Memory & Cognition (2022) 50:1013–1032
https://doi.org/10.3758/s13421-021-01249-z

Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity

Brendan T. Johns[1]

Accepted: 4 October 2021 / Published online: 22 November 2021

Abstract

Contextual diversity modifies word frequency by ignoring the repetition of words in context (Adelman, Brown, & Quesada, 2006, *Psychological Science, 17*(9)*,* 814–823). Semantic diversity modifies contextual diversity by taking into account the uniqueness of the contexts that a word occurs in when calculating lexical strength (Jones, Johns, & Recchia, 2012, *Canadian Journal of Experimental Psychology, 66,* 115–124). Recent research has demonstrated that measures based on contextual and semantic diversity provide a considerable improvement over word frequency when accounting for lexical organization data (Johns, 2021, *Psychological Review, 128,* 525–557; Johns, Dye, & Jones, 2020a, *Quarterly Journal of Experimental Psychology, 73,* 841–855). The article demonstrates that these same findings generalize to word-level episodic recognition rates, using the previously released data of Cortese, Khanna, and Hacker (Cortese et al., 2010, *Memory*, 18, 595–609) and Cortese, McCarty, and Schock (Cortese et al., 2015, *Quarterly Journal of Experimental Psychology*, *68,* 1489–1501). It was found that including the best fitting contextual diversity model allowed for a very large increase in variance accounted for over previously used variables, such as word frequency, signalling commonality with results from the lexical organization literature. The findings of this article suggest that current trends in the collection of megadata sets of human behavior (e.g., Balota et al., 2007, *Behavior Research Methods, 39*(3)*,* 445–459) provide a promising avenue to develop new theoretically oriented models of word-level episodic recognition data.

Keywords Recognition memory · Word frequency · Corpus-based models · Distributional semantics · Computational modeling

The impact of word frequency (WF) on recognition memory has been a cornerstone effect in the study of episodic memory. Words that are low in frequency tend to be correctly recognized and correctly rejected at a greater rate than words that are high in frequency, a phenomenon referred to as the mirror effect of frequency (Glanzer & Adams, 1985, 1990). The mirror effect has become a standard testing point for computational models of recognition memory (e.g., Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Recent studies on item-level effects in recognition memory has confirmed the importance of word frequency in predicting recognition memory performance at the individual word level (Cortese et al., 2010; Cortese et al., 2015).

Word frequency is calculated by counting the number of times that a word occurs across a corpus of natural language. The first large set of word frequency values were collected by Kučera and Francis (1967), who measured the frequency of words in a sample of approximately 1 million words taken from a variety of sources, such as newspaper articles and fiction novels. Kučera and Francis faced an uphill battle in calculating these values, given the technological limitations of the time. Modern researchers are no longer burdened by these problems. It is now possible to calculate word frequency from much larger sources of language, some spanning tens of millions or billions of words, including television and movie subtitles (Brysbaert & New, 2009), newspaper articles (Davies, 2009), fiction books (Johns, Dye, & Jones, 2020a; Johns & Jamieson, 2019), textbooks (Landauer & Dumais, 1997), social media (Herdağdelen & Marelli, 2017; Johns, 2019) and online encyclopaedias (Shaoul & Westbury, 2010), among others.

✉ Brendan T. Johns
brendan.johns@mcgill.ca

[1] Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, Quebec H3A 1G1, Canada

Coupled with the ability to collate large sources of text is the development of corpus-based cognitive models designed to examine different theoretical constructs of lexical strength (for a review, see Jones et al., 2017). A notable development is the contextual diversity (CD) account of Adelman et al. (2006, see also McDonald & Shillcock, 2001, for a similar proposal and Steyvers & Malmberg, 2003, for preceding work in recognition memory). Adelman et al. (2006) proposed that when calculating the strength of a word that the repetitions of a word in a context should be ignored, with context in this work typically defined by relatively small units of language such as a document or paragraph in a corpus (Hollis, 2020; Johns, et al., 2012; Jones et al., 2012). This work is based off ideas formulated within the rational analysis of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991), which proposes that words that appear in a greater number of contexts are more likely to be needed in a future context (and hence should be more available in the lexicon). It has been shown across multiple corpora and data sets that a CD count provides a better fit to lexical organization data compared with a WF count (e.g., Brysbaert & New, 2009). Additionally, the impact of contextual diversity has been established using artificial stimuli in episodic recognition (Nelson & Shiffrin, 2013), cross-situational learning (Kachergis et al., 2017), and lexical decision (Johns et al., 2016a; Mak et al., 2021).

Key to calculating CD is a definition of linguistic context, a subject of recent theoretical work in the study of lexical organization. Johns, Dye, and Jones (2020a) provided a first examination into the impact of operationalizing context at much larger units of analysis than had previously been considered. This work was prompted by the large-scale data collection of Brysbaert et al. (Brysbaert et al., 2019; see also Brysbaert et al., 2016; Keuleers et al., 2015), who measured the proportion of words that a population of language users recognized as a member of their language, entitled word prevalence. Johns et al. (2020a) constructed two new measures of contextual diversity to account for this data, one at the book level and one at the author level, by assembling and organizing a set of approximately 25,000 fiction novels. Johns et al. (2020) labeled the measures book prevalence (BP) and author prevalence (AP). For the BP measure, if a word occurred in a book, the strength of that word was increased (with word repetition within a book being ignored). For the AP measure, if a word occurred in an author's writings, then it was increased in strength (with repeated usage of a word ignored across the totality of an author's written materials). It was found that measuring CD at these levels significantly increased the fit of the measures over WF and smaller definitions of context.

However, using a book or an author as a definition of linguistic context is theoretically muddled, as they fail to satisfy an ecologically valid notion of linguistic context. To overcome these issues, Johns (2021) recently constructed new CD measures from a communicatively-oriented source of language. Specifically, two new measures were proposed, user contextual diversity (UCD) and discourse contextual diversity (DCD), attained from analyzing the communication patterns of hundreds of thousands of individuals across tens of thousands of discourses on the internet forum Reddit (total words analyzed was approximately 55 billion), attained from Baumgartner et al. (2020). UCD is a count of the number of users who had used a word in their communications, while DCD is a count of the number of discourses (subreddits) that a word was used in. It was found these measures provided benchmark fits to a variety of lexical organization data, especially when transformed with the semantic diversity model (Jones et al., 2012; see below for more detail). The results of this work suggest that by using contextual diversity based on properties of the social environment that individuals are embedded in (such as the discourses where communication takes place or the people who one communicates with), a better accounting of the organization of the lexicon can be attained.

These results prompt the question as to whether the importance of new theoretical measures of word strength generalize to item-level effects in recognition memory. Luckily, Cortese et al. (Cortese et al., 2010; Cortese et al., 2015) has published item-level recognition rates for large sets of monosyllabic (Cortese et al., 2010) and disyllabic words (Cortese et al., 2015), which will allow a similar analysis that was done in Johns (2021) with lexical organization data to also be done on recognition memory performance. There is evidence that disyllabic words show a larger effect of semantic variables (e.g., Cortese & Schock, 2013), so it is expected that there should be a bigger effect in the disyllabic word collection.

There will be two types of contextual diversity models tested: (1) count-based models and (2) semantic diversity models. Count-based models are based upon the contextual diversity measure of Adelman et al. (2006), with a modification to the context size of a model. That is, they are counts of the number of contexts that a word occurs in, where the contextual unit is manipulated. The semantic diversity-based models will modify the count-based measures by using the semantic distinctiveness model (SDM; Jones et al., 2012) to modify the weight given to a context. The difference between the two model types is that for count-based models each context that a word occurs in increases that word's strength in memory by 1, while for the semantic diversity-based models each context increases a word's strength in memory with a continuous value between 0 and 1. This continuous value is calculating by weighting unique contexts as being more important to the lexical strength of a word, and has been shown to provide a benefit across a number of different empirical examinations, including artificial language learning (Jones et al., 2012), spoken word recognition (Johns et al., 2012a), natural language learning (Johns et al., 2016a), bilingualism and aging (Johns et al., 2016b; Qiu & Johns, 2020), and large collections of lexical decision, naming, and word prevalence data (Johns, 2021; Johns et al, 2020a; Jones et al., 2012). Determining whether this advantage holds for item-level recognition memory performance will assess the generality of the advantage for contextual diversity measures across different behavioral data types.

Past research in episodic memory reinforces the notion that contextual diversity measures may be the driving force behind frequency effects in recognition memory. Dennis and Humphreys (2001) propose that it is contextual overlap between past experiences and the current episode that drives recognition memory performance (see also Popov & Reder, 2020; Reder et al., 2000, for a similar proposal). Lohnas et al. (2011) found that items with greater contextual variability were better able to recalled, while Steyvers and Malmberg (2003) found that greater contextual variability impairs recognition performance (see also Aue et al., 2018). Qiu and Johns (2020) demonstrated that a word's contextual and semantic diversity impacts paired-associate learning across aging. Taken together, these results suggest a lexical strength measure based on contextual occurrence, rather than frequency, could provide a better fit to word-level recognition memory data.

Indeed, the motivations for the development of the SDM (Johns et al., 2020a; Jones et al., 2012) lie in the concept of the impact of distinctiveness on episodic memory performance. The study of distinctiveness on episodic memory performance dates back to the classic von Restorff effect (von Restorff, 1933; see also Hunt, 1995; MacLeod, 2020), where it was found that a unique stimulus (e.g., a word written in red) in a field of redundant stimuli (e.g., words written in black) has improved memorability. Distinctiveness has played a central role in a variety of theoretical accounts of memory performance (e.g., Brown et al., 2007; Hunt & McDaniel, 1993; Neath & Crowder, 1990). The SDM embodies notions of distinctiveness by ascribing more weight to more unique contextual usages of a word, based on the semantic similarity between words and contexts. At the end of model training, those words that occur more often in semantically distinct contexts are those that have the highest strength in memory. If it is found that the SDM provides a better fit to item-level recognition rates, similar to what has been found with lexical organization data, it would suggest that episodic memory functions are integral to the storage and maintenance of lexical information in memory. Importantly, this article will contrast and compare different implementations of the SDM by utilizing different representations that map onto different properties of the lexical environment in order to determine the underlying best model type.

There are multiple goals of this article. The first is to determine whether the frequency values derived from the Reddit data described by Johns (2021) provide a better accounting of word-level recognition memory patterns compared with the current standard frequency norms that are used, namely the SUBTLEX norms of Brysbaert and New (2009). If it is found that these new norms provide a better ability to account for item-level trends in recognition memory, it would signal a new methodological tool to control stimuli in recognition memory experiments. A second goal is to determine if the pattern of findings on the importance of contextual diversity in lexical organization extends to a different data type—namely, word-level recognition memory patterns, along with the theoretical implications of such a finding (such as the role of episodic distinctiveness in the storage of lexical information in memory). The final goal is to demonstrate the usefulness of collecting word-level data using episodic memory tasks, such as what was done in Cortese et al. (2010; Cortese et al., 2015), in order to spur new theoretical developments in computational cognitive models of recognition memory performance. The first section of the paper will describe the different models, the data, and the analysis technique. Subsequently, a determination of the most parsimonious model will be conducted.

Modeling overview

Reddit data

The Reddit data of Johns (2021) was assembled from a website called pushshift.io (Baumgartner et al., 2020), which collects all Reddit comments for each month using the publicly available Reddit API,[1] and makes them available as database files. For the norms developed in Johns (2021) all comments from January 2006 to September 2019 were downloaded. Two types of corpora were assembled: (1) user corpora and (2) discourse corpora. The user corpora contained all of the individual Reddit users who had publicly available usernames who produced over 3,000 comments. The discourse corpora were the comments that these users made across subreddits. This data assembly resulted in 334,345 user corpora and 30,327 discourse corpora. Importantly, the sum total number of words contained in each set of corpora are identical, as they are composed of the same comments, it is only the organization of that information into different contextual units that differs. In total, there were approximately 55 billion words contained in each corpus

[1] Information on the API can be found at: https://www.reddit.com/dev/api/

set. More information on the information contained in these corpora can be found in Johns (2021).

Vocabulary

The vocabulary of the model was the combined words contained in the word prevalence data from Brysbaert et al. (2019) and the words from the English lexicon project (ELP; Balota et al., 2007) and the British lexicon project (BLP; Keuleers et al., 2012). This resulted in a word list consisting of 81,261 words. The values for all of the models reported here are available online for all of these words (available at btjohns.com/Johns_MC_CDvals.xlsx or https://osf.io/5nr6x/).

Count models

There will be four different contextual diversity count models used in the following analyses: (1) word frequency (WF), (2) contextual diversity (CD), (3) discourse contextual diversity (DCD), and (4) user contextual diversity (UCD). Since the user and discourse corpora contain identical comments, just organized differently, the WF and CD metrics from these corpora are identical. Word frequency is number of occurrences of a word across all comments. To calculate CD, the context size used was an individual comment (roughly analogous to a paragraph), with repetitions within a comment being ignored. Thus, CD is the total number of times a word occurred in a comment. As stated, DCD is the number of discourses a word was used in (thus, has a maximum value of 30,327), while UCD is the total number of users who used a word in their comments (thus, has a maximum value of 334,345).[2] Each variable used in the subsequent analysis (including the semantic diversity models described below) will be reduced with a natural logarithm, consistent with past research (Adelman & Brown, 2008; Jones et al., 2012).

Semantic diversity models

The semantic distinctiveness model (SDM; Jones et al., 2012) was developed in order to explain the effect of semantically diverse contextual information on lexical organization. The impact of semantic diversity on language processing is an active research area (e.g., Hoffman et al., 2013; Hsiao & Nation, 2018; Cevoli et al., 2021). The unique aspect of the SDM is that it examines the diversity of the contexts across learning. The underlying motivation of the SDM is to replace a CD count with a graded measure, such that each new context that a word appears in updates a word's strength in the lexicon relative to how unique that contextual usage of a word is, compared with the word's event history.

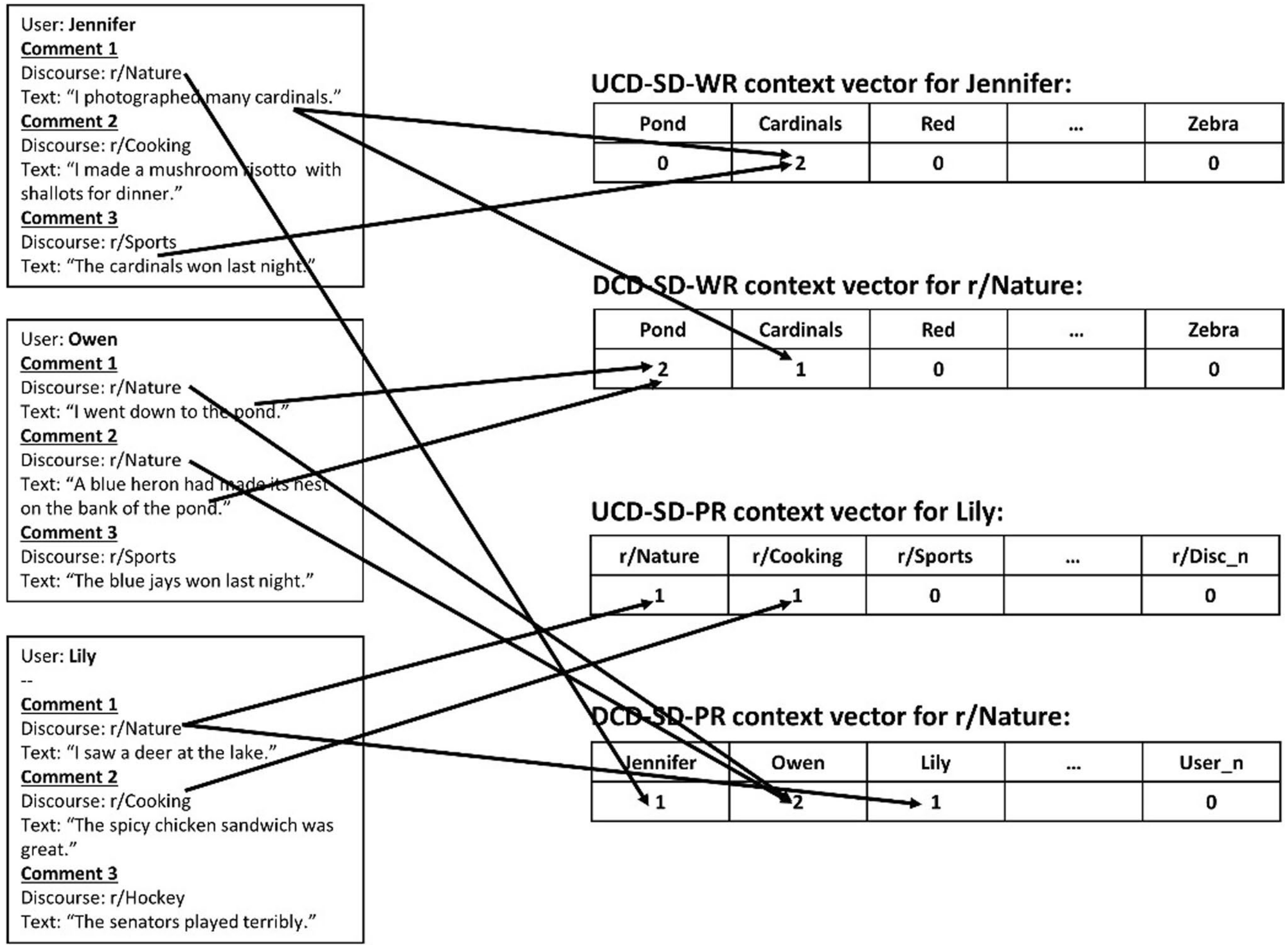
The SDM is a type of cognitive model entitled distributional models of semantics, which learn the meaning of words from word co-occurrence patterns in large text corpora (e.g., Jones & Mewhort, 2007; Landauer & Dumais, 1997; for a review, see Günther et al., 2019, and Kumar, 2021). Most distributional models focus on learning the meaning of words, but research on the SDM has focused on deriving lexical strength measures from this model type. This is accomplished through the use of an expectancy-congruency mechanism, where if a word appears in an unexpected context, then it is updated strongly. However, if the word occurs in an expected (i.e., consistent with past experience) context then it receives only a weak update to its lexical strength. This is similar to the use of an encoding mechanism where only information that was not previously known (e.g., information from non-redundant contexts) is encoded strongly into a word's lexical representation.

There have been multiple implementations of the model (e.g., Johns et al., 2014, 2020a; Jones et al., 2012), all providing a substantial increase in fit over WF and CD counts. In the first implementation of the model, described in detail in Jones et al. (2012), the model's memory store was a Word × Context matrix (equivalent to the classic distributional model type Latent Semantic Analysis; Landauer & Dumais, 1997). At the beginning of learning the matrix is empty. Each new context that the model encounters adds a new column to the matrix. For each word that occurred in a context, a semantic distinctiveness (SD) value was added into the column for that context. An SD value is the similarity between a word's representation and the representation of a context (which is the sum of all the word's representation that occurred in a context), transformed with an exponential density function (in order for high similarity contexts to have a low SD value, and vice versa). Words that did not occur in that context received a value of 0 in that column. An SD value is a graded measure between 0 and 1 that represents how unique that context is compared with the past context that a word occurred in. A word's lexical strength was then the sum of its entries within the matrix.

Although there are advantages to this implementation, such as a word's strength being distributed across its row in memory, it has its drawbacks as well. A primary issue with the model is its use of an ever-expanding matrix, which leads to high computational costs for large corpora that consist of a large number of contexts. The other is that the context representation does not scale to larger context sizes, such as a book or author level definition of context.

To overcome these limitations, Johns, Dye, and Jones (2020a) proposed a new implementation of the model. To scale up to larger context size, the memory representation

[2] In practice, no word hit these maximum values, because some user and discourse corpora only contain links and/or pictures.



User: **Jennifer**
Comment 1
Discourse: r/Nature
Text: "I photographed many cardinals."
Comment 2
Discourse: r/Cooking
Text: "I made a mushroom risotto with shallots for dinner."
Comment 3
Discourse: r/Sports
Text: "The cardinals won last night."

User: **Owen**
Comment 1
Discourse: r/Nature
Text: "I went down to the pond."
Comment 2
Discourse: r/Nature
Text: "A blue heron had made its nest on the bank of the pond."
Comment 3
Discourse: r/Sports
Text: "The blue jays won last night."

User: **Lily**
--
Comment 1
Discourse: r/Nature
Text: "I saw a deer at the lake."
Comment 2
Discourse: r/Cooking
Text: "The spicy chicken sandwich was great."
Comment 3
Discourse: r/Hockey
Text: "The senators played terribly."

UCD-SD-WR context vector for Jennifer:

Pond	Cardinals	Red	...	Zebra
0	2	0		0

DCD-SD-WR context vector for r/Nature:

Pond	Cardinals	Red	...	Zebra
2	1	0		0

UCD-SD-PR context vector for Lily:

r/Nature	r/Cooking	r/Sports	...	r/Disc_n
1	1	0		0

DCD-SD-PR context vector for r/Nature:

Jennifer	Owen	Lily	...	User_n
1	2	1		0

Fig. 1 An illustration of the different context representation types used for the socially-based CD models. The WR models form a context representation which is a count of the number of times that an individual used a word (the UCD-SD-WR model) or the number of times that a word was used within a discourse topic (the DCD-SD-WR model). The context vector of the UCD-SD-PR model is a count of the number of comments that a single user produced in each discourse topic, while the feature values for the DCD-SD-PR is the number of comments that each user produced in that discourse

for the model was changed to a Word x Word matrix. In this implementation, each context is the word frequency distribution of the words that occurred in a specified contextual unit (e.g., a book), while the memory for a word is the frequency distribution of all the contexts that a word occurred in across training. Context in Johns, Dye, and Jones (2020a) was defined either at the book level or at the author level (the combined books that a single author wrote).

Johns (2021) updated this model by changing the type of information that was contained in a word's representation, through the modification of both the UCD and DCD count. There were two representation types used: (1) word representations (WR) and (2) population representations (PR). The word representations were consistent with the implementation of Johns, Dye, and Jones (2020a), where the information contained in the representations was the word frequency distributions of either a user corpus or discourse corpus (for a matrix dimensionality of 81,261 × 81,261). However, the population representations were quite different than the WR models (signified with DCD-SD-WR and UCD-SD-WR), as they contained the communication pattern of users across discourses. For the DCD-SD-PR model, the representation consisted of the number of comments each user made in that discourse (so the model's matrix has a dimensionality of 81,261 × 334,345), while for the UCD-SD-PR the representation consisted of the number of comments a single user made across discourses (for a matrix dimensionality of 81,261 × 30,327). For the DCD models, the models are updated for each discourse, so received a total of 30,327 updates. For the UCD models, the models are updated for each user, and so received a total of 334,345 updates.

To clarify the representations that will drive the WR and PR models, Fig. 1 contains a pictorial demonstration

Table 1 Correlations between the different lexical strength variables

Measure	2	3	4	5	6	7	8
1. WF	.999	.947	.941	.952	.936	.949	.964
2. CD	.	.952	.943	.958	.941	.955	.967
3. DCD	.	.	.978	.993	.992	.999	.975
4. UCD	.	.	.	.962	.984	.977	.943
5. DCD-SD-PR	.	.	.	.	.987	.994	.981
6. UCD-SD-PR	.	.	.	.	.	.991	.964
7. DCD-SD-WR	.	.	.	.	.	.	.978
8. UCD-SD-PR	.	.	.	.	.	.	.

N = 2,897 for the disyllabic data All correlations significant at the p < .001 level.

of how the context vectors for the different model types are derived. The WR models form a representation of a context by counting the number of times each word was produced by an individual user across the discourses that they communicated within (the UCD-SD-WR model) or a count of the number times each word was produced within a discourse topic across all users (the DCD-SD-WR model). The PR models form a context vector by counting the number of comments that a single user produced across the different discourses (the UCD-SD-PR model) or the number of times each user commented within a single discourse (the DCD-SD-PR model). These context representations are used to update the word representations of each word that occurred in the specified contextual unit.

The DCD-SD-PR model measures the consistency of word usage of people within discourses. A word with a relatively high DCD-SD-PR strength would entail that the word is used across many different discourse types but is not produced by a consistent set of language users within those discourses. The UCD-SD-PR model is measuring the consistency of discourse communication patterns across individuals. A word with a relatively high UCD-SD-PR strength would indicate that the word is being used by many individuals across many discourses, but with no consistent pattern of discourse usage (i.e., it would be difficult to predict the discourse that an individual would use a word in). Johns (2021) found that the PR models significantly outperformed the WR models, demonstrating the importance of social and communicative information in lexical organization, with the UCD-SD-PR being the overall best fitting model. The goal of this article is to determine if this same pattern holds for item-level effects in episodic recognition.

Data and analysis technique

There will be two data sets analyzed in this article—the item-level recognition rates for monosyllabic words (Cortese et al., 2010) and disyllabic words (Cortese et al., 2015). Both of these studies used two different encoding conditions, but recognition rates were very similar across those manipulations, so the collapsed item-level rates were analyzed in this article. There were 2,578 words contained in the Cortese et al. (2010) data, while there were 2,897 words in the Cortese et al. (2015) data. It is important to note that these data sets are considerably smaller than the data sets used in lexical organization data, mainly due to the nature of the different tasks, but these smaller sample sizes provide limits in terms of theoretical conclusions compared with previous analyses.

Consistent with past studies (e.g., Adelman et al., 2006) it is necessary to use hierarchical linear regression to separate out the unique contributions of the different lexical strength variables. The end result of this analysis technique is the amount of predictive gain (measured as percentage ΔR^2 improvement) for one predictor over other competing predictors. The percent ΔR^2 is the amount of improvement that one variable causes over other controlled variables. For example, if the inclusion of an additional variable in a regression causes the R^2 of the model to be increased to 25% variance accounted for from 20%, this would represent a ΔR^2 of 20%. There will be nine different lexical strength measures compared in this study: (1) the classic SUBTLEX frequency measures (Brysbaert & New, 2009), (2) Reddit word frequency, (3) Reddit contextual diversity (CD), (4) DCD, (5) UCD, (6) DCD-SD-WR, (7) UCD-SD-WR, (8) DCD-SD-PR, and (9) UCD-SD-PR.

Results

As a first pass at understanding the various lexical strength measures, Table 1 contains the intercorrelations of the eight lexical strength measures derived from the Reddit data for the words contained in Cortese et al. (2015; this data set contained a larger number of words than Cortese et al., 2010). The results in this table are reflective of the results found in Johns (2021), where all variables are highly correlated with each other, and so regression analyses are needed to

Table 2 Correlations between the lexical strength variables and data sources

	Monosyllabic			Disyllabic		
	Hits	FA	Hits-FA	Hits	FA	Hits-FA
SUBTLEX	−.522	−.13	−.33	−.417	.057	−.323
WF	−.555	−.12	−.364	−.46	.204	−.456
CD	−.56	−.114	−.372	−.468	.21	−.466
DCD	−.483	−.03	−.37	−.461	.266	−.502
UCD	−.415	−.007	−.331	−.407	.267	−.466
DCD-SD-PR	−.535	−.051	−.397	−.482	.263	−.514
UCD-SD-PR	−.471	−.019	−.368	−.444	.274	−.497
DCD-SD-WR	−.504	−.038	−.381	−.461	.263	−.5
UCD-SD-WR	−.532	−.105	−.356	−.454	.212	−.459

N = 2,578 for the monosyllabic data; N = 2,897 for the disyllabic data; all correlations significant at the p < .001 level

separate out the individual contributions of each variable. The WF and CD variables are highly correlated to each other, but relatively less correlated to the DCD and UCD derived measures, signaling a divergence between different measures of contextual diversity.

In order to understand the connection between these variables and the data contained in Cortese et al. (2010; Cortese et al., 2015), Table 2 contains the correlations between the different lexical strength variables and the hit rate, false alarm (FA) rate, and hit-FA rate to both the monosyllabic and disyllabic data. For the disyllabic data, there is a considerable advantage for the Reddit-based data over the standard SUBTLEX frequency for all three datatypes. For the monosyllabic data, there is a smaller advantage for the Reddit-based WF values for the hits and hit-FA rate, while having a slight disadvantage for the FA rates (although all correlations are quite low to FA rates in this data set). One possible reason for the smaller fit to the monosyllabic data is that these words have less variability, as the standard deviation in log Reddit WF was 1.95 for the monosyllabic data while it was 2.23 for the disyllabic data. As will be seen in the following analyses, this trend of poorer fit to the monosyllabic data will hold across most variables tested. Subsequent analyses will focus on the hit-FA rate, as this is a summation of the other two data types.

The first regression analysis will determine which word frequency measures (either SUBTLEX derived or Reddit derived) provide the best fit to the hit-FA data for both the monosyllabic and disyllabic data. The results of the regression analysis are contained in Fig. 2, which shows that the Reddit WF values has a considerable advantage over the classic SUBTLEX frequency values. For the monosyllabic data, the Reddit WF values provide a 19.25% gain in variance accounted for over SUBTLEX WF values. For the disyllabic data there was a gain of 50.54% in variance accounted for the Reddit WF values, while the SUBTLEX values account for very little unique variance. There is considerable variability in word frequency distributions (see Johns & Jamieson, 2018, Johns et al., 2020b, for examples of this), so the negligible impact of the SUBTLEX frequency values is somewhat surprising. This finding suggests that the Reddit WF values provide a significantly better account of item-level recognition rates than previously used corpus types, while accounting for similar patterns in the data as the SUBTLEX norms.

An additional test is to determine if the Reddit WF values account for additional levels of variance above and beyond SUBTLEX WF by controlling for other standard information sources in a regression. Following Cortese et al. (Cortese et al., 2010; Cortese et al., 2015), the following

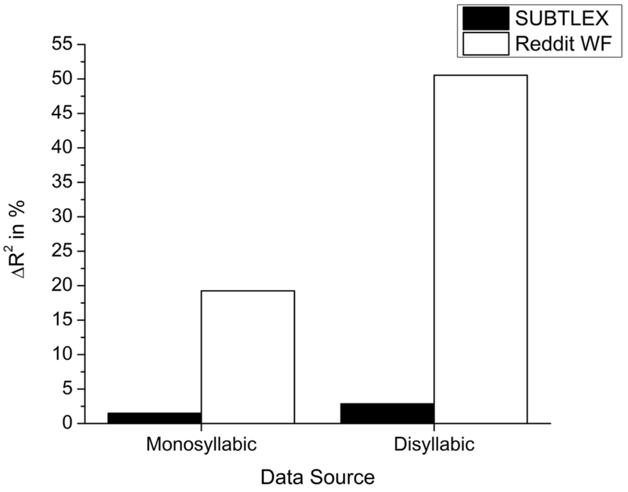


Fig. 2 Improvement in amount of variance accounted for the word frequency variables over each other. This figure demonstrates that the Reddit word frequency values offer a considerable advantage over the SUBTLEX word frequency values in accounting for word-level recognition rates

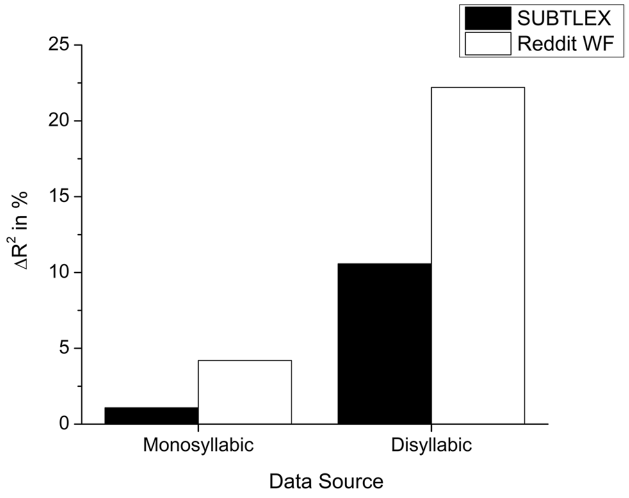


Fig. 3 The amount of unique variance that the Reddit word frequency and SUBTLEX word frequency values account for when five other standard psycholinguistic variables are included as predictors. This figure demonstrates that the Reddit word frequency values account for more unique variance compared with SUBTLEX word frequency values even when variance from other psycholinguistic information sources are accounted for

variables were controlled for: (1) word length, (2) imageability (Schock et al., 2012a), (3) age-of-acquisition (Schock et al., 2012b), (4) phonological distance (PLD), and (5) orthographic distance (OLD). PLD and OLD is the mean edit distance of the 20 closest words, using either an orthographic or phonological representation. A second regression analysis was conducted to determine how much extra variance either SUBTLEX or Reddit WF accounts for over and above these variables. The results of this analysis are contained in Fig. 3 and demonstrate that for both the monosyllabic and disyllabic data, the Reddit WF values explain more variance than the SUBTLEX variables. However, as in the previous analysis, the effect was bigger for the disyllabic data, where the Reddit WF allows for an 22.2% gain in variance accounted for, compared with a 10.57% gain for SUBTLEX WF. Overall, the regression model with the five variables and SUBTLEX WF accounted for 27.2% of variance in the monosyllabic data and 35.0% of the data for the disyllabic data. Comparatively, the five variables with the Reddit WF values accounted for 28.0% of the monosyllabic data and 40.1% of the disyllabic data.

This first result demonstrates that the WF values derived from Reddit provide a considerable advantage over the standard SUBTLEX WF values. A second question concerns whether the various contextual diversity measures allow for additional variance to be accounted for in word-level recognition rates, similar to results in lexical organization data (Johns, 2021), compared with what was seen in the above analysis of Reddit WF values. However, before comparing the different values, one pertinent issue is that recognition rates are not linear, but rather quadratic (Hemmer & Criss, 2013; Wixted, 1992; Zechmeister et al., 1978). In order to test whether the different variables the nontransformed versions of the variables were compared with a quadratic transformation ($x + x^2$, where x is a specified variable). The results of this comparison are contained in Table 3. This table shows that for the monosyllabic data, all variables saw an improvement when using a quadratic transformation, with the improvement offered for the DCD and UCD-derived variables showing a considerable advantage. For the disyllabic the results were more mixed, with a small decline in performance for three variables and a small increase for four variables. Given that overall using a quadratic increase performance on average, this transformation will be used in the remaining analyses.

Table 3 Impact of including a polynomial term on model fit

	Monosyllabic		Disyllabic	
	Linear	Quadratic	Linear	Quadratic
WF	−.364	−.365	−.456	−.453
CD	−.372	−.374	−.466	−.463
DCD	−.37	−.385	−.502	−.507
UCD	−.331	−.346	−.466	−.473
DCD-SD-PR	−.397	−.405	−.514	−.517
UCD-SD-PR	−.368	−.381	−.497	−.504
DCD-SD-WR	−.381	−.389	−.5	−.505
UCD-SD-WR	−.356	−.364	−.459	−.457

N = 2,578 for the monosyllabic data; N = 2,897 for the disyllabic data; all correlations significant at the $p < .001$ level.

To examine the different CD measures, a regression analysis similar to that contained in Fig. 3 was done, where the amount of variance that the CD measures account over the five variables outlined previously was calculated. The results of the analysis are displayed in Fig. 4 and contains the increase in variance found for the Reddit WF as a reference point. The top panel displays the results for the monosyllabic data set, while the bottom displays the results of the disyllabic data. The results again point to a bigger effect for disyllabic words than monosyllabic words. However, all of the CD-derived measures show a large advantage over the typical WF and CD values. Consistent with past results, the DCD and UCD counts provide a considerable improvement above WF and CD. Additionally, there is an improvement for the SD-transformed models, as the DCD-SD-PR and UCD-SD-PR provide an advantage over the count-based alternatives. However, the models utilizing a WR representation did not show an equivalent advantage, similar to the findings of Johns (2021) examining lexical organization data.

So far, the analyses conducted here have demonstrated that the newly proposed contextual diversity models account for much more unique variance in item-level recognition

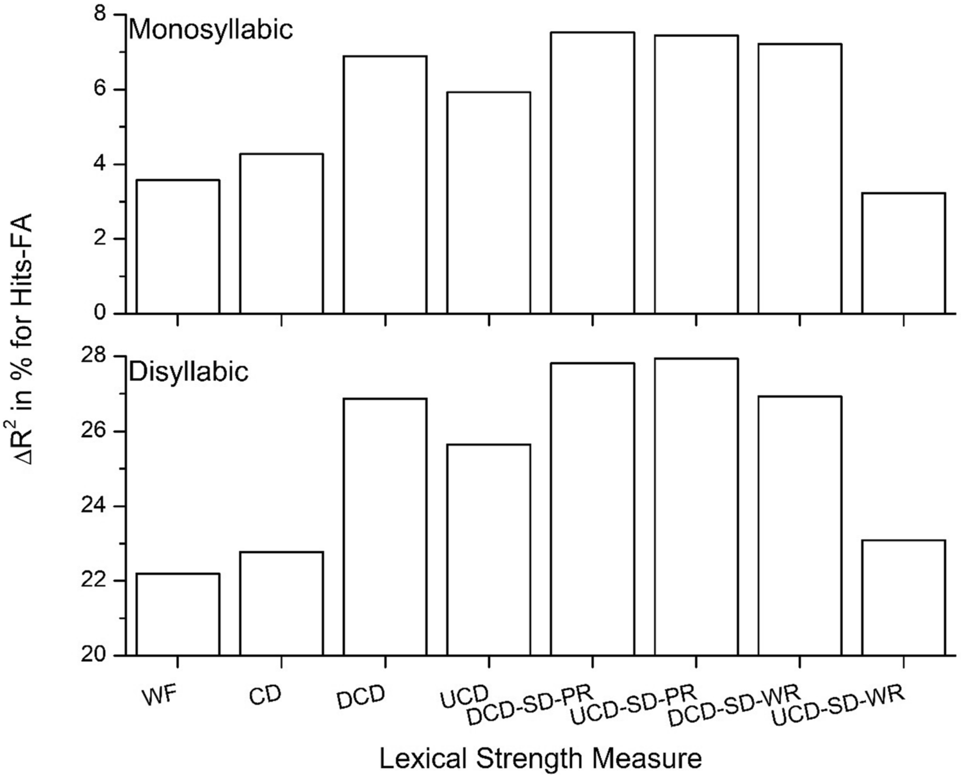


Fig. 4 Improvement of the contextual diversity derived variables over the previously outlined psycholinguistic variables. This figure shows that the contextual diversity measures derived in Johns (2021) offer a considerable advantage in terms of variance accounted for compared with word frequency

rates than word frequency. However, the variables have not been directly compared, an important consideration given that a single lexical strength measure is often used in stimuli selection. Given that the previous analyses demonstrated that the PR models account for the most variance, the focus will be on evaluating these models. Specifically, a regression analysis will be conducted contrasting the PR models to WF, to their count-based alternatives, and to each other, using a similar methodology to the above analyses. The result of this analysis is contained in Fig. 5 for both the monosyllabic (top panel) and disyllabic (bottom panel), with both datatypes showing identical patterns. This figure shows that both the DCD-SD-PR and UCD-SD-PR models account for considerably more unique variance than the WF and the DCD and UCD metrics. When compared with each other the DCD-SD-PR model accounts for more variance than the UCD-SD-PR model, suggesting that the DCD-SD-PR model is the best fitting contextual diversity measure.

The superiority of the DCD-SD-PR model suggests that there is a different pattern of fit for the episodic recognition data compared with the lexical organization data evaluated in Johns (2021), as the UCD-SD-PR model was found to provide the best fit to lexical organization data. The overall amount of variance that the regression model with the five psycholinguistic variables with the DCD-SD-PR strength variance is 29.9% for the monosyllabic data and 43.3% for the disyllabic data. This represents a gain of in variance accounted for of approximately 7% for the monosyllabic data, and 24% for the disyllabic data, compared with using the five previously outlined variables with the SUBTLEX WF values.

To further examine how the PR models operate, Fig. 6 contains the semantic distinctiveness values for the first 1,000 contextual occurrences for the randomly selected words *adore*, *bike*, *large*, and *truck*, while Fig. 7 contains the average semantic distinctiveness across the first 1,000 contextual occurrences for all words in the model's lexicon, both from the DCD-SD-PR model. These figures show rather disparate findings. Figure 7 shows that there is a constant decrease in the average update strength that a word receives for a contextual occurrence (similar to a power function). However, Fig. 6 shows that most of the strength for a word comes from highly distinctive contextual occurrences. For example, for the word *adore,* 53.45% of the word's strength comes from semantic distinctiveness values above 0.5, even though these occurrences account for only 7.28% of all contextual occurrences. The other three words had similar values, with the high SD contexts

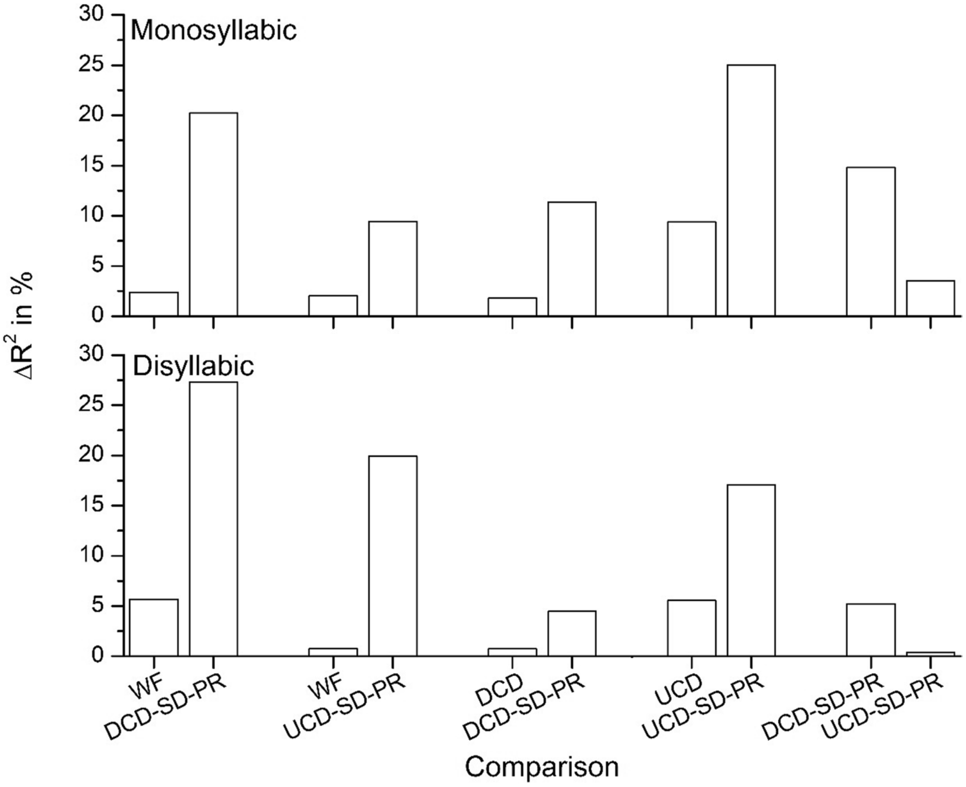


Fig. 5 Amount of unique variance that the different lexical strength measures account for when compared directly with each other, for both the monosyllabic and disyllabic data sets

contributing 50.44% of the lexical strength for the word *bike* (6.32% of all occurrences for that word), 49.99% for the word *large* (6.58% of all occurrences), and 51.65% for the word *truck* (5.29% of all occurrences). This suggests that for the newly proposed SDM models, highly distinct contextual occurrences contribute a disproportional amount to a word's strength in memory.

Thus, the success of these CD models may not necessarily be due to the continuous nature of the model's update function, but instead an ability to discriminate unique from redundant contexts. This suggests that a count of highly distinctive contexts could account for a similar amount of variance as the continuous alternative. From a methodological perspective this may not have a great deal of import. However, it does from a theoretical point of view: If a count of highly distinctive contexts can achieve a similar level of performance to the continuous model, it would suggest that it is unique episodic experiences with a word that leads to a word being strongly encoded in memory. Initial examinations using a count-based metric was recently done by Johns and Jones (2021), who found that a count-based UCD-SD-PR measure accounted for equivalent levels of unique variance in lexical organization data sets to the continuous version of the model.

In order to test whether a count-based version of the PR models could account for a similar level of variance as the continuous updating model, an additional simulation was done. To construct a count-based metric, a criterion was placed such that whenever an SD value exceeded the criterion, a word's strength in the memory was increased by 1. Thus, the model now has two parameters: the λ parameter and the update criterion. The two parameters are interdependent, as when the λ parameter goes up the number of contexts that would exceed a given criterion goes down, due to a higher down-weighting of high similarity contexts. To optimize the count-based metrics, a grid-search algorithm was used to determine the best combination of the two parameter types, with all λ values between 0 and 400 (in steps of 4) and all criterion values between 0 and 1 (in steps of 0.01) being evaluated. The fit to the monosyllabic data and disyllabic data were optimized separately.

The resulting correlations for the count-based models (as well the continuous implementations for comparisons sake) to both the monosyllabic and disyllabic data for the UCD-SD-PR and DCD-SD-PR models is contained in the top panel of Fig. 8. This figure shows that for both model types the count-based implementations have a higher correlation than their continuous counterparts across both datatypes.

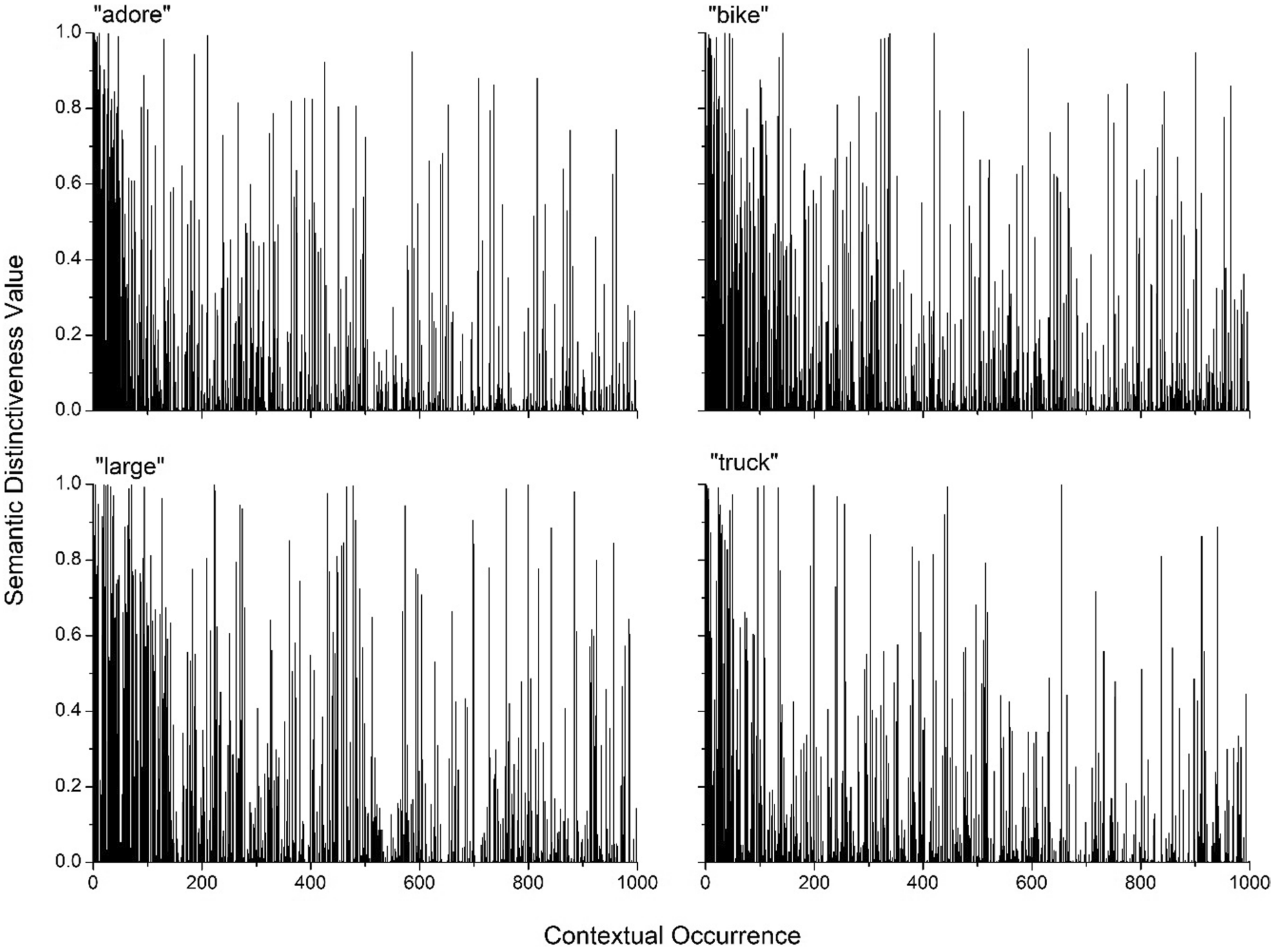


Fig. 6 SD values for the first 1,000 contexts for the DCD-SD-PR model for four randomly selected words. This figure shows that much of the strength of words comes from highly distinct contexts

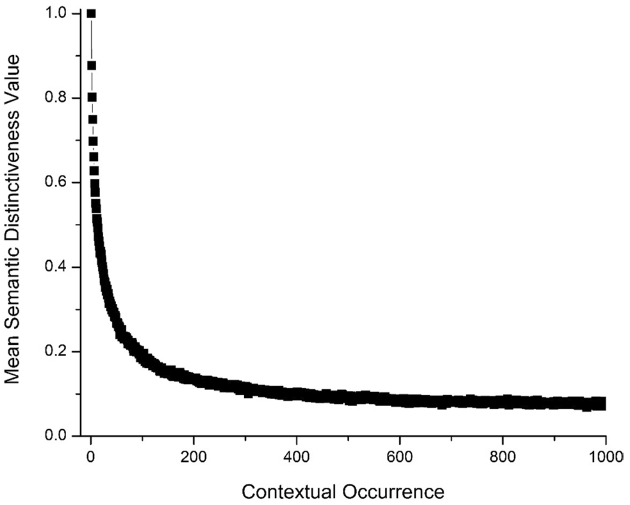


Fig. 7 Average update strength for the DCD-SD-PR model across the first 1,000 contexts for all words in the lexicon. Contrasted with Fig. 6 the average update strength is not indicative of how lexical strength is updated

To shed light on the optimization of the models, Table 4 contains the parameter values for both model types and data sets, as well as the percentage of contexts that was used to update the strength of words for the combined word set of Cortese et al. (2010, 2015). This table shows that there was not a great deal of consistency in the optimal parameter sets, likely due to different models having different distributional properties. However, it does show that in the count-based implementations, most contexts are ignored, with a range of .86% (for the UCD-SD-PR model fit to the disyllabic data) to 6.07% (for the DCD-SD-PR model fit to the disyllabic data) of the total number of contexts being used to update a word's strength in memory. This suggests that the success of the count-based models rely upon the identification of unique contextual usages of a word.

The bottom panel of Fig. 8 contains the amount of unique variance that the continuous and count-based implementations of the UCD-based and DCD-based models account for across the two datatypes. This figure shows that for each

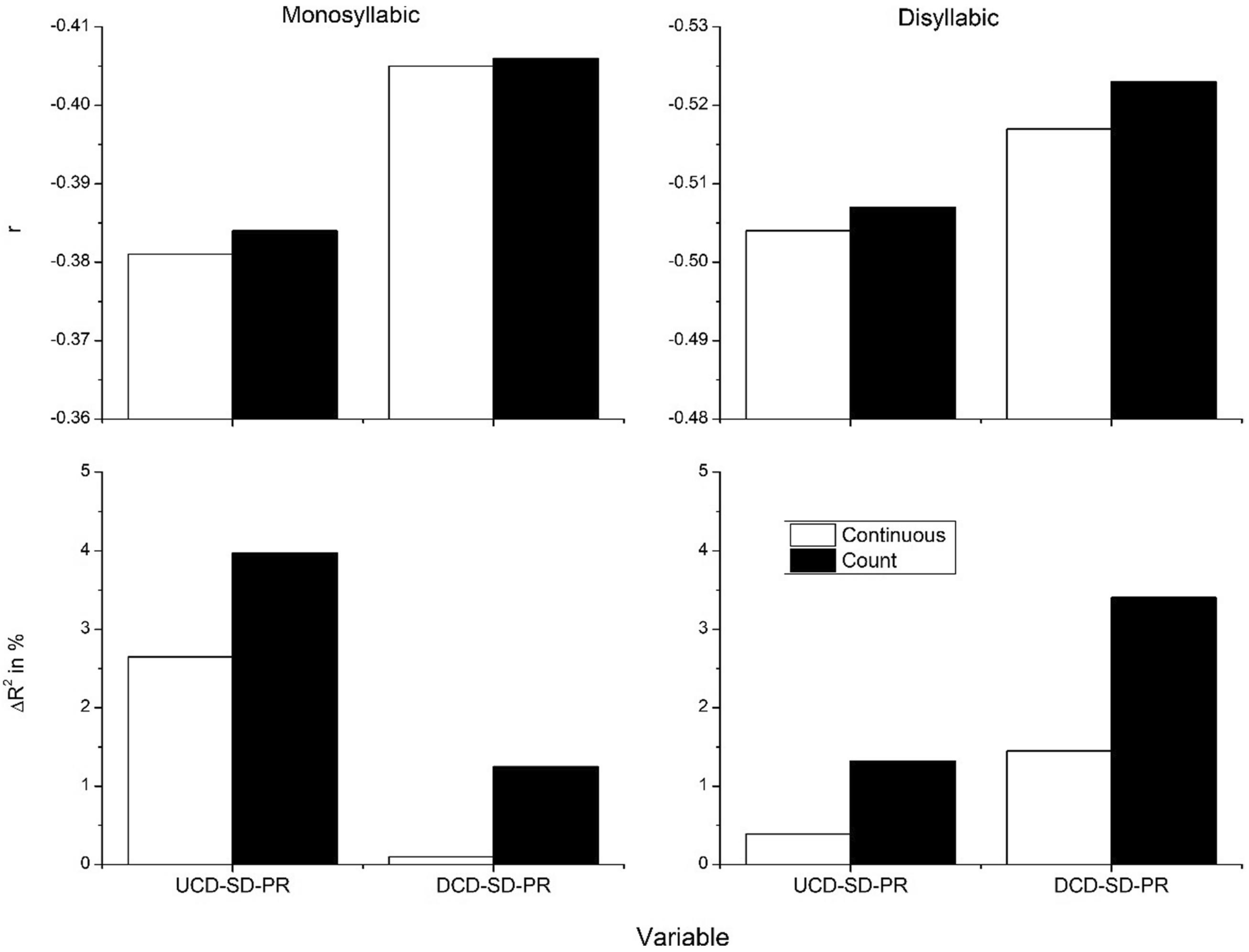


Fig. 8 Correlations (top panels) and amount of unique variance accounted for (bottom panels) contrasting the continuous and count-based versions of the UCD-SD-PR and DCD-SD-PR models to the Hits-FA data for the monosyllabic (left panel) and disyllabic (right panel)

Table 4 Parameters and percentage of contexts used for the DCD-SD-PR and UCD-SD-PR count-based implementations

Model	Data	λ	Crit	% Contexts Used
DCD-SD-PR	Monosyllabic	372	.55	2.75
	Disyllabic	116	.57	6.07
UCD-SD-PR	Monosyllabic	240	.49	.93
	Disyllabic	360	.4	.86

examination the count-based metric accounted for more unique variance than the continuous measure. The final regression, contained in Fig. 9, contrasts the count-based implementations of the UCD-SD-PR and DCD-SD-PR models, in order to determine which alternative provides the best accounting of the data. The result demonstrates that the count-based DCD-SD-PR model accounted for the most unique variance. Overall, this suggests that the best contextual diversity measure of item-level recognition performance is given by a count of highly distinct discourse contexts. Given that the DCD-SD-Count model provides the best accounting to this data, Fig. 10 contains a scatterplot of the fit of this model to both the monosyllabic (top panel) and disyllabic (bottom panel). This finding will be discussed in more detail in the General Discussion.

One analysis that has not been reported in the amount of unique variance that the five other psycholinguistics measures detailed above account for when the best fitting lexical strength variable (the DCD-SD-Count model) is controlled for. This has theoretical consequences for models of recognition, as Cortese et al. (2010, 2015) found that imageability was the most important predictor of item-level recognition performance, a result interpreted in favor of the dual-coding hypothesis of Paivio (1991). As a first glance at understanding the relationship among these variables, Table 5 contains the correlation matrix for the words from

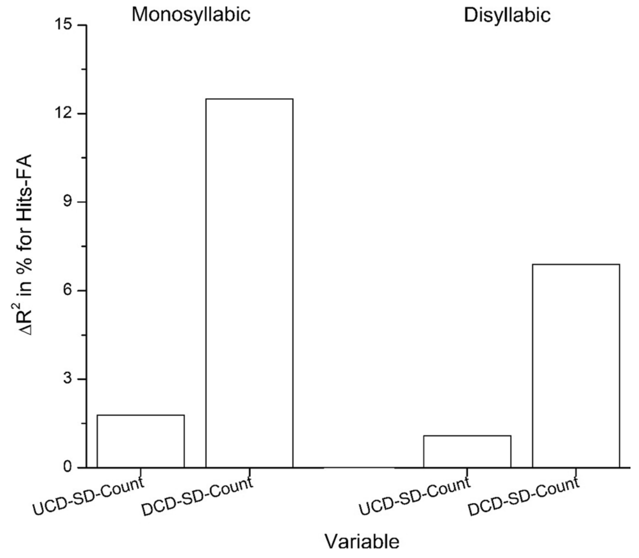


Fig. 9 The amount of unique variance that the UCD-SD-Count and DCD-SD-Count models account for in the monosyllabic and disyllabic data

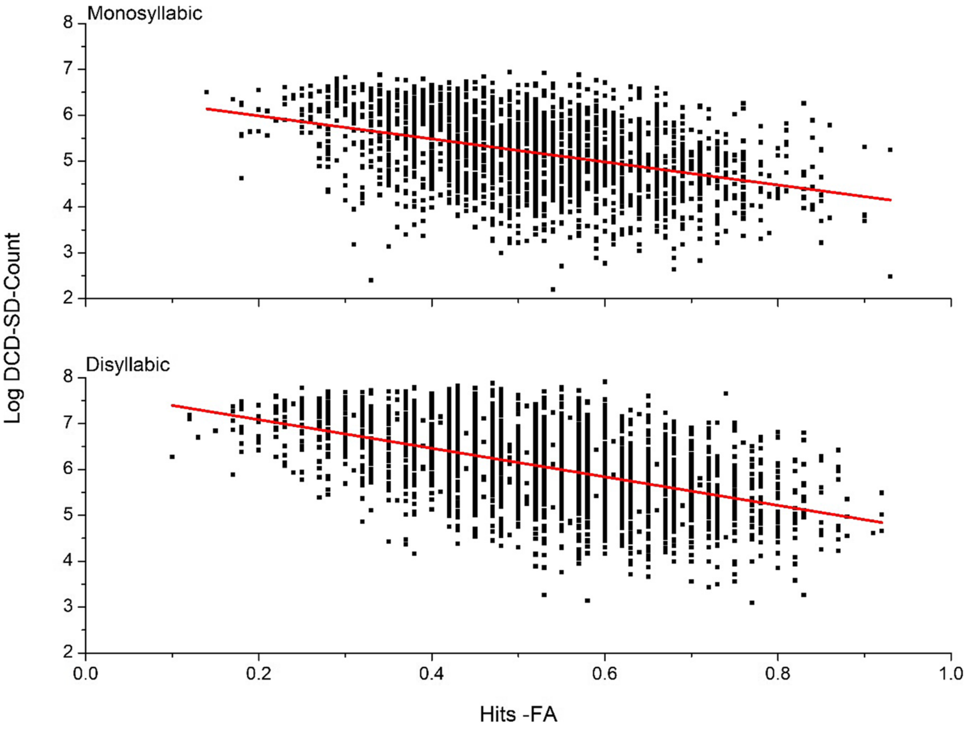


Fig. 10 Linear fit of the DCD-SD-Count model to the monosyllabic and disyllabic data

the monosyllabic and disyllabic data for the six psycholinguistic variables (AoA, Imageability, OLD, PLD, length, and DCD-SD-Count). Figure 11 contains the amount of unique variance that each variable accounts for when compared against each other for both the monosyllabic (top panel) and disyllabic (bottom panel) data. This figure shows that for monosyllabic data, imageability accounted for the greatest amount of unique variance followed by word length and the DCD-SD-Count. However, there is a different pattern for the disyllabic data—the DCD-SD-Count accounted for the most unique variance followed by word length and imageability. Given that both the DCD-SD-Count and imageability still account for significant amounts of unique variance when other variables are controlled for, this suggests that multiple types of lexical information are used when remembering individual words.

Finally, in order to demonstrate that the advantage for the lexical strength variables from Johns (2021) generalizes to a different source of recognition memory data, the correlation between the different lexical strengths variables and false recognition rates in the Deese/Roediger-McDermott (DRM; Roediger & McDermott, 1995) paradigm was calculated. Specifically, the correlation of the different lexical strength variables to the combined 52 false memory lists from Stadler et al. (1999) and Gallo and Roediger (2002) was taken. For a similar analysis examining accounting for false recognition rates, see Johns et al. (2012b) and Johns et al. (2019). The results are displayed in Fig. 12 and demonstrate that the advantage for the Reddit data generalizes to this data, as does the advantage for the contextual and semantic diversity values. The overall best fitting model was the count-based UCD-SD-PR model, although its advantage over the other

Table 5 Correlation tables for the words for the monosyllabic and disyllabic data for the differing psycholinguistic variables and the best fitting contextual diversity measure

	Measure	2	3	4	5	6	7
Monosyllabic	1. Hits-FA	.253	.201	.106	.055	−.093	−.398
	2. AoA	.	−.47	.239	.186	.213	−.696
	3. Imageability	.	.	−.085	−.023 ɸ	−.046	.066
	4. OLD	.	.	.	.654	.699	−.233
	5. PLD	.	.	.	.	.53	−.215
	6. Length	.	.	.	.	.	−.163
	7. DCD-SD-Count	.	.	.	.	.	.
Disyllabic	1. Hits-FA	.109	.336	.146	.041	−.122	−.518
	2. AoA	.	−.505	.186	.206	.123	−.525
	3. Imageability	.	.	−.023ɸ	−.109	−.037	−.058
	4. OLD	.	.	.	.657	.679	−.162
	5. PLD	.	.	.	.	.519	−.106
	6. Length	.	.	.	.	.	−.062
	7. DCD-SD-Count	.	.	.	.	.	.

N = 2,578 for the monosyllabic data; *N* = 2,897 for the disyllabic data; ɸ = *p* > .05.

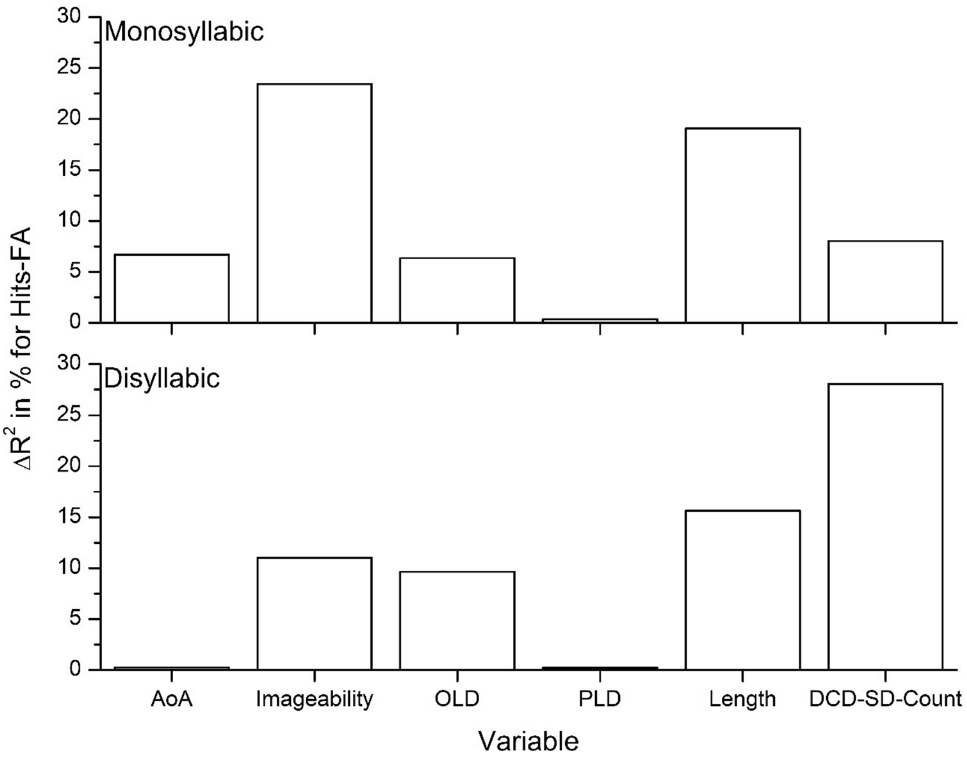


Fig. 11 Amount of unique variance that each psycholinguistic variable account for when compared against each other

CD variables is quite small (likely due to the small sample size). This result, with the caveat of a very small sample size for this type of analysis, reinforces the advantage of the lexical strength variables derived from the Reddit data, as well as the advantage for the contextual diversity-based measures, generalizes to a different type of recognition memory data.

General discussion

There were two main goals of this article. The first was to determine if word frequency measures derived from the internet forum Reddit provides a better accounting

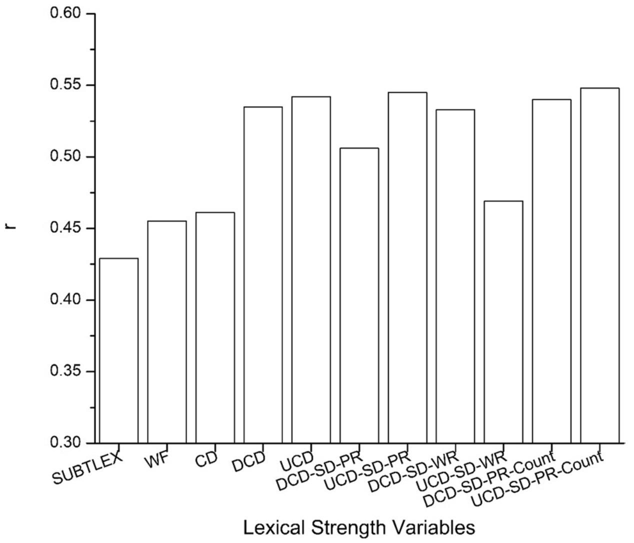


Fig. 12 Correlations between the combined false memory lists of Stadler et al. (1997) and Gallo and Roediger (2002) and the different lexical strength variables

word-level recognition memory performance compared with the classic SUBTLEX word frequency measures (Brysbaert & New, 2009), which are typically used to control behavioral experiments. The second was to determine if the contextual diversity-based measures described in Johns (2021), based on much previous research on the organization of the mental lexicon (e.g., Adelman et al., 2006; Brysbaert & New, 2009; Johns et al., 2020a), also extends to word-level recognition rates. Both of these hypotheses were found to be correct.

Word frequency has a rich history of usage in studies of recognition memory, most clearly in the mirror effect of frequency (e.g., Glanzer & Adams, 1985). An important topic in the study of the impact of word frequency is the corpus from which one derives such measures (e.g., Brysbaert & New, 2009; Johns, 2021; Johns et al., 2020a; Kučera & Francis, 1967). In the word-level data collection effort of Cortese et al. (2010; Cortese et al., 2015) it was found that the word frequency values derived from SUBTLEX (Brysbaert & New, 2009) accounted for significant levels of variance in the data sets. In this article, word frequency values derived from Reddit were found to account for significantly more variance than the SUBTLEX WF values.

There are a number of possible reasons for the advantage of the Reddit WF values. One is that Reddit is a more communicative source of linguistic information than compared with television and movie subtitles, which the SUBTLEX corpus is composed of. The other possibility is that the Reddit corpus is much larger than the SUBTLEX corpus, resulting in more refined frequency values. However, previous studies have found that smaller but more targeted corpora offer a better fit to lexical data than larger but nontargeted corpora (Johns & Jamieson, 2019; Johns, Jones, & Mewhort, 2019), which suggests that it is the ecological validity of language that matters. However, this is an important topic for future research.

The second hypothesis tested determined whether contextual diversity-based measures provide an advantage over word frequency, consistent with past studies in lexical organization (e.g., Adelman et al., 2006; Brysbaert & New, 2009; Johns, 2021; Johns et al., 2012a, 2016, 2020a; Jones et al 2012; see Jones et al., 2017 for a review). It was found that this hypothesis was also correct, as the CD values had a higher correlation and accounted for more variance than the WF values, although the advantage was much more pronounced for disyllabic words than for monosyllabic words. The best models were found to be the CD values transformed by the semantic distinctiveness model (SDM; Johns et al., 2020a; Jones et al., 2012), specifically the models using a population representation (PR) from Johns (2021), namely the UCD-SD-PR and DCD-SD-PR model, with the DCD-SD-PR accounting for the most unique variance between the two.

However, the best accounting of the data was given by the count-based alternatives to the UCD-SD-PR and DCD-SD-PR models, where instead of a continuous strength updating mechanism, only contexts that exceeded a set criterion caused an increase in a word's strength in memory. When fitting the criterion for the count-based models it was found that the best fitting models only use a very small percentage of contexts (i.e., less than 10%), suggesting that it is only very semantically distinct contexts that update a word's strength in memory. The count-based models offer a different theoretical interpretation than the continuously updating model, as it suggests that the lexical storage of a word is based upon its occurrence within distinct context types, rather than a continuous updating across every contextual occurrence of a word.

There are a number of implications of the work described here. One is methodological—the lexical variables from the Reddit data first described in Johns (2021) provide a substantially better fit than the classic word frequency values of Brysbaert and New (2009) and Kučera and Francis (1967). This provides researchers with more power to control their experiments, as well as to avoid any unwanted confounds in their designs.

The other implication is theoretical—that the locus of frequency effects in the study of recognition memory may not actually lie in word frequency, but instead in contextual diversity. A similar proposal has been made by the Source of Activation Confusion (SAC) model of memory (Popov & Reder, 2020; Reder et al., 2000), which proposes that frequency effects in recognition memory are mainly driven by differing patterns of contextual occurrence for high-frequency and low-frequency words. Contextual diversity as first explored by Adelman et al. (2006) offered a relatively small deviation from WF. However, refinement of this initial finding has demonstrated that linguistic context is much larger, and in some ways more abstract, than had been previously considered. In particular, linguistic context may not be confined in the moment-to-moment variability in language usage, but instead may map onto higher level properties of human communication and human experience, such as the discourse one is communicating within or the people who one is communicating with. This perspective is consistent with other perspectives, such as usage-based and adaptive theories of language processing (e.g., Beckner et al., 2009; Christiansen & Chater, 2008; Tomasello, 2003, 2009). The fact that the same lexical strength models described in Johns (2021) generalize to episodic recognition data provides a promising avenue for future research addressing the commonalities between memory and language processing at an item-level of analysis.

An additional theoretical implication is the usage of global distinctiveness to account for differing levels of memorability across items. Local distinctiveness (that is, distinctiveness within a single episodic context) is a well-known and classic effect in list memory performance (von Restorff, 1933), and details how unique stimulus properties increase the strength of an item within context. Global distinctiveness, as measured by the SDM, refers to the uniqueness of the semantic content of the contexts that a word occurs in across learning. In this model, words that occur in more unique semantic contexts have a great strength in memory. Indeed, the results of this article demonstrate that a simple count of the number of very distinct semantic contexts provides the best fit to item-level recognition rates. This result is coherent with the rational analysis of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991), as highly distinct contexts signal the overall types of contexts that a word could occur in. Words that occur in the greatest number of context types should be the most available in memory as they are more likely to be used in any future context.

The overall best fitting model was the count version of the DCD-SD-PR model, which differs from results examining lexical organization (Johns, 2021), where the UCD-SD-PR has been found to offer the best fit to the data. This suggests that for recognition memory, the main contextual information source is the number of different discourse topics that a word had occurred in, rather than the number of people who had used a word previously (which seems to provide the best fit to lexical organization data). This suggests that multiple types of contextual information may be contained in a word's lexical representation, which can be differentially accessed depending on task requirements. How to construct a model that can accomplish this is an important question for future research.

For this type of research on episodic memory to continue to evolve, it is going to be necessary to expand the types of data that is available for researchers to do item-level analyses on. The collection of large sources of item-level data in psycholinguistics is a very important trend in the development of computational models of language (Johns et al., 2020b). Indeed, there are wide and varied collections of psycholinguistic norms available, including such data as modality norms (Lynott & Connell, 2009), taboo words (Roest et al., 2018), humor norms (Engelthaler & Hills, 2018), idiomatic processing (Bulkes & Tanner, 2017), and body–object interaction ratings (Bennett et al., 2011), among many others. The data collection efforts of Cortese et al. (2010; Cortese et al., 2015) provides the basis of the current article, and hence signal the promise of these types of approaches to better understanding theoretical issues in episodic memory.

Of course, there are aspects of recognition memory experiments that make this type of data collection difficult, although it is an active research area within episodic recognition (e.g., Cox et al., 2018). In particular, most previous data collected has focused on word-in-isolation designs, where data for individual words was collected, and where surrounding words were randomized. In contrast, episodic memory is inundated with list composition effects where the impact of surrounding items is considerable, such as the range of word frequency of words in a list (Malmberg & Murnane, 2002), list length effects (Dennis & Chapman, 2010; Murnane & Shiffrin, 1991), and strength and list strength effects (Ratcliff et al., 1990). The effects of semantics are most pronounced in studies of false memory, where there are large effects on number of associations studied (Robinson & Roediger, 1997), type of association (Dewhurst & Anderson, 1999; Hutchison & Balota, 2005), and list study effects (Dewhurst et al., 2009), among others. Thus, in countenance to big data studies in lexical organization, which have used standard tasks such as lexical decision and naming (e.g., Balota et al., 2007), it is difficult to propose an optimal setup to examine word-level variability in recognition performance. However, the work of Cortese et al. (2010, 2015) demonstrates how this type of study is possible, which hopefully informs future data collection efforts.

Current trends in computational cognitive modeling of episodic memory which combine processing mechanisms of episodic memory models with representations derived from distributional models (e.g., Johns et al., 2012b, 2021; Mewhort et al., 2018; Osth et al., 2020) provide a promising pathway to further examine both item-level and group-level effects in recognition memory performance with computational models. These models, in principle, have a lexicon of information available to them to process any word, and should be able to predict item-level effects in recognition memory (see Johns et al., 2012b, 2019, 2021, for examples of this). The work described here sets out constraints for representational considerations for these models, such as context size and representational information, which these models will need to build towards to be better able account for item-levels effects in recognition memory.

In the computational cognitive sciences, episodic memory models (e.g., Dennis & Humphreys, 2001; Howard & Kahana, 2002; Shiffrin & Steyvers, 1997) have proven to be particularly powerful models of human behavior. However, the focus of these classical models has concentrated on explaining memory performance in highly controlled behavioral tasks. In recent years, there has been considerable progress made in integrating other information sources into the processing of different models, including contextual information, into the operations of these models (e.g., Cox & Shiffrin, 2017; Osth & Dennis, 2015; Popov & Reder, 2020) to much success. For future work, a major challenge that this work faces are how to integrate local processing (e.g., how do we make decisions about which word occurred in a list; the problems that traditional episodic recognition models attempt to answer) to global effects in memory (e.g., why some words are more easily recognized even with equal frequency of occurrence, which is the answer that SDM is trying to answer).

Additionally, the finding that the count-based version of the SDM provides the best fit to the suggests that an important aspect of the storage of word information is in the identification of unique contextual occurrences of a word. This entails a de facto recognition process, where the model has to identify unique context types for a given word. However, the SDM is not a process model, but rather a learning and representational model, and is agnostic about the mechanisms that can accomplish this. Given the range of different mechanisms that have been proposed to accomplish episodic recognition (e.g., Dennis & Humphreys, 2001; Hintzman, 1988; Howard & Kahana, 2002; Murdock, 1982; Shiffrin & Steyvers, 1997), it possible that the integration of more sophisticated processing mechanisms with the learning and storage mechanisms of the SDM could provide a superior accounting for item-level data, an important topic for future research. This requires continued integration of corpus-based representational models with processing models developed in the computational cognitive sciences in order to continue to develop more realistic theories of cognition (Johns et al., 2020b).

Overall, this work demonstrates the promise of adopting current big data approaches in the study of cognition to the study of episodic memory, building off of the research of Cortese et al. (2010, 2015). Although the study of recognition memory is considerably more complex than single word studies, the fact that the lexical strength measures of Johns (2021) generalize to item-level variability in episodic recognition signals the promise of these approaches to better understanding episodic recognition at the single word level.

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

References

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

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. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*(4), 703–719. https://doi.org/10.1037/0033-295X.96.4.703

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x

Aue, W. R., Fontaine, J. M., & Criss, A. H. (2018). Examining the role of context variability in memory for items and associations. *Memory & Cognition, 46*(6), 940–954. https://doi.org/10.3758/s13421-018-0813-9

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/BF03193014

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Data set. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 830–839. https://ojs.aaai.org/index.php/ICWSM/article/view/7347. Accessed Feb 2020

Bennett, S. D., Burnett, A. N., Siakaluk, P. D., & Pexman, P. M. (2011). Imageability and body–object interaction ratings for 599 multisyllabic nouns. *Behavior Research Methods, 43*, 1100–1109. https://doi.org/10.3758/s13428-011-0117-5

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Ke, J., Christiansen, M. H., Larsen-Freeman, D., Croft, W., Schoenemann, T., & Five Graces Group. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*(Suppl. 1), 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*(3), 539–576. https://doi.org/10.1037/0033-295X.114.3.539

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

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. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance, 42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods, 49*(2), 772–783. https://doi.org/10.3758/s13428-016-0747-8

Cevoli, B., Watkins, C., & Rastle, K. (2021). What is semantic diversity and why does it facilitate visual word recognition? *Behavior research methods, 53*(1), 247–263.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*(5), 489–509. https://doi.org/10.1017/S0140525X08004998

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*(6), 595–609. https://doi.org/10.1080/09658211.2010.493892

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology*, 68, 1489-1501.

Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *Quarterly Journal of Experimental Psychology*, 66, 946-972.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General, 147*(4), 545–590. https://doi.org/10.1037/xge0000407

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review, 124*(6), 795–860. https://doi.org/10.1037/rev0000076

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*(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language, 63*(3), 416–424. https://doi.org/10.1016/j.jml.2010.06.001

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478. https://doi.org/10.1037/0033-295X.108.2.452

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & Cognition, 27,* 665–673. https://doi.org/10.3758/BF03211560

Dewhurst, S. A., Bould, E., Knott, L. M., & Thorley, C. (2009). The roles of encoding and retrieval processes in associative and categorical memory illusions. *Journal of Memory and Language, 60*(1), 154–164. https://doi.org/10.1016/j.jml.2008.09.002

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior research methods, 50*(3), 1116–1124.

Gallo, D. A., & Roediger, H. L., III. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language, 47*(3), 469–497. https://doi.org/10.1016/S0749-596X(02)00013-X

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8–20. https://doi.org/10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5–16. https://doi.org/10.1037/0278-7393.16.1.5

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*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1947–1952. https://doi.org/10.1037/a0033744

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41*(4), 976–995. https://doi.org/10.1111/cogs.12392

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

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. https://doi.org/10.3758/s13428-012-0278-x

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 https://doi.org/10.1016/j.jml.2020.104146

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

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. https://doi.org/10.1016/j.jml.2018.08.005

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review, 2*(1), 105–112. https://doi.org/10.3758/BF03214414

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*(4), 421–445. https://doi.org/10.1006/jmla.1993.1023

Hutchison, K. A., & Balota, D. A. (2005). Decoupling semantic and associative information in false memories: Explorations with semantically ambiguous and unambiguous critical lures. *Journal of Memory and Language, 52*(1), 1–28. https://doi.org/10.1016/j.jml.2004.08.003

Johns, B. T. (2019). Mining a crowdsourced dictionary to understand consistency and preference in word meanings. *Frontiers in Psychology, 10,* Article 268. https://doi.org/10.3389/fpsyg.2019.00268

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review, 128*(3), 525–557. https://doi.org/10.1037/rev0000265

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science, 42*(4), 1360–1374. https://doi.org/10.1111/cogs.12583

Johns, B. T. & Jamieson, R. K. (2019). The influence of time and place on lexical behavior: A distributional analysis. *Behavior Research Methods, 51*, 2438–2453.

Johns, B. T. & Jones, M. N. (2021). Content matters: Measures of contextual diversity must consider semantic content. PsyArXiv.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012a). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *Journal of the Acoustical Society of America, 132*(2), EL74–EL80. https://doi.org/10.1121/1.4731641

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012b). A synchronization account of false recognition. *Cognitive Psychology, 65*(4), 486–518. https://doi.org/10.1016/j.cogpsych.2012.07.002

Johns, B. T., Dye, M., Jones, M. N. (2014). The influence of contextual variability on word learning. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Johns, B. T., Dye, M. W., & Jones, M. N. (2016a). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review, 23,* 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Johns, B. T., Sheppard, C., Jones, M. N., & Taler, V. (2016b). The role of semantic diversity in lexical organization across aging and bilingualism. *Frontiers in Psychology, 7,* 703. https://doi.org/10.3389/fpsyg.2016.00703

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review, 26*(1), 103–126. https://doi.org/10.3758/s13423-018-1501-2

Johns, B. T., Dye, M., & Jones, M. N. (2020a). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology, 73*(6), 841–855. https://doi.org/10.1177/1747021819897560

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2020b). The continued importance of theory: Lessons from big data approaches to cognition. In S. E. Woo, R. Proctor, & L. Tay (Eds.), *Big data in psychological research*. APA Books.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2021). A continuous source reinstatement model of true and false recollection. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 75*(1), 1–18. https://doi.org/10.1037/cep0000237

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. https://doi.org/10.1037/0033-295X.114.1.1

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(2), 115–124. https://doi.org/10.1037/a0026727

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizational principle of the lexicon. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 64, pp. 239–283). Elsevier. https://doi.org/10.1016/bs.plm.2017.03.008

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science, 41*(3), 590–622. https://doi.org/10.1111/cogs.12353

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*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68*(6), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review, 28,* 40–80. https://doi.org/10.3758/s13423-020-01792-x

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. https://doi.org/10.1037/0033-295X.104.2.211

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language, 64*(3), 249–255. https://doi.org/10.1016/j.jml.2010.11.003

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods, 41*(2), 558–564. https://doi.org/10.3758/BRM.41.2.558

MacLeod, C. M. (2020). Zeigarnik and von Restorff: The memory effects and the stories behind them. *Memory & Cognition, 48*(6), 1073–1088. https://doi.org/10.3758/s13421-020-01033-5

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,* 104203.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 616–630. https://doi.org/10.1037/0278-7393.28.4.616

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

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.

Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review, 25*(3), 932–950. https://doi.org/10.3758/s13423-017-1327-3

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*(6), 609.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 855–874. https://doi.org/10.1037/0278-7393.17.5.855

Neath, I., & Crowder, R. G. (1990). Schedules of presentation and temporal distinctiveness in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 316–327.

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

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311. https://doi.org/10.1037/a0038692

Osth, A. F., Shabahang, K. D., Mewhort, D. J., & Heathcote, A. (2020). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. *Journal of Memory and Language, 111,* 104071.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45*(3), 255–287. https://doi.org/10.1037/h0084295

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review, 127*(1), 1–46.

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.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163–178. https://doi.org/10.1037/0278-7393.16.2.163

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 294–320. https://doi.org/10.1037/0278-7393.26.2.294

Robinson, K., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 231–237. https://doi.org/10.1111/j.1467-9280.1997.tb00417.x

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2018). Dutch taboo norms. *Behavior Research Methods, 50*(2), 630–641. https://doi.org/10.3758/s13428-017-0890-x

Schock, J., Cortese, M. J., & Khanna, M. M. (2012a). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods*, 44, 374-379.

Schock, J., Cortese, M. J., Khanna, M. M., & Toppi, S. (2012b). Age of acquisition estimates for 3,000 disyllabic words. *Behavior Research Methods*, 44, 971-977.

Shaoul, C. & Westbury, C. (2010). *The Westbury Lab Wikipedia Corpus 2010*. University of Alberta.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. https://doi.org/10.3758/BF03209391

Stadler, M. A., Roediger, H. L., III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*(3), 494–500. https://doi.org/10.3758/BF03211543

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. https://doi.org/10.1037/0278-7393.29.5.760

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University Press.

von Restorff, H. (1933). Ueber die Wirkung von Bereichsbildungen im Spurenfeld. Analyse von Vorgängen im Spurenfeld. I. Von W. Köhler und H. v. Restorff [On the effect of field formations in the trace field. Analysis of processes in the trace field. I. By W. Kohler and H. v. Restorff]. *Psychologische Forschung, 18,* 299–342. https://doi.org/10.1007/BF02409636

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 681–690. https://doi.org/10.1037/0278-7393.18.4.681

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society, 11*(6), 371–373. https://doi.org/10.3758/BF03336857

Open practices statement

All model values are available at: btjohns.com/Johns_MC_CDvals.xlsx or https://osf.io/5nr6x/