 438). It seems much more likely that it is a word's distributional properties that influences how it is learned and remembered.

Rather than ignoring the highly structured nature of words and texts, or relying on intuitive categorical metrics, like frequency, that obscure the more complex nature of the underlying distribution (MacCallum, Zhang, Preacher, & Rucker, 2002), scientists should use simple, cognitively plausible algorithms to uncover the underlying statistical structure of language that is relevant to human learners (Halevy, Norvig, & Pereira, 2009; Recchia & Jones, 2009). Such a proposal aligns well with the idea that cognitive models should be constrained by plausible representational assumptions, derived from environmental information (Johns & Jones, 2010).

ACKNOWLEDGMENTS

This work was supported by NSF BCS-1056744 and IES R305A150546. We would like to thank Geoff Hollis for feedback during earlier versions of this chapter.

REFERENCES

Abernethy, E. M. (1940). The effect of changed environmental conditions upon the results of college examinations. *Journal of Psychology, 10*, 293–301.

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*(1), 214–227.

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*(9), 814–823.

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., … Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods, 46*(4), 1052–1067.

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2014). Individual differences in reading aloud: A mega-study, item effects, and some models. *Cognitive Psychology, 68*, 113–160.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing and Management, 39*, 45–65.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One, 4*(11), e7678.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6*(4), 451–474.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*(4), 703–719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408.

Asr, F. T., Willits, J. A., & Jones, M. N. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In *Proceedings of the 37th meeting of the cognitive science society*.

Baayen, R. H. (1996). The randomness assumption in word frequency statistics. In *Research in humanities computing* (Vol. 5, pp. 17–31). Oxford: Oxford University Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Luedeling, & M. Kyto (Eds.), *Corpus linguistics. An international handbook* (pp. 900–919). Berlin: Mouton De Gruyter.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*(3), 436–461.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 1174–1220.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). Word-frequency mirror effect in young, old, & early stage Alzheimer's Disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language, 46*, 199–226.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283–316.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition, 29*(4), 639–647.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Barabási, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature, 435*, 207–211.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology, 49*(2), 229–240.

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

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin and Review, 8*(3), 531–544.
Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*(3), 510–526.
Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711–733.
Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior, 12*(5), 563–576.
Church, K. W., & Gale, W. A. (1995a). Poisson mixtures. *Natural Language Engineering, 1*(02), 163–190.
Church, K. W., & Gale, W. A. (1995b). Inverse document frequency (idf): A measure of deviation from Poisson. In *Proceedings of the third workshop on very large corpora* (pp. 121–130).
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.
Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.
Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*(2), 337–344.
Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478.
Durda, K., & Buchanan, L. (2008). WINDSOR: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods, 40*(3), 705–712.
Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York: Dover.
Ekstrom, A. D., Arnold, A. E., & Iaria, G. (2014). A critical review of the allocentric spatial representation and its neural underpinnings: Toward a network-based perspective. *Frontiers in Human Neuroscience, 8*, 803.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.
Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145–154.
Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift Für Anglistik Und Amerikanistik, 54*(2), 177–190.
Farrell, S., & Lewandowsky, S. (2015). *An introduction to model-based cognitive neuroscience*. New York: Springer.
Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488–496.
Firth, J. R. (1957). *Papers in linguistics 1934–1951*. London: Oxford University Press.
Forster, K., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 627–635.
Gallistel, C. R. (2003). Conditioning from an information processing perspective. *Behavioural Processes, 62*(1–3), 89–101.
Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition, 15*(1), 24–28.
Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*(2), 256–281.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325–331.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.
Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*, 23–29.
Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183–216). New York: Wiley.
Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403–437.
Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science, 18*(12), 1069–1076.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.
Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.
Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*(2), 151–167.
Harris, Z. (1970). *Papers in structural and transformational linguistics*. Dordrecht/Holland: D. Reidel.
Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review, 7*(2), 185–207.
van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190.
Hills, T. T., Maouene, J., Riordan, B., & Smith, L. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language, 63*(3), 259–273.
Hintzman, D. L., Block, R. A., & Summers, J. J. (1973). Contextual associations and memory for serial position. *Journal of Experimental Psychology, 97*(2), 220–229.
Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730.
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*(9), 2432–2446.
Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, 1–19.
Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review, 112*(1), 75.
Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 923.
Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269–299.
Howard, M. W., Shankar, K. H., & Jagadisan, U. K. K. (2010). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science, 3*(1), 48–73.
Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*, 401–410.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics, 40*, 431.

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning and Behavior, 40*(1), 61–82.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13*(2), 278–305.

Johns, B. T., Dye, M. W., & Jones, M. N. (2014). The influence of contextual variability on word learning. *Proceedings of the 35th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review, 23*(4), 1214–1220.

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*(2), EL74–EL80.

Johns, B. T., & Jones, M. N. (2008). Predicting word-naming and lexical decision times from a semantic space model. In *Proceedings of the 30th annual cognitive science society*.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin and Review, 17*(5), 662–672.

Johns, B. T., Sheppard, C., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in lexical organization across aging and bilingualism. *Frontiers in Psychology, 7*, 703.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*(2), 115–124.

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

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

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, & J. T. Townsend (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254).

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*. http://dx.doi.org/10.1111/cogs.12353.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation, 1*(2), 139–159.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering, 2*(1), 15–59. Chicago.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1*.

Kinsbourne, M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 13*(1), 63–69.

Kliegl, R., Olson, R. K., & Davidson, B. J. (1982). Regression analyses as a tool for studying reading processes: Comment on Just and Carpenter's eye fixation theory. *Memory and Cognition, 10*(3), 287–296.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language, 64*(3), 249–255.

Lovelace, E. A. (1988). On using norms for low-frequency words. *Bulletin of the Psychonomic Society, 26*(5), 410–412. http://dx.doi.org/10.3758/bf03334899.
Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1*(4), 309–317.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.
MacDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd annual conference of the cognitive science society, Edinburgh, Scotland*.
Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, and Computers, 36*(3), 421–431.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
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*(Pt 3), 295–323.
McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review, 39*(4), 352.
McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547–559.
Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227.
Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*(1), 1–28.
Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., & Kahana, M. J. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science, 342*(6162), 1111–1114.
Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393–424.
Monsell, S. (1991). The nature and locus of word frequency effects in reading. In *Basic processes in reading: Visual word recognition* (pp. 148–181).
Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science, 26*(9).
Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165–178.
Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128*(4), 403–415.
Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 721–756.
Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory and Cognition, 28*(4), 509–522.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers, 36*(3), 402–407.
Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120*(2), 356–394.
Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In *Cognitive skills and their acquisition*.
Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527.

Paivio, A. (1971). *Imagery and verbal processing*. New York: Holt, Rinehart & Winston.
Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1), 1–25.
Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word identification times in young readers. *Journal of Experimental Child Psychology, 116*(1), 37–44.
Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin and Review, 15*(1), 161–167.
Pexman, P. M., Siakaluk, P. D., & Yap, M. J. (2013). Introduction to the research topic meaning in mind: Semantic richness effects in language processing. *Frontiers in Human Neuroscience, 7*, 723.
Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research* (pp. 433–462). Timonium, MD: York Press.
Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 275–283.
Polyn, S. M., & Kahana, M. J. (2007). Memory search and the neural representation of context. *Trends in Cognitive Sciences, 12*(1), 24–30.
Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science, 6*(1), 5–42.
Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*, 909–957.
Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*(3), 191–201.
Recchia, G., Johns, B. T., & Jones, M. N. (2008). Context repetition benefits are dependent on context redundancy. In *Proceedings of the 30th annual cognitive science society*.
Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*(3), 647–656.
Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.
Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *The American Psychologist, 43*(3), 151–160.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.
Riordan, B., & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. In D. S. MacNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual cognitive science society* (pp. 599–604).
Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3*(2), 303–345.
Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2016). From words to behaviour via semantic networks. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2207–2212). Austin, TX: Cognitive Science Society.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487–494.
Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior, 19*(6), 736–755.
Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* (Doctoral dissertation, Institutionen för lingvistik).
Salasoo, A., Shiffrin, R. M., & Feustel, T. C. (1985). Building permanent memory codes: Codification and repetition effects in word identification. *Journal of Experimental Psychology: General, 114*(1), 50–77.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.
Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill computer science series.
Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*(1), 1–17.
Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27*(5), 499–520.
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*(1), 82–102.
Serrano, M.Á., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. *PLoS One, 4*(4), e5372.
Shannon, C. E. (1948). A note on the concept of entropy. *Bell System Technical Journal, 27*, 379–423.
Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*(2), 393–413.
Shillcock, R. C., McDonald, S., Hipwell, P., & Lowe, W. (1998). *Distinctiveness of spoken word context predicts visual lexical decision time*. Research Paper EUCCS-RP-1998-4.
Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin and Review, 3*(3), 314–321.
Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.
Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568.
Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460–471.
Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory and Cognition, 6*(4), 342–353.
Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.
Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 760–766.
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41–78.
Thompson-Schill, S., Ramscar, M., & Chrysikou, E. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science, 8*(5), 259–263.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138–146.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352–373.
Whaley, C. P. (1978). Word—Nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 143–154.
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.
Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends, 30*(1), 53–64.
Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin and Review, 18*(4), 742–750.
Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

FURTHER READING

Jones, M. N. (2016). *Big data in cognitive science*. Routledge: Psychology Press.
Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on visual-word recognition. In P. M. A. Rabbitt, & S. Dornic (Eds.), *Attention and performance V* (pp. 98–118). London.
Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience, 6*(315), 1–16.
Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*(4), 893–912.
Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 796–800.