Determining the Importance of Frequency and Contextual Diversity in the Lexical Organization of Multiword Expressions

Marco S. G. Senaldi, Debra Titone, and Brendan T. Johns
McGill University

In press, *Canadian Journal of Experimental Psychology*

Corresponding Address
Marco S. G. Senaldi
Department of Psychology, McGill University
2001 McGill College Avenue
Montreal, Quebec, Canada
H3A 1G1
Email: marco.senaldi@mcgill.ca

Acknowledgements: This research was supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2020-04727 to BTJ and NSERC Discovery Grant 261769 to DT.

Abstract

Corpus-based models of lexical strength have called into question the role of word frequency as an organizing principle of the lexicon, revealing that contextual and semantic diversity measures provide a closer fit to lexical behavior data (Adelman, Brown & Quesada, 2006; Jones, Johns, & Recchia, 2012). Contextual diversity measures modify word frequency by ignoring word repetition in context, while semantic diversity measures consider the semantic consistency of contextual word occurrence. Recent research has shown that a better account of lexical organization data is provided by socially-based measures of semantic diversity, which encode the communication patterns of individuals across discourses (Johns, 2021). While most research on contextual diversity has focused on single words, recent corpus-based and experimental evidence suggests that an integral part of language use involves recurrent and more structurally complex units, such as multiword phrases and idioms. The aim of the present work was to determine if contextual and semantic diversity drive lexical organization at the level of multiword units (here, operationalized as idiomatic expressions), in addition to single words. To this end, we analyzed normative ratings of familiarity for 210 English idioms (Libben & Titone, 2008) using a set of contextual, semantic and socially-based diversity measures that were computed from a 55-billion word corpus of Reddit comments. Results confirm the superiority of diversity measures over frequency for multiword expressions, suggesting that multiword units, such as idiomatic phrases, show similar lexical organization dynamics as single words.

Keywords: Lexical organization; semantic diversity; idioms; multiword expressions; distributional semantics

Public Significance Statement

Corpus-based evidence indicates that the ease with which we access single words in the lexicon depends on their contextual and social diversity, rather than their frequency.

However, an integral part of our language environment consists also of conventional multiword units like idioms.

We demonstrate that contextual and social diversity shape the lexicon also at the level of multiword units and that idioms thus exhibit the same lexical organizational dynamics as single words.

Determining the Importance of Frequency and Contextual Diversity in the Lexical Organization of Multiword Expressions

Research into the organization of the mental lexicon revolves around two core issues. The first is to determine the environmental information sources that drive the organization of language and shape lexical behavior. The second issue pertains to the granularity at which such principles operate, namely whether they only involve single words or extend to more structurally complex units such as multiword phrases and idioms.

Classical accounts of lexical organization assign a central role to word frequency in determining the relationship between the language environment and lexical behavior (Broadbent, 1967; Brysbaert, Mandera, & Keuleers, 2018; Forster & Chambers, 1973). However, corpus-based models of lexical strength have recently called into question the importance of word frequency, demonstrating that contextual diversity provides a better quantitative fit to lexical behavior data (Adelman, Brown & Quesada, 2006; Brysbaert & New, 2009; see McDonald & Shillcock, 2001 for earlier work on this issue and Jones, Dye, & Johns, 2017 for a review). In these studies, contextual diversity is operationalized as the number of documents in a corpus that a given word occurs in, ignoring word repetition within context. Theoretical support for the role of contextual diversity in lexical organization comes from the principle of likely need, rooted in the rational analysis of memory (Anderson & Milson, 1989; Anderson & Schooler, 1991). Under this view, human memory and the lexicon are proposed to be complex adaptive systems, whose task is to maximize the accessibility of needed information. Words that have occurred in many different contexts in the past are likely to appear in future contexts, and hence should be the most accessible.

The Semantic Distinctiveness Model (Johns, Dye, & Jones, 2016, 2020; Jones et al., 2012) modifies contextual diversity by accounting for the semantic diversity of the contexts in which words occur. In this model, each contextual occurrence of a word increases its lexical strength in a graded fashion based on an expectancy-congruency mechanism: the more the current context is semantically dissimilar to past contextual uses of a word, and thus unexpected, the greater the increase in a word's strength that the context provides. The Semantic Distinctiveness Model has repeatedly been shown to provide a superior fit than a contextual diversity count across many data types, including visual word processing (Jones et al., 2012; Johns et al., 2020), spoken word recognition (Johns et al., 2012), natural language learning

(Johns, Dye, & Jones, 2016) and bilingualism and aging (Johns, Sheppard, Jones, & Taler, 2016; Qiu & Johns, 2020).

While early studies operationalized contextual diversity in terms of small contextual units, such as documents, paragraphs, or moving window in a corpus, recent work by Johns et al. (2020) demonstrated that measuring contextual and semantic diversity in quantitatively larger contexts (i.e., number of books a word occurs in) provides a better account of whether a speaker of the general public knows a word. Following up on this work, Johns (2021a) used a 55-billion-word collection of comments from the online discussion forum Reddit to implement a series of socially informed lexical strength measures. The goal of this work was to construct more ecologically valid notions of contextual diversity, which is a recent criticism of contextual diversity measures as put forth by Hollis (2020; see also Johns & Jones, 2021 for a discussion of these issues). Contextual and semantic diversity measures were recorded at the level of single comments (roughly equivalent to the claims of Adelman et al., 2006), subreddits (Discourse Contextual Diversity) and Reddit users (User Contextual Diversity). Subreddits contain comments surrounding a given topic, and roughly map onto a discourse topic.

When measuring semantic diversity at the discourse level and at the user level using the Semantic Distinctiveness Model, two context representations were proposed – word-based contextual representations and population-based contextual representations. In models using word-based contextual representations, context is represented by the frequency of the words used in each discourse or by a user and are consistent with previous instantiations of the model (Johns et al., 2020). In models using population-based contextual representations, contexts are defined by the commenting patterns of users across discourses, a more socially-oriented representation of linguistic context. For example, when computing Discourse Semantic Diversity or User Semantic Diversity for a word form like *cardinals*, we are accounting for the fact that this word form is employed across different subreddits (e.g. *r/nature* and *r/sports*) or by different specific users in the corpus (e.g., *Jennifer*, *Owen* and *Lily*). Word-based contextual representations for the two measures would consist of the frequency distributions of the words used in each subreddit (e.g. *r/nature*) or by each user (e.g., Jennifer) in which *cardinals* occurred. Population-based contextual representations for Discourse and User Semantic Diversity would instead measure how often each user comments in each subreddit (e.g. *r/nature*) containing *cardinals* and how often each user employing *cardinals* comments in each subreddit, respectively.

It was found that population-based models provided benchmark fits to a variety of large datasets examining lexical organization. This finding revealed that psycholinguistic measures of lexical strength can be built from socially-informed usage measures indicating to what extent different users/speakers employ a specific word across different communicative contexts. Lexical strength here is roughly intended as the resting-activation level or ease of lexical access of a given word or phrase, as indexed by on-line measures of lexical decision and naming, or off-line familiarity scores. Such definitions are in line with a usage-based view of the lexicon, which assigns language experience a primary role in shaping lexical representations (Bybee, 1985, 2010; Tomasello, 2003, 2009).

Johns (2021a) proposed that the role of contextual and semantic diversity is consistent with predictive accounts of language processing (e.g., Altmann & Mirkovic, 2009; Kutas & Federmeier, 2011). A recording of the types of contexts that a word occurs in, as is done in the Semantic Distinctiveness Model, allows for expectations to be constructed about the type of upcoming contexts that a word could occur in. The semantic diversity transformations employed by the model allows for a normalization of sorts, where new contextual occurrences are weighted more strongly if they are not redundant with past experiences, as this signals a new type of context that a word could occur in. The superiority of the models using population-based contextual representations over the ones using word-based representations suggest that linguistic contexts are not just composed of the words that occurred in a context, but extra-linguistic information is also included, such as the communicative discourse that words were used in and who produced them. From a predictive point of view, this suggests that the expectations that are being formed by the language processing system involve who one is going to be communicating with and under what circumstances. Given the central role given to communication in this account, the use of the population-based models is strongly related to usage-based and adaptive theories of language processing (Beckner et al., 2009; Christiansen & Chater, 2008; Tomasello, 2003, 2009).

Additionally, Johns (in press) has recently demonstrated that the advantage of discourse and user contextual diversity measures generalizes to accounting for word-level recognition memory rates, while Johns (2021b) showed that a distributional model trained with the communication patterns of users on Reddit allowed for a unique signal of word meaning to be constructed. For example, distributional models trained with word-based contextual

representations appeared to extract subordinate semantic category members as nearest neighbors of a word (e.g., different dog breeds for *poodle*, different sports for *basketball*). By contrast, models encoding users' communication patterns populated a word's semantic neighborhood with properties of that word (e.g., *grooming* and *matting* for poodle, *rebounds* and *athleticism* for basketball. Combined, the work of Johns (2021a,b; in press) suggests that communicative information is an integral part of lexical representations across a number of domains.

While the impact of semantic and social diversity on lexical strength has been mainly addressed at the single-word level, a wealth of corpus-based and experimental evidence suggests that an integral part of language usage involves fixed and recurrent multiword expressions (Christiansen & Arnon, 2017; Erman & Warren, 2000; Siyanova-Chanturia & Martinez, 2015), which could therefore serve as the building blocks of lexical organization beside single words. Multiword units constitute a heterogeneous class of expressions, including, among others, collocations (e.g., *torrential rain*, *open a bank account*), phrasal verbs (e.g., *look after*, *catch up*), irreversible binomials (e.g., *black and white*, *salt and pepper*), and idioms (e.g., *spill the beans*, *bury the hatchet*). Although all multiword language exhibits some degree of formal rigidity and semantic idiosyncrasy, idioms are usually regarded as the prototypical instance of this class of expressions, given their non-compositional nature and word-like processing with respect to matched literal phrases (e.g. *cook the beans*, *throw the hatchet*; Cacciari & Tabossi, 1988; Carrol & Conklin, 2019; Cronk & Schweigert, 1992; Siyanova-Chanturia, Conklin, & Schmitt, 2011; Titone, Lovseth, Kasparian & Tiv, 2019). As well, idioms form a heterogeneous class of expressions, exhibiting varying levels of semantic decomposability, literal plausibility, familiarity and formal rigidity, which are all defining features of multiword units (Wulff, 2008). Therefore, they will represent the focus of this preliminary investigation on contextual diversity and multiword language. Of note, hybrid models of idiom processing (Caillies & Butcher, 2007; Cutting & Bock, 1997; Libben & Titone, 2008; Smolka, Rabanus & Rösler, 2007; Sprenger, Levelt & Kempen, 2006; Titone & Connine, 1999; Titone & Libben 2014; Titone, Lovseth, Kasparian, & Tiv, 2019) underline the importance of variables like familiarity in mediating early-stage holistic retrieval of idiomatic forms from the mental lexicon, while compositional word-by-word parsing is expected to come into play only at a later stage.

The finding that contextual and semantic diversity measures provides a superior account to word frequency has obvious theoretical implications. However, it also has important applied

consequences for language learning and processing. The use of contextual and semantic diversity has recently been explored in relation to determining 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), speech therapy (Plante et al., 2014), reading (Joseph & Nation, 2018; Perea, Soares, & Comesana, 2013), and second-language acquisition (Frances, Martin, & Dunabeitia, 2020). Thus, determining the generality of these effects to multiword phrases could also inform the learnability of this information in the classroom.

In the present study, the socially-based diversity measures proposed by Johns (2021a) will be used to predict normative ratings of familiarity for a set of English idioms collected by Libben & Titone (2008). The models will be trained by treating idioms as equivalent to single words and determining the lexical strength of the idioms with a very large corpus of Reddit comments, totaling more than 55 billion words. If it is found that contextual diversity measures provide a similar advantage for idiomatic processing, it would demonstrate the intriguing possibility that contextual, semantic, and social diversity operate as organizing forces of the mental lexicon beyond the single-word level, signaling a new theoretical domain to explore the dynamics of lexical organization with.

The use of self-reported familiarity ratings as an index of idioms' lexical storage calls for a methodological clarification. Familiarity as used here refers to the subjective frequency with which a speaker has encountered a given idiom in written or spoken form, and whether they are confident about its meaning or not (Libben & Titone, 2008). Undoubtedly, off-line metalinguistic ratings do not reflect the time course of lexical access as closely as on-line evidence such as what is collected in eye-tracking and EEG data, but rather familiarity reflects the outcome of the comprehension process. Nonetheless, previous experimental research on on-line idiom processing has suggested that familiarity is the variable that predicts and modulates idioms' early-stage retrieval from the lexicon (Carrol & Conklin, 2020; Carrol & Littlemore, 2020; Cronk & Schweigert, 1992; Titone & Libben, 2014; Titone et al. 2019). In addition, while on-line (e.g., eye-tracking) idiom processing data are usually measured within a specific sentential context and are thus less generalizable, idiom familiarity ratings are collected for the expressions in isolation. Thus, they may have more general validity when used to predict all the corpus occurrences of an idiom across very diverse contexts. The first part of our analysis will be

devoted to demonstrating that single-word familiarity ratings exhibit the same patterns as lexical decision and naming latencies when correlating with corpus-based diversity measures, and hence can be used as an approximation of lexical access data.

The theoretical motivation of this research is to establish that higher-level principles of lexical organization, such as contextual, semantic, and social diversity, operate as organizing forces of the mental lexicon beyond the single-word level. By establishing that organizational principles built around social and contextual environmental information sources apply at the multiword level would open new pathways in the development of computational models of lexical organization.

Methods

Reddit data. The Reddit corpus was assembled from a website entitled pushshift.io (Baumgartner et al., 2020), which collects all Reddit comments for each month using the publicly available Reddit API[1]. All comments from users who had publicly available usernames were assembled from January 2006 to September 2019. Two types of corpora were assembled: user and discourse corpora. A criterion was set on the user corpora, where only users who had produced more than 3,000 comments were included in the resulting corpora. This resulted in 334,345 user corpora and 30,327 discourse corpora. Each user corpus contained 166,594 words (a sizeable sample of language; roughly equivalent to two fiction novels as calculated by Johns & Jamieson, 2019), while each subreddit corpora contained 1,838,334 words. The total number of single words contained in both corpora was approximately 55.7 billion words (the sum total number of words is the same for the user and discourse corpora, as each contain the same comments but organized differently).

Idioms. Libben and Titone (2008) collected normative ratings from 160 native English speakers for a set of 210 English idioms with a verb-determiner-noun structure (e.g. *she spilled the beans, he battled the storm*). Rated variables include *familiarity*, i.e. the subjective frequency with which subjects encounter an idiom, whether they know its meaning or not, *plausibility* of a literal meaning for an idiom string (e.g. *pulled the plug* vs *ate her words*) and *predictability* of an idiom's final word given the previous context (e.g. *twiddled her… thumbs* vs *bent the … law*). Other ratings measured whether an idiom's component words were semantically related to the

[1] API information at: https://www.reddit.com/dev/api/

overall figurative meaning (*global decomposability*), whether this relationship was literal (*normal decomposability*), and whether the verb and the noun were specifically related to the overall phrase meaning (*verb/noun relatedness*). Familiarity, literal plausibility and verb/noun relatedness were rated on a 1-5 Likert scale, while the other measures were expressed as a 0-1 proportion. Summary statistics on idioms' normative ratings are reported in Table 1.

The 210 idioms in this dataset were extracted from the Reddit corpus through a string-searching algorithm. To ensure that the extraction procedure included every possible morphological variant of the 210 idioms, for each idiom we generated a list of verb-inflected forms (e.g. *spills the beans, beans were spilled*) and replaced pronouns with a wildcard (e.g. *lose _ head*, *spill _ guts*). The string-searching algorithm was then instructed to consider every variant of each idiom.

When dealing with idiomatic phrases that are automatically extracted from corpora, a common and reasonable criticism is that some of them might also occur with a literal meaning depending on the context (e.g., *see the light, take the cake*). It is virtually impossible to ensure that all the corpus-derived occurrences of an idiom are indeed figurative, however, a few precautions in the extraction procedure can attenuate this risk of ambiguity. For example, most idioms occur with a fixed formal configuration and a fixed determiner with respect to the corresponding literal phrases (Fellbaum, 1993). When word sequences such as *see the light* or *go to town* are used literally, they can occur with different determiners, like *see a/the/that light* or *go to the/that town.* By contrast, when they are intended idiomatically, they almost exclusively occur with a definite determiner and no determiner, respectively (*see <u>the</u> light* and *go <u>to town</u>*). When extracting these idioms we were thus maximally conservative, only searching for different verbal inflections while keeping the rest of the wording fixed.

Count models. There will be three count models analyzed for the idiom data: 1) Frequency, 2) Discourse Contextual Diversity, and 3) User Contextual Diversity. Frequency is a count of the number of times that an idiom occurred within the Reddit data. Typically, a contextual diversity measure is contained in these analyses. The contextual diversityJohns (2021a) was a count of the number of comments a word appeared in (with repetitions within comments ignored). It was found that idioms are very rarely repeated within comments and so a resulting contextual diversity measure was virtually identical to frequency. Thus, this variable was not included in the analysis. Discourse Contextual Diversity is a count of the number of

discourses an idiom occurred in (with repetitions ignored within discourse). User Contextual Diversity is a count of the number of users who used an idiom (with repetitions again ignored).

Semantic diversity models. In the Semantic Distinctiveness Model, each occurrence of a word in a context results in a graded strength increase between 0 and 1 (repeated occurrences are ignored). To determine this encoding strength, the model uses an expectancy-congruency mechanism. When a word appears in a redundant context, compared to previous experience (meaning that the word would be expected to occur in such a context), the word is given a low encoding strength for that context. However, when the word appears in a unique context (meaning it would not have been expected to occur in that context), the word is given a high encoding strength for that context. In the version of the model used in Johns (2021a), the lexical strengths of words are accumulated in an external counter.

The two main components of the Semantic Distinctiveness Model are the representations used for words and contexts. Context representations contain information about the construction of the current context being processed, while the memory representations for words are the sum of the past contexts a word has occurred in. The update strength for a word is a transformation of the similarity between the word's representation in memory to the current context representation, with high similarity values (signaling a redundant context) being transformed into a low update strength and low similarity values (signaling a unique context) being transformed into a high update strength. Similarity values are transformed with an exponential transformation. The level of transformation is controlled with a parameter, λ, with larger parameter values decreasing the impact of high similarity contexts and increasing the strength of low similarity contexts.

Johns (2021a) sets out four new models based in the Semantic Distinctiveness Model architecture – two modifications of the Discourse Contextual Diversity and User Contextual Diversity counts. The divergences of the models were due to whether they used a word-based contextual representation or a population-based contextual representation. The models computing Discourse Semantic Diversity and User Semantic Diversity with a word-based representation are consistent with past implementations of the Semantic Distinctiveness Model (e.g., Johns et al., 2020), where the context representation is a vector of the word frequencies of that discourse (in this case, a subreddit or user corpus), and the memory representation of words was the sum of the word frequencies of the contexts that a word occurred in. The dimensionality

of the representation was the training word list from Johns (2021; *n*=81,261) plus the idiom data described previously (*n*=210) for a total dimensionality of 81,471.

In contrast, the models measuring Discourse Semantic Diversity and User Semantic Diversity with a population-based contextual representation do not employ a direct linguistic representation, but instead utilize a representation based in a count of the commenting pattern of words across users or discourses. For the population-based model capturing discourse-level semantic diversity, the context representation consisted of the number of comments each user made in that discourse (for a dimensionality of 334,345), while the context representation when measuring user-based semantic diversity consisted of the number of comments a single user made across all discourses (for a dimensionality 30,327).

For both representation types, the words' memory representations are updated for each word that occurred in a context. The update consists of summing the context representation into each word's memory representation. For the word-based representation models, the context representation is normalized (due to the magnitudes of high frequency words), while for the population-based representation models there is no normalization done.

The model computing Discourse Semantic Diversity with a population-based context representation measures how consistent language usage is by individuals within discourses. If a word has a relatively large strength in the model, then it would signal that the word is used across many discourses by an unpredictable set of users. In contrast, the model computing User Semantic Diversity with a population-based context representation measures how consistently a word is used across discourses by individuals. A word's relatively high strength in this model would signal that the word is produced by many individuals but with no predictable discourse pattern. Johns (2021a) found that the population-based models accounted for significantly more variance than the word-based models, suggesting that social and communicative information is an important source of information used in lexical organization. The following analysis will attempt to determine if this pattern is replicated with familiarity data for multiword expressions.

We will now provide a formal sketch of the Semantic Distinctiveness Model. For complete details of the implementation of the model, please see Johns (2021a). A word's strength in the Semantic Distinctiveness Model is updated with a semantic distinctiveness value, which is a transformation of the similarity between a word's representation and a context that the

word occurred in. Similarity is taken with a vector cosine (normalized dot product) between the two vectors:

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

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

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

Where *i* is the word being processed in context *j*, $\mathbf{M}_\mathrm{i}$ is the memory vector for that word, **c** is the context vector, and λ is a scaling parameter. λ controls the differential weight given to high versus low variability contexts, and it is the only free parameter in the model. A semantic diversity value signifies how unique the contextual occurrence of a word is, compared to the past contexts that a word has occurred in (as encoded in its memory representation). Finally, each word has its memory representation updated with the context representation:

$$\mathbf{M}_i = \mathbf{M}_i + \mathbf{c} \quad (3)$$

In the Semantic Distinctiveness Model, the λ parameter controls the amount of discounting applied to high similarity contexts and the amount of strengthening applied to low similarity contexts, and is a central component of the model. The operation of the word-based and population-based version of the models are identical, with only the content of the representations changing. In Johns (2021a) it was found that the population-based models were optimized with a maximized λ parameter and was set at 400. The same parameter was used here, and it was confirmed that the population-based model's performance was maximized at this parameter level for idioms. In contrast, Johns (2021a) found that the word-based models typically performed best with a low λ value of 1 (consistent with past implementations of the Semantic Distinctiveness Model, see Johns et al., 2020), and so this value was used for the word-based models in this article.

Table 2 briefly summarizes the frequency and diversity models that will be compared in the present analysis.

Analysis technique. Consistent with past studies (e.g., Adelman, et al., 2006; Johns, et al., 2016) 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 percent ΔR^2 improvement) for one predictor over other competing predictors. All strength variables were transformed with a natural logarithm.

Results

Given that idiom familiarity is the target measure of this article, it is first necessary to demonstrate that the advantage of the contextual diversity measures extend to word-level familiarity values, similar to the advantages that were seen in the lexical decision and naming data examined in Johns (2021a) and related studies. To accomplish this, the familiarity data from Clark and Paivio (2004; *n*=2,298; these data are an extension of the classic Paivio et al., 1968 norms) were analyzed.

Figure 1 (left panel) contains the correlation table of this data to word frequency and the different contextual diversity measures. This figure demonstrates that the correlations trends in the familiarity data mostly replicates the finding from lexical decision and naming data seen in Johns (2021a), with the small exception that the User Contextual Diversity count measure does not provide an advantage over word frequency, unlike what is found in previous lexical organization datasets.

To test whether the diversity variables account for more variance above and beyond word frequency, a regression analysis was done where the amount of unique variance that each diversity measure accounted for when frequency was controlled for (and vice versa) was assessed. The results of this analysis are contained in Figure 2, which shows that all of the diversity measures (with the exception of User Contextual Diversity) account for more variance than word frequency, while minimizing the amount of unique variance that word frequency accounts for. The measures that accounted for the most unique variance were the semantic-diversity models using a population-based representation, consistent with the results of Johns (2021a). This analysis establishes that familiarity data shows the same patterns as lexical decision and naming time.

Correlations between the different lexical strength variables and the idiom familiarity data are contained in Figure 3 (left panel). This figure shows that correlation patterns replicate the findings of Johns (2021a) and the word-level familiarity values analyzed above. As was seen in Figure 1 (left panel), the lexical strength measures are all highly intercorrelated with each other, although even more so for the idioms, likely due to less variance in the amount of

repetition in idiom usage and the smaller sample size for idioms. The fit to the correlations also follows previous findings, as the contextual diversity measures all have a higher correlation to this data compared to frequency, with the discourse-based and user-based semantic diversity models using a population-based contextual representation being the variables with the best fit, equivalently to past findings.

The high correlation among frequency, Discourse Semantic Diversity with a population-based representation and familiarity data is likely an example of a third variable problem, where the correlation between two variables is due to the presence of a third variable. To isolate the correlation between frequency and familiarity, and frequency and the Discourse Semantic Diversity measure, partial correlations were computed between these variables when the other competing variable was controlled for (e.g., frequency was controlled for when calculating the correlation between familiarity and population-based Discourse Semantic Diversity). Figure 1 (right panel) and Figure 3 (right panel) report partial correlations for single-word data from Clark and Paivio (2004) and idiom data when frequency is controlled for. For idioms, the partial correlation between population-based Discourse Semantic Diversity and familiarity is $r=.382$, $p<0.001$, a small reduction in the fit displayed in the left plot of Figure 3. The partial correlation between frequency and familiarity is $r=-.294$, $p<0.001$, a switch in the sign of the correlation for frequency, suggesting that when population-based Discourse Semantic Diversity is controlled for, frequency has a different role to play in the lexical organization of multi-word phrases. The following analyses will attempt to tease apart the role of these variables using hierarchical linear regression.

To determine whether the diversity measures account for more variance over frequency for the idiom familiarity data, a regression analysis was conducted calculating the amount of unique variance that the diversity measures account for over frequency (and vice versa), equivalent to the analysis contained in Figure 2. The results of this regression are contained in Figure 4 and show that similar trends were found in the idiom dataset as was found for single words. In particular, it was found that all of the diversity-derived measures accounted for more unique variance than frequency. However, unlike what was found in the single word analysis, frequency still accounted for significant amounts of unique variance, suggesting that frequency of occurrence is important in computing the lexical strength of multiword expressions. The best

overall model contained frequency and population-based Discourse Semantic Diversity, which accounted for 31.3% of the variance in this dataset.

However, it is possible that the population-based Discourse Semantic Diversity model is accounting for variance that other previously proposed psycholinguistic variables account for. To this possibility, an additional regression was conducted calculating the amount of unique variance that the Discourse Semantic Diversity and frequency variables account for when the following variables are also used as predictors: 1) literal plausibility (Titone & Connine, 1994), 2) verb relatedness (Titone, Lovseth, Kasparian & Tiv, 2019), 3) noun relatedness (Titone et al., 2019), 4) global decomposability (Gibbs & Nayak, 1989), 5) normal decomposability (Nunberg, 1978), and 6) predictability of final word (Cacciari & Tabossi, 1988). If the Discourse Semantic Diversity variable still accounts for unique variance in this regression, it would signal that the measure is accounting for previously unknown properties of lexical storage of multiword expressions.

The results of this analysis are contained in Figure 5, which shows that both the frequency and population-based Discourse Semantic Diversity measures account for significant levels of unique variance when these other psycholinguistic variables are contained in the regression – with 6.2% and 13% variance accounted for, respectively. The only other two variables that accounted for significant levels of unique variance were global decomposability and final word predictability. As reported in Figure 5, comparison of hierarchically nested models indicated that the addition of global decomposability and final word predictability resulted in significant predictive gains of 1.64% ΔR^2 and 5.54% ΔR^2 respectively. The overall best predictor was the population-based Discourse Semantic Diversity variable, indicating the importance of contextual information in the lexical storage of multiword expressions. Together, all variables accounted for 48.4% of the variance in the idiom familiarity data.

General Discussion

Previous studies have demonstrated the superiority of contextual diversity measures of lexical strength over the classic word frequency measure (Adelman et al., 2006; Johns et al., 2020; Johns, 2021). The goal of this study was to determine if these findings extend to multiword expressions, through the analysis of idiom familiarity values from Libben and Titone (2008). Frequency and contextual diversity measures were derived for idioms from a recently

constructed corpus of over 55 billion words from the online discussion forum Reddit (Johns, 2021). A similar advantage for the contextual diversity measures over frequency was found, indicating that multiword expressions show similar lexical organization dynamics as single words do.

The best fitting measures were found to be contextual diversity measures transformed with the Semantic Distinctiveness Model (Jones et al., 2012), with the models utilizing population-based representations offering the best fit. Population-based representations measure the communication patterns of users across or within discourses and offer a more socially-oriented view of lexical organization than purely linguistic proposals. In Johns (2021a) the model capturing User-level Semantic Diversity with a population representation was found to offer the best fit to single word data, however here the model capturing Discourse-level Semantic Diversity with a population representation performed best with idiom familiarity. The former model measures the consistency of word or phrase usage of individuals across discourses – a high strength value in this model would signal that the word or the idiom is being used by many individuals across a diverse set of discourses, while a low value would signal that they are used by only a subset of the population in a set number of discourses. The latter model measures the consistency of user communication within discourse – a high strength value in this model signals that a word or an idiom is used across discourses by a diverse set of individual language users, while a low value would signal that a word is used in a limited number of discourses by a predictable set of individuals. The finding that the population-based Discourse Semantic Diversity model provides the best fit to idiom familiarity suggests that the idioms that are stored strongest in memory are those that are used across many different discourses, with there being little predictability of who would be likely to use that expression.

The main theoretical debate in the construction of contextual diversity measures is the role of repetition and likely need in determining lexical organization (see Jones, Dye, & Johns, 2017 for an in-depth discussion of these issues and Westbury, 2021 for a related discussion in a different area of psycholinguistics). The findings of the present work are in line with Johns (2021a), as they support a lexical organization mechanism based upon lexical need calculations. In both works, likely need, as represented by contextual diversity indices, is shown to be a stronger determinant of idioms' lexical access as compared to repetition. These results confirm that even more complex structures like multiword units are to be included among the basic

combinatorial units of the lexicon beside single words. To illustrate the different type of distributional information that is encoded by a contextual diversity index such as population-based Discourse Semantic Diversity with respect to a repetition-based frequency index, it can be instructive to compare two idioms such as *drop a brick* and *hit the headlines.* Despite their comparable frequency values (1,294 and 1,293 respectively), the latter is more diverse in its use, as it has a higher Discourse Semantic Diversity score (40.30 *vs* 31.68), which indicates that its meaning can apply to a larger and more unpredictable set of contexts compared to *drop a brick.* Vice versa, two idioms like *hit the hay* and *face the music* have roughly equivalent diversity values (137.14 and 137.12 respectively), despite *hit the hay* being much less frequent (3,815 vs 11,157), as its meaning denotes a basic everyday event and can thus occur in a diverse set of contexts.

On the other hand, the present contribution diverges slightly from Johns' (2021) results on single words in that Discourse-level Semantic Diversity appears to play a bigger role here in determining idioms' lexical storage with respect to User-level Semantic Diversity. Although the limited sample size of the data analyzed in this study makes it difficult to draw any definitive conclusions on this issue, we can suggest two tentative explanations to be evaluated with future research. First off, longer word sequences such as idioms and multiword units tend to be more constrained in their semantics as they encapsulate more complex meanings with respect to distributionally freer and thus more polysemous single words. Hence, while different individuals can employ a given word (e.g., *get* or *bear*) with a different meaning and connotation depending on the context, we can expect an idiom such as *kick the bucket* to be intended quite consistently across different users.

Related to this point, a second explanation of the greater role of Discourse-based rather than User-based Semantic Diversity for idioms relates to the role of culturally-shared conceptual metaphors (Lakoff & Johnson, 1980; Reid & Katz, 2018a) in the perception of idioms' semantics. Since most idioms diachronically stem from the crystallization of linguistic and conceptual metaphors that are commonly shared by speakers within a given culture, we can imagine that they will be employed quite uniformly at the single-user level and that they will rather vary in their occurrence depending on the discourse.

An additional difference between the findings of this article and Johns (2021a) is that frequency here still accounts for unique variance in the idiom familiarity data. This suggests that

frequency of occurrence stills matters in idiomatic processing, with frequency combined with contextual factors determining speakers' familiarity with different multiword expressions. This may be due to idioms being relatively low in frequency compared to most single words, with that leading to a less consistent contextual representation for multiword expressions. How exactly these different environmental variables combine to determine idiomatic storage needs to be determined with future modeling efforts.

From the perspective of idiom processing theories, the results obtained in the present work call for a few considerations. Of note, discourse-based contextual diversity accounted for the most variance in idiom familiarity ratings even when controlling for other standard idiom-related variables, demonstrating the importance of taking the role of this measure into greater consideration in future idiom processing research. Secondly, the fact that idioms' lexical strength appears to be influenced by the same variables that impact the lexical organization of single words confirms that idioms at least partly possess a unitary word-like status in the lexicon, as hypothesized by non-compositional (Bobrow & Bell, 1973) and hybrid models of idiom processing (Libben & Titone, 2008). One could hypothesize that such word-like behavior of idioms in lexical organization is mostly driven by their semantic non-compositionality. While the fact that contextual diversity remained significant even after controlling for decomposability is at odds with this hypothesis, median-splitting the idiom set for global decomposability revealed a higher correlation between population-based Discourse Semantic Diversity and Familiarity for less decomposable ($r = .6$, $p<.001$) rather than more decomposable ($r =.41$, $p<.001$) idioms. These preliminary results suggest that Discourse-based Semantic Diversity is more predictive of perceived familiarity for those idioms that are more opaque in their semantic structure and thus behave more similarly to single words[2]. Further clarification could come from future experiments on collocations (e.g., *torrential rain*, *strong tea*), which, despite being conventional and lexically rigid, preserve their semantic compositionality. Finally, future research addressing the role of verb contextual diversity and noun contextual diversity individually could shed

[2] To further verify that frequency overrides decomposability, as predicted by idiom processing models, we median-split our dataset for familiarity and ran two separate linear regressions where we predicted familiarity scores for high-familiar vs low-familiar idioms from Discourse Semantic Diversity, global decomposability and their interaction. Notably, both diversity ($t = 4.57$, $p <.001$) and decomposability ($t = 2.16$, $p <.05$), but not their interaction, appeared to be significant only when predicting low familiarity scores. Nonetheless, since the interaction between diversity and decomposability was not significant, further investigations will need to clarify if decomposability affects the way in which computational scores of diversity can model the lexical strength of idiomatic and multiword expressions.

further light on which components of a multiword phrases are more salient in determining its lexical strength.

As discussed in the Methods section, the potential ambiguity of some idioms between a literal and a figurative meaning depending on the context (e.g. *see the light*) was accounted for by a formally conservative extraction procedure, which only looked at different verbal inflections while keeping the rest of the target phrase intact. Interestingly, when median-splitting the idiom set by literal plausibility, a higher correlation between familiarity and population-based Discourse Semantic Diversity was obtained for more literally plausible ($r=.6$, $p<.001$) than for less literally plausible ($r=.4$, $p<.001$) idioms. We can thus hypothesize that sensitivity to discourse context is even more reliable as a predictor of familiarity for phrases that can have very different (literal vs figurative) meanings depending on the context.

The analyses reported above on the role of global decomposability and literal plausibility dovetail interestingly with previous on-line evidence supporting a multidetermined view of idiom processing (Libben & Titone, 2008). Both in the present work and in past studies using cross-modal priming (Titone & Libben, 2014) and eye-tracking (Titone et al., 2019), lower semantic decomposability facilitates the identification of idiomatic strings as holistic units. As well, the data presented here confirm that literal plausibility is an important determinant of idiom processing, whose role needs to be clarified on a contextual basis. Future investigations will assess the potential of corpus-based diversity indices to predict on-line idiom processing data such as semantic priming latencies and eye-movement measures of reading.

The work described here complements previous studies on the lexical organization of single words and contributes interesting evidence to ongoing research on cognitively plausible computational models of multiword language (Constant et al. 2017; McCauley & Christiansen, 2017) and nonliteral language phenomena, including metaphors (Reid, Al-Azary & Katz, 2020; Reid & Katz, 2018b). Overall, the finding that higher level measures of language usage, such as contextual, semantic, and social diversity, which have been established to be important in lexical organization at the single word level, also generalize to multi-word expressions suggests a much more dynamic lexicon than has previously been proposed. In particular, it suggests that multi-word expressions, and idioms in particular, are also included in the organization of language, and are controlled by similar principles as single words. This result indicates that new theoretical

accounts of lexical organization need to be constructed that also include multi-word units in a model's lexicon and organizational principles, a substantial task for future development.

Open practices statement. Data and model values used in this article are available at https://osf.io/y93ar/.

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.

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. https://psycnet.apa.org/doi/10.1037/0033-295X.96.4.703

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

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.

Bobrow, S. A., & Bell, S. M. (1973). On catching on to idiomatic expressions. *Memory & Cognition, 1*, 343-346. https://doi.org/10.3758/bf03198118

Broadbent, D. E. (1967). Word-frequency Effect and Response Bias. *Psychological review, 74*(1), 1-15. https://doi.org/10.1037/h0024206

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. https://doi.org/10.1177/0963721417727521

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

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins Publishing.

Bybee, J. (2010). *Language, usage and cognition.* Cambridge University Press.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, *27*, 668-683. https://doi.org/10.1016/0749-596X(88)90014-9

Caillies, S., & Butcher, K. (2007). Processing of idiomatic expressions: Evidence for a new hybrid view. *Metaphor and Symbol, 22*(1), 79-108. https://doi.org/10.1080/10926480709336754

Carrol, G., & Conklin, K. (2019). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech, 63*, 95-122. http://dx.doi.org/10.1177/0023830918823230

Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech, 63*, 95-122. https://doi.org/10.1177%2F0023830918823230

Carrol, G., & Littlemore, J. (2020). Resolving figurative expressions during reading: The role of familiarity, transparency, and context. *Discourse Processes*, 57, 609-626. https://doi.org/10.1080/0163853X.2020.1729041

Christiansen, M. H., & Arnon, I. (2017). More Than Words: The Role of Multiword Sequences in Language Learning and Use. *Topics in Cognitive Science, 9*, 542-551. https://doi.org/10.1111/tops.12274

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*, 489-509.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behavior Research Methods, Instruments, & Computers, 36, 371–383. https://doi.org/10.3758/BF03195584

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics, 43*, 837-892.

Cronk, B. C., & Schweigert, W. A. (1992). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics, 13,* 131–146. http://dx.doi.org/10.1017/S0142716400005531

Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & cognition, 25*(1), 57-71. https://doi.org/10.3758/BF03197285

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk, 20*(1), 29-62. https://doi.org/10.1515/text.1.2000.20.1.29

Fellbaum, C. (1993). The determiner in English idioms. *Idioms: Processing, structure, and interpretation*, 271-295.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627-635. https://doi.org/10.1016/S0022-5371(73)80042-8

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.

Gibbs, R. W., Jr., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21,* 100–138. https://doi.org/10.1016/0010-0285(89)90004-2

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, 104146.

Johns, B. T. (2021a). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*

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., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73, 841-855. https://doi.org/10.1177%2F1747021819897560

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. https://dx.doi.org/10.1121%2F1.4731641

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

Johns, B. T. (2021b). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*.

Johns, B. T. (in press). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. *Memory & Cognition*.

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across aging and bilingualism. Frontiers in Psychology, 7, 703:1–11. https://doi.org/10.3389/fpsyg.2016.00703

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizational principle of the lexicon. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 67, 239–283). San Diego, CA: Elsevier Academic Press

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. https://doi.org/10.1037/a0026727

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.

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.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* University of Chicago press.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36,* 1103–1121. http://dx.doi.org/10.3758/MC.36.6.1103

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.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science, 9*, 637-652.

McDonald, S., & Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language & Speech*, 44, 295–323.

Nunberg, G. (1978). *The pragmatics of reference*. Bloomington, IN: Indiana University Linguistics Club.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*, 1-25. https://doi.org/10.1037/h0025327

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., ... & Burton, R. (2014). Variability in the language input to children enhances learning in a treatment context. *American Journal of Speech-Language Pathology, 23*, 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. https://psycnet.apa.org/doi/10.3758/s13423-019-01691-w

Reid, J. N., Al-Azary, H., & Katz, A. N. (2020). Metaphors: Where the neighborhood in which one resides interacts with (interpretive) diversity. *Proceedings of CogSci.*

Reid, J. N., & Katz, A. N. (2018a). Something false about conceptual metaphors. *Metaphor and Symbol*, *33*, 36-47. https://doi.org/10.1080/10926488.2018.1407994

Reid, J. N., & Katz, A. N. (2018b). Vector space applications in metaphor comprehension. *Metaphor and Symbol, 33*, 280-294. https://doi.org/10.1080/10926488.2018.1549840

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, 105312.

Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS One*, *12*(6), e0179004.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 251–272. http://dx.doi.org/10.1177/0267658310382068

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics, 36*, 549-569. https://doi.org/10.1093/applin/amt054

Smolka, E., Rabanus, S., & Rösler, F. (2007). Processing Verbs in German Idioms: Evidence Against the Configuration Hypothesis. *Metaphor and Symbol*, 22(3), 213-231. 10.1080/10926480701357638

Sprenger, S. A., Levelt, W. J., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of memory and language*, *54*(2), 161-184. https://doi.org/10.1016/j.jml.2005.11.001

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*.

Titone, D. A., & Connine, C. M. (1994). Comprehension of idiomatic expressions: Effects of predictability and literality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1126–1138. http://dx.doi.org/10.1037/0278-7393.20.5.1126

Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of pragmatics*, *31*(12), 1655-1674. https://doi.org/10.1016/S0378-2166(99)00008-9

Titone, D., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A crossmodal priming investigation. *The Mental Lexicon, 9,* 473–496. https://doi.org/10.1075/ml.9.3.05tit

Titone, D., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Canadian journal of experimental psychology/Revue canadienne de psychologie experimentale*, *73*, 216-230. https://psycnet.apa.org/doi/10.1037/cep0000175

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. (2021). Prenominal adjective order is such a fat big deal because adjectives are ordered by likely need. *Psychonomic Bulletin & Review*, 28, 122-138.

Wulff, S. (2008). Rethinking idiomaticity: A usage-based approach. A&C Black.

Dimension	Mean	SD	Min	Max
Familiarity (/5)	3.34	0.90	1.57	4.97
Literal Plausibility (/5)	2.99	1.17	0.77	5
Global Decomposability (/1)	0.53	0.25	0.03	0.98
Normal Decomposability (/1)	0.48	0.26	0	1
Verb Relatedness (/5)	2.73	0.97	0.56	4.93
Noun Relatedness (/5)	2.82	1.13	0.54	4.93
Final Word Predictability (/1)	0.17	0.23	0	0.96

Table 1. Normative characteristics of the 210 idioms in the dataset (from Libben & Titone, 2008).

Model	Abbreviation	Description
Word Frequency / Frequency	WF / Freq	Total occurrences of a word/phrase in the corpus
Contextual Diversity	CD	Number of comments a word/phrase occurs in
Discourse Contextual Diversity	DCD	Number of subreddits a word/phrase occurs in
User Contextual Diversity	UCD	Number of users employing a word/phrase
Discourse Semantic Diversity (word-based context)	DCD-SD-WR	Measures the extent to which a word/phrase is used across many subreddits containing different word frequency distributions
User Semantic Diversity (word-based context)	UCD-SD-WR	Measures the extent to which a word/phrase is used by many users who in turn differ in their word frequency distributions
Discourse Semantic Diversity (population-based context)	DCD-SD-PR	Measures the extent to which a word/phrase is used across many subreddits that are commented on by an unpredictable pool of users
User Semantic Diversity (population-based context)	UCD-SD-PR	Measures the extent to which a word/phrase is used by many users who comment on an unpredictable set of subreddits

Table 2. Summary of the frequency and diversity models compared in the analysis. Abbreviations are used to refer to the models in the plots and in the supplementary materials.

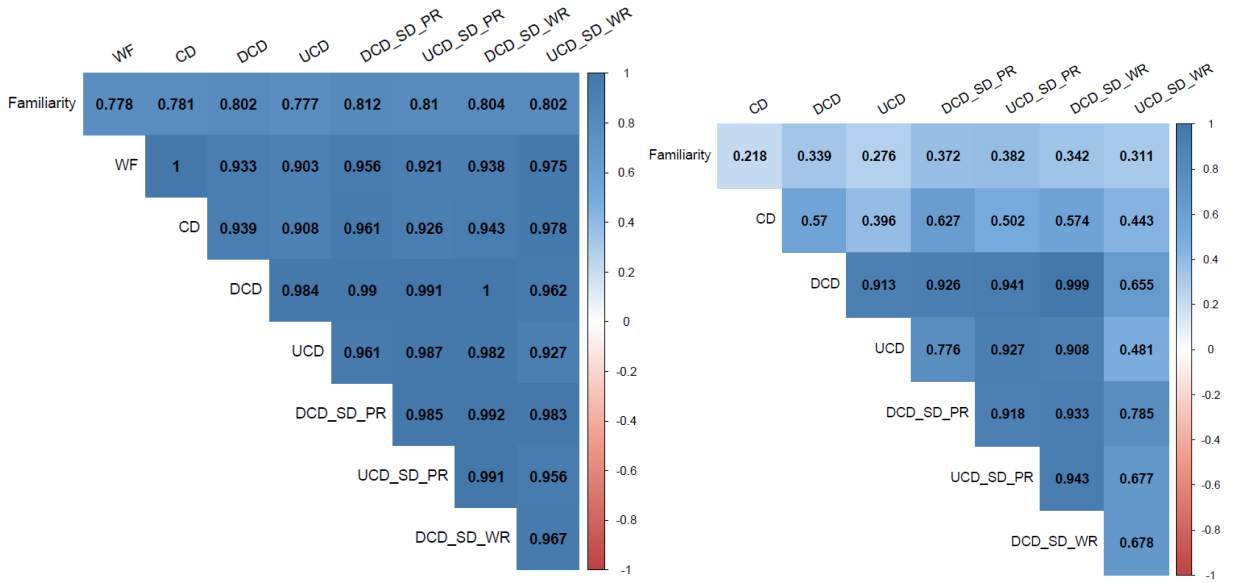

Figure 1. The left plot reports correlations among the different lexical strength variables and the Clark and Paivio (2004) word familiarity data. All correlations are significant at the $p<0.001$ level. The right plot reports partial correlations when word frequency is controlled. WF = word frequency; CD = contextual diversity; DCD = Discourse Contextual Diversity; UCD = User Contextual Diversity; DCD-SD-WR = word-based Discourse Semantic Diversity; UCD-SD-WR = word-based User Semantic Diversity; DCD-SD-PR = population-based Discourse Semantic Diversity; UCD-SD-PR = population-based User Semantic Diversity

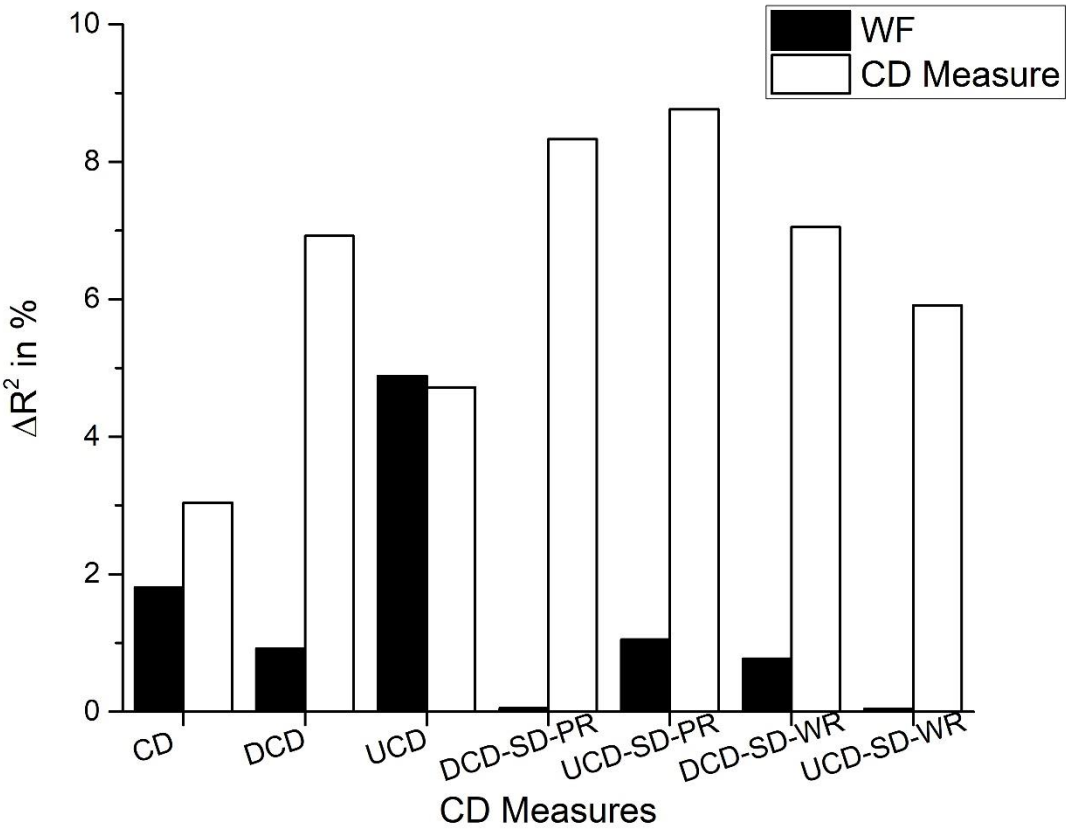

Figure 2. The amount of unique variance that WF and the various CD measures account for in the word familiarity dataset of Clark and Pavio (2004). This finding replicates the results of Johns (2021a) where the SD-PR models account for the most variance in large lexical decision and naming datasets. WF = word frequency; CD = contextual diversity; DCD = Discourse Contextual Diversity; UCD = User Contextual Diversity; DCD-SD-WR = word-based Discourse Semantic Diversity; UCD-SD-WR = word-based User Semantic Diversity; DCD-SD-PR = population-based Discourse Semantic Diversity; UCD-SD-PR = population-based User Semantic Diversity

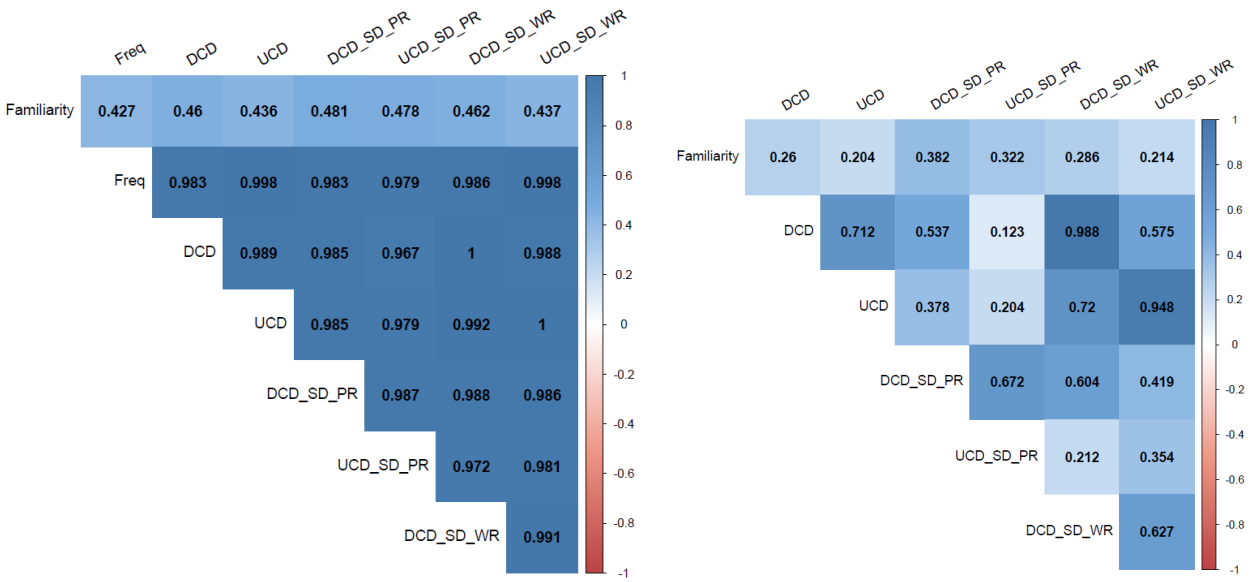

Figure 3. The left plot reports correlations among the different lexical strength variables and the idiom dataset. All correlations are significant at the $p<0.001$ level. The right plot reports partial correlations when frequency is controlled. Freq = frequency; DCD = Discourse Contextual Diversity; UCD = User Contextual Diversity; DCD-SD-WR = word-based Discourse Semantic Diversity; UCD-SD-WR = word-based User Semantic Diversity; DCD-SD-PR = population-based Discourse Semantic Diversity; UCD-SD-PR = population-based User Semantic Diversity

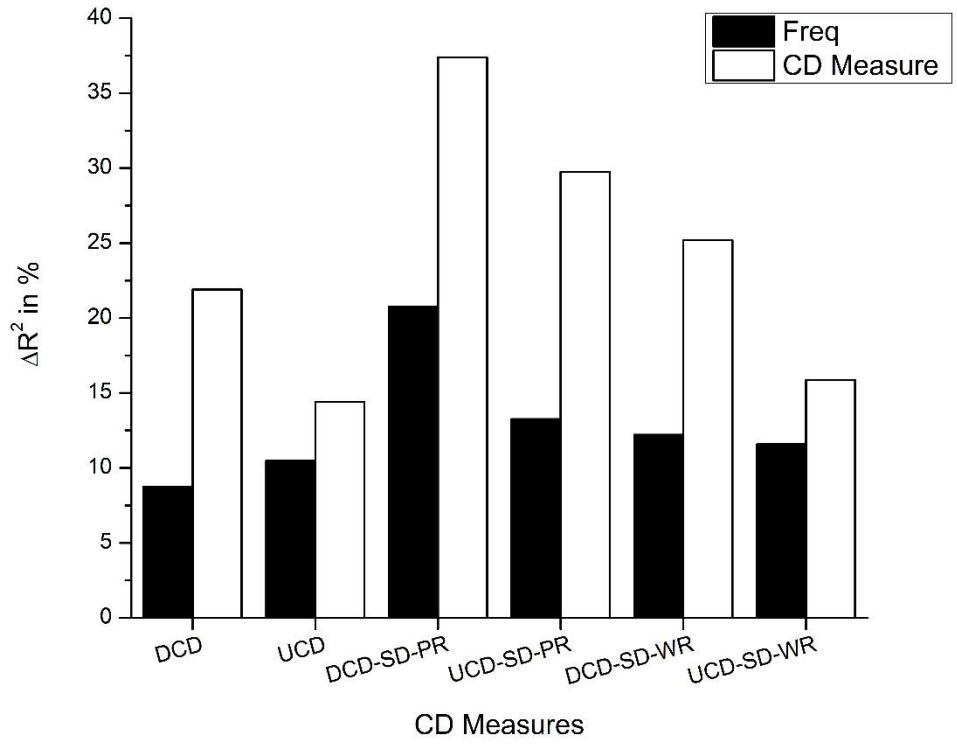

Figure 4. The amount of unique variance that frequency and the various diversity measures account for in the idiom familiarity data. This finding demonstrates that the various diversity measures account for considerable levels of variance above and beyond frequency, similar to the results on single word data. However, unlike single word data, frequency still accounts for significant levels of variance. Freq = frequency; CD = contextual diversity; DCD = Discourse Contextual Diversity; UCD = User Contextual Diversity; DCD-SD-WR = word-based Discourse Semantic Diversity; UCD-SD-WR = word-based User Semantic Diversity; DCD-SD-PR = population-based Discourse Semantic Diversity; UCD-SD-PR = population-based User Semantic Diversity

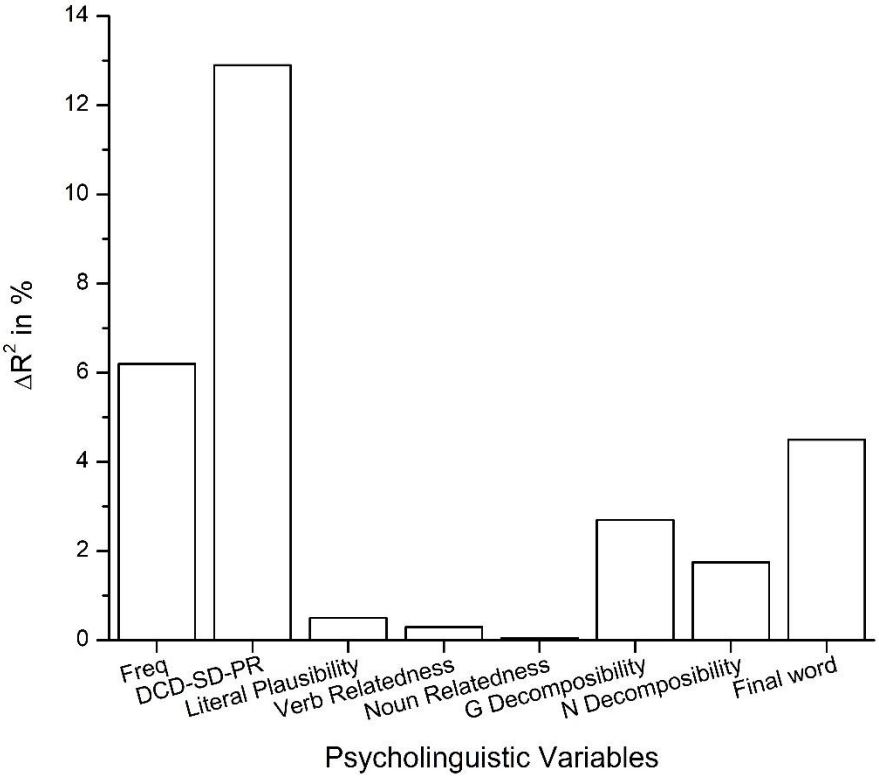

Figure 5. The amount of unique variance that the various psycholinguistic variables account for in the idiom familiarity data. Freq = frequency; DCD-SD-PR = population-based Discourse Semantic Diversity