Behavior Research Methods
https://doi.org/10.3758/s13428-019-01234-0

Gender bias at scale: Evidence from the usage of personal names

Brendan T. Johns [1] · Melody Dye [2]

Abstract
Recent research within the computational social sciences has shown that when computational models of lexical semantics are trained on standard natural-language corpora, they embody many of the implicit biases that are seen in human behavior (Caliskan, Bryson, & Narayanan, 2017). In the present study, we aimed to build on this work and demonstrate that there is a large and systematic bias in the use of personal names in the natural-language environment, such that male names are much more prevalent than female names. This bias holds over an analysis of billions of words of text, subcategorized into different genres within fiction novels, nonfiction books, and subtitles from television and film. Additionally, we showed that this bias holds across time, with more recent work displaying the same patterns as work published tens or hundreds of years previously. Finally, we showed that the main cause of the bias comes from male authors perpetuating the bias toward male names, with female authors showing a much smaller bias. This work demonstrates the potential of big-data analyses to shed light on large-scale trends in human behavior and to elucidate their causes.

Keywords Big data · Gender bias · Computational social science · Lexical organization · Distributional modeling · Corpus studies

Cognitive models of lexical organization and lexical semantics point to the structure of the natural-language environment as being the main organizer of these systems (e.g., Brysbaert, Mandera, & Keuleers, 2018; Jones, Dye, & Johns, 2017; Landauer & Dumais, 1997). The implication of these models is that the structure and content of the language that people experience directly impacts their knowledge and lexical-processing systems. Thus, if human experience with natural language is biased in systematic ways, there will be a corresponding effect on the human language and memory systems.

Indeed, recent research using advanced computational methods has demonstrated that the natural-language environment embodies many of the implicit biases that are seen in human behavior (Caliskan, Bryson, & Narayanan, 2017). As the basis of this study, the researchers derived word meaning representations from a distributional model of semantics, a class of model that learns the meaning of words from large natural-language corpora (see Jones, Willits, & Dennis, 2015, for a review). Using stimuli similar to those in the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998), a standard measure of the automatic associations that people have between concepts (such as gender or race), Caliskan et al. showed that many of the implicit biases that people have were also encoded within their model's representations. The results of this study suggest that one cause of implicit bias in human behavior comes from the content of the language that people experience.

One of the biases explored by Caliskan et al. (2017) was gender bias, which was examined through the associations that the model formed to male versus female personal names. Gender biases are common in many aspects of modern life. For example, males receive higher wages (Kilbourne, England, Farkas, Beron, & Weir, 1994), receive more support as students (Steele, 1997), and receive higher peer-reviewed scores on academic applications (Wennerås & Wold, 1997), to name just a few common differences. The accumulation of such biases likely results in a myriad of prejudices against females, leading to, for instance, the lack of female advancement in academia (Barres, 2006).

Although Caliskan et al. (2017) focused on the meaning and valence associated with personal names, here we will focus on the word frequencies of male versus female names.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-019-01234-0) contains supplementary material, which is available to authorized users.

✉ Brendan T. Johns
btjohns@buffalo.edu

[1] University at Buffalo, Buffalo, NY, USA

[2] University of California, Berkeley, CA, USA

Published online: 22 April 2019

Word frequency has ubiquitous effects on the human language and memory systems, with high-frequency words being retrieved more easily and identified faster (Forster & Chambers, 1973; Morton, 1969; Scarborough, Cortese, & Scarborough, 1977) and recalled at a greater rate (Gregg, 1976), and with word frequency being a main information source used in decision making (Tversky & Kahneman, 1973). Frequency and related derived variables are central to modern theories of lexical organization (e.g., Adelman, Brown, & Quesada, 2006; Brysbaert et al., 2018; Jones et al., 2017; Jones, Johns, & Recchia, 2012). Thus, systematic discrepancies in frequency distributions could have considerable consequences for the lexical organization of words—in the case of this study, personal names.

However, to determine whether there is a bias in the natural-language environment for personal names, it is first necessary to determine what the actual prevalences of those names are in the real world. Luckily, the Social Security Administration of the US government recently released a database of personal names from 1880 onward (this data source will subsequently be referred to as the *SSA data*).[1] This database has a total of approximately 330 million names, organized by date of birth, frequency, and gender.

The SSA dataset provides a unique opportunity for the study of natural language. The assumption of lexical organization and lexical semantic models is that language provides an accurate snapshot of the overall environment that humans are embedded in. For example, one theory put forth to explain the importance of contextual information in lexical organization is the principle of "likely need" (Adelman et al., 2006; Anderson & Schooler, 1991; for a review, see Jones et al., 2017). In terms of lexical organization, the principle of likely need states that a word that has been experienced in many contexts during learning is more likely to be needed in unknown future contexts; hence, it should be more accessible in the lexicon (when compared to a word that has been experienced in fewer contexts). However, if a word's (e.g., a personal name's) occurrence pattern is not actually diagnostic of that word's referent's (e.g., a person's) probability of occurrence in the world, this would lead to an inefficient and poorly organized lexical system for that class of words. The SSA data allow for a determination as to whether the natural-language environment systematically deviates from the structure of real-world demographics, by assessing whether linguistic corpora have a similar distributional structure for personal names.

The overarching goal of this article was to determine whether there is a systematic bias away from female names in the natural-language environment, exposing an additional source of gender bias. This analysis was done over billions of words of text, subcategorized into different genres within fiction and nonfiction books and subtitles from television and film. Additionally, a subset of these text sources were categorized by time of publication and author gender, to determine whether author characteristics influence the usage of personal names.

Method

Materials

Three text sources were used in this study: (1) nonfiction books, (2) fiction books, and (3) subtitles from films and television shows. These different text sources were categorized by their source genre. There were ten genres of fiction books, eight genres of nonfiction books, and six genres of subtitles. The reason why the lexical sources were split by genre was to ensure that any biases found were consistent across the language materials, and not due to the composition of a particular genre.

The nonfiction book collection was composed of textbooks or common writings from six academic fields (*history*, *psychology*, *chemistry*, *physics*, *classics*, and *political science*). This book collection was sorted using the cataloging system from the library at the University at Buffalo, and the books were transformed from ebooks to a machine-readable format. The other two genres analyzed were travel guides and how-to (reference) books.

There were six genres of subtitles (*drama*, *action*, *comedy*, *crime*, *family*, and *documentary/reality*). The subtitles of television shows and films were sorted as to their genre using the Internet Movie Database (IMDB) website.

The largest group of texts tested here was fiction books, which were sorted into ten genres (*literature*, *historical fiction*, *romance*, *fantasy*, *science fiction*, *thriller*, *young adult*, *crime*, *horror*, and *mystery*). This book set was organized by author. To attain genre information for the fiction books, the dominant genre that an author wrote in was recorded by using the most frequent tag on the book review website Goodreads and online retailer Amazon. The books written by that author were then labeled as having being written in that genre. Although this is less precise than in author studies examining the impact of genre on writing (see Johns & Jamieson, 2018), tagging each book by its genre was infeasible for such a large collection.

The fiction book set is described in Table 1 and was produced by 3,208 authors. To understand the effect of author characteristics on personal name usage, the following information about the different authors was assembled: (1) author gender, (2) date of birth, and (3) place of birth. This information was collected by labeling each author with publicly available information from Goodreads, Amazon, or the online encyclopedia Wikipedia. This information was collected

[1] The data can be accessed at https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data.

Table 1 Characteristics of language sources across genres

Type	Genre	# of Sources	Words per Source	Total # of Words
Fiction	Fantasy	2,231	98,972	220,806,532
	Historical fiction	1,312	108,995	143,001,440
	Mystery	2,505	72,928	182,684,640
	Romance	5,004	74,835	374,474,340
	Thriller	1,245	99,249	123,565,005
	Young adult	2,785	40,066	111,586,014
	Crime	643	73,841	47,479,763
	Literature	2,740	88,729	243,117,460
	Horror	728	86,202	62,755,056
	Science fiction	5,336	75,225	401,400,600
	Total/average	**24,529**	**81,904**	**1,910,870,850**
Nonfiction	Chemistry	479	97,628.88	46,764,234
	Politics	101	101,346.72	10,236,019
	How-to	714	115,847.23	82,714,919
	Philosophy	656	103,721.11	68,041,051
	Psychology	338	144,160.99	48,726,415
	Travel	553	145,320.08	80,362,002
	History	475	186,386.41	88,533,544
	General nonfiction	2,100	96,049.54	201,752,069
	Total/average	**5,596**	**123,238.57**	**648,493,757**
Subtitle	Crime	2,000	1,515.88	3,031,768
	Drama	2,000	1,656.18	3,312,363
	Comedy	2,000	1,030.48	2,060,896
	Family	2,000	994.84	1,989,673
	Action	2,000	1,283.65	2,567,306
	Adventure	2,000	1,150.75	2,301,489
	Documentary/reality	1,000	1,664.84	1,664,839
	Total/average	**13,000**	**1,328.09**	**16,928,334**

manually for each author from these sources, and author gender, date of birth, and place of birth were recorded by assessing biographical information from the above-described websites. Author gender and place of birth were recorded for all authors, whereas date of birth was only available for 2,090 authors. The collection of author gender information allowed for determining whether there was a difference in the use of personal names across male and female authors. Date-of-birth information provided knowledge about whether the bias toward male name usage is a historical phenomenon or is still prevalent in modern authors. Place-of-birth information allowed for determining whether a bias toward male names is isolated to a certain geographical location.

The characteristics of these language sources are contained in Table 1. The table shows that this analysis was done over a substantial amount and unique range of natural language, totaling over 2.5 billion words of text. More than 55 million of these words are personal names, signaling that personal names are a relatively frequent part of natural language, and hence should be targets for corpus-based analyses.

SSA data

As we stated previously, the SSA name data provide an opportunity to determine whether natural language is biased for personal names, by assessing whether personal names in natural language have a different distributional structure than the demographic information contained in the SSA data. The SSA data provide a benchmark that the frequency distributions derived from the various corpora can be compared to, enabling an examination of whether the natural-language environment deviates from real-world frequencies in terms of personal name usage.

The first step to analyzing these data was to remove any names that overlapped with other words. This was done by using a list of words that had proper names identified. Specifically, we used the word list from the English Lexicon Project (Balota et al., 2007), which includes mostly common nouns, but also proper nouns (e.g., *Abe* and *Aaron*). In total, the word list from the ELP contains 40,482 words. If a name was found to contain an analogue common noun in the ELP (e.g., *will*, *may*, . . .), it was removed from the analysis, so that these words would not bias the resulting counts.

Additionally, only names that had 250 or more references in the SSA data were used in the analysis, to ensure that a name had an actual correspondence within the social environment.

Some names are used by both males and females (e.g., *Taylor*). If a word had 1,000 or more occurrences in one gender than in the other, then it was added to that gender's name list. If there was a difference of less than 1,000 in the male and female counts for a name, then that name was removed from the analysis.

After removing these names from the database, this left 14,053 female names with a total of 153,705,534 occurrences in the SSA dataset, as well as 6,995 male names with a total of 157,841,547 occurrences. That is, there were considerably more female names (a trend documented elsewhere; e.g., Dye, Johns, Jones, & Ramscar, 2016), but with a relatively equal number of occurrences within the SSA dataset, as

compared to males (i.e., 50.6% of all occurrences from the assembled names are male).

Visualization technique

Zipf (1935) scales were used to visualize the frequency distribution of female and male names, a standard way of examining frequency distributions within quantitative linguistics (see Ferrer i Cancho & Solé, 2003; Piantadosi, 2014; Zipf, 1949). A Zipf scale is a log–log scale, where the *x*-axis corresponds to the log of the rank of that word's frequency within the total distribution, and the *y*-axis corresponds to the log of a word's total frequency. Typically, there is a linear relationship between log word frequency and log rank word frequency. However, this would not necessarily be the case with names, for reasons to do with the contextual dependency of names described below.

Figure 1 contains the Zipf scale of personal names from the SSA dataset, split by gender. This figure shows a greater number of high-frequency male names (e.g., *James*, *John*, *Robert*, and *Michael* are all more frequent than the highest-frequency female name, *Mary*), but with female names having a longer tail of low-frequency names. This pattern of occurrence allows for a determination of whether there is a deviation in personal name usage in natural language. If the frequency distributions derived for personal names from the various text corpora resembled the pattern seen in the SSA data, there would be no bias, and it could be concluded that natural language provides an accurate view of the distribution of personal names in the real world. However, if there were a systematic deviation across the text sources, it would provide evidence of a bias within natural language for personal name usage.

To visualize the frequency distributions of male and female names from the various corpora, all of the names from the SSA data were assembled. For each name from this dataset, the frequency of that name within each text source was computed, separated by gender. The male and female distributions were then ranked and transformed with a logarithm and plotted on a log–log scale. The goal of using this visualization technique is to allow for an examination of the distributional structure of male and female personal names from the various corpora and to determine whether the distributions built from the corpora accurately reflect the patterns seen in the SSA data.

However, Piantadosi (2014) pointed out that the standard method of plotting word frequency distributions, described above, is flawed, because both the frequency of a word and its corresponding rank are calculated within the same corpus. Piantadosi demonstrated that this method results in spurious regularity that is not actually contained in the frequency distribution. To overcome this problem, Piantadosi suggested splitting a corpus in two and calculating the overall word frequency from one half, and the rank of words from the other half. This method will be used here to visualize the word frequency distributions attained from the various corpora, and will be referred to as the *split-corpus method*.

One limitation of the split-corpus method for the purposes of this article is that names are likely to be more contextually dependent than other word classes. For example, if the main character of book *x* is *Jennifer*, this does not mean that the

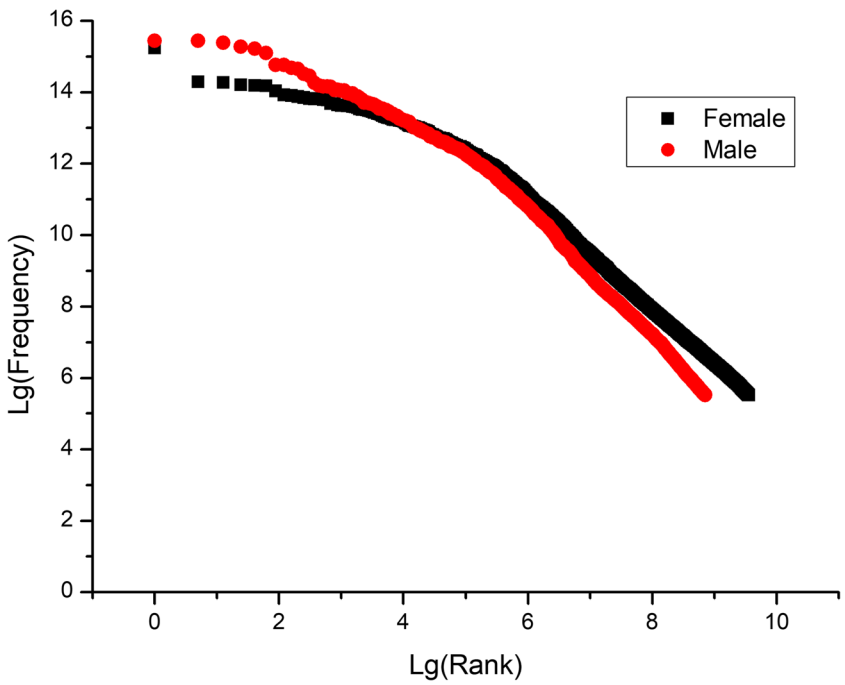

Fig. 1 Zipf scale of personal name frequencies from the SSA data. This figure shows that high-frequency male names are much more frequent than high-frequency female names, but that female names have a longer, low-frequency tail

main character of book $x + 1$ will also be *Jennifer*. Thus, if a concatenated corpus of books is simply split in half, the resulting correlation between rank and frequency will likely be underestimated. To overcome this problem, a corpus will instead be split at the sentence level, with each even-numbered sentence updating overall frequency, and each odd-numbered sentence updating rank frequency. Using this method allows for the contextual dependency of names to be included in the counts, since splitting a corpus at the sentence level does not eliminate the contextual dependency of names.

Results

We performed five analyses in all. The first three analyses examined personal name usage across the different types of texts described in Table 1 (nonfiction books, subtitles, and fiction books). The goal of these analyses was to determine whether there is a consistent bias in the frequency distributions of personal names across the different text sources. In the fourth analysis, we examined how author characteristics impact the usage of female and male personal names. The final analysis contrasted a word frequency count with a contextual diversity count (e.g., Adelman et al., 2006) when examining personal name usage.

Nonfiction books

Each of the eight genres of nonfiction books will be visualized separately, to ensure that each individual genre shows roughly the same pattern of occurrence. Additionally, all the individual genre corpora will be collapsed into a single Zipf graph, to visualize the total distribution of name usage in nonfiction books. To produce the Zipf scales that correspond to the distributions of personal names in the SSA data that are contained in Fig. 1, the word frequencies of the names contained in the SSA data were counted, for both male and female names, and plotted in Zipf scales using the split-corpus method described previously.

Figure 2 contains the Zipf scales for the eight nonfiction genres, whereas the top panel of Fig. 3 contains the collapsed distribution across all eight genres. These figures show that for most genres there is a large and systematic advantage for male over female names. That is, at virtually every point along the frequency distribution, male names are more frequent than female names.

Overall, approximately nine million personal names are used in the nonfiction corpus, of which 61.2% are male. This represents a large, and systematic, bias in this source of language, such that female names are systematically underrepresented in word frequency, when compared to male names.

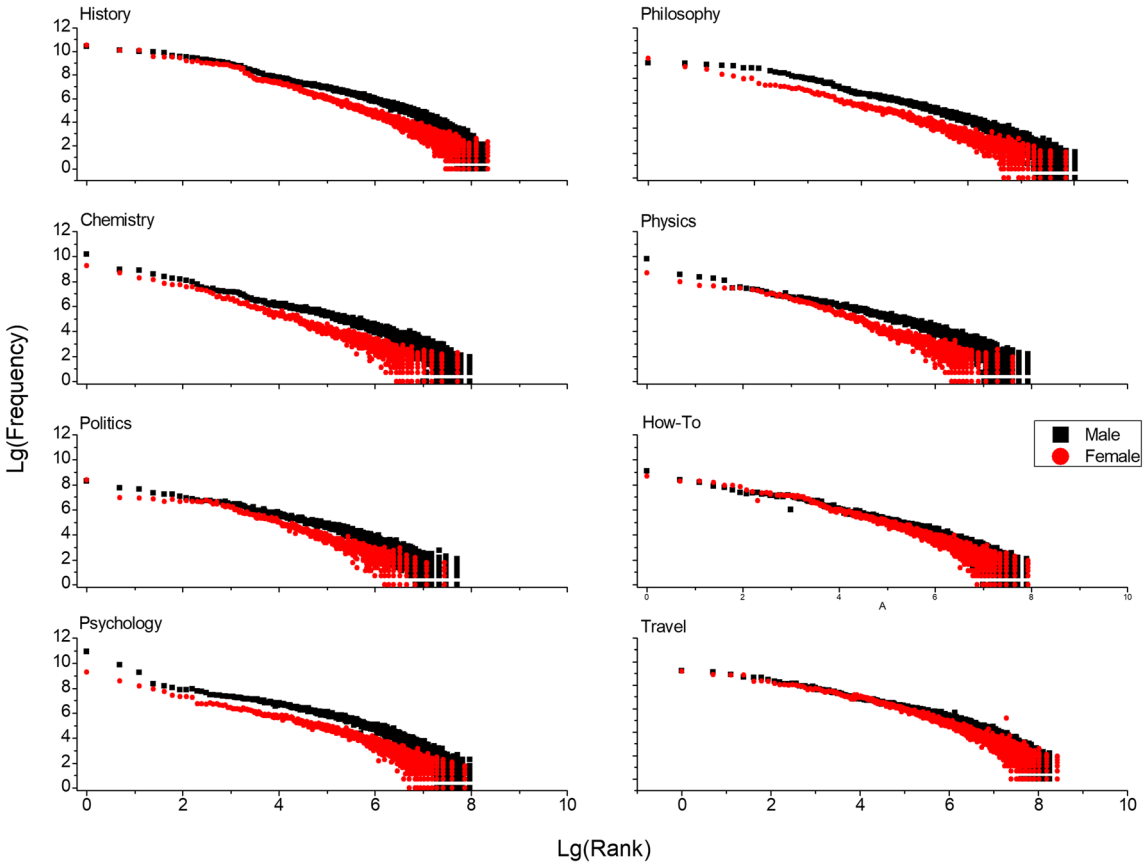

Fig. 2 Zipf scales of male and female name frequencies from the eight nonfiction genres

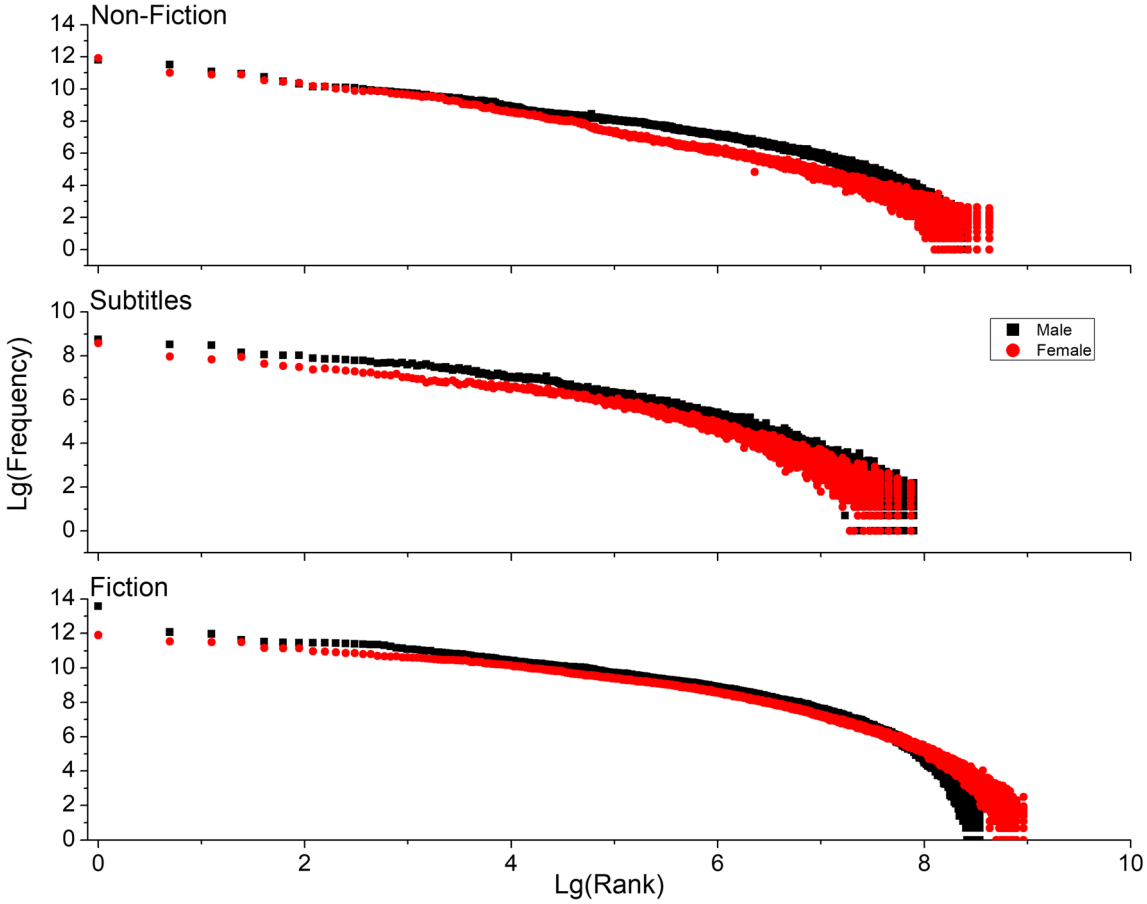

Fig. 3 Collapsed Zipf graphs for the texts from the nonfiction book collection, the film and television subtitles, and the fiction book collection

Two of the genres, however—*travel* and *how-to*—show a smaller bias, with only 52.8% and 53.4%, respectively, of the names in these corpora being male. This suggests that genre-specific characteristics impact name usage, which will be analyzed further for the fiction genres.

This first result indicates that the natural-language environment in works of nonfiction does not provide an accurate portrayal of personal name usage in the real world, especially for more academically oriented texts. For the six academic genres (history, philosophy, psychology, chemistry, physics, and politics), this may be due to male researchers historically being more prevalent in these fields, and thus being discussed more often. A direct test of this reasoning will be given by examining the distributions contained in subtitles and fiction books, since these are not describing real people, but typically, instead, fictional characters constructed to fill a role in a narrative. If this bias is still found in these text sources, it would suggest that there is a general bias in the natural-language environment away from female names.

Subtitles

Figure 4 contains the Zipf scales for the six subtitle genres, and the middle panel of Fig. 3 displays the collapsed scale across all subtitle genres. This figure shows that the distribution of male and female names in the subtitle corpus replicates the results from the nonfiction books, with male names being much more frequent across every genre and at practically each point along every distribution. Overall, approximately 1.1 million personal names are used in the subtitle texts, and approximately 61.5% of these names are male. Given that five of these genres are television shows or movies that feature fictional characters, this invalidates the hypothesis that the personal-name bias seen in the nonfiction books was due to references to historically important figures, who may have tended to be male. Instead, it seems that there is a general preference for using male over female names when constructing fictional characters. The results from the fiction novel collection, by far the largest corpus here, would provide a stronger test of this assumption.

Fiction novels

The Zipf scales for the ten fiction genres are contained in Fig. 5, and the bottom panel of Fig. 3 displays the collapsed scale across all genres of fiction books. Figure 5 shows that seven of the ten genres have the same pattern as the nonfiction books and subtitles (with the exceptions being *romance*, *young*

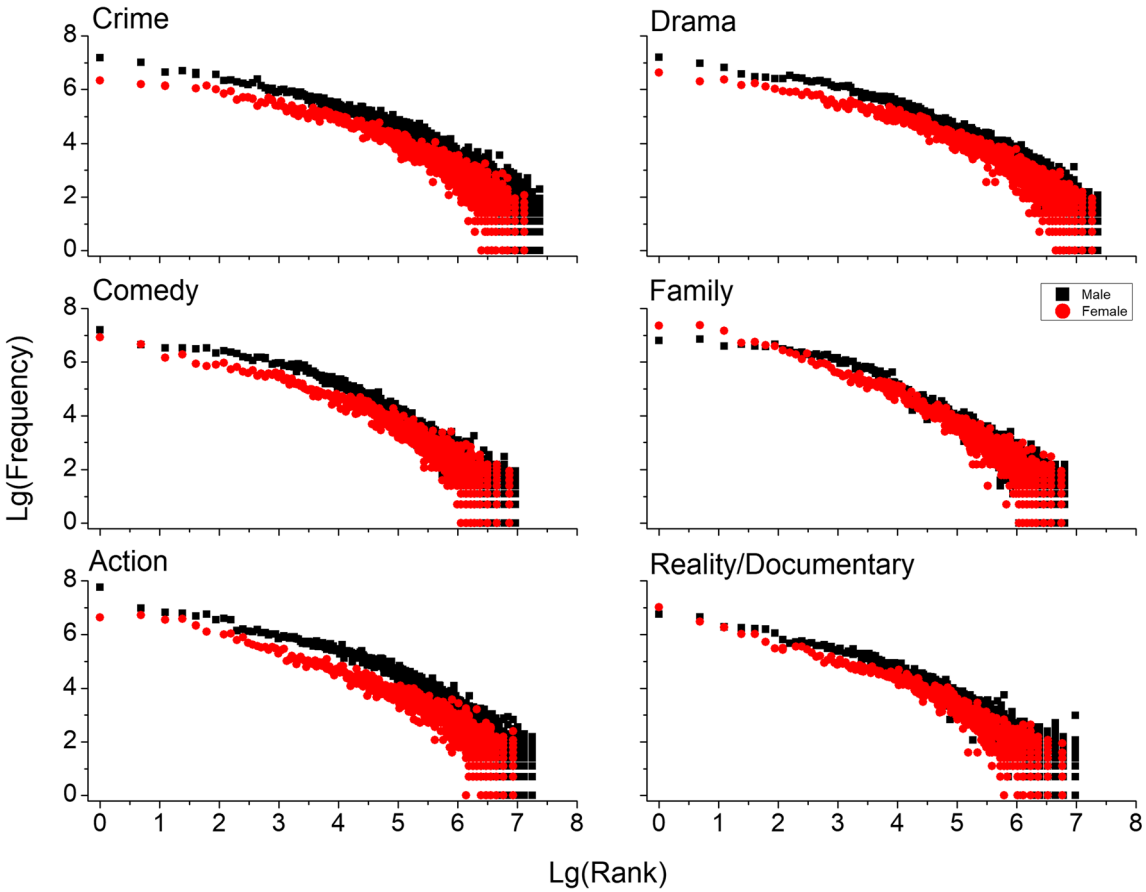

Fig. 4 Zipf scales of male and female name frequencies from the six subtitle genres

adult, and *historical fiction* novels), in which male names are more prevalent than female names across the entire frequency distribution. In this book collection, 58.9% of the names produced are male names. Again, the majority of the genres show a systematic bias toward using male names, similar to the nonfiction books and subtitles.

To sum up the findings of the analyses above, Fig. 6 contains the proportions of male names used across the 24 subcorpora for the three types of texts. This figure shows that for every genre of text, each one contains more male than female names. Using a one-sample *t* test, we confirmed that the proportions of male names used were significantly different from .5 for nonfiction books [$t(7) = 5.166$, $p = .001$], subtitles [$t(5) = 5.67$, $p = .002$], and fiction books [$t(9) = 6.1$, $p < .001$]. This confirms that there is a consistent, and strong, bias toward using male names across the different sources of language.

Author characteristics

Figures 5 and 6 show significant variability in the prominence of male over female names across the fiction genres. For example, three of the genres—*romance*, *young adult*, and *historical fiction* novels—show a much smaller bias toward male names than the other genres. Indeed, male names were used only 52.6% of the time in *romance* novels, 53.6% in *young adult* novels, and 55.9% in *historical fiction* novels, as compared to 65.8% in *thriller* novels, 63.9% for *science fiction* novels, and 67.2% for *horror* novels. We hypothesized that author effects were likely at play to explain why there would be such variability in male name usage across genres.

The effect of author gender on personal name production was the first aspect of the data to be analyzed. Table 2 shows the lexical characteristics of the male and female authors, which shows that there are more female than male authors, but that the male authors have produced slightly more books, leading to the male corpus being larger than the female corpus. However, the female corpus is still of considerable size, consisting of more than 950 million words. Indeed, our sample likely has an overrepresentation of female authors, as recent research has shown that in 1950 only 25% of fiction books were written by female authors, down from 50% in 1850 (Underwood, Bamman, & Lee, 2018). The totals in Table 2 are slightly higher than those in Table 1, because this sample also includes nonfiction books from 374 of the fiction authors (169 female and 205 male authors).

To examine personal name production by author gender, Fig. 7 contains the Zipf scales for all the books produced by

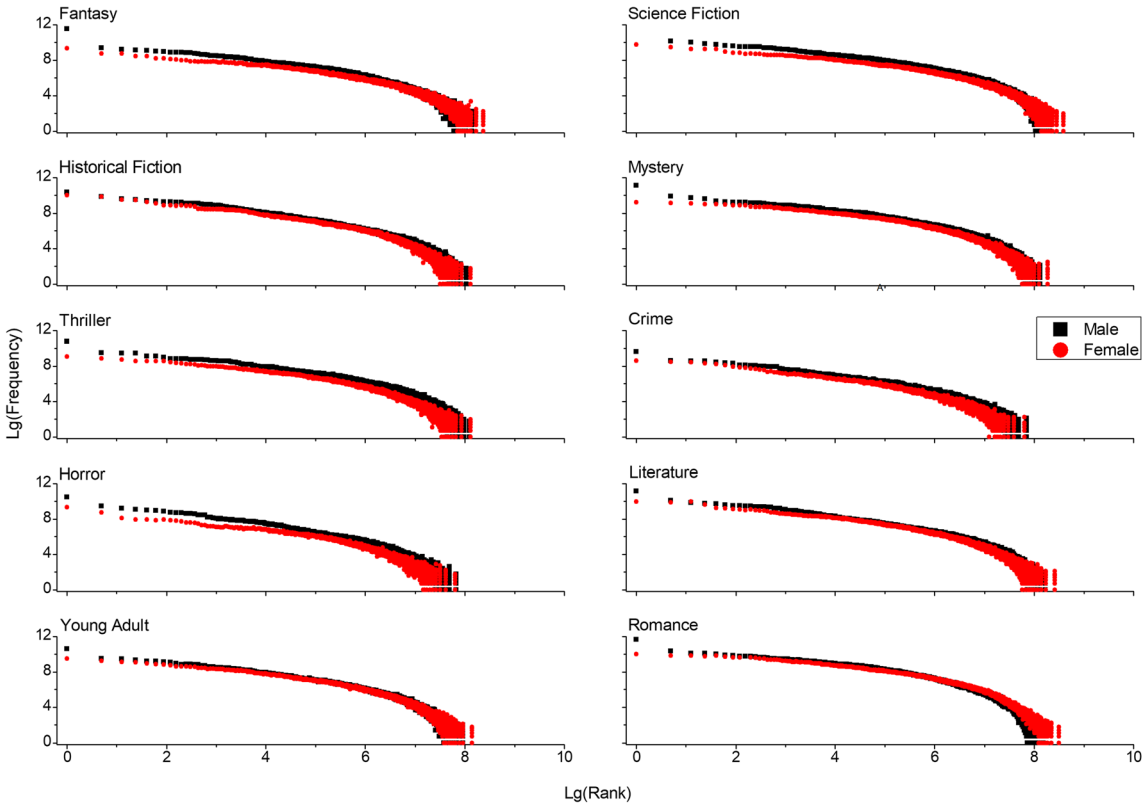

Fig. 5 Zipf scales of male and female name frequencies from the ten fiction genres. All genres except for the romance, historical fiction, and young adult genres show the same bias toward male over female names that was seen in the nonfiction and subtitle corpora

male authors (top panel) and all the books produced by female authors (bottom panel). This figure shows that the bias toward using male personal names is caused mainly by male authors, who have a much greater usage of male names across the frequency spectrum. Female authors still have a bias toward male names, but it is much smaller, similar to the distributions

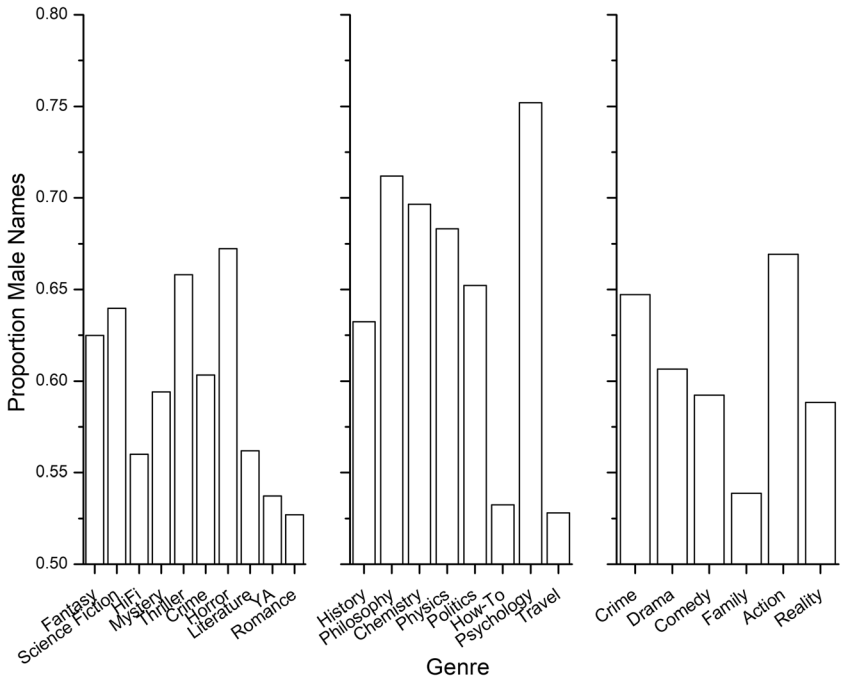

Fig. 6 Proportions of male names used across the 24 subcorpora. This figure shows that for all genres analyzed, a greater proportion of male names are used

Table 2 Characteristics of language sources by author gender

Author Gender	# of Authors	# of Books	# of Words
Female	1,696	13,022	983,793,277
Male	1,512	13,618	1,111,014,271

from the *romance* and *young adult* genres. Male authors produced approximately 21 million names, with 65.6% of these names being male. Female authors produced over 23 million names, of which only 52.9% were male names, demonstrating a large gender discrepancy in the usage of personal names. Given that the male corpus consists of almost 150 million more words than the female corpus, this also demonstrates that female authors tend to use personal names to a greater degree in their writings.

The gender difference in personal name usage also explains the lack of a major bias in the *young adult*, *historical fiction*, and *romance* genres, as well as the greater percentage of male names in the *thriller*, *science fiction*, and *horror* novels. For the *romance* genre, 98.8% of the authors were female; for the *young adult* books, 70.1% of the authors were female; and 64.8% of the authors of *historical fiction* novels were female. Comparatively, 86.6% of the *thriller*, 78.7% of the *science fiction*, and 88.3% of *horror* authors were male. To assess to what degree the difference in personal name usage across the different genres could be explained by the difference in author gender, the proportion of male names used in a genre (data contained in Fig. 6) was correlated with the proportion of male authors in that genre. A rank correlation found a very strong relationship between these variables, $r(9) = .94$, $p < .001$, signaling that when a genre has a greater degree of male authorship, there is a corresponding increase in the frequency of male names. When a genre is mainly written by female authors, there is a much more egalitarian use of male and female names.

An additional aspect of the data that needs to be understood is the extent to which the bias toward male names is a historical artifact, or whether it holds for modern authors as well. To answer this question, the dates of birth of as many authors as possible were recorded. As we previously stated, a total of 2,090 authors had a date of birth publicly available. To visualize how personal name usage was changing as a function of author date of birth, the proportion of male names used was recorded for each individual author. This proportion was then plotted against the author's date of birth, split by the gender of the author, in a scatterplot. The plot is contained in Fig. 8 and shows that there has been very little change in the personal name bias across time. That is, the proportions of male names produced are stable across time for both male and female authors, with the larger bias toward male names for male authors holding across time. We observed no significant correlation between author date of birth and the proportion of male names used by male authors [$r(1, 239) = -.001$, n.s.] or female

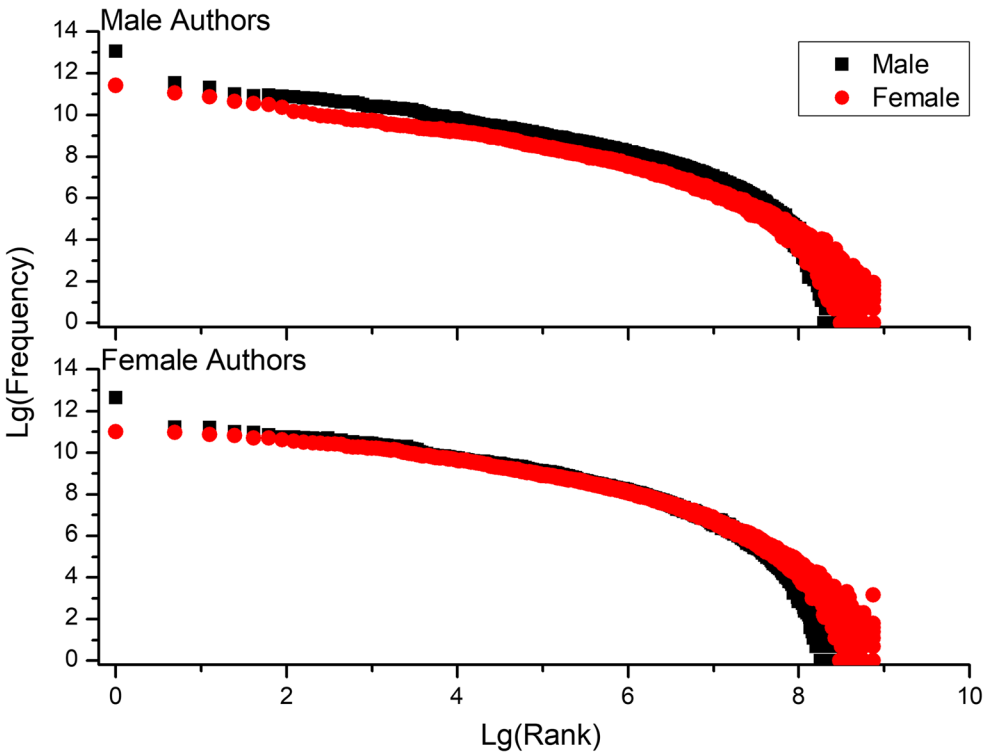

Fig. 7 Zipf scales from the writings of male and female authors. This figure shows that most of the bias toward male names is coming from male authors, with female authors showing a much smaller bias

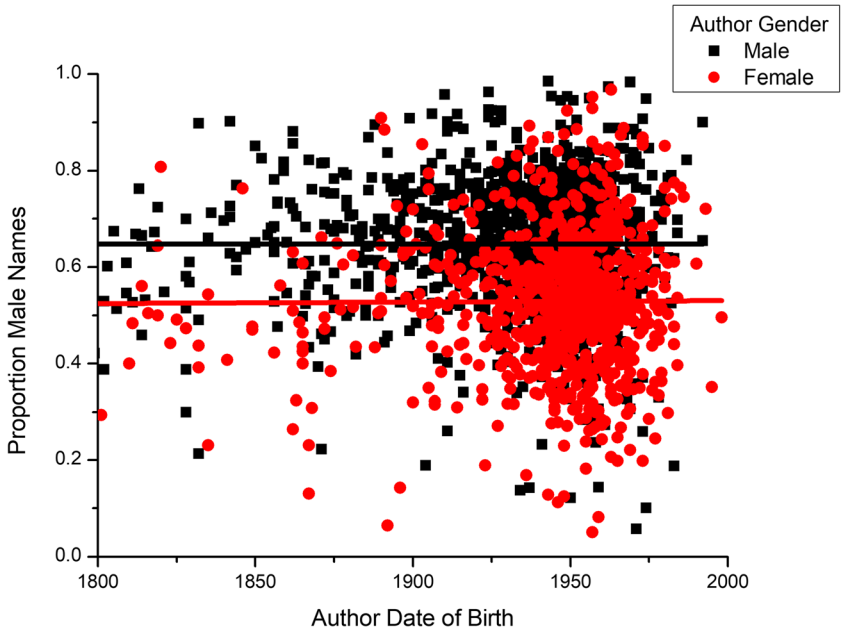

Fig. 8 Proportions of male names used by an author, as a function of the author's date of birth. This figure shows that the usage of male names has not changed across time for either male or female authors. The lines in the figure are linear trend lines

authors [$r(851) = .015$, n.s.]. Overall, this result indicates that the preference for male names in written language has seen very little change over the last 200 years, with modern authors showing the same bias toward using male names as authors born centuries earlier.

A question not yet addressed is whether an author's country of birth has an effect on the gender name bias seen in the above analyses. Answering this question would assess whether there are cultural differences in the usage of male versus female names in English-speaking nations. In our sample, most of the authors were from one of the following four countries: (1) the United States, (2) the United Kingdom, (3) Canada, and (4) Australia. The corpora were thus split for these four countries, by male and female authors, to determine whether the overall preference for male names is consistent across these countries, and also that the lesser preference for male names for female authors is also present. Table 3 contains the characteristics of these corpora and shows that each of the four countries has a sizeable representation in terms of corpus size. Authors from countries other than the USA were not excluded previously because it is not clear how much the name distributions of people in these countries would differ, since data equivalent to the SSA data are not available for them.

The Zipf scales for male and female authors across the four countries are shown in Fig. 9. The scales show a remarkable consistency: Across all four countries, the male authors have a large preference for male names, whereas female authors have a much more equal usage of male and female personal names. This demonstrates that the gender bias toward using male personal names in the natural-language environment is not isolated to one country, but is ubiquitous across multiple English-speaking countries. Additionally, the finding that the female authors have a smaller bias was also found to be consistent across these countries. Overall, name usage across these countries seems equivalent, with authors born in different countries having roughly equivalent biases.

Table 3 Characteristics of language sources split by author gender and author place of birth

Author Place of Birth	Author Gender	# of Authors	# of Books	# of Words
USA	Female	1,119	8,735	652,904,636
	Male	879	8,632	679,389,049
UK	Female	322	2,596	215,741,498
	Male	416	3,575	289,495,440
Canada	Female	69	448	29,414,707
	Male	45	347	31,811,458
Australia	Female	85	552	41,806,687
	Male	51	313	22,063,006

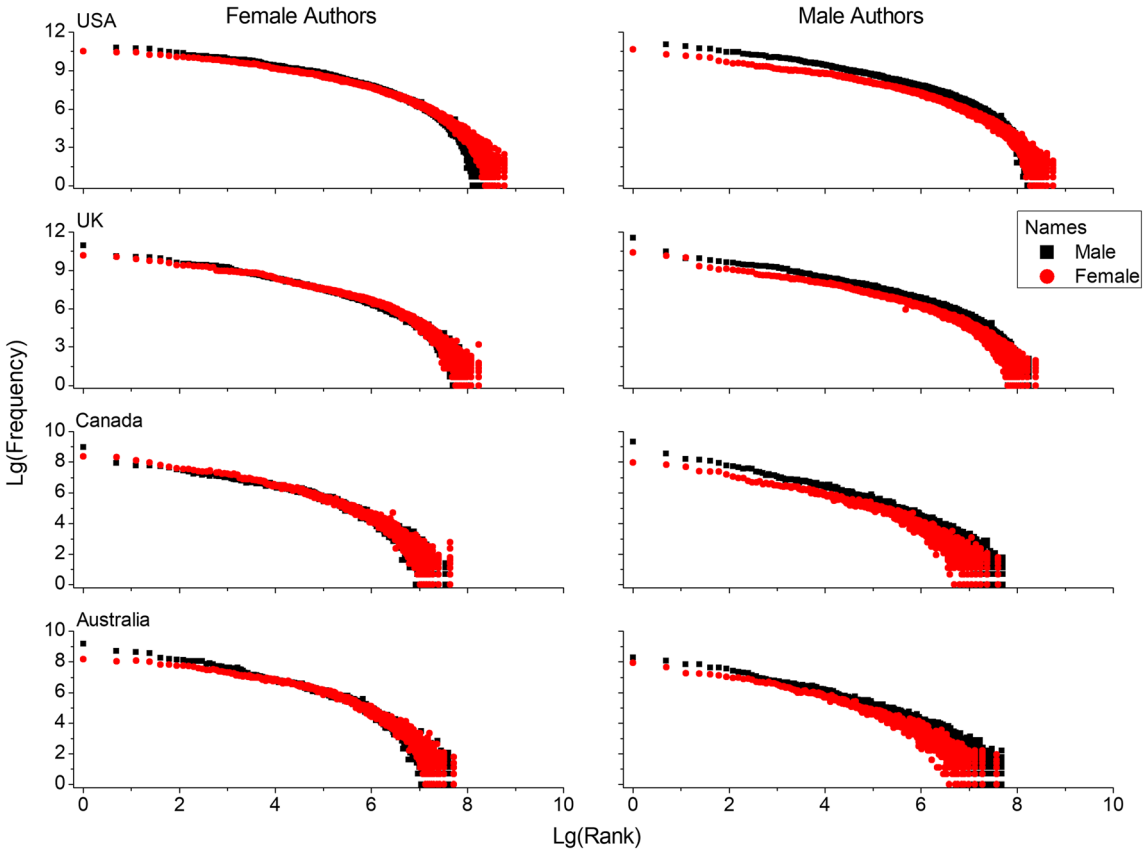

Fig. 9 Zipf scales of male and female name frequency for male and female authors across four countries of birth

So far, the comparison between the prevalence of male and female names has focused on the total number of tokens of names, but not on the number of types. As we discussed in the Method section, there are more female than male names in the SSA data (see Dye et al., 2016), by a ratio of about 2:1 in the sample of names used here. This can be seen in Fig. 1, where the SSA data have a longer tail for female names. However, this same tail is not seen in the various Zipf scales from the text sources, suggesting that in the natural-language environment, the number of female name types is vastly underrepresented. Figure 10 contains the ratios of female to male name types in the subcorpora across the nonfiction, subtitle, and fiction corpora. This figure shows that the various subcorpora have mostly equal ratios of female and male types, which is not reflected in the SSA data. This finding suggests that not only are female names underrepresented in terms of overall frequency, but also that many female names are not represented in the natural-language environment at all.

Word frequency vs. contextual diversity

So far, word frequency (WF) has been the only variable used to examine the occurrence patterns of personal names. Recent research in lexical organization has pointed to an alternative form of counting word occurrences, namely contextual diversity (CD), as providing an account of lexical behaviors superior to that derived from word frequency counts (Adelman et al., 2006; Brybaert & New, 2009; Johns, Gruenenfelder, Pisoni, & Jones, 2012a; for a review, see Jones et al., 2017). To calculate the CD of a word, repetitions within a context are ignored. Thus, a CD count measures how many different contexts a word occurs in, not the total number of times that a word occurs. Context is defined differently depending on the corpus, but it is typically assessed at the paragraph or document level.

Here context will be considered at the book level. By assessing CD at the book level for personal names, it provides insight into how contextually dependent name usage is, an issue we discussed in the Method section of this article. One characteristic of CD is that it ignores the contribution of word *burstiness* (see Altmann, Pierrehumbert, & Motter, 2009), which is the finding that if a word occurs once, it is more likely to be used many more times in that context. CD ignores this phenomenon by only counting one occurrence. Names at the book level are likely to look very bursty, since character names likely differ significantly across books, leading to some names only occurring in very few books, but they still could have relatively high frequency values if the names are used frequently in that small number of books. For example, if *Jennifer* is the main character of three books but is not a

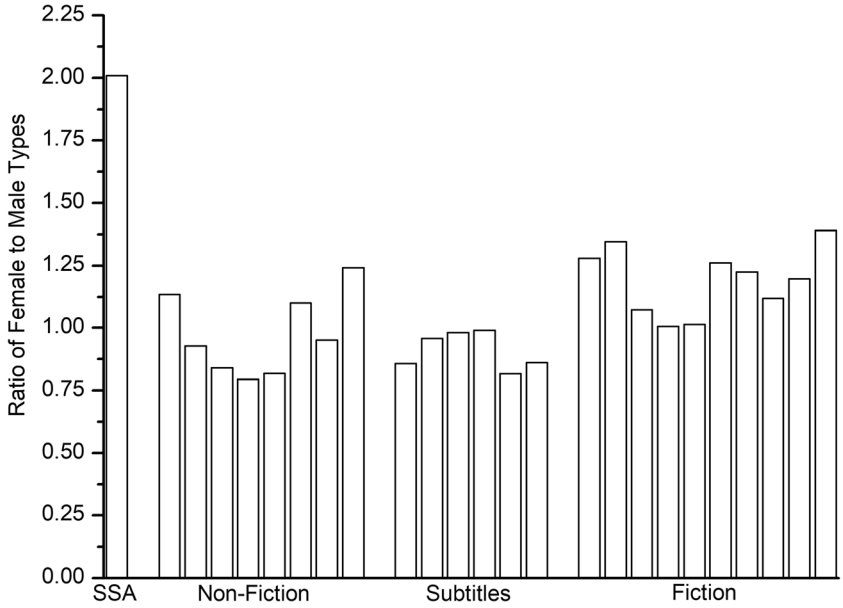

Fig. 10 Ratios of female to male name types for both the SSA data and the various subcorpora

character in any other book, *Jennifer* would have a relatively high WF but a low CD. Thus, it is possible that the frequency differences seen so far between the usage of male and female names were due to different contextual usages of names, with male names being more bursty than female names but having the same number of contextual usages.

To illustrate the contextual dependency of personal names, consider Fig. 11. This figure contains the word frequencies of the name *Jennifer* and the word *adult* for all books contained in the author corpus. These words occur approximately equally, with *Jennifer* having a total frequency of 36,862, whereas *adult* has a frequency of 38,348. However, as Fig. 11 shows, the distributions of occurrences are vastly different for these words; *adult* occurs relatively equally across all books, whereas *Jennifer* has bursts of high frequency in a smaller number of books. This results in a large divergence in the CD counts of these words: *Jennifer* occurred in only 2,026 books, whereas *adult* occurred in 13,099 books. This demonstrates that a CD count of names could offer a considerably different distribution of personal name prevalence than a WF count does.

This possibility was tested by calculating CD values from the book corpus split by gender, since this would allow for a

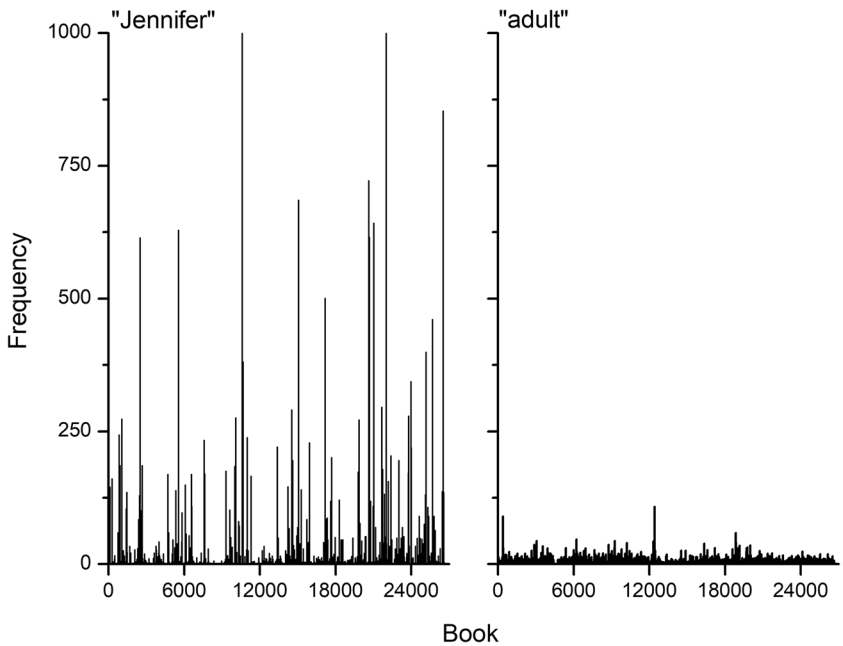

Fig. 11 Example of the burstiness of personal names. The words *Jennifer* and *adult* occur approximately equal numbers of times in the author corpus, but *Jennifer* has a much more contextually dependent usage, in that it occurs in bursts across books, as compared to the word *adult*

determination of whether the author gender differences found previously hold for a CD count as well as for a WF count. CD values were calculated by only increasing a name's count the first time it occurred in a book, with repetitions of a name within a book not increasing a name's count.

To visualize the resulting distributions, Zipf graphs using the split-corpus methodology were used. Unlike in the visualizations of WF we used previously, in which the corpora were split at the sentence level to avoid contextual dependency of name usage within books, for the CD visualizations, the corpora were split at the book level.

Figure 12 contains the results of this analysis for the combined male/female corpus (top panel), the male corpus (middle panel), and the female corpus (bottom panel). This figure shows that the CD count gives a distribution very similar to that from frequency counting, with the main difference being that the graphs are noisier than the WF graphs visualized previously, likely due to the contextual dependency of name usage. Overall, little change in the name bias was apparent from using a CD count. For the combined corpus, 57.6% of all contextual occurrences of personal names were male. For the male corpus, this increased to 61.0%, whereas for the female corpus the percentage dropped to 53.3%, similar to the overall occurrences calculated using frequency.

As is shown in Fig. 11, it is likely that personal name usage is considerably contextually dependent, with names being more bursty than other words. One way to quantitatively test the assumption of the contextual dependency of personal names would be to take the correlation between CD and WF values across names. It is typically the case that WF and CD counts are highly correlated, because it is somewhat rare for words to occur many times in a context, but not across contexts, resulting in WF and CD values being very similar. For example, the correlation between CD and WF (transformed with a logarithm) from the standard SUBTLEX corpus (Brysbaert & New, 2009) for the words contained in the English Lexicon Project (ELP; Balota et al., 2007) is r = .987, p < .001.

Of course, we were using a different corpus here, so we recalculated WF and CD values for the author book corpus using the words from the ELP, with context being defined at the book level (i.e., repetitions of words within a book were ignored). A correlation of r = .97, p < .001, was found, suggesting that counting at the book level does decrease the

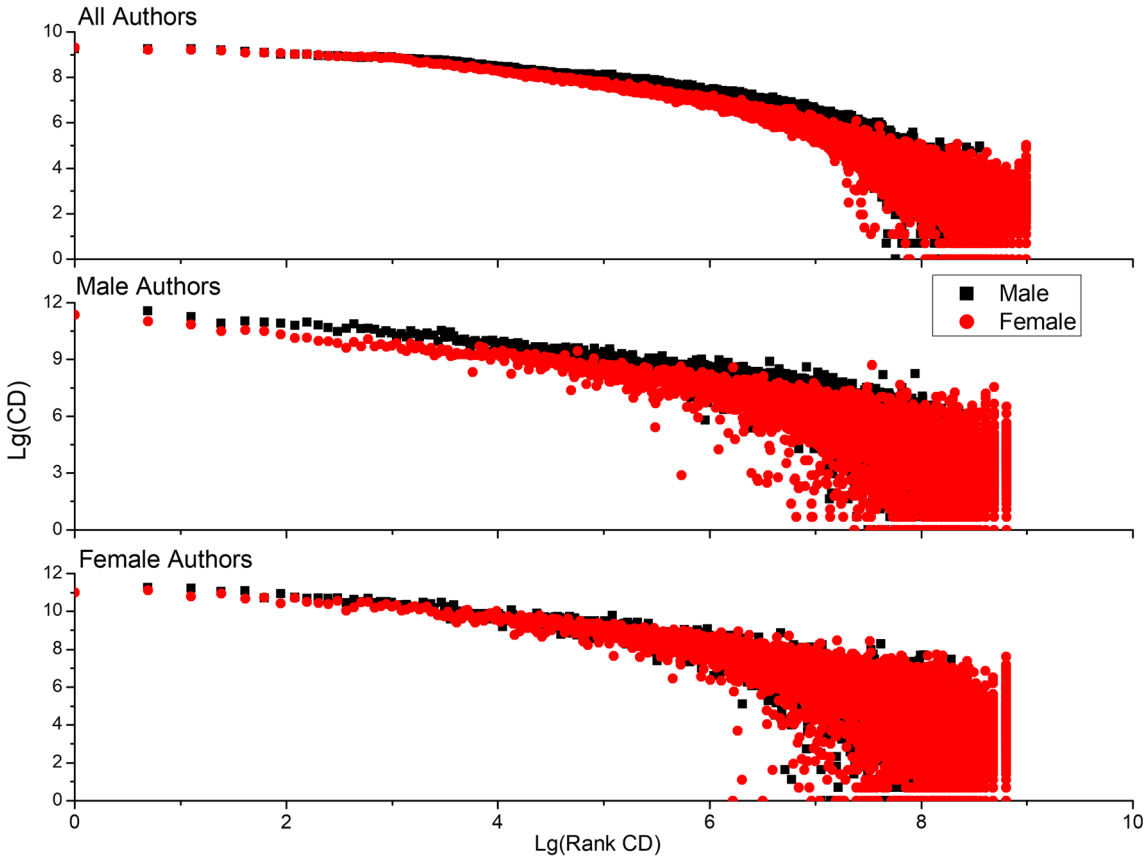

Fig. 12 Zipf scales using a contextual diversity count for the combined author corpus (top panel), the male corpus (middle panel), and the female corpus (bottom panel). This figure shows that a contextual diversity count shows the same trends that a word frequency count does, but produces noisier Zipf scales

correspondence between WF and CD somewhat, but the two measures are still very similar.

To calculate the correlation between WF and CD for names, only names that occurred in the corpus were included (i.e., names that had a WF and CD of 0 were not included in the correlation). For the combined male/female corpus, the correlation between WF and CD for female names was r = .85, p < .001, whereas for male names it was r = .91, p < .001. For the male author corpus, the correlation between WF and CD for female names was r = .84, p < .001, whereas for male names it was r = .9, p < .001. For the female author corpus, the correlation between WF and CD for female names was r = .82, p < .001, whereas for male names it was r = .87, p < .001.

These correlations suggest that personal names, as compared to other words, are relatively more contextually dependent, meaning that names can occur highly within a context but not across contexts. This results in a relatively greater disparity between the WF and CD counts for personal names. Additionally, female names have a larger disconnect, suggesting that female names tend to be more bursty within contexts than male names are, resulting in a greater divergence in CD and WF counts for female personal names. There is a small difference between male and female authors, in which female authors have a smaller correlation between WF and CD, suggesting that female authors have a greater tendency to use personal names more frequently within a context. This might be related to the previous finding that female authors have a greater tendency to use personal names overall, with those additional occurrences taking place within a book rather than across books (e.g., female authors use specific character names more often within a book than male authors do).

One aspect of the SSA data that had not yet been tested is whether authors use names in proportion to their real-world frequency, and whether there is a gender difference in this mapping. Given that CD has been shown to be a better predictor of behavioral patterns in lexical organization (e.g., Adelman et al., 2006), it is possible that it could provide a better fit to the SSA data as well. To test this possibility, correlations were calculated between the name counts from the SSA data and the WF and CD measures from the combined male/female author corpus, the male author corpus, and the female author corpus. The SSA counts and the WF and CD values were transformed to a log scale. To enable a fair comparison across the different counts, only names that occurred in each corpus were included in this analysis. This resulted in 4,592 male names and 6,329 female names.

The correlations between the SSA counts and frequency counts are contained in Table 4. This table shows that the WF and CD values from female authors have a closer correspondence to the SSA data for both male and female names, suggesting that female authors produce names closer to the real-world frequency of names than do male authors, across both male and female names. Male authors have a higher correlation to male than to female name counts, suggesting that they have a better understanding of the distribution of male than of female names. Across all three corpora (combined, male, and female), the WF measure has the stronger correlation, suggesting that in terms of accounting for the distribution of personal names in the social environment, it provides a more accurate accounting.

Overall, the analysis of a CD measure of personal names suggests that the bias toward male names still exists at higher units of counting occurrence, with the overall bias toward male name usage still being the result of male authors preferring to use male names. Additionally, we showed that names are more contextually dependent than other words.

Supplementary materials

To aid other researchers who are interested in examining personal name usage, this article is linked to supplementary material that contains the SSA data and the various name counts used in the analyses above. The criteria for inclusion of a name into the supplementary material were relaxed, as compared to those we used in the analyses above, to aid in answering alternative questions (e.g., if one wished to examine the impact of having a very rare name, or the impact of having a more androgynous name). To be included in the supplementary materials, a name just needed to appear 25 times in the SSA for a gender. If a name appeared for both males and females, the name was not excluded. This resulted in 22,845 male names and 38,346 female names. Many of these serve as both real words and personal names, so researchers have to be sure to control the names appropriately. Additionally, the specific names that were used in the analyses above are contained in the supplementary materials as well, to allow for replication of the results in this article. The total name frequency counts from the nonfiction, subtitle, and fiction corpora were included. Additionally, the female and male author counts were included, as well as the counts split by author place of birth.

Table 4 Correlations of name counts from the SSA data to the frequency data from the female and male corpora

	Female SSA Count	Male SSA Count
Combined WF	.541	.552
Combined CD	.484	.524
Male WF	.477	.509
Male CD	.421	.491
Female WF	.551	.554
Female CD	.523	.535

All correlations are significant at the p < .001 level. WF, word frequency; CD, contextual diversity.

Finally, the CD measures from the author corpus were also included.

General discussion

This article has demonstrated a large and persistent gender bias in the natural-language environment for the usage of personal names. Specifically, it was found that male personal names are used at a much greater rate than female personal names. This bias was shown to hold across nonfiction books, television and film subtitles, and fiction novels. Across these three sources of texts, there were 24 different genres, and each one contained more male than female names. When the collection of fiction novels used in this analysis was organized by author gender, date of birth, and place of birth, the underlying cause of this bias was determined. We found that the majority of the bias toward using male names came from male authors, with female authors showing a much smaller bias. We also showed that this bias has not changed over the last 200 years and is consistent across authors from multiple English-speaking nations. Additionally, the greater prevalence of male names was also confirmed using a contextual diversity count, demonstrating that the bias toward male names holds at different levels of counting word occurrences.

The results contained in this article point to the continued power of big-data analyses of behavioral data, and in this particular case, large natural-language corpora. Not only can these corpora be used to develop new models of language processing (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Jamieson, Avery, Johns, & Jones, 2018; Johns & Jones, 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997), but can also serve to examine large scale trends in human behavior (e.g., Johns & Jamieson, 2018). Language is a central organizer of human cognition (Brysbaert et al., 2018; Johns, Jones, & Mewhort, 2012b; Jones et al., 2017), and a large percentage of our everyday experience consists of linguistic stimuli, with a typical human being reading millions of words per year (Brysbaert, Stevens, Mandera, & Keuleers, 2016). Thus, any biases away from reality in natural language can have a major impact on a person's lexical behavior.

A good example of the impact of frequency biases is given by the availability heuristic (Tversky & Kahneman, 1973), which reflects the finding that people depend on base rate information when making decisions, even when more informative data are available. A prominent example of this phenomenon is given by Combs and Slovic (1979), who demonstrated that newspaper articles overrepresent certain causes of death (e.g., murders) and underrepresent others (e.g., lung cancer), and found that subjects made a corresponding error in judgment about the likelihood of someone dying from those events. This suggests that the information that is present in natural language has consequences for the decisions that people make in everyday life. Given that personal names are a fundamental aspect of human life, and that the use of personal names is biased toward males, this suggests that general decision-making biases may favor males, for no other reason than that male names are more frequent in everyday experiences with natural language.

Studies using traditional methodologies have typically found very small, or nonexistent, gender differences in language-processing or verbal abilities using behavioral and neuroimaging paradigms (for reviews, see Hyde & Linn, 1988; Wallentin, 2009). However, the present study shows that big-data methodologies can show language differences on a scale that is not possible with traditional approaches. For example, Johns and Jamieson (2018) recently showed that much of the variability in written language lies at the level of the individual author, signaling that different individuals can use language in very different ways. Schwartz et al. (2013) used Facebook messages to analyze how language usage differs by the personality of an individual. Park et al. (2016) also used Facebook messages to demonstrate gender differences in the usage of language. Taken together, these studies demonstrate the promise of big-data methodologies to understanding diverse problems in natural-language processing and usage.

This work has a number of theoretical implications for models of lexical organization. Although models differ in their mechanisms (e.g., Goldinger, 1998; Jones et al., 2012; Murray & Forster, 2004; Norris, 2006), all models of lexical organization are organized around environmental occurrence of word forms (e.g., word frequency, contextual diversity, or some related measure). The predictions of these models is that words that are experienced more often should be more available in the lexicon. In receptive tasks, this emerges as lower response times for highly frequent words. In expressive tasks, this entails that high-frequency words should, in turn, be produced at a greater rate. The results of this article, and especially the results in Fig. 8, show that there has been little change in the bias toward male names across time, demonstrating that these predictions seem to be borne out in the usage of personal names at scale. If people produce and understand language on the basis of how they experience it, they should then embody the same biases that language users of a previous generation did. The results of Fig. 8 show exactly this, that there has been no change in the bias toward male names over the past 200 years.

By assessing the structure of the language environment at a large scale, it is possible to make predictions about how certain lexical classes (such as names) should be organized in an individual's mental lexicon based on a person's exposure to different types of lexical material. To make this claim more explicitly, we ran a Monte Carlo simulation in which the proportion of male names contained in a set of books was calculated. The book sets were composed of 1,000 books and were sampled at 10%, 30%, 50%, 70%, and 90% female authorship. To simulate the previous results to a point, 5,000

resamples were done of the book sets at each level, and the mean proportion of male names in the book set was calculated. Figure 13 contains the results of this simulation, which shows that the bias toward male names decreases substantially as a function of the number of female authors that the book set contains. This shows that it is possible to make explicit predictions about lexical organization based on one's reading habits. For example, if a person were a heavy reader of *romance* novels, that person would likely have a very different distribution of name frequencies than someone who only read *thriller* novels. To continue to develop and refine models of lexical organization, it will be essential to show how these models scale, to capture large-scale trends in human behavior and provide parsimonious explanations of that behavior.

An additional theoretical issue that this article raises is how different sources of statistical information are balanced. This article shows that lexical experience is biased away from using female names. However, presumably the social environment is not biased in such a fashion—people likely meet equal numbers of males and females in their everyday life. Given the bandwidth of natural language, it is likely that people have more experience with names through books, television shows, or movies then they do through social experience. These experiences may not be weighted equally, however, with names of spouses, parents, and friends likely being more important to a person than a character from a book or film. How these different statistical sources are used and integrated in our lexical organization system is an important area that needs to be understood.

A different theoretical contribution of this work is the finding of the contextual dependency of personal names. The influence of context on lexical organization and word learning is a burgeoning research area (e.g., Adelman et al., 2006; Hsiao & Nation, 2018; Johns, Dye, & Jones, 2016a; Johns, Gruenenfelder, Pisoni, & Jones, 2012a; Johns, Sheppard, Jones, & Taler, 2016b; Jones et al., 2012; Joseph & Nation, 2018; Rosa, Tapia, & Perea, 2017; Vergara-Martínez, Comesaña, & Perea, 2017). To empirically validate contextual approaches to lexical organization, it has been necessary to devise artificial language experiments (e.g., Jones et al., 2012) or highly controlled natural-language experiments (e.g., Johns, Dye, & Jones, 2016a), due to the highly correlated nature of word frequency and context counts. As has been shown here, personal names provide a stimuli type that have a natural divergence between overall frequency and contextual occurrence, as is demonstrated in Fig. 11. The materials contained in the supplementary materials of this article provide an ability to construct these comparisons, which will hopefully lead to a better understanding of how frequency and context interact to organize the mental lexicon.

This work also has implications for distributional models of language (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Johns & Jones, 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997), which aim to build knowledge from the statistical structure of natural language. In the study that has defined this field, Landauer and Dumais proposed that this class of models offers a solution to many problems in the field of knowledge acquisition. However, the present article, along

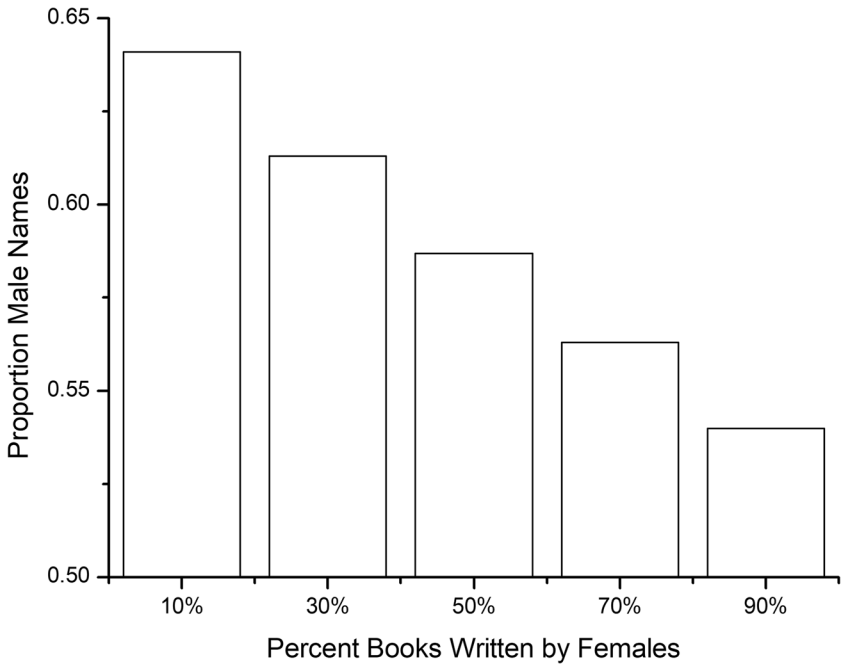

Fig. 13 Monte Carlo simulation of the effect of reading a greater percentage of female authors on the prevalence of male names. As one reads more female authors, the usage of male and female names becomes more egalitarian

with others (e.g., Caliskan et al., 2017), presents clear evidence that natural language is biased in systematic ways. This entails that the representations that a distributional model forms when learning from large text bases will also be biased. If one wishes to build a model that does not have such biases, it may be necessary to have a deliberately curated corpus (see Johns, Jones, & Mewhort, in press, for a promising method to accomplish this).

Much of language is an accumulation of cultural evolution (Christiansen & Chater, 2008). People are exposed to the beliefs of language users of a previous generation, who were exposed to the beliefs of their preceding generation. The accumulation of these beliefs likely result in systematic biases being embedded in the statistical structure of our natural-language environment. As big-data approaches to cognition continue to develop, and as the collection and curation of natural-language sources continues, there is going to be a continued opportunity to reveal and understand the systematic biases that are a part of human experience.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision time. *Psychological Science*, *17*, 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

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. https://doi.org/10.1371/journal.pone.0007678

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

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

Barres, B. A. (2006). Does gender matter? *Nature*, *442*, 133–136.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*, 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many 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. https://doi.org/10.3389/fpsyg.2016.01116

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–508, disc. 509–558. https://doi.org/10.1017/S0140525X08004998

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly*, *56*, 837–849.

Dye, M., Johns, B. T., Jones, M. N., & Ramscar, M. (2016). The structure of names in memory: Deviations from uniform entropy impair memory for linguistic sequences. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 1763–1768). Austin, TX: Cognitive Science Society.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*, 788–791.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Gregg, V. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), Recall and recognition (pp. 183–216). Oxford, UK: Wiley.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244. https://doi.org/10.1037/0033-295X.114.2.211

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.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69. https://doi.org/10.1037/0033-2909.104.1.53

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*, 119–136. https://doi.org/10.1007/s42113-018-0008-2

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

Johns, B. T., Dye, M., & Jones, M. N. (2016a). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *4*, 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

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

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. (2012b). A synchronization account of false recognition. *Cognitive Psychology*, *65*, 486–518.

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

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*, 703:1–14. https://doi.org/10.3389/fpsyg.2016.00703

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

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

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37. https://doi.org/10.1037/0033-295X.114.1.1

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, NY: Oxford University Press.

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, *166*, 190–211.

Kilbourne, B. S., England, P., Farkas, G., Beron, K., & Weir, D. (1994). Returns to skill, compensating differentials, and gender bias: Effects of occupational characteristics on the wages of white women and men. *American Journal of Sociology*, *100*, 689–719.

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

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178. https://doi.org/10.1037/h0027366

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*, 721–756. https://doi.org/10.1037/0033-295X.111.3.721

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357. https://doi.org/10.1037/0033-295X.113.2.327

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., . . . Seligman, M. E. (2016). Women are warmer but no less assertive than men: gender and language on Facebook. *PLoS ONE*, *11*, e0155885. https://doi.org/10.1371/journal.pone.0155885

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*, 1112–1130.

Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS ONE*, *12*, e0179004. https://doi.org/10.1371/journal.pone.0179004

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17. https://doi.org/10.1037/0096-1523.3.1.1

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*, e73791. https://doi.org/10.1371/journal.pone.0073791

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. *Journal of Cultural Analytics*. Advance online publication. https://doi.org/10.22148/16.019

Vergara-Martínez, M., Comesaña, M., & Perea, M. (2017). The ERP signature of the contextual diversity effect in visual word recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *17*, 461–474. https://doi.org/10.3758/s13415-016-0491-7

Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: A critical review. *Brain and Language*, *108*, 175–183.

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *387*, 341–343.

Zipf, G. K. (1935). The psycho-biology of language: An introduction to dynamic philology. Boston, MA: Houghton Mifflin.

Zipf, G. K. (1949). Human behaviour and the principle of least-effort. Cambridge, MA: Addison-Wesley.