Estimating the Prevalence and Diversity of Words in Written Language

Brendan T. Johns[1], Melody Dye[2], & Michael N. Jones[3]

[1] *University at Buffalo*

[2] *University of California, Berkeley*

[3] *Indiana University, Bloomington*

In Press, *Quarterly Journal of Experimental Psychology*

Correspondence
Dr. Brendan Johns
Dept. of Communicative Disorders and Sciences
122 Carey Hall
University at Buffalo
Buffalo, NY, 14214

Email: btjohns@buffalo.edu
Phone: (716) 829-2797
Fax: (716) 829-3979

Abstract

Recently, a new crowd-sourced language metric has been introduced, entitled word prevalence (Brysbaert, Mandera, McCormick, & Keuleers, in press; Brysbaert, Stevens, Mandera, & Keuleers, 2016; Keuleers, Stevens, Mandera, & Brysbaert, 2015), which estimates the proportion of the population that knows a given word. This measure has been shown to account for unique variance in large sets of lexical performance (Brysbaert, et al., 2016, in press; Keuleers, et al., 2015). This article aims to build on the work of Brysbaert, et al. (2016) and Keuleers, et al. (2016) by introducing new corpus-based metrics that estimate how likely a word is to be an active member of the natural language environment, and hence known by a larger subset of the general population. This metric is derived from an analysis of a newly collected corpus of over 25,000 fiction and non-fiction books, and will be shown that it is capable of accounting for significantly more variance than past corpus-based measures.

Estimating the Prevalence and Diversity of Words in Written Language

In the study of lexical processing and retrieval, a host of explanatory variables have been proposed to account for human performance in tests of language processing. These variables include objective measures, based on environmental occurrence (such as frequency and contextual diversity) and surface characteristics (such as word length and orthographic neighborhood density), as well as subjective measures, derived from human judgments (such as valence and concreteness; for a comprehensive review, see Baayen, Ramscar, & Milin, 2016). In large-scale regression analyses incorporating these variables, word frequency has proven to be one of the most robust behavioral predictors across an array of tasks, including lexical decision and word naming (Adelman, Sabatos-DeVito, Marquis, & Estes, 2014; Baayen, Feldman, & Schreuder, 2006; Balota et al., 2007; Brysbaert, Mandera, & Keuleers, 2018; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Yap & Balota, 2009).

In the simplest case, a word's normative frequency is a register of the number of separate occurrences of that word in a ***corpus***—i.e., a collection of text or speech comprising a large set of different documents. As an explanatory variable, word frequency (WF) acts as a kind of proxy for linguistic experience. A key simplifying assumption is that the corpus from which the counts are drawn is representative of the 'average' speaker's experience with the language. However, while this is almost certainly true for common words, which fall in the higher ranges of the frequency spectrum, it is far less likely to be true of lower frequency words, where individual exposure is highly variable, and not well-reflected by averages. Indeed, beyond grammar school, speakers sample language in increasingly directed and idiosyncratic ways (Gardner, et al., 1987), making it impossible to reliably estimate vocabulary size from small word samples (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). It has thus been argued that for certain psychometric tests, experiential familiarity ratings may prove a better measure than those that are corpus-based, particularly when sampling in the tail of the distribution (Allen & Garton, 1968; Gernsbacher, 1984).

In line with this, a novel measure has been proposed to help account for speaker variability at scale: ***word prevalence*** (Brysbaert, Stevens, Mandera, & Keuleers, 2016; Brysbaert, Mandera, McCormick, & Keuleers, in press; Keuleers, Stevens, Mandera, & Brysbaert, 2015). Word prevalence is a measure of the proportion of the population that knows a given word, and

relies on large samples of words and participants for its validity. To establish word prevalence scores for Dutch, Keuleers et al. tested items from a 50,000+ word master list on over 275,000 native speakers from Belgium and the Netherlands. Subjects were presented with a sequence of words and non-words, and asked to identify which words were known to them, selecting from either "Yes, I know this word" or "No, I do not know this word". One of the principal aims of the study was to establish whether word prevalence would have explanatory value over and above corpus-based frequency measures.[1] On the face, word prevalence[2] offers a complementary representation of the distribution of lexical experience: Words that are universally known nevertheless vary widely in their frequency of use, and words that are used rarely vary widely in their prevalence.

When word prevalence was pitted against frequency in an analysis of visual lexical decision times in the Dutch Lexicon Project (DLP), it was found to explain similar amounts of unique variance, and to be the best overall predictor of response times (RT; see also Brysbaert, Stevens, Mandera, & Keuleers, 2016). A closer assessment of the contributions of each variable indicated that the better known a word was, the more its frequency mattered as a predictor; conversely, for rare words, in the lower half of the frequency spectrum, the effect size of frequency was minimal. Keuleers, et al. concluded that prevalence provides the better estimate of rare words, and frequency of widely known words. More recently, Brysbaert, Mandera, McCormick, & Keuleers (2019) demonstrated that these findings also hold for data from the English lexicon project (Balota, et al., 2007).

While this finding is striking, the practice of using prevalence to predict lexical decision response times should give room for pause. In a typical LDT, subjects are asked to determine whether an item is a real word in their language (rather than whether they know it), and are judged on both speed and accuracy (rather than accuracy alone). Word prevalence can thus be classified as a modified lexical decision task: The test design closely resembles that of a standard

[1] Another goal of the study was to assess vocabulary size in the Dutch-speaking population and establish its determinants. Individual vocabulary scores were computed as the percentage of hits (correct identifications of real words) minus the percentage of false alarms (incorrect identifications of non-words). Among the major positive contributions to vocabulary size were age, educational attainment, and number of foreign languages spoken.

[2] Word prevalence was estimated from ~250 measurements per word, using a fitted explanatory item response model to predict 'item difficulty' (Keuleers et al., 2015). In this article, we used the probit transformed prevalence measures from Brysbaert, et al. (2019).

LDT, and the relevant variations – in the precise form of the question asked and in the variable(s) of interest – are minor.

This is where a potential issue arises. In lexical decision, RT and accuracy are tightly coupled, with a canonical trade-off between speed and accuracy. In a model of LDT response latencies, accuracy would thus invariably be a strong predictor. However, it is not common practice to predict response time from accuracy, because this has relatively little explanatory value. For the most part, cognitive modelers are less interested in characterizing the relationship between behavioral outputs, and more in the environmental inputs that produce one or the other behavior.[3] Yet using word prevalence to explain lexical decision falls prey to precisely this criticism: One performance measure is being used to explain another.

In constructing a model of human behavior, a cognitive modeler must specify both the representational input (the relevant environmental structure) and the cognitive process that operates over that input (Estes, 1975). For instance, in a highly simplified model of word recognition, each repetition of an item in experience serves to lower its resting state threshold; as a result, more frequently experienced items are processed more efficiently (Morton, 1969). In such a model, the relevant input for an item is its frequency, and the mechanism that produces behavior is an internal counter that adjusts the item's resting state. Clearly, representation and process are interdependent: The form and complexity of the assumed process is contingent on the choice of representation, and vice versa. The explanatory value of a model thus depends on the judicious choice of representation and process, and the validity of the assumptions underlying those choices (Johns & Jones, 2010; Johns, Jones, & Mewhort, 2012; Johns, Mewhort, & Jones, 2017).

Given this approach to modeling, a problem arises when representation and process are conflated, as occurs when one type of behavior is used to predict a highly similar form of behavior (for an extended critique, see Jones, Hills, & Todd, 2015 and Johns, Jamieson, & Jones, in press). When word prevalence is used as an independent variable it is subject to this critique: The vocabulary test devised by Brysbaert, et al. (2019) and Keuleers et al. (2015) does not deliver a pure readout of the prior lexical experience of speakers; instead, what it delivers is an introspective human judgment (a behavior) that reflects retrieval (process) from semantic

[3] When both speed and accuracy are considered jointly, patterns of responding can be informative about how participants set decision thresholds (see e.g., Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

memory (representation)[4]. The process at work in lexical decision is thus already partially embedded in the word prevalence measure. Jones et al. (2015) refer to this as a 'Turk problem', in reference to the famous 18th century chess-playing machine, which appeared to function as a self-operating automaton, but in fact, concealed a human chess master within. In cognitive modeling, a Turk problem arises when the representational input is derived directly from human behavioral data, hiding the requisite process complexity—the "man in the machine"—within the representation itself.

This is not to say that the data collected by Brysbaert, et al. (2019) and Keuleers, et al. (2015) is not valuable or important. Indeed, the data from these studies will likely play an important role in theory development in word processing for years to come. However, word prevalence measures are still data, and data needs to be understood with theory.

Corpus-Based Measures of Prevalence

Psychologists studying lexical processing face a conundrum. Introspective measures of familiarity have revealed variability in lexical knowledge that is not well-captured by aggregate corpus-based measures, like frequency (Allen & Garton, 1968; Gernsbacher, 1984; Keuleers et al., 2015). However, objective, environmentally-derived measures have the advantage of being interpretable, replicable, and more straightforwardly incorporated into computational models (McDonald & Shillcock, 2001). Deriving psycholinguistic measures from corpora offer a more objective measurement of the types of materials that one might have been exposed to, as they are not obtained through introspective judgements or other psycholinguistic tasks, but instead are measurements of a type of language that a wide variety of authors have used. An open question, then, is whether it might be possible to develop a prevalence measure that is estimated from the linguistic environment, rather than from introspective ratings.

In recent years, a number of corpus-based measures have been developed that improve on raw frequency counts. For instance, rather than computing a word's overall occurrence rate, measures of ***contextual diversity*** (CD) return the number of separate documents a word occurs in (Adelman et al., 2006; McDonald & Shillcock, 2001). Like CD, measures of ***semantic diversity***

[4] Word prevalence is a more objective measure of word knowledge than other psycholinguistic variables, such as concreteness or familiarity, as it is a proxy for one's receptive vocabulary size. However, when using it to explain lexical decision or naming it is unclear why it is explaining additional variance, given the overlapping similarities between the tasks.

(SD) yield a weighted document count, with the weight determined by a similarity distance-metric between documents: The more similar the contexts in which a word occurs, the less each separate occurrence is weighted (see Hoffmann, Lambon Ralph, & Rogers, 2012; Hsiao & Nation, 2018; Jones, Johns, & Recchia, 2012; Johns, Gruenenfelder, Pisoni, & Jones, 2012; Johns, Dye, & Jones, 2016; Johns, Sheppard, Jones, & Taler, 2016). In terms of predictive power, context-based measures have been found to consistently outperform frequency (see Jones, Dye, & Johns, 2017 for a review).

Here, we outline and contrast two different classes of environmental corpus-based variables: 1) occurrence-based variables, and 2) prevalence-based variables. Occurrence-based variables include WF, CD, and SD models. The central principle of these variables is that the weight of a word is updated with each occurrence of a word (such as a WF count) or each occurrence within a limited size context (such as a paragraph or document in the case of the CD and SD variables).

The prevalence-based measures are adapted contextual diversity measures, taking place at much larger units of language. Specifically, prevalence will be measured at two levels: 1) ***book prevalence*** (BP), and 2) ***author prevalence*** (AP). BP measures the number of books that a word appears in, while AP measures the number of authors that used a word in their writings. We entitle these prevalence measures as they are simply measuring whether a word is used across large swathes of language. If all authors use a word, regardless of the frequency of that word, it is likely that a person would have experienced that word. However, if a word only occurs in certain writings (such as only in *fantasy* novels), then only a subset of the population may have encountered that word. Johns and Jamieson (2018) recently demonstrated that there is meaningful semantic variation at both the book and individual author level. Additionally, more recent research has demonstrated that there are systematic differences in language usage based on the demographic characteristics of authors, such as gender (Johns & Dye, 2019) and time and place of birth (Johns & Jamieson, in press). Thus, both measures will be modified with a semantic diversity transformation, using computational techniques adapted from the semantic distinctiveness model (Jones, et al., 2012; Johns, et al., 2012). This will yield two more measures: 1) ***semantic diversity-book prevalence*** (SD-BP), and 2) ***semantic diversity-author prevalence*** (SD-AP).

As stated, this study will contrast occurrence-based counting (counting at small units of context; the WF, CD, and vSDM measures) and prevalence-based counting (counting at large units of measurements; the BP, AP, SD-BP, and SD-AP measures). However, there will also be a contrast of diversity measures – CD vs. vSDM for the occurrence-based counts, and BP, AP vs. SD-BP, SD-AP for the prevalence-based counts. The occurrence/prevalence comparisons will allow for a determination of the effect of measuring language at different levels, while the diversity measures will determine whether the semantic diversity transformations extend to new datasets, corpora, and levels of analysis, consistent with past results (Hsiao & Nation, 2018; Hoffmann, et al., 2012; Jones, et al., 2012; Johns, et al., 2012, 2016).

The results of Brysbaert, et al. (2016, 2019) and Keuleers, et al. (2015) demonstrates that there is significant variability in terms of people knowing that a given word is a part of their language. The motivation for the current work is determining whether better measures can be constructed to examine the reasons for this variability. Across the occurrence-based and prevalence-based variables there is a total of seven different variables. These variables will be contrasted on both lexical decision and naming data from lexicon projects (Balota, et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and the word prevalence measures of Brysbaert, et al. (2019). Additionally, the language materials used to train the lexical variables are organized by author's country of birth and gender, enabling an analysis on the effects of differential linguistic experience on lexical behavior, an important goal in the understanding on the interaction between experience and behavior on language processing (see Johns & Jamieson, 2018; Johns, Jones, & Mewhort, 2019; van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

Materials

The lexical materials assembled here consists of books organized by author and genre. To organize the book set, the dominant genre that an author wrote in was recorded by using the most frequent tag on the book review websites *GoodReads* and online retailer *Amazon*. The books written by that author were then labelled as having being written in that genre. Although this is less precise than author studies examining the impact of genre on writing (see Johns & Jamieson, 2018), tagging each book by its genre was unfeasible for such a large collection.

The characteristics of the book set, organized by genre, is contained in Table 1. Overall, there were thirteen different genres of books, consisting of over 26,000 books written by 3,200 authors and containing approximately 2.1 billion words. There was some variability in terms of

the number of words per book, with young adult novels having the lowest average number of words, and historical fiction books having the highest. However, all book types comprised large amounts of text, and should therefore offer a fair test of the different models of lexical organization used here.

Given that the book set assembled here is organized by author, it was possible to organize the materials by author characteristics. This is a benefit of using books as a lexical source, compared to subtitles or newspaper articles, as it is possible to isolate some personal characteristics of the person who produced the language source. Specifically, the country of birth and gender of each author was recorded, to determine how these factors influence a model's fit to lexical behavior. For country of birth, most authors came from the USA or UK, with smaller collections from Canadian and Australian authors. Characteristics of the book set split by country of birth of the authors is contained in Table 2.

Finally, Table 3 contains the characteristics of the book set, split by author gender. This table shows that there are slightly more female authors, but on average male authors produced more books that were slightly longer. This led to male authors producing about 130 million words more than female authors in this sample. The splits described in Table 2 and 3 will be used to determine the effects of author demographics on fits to lexical behaviors.

Models

As previously discussed, there will be both occurrence-based and prevalence-based variables used in this study. There are three occurrence-based models: WF, CD, and SDM. There are four prevalence-based models: BP, AP, SD-BP, and SD-AP. Each model will be described in turn.

WF and CD. As in previous studies (e.g. Jones, et al., 2012; Johns, et al., 2012), all models were compared against a word frequency (WF) and contextual diversity (CD; Adelman, et al., 2006) baseline. A WF count computes the number of occurrences of a word across the entire corpus. A CD count computes the number of different contexts that a word occurs in. In a standard CD count, context is typically operationalized as a paragraph (Adelman, et al., 2006; Jones, et al., 2012) or a moving window (Johns, et al., 2016b). Due to the difficulty of parsing paragraphs within electronic books, a moving window of 20-sentences is used here. Historically, operationalisations of context differ greatly across studies. For example, some studies have defined context as the list a word is contained in, to changes in time, or the room in which learning took place (Schmidt, 1991; Verkoeijen, Rikers, & Schmidt, 2004; Wickens,1987).

Thus, the definition of context is likely model-dependent (see Jones, et al., 2017 for a review of context effects in language and memory).

SDM. A variation of the SDM (Jones, et al., 2012) will be included in the analysis. The variation of the model is entitled the vector-space SDM (vSDM), first described by Johns, Dye, & Jones (2014), and empirically validated in Johns, Dye, & Jones (2016). In the original SDM a given word's strength in memory is represented in a Word x Document matrix, which can be trained over any large corpus of documents. For each new document that is encountered, a new column is added to the matrix. If a word occurred in that document, then it is assigned a semantic distinctiveness (SD) value in that column, which signals how redundant the new context is, compared to past experience (for details about how this is calculated, see Jones, et al., 2012). A word's strength in memory (corresponding to how easily a word is to retrieve from memory) is then just the summed semantic distinctiveness values across its row in the matrix.

While this implementation of the model has proven successful, it is also difficult to scale due to the computational resources required for an ever-expanding matrix. A more tractable approach is to use a set vector size, similar to that employed in vector accumulation models, such as the BEAGLE model of semantics (Jones & Mewhort, 2007). The vSDM model uses this architecture, which is based on the same mechanisms as the original model, but is less computationally expensive. Since the vSDM does not involve an increasing amount of computation as more documents are processed, it can be run over much larger corpora, which is necessary given the amount of language materials contained in Table 1.

The fundamental operation of the SDM is that it utilizes an expectancy-congruency mechanism to build a word's semantic representation: The encoding strength for a word in a given context is relative to the information overlap between the context and the memorial representation of the word. This mechanism is very similar in principle to models that adjust attention across learning to dimensions that are more diagnostic. The vSDM will have underlying differences in terms of implementation, but this mechanism is the basic assumption about the importance of contextual variability.

In the vSDM, each word is represented by an initially empty distributed vector, representing the meaning of a word. When a word occurs in a context, that word's vector is updated. How strongly that word is updated depends on the similarity between context and the

word's representation. To encode the representation of a context, the memory vectors of each word in the context are summed:

$$\mathbf{Context} = \sum_{i=1}^{n} \mathbf{M}_i \qquad (1)$$

Where $\mathbf{M}_i$ is the memory vector for word *i*, and **Context** is the vector representing the meaning of the current context.

This context vector allows us to assess the similarity between the current document context and a word's semantic representation. As in the original SDM model, we use the vector cosine (a normalized dot product) as our similarity metric, and subject it to an exponential transformation, so as to properly weight redundant contexts over distinctive ones. Semantic distinctiveness for a given word is thus computed as a function of:

$$SD_i = e^{-\lambda * \cos(\boldsymbol{M_i}, \boldsymbol{Context})} \qquad (2)$$

Where λ is a free parameter that scales the differences between high and low similarity contexts[5].

In the original SDM, each new context is encoded as a new column in a Term x Document matrix. To replicate this process, in the vSDM a random Gaussian vector[6] is generated and added into each of the word's memory vectors that occurred in the context. This is meant to be analogue to the original SDM model's use of a new context being encoded as a new column. Only the words that occur in the context are update with context vector. The strength with which the memory vector is updated is modulated by the SD value:

$$\mathbf{M_i} = \mathbf{M}_i + (\mathbf{RG} * SD_i) \qquad (3)$$

Where **RG** is the randomly generated Gaussian vector and SD_i is the SD value for word *i* in that context. Words that do not occur in the context are not updated.

In this version of the model, the strength of a word is an external counter that accumulates the SD values of a word across its occurrences. As with the WF and CD variables, the SD counts will be reduced with a natural logarithm.

BP and AP. The book prevalence measure (BP) is similar to the CD measure, but defines context at the level of the book, rather than at the level of the paragraph (or paragraph-sized chunk). Specifically, BP computes the number of different books that a word occurs in. The

[5] Consistent with past simulations (e.g. Jones, et al., 2012) the λ parameter is set at 5.

[6] Each value in the vector has a mean of 0.0 and a standard deviation of $1/\sqrt{N}$ where N=2,048.

words *shine* and *plasma* offer a simple demonstration of how the CD and BP measures diverge. While both words occur in a similar number of paragraph contexts, *shine* occurs in nearly four times as many books as *plasma*. The two models thus make markedly different predictions of the lexical strength of these words, with CD yielding an even count, and BP yielding a notable asymmetry.

The author prevalence (AP) measure builds upon the BP measure by counting only if a specified author used a word or not. The AP measure will provide a proxy of how widely used a certain word is across authors, and thus how likely it is to be an active member of the language environment, such that if a word is used by a wide variety of authors, it is likely that the general population would have experienced it[7]. In some ways, the AP measure could be conceptualized as being an account of an author's productive vocabulary, as some authors have millions of tokens contained in their book collection.

To gain an understanding of the scale that the BP and AP model are operating at, Figure 1 contains a histogram of the number of unique words that each book and author use. Most books contain between 2,500 and 15,000 unique words, with an average of 5,986 unique words per book. Most authors use between 5,000 and 30,000 unique words across their writings, with an average of 14,213 unique words per author. In contrast to this, using a window of 20 sentences to calculate the CD and vSDM variables, there is an average of 136 unique words contained in each window. That is, the BP and AP measures are operating at a much greater unit of word occurrence than past measures.

SD-BP and SD-AP. To justify applying the semantic transformations described by the SDM to the BP and AP, it is worth considering recent work in distributional semantics examining language at both the book and author level. Specifically, in Johns and Jamieson (2018), a sample of fiction books was organized by author and genre. There was a small genre effect, where authors who wrote in the same genre had a small increase in the similarity of their writings when compared to authors who wrote in different genres. However, the biggest difference emerged at the individual-author level: each author had a unique signature of language usage.

[7] We do not mean to insinuate that we believe that individual subjects would have read these exact books or authors, but instead that if a word is used almost universally across authors, even if with low frequency, it is more likely that an individual would have experienced that word.

However, Johns & Jamieson (2018) used a much smaller set of books than is used here (they used a collection of 1,850 books). To ensure that the genre- and author- signature effect is replicated with the larger set of books described in this article, the same methodology used to assess book similarity by Johns and Jamieson (2018) was used here. Specifically, to measure the similarity of the book set used here, we (a) identified the 100,000 highest-frequency words across the corpus, (b) constructed a vector for each book which recorded the number of times that each of the 100,000 highest-frequency words appeared in a book, (d) converted all word frequencies to their log equivalents (i.e., $n' = \ln(n + 1)$, where n is the count from the book), (e) computed the cosine similarity of each author's word vector to every other author's word vector. All similarity values were then organized into one of three groups: (a) within author (i.e., similarity of books written by the same author), (b) within genre (i.e., similarity of books from the same genre), and (c) across genres (i.e., similarity of books from different genres). The within-author distribution was composed of over 280,000 comparisons, the within-genre distribution was composed of over 42 million comparisons, and the genre distribution was composed of over 312 million comparisons.

Figure 2 displays the results of this simulation, which shows that this book set shows an identical pattern to what was found by Johns and Jamieson (2018): there is a small positive shift in similarity for books written in the same genre (relative to books written in different genres), but there is a much larger positive shift for books written by the same author. Johns and Jamieson (2018; in press) refer to this as the author signature effect. This finding demonstrates that there is semantic variance at both the book and author level, which suggests that applying the SDM transformations to the BP and AP count will likely increase the performance of that count, similar to how the SDM improves upon a CD count.

To accomplish this, a very similar model described in equations 1-3 will be used. However, instead of forming a context representation by summing Gaussian representations, the context representation will be the frequency distribution of a book (in the case of the SD-BP variable) or all of the books written by a single author (in the case of the SD-AP variable). A word's representation will be incrementally constructed by summing the distribution of each book or author into a word's memory representation, added in according to the strength of the SD signal. The SD value for a word is calculated by taking the cosine similarity between the

context representation and the word's representation. However, unlike the SDM and vSDM, it was found for both the SD-BP and SD-AP a simple linear transformation performed best:

$$SD = 1 - cosine(\mathbf{M_i}, \mathbf{context}) \qquad (4)$$

Where $\mathbf{M_i}$ is the memory representation for word *i*.

To conceptualize the results of the SD transformations on the BP and AP variables, it is worth considering the nature of lexical experience and the content of books and the writings of different authors. As Figure 2 shows, the average book is quite similar to each other (i.e., the average cosine across books is above 0.7). This suggests that there is much semantic redundancy across books. By weighting each book (or each author's writings) by how unique that writing is (compared to past experience), it allows for books with a large amount of overlap to be reduced in importance. Past research (e.g., Jones, et al., 2012, 2016) has shown that this type of operation is important in lexical organization when applied to smaller units. Here we will determine if the operations also work at large units of language as well.

Training methodology. The vSDM, SD-BP, and SD-AP models are sensitive to the word list used in training, as it impacts the context representation that is formed. The word list used to train these models are the 81,276 unique words contained in the prevalence norms of Brysbaert, Mandera, McCormick, & Keuleers (2019), the English lexicon project (Balota, et al., 2007), and the British lexicon project (Keuleers, Lacey, Rastle, & Brysbaert, 2012). Of those words, 78,033 occurred at least once in the corpus, demonstrating the diversity of the language that is contained in the book collection assembled here. The CD and vSDM variables will be derived from a moving window of 20 sentences across the entire corpus, similar to past research (e.g., Johns, Sheppard, Jones, & Taler, 2016).

Data

The two main data sources that will be used to assess the above described variables will be the recently released word prevalence (WP) data of Brysbeart, et al. (2019) and lexical decision time (LDT) data from two lexicon projects – the English lexicon project (ELP; Balota, et al., 2007) and the British lexicon project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012). Additionally, naming time (NT) will be analyzed from the ELP. For both the ELP and BLP, the lexical decision time that will be used is the z-transformed reaction times. Only words that had occurred one or more times in the corpus were included in the following analyses. We did not set an

accuracy threshold as we are equally interested in explaining lexical decision and naming accuracy data as we are in explaining lexical decision and naming reaction time.

The advantages of using books to derive lexical statistics is that they contain a greater amount of low frequency words compared with subtitles (Brysbaert & New, 2009) or social media (Herdağdelen & Marelli, 2017), enabling a more complete analysis of the data space. For example, the social media norms of Herdağdelen and Marelli (2017) contains roughly 24,883 of the 61,855 words from the Brysbaert, et al. (2019) WP norms, while the book norms derived here contains 58,711 of the words from these norms. This also makes it difficult to compare performance of the different frequency values.

Results

As a first pass at understanding the differences and similarities of the seven lexical variables, Table 4 contains the pairwise correlations of the variables. The first important aspect of this analysis is to notice that all variables are fairly redundant, similar to past work on contextual and semantic diversity (e.g., Adelman, et al., 2006; Jones, et al., 2012). Similar to these past studies, it will be necessary to use regression analyses to separate the unique contributions of these variables to the datasets analyzed here. However, it is clear from these correlations that the occurrence-based variables cluster together, as do the prevalence-based variables. This suggests that these two classes are assessing relatively different types of lexical information.

To examine the fit of the different lexical variables to the different lexical databases, Table 5 contains the correlation between the seven lexical variables and the seven different lexical behaviors (WP, ELP lexical decision time, ELP lexical decision accuracy, ELP naming time, ELP naming accuracy, BLP lexical decision time, and BLP lexical decision accuracy). For the occurrence-based variables, the results are consistent with past results (e.g., Jones, et al., 2012) where the vSDM has a higher correlation to all datatypes, compared to the WF and CD variables. For the prevalence-based variables, this table shows that applying the semantic diversity transformations to the BP and AP count substantially increases the fit of these variables to all lexical behaviors.

Importantly, there are differences in the fits across WP, lexical decision time, naming time, and lexical decision accuracy. Specifically, the prevalence-based provide a very substantial increase in the fit to the WP data, compared to the occurrence-based variables. This is also the case for lexical decision and naming accuracy, where the prevalence-based variables

substantially outperform the occurrence-based data. This is not the case for lexical decision time, where the best performing occurrence-based variable (vSDM) is roughly the same as the best performing prevalence-based variable (SD-BP). For naming time, the best overall predictor is the SD-AP variable.

In line with previous research (e.g. Adelman, et al., 2006; Brysbaert & New, 2009; Jones, et al., 2012), regression analyses were conducted to isolate the unique contribution of each variable to lexical processing. The analysis we conducted is standard and provides a measure of the predictive gain (i.e., measured as percent ΔR^2 improvement) for one predictor over another competing predictor (see Adelman, Brown, & Quesada, 2006; Johns, Sheppard, Jones, & Taler, 2016). For the ELP and BLP, polynomial regression was used, such that each variable had two predictions: log(variable) + $\log^2$(variable). The inclusion of the square of the logarithm did not impact a variable's fit to the WP data, and so was not included when analyzing this data.

Given that there is likely little unique variance to explain across seven redundant lexical variables, the analysis was simplified by removing three variables: CD, BP, and AP. These were removed because the vSDM, SD-BP, and SD-AP are semantically weighted versions of these variables, respectively, and thus have overlapping theoretical conceptualizations. To justify the exclusion of these variables, a regression was done calculating the unique variance accounted for by the count variable (i.e., CD, BP, AP) over the SD-transformed variable (i.e., SD, SD-BP, SD-AP), and vice versa. Figure 3 contains the results of this analysis, and shows that the SD transformed variable explain more variance than the count variable for each dataset.

Table 6 contains the results of the regression analysis for the four variables (WF, SD, SD-BP, and SD-AP). This table shows that there is high agreement across all datatypes, where the SD-AP variable accounts for the most variance, the SD-BP variable accounts for some, while the WF and SD variable accounts for very little. This is especially true for the WP and lexical decision and naming accuracy, with smaller effects for the lexical decision and naming reaction time data. The finding that the SD-BP variable still accounts for large amounts of variance in some datasets suggests that there are differences in the semantic content of books versus the writings of an individual author.

Split by country of birth. As table 2 shows, the book collection used here has a large number of authors from both the USA and UK. The WP data from Brysbaert, et al. (2019) contains word prevalence measures of subjects from the USA and UK. Additionally, the ELP

was collected on subjects from the USA, while the BLP was collected from subjects in the UK. Thus, it is possible to determine whether the lexical statistics derived from the writings of an author born in the same country as where the data was collected provides a better accounting for that data. Recent research by Johns, Jones, and Mewhort (in press) suggests that lexical behavior is strongly influenced by differential experience with language, thus splitting a corpus by place of birth offers another test of this hypothesis, and this hypothesis was validated by Johns and Jamieson (in press).

Table 7 contains the correlation between the WF and SD-AP variables, derived from a corpus of either USA or UK authors, to the different datasets (the WP data was split into USA or UK data). Although only the WF and SD-AP variables was included in this table, all variables showed very similar trends. As a comparison, the SD-AP variable trained on the entire corpus was included in this table. For the data collected in the USA (the USA_WP and ELP data), the SD-AP variable trained on the corpus of authors born in the USA (USA_SD-AP) offered the best fit across each datatype. In comparison, the variables derived from the authors born in the UK provide a poor accounting for the USA data, suggesting that the lexical statistics of these writings do not map onto the lexical experience of American subjects. Indeed, the USA_SD-AP exceeds the correlation of the SD-AP variable, which was trained on a considerably larger amount of language.

The UK_SD-AP provides a solid fit to the data collected in the UK (the UK_WP and BLP data). However, it does not provide as large an advantage as is seen in the USA data for variables derived from the writings of Americans. Indeed, for the UK_WP and the BLP lexical decision accuracy data, the UK_SD-AP actually provides a worse fit than both the USA_SD-AP and the SD-AP.

Overall, the trends in the data contained in Table 7 suggests that subjects in the USA are relatively unfamiliar with the language used by British authors (in comparison to the lexical statistics derived from American authors), hence leading to a poor fit between the lexical statistics contained in the UK_WF and UK_SD-AP variables to the data collected in the USA. However, the UK subjects seem to have equal amounts of familiarity to both American and British authors, leading to neither corpus offering a large advantage. Similar results were found in Johns and Jamieson (in press) for word familiarity and category production data, suggesting a general trend. However, an alternative possibility for this trend is that there are other types of

lexical experience that better account for the lexical experience of people from the UK, which do not seem to be captured with books, a question for future research.

To get a better understanding of the unique variance that each variable is accounting for, another regression analysis was conducted. The top panel of Figure 4 displays the amount of unique variance that the USA_SD-AP accounts for over the UK_SD-AP (and vice-versa) across the eight data sources. This figure shows that the USA_SD-AP variable accounts for the most variance across seven of the eight datatypes, with the only exception being BLP lexical decision reaction times. The advantage for the USA_SD-AP is quite striking for the data collected in the USA, where it offers a considerable advantage over the UK_SD-AP variable. This advantage was smaller for the UK_WP and BLP lexical decision accuracy data.

To determine whether the variables derived from the country-specific corpus offers an advantage over the SD-AP variable that was trained across all of the writings contained in the book collection, an additional regression was done where the amount of unique variance that the SD-AP variable trained on a country-specific corpus (USA_SD-AP or UK_SD-AP) over the SD-AP variable trained on the entire book collection was calculated, and vice-versa. The results of this analysis is contained in the bottom panel of Figure 4. For the UK data, the SD-AP variable offers the best fit to all datatypes. However, for the USA data, the USA_SD-AP variable accounts for the most unique variance, suggesting that including authors from different countries actually harms the fit of lexical variables to the USA data.

One possibility not yet discussed is that the differences in corpus size may be driving the differences seen in Table 7 and Figure 5. As can be seen in Table 2, the book collection contains almost three times more American authors than British authors. To test what effect this differential levels of information is having on the fits to the WP data, a Monte Carlo simulation was done. In this simulation, 700 authors who were born in the USA and 700 authors born in the UK were randomly selected. UK_AP and USA_AP variables were then computed from these randomly selected authors. The AP measure was chosen because it is computationally inexpensive compared to the SD-AP measure, but still offers a good fit to this data. Given that there are only 738 authors from the UK, there is not likely to be much variance in the computed statistics for the UK_AP variable, but there should be for the authors from the USA_AP variable. Resampling was done 5,000 times and the average correlation to the USA_WP and UK_WP data was computed.

The results of this simulation is contained in Figure 5, which displays the USA_AP and UK_AP fits to the USA_WP and UK_WP data when the measures are either sampled or computed from all of the authors from a certain country. This figure shows that there is a small drop in the correlation when only 700 authors are used, but there is not a change in trends. The USA data is still well accounted for by the AP variable when trained on the USA corpus, and still has a poor fit the AP variable when it is trained on the UK corpus. The UK data is still well accounted for by either corpus. This simulation demonstrates that the different patterns in lexical statistics are causing the discrepant fits to the lexical behavior, and not the overall size of the different corpora.

Gender split. The word prevalence data of Brysbaert, et al. (2019) includes data from both male and female subjects. Given that the book collection used here has also been categorized by male and female authors (see Table 3), the goal of the final analyses is to determine whether gender-specific corpora offer an advantage in accounting for lexical behavior.

The correlations between the Female_WF, Female_SD-AP, Male_WF, and Male_SD-AP variables, to the Female_WP and Male_WP data is contained in Table 8. Again, the SD-AP variable trained on all materials was included in the table to serve as a comparison. Table 7 shows that, overall, the data collected from female subjects are better accounted for by the lexical variables. Additionally, the female data is best accounted for by the lexical variables trained on the corpus of female authors, compared to the lexical variables trained on the corpus of male authors. For the data collected from male subjects, only the Male_SD-AP variable shows a slightly higher correlation to the Male_WP data, compared to the Female_SD-AP variable. However, the highest correlation to both the Female_WP and Male_WP data is from the SD-AP trained on all materials, suggesting that gender-specific corpora do not provide an overall advantage in accounting for gender-specific data.

To determine how much unique variance each variable accounts for, two more regression analyses were done. The top panel of Figure 6 displays the amount of unique variance the Female_SD-AP variable accounts for over the Male_SD-AP variable (and vice-versa) for the Male_WP and Female_WP data. This figure shows that for the Female_WP data the Female_SD-AP accounts for the most variance, while the Male_SD-AP variable accounted for very little. The opposite was also true for the Male_WP, although the Female_SD-AP still accounted for some variance in this data. The bottom panel of Figure 6 contains the amount of

unique variance that the gender-specific SD-AP variables accounts for over the SD-AP variable computed on the entire corpus. This figure shows that the SD-AP trained on all materials accounts for more variance in the data than the gender-specific corpus, demonstrating that there is no overall advantage to using gender-specific corpora.

The results of this analysis suggests that there are gender differences in the usage of language (at least at a very large scale; see Johns & Dye, 2019 for a more specific case of gender differences in language usage), and that there are some differences in lexical behavior that are reflective of the gender discrepancy, such that a variable trained on a corpus of female authors outperforms a variable trained only on male authors, for female subjects. However, the best fitting model is still the one trained on all authors, suggesting that people have a mix of experience with the writings of both female and male authors.

Supplementary Materials

The supplementary materials to this article contains the seven lexical variables (WF, CD, vSDM, BP, AP, SD-BP, and SD-AP) trained on both the complete book collection and also splits on place of birth (authors from the USA and UK) and gender (male and female authors). It is our hope that these variables can be of use to other researchers examining lexical processing. The word list for the materials contains the lemmas from the Brysbaert, et al. (2019) study, as well as inflected forms from the English Lexicon Project (Balota, et al., 2007) and the British Lexicon Project (Keuleers, et al., 2012).

General Discussion

Subjective ratings (e.g., AoA, familiarity, concreteness, meaningfulness, etc.) give excellent predictions of human behavioral data, and encourage the field to improve focus on valid psychological constructs. But in addition to being good predictors, they are ultimately dependent variables that need to be themselves explained, and provide a challenging target for mechanistic explanations of learning and processing (Baayen et al., 2016; Gernsbacher, 1984; Jones et al., 2015; Westbury, et al., 2013; Recchia & Jones, 2012). A necessary intermediate step to explanation is the ability to link subjective ratings to objective environmental statistics.

This article introduced four novel measures of lexical strength, derived from much larger units of linguistic context than have previously been reported in the literature. Specifically, instead of relying on paragraph or document contexts (Adelman, et al., 2006; Jones, et al., 2012), whole books and the combined books of an individual author were used as the basic unit of

measurement. When applied to various lexical behaviors, book prevalence (BP; a measure of the number of books a word occurred in) and author prevalence (AP; a measure of the number of authors that used a word) provided a better fit than occurrence-based variables (WF, CD, and SD). When the AP and BP measure was then combined with the machinery of the semantic distinctiveness model (SDM; Jones, et al., 2012, 2017; Johns, et al., 2012, 2016a, b), the resulting SD-BP and SD-AP variables were found to account for the most unique variance across multiple datasets, with the SD-AP being the overall best predictor.

The results reported here were inspired by the large-scale behavioral collection efforts of Keuleers et al. (2015) and Brysbaert et al. (2016, 2019), who used a crowd-sourcing methodology to estimate the proportion of the population that was familiar with a given word. The resulting 'word prevalence' measure explained significant unique variance in lexical decision performance, and revealed considerable variability in lexical knowledge across the population. We applaud their work as a huge step in the right direction; however, a potential limitation of that work is that it seeks to explain behavior in terms of similar behavioral data, which raises a host of thorny theoretical issues (Jones, et al., 2015). To address this problem, we sought to derive a similar measure directly from large-scale natural language materials, which lent the measure the added benefits of being objective and readily interpretable in a modeling context.

In most other respects, however, Keuleers et al.'s (2015) word prevalence measure and our environmentally-derived BP, AP, SD-BP and SD-AP measures are complementary to one another. Measuring the probability that a word will occur in a particular book or be used by a particular author is not so different from measuring the probability that a word will be known to a particular speaker—words used across all discourse topics are more likely to be encountered, and thus more likely to be known by a larger proportion of speakers. Both measures allow for better estimation of how widely used a word is in the language, and the relationship between measures of prevalence and measures of frequency sets important constraints on theory construction, posing fresh challenges for language theorists.

The findings of this article suggest that by assessing the occurrence of words at quite high levels of measurement (e.g. whether it occurs in a book or whether an author used a word) provides important insight into how likely it is that a person had experienced that word before. This is borne out by the fact that the prevalence-based metrics offer a very significant

improvement to lexical decision and naming accuracy data, and word prevalence data. Likewise, occurrence-based metrics may over inflate the likelihood of a person knowing that a word is a word, as a word may have a relatively high word frequency, but if that word is only used by a couple of authors, it is unlikely that many would have experienced that word before.

One advantage of using books as a source of lexical information is that it allows for control and understanding of the materials that are being used. In this article, we used author place of birth and gender to try to understand the influence of culture and gender on word processing. This follows previous work (e.g., Johns, Jones, & Mewhort, 2019; Johns & Jamieson, in press) in trying to understand the effects of differential experience on language processing. In this article, we found that using a country-specific corpus allows for a better accounting of behavioral data collected in the USA, such that a corpus derived from only American authors provided a better fit to this data than a corpus derived from the complete book set, similar to past work on the SUBTLEX corpus (van Heuven, et al., 2014). Unique effects were also found for subjects from the UK, and also to male and female data. Additionally, some effects of author gender were found, with a female author corpus providing a superior fit to word prevalence data collected from female subjects, compared to a male author corpus. However, the overall corpus provided the best fit to the female subjects. These findings suggest that the fits of models are sensitive to the composition of the corpus that was used for training, and that by tailoring corpora to individual sets of data (depending on the subjects that were collected from), sometimes better and more insightful models can be developed.

However, even though books were the main unit of measurement used in the current study, these norms can likely be modulated with other lexical information. The language contained in books likely differs significantly from other sources of language, such as subtitles, spoken language corpora, or social media posts. The issue becomes a question of scale of context – books provide a simple method of looking at word occurrence at quite large levels. Other sources are not so easily segmented, so determining the correct method of parsing other corpus types to estimate word prevalence is a question for future research.

This work highlights the need for the continued evolution of corpus-driven analyses of lexical behavior. Different collections of language contain different information, and have a different probability of being encountered by members of the language-speaking population. Determining the success of language materials in predicting performance, and establishing their

connection to linguistic experience, will facilitate the development of better measures of verbal behavior, and advance our understanding of lexical organization in memory.

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(9), 814–823.

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

Allen, L. R., & Garton, R. F. (1968). The influence of word-knowledge on the word-frequency effect in recognition memory. *Psychonomic Science*, 10, 401-402.

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55, 290-313.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30, 1174-1220.

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.

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, 441.

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.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51, 467-479.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, 12, 263-282.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of experimental psychology: General*, *113*(2), 256.

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, 718-730.

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, 103, 114-126.

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

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

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, 65, 486-518.

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

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.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2017). Small worlds and big data: Examining the simplification assumption in cognitive modeling. In Jones, M. N. (Ed.) *Big Data in Cognitive Science: From Methods to Insights*, Taylor & Francis.

Johns, B. T. & Dye, M. (2019). Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods*, 51, 1601-1618.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, 26, 103-126.

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

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

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.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, & Griffiths. *Psychological Review*, 122, 570-574.

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*.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44, 287-304.

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, 1665-1692.

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

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.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76, 165.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6, 5-42.

Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, 6, 315:1-16.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, 19, 523–542.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, 58, 140-159.

Westbury, C. (2016). Pay no attention to that man behind the curtain. *The Mental Lexicon*, *11*, 350-374.

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67, 1176-1190.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 30,796 – 800.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. Journal of Memory and Language, 60(4), 502-529.

Wickens, D. D. (1987). The dual meanings of context: Implications for research, theory, and applications. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and learning: The Ebbinghaus Centennial Conference* (pp. 135–152). Hillsdale, NJ: Erlbaum.

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, 4, 991.

Table 1.
Characteristics of books across genres.

Genre	# of Authors	# of Books	Words per Book	Total # of Words
Christian	51	410	65,239	26,747,990
Crime	82	643	73,841	47,479,763
Fantasy	264	2,231	98,972	220,806,532
Historical Fiction	198	1,312	108,995	143,001,440
Horror	75	728	86,202	62,755,056
Literature	357	2,740	88,729	243,117,460
Mystery	321	2,505	72,928	182,684,640
Non-fiction	248	1,189	105,745	125,730,805
Romance	615	5,004	74,835	374,474,340
Science Fiction	455	5,336	75,225	401,400,600
Thriller	169	1,245	99,249	123,565,005
Western	47	498	59,967	29,863,566
Young Adult	325	2,785	40,066	111,586,014
Total/Average	3,207	26,626	80,769	2,093,213,211

Table 2.
Characteristics of books organized by author country of birth.

Country	# of Authors	# of Books	Words per Book	Total # of Words
USA	2,000	17,130	77,792	1,332,576,960
UK	738	6,171	81,873	505,238,283
Australia	136	865	73,837	63,869,005
Canada	114	795	77,014	61,226,130
Other/Unknown	222	1,692	78,121	132,180,732

Table 3.
Characteristics of books organized by author gender.

Gender	# of Authors	# of Books	Words per Book	Total # of Words
Male	1,514	13,651	81,408	1,111,300,608
Female	1,696	13,002	75,665	983,796,330

Table 4.
Correlations between the lexical variables.

Measure	2	3	4	5	6	7
1. WF	.998	.994	.977	.961	.961	.95
2. CD	.	.996	.975	.959	.961	.949
3. vSDM	.	.	.979	.967	.959	.951
4. BP	.	.	.	.994	.977	.964
5. AP	.	.	.	.	.958	.964
6. SD-BP	.	.	.	.	.	.995
7. SD-AP	.	.	.	.	.	.

Note. N=78,033; all correlations significant at the $p<0.001$ level.

Table 5.
Correlations between the lexical behaviors and lexical variables.

	WF	CD	vSDM	BP	AP	SD-BP	SD-AP
WP	.656	.665	.669	.7	.703	.709	.725
ELP_LDT	-.674	-.673	-.676	-.664	-.651	-.677	-.675
ELP_LDT_Acc	.447	.453	.459	.472	.472	.483	.494
ELP_NT	-.582	-.582	-.586	-.581	-.578	-.59	-.594
ELP_NT_Acc	.372	.376	.382	.394	.401	.398	.412
BLP_LDT	-.61	-.614	-.615	-.62	-.6	-.639	-.633
BLP_LDT_Acc	.564	.575	.587	.635	.651	.639	.662

Note. N=58,711 for WP data; N=40,306 for ELP data; N=28,710 for BLP data; all correlations are significant at the $p<0.001$ level.

Table 6.
Unique effects of WF, SD, SD-BP and SD-AP in % change in ΔR^2.

	WF	SD	SD-BP	SD-AP
WP	0.07	0.0 *n.s.*	1.015	4.58
ELP_LDT	1.147	0.532	1.509	3.24
ELP_LDT_Acc	0.0 *n.s.*	0.0 *n.s.*	2.512	6.689
ELP_NT	0.815	0.291	2.854	5.311
ELP_NT_Acc	0.0 *n.s.*	0.0 *n.s.*	6.56	12.165
BLP_LDT	0.41	0.63	.194	1.856
BLP_LDT_Acc	0.089	0.076	.955	5.557

Note. N=58,711 for WP data; N=40,306 for ELP data; N=21,911 for BLP data; all correlations are significant at the $p<0.001$ level unless otherwise noted.

Table 7.
Correlations of variables derived from authors born in either USA or UK to lexical behaviors.

	USA_WF	UK_WF	USA_SD-AP	UK_SD-AP	SD-AP
USA_WP	.639	.584	.711	.643	.701
UK_WP	.622	.615	.69	.672	.697
ELP_LDT	-.659	-.649	-.678	-.635	-.668
ELP_LDT_ACC	.434	0.426	.497	.45	.488
ELP_NT	-.567	-.56	-.597	-.55	-.585
ELP_NT_ACC	.361	.356	.42	.366	.406
BLP_LDT	-.601	-.596	-.628	-.637	-.633
BLP_LDT_ACC	0.564	.548	.652	.638	.662

Note. N=52,734 for WP data; N=39,766 for ELP data; N=28,729 for BLP data; all correlations significant at the $p<0.001$ level.

Table 8.
Correlations of variables derived from male and female authors to male and female WP data.

	Female_WF	Male_WF	Female_SD-AP	Male_SD-AP	SD-AP
Female_WP	.643	.621	.706	.695	.725
Male_WP	.619	.612	.68	.682	.705

Note. N=53,247; all correlations significant at the $p<0.001$ level.

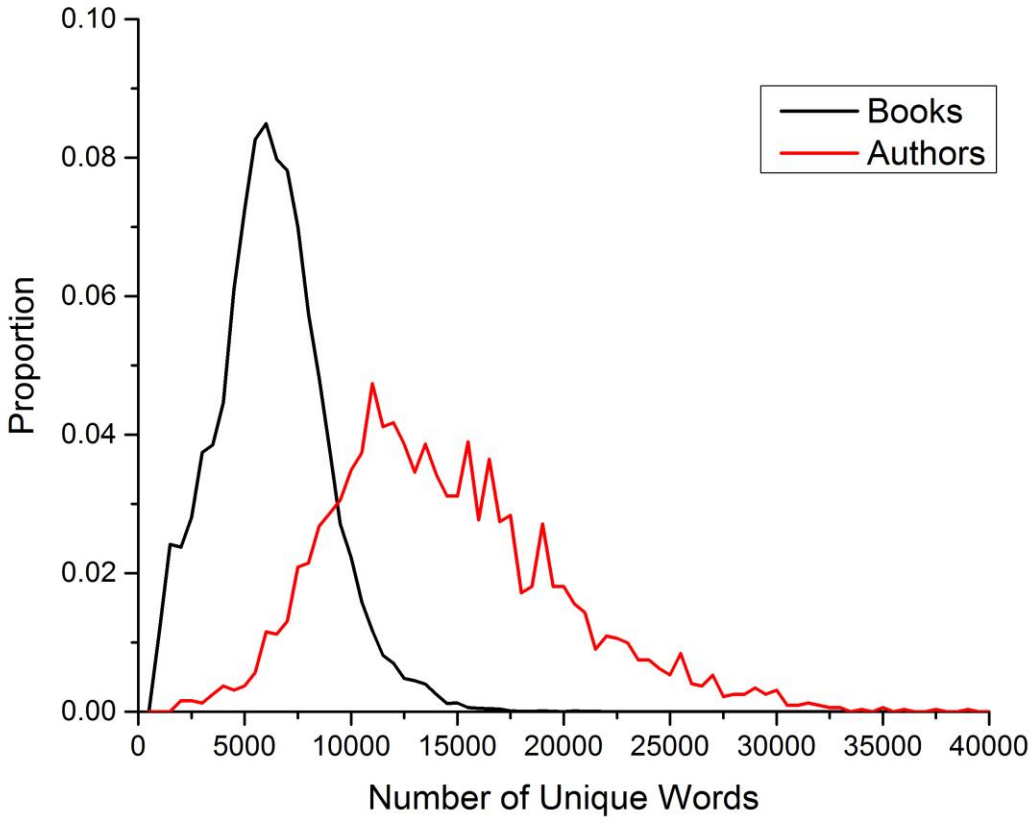

Figure 1. Histogram of the number of unique words that each book contains and each author uses.

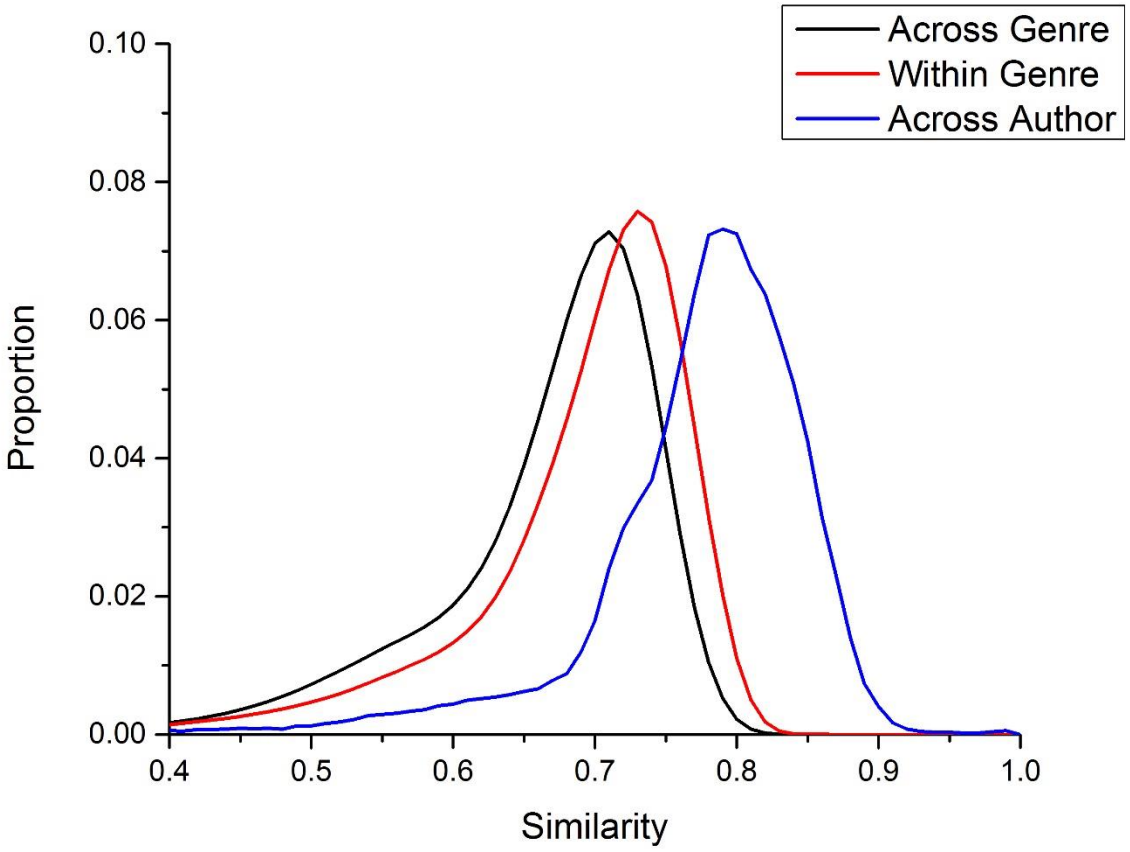

Figure 2. Similarity distributions across three levels: 1) books written by different authors across genres (black line), 2) books written by different authors within a genre (red line), and 3) books written by the same author (blue line). Similarity is the vector cosine between the frequency distribution of two books. This simulation replicates the findings of Johns and Jamieson (2018) with a much larger book set.

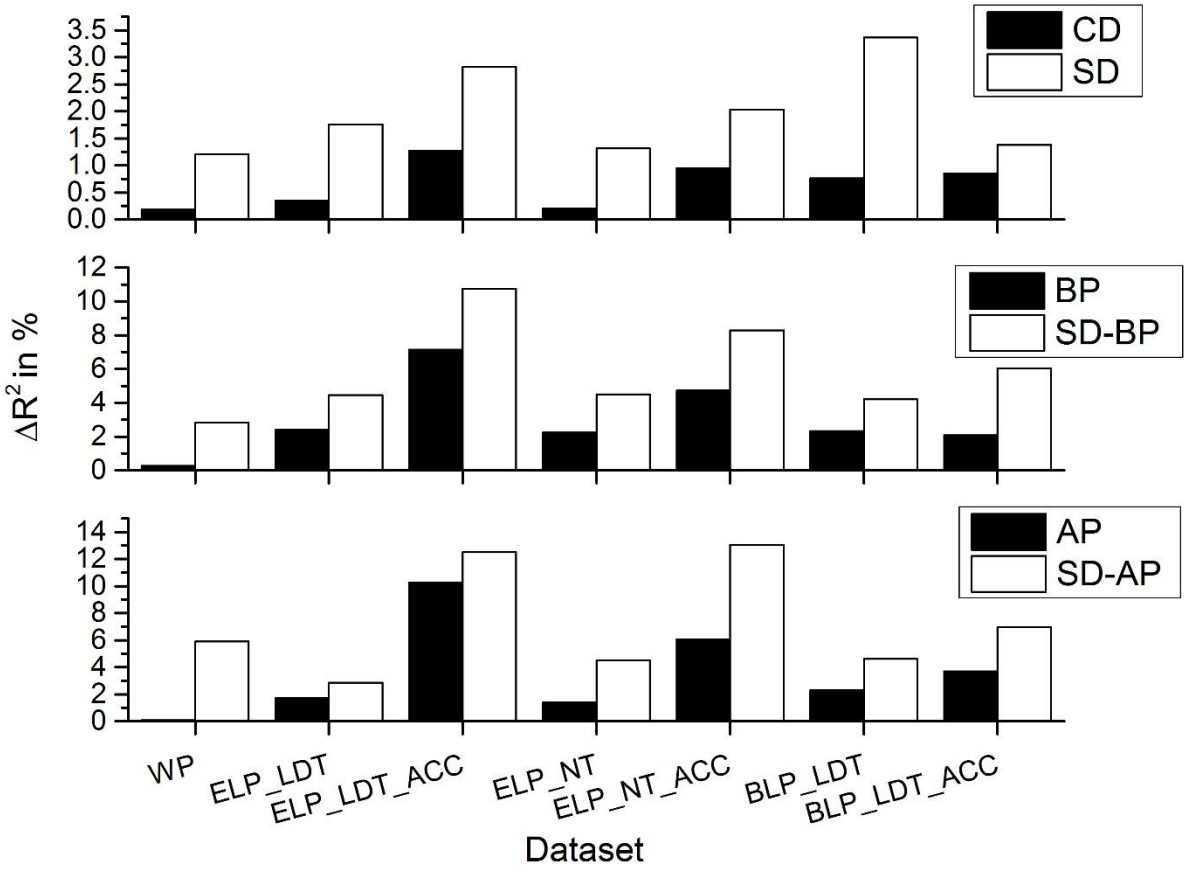

Figure 3. Results of a regression analysis demonstrating that the SD transformed variables accounts for more unique variance than the count variables for each set of data and across multiple levels, including counting in paragraphs (SD vs. CD), books (BP vs. SD-BP), and at the author level (AP vs. SD-AP).

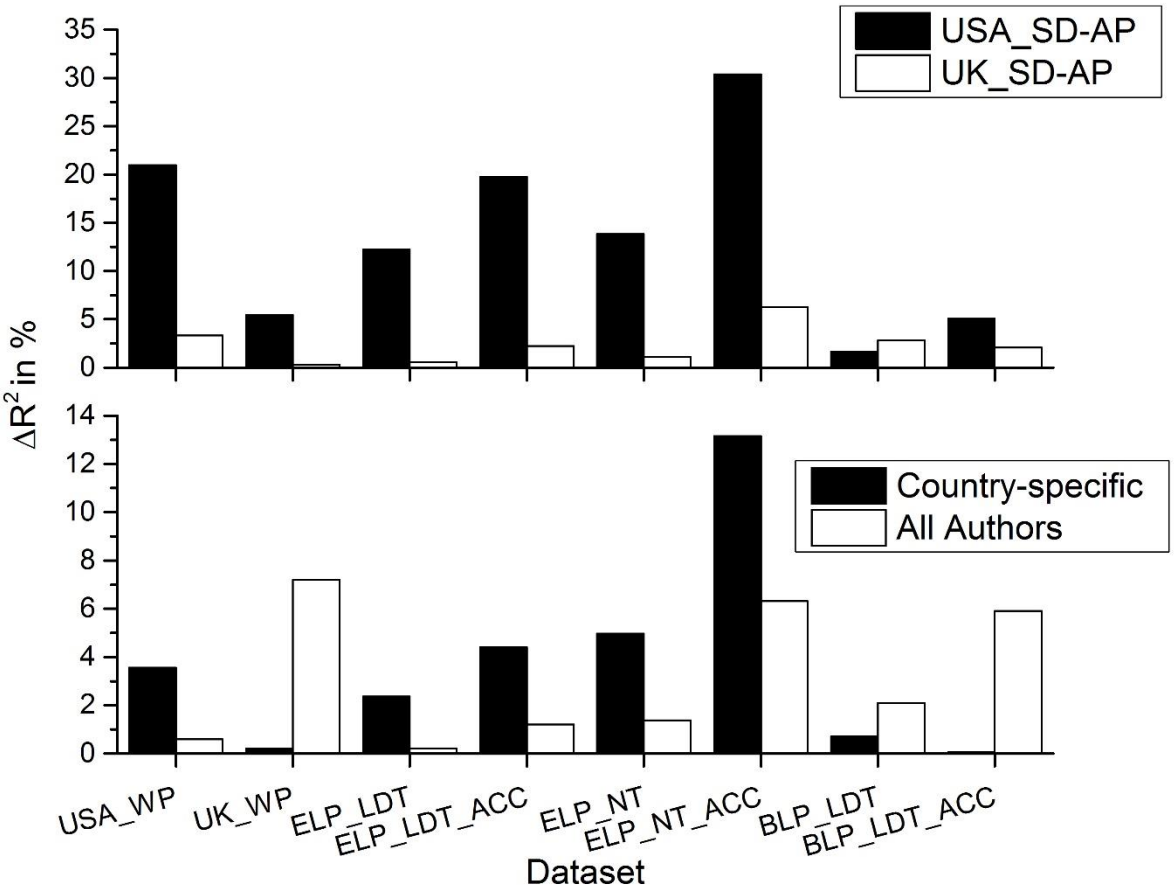

Figure 4. Results of regression analyses testing the amount of unique variance that the SD-AP accounts for when trained on authors from the UK versus authors from the USA (top panel), and when trained on a country-specific corpus versus all available authors (bottom panel).

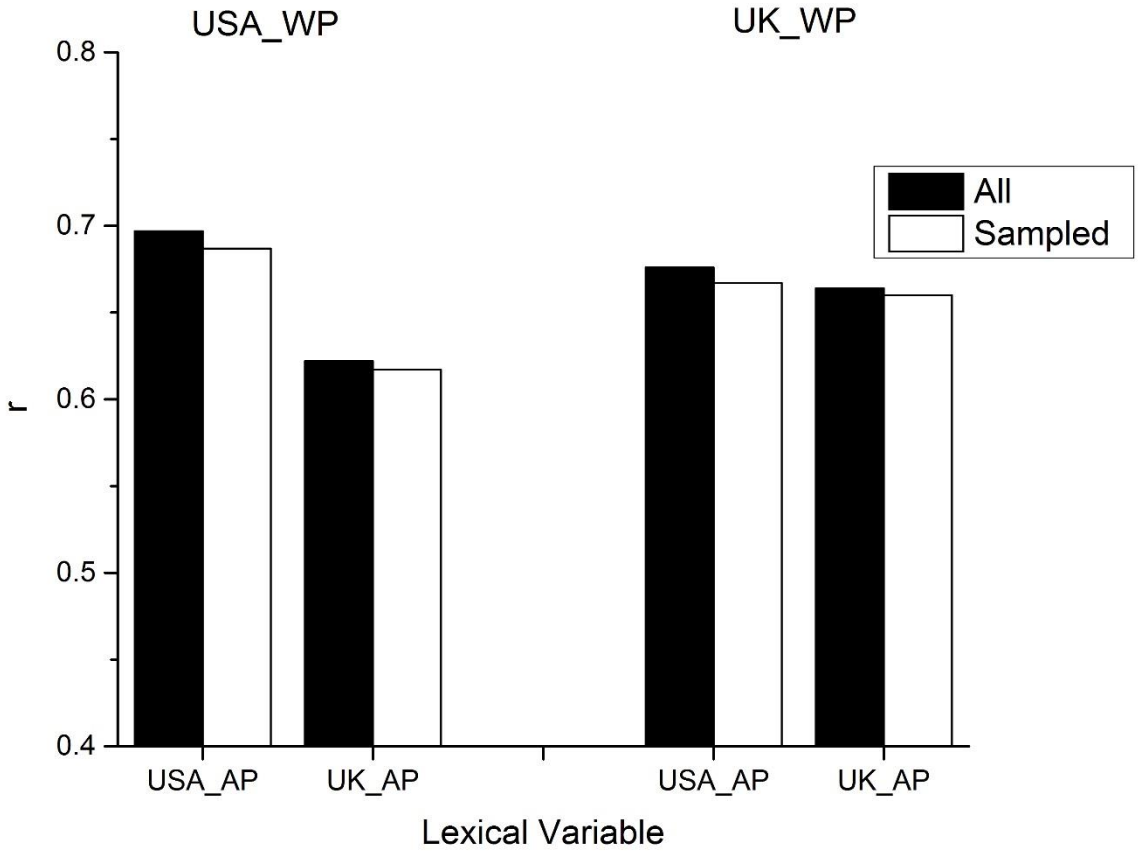

Figure 5. Results of a Monte Carlo analysis testing whether the superiority of the USA_AP variable results from the greater level of materials assembled from American authors.

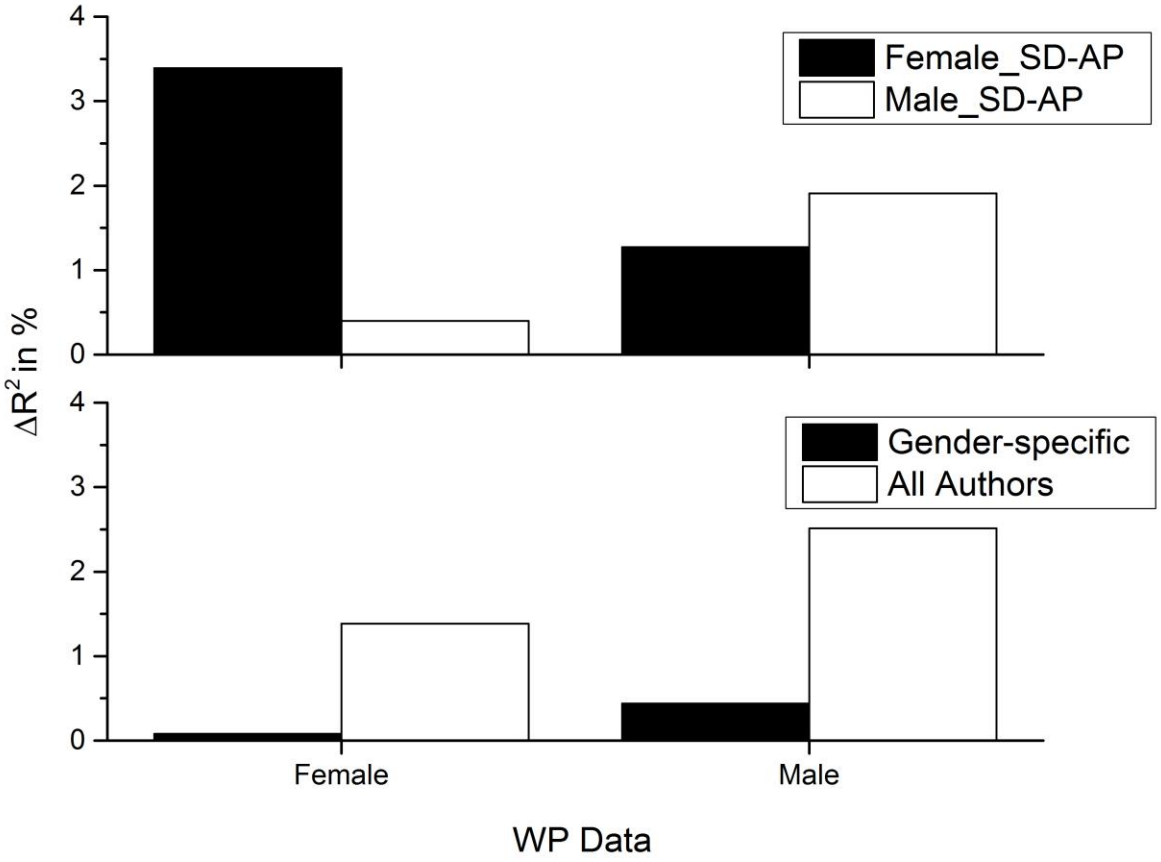

Figure 6. Results of regression analyses testing whether female or male WP data is better accounted for by the SD-AP variable when trained on male or female authors (top panel), and whether the gender-specific SD-AP variable accounts for more variance than the SD-AP variable trained on all authors (bottom panel).