COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2018) 1–15
Copyright © 2018 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12583

A Large-Scale Analysis of Variance in Written Language

Brendan T. Johns,[a] Randall K. Jamieson[b]

[a]*Department of Communicative Disorders and Sciences, University at Buffalo*
[b]*Department of Psychology, University of Manitoba*

Received 4 April 2017; received in revised form 15 November 2017; accepted 21 November 2017

Abstract

The collection of very large text sources has revolutionized the study of natural language, leading to the development of several models of language learning and distributional semantics that extract sophisticated semantic representations of words based on the statistical redundancies contained within natural language (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The models treat knowledge as an interaction of processing mechanisms and the structure of language experience. But language experience is often treated agnostically. We report a distributional semantic analysis that shows written language in fiction books varies appreciably between books from the different genres, books from the same genre, and even books written by the same author. Given that current theories assume that word knowledge reflects an interaction between processing mechanisms and the language environment, the analysis shows the need for the field to engage in a more deliberate consideration and curation of the corpora used in computational studies of natural language processing.

Keywords: Distributional semantics; Cognitive modeling; Natural language processing; Big data analytics

1. Introduction

People read millions of words each year (Brysbaert, Stevens, Mandera, & Keuleers, 2016), a scenario that makes an analysis of language learning difficult. To deal with the problem at a meaningful scale, researchers have developed and relied on computational models of language (Brysbaert, Mandera, & Keuleers, 2017; Jones, 2016; Landauer & Dumais, 1997).

Correspondence should be sent to Brendan Johns, Department of Communicative Disorders and Sciences, University at Buffalo, 122 Cary Hall, Buffalo, NY 14214. E-mail: btjohns@buffalo.edu

The general approach—called *distributional semantics*—is represented in several models, including the LSA, BEAGLE, Topic, retrieval, and neural embedding models (e.g., Griffiths et al., 2007; Johns & Jones, 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013). Although the models differ in important ways (see Jones, Willits, & Dennis, 2015, for a review), they share a common spirit: Simple learning mechanisms applied to a corpus of natural language can explain how people learn word meanings. Models of this type use different underlying mechanisms, but they all propose an account of language learning in which a word's meaning derives from the company it keeps—consistent with a more general approach dating back to Wittgenstein (1953), who stated, "The meaning of a word is its use in the language" (p. 43).

In combination with frameworks for memory and decision, the theories have been established as productive accounts of language behavior, including false memory, word classification, lexical organization, bilingualism, semantic search, and also cognitive technologies for diagnosis of memory impairment, text analysis, and semantic search engines (e.g., Chubala, Johns, Jamieson, & Mewhort, 2016; Hills, Jones, & Todd, 2012; Jamieson, Aujla, & Cook, 2017; Johns, Dye, & Jones, 2016; Johns et al., in press; Johns & Jones, 2010; Johns, Jones, & Mewhort, 2012; Johns, Sheppard, Jones, & Taler, 2016; Jones, Johns, & Recchia, 2012; Rubin, Koyejo, Gorgolewski, Jones, Podrack, & Yarkoni, in press; Taler, Johns, Young, Sheppard, & Jones, 2013). But there is a problem lurking underneath.

Although distributional models acknowledge and even emphasize the interaction between learning mechanisms and the environment, analyses have focused on the learning mechanisms. To represent the environment, researchers have used large and varied corpora aimed at averaging out idiosyncrasies and artifacts from a selective reading history. The solution is consistent with wisdom from sampling theory: A particular corpus can force idiosyncrasies into the semantic space, but if you include a large random sample of text, the idiosyncrasies wash out (i.e., you obtain a reliable point estimator). It is also consistent with sound experimental logic: Vary one factor at a time so that causes can be inferred properly.

Although the random sampling strategy is statistically sound, it neglects an interesting psychological question. Peoples' word knowledge depends on what they have read. In the work presented here, we ask is if there is sufficient variability in written language to motivate an analysis of semantics conditional on a selective reading history. To answer the question, we compiled a corpus of fiction books and compared the structure of language use between genres, within genres, within authors, between authors, and as a function of publication date. We expect to show differences in language use between genres, and to a lesser extent within genres. We expect individual authors to use language in a consistent manner, but with some variation over time. If we are correct, the results will provide proof of concept that variation in a person's experience with language is sufficient to motivate a deliberate curation of the documents included in corpora.

In contrast to standard analysis that uses encyclopedias and newspapers, we conducted our analysis using fiction books. We made the decision for several reasons. First, fiction is a well-formed domain with an established set of genres, and we will use the established categories to define our comparisons. Second, people willingly declare the genres that they read and the genres they neglect. Therefore, the results will speak to meaningful differences in peoples' selective language histories. Finally, each genre includes a number of authors and, thereby, supports a comparison of the variation in language use between categories and within categories.

2. Method and materials

2.1. Materials

We compiled fiction novels from seven genres: *fantasy*, *mystery*, *thriller*, *historical fiction*, *science fiction*, *romance,* and *literature*. These were the seven most popular fiction genres on the book review website *GoodReads*, and books that were highly cited in multiple genres were not used. All texts were taken from a book's digital publication.

Table 1 shows the number of books in each genre, the number of authors per genre, the number of books per author, and the number of words per book. Because novels vary in length, the number of books from each genre varied with the overall goal of the sample to keep the amount of linguistic information (i.e., number of words) equivalent across the collections. In total, the corpus included 1,850 books totaling approximately 240 million words (i.e., a large sample of natural language).

All books were categorized into a genre based on the classifications published by *GoodReads* and the online retailer *Amazon*. All genres assigned were the first tagged genre on both websites. Books that had a mixed genre (i.e., were tagged as belonging to a particular genre by one site but not the other) were removed from the analysis. We also recorded each book's publication date so that we could conduct an analysis of language use as a function of publication date.

2.2. Comparison method

We used two computational methods to analyze differences in language usage. The first was a bag-of-words model and the other was the BEAGLE model of semantics

Table 1
Characteristics of corpora

Genre	Total Number of Books	Total Number of Authors	Average Number of Books per Author	Average Number of Types per Book	Average Number of Words per Book
Fantasy	200	54	3.7	9,913	167,582
Mystery	352	31	11.35	7,769	91,343
Science fiction	265	32	8.28	9,731	121,768
Romance	300	53	5.66	7,276	102,996
Historical fiction	201	31	6.48	10,718	155,650
Literature	236	32	7.38	10,252	112,659
Thriller	244	22	11.09	10,319	133,227

(Jones & Mewhort, 2007). These two representation types were selected because they are complementary in terms of the information they provide. A bag-of-words analysis provides a high-level examination of the content of books, whereas BEAGLE provides a more fine-grained look into the lexical statistics of language use.

In the bag-of-words model, a book is represented by the frequency distribution of its constituent words, with no attention paid to the order in which words occurred. It is considered a simple method of constructing semantic representations and is often the starting point for more complicated machine learning algorithms. By analyzing what words are used at the book level, we can examine the differences in the word use and semantic content of any two books. To conduct the bag-of-words analysis, the most frequent 80,000 words from all the books collected were counted and each book's representation was equal to the count (i.e., the raw frequency) of each word in that particular book; thus, each book was represented by a large vector representation.[1] A vector cosine was used to compute the similarity between books. A vector cosine is a normalized dot product and returns a value between 1 (completely similar) and −1 (completely different), and it is calculated with the following equation:

$$\mathrm{Cos} = \frac{X \cdot Y}{\sqrt{X^2} * \sqrt{Y^2}}$$

where **X** and **Y** are two vectors.

To build upon the bag-of-words analysis, we re-conducted the analysis using the BEAGLE model of semantics (Jones & Mewhort, 2007). BEAGLE constructs a lexical semantic representation of each word in a corpus by exploiting statistical redundancies in word co-occurrence in the corpus. In this model, words are initially represented by randomly generated static environmental vectors, which are assumed to represent perceptual properties of a word.[2] Each word has a different environmental vector. These environmental vectors do not change across learning and are used as the building blocks to learn different types of lexical relations. The model works at the sentence level and records the usage of a word by learning two types of statistical information: context (i.e., the words that co-occur with a word in language such as *cat-mouse*) and order (i.e., the shared syntagmatic roles of words with respect to other words such as both *cat* and *panther* pounce on prey).

The information extracted from the model's learning processes is stored in large, distributed vectors using principles of holographic memory (Gabor, 1968, 1969; Longuet-Higgins, 1968; Murdock, 1982; Murdock, 1983, 1995, 1997; Poggio, 1973). Each word has its own context and order representation, and each representation is updated every time a word occurs in a sentence. A word's context vector is updated by summing the environmental vectors of the other words in a sentence into the overall context representation. This means that the context representation is learning direct co-occurrence information. A word's order representation is update by constructing the n-grams that surround a word in a sentence (up to a certain size) and summing these n-grams into the overall order representation. The order representation is learning how a word is used in relation to the words that it is surrounded by within a sentence. In sum, a word's context vector represents pure co-occurrence information, while order information is encoding simplified syntactic relations. The similarity of two words was again assessed with a vector cosine.

Typically, distributional models are used to construct representations of individual word meanings. Even though a single book includes a significant amount of lexical information, it does not contain enough information to derive consistent lexical semantic representations at the single word level. To overcome this limitation, a modified version of the BEAGLE model was used to construct representations at the level of an individual book, rather than a word. Instead of summing context or order information into a word's semantic representation, we summed the vectors into a single composite vector that represents the book. Thus, each book was represented as a sum of the lexical statistics that defines how words were used in that book. Context and order information were stored in separate vectors. Similarity between two books was assessed by taking the vector cosine between their respective vectors.

More formally, a book's context representation will be equal to the sum of each word's environment vector into a composite representation:

$$\mathbf{Book}_{\mathbf{Context}} = \sum_{i=1}^{S}\sum_{j=1}^{W}\mathbf{e}_{i,j}$$

where **Book** is the composite vector representing a single book, S is the number of sentences in the book, W is the number of words in the sentence, and $\mathbf{e}_{i,j}$ is the environmental vector for the word in sentence i in position j within that sentence. This tends to return a noisier distribution than the bag-of-words analysis (due to the use of random vectors; Johns & Jones, 2010). But it provides an index for analysis from a known and well-used theory of semantic knowledge.

The key benefit from BEAGLE is its ability to form order representations by constructing n-gram representations. As described in the original paper, BEAGLE constructs n-gram information by binding together the environmental vectors for a specified number of the words. This binding is accomplished with directional circular convolution, a standard technique in mathematical models of memory (see Jones & Mewhort, 2007; Plate, 2003). For each book, a separate representation will be constructed for n-grams from size 2 to 4, which will be computed with the following equations:

$$\mathbf{Book}_{\mathbf{Bigram}} = \sum_{i=1}^{S}\sum_{j=2}^{W}\mathbf{e}_{i,j-1} \circledast \mathbf{e}_{i,j}$$

$$\mathbf{Book}_{\mathbf{Trigram}} = \sum_{i=1}^{S}\sum_{j=3}^{W}\mathbf{e}_{i,j-2} \circledast \mathbf{e}_{i,j-1} \circledast \mathbf{e}_{i,j}$$

$$\mathbf{Book}_{\mathbf{Quadgram}} = \sum_{i=1}^{S}\sum_{j=4}^{W} \mathbf{e}_{i,j-3} \circledast \mathbf{e}_{i,j-2} \circledast \mathbf{e}_{i,j-1} \circledast \mathbf{e}_{i,j}$$

where **Book** is the composite vector representing a single book and $\circledast$ signifies directional circular convolution. These different representations allow us to inspect how individual books differ in their use of increasingly long strings of words. Encoding of n-gram information (a sequence of *n* items within a sentence) will also provide a meaningful contrast to the distributions constructed with the bag-of-words and BEAGLE context vectors, as it captures word order information. The BEAGLE analysis has multiple advantages, with the first being that context vectors allow for a direct replication of the results of the bag-of-words model, strengthening the results of the combined analyses. Secondly, it offers a coherent model that can construct increasing units of language. These statistics can be integrated into a single vector to represent a "bundle" of language statistics, enabling a direct analysis of word order in the same way that co-occurrence information is analyzed.

The similarity values for both models were computed at three levels: (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). With the number of books collected in this analysis, these three levels of analysis provide a good deal of information to distinguish differences in word usage. The within-author distribution was composed of 11,262 comparisons (i.e., one similarity measurement between every pair of books by the same author), the within-genre distribution was composed of 249,647 comparisons (i.e., one similarity measurement between every pair of books within the same genre), and the genre distribution was composed of 1,371,484 comparisons (i.e., one similarity measurement between every pair of books in different genres). These values were calculated by assessing the number of pairwise comparisons within the different levels, given the number of books available for the different comparisons. The distributions are sufficiently populated to allow for an examination of how language usage changes as a function of author, genre, and publication date.

3. Results

3.1. Bag-of-words analysis

Fig. 1 shows the similarity distributions from the within-author, within-genre, and across-genre comparisons using the bag-of-words method. As shown, language use varies about equally at all levels of analysis, as indicated by the large but approximately equal spread in all three of the similarity distributions. However, the figure also shows a difference between all of the distributions with the similarities between books written by the same author (M = 0.772, SD = 0.13) being higher than the similarities between books belonging to either the same (M = 0.712, SD = 0.054) or across genres (M = 0.662,

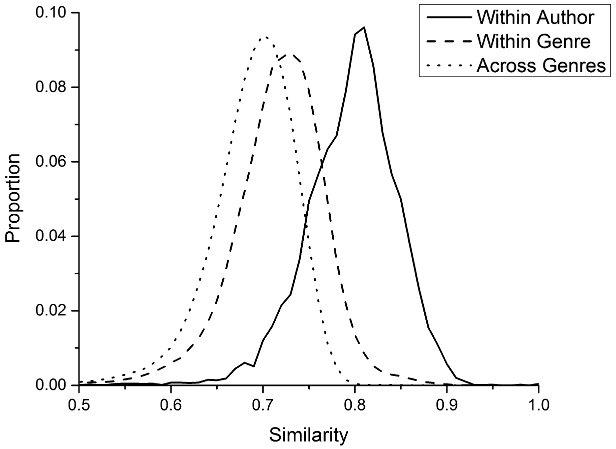

Fig. 1. Similarity distributions of books across three levels for the bag-of-words analysis: (1) books written by the same author, (2) books written within the same genre, and (3) books written in different genres. This figure shows that books written by the same author have much more distinct usage of language when compared with books by different authors.

SD = 0.128). As also shown, similarity calculations within genres are shifted positively relative to the across genres distribution, indicating that books within the same genre have some systematic relation to one another. However, the most striking aspect of Fig. 1 is how positive the similarity calculations are for books written by the same author. That difference presents quantitative evidence that books written by the same author are much more similar to one another than they are to books written by different authors—a result that indicates high individual variability in the usage of written language. This finding suggests that individual authors exhibit unique patterns of word usage across their book sets.

Fig. 2 examines the results in Fig. 1 at a finer scale by showing the average similarity at each of three levels of comparison, but split by genre. As shown, there is very little change in the average values across genres, whereas there is some variability in the average within-author and within-genre comparisons. Different genres show different average similarity values, for example, the *literature* genre has both the lowest within-author and within-genre similarity values, suggesting that authors within this genre have the most unique books compared to the other books they have written and the other books within the same genre. In contrast, some genres have both a high within-author and within-genre similarity; for example, the *fantasy* genre has high average similarities for these comparisons. These differences could be due to particular characteristics of writing within a genre, such as the fact that most fantasy books are written in multi-volume anthologies, leading to greater within-author similarity.

An additional analysis for the distributions in Fig. 1 is to examine similarities by date of publication. This allows one to look at how language use changes across time within authors, across genres, and between authors. Results of that analysis are presented in

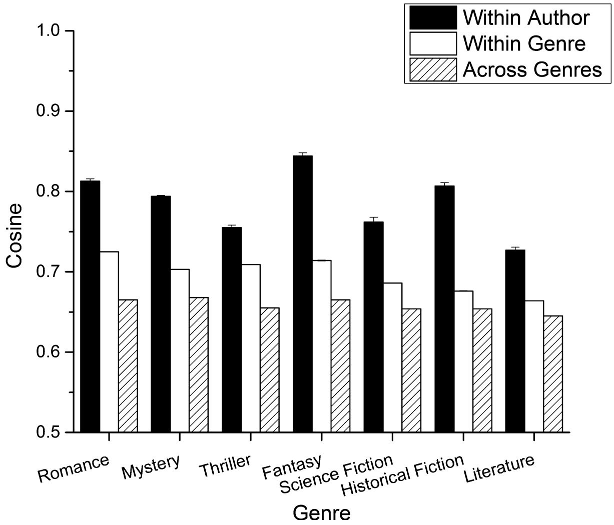

Fig. 2. A breakdown by genres of the similarity distributions contained in Fig. 1. This figure plots the average similarity values for the within-author, within-genre, and across-genres comparisons across the seven different genres. Error bars represent standard error.

Fig. 3 for books published up to 30 years apart—a range selected because it allowed for at least 50 similarity values at each comparison level.

Fig. 3 shows the similarity between all pairs of books plotted against the absolute difference of their publication dates (*x*-axis) for within-author, within-genre, and across-genre comparisons. As shown, all three comparison levels exhibit a negative slope, presenting evidence that books grow less similar as the lag in publication dates increases. The data show an unintuitive pattern where the negative shift in similarity across time is the greatest at the within-author level, demonstrating that the largest change in language use occurs at the individual level (slope = −0.0021). In contrast, there is a smaller change at the within-genre comparison (slope = −0.0011), while there is very little change at the across-genre level (slope = 0.0006).

In Fig. 3, a window of 30 years was used, a consequence of the length of an individual author's window of productivity. But the constraint is loosened when looking at the change in language use within and between genres. Fig. 4 is a companion to Fig. 3, but for a maximum of a 100-year publication lag. This figure shows that there is a constant decrease in similarity as the lag in publication date increases; the trend is true for both the within- and across-genre comparisons. Indeed, with a 100-year difference in publication date between any two books, there is little difference in average similarity for books written in the same genre (*M* = 0.662, *SD* = 0.043) compared to books written in different genres (*M* = 0.665, *SD* = 0.038). At the opposite end of the distribution, books written in the same genre in a similar period have a much greater similarity than books written in the same period across different genres.

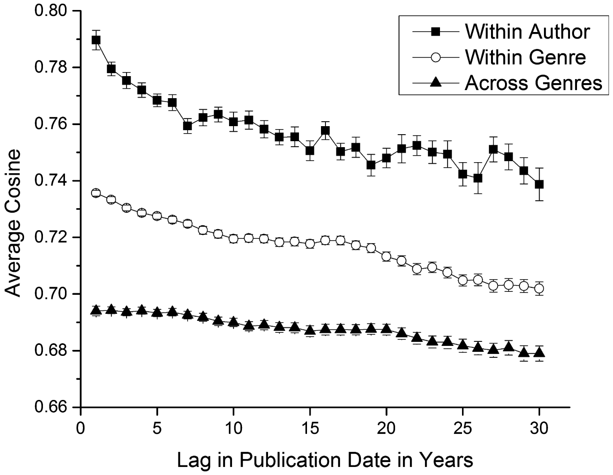

Fig. 3. Similarity of books sorted by difference in the publication date of the two books, with a maximum difference in publication date up to 30 years. This figure shows that there is a downward trend in similarity for all levels of analysis, but that there is a greater change at the individual author level. Error bars represent standard error.

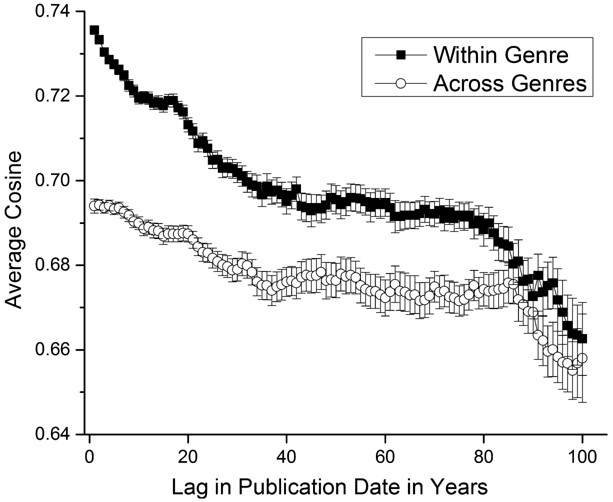

Fig. 4. Change in similarity of books up to 100 years in difference in publication date for books written within the same genre and books written in different genres. This figure shows that there is a large period effect for genre similarity, where books written in the same genre in the same time period are much more similar than books written across genres. Error bars represent standard error.

In summary, the bag-of-words analysis provides strong evidence for variation in language use in fiction between genres, within genres, within authors, and as a function of publication lag. If people read selectively rather than globally, the analysis recommends a careful consideration of the books included in a corpus intended to model language learning of a given cohort of learners.

3.2. BEAGLE analysis

Next, we applied the BEAGLE model of semantics to construct context and n-gram representations. The distribution for representations of context and the four n-gram representations are contained in Fig. 5. As expected, the distributions for the context representation are quite similar to the results of the bag-of-words analysis (see Fig. 1), as they both measure the same information (word occurrence overlap). As shown, the distributions have a greater degree of overlap than what was seen in the bag-of-words analysis, likely due to the noisy nature of distributed random vector representations. However, for the bigram representations, there is still a large increase in the similarity distribution for books written by the same author—a result that suggests it is not just the types of words that are unique to an individual's language use but also a change in the word order that individuals use. However, this effect is reduced in trigrams, and almost eliminated at the quadgram level, suggesting that there are fewer unique signatures of individual differences in language, the longer the unit of analysis. The same pattern was found for the differences in books written in the same genre versus those written in different genres with the effect being virtually eliminated at the trigram and quadgram levels.

To examine how these different representation types change across time, Fig. 6 displays the change across 30 years for the three comparison and four representation types. As expected, context representation once again closely resembles the bag-of-words representation. The bigram representation shows a similar effect for the within-author comparison as found previously in Fig. 3 with a negative slope, suggesting that word order is also changing as a function of publication date. However, unlike the bag-of-words and context representation, there was no difference in the within- and across-genre comparisons. For the trigram representation, the change across time for the within-author comparison was less pronounced. This result was entirely absent in the quadgram representation comparison.

In summary, the BEAGLE method corroborates the results from the bag-of-words analysis: There is strong and now consistent evidence for variation in language use in fiction between genres, within genres, within authors, and as a function of publication lag, and that this applies to local word order (bigrams and trigrams).

4. General discussion

The substance of this article is an analysis of variance in written language. The aim of the analysis was to present those measurements as motivation for a more deliberate curation of the language environment in semantic and language modeling. To accomplish this, we compiled a substantial collection of books from seven genres written by hundreds of authors. We applied a bag-of-words and a semantic model to measure how word use differs as a function of author, genre, and publication date. Both analyses converged on common conclusions. Books written in the same genre are more alike compared to books written in different genres and books written by the same author are much more similar

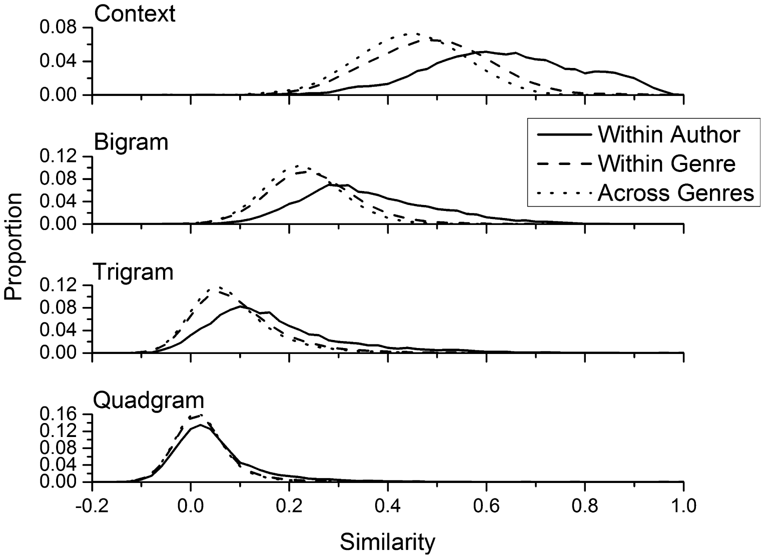

Fig. 5. Similarity distributions derived from the BEAGLE model of semantics using context, bigram, trigram, and quadgram representation types.

to one another than books written by different authors. The results demonstrate that authors have their own unique usage of language.

An analysis of differences in publication date revealed that the greatest difference in language use occurred within an individual author compared to within or between genres. That result suggests that on a short time scale (i.e., a 30-year period), a great deal of the change in language use reflects changes within the individual author. However, when the timescale is increased (i.e., a 100-year period), changes were revealed both within and between genres. Specifically, there is a substantial effect of publication period on genre, suggesting that books written in the same genre are most similar when they are written in the same time period.

By one view, our analysis offers a glimpse into the distributional structure of the written word and offers an interesting method for computational linguistics and humanities. More generally, it shows the importance of measuring and understanding the variance of language usage when considering the comparison of models for natural language processing. Contrary to intuition, the quantitative analysis of language use presented here shows that the majority of the uniqueness of language use is contained at the individual author level rather than the genre level. However, in the domain of cognitive science, the analysis points to a more important contribution.

The results of this article represent an evolution in the distributional approach to language. Typically, models of distributional semantics are trained on a standard corpus, and the capabilities of different learning algorithms are assessed on how well they are able to perform on language tasks. However, these corpora are composed of randomly sampled documents including many different authors writing on many different topics. But, people do not read as randomly as random sampling supposes and, to the extent that a model's

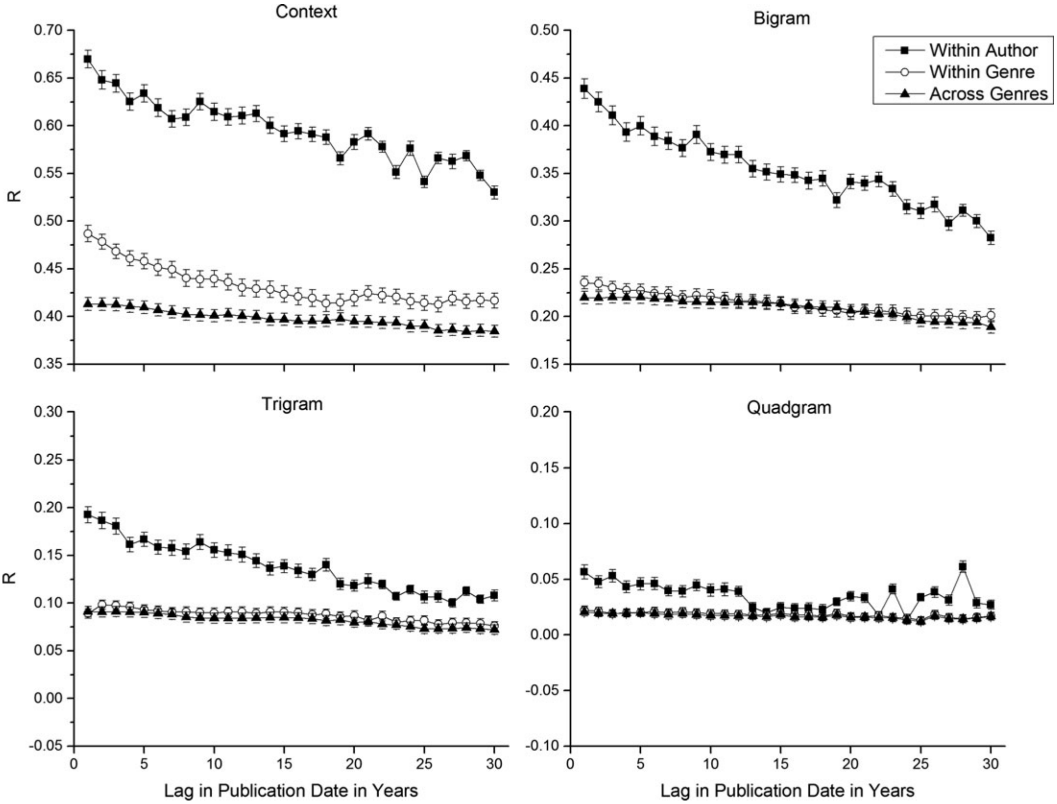

Fig. 6. Differences across time for the BEAGLE model of semantics using context, bigram, trigram, and quadgram representation types. Error bars represent standard error.

fit to human behavior depends on the interaction between the processing model and the corpus to which the processing model is applied, it is equally important to model the language environment of the target group.

The results present a meaningful challenge for corpus-based models of language. For one, the assumption that a random sample of texts can stand for language experience on the whole might be incorrect and a more deliberate curation of language experience is important for making progress. Indeed, our results point to the composition of language being quite variable. If models are only trained on either a single genre (i.e., corpora composed of only newspaper articles, movie abstracts, books, Wikipedia) or too many sources (i.e., random sampling of text), the model represents an experimental participant with an unusual or even distorted view of the natural language environment. In that case, the wisdom of the less-is-more principle of ecological rationality applies (e.g., Goldstein & Gigerenzer, 2002). Using a randomly sampled corpus to model language induction is a good practical research strategy. But it misconstrues the more selective nature of peoples' reading histories. To the extent that language learning is an interaction between learning mechanisms and the language environment, attention needs be given to both factors.

In closing, our analysis motivates an interesting problem for work in the field of natural language processing and semantics. The method of randomly sampling texts to a corpus might be a sensible strategy for modeling language behavior averaged over a large number of individuals. But an ideal model would model language behavior in individuals or subgroups of the larger population. The fact that language differs by genre and author presents a way to test if current models are capable of that precision. If a corpus was constructed from the reading history of two different individuals, and language models were applied to derive the meaning of words from those histories, the word meanings derived from one individual's history should match that individual's language judgments better than the words meanings derived from another person's history. If true, the result provides a formal basis for understanding and predicting individual differences in language comprehension and semantics. It also provides the basis for techniques to infer a person's preferences, behaviors, and psychology (see Johns, Jones, & Mewhort, 2016). Indeed, as more and more reading and language behavior is tracked online, the possibility of curating personalized corpora as the input for psychological models becomes increasingly feasible (Griffiths, 2015).

Notes

1. Function words were included, as removing them did not have a large effect on the results.
2. Environment vectors are random Gaussian vectors, constructed with a mean of 0 and a standard deviation of $1/\sqrt{n'}$, where n represents the number of elements in the vector. In this study, vectors had a size of 2,048.

References

Brysbaert, M., Mandera, P., & Keuleers, E. (2017). Corpus linguistics. In A. M. B. De Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics* (pp. 230–246). London: Wiley.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How any words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116.

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to the implicit rule-learning task: Implicit learning of semantic structure. *Quarterly Journal of Experimental Psychology*, *69*, 1049–1055.

Gabor, D. (1968). Improved holographic model of temporal recall. *Nature*, *217*, 1288–1289.

Gabor, D. (1969). Associative holographic memories. *IBM Journal of Research and Development*, *13*, 156–159.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Hills, T., Jones, M., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*, 431–440.

Jamieson, R. K., Aujla, H., & Cook, M. T. (2017). A psychologically inspired search engine. In *Lecture Notes in Computer Science: High Performance Computing Systems and Applications*. Berlin: Springer.

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

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., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, *69*, 233–251.

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., Jones, M. N., & Mewhort, D. J. K. (2016). Experience as a free parameter in the cognitive modeling of language. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J. C. (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

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

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F. W., & Jones, M. N. (in press). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology*.

Jones, M. N. (2016). Developing cognitive theory by mining large-scale naturalistic data. In M. N. Jones (Ed.), *Big data in cognitive science* (pp. 1–12). New York: Taylor & Francis.

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

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

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

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

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *In the International Conference on Machine Learning*, *14*, 1188–1196.

Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature*, *217*, 104.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

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

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review*, *90*, 316–338.

Murdock, B. B. (1995). Developing TODAM: Three models for serial-order information. *Memory & Cognition*, *23*, 631–645.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*, 839–862.

Plate, T. A. (2003). *Holographic reduced representations (CSLI Lecture Notes No. 150)*. Stanford, CA: CSLI Publications.

Poggio, T. (1973). On holographic models of memory. *Kybernetik*, *12*, 237–238.

Rubin, T. N., Koyejo, O., Gorgolewski, K. J., Jones, M. N., Poldrack, R. A., & Yarkoni, T. (2017). Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *PLoS Computational Biology*, *13*(10), e1005649.

Taler, V., Johns, B. T., Young, K., Sheppard, C., & Jones, M. N. (2013). A computational analysis of semantic structure in bilingual fluency. *Journal of Memory and Language*, *69*, 607–618.

Wittgenstein, L. (1953). *Philosophical investigations*. Hoboken, NJ: John Wiley & Sons.