Journal of Memory and Language 123 (2022) 104313

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Content matters: Measures of contextual diversity must consider semantic content

Brendan T. Johns [a,*], Michael N. Jones [b]

[a] McGill University, Canada
[b] Indiana University, United States

ARTICLE INFO

Keywords:
Lexical organization
Contextual diversity
Word frequency
Distributional modeling
Lexical semantics

ABSTRACT

Measures of contextual diversity seek to replace word frequency by counting the number of different contexts that a word occurs in rather than the total raw number of occurrences (Adelman, Brown, & Quesada, 2006). It has repeatedly been shown that contextual diversity measures outperform word frequency on word recognition datasets (Adelman & Brown, 2008; Brysbaert & New, 2009). Recently, Hollis (2020) demonstrated that the standard operationalization of contextual diversity as a document count accounts for relatively little unique variance over word frequency when other variables of contextual occurrences are controlled for. One aspect of the analysis conducted by Hollis (2020) that was not taken into account was the semantic content of the contexts that words occur in. Johns, Dye, and Jones (2020) and Johns (2021) have recently shown that defining linguistic contexts at larger, and more ecologically valid, levels lead to contextual diversity measures that provide very large improvements over word frequency, especially when implemented with principles from the Semantic Distinctiveness Model of Jones, Johns, and Recchia (2012). Across a series of simulations, we demonstrate that the advantages of contextual diversity measures are dependent upon the usage of semantic representations of words to determine the uniqueness of contextual occurrences, where unique contextual occurrences provide a greater impact to a word's lexical strength than redundant contextual occurrences. The results of the simulations suggest that for better theoretical accounts of lexical strength to be developed, attention needs to be paid to the representation of linguistic contexts. Code and data associated with this article is available at https://osf.io/r5ec2/.

Theories of lexical organization are dependent on measures of word occurrence in the natural language environment, as it has been repeatedly shown that words that occur more often in natural language are processed faster (Broadbent, 1967; Forster & Chambers, 1973; Krueger, 1975). Classically, *word frequency* (WF) has been the central lexical strength measure used. Word frequency is measured by simply counting the number of times that a word occurs in a corpus of natural language, and it has served as a central theoretical and methodological measure in psycholinguistics for decades across many domains (see Brysbaert, Mandera, & Keuleers, 2018 for a review).

However, the role of word frequency has been questioned by researchers examining contextual diversity accounts of lexical organization (Adelman, Brown, & Quesada, 2006; see Jones, Dye, & Johns, 2017 for a review). Adelman et al. (2006; see also McDonald & Shillcock, 2001 for a similar proposal and Schwanenflugel, Harnishfeger, & Stowe, 1988 for an early evaluation of these ideas and Caldwell-Harris, 2021 for a recent review) proposed a measure they entitled *contextual diversity* (CD) which measures the number of different contexts that a word occurs in. In this account, a linguistic context was operationally defined as a document (with varying definitions of a document across different corpus types), with lexical strength being determined by the number of documents that a word occurred in (with repetition within context being ignored). Measures based on contextual diversity have repeatedly been shown to provide a superior account of word recognition data than word frequency (e.g., Adelman et al., 2006; Adelman & Brown, 2008; Brysbaert & New, 2009). When we refer to CD in this article, we are referring to the operationalization of counting the number of times that a word occurs in a specified contextual unit in a corpus.

The theoretical justification for using CD as an organizational principle of the lexicon was hypothesized by Adelman et al. (2006) to be explained by the principle of likely need from the rational analysis of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991; see

* Corresponding address at: Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, Quebec H3A 1G1, Canada.
E-mail address: brendan.johns@mcgill.ca (B.T. Johns).

https://doi.org/10.1016/j.jml.2021.104313
Received 6 April 2021; Received in revised form 20 December 2021; Accepted 21 December 2021
Available online 29 December 2021

Jones et al., 2017 for a more in-depth discussion of these issues and Westbury, 2020 for related ideas). This principle proposes that memory is adaptive and should be organized such that information that is most likely to be required in a future context should be the most available in memory. In terms of lexical organization, this principle states that words that occur in a wide variety of contexts should be the most readily available in the lexicon, as they are most likely to be accessed in a given future context.

Recently, the superiority of contextual diversity over word frequency has been questioned by Hollis (2020). Among other examinations into the impact of contextual diversity on lexical organization, Hollis noted that definitions of context in traditional measures of CD have been poorly operationalized. Further, Hollis demonstrated through multiple simulations that CD measures seem to account for relatively small amounts of variance once other aspects of word contextual occurrences, such as word burstiness (the propensity of words to repeat in context; Madsen, Kauchak, & Elkan, 2005), and the nonlinear form of the relationship between the variables, are controlled for.

However, the debate is more profound than simply which measure gives a superior fit to the human data—word frequency and contextual diversity measures imply radically different mechanisms of lexical learning and organization. If word frequency offers the best explanation of the human data, it implies a classic repetition-based mechanism of encoding and organization of the lexicon. In contrast, if contextual diversity measures offer a superior fit to the human data, this implies a need-based rational mechanism of lexical encoding and organization. Finding the correct generating mechanism has far-reaching consequences ranging from how we optimally teach children vocabulary in the classroom (Mak, Hsiao, & Nation, 2021; Rosa, Tapia, & Perea, 2017; Rosa, Salom, & Perea, 2022; Tapia, Rosa, Rocabado, Vergara-Martínez, & Perea, in press) to speech therapy (Plante et al., 2014), reading (Joseph & Nation, 2018; Perea, Soares, & Comesana, 2013), and second-language acquisition (Frances, Martin, & Dunabeitia, 2020).

One of the primary criticisms that Hollis (2020) leveled against theories of contextual diversity is that they do not accurately operationalize notions of linguistic context. That is, by using somewhat arbitrary notions of context (e.g., a moving window, paragraph, or a document within a corpus) that do not accurately map onto ecologically valid notions of what a linguistic context is, the theoretical utility of contextual diversity is questionable. Thus, in order to show the limitations of these operationalizations of context, the CD measures Hollis (2020) tested were along the same lines as previous definitions used, such as Wikipedia articles, webpages, or textbook paragraphs. The results of Hollis (2020) demonstrate that deriving contextual diversity measures at these levels have questionable theoretical usefulness[1].

The lack of ecological validity of previously used definitions of context was also one of the motivating factors for the parallel work to Hollis (2020) contained in Johns, Dye, and Jones (2020) and Johns (2021), which we believe to be complimentary to the work of Hollis (2020) and contain potential clues to answer to some of the concerns raised. In Johns et al. (2020) and Johns (2021), CD measures at much larger units of language than previously considered measures were constructed. For example, Johns et al. (2020) used a large corpus of fiction novels, and constructed CD measures at the single book (i.e., a word's strength was increased when it occurred in a book, with repetitions within book being ignored) or author (i.e., a word's strength was increased when it was contained in an author's collective writings). Johns et al. (2020) found that CD measures constructed at these levels provided a large improvement over WF and traditional measures of CD (i.e., CD measured at the paragraph or document level) for lexical decision and naming accuracy data from the English lexicon project (Balota, et al., 2007) and the recently released word prevalence data of Brysbaert, Mandera, McCormick, and Keuleers (2019).

However, using a book or author as a definition of context also suffers from the same issues that previous definitions of context have, in that they are simply convenient criteria to divide a text corpus into discrete units for quantitative analyses. That is, using a book as a definition of context suffers from similar ecological validity criticisms as a document; they are both unlikely organizational points for lexical memory. To overcome these issues, Johns (2021) recently proposed a redefinition of linguistic context based in the communication patterns of individual language users across discourses. These measures were operationalized and constructed through the analysis of a very large internet discussion forum, namely Reddit, attained from the website *pushshift.io* which aggregates Reddit posts (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020), using a collection of over 55 billion words produced by hundreds of thousands of individuals.

Specifically, two new CD measures were proposed by Johns (2021), entitled user contextual diversity (UCD) and discourse contextual diversity (DCD). UCD is based on the number of users who had produced a word during their communications, while DCD is based on the number of discourses that a word was produced in (with discourse being defined as a subreddit, which are comments organized around a specific topic; e.g., r/CogSci is a discussion forum focused on cognitive science). The advantage of these definitions is that they map onto everyday notions of language usage better than previous definitions, as the UCD measures how likely it is for someone to use a word, while DCD measures the propensity of encountering a word across different, and more naturalistic, discourse types.

It was found that measuring contextual occurrence at these levels provided a large and systematic advantage over WF across a number of datasets, for both reaction time and accuracy data (a possibility hinted at by Hollis, 2020). This suggests that the advantage of a CD measure lies at much larger units of natural language, which map onto communicative aspects of language usage rather than purely linguistic notions, such as moment-to-moment differences in language usage within relatively small units of language (i.e., a paragraph or a document), in coherence with usage-based theories of language (Tomasello, 2003).

However, in both Johns et al. (2020) and Johns (2021), the real advantage of the redefined notions of CD emerged when the measures were transformed with the Semantic Distinctiveness Model (SDM), first described in Jones, Johns, and Recchia (2012). The underlying conception of the SDM is that each context that a word occurs in is not equally informative about the types of future contexts that a word could possibly occur in. If a word only occurs in one context type, it is easy to predict the future context that a word will occur in. However, if a word occurs in many different context types then it is difficult to predict that word's future contextual occurrence pattern. The SDM is a type of distributional model of semantics (see Günther, Rinaldi, & Marelli, 2019; Kumar, 2020 for recent reviews), which learn the meanings of words through the processing of large text corpora. However, instead of focusing on deriving semantic representations of word meanings (which it is capable of; Johns & Jones, 2008) the emphasis of the SDM is placed on deriving accurate measures of lexical strength.

The SDM implements this learning mechanism with an expectancy-congruency mechanism. In the model, a word's lexical strength is updated each time it occurs in a context. The update strength is determined by how dissimilar the current context is compared to the past contextual usages of a word; essentially, how predictable the currently experienced context is given the current contents of memory. The more unique a context is for a word, compared to past usages, the greater the update strength that is applied to the word. The SDM has been

[1] The original motivation of the CD measure as developed by Adelman et al. (2006) was to examine temporal aspects of contextual word occurrence, not lexical semantic, somewhat analogous to examining spacing effects in episodic memory, and so the criticism leveled by Hollis (2020) do not necessarily map onto the initial theoretical motivations for the development of the original CD measure. However, in this article we interpret the success of CD measures as being lexical semantic in nature, and so the criticisms employed by Hollis (2020) do apply to this work.

developed with targeted experimentation using artificial (Jones et al., 2012) and natural (Johns, Dye, & Jones, 2016) language experiments. Additionally, the SDM has been shown to provide unique insights into spoken word recognition (Johns et al., 2012), bilingualism (Johns, Sheppard, Jones, & Taler, 2016; Hamrick & Pandža, 2020), and aging (Qiu & Johns, 2020). The success of the SDM suggests that the content of the contexts that a word is experienced in are important to lexical organization.

Relatedly, in a key, and clever, simulation Hollis (2020) constructed random contexts through the use of WF values (i.e., assembled randomized documents by sampling from the frequency of words), and the resulting CD metric was entitled CD_rand. When contrasted with WF in a regression, it was found that the CD_rand variable accounted for a similar level of variance as standard CD variables. This result calls into question the validity of a CD count as measuring the semantic variability of the contexts that a word occurs in, as randomly assembled contexts do not have any semantic cohesion.

However, this simulation does not actually take into account the semantic structure of the contexts that a word occurs in, which is the source of information that the SDM capitalizes upon. In distributional models, the meaning of a word is derived from the surrounding linguistic context in which it occurs. To calculate a CD or CD_rand measure, the words surrounding a target word in context are ignored – only the occurrence of the target in a contextual unit is considered. Thus, randomly constructed contexts are not semantically coherent from a bird's eye view, but that lack of coherence does not impact the CD count model – it is only the fact that a word occurred in a context that matters to the model.

The SDM does consider the semantic content of the context that a word occurs in, as the update strength that a word receives from a contextual occurrence is dependent on the overlapping similarity between a word's representation in memory (learned from past contextual occurrences of that word) and the current context. Johns (2021) demonstrated that a UCD measure modified with the SDM accounted for considerably more variance than word frequency across a range of lexical organization datasets. In Johns (2021), two different representation types were contrasted: word representations and population representations. Word representations used the word frequency distribution of a contextual unit as a representation (in the case of the UCD measure, it was the frequency of all the word's that a user produced on Reddit), which was consistent with previous implementations of the SDM (e.g., Johns et al., 2020). In population representations, contexts are represented by the commenting pattern of users across discourses. In the case of the UCD measure, the population representation was the number of comments that an individual made in each discourse on Reddit. It was shown that the best implementation of the SDM utilized a population representation, signalling the importance of communicative information in lexical organization. We will refer to this model as the UCD-SD model in this article.

To demonstrate the scale that the UCD-SD improves upon WF, Fig. 1 shows the amount of unique variance that the UCD-SD and WF measures account for across four datasets: 1) English Lexicon Project (ELP) lexical decision and accuracy data (Balota, et al., 2007), 2) British Lexicon Project (BLP) lexical decision and accuracy (Keuleers, Lacey, Rastle, & Brysbaert, 2012), 3) the word prevalence (WP; a modified lexical decision task) data of Brysbaert, Mandera, McCormick, and Keuleers (2019), and 4) the recently released response times using a similar task as utilized to collect the WP data (Mandera, Keuleers, & Brysbaert, 2020; referred to as WP_RT data in various places of this article). Reaction times were z-transformed while the WP data was probit transformed, and both variables were transformed with a logarithm. The regression calculated the amount of predictive gain (measured as percent ΔR^2 improvement) for one predictor over another competing predictor, a standard analysis technique (see Adelman, et al., 2006; Johns, Sheppard, Jones, & Taler, 2016). This figure shows that UCD-SD provides a very substantial improvement over WF, from an 8.5% improvement for

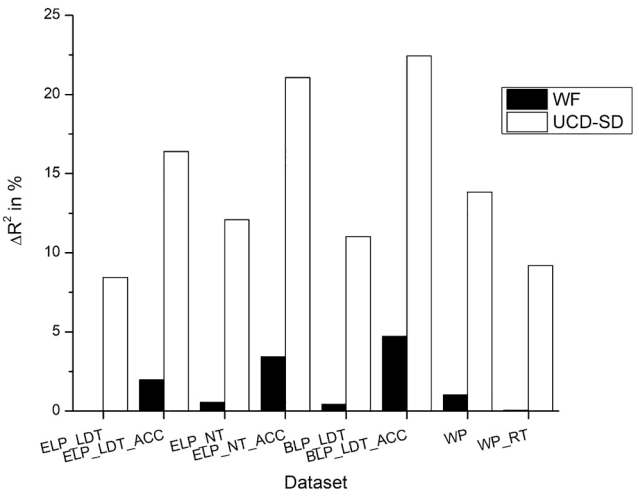

Fig. 1. A regression analysis calculating the amount of unique variance that the UCD-SD and WF variables from Johns (2021) account for across the various lexical organization datasets. N = 40,460 for ELP data; N = 28,710 for BLP data; N = 58,711 for WP data.

ELP lexical decision RT data to a 22.5% improvement for BLP lexical decision accuracy data, while minimizing or eliminating the amount of unique variance accounted for by WF. To put these improvements into perspective, a typical advantage for CD usually lies in the 1–3% range (Adelman et al., 2006; Adelman & Brown, 2008; Brysbaert & New, 2009; Jones et al., 2012). Given that all the datasets contained in this analysis number in the tens of thousands of data points, the improvements of the UCD-SD model over previous theoretical accounts are substantial.

Additionally, Johns (in press a) has recently demonstrated that the improvement offered by the DCD-SD and UCD-SD measures over their count-based alternatives generalizes to word-level episodic recognition rates, while Johns (in press b) demonstrated that a model of distributional semantics trained with communication patterns of users allowed for a unique signal about word meaning to be constructed, and Senaldi, Titone, and Johns (in press) demonstrated the advantages of these measures extends to idiom familiarity datasets. Combined, these results (Johns, 2021; in press a; in press b; Senaldi et al., in press) suggest that communicative and social information are integral parts of linguistic representation across multiple data types and levels of analysis.

Hollis (2020) was correct in calling out many issues in the theoretical development of contextual diversity accounts of lexical organization. For experienced-based theories of cognition to be plausible, the types of materials that a model is trained upon must be coherent with the types of experience that a typical person receives (Johns, Jones, & Mewhort, 2019). Traditional models of contextual diversity have failed to meet this criterion. However, we believe that many of these concerns about measures of contextual diversity are answered by Johns (2021), not just because of the superior fits that the models provide but because of their increased theoretical validity, in particular the UCD-SD model which operationalizes context at the single individual level and has been shown to explain a wide range of behavioral data.

The goal of this article is to determine the underlying reasons for the success of the UCD-SD model. In turn, the results of this article will demonstrate that by producing lexical strength measures that are coherent with more ecologically valid notions of linguistic context, superior models of lexical strength can be constructed. This will be done by examining five aspects of the UCD-SD model: 1) the scale of training materials, 2) the semantic coherency of the training materials, 3) the importance of highly semantically distinct contextual usages of words, 4) the number of user comment used to construct a contextual representation, and 5) the number of discourses that are used to form a contextual representation. Each simulation will shed light on the

plausibility of the proposals put forth by Johns (2021), with the hopes of alleviating concerns about the nature of contextual diversity measures raised by Hollis (2020). Overall, the results of this article will provide theoretical insights into the cognitive mechanisms that are at play in deriving lexical strength, and how those mechanisms interact with the structure of the natural language environment. A secondary goal is to release code and materials pertaining to the SDM that will allow other researchers to explore context effects in lexical organization.

Model and materials

This section will describe the model used in this article, as well as the training materials that will be released, based upon the findings of Johns (2021).

The semantic distinctiveness model (SDM)

The SDM has both processing and representational components. The representational elements of the model are the context and word representations. The context representation signals the meaning of the current context being processed. A word representation is a recording of all the contexts that a word had occurred in, and each word has its own unique representation, assumed to encode a word's meaning. In the model, a word's lexical strength is updated for each context that a word occurred in. The update strength that a word receives for each contextual occurrence is determined through a transformation of the similarity between the current context that a word is occurring in and the past contexts that a word was used in (as is encoded in its lexical representation).

However, the underlying representational assumptions of the model have changed across the various implementations of the model. In the initial implementation of the model of Jones et al. (2012), words were represented in a Word-by-Document matrix, similar to the representational assumptions of the Latent Semantic Analysis model of Landauer and Dumais (1997). In this implementation, each time the model processed a document, a new column was added to the matrix. If a word occurred in the document, it was given an encoding strength for that document (with the words that did not occur in that context getting a value of 0). A context representation was formed by summing the rows of each word that occurred in the context.

This initial implementation had its advantages, such as a word's strength being contained directly in its representation (as the strength of a word was determined by summing across its row in the matrix), but it was computationally expensive to scale up to larger corpora. Johns et al. (2020) switched to using a word frequency distributional representation, where the context representation was a count of each word that occurred in a context (defined at a much larger scale than previous implementations, namely a whole book or the writings of an individual author). A word's representation was then the sum of the contexts (word frequency distributions) that a word occurred in, similar to count-based models of distributional semantics (Johns, Mewhort, & Jones, 2019). Johns (2021) further modified this approach by using population representations (PR) which contain the communication patterns of individuals across discourses. As discussed previously, the best fitting model tested by Johns (2021) was an SDM that was organized around user contextual diversity (UCD), where each context was an individual language user. In Johns (2021) this was referred to as the UCD-SD-PR model, and here as the UCD-SD model.

In the UCD-SD, the lexical strength of a word is stored in an external counter. When a word was produced by a user in the reddit data, that word's strength in the lexicon was increased. Repeated usages by a user were ignored. Thus, words that are used by the majority of the population of language users are the words that have the strongest lexical strength. The population representation that the model used was the count of the number of comments that an individual made across all discourse types. Thus, the context representation had a representational dimensionality corresponding to the number of discourses contained in the dataset which was 30,327, and each element in the vector is a count of the number of comments that the individual user made in each of the discourse topics. The word representation that the model utilized was the sum of the discourse communication pattern of every user who produced a given word.

Since the UCD-SD-PR model will be used in this article to explore different aspects of contextual diversity, Fig. 2 contains a demonstration of how a linguistic context is constructed in this model for a hypothetical user named *Jennifer*. This figure shows that each user is represented by the discourse communication pattern that the user engages in, not the specific words that they produce, unlike previous implementation of the model (e.g., Johns et al., 2020). However, the words that the user produced have their lexical strength updated, with repetitions of words being ignored (e.g., the word *language* was produced by the user 3 times, but it is still updated only one time for that user). The context vector is also added into the memory representations for each of the words that were produced. This process is repeated for each of the user corpora that are being used in model training, which in Johns (2021) was for over 330,000 individual users.

In the SDM, a word's strength is increased according to a semantic distinctiveness (SD) value. The first step to calculating this strength increase is to take the similarity between a word's representation and the current context. Similarity is assessed with a vector cosine (normalized dot product):

$$S(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{N} \mathbf{x}_j \times \mathbf{y}_j}{\sqrt{\sum_{j=1}^{N} \mathbf{x}_j^2}\sqrt{\sum_{j=1}^{N} \mathbf{y}_j^2}} \tag{1}$$

where N is the size of the vector. An SD value is calculated with an exponential transformation of the similarity between a word and context (based on Shepard's (1987) law of psychological distance):

$$SD_{i,j} = e^{-\lambda * S(\mathbf{M}_i,\mathbf{c})} \tag{2}$$

Where i is the word being processed in context j, $\mathbf{M}_i$ is the memory vector for that word, $\mathbf{c}$ is the context vector, and λ is a scaling parameter. λ controls the differential weight given to high versus low variability contexts, and is an important part of the model, which will be discussed subsequently. It is the only free parameter in the model. An SD value signals how unique a contextual occurrence is for a word. Each word that occurred in the context is updated by summing the context representation into their representations:

$$\mathbf{M}_i = \mathbf{M}_i + \mathbf{c} \tag{3}$$

In the SDM, the λ parameter controls the amount of discounting applied to high similarity contexts and the amount of strengthening applied to low similarity contexts. In the implementations of Jones et al. (2012) and Johns et al. (2020) it was found that a relatively small λ value, typically between 1 and 6, optimizes the model's fit. However, in Johns (2021) it was found that for the PR-based models the models were optimized by maximizing λ (which was set at 400 in the resulting simulations). Johns (2021) demonstrated that with a λ at this level, much of the strength of a word comes from very distinctive contextual occurrences of a word, with redundant experiences having a modulating impact on lexical strength. From a likely need perspective, this large discounting of redundant contexts makes perfect sense: only very unique experiences signal a new type of context (in the case of the UCD-SD model, a new type of person) that a word can occur in. That is, the results of the modeling results suggests that it is total *number of types* of contexts that a word occurs in, not the overall total *number of contexts* (as the original CD proposes), that matters in lexical organization.

Materials and data availability

One of the goals of this article is to disseminate both code and

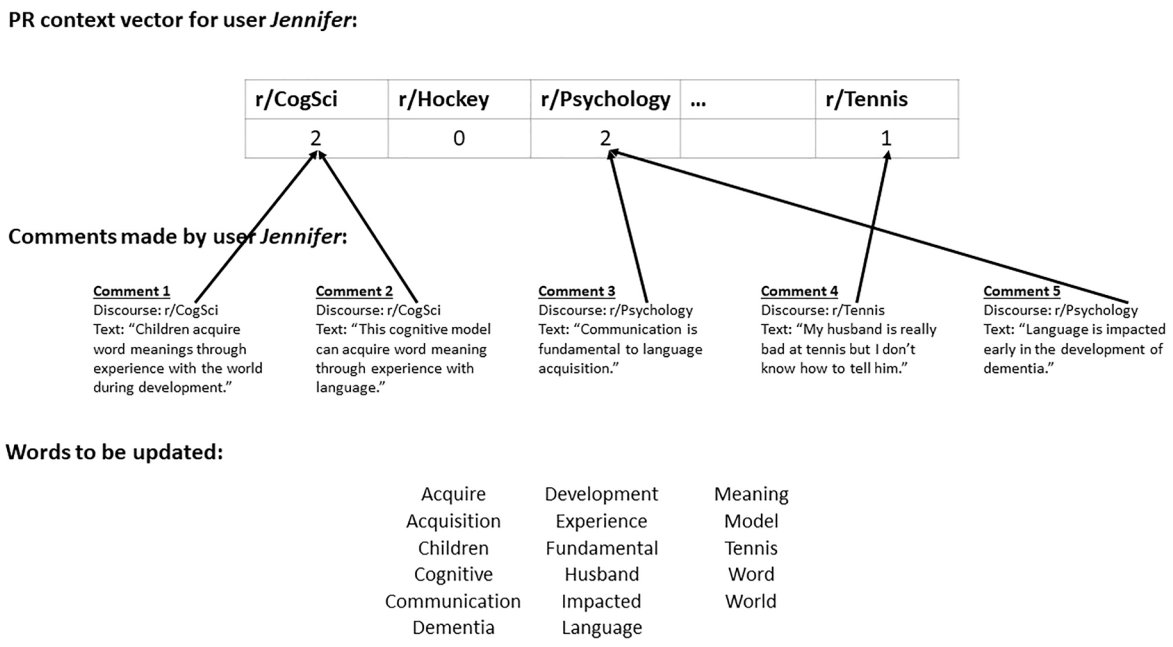

Fig. 2. A demonstration of how the PR context vector is formed for the hypothetical user *Jennifer*. Each discourse that the user communicates in is used to update the strength of that discourse feature for that word. Each word that is produced by the user is used to update the word's lexical strength and its memory representation, with repetitions of words being ignored. In this figure, the updated words do not include function words.

materials that Johns (2021) used to calculate lexical strength values for the UCD-SD-PR. However, the original materials that were used to calculate these values required a dedicated hardware setup with more storage capacity than could be realistically shared. To minimize the required file size, here we sample 20,000 random users, and pre-processed the resulting corpora to make the resulting filesizes small enough to be disseminated. Although reducing the number of users included in the analysis will likely decrease model performance, it will allow for the materials to be shared more easily.

To reduce the file size of the shared materials, a number of pre-processing steps were taken. The first step taken was to remove very high frequency words from the resulting corpora. To accomplish this, the stoplist from Landauer & Dumais (1997) was used. Any word that occurred on this stoplist was excluded from the resulting materials. In the PR models of Johns (2021), context representations (and thus word representations) are not directly tied to word usage, so the inclusion or exclusion of words does not impact the representations formed by the model, unlike the representational assumptions of Jones et al. (2012) and Johns et al. (2020). In Johns (2021) these words were included in the analysis, however the only difference caused will be a smaller number of words included in the resulting analyses. Additionally, each word was transformed into a location value, according to its rank frequency (e.g., the most frequent word was given location 0, the second most frequent word given location 1, etc…), which further reduced filesize. Each line in a corpus file represents a single comment made by the user. The results of this preprocessing reduced the compressed file size to slightly<5gb, which is the shareable limit on OSF.

The random sample used in this article represents a relatively small portion of the data analyzed by Johns (2021), as the analysis contained in that article examined the language usage patterns of 334,345 individuals. However, the sample used here still contains a significant amount of language materials even with very high frequency words removed, as it contains 2,264,109,494 words from 122,780,409 comments across the 20,000 randomly selected users.

All materials are available at https://osf.io/r5ec2/. In the resulting materials, all 20,000 users have individual corpora stored as text files. Each line in a user's corpus is a comment made by that user, with stoplist words removed. The first number in the line represents the discourse where the comment was made, with values ranging from 0 to 30,327 (the total number of discourses contained in the materials analyzed by Johns, 2021). Discourses and users were anonymized in order to reduce the possibility of any identifiable information being included in the resulting material. The code contained in this folder, written in the Java programming language, computes the UCD_SD_PR used in this article.

Results

As stated previously, there are five manipulations of the UCD-SD model that will be done in order to better understand the model's behavior, and to demonstrate that it is a more plausible operationalization of contextual diversity than previously used measures. The five manipulations are: 1) the scale of training materials, 2) the semantic coherency of the training materials, 3) the importance of highly semantically distinct contextual usages of words, 4) the number of user comment used to construct a contextual representation, and 5) the number of discourses that are used to form a contextual representation. Each will be described in turn, and each will provide unique information about the modeling approach.

Scale of training materials

The first aspect of this analysis that needs to be evaluated is the impact of the smaller randomly sampled Reddit corpora on the WF and UCD-SD model performance, compared to the much larger material set used by Johns (2021). Table 1 contains the Pearson correlation coefficients for the WF and UCD-SD models trained on the sampled corpora[2] and the full training materials from Johns (2021), for the various datasets used in Fig. 1. This table shows that for WF there were

[2] The λ parameter for the UCD-SD model was independently fit to the ELP, BLP, and WP datasets. The λ parameter for the ELP and BLP data was set at 200, while the λ parameter was set at 50 for the WP data. These parameters are lower than the optimal λ parameters found in Johns (2021) but are still much larger than previous implementations of the model. The lower λ parameter likely reflects the smaller amount of training materials used in this article.

Table 1
Correlations between the WF and UCD-SD models for both the full and sampled Reddit corpora.

	WF		UCD-SD	
	Full	Sampled	Full	Sampled
ELP_LDT	-.663	-.66	-.691	-.681
ELP_LDT_Acc	.504	.506	.543	.54
ELP_NT	-.559	-.556	-.595	-.585
ELP_NT_Acc	.4	.4	.44	.433
BLP_LDT	-.644	-.645	-.681	-.673
BLP_LDT_Acc	.614	.619	.677	.661
WP	.698	.704	.755	.748
WP_RT	-.728	-.73	-.772	-.762
Average |r|	.601	.603	.644	.635

Note. N = 39,948 for ELP data; N = 28,065 for BLP data; N = 57,716 for WP data; all correlations are significant at the $p < 0.001$ level.

negligible differences between the full and sampled corpora. For the UCD-SD variable, the model trained on the full user corpus outperformed the model trained on the sampled corpora across every dataset, suggesting that more training materials boosts the UCD-SD model performance. However, the sampled UCD-SD model still provides a much better fit to all data sources than WF, consistent with the previous findings.

To ensure that the sampled UCD-SD model still accounts for more unique variance than WF does, a replication of the regression analysis contained in Fig. 1 was conducted using the SDM trained on the sampled training materials. The results of this analysis are contained in Fig. 3 and demonstrates that the sampled UCD-SD variable still accounts for significantly greater levels of unique variance compared to WF, while minimizing or eliminating the unique variance that WF accounts for. This finding suggests that even though the sampled UCD-SD variable has a reduced fit, as shown in Table 1, it still retains the same pattern of large improvement over WF as was found when the model was trained on the complete training set.

The primary purpose of the above simulation is practical, since by ensuring that the same pattern of superiority of the UCD-SD model holds at a smaller number of contexts it makes it more computationally feasible to manipulate different aspects of the model's framework to better understand its performance. However, there are theoretical consequences as well - namely that the model does seem to benefit somewhat from processing a greater number of different contexts (or users, in this case). To better understand the function of this improvement, an additional simulation was done where the number of contexts used by the model was manipulated from 40,000 to 320,000 users in steps of 40,000, and the amount of improvement over WF was calculated using linear regression. The WF values were derived from the total corpus, rather from the limited samples. The λ parameter was fit independently at each context size. The result of the simulation is contained in Fig. 4 and shows that there is a rather limited improvement in fit over word frequency as a function of the number of contexts studied (an average improvement from 12.12% to 13.31% across the different datasets). This suggests that after a certain number of contextual experiences, the model only benefits very slightly from additional contexts being studied.

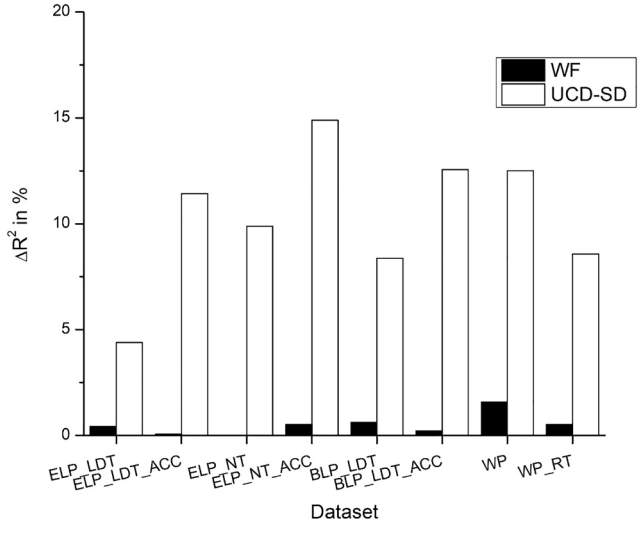

Fig. 3. A regression analysis calculating the amount of unique variance that the UCD-SD and WF variables trained on the sampled training materials account for across the various lexical organization datasets.

Semantic consistency of training materials

The following simulations will attempt to tease apart the reasons for the superiority of the UCD-SD model over WF. Specifically, we will attempt to determine the impact of the semantic consistency of the training materials that the model is trained on. This will be done by constructing randomized comparison training materials to compare to the intact training materials used in the fits contained in Table 1 and Fig. 3. Specifically, each user corpus will be randomized to include a certain percentage of other user's comments, thus reducing the semantic consistency of the resulting context representation that the model constructs. This is similar to the simulation ran by Hollis (2020) with the CD_rand variable, however here we will be randomizing the user corpora at the comment level, rather than the word level. The first comparison training materials will retain 75% of a user's comments, with the other 25% of that user's comments being replaced with comments from other users. The replaced comments will be randomly placed in other user's corpora, so the total size of all corpora will be identical to the 100% intact condition. For example, if a user had produced 1,000 comments during their time on the forum, in the resulting training materials 750 of their comments would be retained in that user's corpus, but 250 of their comments would be randomly replaced with comments produced by other users. Three additional comparison training corpora were constructed at 50%, 25%, and 0% of a user's comment being retained. In the 0% condition, all comments were randomized across users, however each user corpus had the same number of comments as the intact corpora (i.e., the distribution of the number comments across the corpora are preserved, but the content is completely randomized). The 0% condition is the most analogous to the CD_rand measure constructed by Hollis (2020).

The first aspects of the UCD-SD model's performance on randomized corpora to understand is its use of the λ parameter, as the theoretical

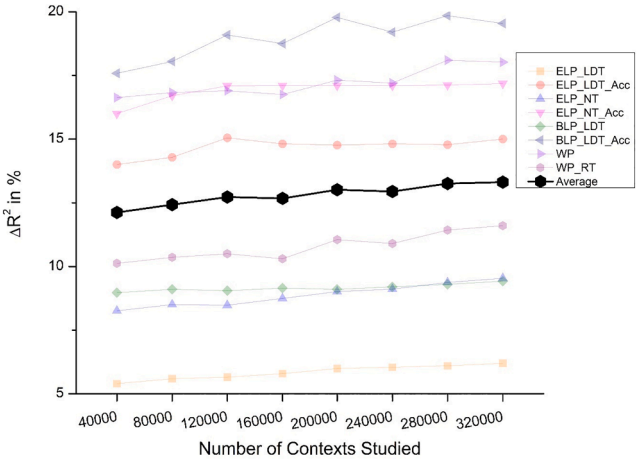

Fig. 4. Increase in fit across the different datasets as a function of the number of contexts studied by the model.

interpretation of this parameter is that it is used in modulating the impact of contextual redundancy on lexical strength. If it turns out that the λ parameter does not impact model performance across the different training material sets, then it would suggest that it is the surrounding architecture of the SDM that is improving performance and not the semantic construction of the different user corpora. As a demonstration of the impact of the λ parameter on model performance, fits of the model across the different training sets with λ values from 0 to 150 to lexical decision RT data from the ELP was computed. The results are displayed in Fig. 5, showing that the optimal λ value modulates by the semantic consistency of the training materials. For the 100% intact training materials the model showed a consistent increase in performance as the λ parameter was increased, consistent with the results of Johns (2021). However, as the training corpora are increasingly randomized, the optimal value of λ is reduced, suggesting that a change in the semantic structure of the materials is detected. For example, the optimal λ for the 0% intact training materials was 2, by far the lowest of all the training materials (after this parameter value there is a rapid drop off in model performance). Additionally, the level of fit with the optimal parameter value was reduced significantly across the different training materials, from an average $r = -.681$ for the 100% intact corpus to $r = -.66$ for the 0% intact corpus, with the correlation for the 0% intact performance being equivalent to WF.

This initial analysis is a first demonstration that the construction of the user corpora significantly impacts model performance. To examine the extent of the impact, the correlations between the model trained on the various training materials to the datasets was assessed. The λ parameter was optimized separately for each dataset (ELP, BLP, and WP) for each corpus type, similar to the results displayed in Table 1. The correlations are contained in Table 2, and the WF correlations are also contained as a comparison. This table shows that as the user corpora are increasingly randomized, the fit of the model decreases. With completely randomized training materials, the average correlation for the SDM is less than WF. However, the SDM trained on the 25% retained training materials still holds an advantage over WF, which speaks to the ability of the SDM to account for semantic changes in context types even with minimal semantic consistency across contexts. Additionally, the SDM with completely randomized materials still retains a significant advantage over WF for accuracy and prevalence data, reflecting the superior fit to this type of data due to the larger count-based measures of context diversity explored by Johns et al. (2020) and Johns (2021).

To determine the impact of random sampling on the amount of unique variance accounted for by the SDM and WF, a regression analysis was conducted mirroring those in Figs. 1 and 3. The top panel of Fig. 6

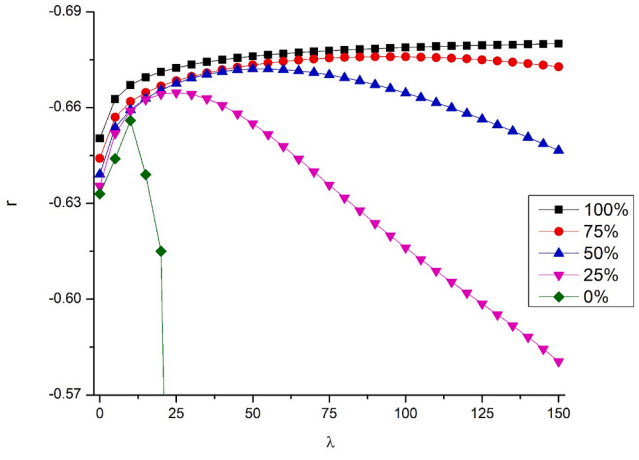

Fig. 5. The impact of the λ parameter on model fit to ELP lexical decision RT for the various training material sets. The 100% intact training materials signals that no randomization of a user's comments had taken place, while the 0% intact training materials have completely randomized training materials.

Table 2
Correlations between the SDM model trained on the intact and randomized corpora and the various lexical datasets.

	100%	75%	50%	25%	0%	WF
ELP_LDT	-.681	-.676	-.672	-.665	-.565	-.66
ELP_LDT_Acc	.54	.532	.528	.526	.519	.506
ELP_NT	-.585	-.58	-.572	-.565	-.555	-.556
ELP_NT_Acc	.433	.428	.421	.418	.41	.4
BLP_LDT	-.673	-.669	-.664	-.659	-.65	-.645
BLP_LDT_Acc	.661	.655	.652	.638	.619	.619
WP	.748	.738	.733	.726	.721	.704
WP_RT	-.762	-.754	-.748	-.743	-.737	-.73
Average |r|	.635	.629	.624	.618	.597	.603

Note. N = 39,948 for ELP data; N = 28,065 for BLP data; N = 57,716 for WP data; all correlations are significant at the $p < 0.001$ level.

displays the amount of unique variance accounted for by the SDM when trained on the different training material sets, while the bottom panel contains the amount of unique variance accounted for by WF from the same training sets. This figure shows that as the SDM is trained with increasingly random training materials, the amount of unique variance accounted for by the model decreases substantially and systematically. As the training materials are increasingly randomized, the amount of unique variance that the SDM accounts for over WF is reduced. The effects on the amount of unique variance accounted for by WF was more varied. Typically, the WF variable accounted for more variance when it was contrasted with the SDM trained on randomized materials, although that was not always the case, especially for accuracy data.

Together, the results contained in Table 2 and Fig. 6 demonstrate that much of the advantage that the SDM has over WF is due to the natural semantic construction of the materials that the model is trained on. To get a better understanding of the impact of the randomization of training materials on model performance, an additional simulation was conducted. In this simulation, fifty random word-context similarity ratings were recorded for each of the 20,000 contexts that the model was trained on, for the 100%, 75%, 50%, and 25% intact corpora. That is, for each of the 20,000 contexts that are apart of a training material set, the similarity between 50 words that were used by a user and the context representation for that user was recorded. These values map onto $S(\mathbf{M}_i, \mathbf{c})$, the similarity between word representation $\mathbf{M}_i$ to context representation $\mathbf{c}$, in equation (2), and are assessed with a cosine (meaning they range between 0 and 1, because there are no negative values in the representation). This led to samples of one million similarity values for each set of training materials. The result of this simulation will show the distribution of word-context similarity that the model is using in constructing its lexical strength measures. Fig. 7 contains the histogram of these values for the four training sets. This figure shows that there is a consistent shift in the similarity distributions as the training materials are randomized. With the non-randomized materials, the similarity distributions are positively skewed, where most of the similarity values are relatively low in similarity. This suggests that the communication patterns of word usage are quite unique across individuals (see Johns & Jamieson, 2018 for a similar result using fiction authors). As the randomness of the materials is increased, the distributions become increasingly negatively skewed, where most similarity values have high similarity (the 0% training materials were not used in this demonstration because all similarity values a were greater than .95 and thus did not have an interesting distributional pattern).

The distributional differences in word-context similarity displayed in Fig. 7 have theoretical consequences for lexical organization. This figure shows that the contextual structure that the SDM is dependent upon is rich in low similarity contexts, meaning individuals with unique communication patterns. This contextual distinctiveness leads to unique contexts having a large impact on a model's performance (Johns, 2021), and indicates from a likely need perspective a new type of context that word could occur in (e.g., signals previously unknown knowledge about

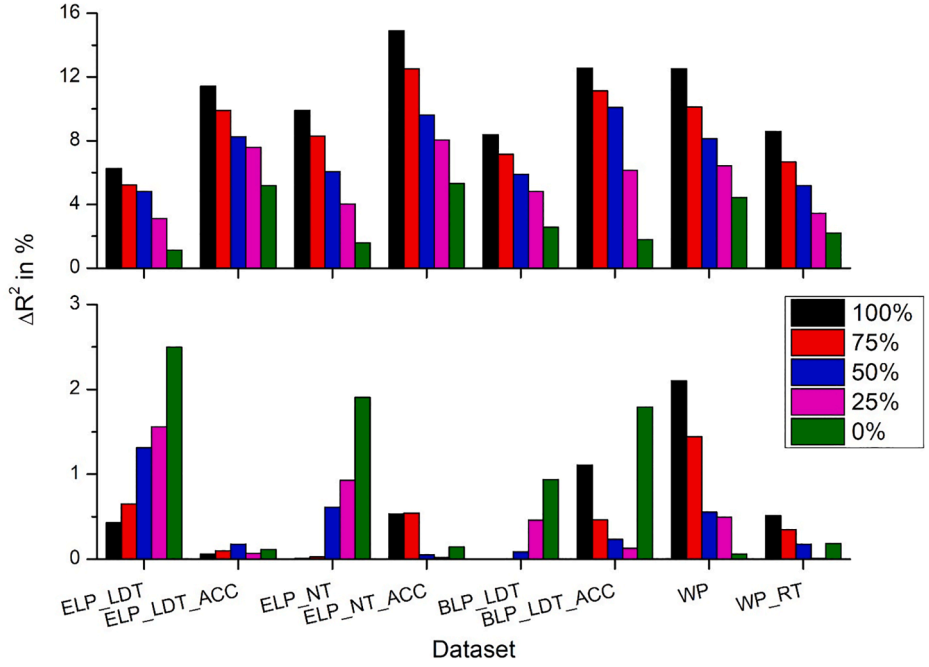

Fig. 6. The amount of unique variance that the SDM (top panel) and WF (bottom panel) account for across the various levels of training material randomization, with 100% signalling that each corpus is completely intact and 0% signalling a completely randomized user corpus.

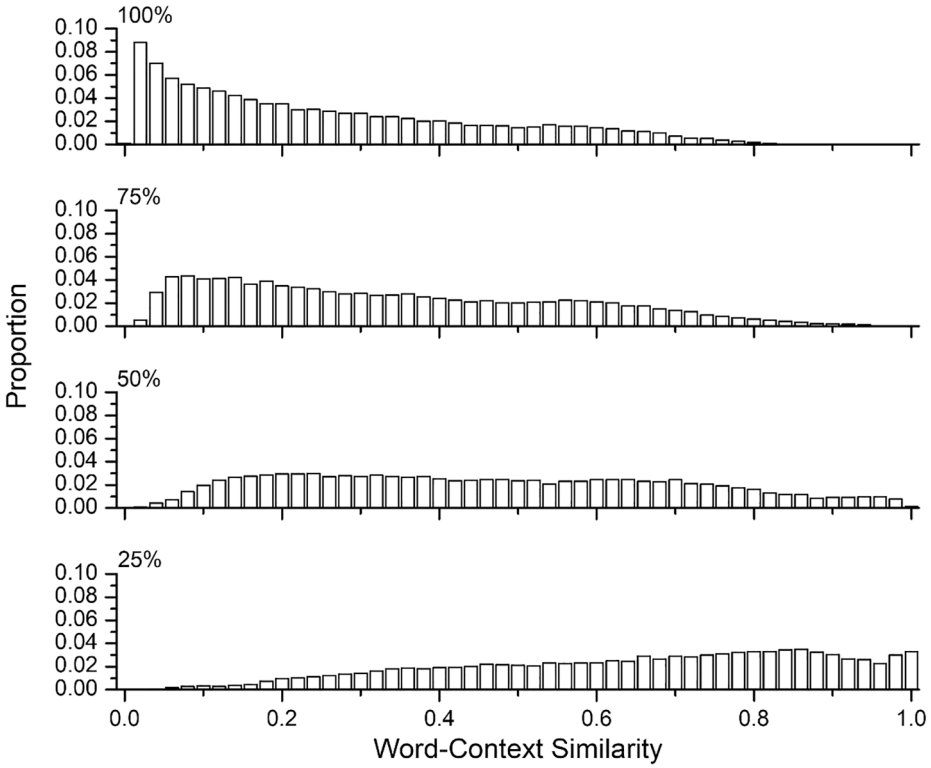

Fig. 7. Histograms of one million word-context similarity values attained from 50 random samples from each of the 20,000 user contexts across the 100%, 75%, 50%, and 25% intact training materials.

how a word is used). Indeed, by taking away the distinctiveness of each user's corpora (through a reduction of its intactness), the model's ability to isolate unique context types is reduced. As the randomness of the corpora are increased, they all drift toward average samples of language, rather than containing the uniqueness of an individual person's communication pattern. This structure also informs the differences in the optimal λ parameter that was found in Fig. 5. As the similarity distribution is increased, there are fewer distinctive contexts. A high λ parameter would reduce these values to be too small to have an impact on a word's lexical strength. Thus, λ must be decreased to still weight distinctive events strongly. As the distribution becomes more negatively skewed, the number of distinctive events disappears (hence the small λ

values for the 25% and 0% intact training materials in Fig. 5).

Importance of highly distinct contexts

To explore the importance of low similarity contexts, we conducted an additional simulation to determine the fit that a count of low similarity contexts would provide to the different lexical organization datasets. This was accomplished by setting a threshold on the SD values that the model calculates (contained in equation (2)), such that if any SD value for a word exceeds the criterion, then the word's strength is incremented by 1 for that context (rather than a continuous value in the standard implementation). The resulting model will be referred to as the SDM_Count model. If it is found that the SDM_Count model provides a comparable fit to the standard implementation, it would suggest that most of the power of the UCD_SD model comes from the occurrence of highly distinctive contexts. The SDM_Count model was fit by optimizing both the λ parameter and the SD criterion using a grid search algorithm, where all λ values between 0 and 300 were tested in steps of 1 and all criterion values were tested between 0 and 1 in steps of 0.005, conjunctively. The three datasets (ELP, BLP, and WP) were optimized separately. For the ELP data the optimal λ parameter was 135 and the optimal criterion was .425, for the BLP data the optimal λ parameter was 84 and the optimal criterion was .425, while for the WP data these values were 42 and .41 respectively.

The results of this simulation are displayed in Table 3 alongside the fits of the standard SDM trained on the sampled user corpora for comparison. This table shows that the SDM_Count achieves a comparable fit to the standard implementation of the model, signaling that the count-based approach to counting distinctive context types works. However, the interesting aspect of this simulation is that for the ELP data only 2.2% of the contexts that a word occurred in were used to update a word's strength for the SDM_Count model, while for the BLP data it was 4.26% and for the WP it was 8.32%. That is, an equivalent performance was achieved for the SDM_Count model by using only a small minority of the contexts that the model processed to update the strength of words in the lexicon. This suggests that the power of the SDM arises from its ability to index the discrete types of contexts that a word occurs in and not a continuous updating of a word's strength across all of its contextual occurrences. However, the standard SDM implementation contains one less parameter and has an equivalent level of fit, and so is more parsimonious, but this result demonstrates the underlying reason for the SDM's success.

The ability to identify highly distinct contexts, the information source that the SDM_Count model is capitalizing upon, relies on an accurate representation of context (and in turn, words). There are two main aspects of this contextual representation that have not yet been explored: the number of user comments that are included in constructing a context representation and the number of discourse types are used as contextual features.

Table 3
Correlations between the SDM_Count and SDM and the lexical datasets.

	SDM_Count	SDM
ELP_LDT	-.679	-.681
ELP_LDT_Acc	.54	.54
ELP_NT	-.585	-.585
ELP_NT_Acc	.433	.433
BLP_LDT	-.671	-.673
BLP_LDT_Acc	.672	.661
WP	.748	.748
WP_RT	-.761	-.762
Average |r|	.636	.635

Note. N = 39,948 for ELP data; N = 28,065 for BLP data; N = 57,716 for WP data; all correlations are significant at the $p < 0.001$ level.

Impact of number of user comments

The first aspect of manipulating the context representation of the model that will be explored is the number of user comments that are used to form a context representation. This manipulation matters because the number of user comments has a direct consequence on the fidelity of a contextual representation – as more comments are used to form a context representation, a more accurate representation of a user's communication pattern is attained. There are also practicality concerns about manipulating this factor, as it is unlikely that an individual person has a perfect memory of the tens of thousands of people that one has interacted with previously. Thus, if it is found that the model does not provide an advantage over WF when only a moderate amount of user comments is included in a user's context representation, it would suggest that the operationalization of context at the individual level is implausible. This simulation is the closest possible to the context size manipulation conducted by Hollis (2020) – with the UCD-SD model it is impossible to manipulate the size of a context, as each context is a single individual. However, by manipulating the number of comments in the construction of the model's context representation, the amount of information that is contained in its representation is changed.

To determine if the number of user comments included in forming a context representation significantly impacts model performance, a simulation was conducted manipulating the number of comments included from 10% to 100% in steps of 10%. At the 10% level, only 10% of a user's comments were included in forming the context representation, while at 100% all user comments were included (the standard model as previously tested). Performance of the model will be assessed using the same regression technique as previously utilized, across all of the eight previously used datasets. The λ parameter was fit independently at each sampling level, as manipulating the context representation changes the distributional structure of the of the similarity values that are derived, which impacts what the best fitting λ parameter is. The model was trained on the smaller 20,000 user corpora described previously.

The results of the simulation are displayed in Fig. 8. This figure shows that even at the lowest number of comments included, 10%, the model still offers a considerable improvement over WF across the different datasets - roughly a 7% increase in variance accounted for, which increases to approximately 11% when all comments are included. Even with a minimal amount of information from which to derive

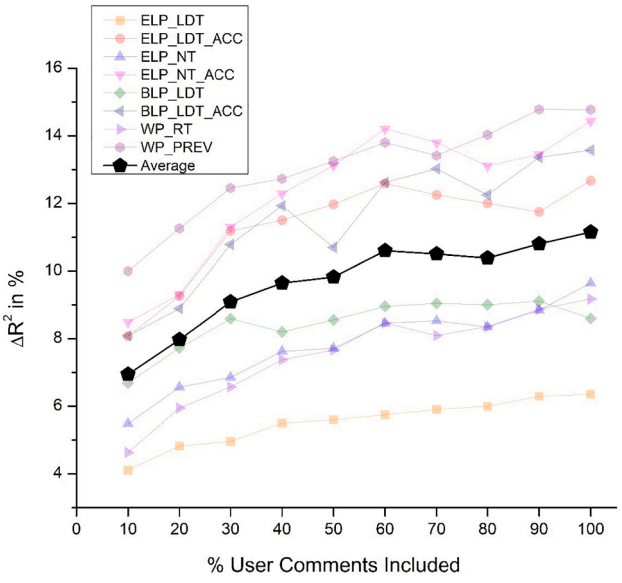

Fig. 8. Performance of the UCD-SD-PR model as a function of the percentage of user comments included in the formation of a context representation.

context representations, the model is still able to substantially outperform WF. This suggests that the model does not need all of a user's comments to form coherent context representations, but that even with a limited subset the communication pattern of a single individual can be represented. As more information about a user is considered, performance increases, suggesting that more complete information about a person's communication patterns allows for better accounting of lexical strength.

Impact of number of discourse features

The final aspect of the model that will be manipulated is the number of discourse features that are used to form a context representation. In the standard version of the UCD-SD-PR model, each of the roughly 30,000 subreddits that were communicated in by the selected users were included as features of a context. Thus, a context representation for a user is the number of comments the user produced in each possible discourse. By limiting the number of discourses, the resolution of the contextual representation is reduced, as the variability of an individual's communication pattern is more limited. To determine the impact of the discourse dimensionality on model performance, the discourse representation was rank-ordered such that the first vector dimension was the subreddit with the most user comments (r/AskReddit), followed by the second most popular subreddit, etc… and the final location in the context representation was the subreddit with the smallest number of user comments. Then, the performance of the UCD-SD-PR model was sampled at 100, 500, 1000, 2000, 5000, 10000, and 30,000 vector lengths. Model training was consistent with the previous simulation.

The result of this simulation is presented in Fig. 9 for reaction time data (top panel) and accuracy data (bottom panel). This figure shows that when only a small number of discourses are included, model performance suffers. However, once there is a sufficient level of discourses (i.e., 1,000 discourse features) included in a context's representation there is only a moderate level of improvement as more discourses are added into the model's representation. Combined with the simulation contained in Fig. 8, this suggests that the UCD-SD-PR model is relatively resilient to a decay in the resolution of the context representation that the model forms. Even with only partial knowledge about a user's communication patterns, the model is still able to form accurate predictions about a word's strength in memory that provides a large improvement over word frequency across all datasets evaluated.

General discussion

The goal of this article was to elucidate the role of semantic content on measures of contextual diversity. Measures of contextual diversity aim to replace the classic word frequency measure with an estimate of contextual occurrence (Adelman et al., 2006; Jones et al., 2017; Caldwell-Harris, 2021). Doing so would greatly constrain the field of possible mechanisms that humans use to organize their lexical knowledge. The importance of contextual diversity was recently evaluated by Hollis (2020), who demonstrated that when other aspects of linguistic contexts (such as word burstiness) are accounted for, contextual diversity offers a very small advantage (if any at all) over word frequency when using the standard operationalization of contextual diversity as a document count. This result seems to be in opposition to the recent findings of Johns et al. (2020) and Johns (2021) who demonstrated large and systematic advantages of diversity counts over word frequency when more naturalistic and ecologically valid notions of linguistic contexts are used to measure contextual diversity. Additionally, in these previous studies very substantial advantages were found for contextual diversity measures over word frequency across many large datasets when these measures were transformed with the Semantic Distinctiveness Model (SDM) of Jones et al. (2012; see Fig. 1).

The SDM provides a weight ranging from 0 to 1 for each context that a word occurs in, through the use of an expectancy-congruency mechanism based on semantic similarity. This signal represents surprisal at the current semantic content in the environment relative to the current contents of memory. A value of 0 signals that a context is completely redundant with past experience, while a value of 1 signals a completely new type of context that a word occurred in. Although there have been numerous previous instantiations of this model, the best fitting current implementation of the model is the UCD-SD-PR model described by Johns (2021; entitled UCD-SD in this article). In this model, a context is a single person (operationalized as a high-volume commenter on the internet forum Reddit), and the context representation that the model uses is the pattern of commenting that a user produced across different discourses. This model provides large and systematic improvements

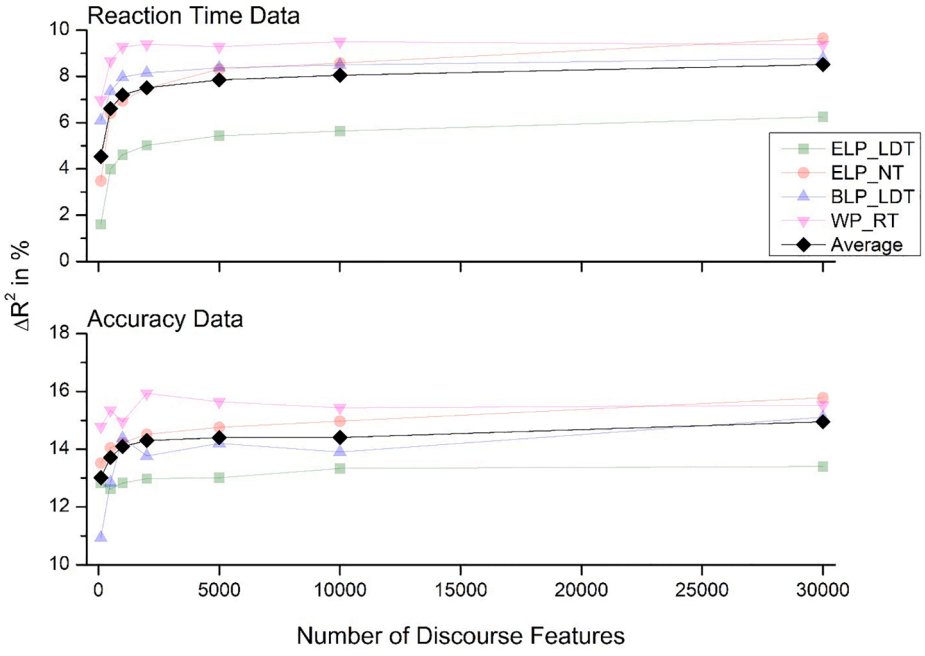

Fig. 9. Impact of the number of discourse features on the performance of the UCD-SD-PR model.

over a word frequency count (see Fig. 1), especially when compared to past contextual diversity measures, while reducing or eliminating the amount of unique variance that word frequency accounts for.

To determine how the semantic construction of a context impacts a contextual diversity measure, Hollis (2020) constructed randomized contexts and measured contextual diversity from these randomized contexts and found that there was little reduction in explained variance by the resulting CD measure. This raises the question of how the content of a context affects the calculation of a word's lexical strength. In the traditional, count-based, operationalization of contextual diversity, the contextual content is unimportant, as it is only the occurrence of a word within a contextual unit that matters. However, in the SDM the strength that a contextual occurrence provides a word is directly dependent on the distributional overlap between a word and context representation. This means that for the SDM, unlike a count-based CD measure, the composition of the training materials is integral to the operation of the model. Thus, the results of Hollis (2020) left an open question as to what the impact of the composition of training materials has on a model calculating contextual strength measures based on the semantic overlap of words and contexts - a question this article aimed to answer.

For our first simulation, we took 20,000 user corpora from Johns (2021) and randomized them at different levels. Each user corpus consists of the comments made by that user on Reddit. To randomize these corpora, a set number of comments were retained, and the rest were replaced with comments from other random users. It was found that as the SDM was trained with increasingly random materials, the explanatory power of the model decreased. When trained with completely randomly constructed user corpora, the model offered little to no improvement over word frequency for most datasets. This result demonstrates that in order to get an accurate representation of contextual occurrence, it is necessary to have consistent semantic representations of linguistic context.

Subsequently, we analysed the word-context similarity structures of the intact and randomized corpora. It was found that for the intact training materials, the similarity distributions were negatively skewed, with most similarity values being relatively low. As the training materials were increasingly randomized, this structure was lost as the training materials became positively skewed. This suggests that the users who produced a word in their communications are relatively dissimilar in their overall communication patterns, and this diversity in contextual variation is what the SDM capitalizes upon.

These simulations demonstrate that the model is dependent on low similarity (or, semantically distinct) contexts in a word's event history, as these provide the bulk of a word's lexical strength (Johns, 2021), due to the optimal λ parameter being relatively high. The λ parameter controls the weighting that is given to high versus low similarity contexts, and at high levels vastly reduces the contribution of redundant contexts to a word's lexical strength. For example, Johns (2021) demonstrated that highly distinct contexts make up approximately 1% of a word's total occurrences but contribute approximately 50% to a word's lexical strength. The importance of highly distinct events makes sense from a likely need perspective, as these events signal a new type of context that a word could be used in. However, the ability to determine these highly distinct events requires a semantic representation of a context, which is not possible with a count-based measure, as a standard contextual diversity count measure interprets each occurrence of a word in a context as equal in importance. When the semantic diversity of the contexts that a word occurs in is removed through randomization, the model no longer provides a significant advantage over count-based measures.

In order to demonstrate the importance of highly distinct events in lexical organization, we constructed the SDM_count model, where a word's lexical strength was increased only when it exceeded a set criterion. We found that by fitting both the λ and criterion parameter, the SDM_count model achieved similar levels of fit as the UCD-SD model, even though the model used only a small percentage of contexts of a word's contextual occurrences to update a word's strength. This simulation demonstrates that the advantage that the SDM capitalizes on is the ability to discriminate between the types of contexts that a word occurs in (in the case of the UCD-SD, the type of person who would use a word). When only highly distinct contexts are used to increase a word's strength there is little drop in model performance, suggesting that it is the occurrence of a word in a new type of context that matters to a word's lexical strength, not just a count of contextual occurrences as the original contextual diversity measure of Adelman et al. (2006) assumed. Additionally, this suggests that lexical organization models based on word frequency vastly overestimate the contribution of individual word occurrences to a word's lexical strength. Furthermore, Johns (in press a) recently used the SDM_Count model to account for word-level variability in episodic recognition and found that the SDM_Count model exceeded performance of the continuous implementation, suggesting that this is a viable method of calculating a word's lexical strength in memory across multiple domains.

The final set of simulations reported the impact that manipulating multiple components of a model's contextual representation had on model performance. In particular, the number of user comments that were used to form a context representation and the number of discourse features that were used in a context representation were manipulated. It was found that both sources of information modulated the performance of the model. However, the model still accounted for much greater levels of variance over word frequency even at the most limited levels of manipulation. These findings suggest that the model is resilient to limited amounts of information contained in a contextual representation. This is an important finding for the ecological validity of the UCD-SD model – given that the model is interpreting a context at the individual level, it is unlikely that people have perfect memory of the communication patterns of all the people that they have interacted with previously. Indeed, these simulations show that even with a relatively small amount of information about an individual's communication pattern, the model is able to capitalize on this information to construct superior estimates of a word's strength in memory.

The theoretical justification given for the importance of contextual diversity comes from the principle of likely need as proposed by the rational analysis of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991), as first proposed by Adelman et al. (2006, 2008). This theorical position was subsequently explored by Jones et al. (2017) and Johns (2021; see Westbury, 2020 for a related proposal in a different area of psycholinguistics) in relation to lexical organization. The principle of likely need states that words that occur in more contexts are more likely to be needed in any future context, and so should be the most accessible in memory. However, this instantiation of this principle ignores the use of prediction, which has been shown to play a fundamental role in language processing (e.g., Altmann & Mirkovic, 2009; Federmeier, 2007; Kutas & Federmeier, 2011). From a combined predictive and likely need perspective, the important aspect of a word's occurrence is the type of contexts that a word occurs in, not the total number, as the occurrence of a word could be predicted based upon the past contexts a word has occurred in. The SDM_count model demonstrates that a direct operationalization of this theory results in a similar level of performance to a continuously updating model. This finding suggests that the lexical experiences that are used to update a word's strength in the lexicon are relatively few, and based on the underlying diversity of the context space that a word occurs in.

From a mechanistic perspective, the finding of the importance of highly distinctive contextual occurrences suggests a need for an updating of models of lexical organization. The majority of previous models of lexical organization and word recognition utilize word frequency in their organizational frameworks in some fashion (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Murray & Forster, 2004; Norris, 2006). For these models to integrate the importance of semantically distinct contextual occurrences, the models will need to be integrated with models of semantic processing, an important area for future research. The ability to accomplish this suggests a requirement for models of

lexical organization to be integrated with distributional models of semantics in order to generate more unified computational models of language processing.

Hollis (2020) criticized the operationalization of traditional CD measures as being ecologically questionable, in that the notions of context that were used in previous CD measures (e.g., paragraphs, documents, moving windows) did not map onto realistic aspects of human experience. Although Adelman et al. (2006, 2008) interpreted their original CD measure as examining temporal effects of contextual occurrences, rather than lexical semantic, the work of Jones et al. (2012) and Johns et al. (2012, 2016, 2020, 2021) interpreted CD in lexical semantic terms, so this was the argument examined here. We believe that the new large-scale measures of context described in Johns (2021) provide a better, more ecologically valid, notion of linguistic context than previous measures do. Indeed, this article has explored the underlying reasons for the success of this approach and has validated the underlying principles of the modeling approach. Specifically, the UCD-SD model explored here operationalizes context at the individual language user level, a more interpretable notion of context, and one that is theoretically justified by adaptive and usage-based theories of language processing (e.g., Beckner, et al., 2009; Christiansen & Chater, 2008; Tomasello, 2003, 2009).

The original motivation for the development of the SDM in Jones et al. (2012) was to determine the importance of the diversity of the semantic content of the contexts that a word appears in. This question has been examined in similar work elsewhere (e.g., Cevoli, Watkins, & Rastle, 2020; Hoffmann, Lambon Ralph, & Rogers, 2012; Hsiao & Nation, 2018), however the unique aspect of the SDM is that it generates a superior measure to word frequency by simply considering the semantic redundancy of the contexts that a word appears in across its event history. Given the magnitude of the advantage for the SDM over word frequency in Figs. 1 and 3, it is a worthwhile question to determine what role word frequency plays in lexical organization, which seems to be relatively small in the results reported here. However, other datatypes may see an increased contribution of word frequency, for example in idiomatic processing (Senaldi et al., in press), where it was found that a CD measure accounted for the most variance, but WF still accounted for a significant amount.

In addition, the mechanisms used by the model used here are comparatively simple. The results of this article, combined with the previous results in this line of research (e.g., Johns, 2021; Johns et al., 2020; Jones et al., 2012, 2017) demonstrate that the content of linguistic experience matters in lexical organization. It is our hope that with the materials released here, combined with new advances in distributional modeling techniques (see Günther et al., 2019; Kumar, 2020), better models of contextual diversity can be generated by refining the semantic representation of linguistic contexts. This should lead to increasingly better explanations of large sets of lexical organization data, an important trend in the computational cognitive sciences (Johns, Jamieson, & Jones, 2020).

CRediT authorship contribution statement

Brendan T. Johns: Conceptualization, Methodology, Software, Formal analysis, Resources, Writing – original draft, Writing – review & editing. **Michael N. Jones:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2020-04727 to BTJ.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*, 814–823.
Adelman, J. S., & Brown, G. D. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*, 214.
Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*, 583–609.
Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703–719.
Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396–408.
Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.
Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *arXiv preprint* arXiv:2001.08435.
Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*, 1–26.
Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review, 74*, 1–15.
Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*, 45–50.
Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*, 467–479.
Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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*, 977–990.
Caldwell-Harris, C. L. (2021). Frequency effects in reading are powerful–But is contextual diversity the more important variable? *Language and Linguistics Compass, 15*(12), Article e12444.
Cevoli, B., Watkins, C., & Rastle, K. (2020). What is semantic diversity and why does it facilitate visual word recognition? *Behavior Research Methods*.
Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*, 489–509.
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.
Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627–635.
Frances, C., De Bruin, A., & Duñabeitia, J. A. (2020). The influence of emotional and foreign language context in content learning. *Studies in Second Language Acquisition, 42*, 891–903.
Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*, 1006–1033.
Hamrick, P., & Pandža, N. B. (2020). Contributions of semantic and contextual diversity to the word frequency effect in L2 lexical access. *Canadian Journal of Experimental Psychology, 74*, 25–34.
Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2012). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*, 718–730.
Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language, 114*, Article 104146.
Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language, 103*, 114–126.
Johns, B. T., Dye, M. W., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review, 23*, 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., Jamieson, R. K., & Jones, M. N. (2020). The continued importance of theory: Lessons from big data approaches to cognition. In S. E. Woo, R. Proctor, & L. Tay (Eds.), *Big Data Methods for Psychological Research: New horizons and Challenges*. APA Books.
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.
Johns, B. T., & Jones, M. N. (2008). In *Predicting word-naming and lexical decision times from a semantic space model* (pp. 279–285). Austin, TX: Cognitive Science Society.
Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology, 73*, 841–855.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science, 43*, Article e12730.
Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*.
Johns, B. T. (in press a). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*.
Johns, B. T. (in press b). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. *Memory & Cognition*.
Jones, M. N., Dye, M., & Johns, B. T. (2017). *Context as an organizational principle of the lexicon*. The Psychology of Learning and Motivation.
Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*, 115–124.
Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology, 166*, 190–211.
Krueger, L. E. (1975). Familiarity effects in visual information processing. *Psychological Bulletin, 82*, 949–974.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.
Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287–304.
Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 1–41.
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*, 211–240.
Madsen, R. E., Kauchak, D., & Elkan, C. (2005). August). Modeling word burstiness using the Dirichlet distribution. In *In Proceedings of the 22nd international conference on Machine learning* (pp. 545–552).
Mak, M. H., Hsiao, Y., & Nation, K. (2021). Anchoring and contextual variation in the early stages of incidental word learning during reading. *Journal of Memory and Language, 118*, Article 104203.
Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods, 52*, 741–760.
McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*, 295–322.
Murray, W. S., & Forster, K. I. (2008). The rank hypothesis and lexical decision: A reply to Adelman and Brown (2008). *Psychological Review, 115*, 240–251.
Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.
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*, 37–44.
Plante, E., Ogilvie, T., Vance, R., Aguilar, J. M., Dailey, N. S., Meyers, C., et al. (2014). Variability in the language input to children enhances learning in a treatment context. *American Journal of Speech-Language Pathology, 23*(4), 530–545.
Qiu, M., & Johns, B. T. (2020). Semantic diversity in paired-associate learning: Further evidence for the information accumulation perspective of cognitive aging. *Psychonomic Bulletin & Review, 27*, 114–121.
Rosa, E., Salom, R., & Perea, M. (2022). Contextual diversity favors the learning of new words in children regardless of their comprehension skills. *Journal of Experimental Child Psychology, 214*, Article 105312.
Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS One, 12*(6), Article e0179004.
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*, 499–520.
Senaldi, M. S. G., Titone, T., & Johns, B. T. (in press). Determining the importance of frequency and contextual diversity in the lexical organization of multiword expressions. *Canadian Journal of Experimental Psychology*.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
Tapia, J. L., Rosa, E., Rocabado, F., Vergara-Martínez, M., & Perea, M. (in press). Does narrator variability facilitate incidental word learning in the classroom?. *Memory & Cognition*.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge. MA: Harvard University Press.
Tomasello, M. (2009). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
Westbury, C. (2020). Prenominal adjective order is such a fat big deal because adjectives are ordered by likely need. *Psychonomic Bulletin & Review, 28*, 122–138.