Journal of Data Science xx (xx), 1–19
July 2020

DOI: xx.xxxx/xxxxxxxxx

Supplementary Material for "Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing"

Mayank Kejriwal*[1] and Akarsh Nagaraj[1]

[1] *Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Marina del Rey, CA, United States of America*

This supplementary material includes details on data preprocessing, character extraction and gender classification in Section 1. Complete statistical significance results for Hypothesis 1 are provided in Section 2. Additional details and precise quantitative estimates for Fig 2 in the main text (exploring the results of Hypothesis 2) are provided in Section 3, and a full set of statistical significance results for Hypothesis 2 are provided in Section 4. Quantitative linear regression results (including supporting statistics such as the goodness of fit or R^2) underlying Hypothesis 3 are provided in Section 5. Secondary analyses (including methodology and results) that use computational tools from NLP to make a qualitative assessments of the kinds of words associated with male and female character occurrences are provided in Section 6. We also provide a more detailed analysis of limitations of the study in Section 7. Code for reproducing the raw data, and replicate the analyses is available as separate supplementary material, along with Excel workbooks that were used for generating the results and statistical significance tests for Hypotheses 1-3 in the main text.

1 Data Preprocessing, Character Extraction and Gender Classification

An important contribution of this study is the quantification of gender-specific character prevalence in literature within a sufficiently broad corpus. Since manually extracting characters and character occurrences from a corpus of 2,443 texts is not feasible, we propose to use high-performance NLP methods to extract characters in various ways, and to automatically classify whether they are male or female. Once extracted, and tagged with gender, robust analysis of character prevalence becomes feasible.

The primary NLP method that we rely on to extract characters is Named Entity Recognition (NER) (Nadeau and Sekine, 2007), which also goes by *named entity identification, entity chunking,* and *entity extraction* in the literature (Sun et al., 2002; Sassano and Utsuro, 2000; Daiber et al., 2013). NER is a specific type of the broader *information extraction* problem that has witnessed considerable advances in recent years, including for domain-specific corpora (Han and Wang, 2021; Kejriwal and Szekely, 2017; Kejriwal, 2019; Kejriwal et al., 2019). NER aims to locate and classify named entities mentioned in natural language text into pre-defined categories, such as person-names, organizations, locations, and even monetary values. In this study, the primary motivation behind using NER is to extract person-names from each text.

In order to apply NER, the input text first needs to be split into sentences. Identifying sentences is important because they form logical units of thought and represent the 'borders' of many grammatical effects. To do so, we employ *sentence segmentation* as the first pre-processing step. Sentence segmentation is an important early step in many NLP pipelines (Palmer, 2000).

[*] Corresponding author. Email: kejriwal@isi.edu

Received April, 2020; Accepted May, 2020

For example, character extraction (an instance of NER) algorithms tend to be more accurate and efficient when they are executed on shorter self-contained spans of text (such as sentences) rather than on entire corpora. Despite its seeming simplicity, a generalizable implementation of sentence segmentation is non-trivial for various language-specific reasons. One example is the use of periods in abbreviations and numbers, in addition to its more common use at the end of sentences.

Recent NLP software packages have achieved impressive results in a variety of tasks, including sentence segmentation, primarily due to the advent of deep learning and maturity of 'language representation' models (Devlin et al., 2018). For our purposes, we used the sentence segmentation module from a Python library called SegTok (Leitner, 2015), developed to process orthographically regular Germanic languages, of which English is an example. SegTok is capable of identifying sentence terminals such as '.', '?' and '!' and *disambiguating* them when they appear in the middle of a sentence (like in the case of abbreviations and website links), which significantly reduces the probability that a sentence is segmented before it has truly concluded. After executing SegTok on each book in our corpus, we also manually assessed its performance by sampling 'challenging' sentences (that contained inconsistent sentence terminals, including periods in the middle of the sentence, as in the cases noted above) and verifying that the full sentence was correctly segmented by the software.

We also conducted a small, but formal, evaluation whereby we randomly sampling 110 sentence outputs that were segmented, and manually tagging them as being correctly segmented with respect to the paragraph in which the sentence was originally embedded. We found that, of these 110 sentences, only two were incorrectly segmented, yielding an accuracy of 98.18%. Of the two sentences that were incorrectly segmented, the error was minor. In both cases, there were words in all caps either at the beginning or end of the sentence. This may have been due to (for example) a chapter header, or slight formatting discrepancies in the underlying corpus itself. In either case, the 'correct' segmentation was always a subset of the actual output.

For these reasons, we selected SegTok as our segmentation package of choice. However, there are also other viable alternatives that future research could consider, especially if SegTok is found not to be as high-performing for other languages. One such package that is also well-known in the NLP literature is *PunktSentenceTokenizer* (Natural Language Tool Kit Team, 2022b), which is part of the Natural Language Tool Kit (NLTK) (Loper and Bird, 2002), a well-known suite of libraries and packages for symbolic and statistical natural language tasks like text classification, tokenization, stemming, tagging, and parsing. In our early assessment of PunktSentenceTokenizer, we found that it was incorrectly splitting sentences by abbreviation-periods in some of our sample texts. However, we did not conduct a systematic evaluation of the package, and there is a possibility that, in a broader evaluation, it could outperform SegTok. Because the performance of SegTok was already found to be quite good on our random sample, and we also found it to be reasonably efficient and easy to use, we selected it for the purposes of this study.

1.1 Character Extraction, Disambiguation and Gender Classification

In order to measure gender-specific character prevalence, which is required for all three of our hypotheses, we need to count the numbers of male and female characters in each of the books in the corpus. As noted earlier, extracting person names from the text is an instance of the NER problem. Similar to the methodology for selecting a viable sentence segmentation module, we compared the performance of two popular NER libraries in the NLP community - *SpaCy*

(Vasiliev, 2020; Explosion.ai, 2022), an industrial-scale open-source software library for advanced NLP, typically relying on neural network models for part-of-speech tagging, dependency parsing, text categorization and NER; and *NE_ Chunk* (Natural Language Tool Kit Team, 2022a), which is also part of NLTK. Named entity chunking extracts 'chunks' (akin to phrases) from sentences and assigns them semantic tags, such as person, location and organization. We found that NE_Chunk consistently outperformed SpaCy by achieving near 100% precision in extracting characters ('person' entities) from a sample set of books. Specifically, we randomly chose four books and manually identified a total of 72 main characters in those books. We then compared these 72 'gold standard' characters to the ones extracted by NE_Chunk. We found that only one (male) character was not correctly extracted. Hence, we chose it as the character extraction tool.

However, a similar caveat applied as mentioned before for SegTok; namely, in a larger evaluation, a different time-period, or texts of different genres and styles, it may not necessarily be the case that NE_Chunk outperforms SpaCy. Also, SpaCy is customizable, and regularly undergoes updates. Future studies seeking to achieve similar aims should therefore re-evaluate these possible choices in the context of their research needs.

Since characters are independently extracted from each sentence, multiple occurrences of the same character can be extracted across all sentences in a book due to artifacts such as slightly different spelling usage or use of only first or last names, instead of the full name. To discover different occurrences (of the same character), we used a relatively simple Python library called *difflib* (DiffLib, 2022), which provides classes and functions for comparing sequences. Specifically, the *SequenceMatcher* class from the difflib library compares two strings and provides a similarity score between 0 (no match at all) to 1 (complete match, i.e., strings are the same). We used the SequenceMatcher class to perform disambiguation of the characters and link the different versions of the same character together. String pairs with a similarity score of 0.7 or above were treated as duplicates. This threshold was selected after some sampling and manual verification.

This deduplication allows us to count the number of *unique* characters extracted from the text of each book. To assess its accuracy, we randomly sampled 76 character pairs that were disambiguated as duplicates by this heuristic technique, and found that 72 were correctly disambiguated, yielding an accuracy of 94.74%. The errors in disambiguation primarily arose from false positives and mostly involved the names of monarchs e.g., George III and George IV were incorrectly identified as duplicates due to high string similarity.

To tag the genders of the extracted characters as male and female, we used *Gender_ Detector* (PyPi, 2015), a Python library developed using data from the *Global Name Data* project (OpenGenderTracking, 2013), which is able to determine the gender of a character from the first name. Using this library, we were able to heuristically tag the gender of each extracted character. We evaluated the accuracy of this method by randomly sampling 100 extracted characters (50 male, and 50 female, to avoid gender-specific skewness in computing accuracy estimates) and manually checking their gender against the predicted gender. There was only one error, which was due to a female character named 'Captain Leslie' being erroneously tagged as male, yielding an accuracy of 99%.

We end this section with a comment on the choice of using packages, such as NLTK, for the experimental methods in this paper. While such packages are established, and with some tuning, demonstrated high enough performance for the empirical study, even better performance could potentially be achieved in future research (including for more challenging texts, and also texts in other languages) due to the impressive performance of neural transformer-based language models (Devlin et al., 2018). Such language models are now considered state-of-the-art in the

NLP literature and have been applied to an impressive range of problems. A specific example that we cite is the work by Wang et al. (2020), who used it for structured prediction. Other examples include work by Wang et al. (2021); Hong et al. (2021); Schick and Schütze (2020). Many language models are also publicly available in the HuggingFace repository (Wolf et al., 2020). While we do not use language models in this current work, an interesting avenue for future research may be to compare results from these language models to the results reported herein, as a means of both replicating the key findings herein, and achieving greater robustness through an alternative methodological pipeline.

2 Hypothesis 1: Statistical Significance Results

Tables 1, 2, and 3 provide additional statistical testing information on the comparison between female-specific character mentions and male-specific character mentions across all texts in the corpus.

Table 1: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 1. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence is zero, when using the character count measure. The research hypothesis is that male-specific character count is higher than female-specific character count; hence, the one-sided P value was reported in the main text.

	Male Character Count	**Female Character Count**
Mean	34.51289398	9.654523127
Variance	1746.860092	172.5366165
Observations	2443	2443
Pearson Correlation	0.809428273	
Hypothesized Mean Difference	0	
df	2442	
t Stat	38.27179386	
P(T≤t) one-tail	8.9741E-252	
t Critical one-tail	1.645477849	
P(T≤t) two-tail	1.7948E-251	
t Critical two-tail	1.960935904	

3 Hypothesis 2: Additional Statistical Details on Reported Means

Table 4 provides additional statistical details on the means reported in Fig 2 in the main text, using the Student's t-test without assuming equal variance. As noted in the main text, female-specific character prevalence (using the Hypothesis 1 definition of prevalence as means) in female-authored texts was found to be significantly lower than male-specific character prevalence, although the magnitude of the difference was smaller than in male-authored texts. The significance of the effect (P value) is also diminished by more than two orders of magnitude, and in the case of the Pronoun Count measure, is only moderately significant (and with the Bonferroni correction, would become insignificant even at the 90% confidence level). To conclude, male-specific

Table 2: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 1. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence is zero, when using the character occurrence count measure. The research hypothesis is that male-specific character occurrence count is higher than female-specific character occurrence count; hence, the one-sided P value was reported in the main text.

	Male Character Occurrence Count	**Female Character Occurrence Count**
Mean	158.8178469	49.50961932
Variance	37960.82553	7901.734858
Observations	2443	2443
Pearson Correlation	0.495011658	
Hypothesized Mean Difference	0	
df	2442	
t Stat	31.88246232	
P(T≤t) one-tail	4.2584E-187	
t Critical one-tail	1.645477849	
P(T≤t) two-tail	8.5168E-187	
t Critical two-tail	1.960935904	

character mentions increase in female-authored texts (compared with male-authored texts), but female-specific character mentions increases much more.

4 Hypothesis 2: Statistical Significance Results

Tables 5, 6, and 7 provide additional statistical testing information on the comparison of the proportion of female-specific character mentions in male-authored versus female-authored texts.

Tables 8, 9, and 10 provide additional statistical testing information on the comparison of female-specific character mentions versus male-specific character mentions in male-authored texts.

Tables 11, 12, and 13 provide additional statistical testing information on the comparison of female-specific character mentions versus male-specific character mentions in female-authored texts.

5 Hypothesis 3: Details on Linear Regression

Figures 1, 2 and 3 show detailed linear regression statistics, as well as analysis of variance (ANOVA) for the relation between proportion of female-specific character, character occurrence, and pronoun counts (dependent variable) respectively, versus year index (independent variable), across male-authored texts.

Figures 4, 5 and 6 show detailed linear regression statistics, as well as analysis of variance (ANOVA) for the relation between proportion of female-specific character, character occurrence, and pronoun counts (dependent variable) respectively, versus year index (independent variable), across female-authored texts.

Table 3: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 1. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence is zero, when using the pronoun count measure. The research hypothesis is that male-specific pronoun count is higher than female-specific pronoun count; hence, the one-sided P value was reported in the main text.

	Male Pronoun Count	**Female Pronoun Count**
Mean	1910.501842	838.1387638
Variance	3229045.972	1394435.717
Observations	2443	2443
Pearson Correlation	0.700126883	
Hypothesized Mean Difference	0	
df	2442	
t Stat	41.23553515	
P(T≤t) one-tail	7.5289E-283	
t Critical one-tail	1.645477849	
P(T≤t) two-tail	1.5058E-282	
t Critical two-tail	1.960935904	

6 Secondary Analysis

While the three hypotheses in the main text are quantitative in nature, we supplement the analysis by conducting a qualitative assessment of the kinds of words associated with male and female character occurrences, using computational techniques from NLP. To do so, we first extracted 5 sentences around the first occurrence of each character (specifically, 1 sentence before and 4 after), which is where the description and introduction of the character is generally present. Then, we filtered all the words in these sentences and retained only the adjectives. We used part-of-speech (POS) tagging to accomplish this step (Voutilainen, 2003), which is a process of converting a sentence to a list of tuples, where each tuple comprises a word in the sentence, with a corresponding tag indicating whether the word is a noun, adjective, verb, and so on. Using POS tagging, we only retained words that were adjectives. In extracting adjectives around the first occurrence of each of the characters, our intent was to understand the descriptive theme and topics associated with male and female characters when they are first introduced in the book.

To find these topics and themes, we relied on an NLP technique called *word embeddings* (Mikolov et al., 2013), where a neural network is used to 'embed' each word in a corpus as a continuous, real-valued vector with a few hundred dimensions. The original neural network-based systems for word embeddings, such as word2vec, operated by sliding a window of a pre-specified size over the sequences of words in the corpus. An objective function is then optimized, such that vectors of words that tend to occur in the same window frequently are 'close' to each other in the embedding space. Empirically, it was found that words that have similar meaning and semantic relations tended to be embedded closer together when a sufficiently large and representative corpus of text is used (such as Wikipedia, or the Google News corpus). As further validation of these embeddings, a number of operations, including analogies, are found to hold naturally in the vector space. For example, the resulting vector for $\vec{King} - \vec{Man} + \vec{Woman}$ was found to lie very close to $\vec{Queen}$ (Mikolov et al., 2013). With the advent of transformer neural networks (Wolf

Table 4: Additional statistical measures and summary statistics in support of Hypothesis 2, reported separately for male- and female-authored texts. The sample size for male- and female-authored texts is 2,278 and 165 respectively. Detailed statistics are provided in the supplementary material. Results are reported to two significant digits.

Author Gender	**Statistic**	**Character Count**	**Character Occurrence Count**	**Pronoun Count**
Male	Mean (with 95% Confidence Interval) for male characters	34.13 (32.39 to 35.86)	158.62 (150.48 to 166.76)	1909.09 (1834.89 to 1983.29)
	Mean (with 95% Confidence Interval) for female characters	8.83 (8.31 to 9.34)	45.37 (41.86 to 48.89)	772.27 (726.52 to 818.02)
	One-sided P value (unpaired Student's t-test)	6.96e-240	2.54e-182	1.87e-297
Female	Mean (with 95% Confidence Interval) for male characters	39.85 (34.33 to 45.36)	161.58 (139.82 to 183.33)	1929.96 (1672.73 to 2187.18)
	Mean (with 95% Confidence Interval) for female characters	21.10 (18.79 to 23.41)	106.61 (89.37 to 123.86)	1747.51 (1497.03 to 1997.99)
	One-sided P value (unpaired Student's t-test)	4.09e-14	2.44e-07	0.038

et al., 2020), language representation learning has become more complex and context-sensitive (Devlin et al., 2018; Khashabi et al., 2020), although for this preliminary experiment we only consider a robust version of the word2vec model, described below.

An advantage of words being represented as vectors, capturing some notion of natural semantics in the embedding space, is that we can use the vector representations of the adjectives within an unsupervised *clustering* framework to recover themes and topics. We used the publicly available pre-trained Wiki News word embeddings that were derived by executing the popular fastText package on a large corpus of text (Facebook Research, 2017). The fastText package can be used for learning of word embeddings while being robust to misspellings and minor variations, and was originally created and released by Joulin et al. (2016). This model can be used to embed words in new text into vectors that capture semantic properties of words in a continuous-dimension space.

Specifically, we obtained two sets of vectors, containing the embeddings of adjectives extracted around male and female characters, respectively. Next, we clustered these words into eight clusters using the classic k-Means algorithm. The number of clusters was chosen as eight, since it was found to provide meaningfully different clusters without much overlap. We initially started with a smaller value for k but incrementally increased it until qualitatively meaningful clusters were visible. In the results, we comment further on how a more systematic approach

Table 5: Details on statistical significance testing (using the unpaired two-sample Student's t-test, assuming unequal variances) for Hypothesis 2. The null hypothesis is that the mean difference between the proportion of female-specific character mentions in male-authored texts and the same proportion in female-authored texts is zero, when using the character count measure. The research hypothesis is that the proportion in female-authored texts is higher than in male-authored texts; hence, the one-sided P value was reported in the main text.

	Proportion in Male-authored Texts	**Proportion in Female-authored Texts**
Mean	0.213821153	0.362475856
Variance	0.01902573	0.019940628
Observations	2278	165
Hypothesized Mean Difference	0	
df	187	
t Stat	-13.07796839	
P(T<=t) one-tail	1.76892E-28	
t Critical one-tail	1.653042889	
P(T<=t) two-tail	3.53784E-28	
t Critical two-tail	1.972731033	

could be used in future work.

Once the clusters were obtained, we took the 'midpoint' or centroid of each cluster, and obtained the five words nearest to the centroid in the vector space. In this manner, we use these five words (per cluster) to approximately represent the main theme of that cluster. We then visualize these 'representative' words and comment on the results qualitatively.

6.1 Results

Recall that the goal of the secondary analysis was a qualitative assessment of the kinds of words associated with male and female character mentions in the text using computational techniques, such as word embeddings and k-Means clustering. For ease of visualization, the outputs of six out of the eight obtained clusters are illustrated in Fig 7. The remaining two clusters covered themes such as nationality (e.g., 'British', 'American'), and were excluded from the visualization as they were largely similar between the two genders.

Although there are some similarities between the representative words across genders, we also found that while male-adjectives clusters contain words like 'strongest', 'largest', 'obnoxious', and 'sensible', female-adjectives clusters tended to contain words like 'beautiful', 'amiable', 'gentle' and 'frightened'. These results suggest that the differences between male-specific and female-specific character measures are not just quantitative (as measured through prevalence statistics) but may also be different qualitatively. Other historical and critical appraisal of that era, particularly the Victorian era that largely coincides with the publication period of the texts in the corpus, have come to a similar conclusion (Hughes, 2014).

At the same time, it is worth noting that this method is qualitative and heuristic, and that more experiments are needed to understand differences in descriptions of male versus female characters. Furthermore, there may also be overlap between words when male and characters

Table 6: Details on statistical significance testing (using the unpaired two-sample Student's t-test, assuming unequal variances) for Hypothesis 2. The null hypothesis is that the mean difference between the proportion of female-specific character mentions in male-authored texts and the same proportion in female-authored texts is zero, when using the character occurrence measure. The research hypothesis is that the proportion in female-authored texts is higher than in male-authored texts; hence, the one-sided P value was reported in the main text.

	Proportion in Male-authored Texts	**Proportion in Female-authored Texts**
Mean	0.217918342	0.382276566
Variance	0.039420612	0.040123424
Observations	2278	165
Hypothesized Mean Difference	0	
df	188	
t Stat	-10.18372339	
P(T≤t) one-tail	5.61609E-20	
t Critical one-tail	1.652999113	
P(T≤t) two-tail	1.12322E-19	
t Critical two-tail	1.972662692	

are introduced in the same paragraph. However, the experiment also helps us to understand differences in themes when describing male versus female characters. By using a similar methodology, other questions may also be explored, including equivalent versions of Hypotheses 2 and 3. For example, a version of the methodology could be used to explore the question of whether descriptions have changes over time, or are different between cross-section samples of male-authored versus female-authored books.

In future work, one could automatically set the value for k (rather than qualitatively increasing it until discovering the value of eight by trial-and-error) by using one of several established heuristic methods, such as the elbow method and the silhouette method (Kodinariya and Makwana, 2013). While the former is an excellent diagnostic tool, the latter provides a solution that is more quantitative. Follow-up research could consider whether using these methods results in different numbers of clusters, and also conduct a more quantitative version of the (currently) proposed qualitative analysis. Finally, it may be possible to study this qualitative finding through a more rigorous and quantitative lens by analyzing the word embeddings and the distance between them in vector space.

7 Limitations of Study and Ethical Issues

Gender, and gender identity, are important and complex issues in society, on which our understanding continues to evolve (O'Brien, 2009; Oakley, 2016). In light of this complexity, it is important to highlight both the limitations of this study, and the limitations of our findings therein. We also discuss ethical caveats that future researchers must bear in mind when interpreting our findings, including when using the data and methods described in this work to obtain their own findings.

Table 7: Details on statistical significance testing (using the unpaired two-sample Student's t-test, assuming unequal variances) for Hypothesis 2. The null hypothesis is that the mean difference between the proportion of female-specific character mentions in male-authored texts and the same proportion in female-authored texts is zero, when using the pronoun count measure. The research hypothesis is that the proportion in female-authored texts is higher than in male-authored texts; hence, the one-sided P value was reported in the main text.

	Proportion in Male-authored Texts	**Proportion in Female-authored Texts**
Mean	0.250288271	0.464174068
Variance	0.026360036	0.036502647
Observations	2278	165
Hypothesized Mean Difference	0	
df	182	
t Stat	-14.01815012	
P(T≤t) one-tail	4.69976E-31	
t Critical one-tail	1.653269024	
P(T≤t) two-tail	9.39951E-31	
t Critical two-tail	1.973084077	

First, and perhaps most importantly, we openly acknowledge a fundamental limitation of this study as one that only considered a dichotomous male-female gender categorization. Unfortunately, despite the best of our efforts, we did not find methods in the NLP literature that would allow us to detect non-binary, non-conforming and transgender individuals with the necessary accuracy. The *Gender_Detector* package that was previously described does not offer such a capability. Accuracy is paramount because non-conforming genders have already faced high levels of oppression historically, and it is not evident that they have been conveyed in literature as directly or representative of the populace as male characters. Considering the large disparity we already witness in our findings for female characters, we also cannot rule out complete suppression of (traditional and dichotomous) gender non-conformity. Indeed, we hope that future studies will make direct use of our data to study this issue in depth, in the same way that this study sought to convey the high levels of female character under-representation in pre-modern English literature.

Second, our study is obviously confined to the subset of books that we considered. While the set we did consider withstood the test of time among the books in that period, the population in that period was exposed to a broader set of literature (including books, pamphlets, plays and so on), which may yield different statistics compared to this study. However, as we showed in the related work, our estimates of female character under-representation agree to some extent with recent statistics on female character representation on screen, or in scenes with meaningful dialogue.

The other limitations noted here are related to the processing of the dataset. We summarize them below:

1. **Gender from Name Assumption:** An assumption made by our pipeline was that gender could be determined from the names of book authors or characters. Although we made some

Table 8: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in male-authored texts is zero, when using the character count measure. The research hypothesis is that male-specific character count is higher than female-specific character count (in male-authored texts); hence, the one-sided P value was reported in the main text.

	Male Character Count	**Female Character Count**
Mean	34.12642669	8.82572432
Variance	1778.557571	158.6068566
Observations	2278	2278
Pearson Correlation	0.844836614	
Hypothesized Mean Difference	0	
df	2277	
t Stat	37.44958128	
P(T≤t) one-tail	6.9574E-240	
t Critical one-tail	1.645523101	
P(T≤t) two-tail	1.3915E-239	
t Critical two-tail	1.96100637	

effort to verify the gender from other sources of information, such as Wikipedia articles, and have a near-perfect accuracy estimate based on sampled manual annotations, we could not do so for all authors and characters. There may be some bias of which we may not be fully aware. Certainly, we caution other scholars on solely relying upon this finding as their one source for determining the genders from names. It was also for this reason that we have released as much of the underlying data used in this study as possible in a repository (Nagaraj and Kejriwal, 2022), along with releasing much of the study-specific findings, statistical analysis, and code as supplementary material.

2. **Small-sample Accuracy Estimates:** Accuracy estimates of the various NLP steps noted in earlier sections were derived from fairly small samples and may be susceptible to bias. Future researchers should not quote or trust those estimates blindly, but aim to do their own sampling and annotation to expand the annotated sample set, discover potential biases in our own sample set, and derive accuracy estimates with higher statistical power.
3. **Possible Methodological Bias:** We advocate for more scrutiny into whether our methods, and the manner in which we investigated our hypotheses (or even the formulation of the hypotheses) might have been skewed or biased in a way that is not apparent to us at present. For instance, there is always the possibility that if the hypothesis had been stated a different way, or if we had used other measures of character prevalence, that the findings may have indicated a different degree of gender bias than what we reported. Another problem is that, as recent work as shown, there is considerable gender bias in seemingly unbiased computational systems, including NLP systems (Chen et al., 2021). Therefore, we hope that future researchers will consider alternative ways of formulating gender-relevant hypotheses, deriving intermediate data structures, and replicating the study using the dataset we have made available.

Table 9: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in male-authored texts is zero, when using the character occurrence count measure. The research hypothesis is that male-specific character occurrence count is higher than female-specific character occurrence count (in male-authored texts); hence, the one-sided P value was reported in the main text.

	Male Character Occurrence Count	**Female Character Occurrence Count**
Mean	158.618086	45.37357331
Variance	39267.83563	7314.324589
Observations	2278	2278
Pearson Correlation	0.511062382	
Hypothesized Mean Difference	0	
df	2277	
t Stat	31.59796208	
P(T≤t) one-tail	2.5373E-182	
t Critical one-tail	1.645523101	
P(T≤t) two-tail	5.0745E-182	
t Critical two-tail	1.96100637	

References

Chen Y, Mahoney C, Grasso I, Wali E, Matthews A, Middleton T, et al. (2021). *Gender Bias and Under-Representation in Natural Language Processing Across Human Languages*, 24–34. Association for Computing Machinery, New York, NY, USA.

Daiber J, Jakob M, Hokamp C, Mendes PN (2013). Improving efficiency and accuracy in multilingual entity extraction. In: *Proceedings of the 9th International Conference on Semantic Systems*, 121–124.

Devlin J, Chang MW, Lee K, Toutanova K (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DiffLib (2022). Documentation for difflib. https://docs.python.org/3/library/difflib.html. Accessed: 2022-09-29.

Explosionai (2022). Library architecture - spacy. https://spacy.io/api. Accessed: 2022-09-29.

Facebook Research (2017). English word vectors - fasttext. https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip. Accessed: 2022-09-29.

Han J, Wang H (2021). Transformer based network for open information extraction. *Engineering Applications of Artificial Intelligence*, 102: 104262.

Hong Z, Ward L, Chard K, Blaiszik B, Foster I (2021). Challenges and advances in information extraction from scientific literature: a review. *JOM*, 73(11): 3383–3400.

Hughes K (2014). Gender roles in the 19th century. *British Library*, 15.

Joulin A, Grave E, Bojanowski P, Mikolov T (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kejriwal M (2019). Information extraction. In: *Domain-Specific Knowledge Graph Construction*, 9–31. Springer.

Table 10: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in male-authored texts is zero, when using the pronoun count measure. The research hypothesis is that male-specific pronoun count is higher than female-specific pronoun count (in male-authored texts); hence, the one-sided P value was reported in the main text.

	Male Pronoun Count	**Female Pronoun Count**
Mean	1909.092625	772.2712906
Variance	3261324.724	1239970.289
Observations	2278	2278
Pearson Correlation	0.724912466	
Hypothesized Mean Difference	0	
df	2277	
t Stat	43.08740237	
P(T≤t) one-tail	1.8737E-297	
t Critical one-tail	1.645523101	
P(T≤t) two-tail	3.7473E-297	
t Critical two-tail	1.96100637	

Kejriwal M, Shao R, Szekely P (2019). Expert-guided entity extraction using expressive rules. In: *Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1353–1356.

Kejriwal M, Szekely P (2017). Information extraction in illicit web domains. In: *Proceedings of the 26th International Conference on World Wide Web*, 997–1006.

Khashabi D, Min S, Khot T, Sabharwal A, Tafjord O, Clark P, et al. (2020). Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700.*

Kodinariya TM, Makwana PR (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6): 90–95.

Leitner F (2015). Segtok - a segmentation and tokenization library. http://fnl.es/segtok-a-segmentation-and-tokenization-library.html. Accessed: 2022-09-29.

Loper E, Bird S (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028.*

Mikolov T, Chen K, Corrado G, Dean J (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Nadeau D, Sekine S (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26.

Nagaraj A, Kejriwal M (2022). Dataset for studying gender disparity in english literary texts. *Data in Brief*, 107905.

Natural Language Tool Kit Team (2022a). Documentation for nltk.chunk package. https://www.nltk.org/api/nltk.chunk.html. Accessed: 2022-09-29.

Natural Language Tool Kit Team (2022b). Source code for nltk.tokenize.punkt. https://www.nltk.org/_modules/nltk/tokenize/punkt.html. Accessed: 2022-09-29.

Oakley A (2016). *Sex, Gender and Society.* Routledge.

O'Brien J (2009). *Encyclopedia of Gender and Society*, volume 1. Sage.

Table 11: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in female-authored texts is zero, when using the character count measure. The research hypothesis is that male-specific character count is higher than female-specific character count (in female-authored texts); hence, the one-sided P value was reported in the main text.

	Male Character Count	**Female Character Count**
Mean	39.84848485	21.0969697
Variance	1286.702513	225.7222469
Observations	165	165
Pearson Correlation	0.595581262	
Hypothesized Mean Difference	0	
df	164	
t Stat	8.163921093	
P(T≤t) one-tail	4.08563E-14	
t Critical one-tail	1.654197929	
P(T≤t) two-tail	8.17125E-14	
t Critical two-tail	1.974534576	

OpenGenderTracking (2013). Github repository for global name data. https://github.com/OpenGenderTracking/globalnamedata. Accessed: 2022-09-29.

Palmer DD (2000). Tokenisation and sentence segmentation. *Handbook of natural language processing*, 11–35.

PyPi (2015). Gender detector v0.1.0. https://pypi.org/project/gender-detector/. Accessed: 2022-09-29.

Sassano M, Utsuro T (2000). Named entity chunking techniques in supervised learning for japanese named entity recognition. In: *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Schick T, Schütze H (2020). Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Sun J, Gao J, Zhang L, Zhou M, Huang C (2002). Chinese named entity identification using class-based language model. In: *COLING 2002: The 19th International Conference on Computational Linguistics*.

Vasiliev Y (2020). *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press.

Voutilainen A (2003). Part-of-speech tagging. *The Oxford Handbook of Computational Linguistics*, 219–232.

Wang C, Liu X, Chen Z, Hong H, Tang J, Song D (2021). Zero-shot information extraction as a unified text-to-triple translation. *arXiv preprint arXiv:2109.11171*.

Wang X, Jiang Y, Bach N, Wang T, Huang Z, Huang F, et al. (2020). Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. (2020). Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical*

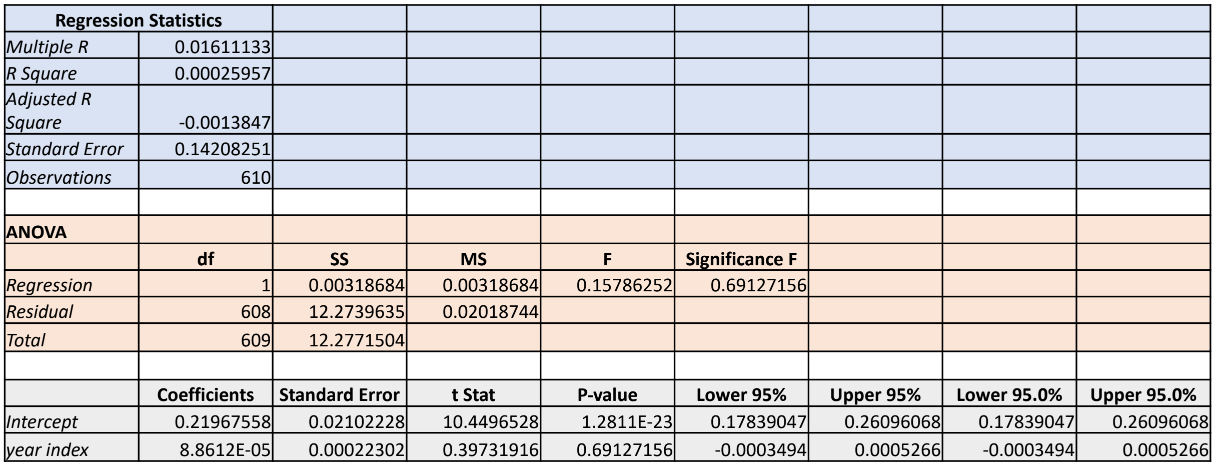
Table 12: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in female-authored texts is zero, when using the character occurrence count measure. The research hypothesis is that male-specific character occurrence count is higher than female-specific character occurrence count (in female-authored texts); hence, the one-sided P value was reported in the main text.

	Male Character Occurrence Count	**Female Character Occurrence Count**
Mean	161.5757576	106.6121212
Variance	20037.3677	12587.40961
Observations	165	165
Pearson Correlation	0.455488957	
Hypothesized Mean Difference	0	
df	164	
t Stat	5.239540615	
P(T≤t) one-tail	2.44491E-07	
t Critical one-tail	1.654197929	
P(T≤t) two-tail	4.88983E-07	
t Critical two-tail	1.974534576	

Methods in Natural Language Processing: System Demonstrations, 38–45.

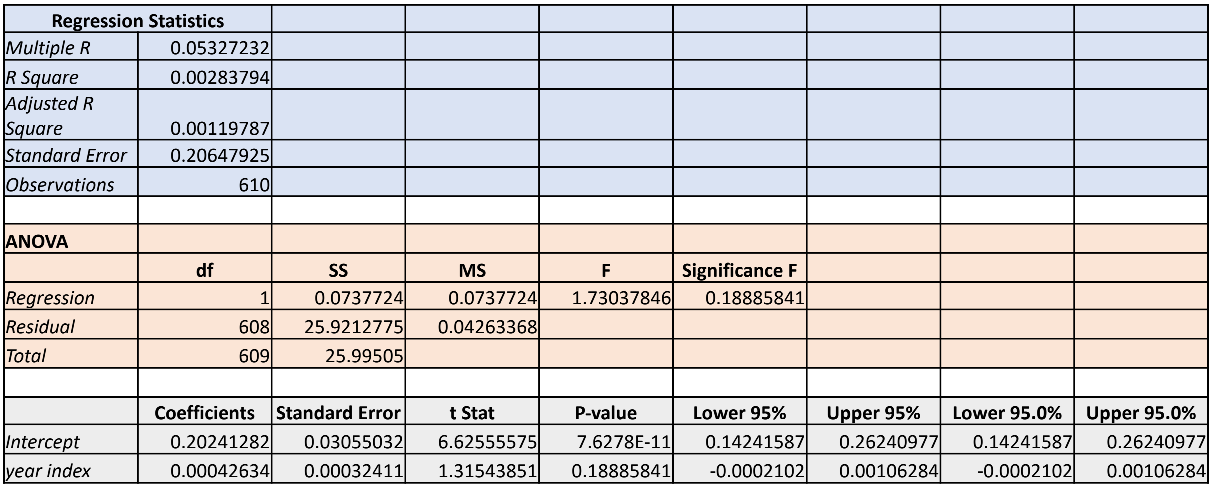
Table 13: Details on statistical significance testing (using the paired two-sample Student's t-test) for Hypothesis 2. The null hypothesis is that the mean difference between male-specific and female-specific character prevalence in female-authored texts is zero, when using the pronoun count measure. The research hypothesis is that male-specific pronoun count is higher than female-specific pronoun count (in female-authored texts); hence, the one-sided P value was reported in the main text.

	Male Pronoun Count	**Female Pronoun Count**
Mean	1929.957576	1747.509091
Variance	2800163.943	2655299.02
Observations	165	165
Pearson Correlation	0.683069363	
Hypothesized Mean Difference	0	
df	164	
t Stat	1.781637511	
P(T≤t) one-tail	0.038329491	
t Critical one-tail	1.654197929	
P(T≤t) two-tail	0.076658982	
t Critical two-tail	1.974534576	



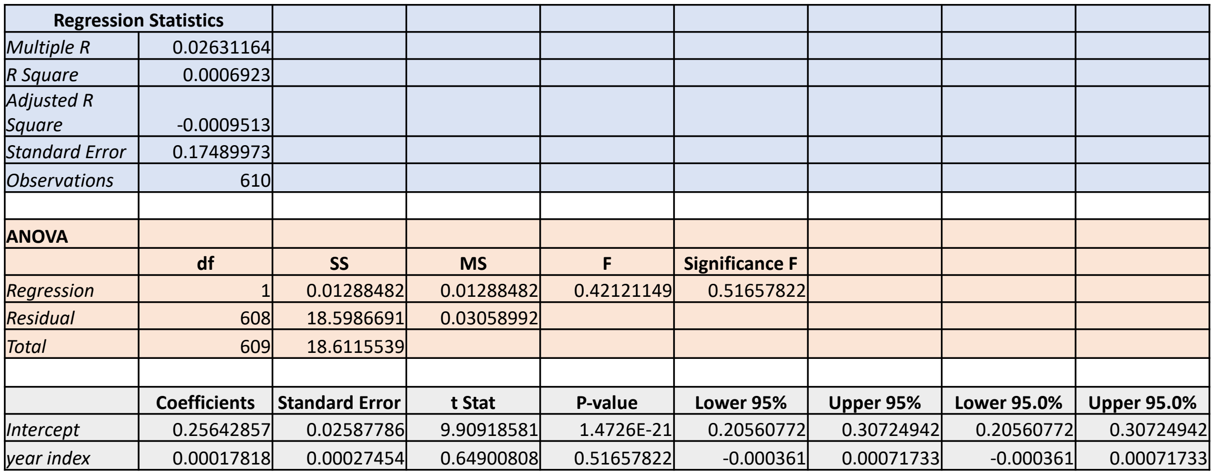
Regression Statistics								
Multiple R	0.01611133							
R Square	0.00025957							
Adjusted R Square	-0.0013847							
Standard Error	0.14208251							
Observations	610							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.00318684	0.00318684	0.15786252	0.69127156			
Residual	608	12.2739635	0.02018744					
Total	609	12.2771504						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.21967558	0.02102228	10.4496528	1.2811E-23	0.17839047	0.26096068	0.17839047	0.26096068
year index	8.8612E-05	0.00022302	0.39731916	0.69127156	-0.0003494	0.0005266	-0.0003494	0.0005266

Figure 1: Detailed linear regression and ANOVA results for proportion of female-specific character counts versus year index (male-authored texts only).



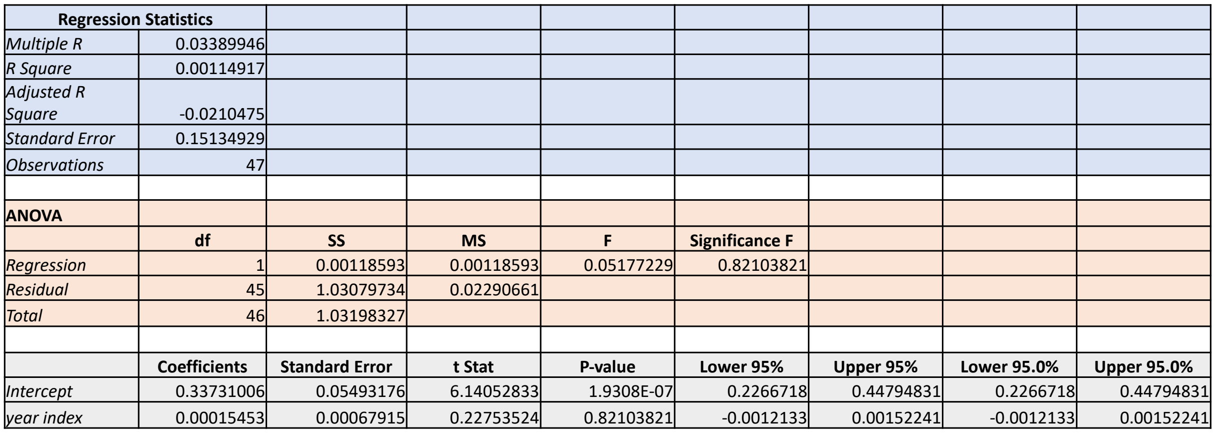
Regression Statistics								
Multiple R	0.05327232							
R Square	0.00283794							
Adjusted R Square	0.00119787							
Standard Error	0.20647925							
Observations	610							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.0737724	0.0737724	1.73037846	0.18885841			
Residual	608	25.9212775	0.04263368					
Total	609	25.99505						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.20241282	0.03055032	6.62555575	7.6278E-11	0.14241587	0.26240977	0.14241587	0.26240977
year index	0.00042634	0.00032411	1.31543851	0.18885841	-0.0002102	0.00106284	-0.0002102	0.00106284

Figure 2: Detailed linear regression and ANOVA results for proportion of female-specific character occurrence counts versus year index (male-authored texts only).



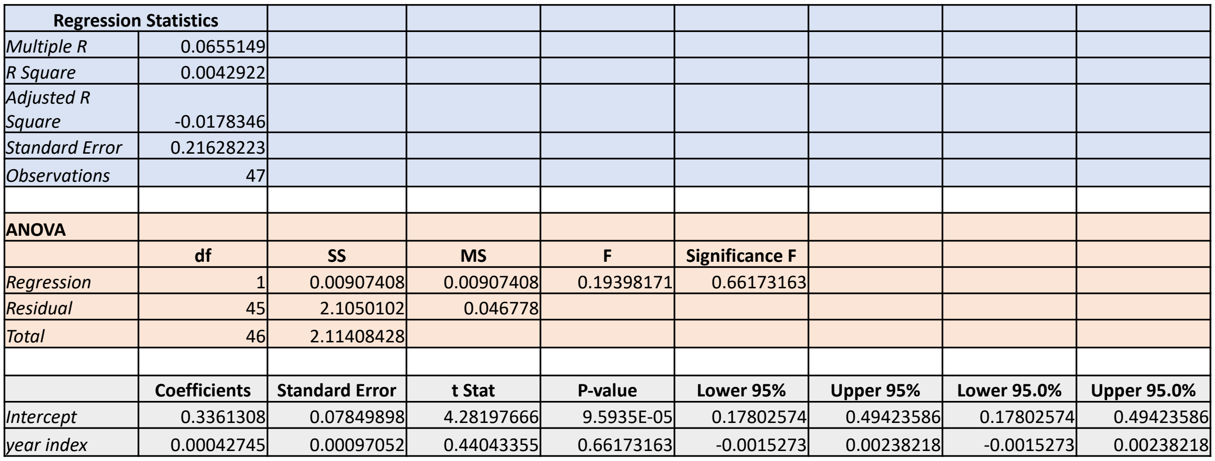
Regression Statistics								
Multiple R	0.02631164							
R Square	0.0006923							
Adjusted R Square	-0.0009513							
Standard Error	0.17489973							
Observations	610							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.01288482	0.01288482	0.42121149	0.51657822			
Residual	608	18.5986691	0.03058992					
Total	609	18.6115539						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.25642857	0.02587786	9.90918581	1.4726E-21	0.20560772	0.30724942	0.20560772	0.30724942
year index	0.00017818	0.00027454	0.64900808	0.51657822	-0.000361	0.00071733	-0.000361	0.00071733

Figure 3: Detailed linear regression and ANOVA results for proportion of female-specific pronoun counts versus year index (male-authored texts only).



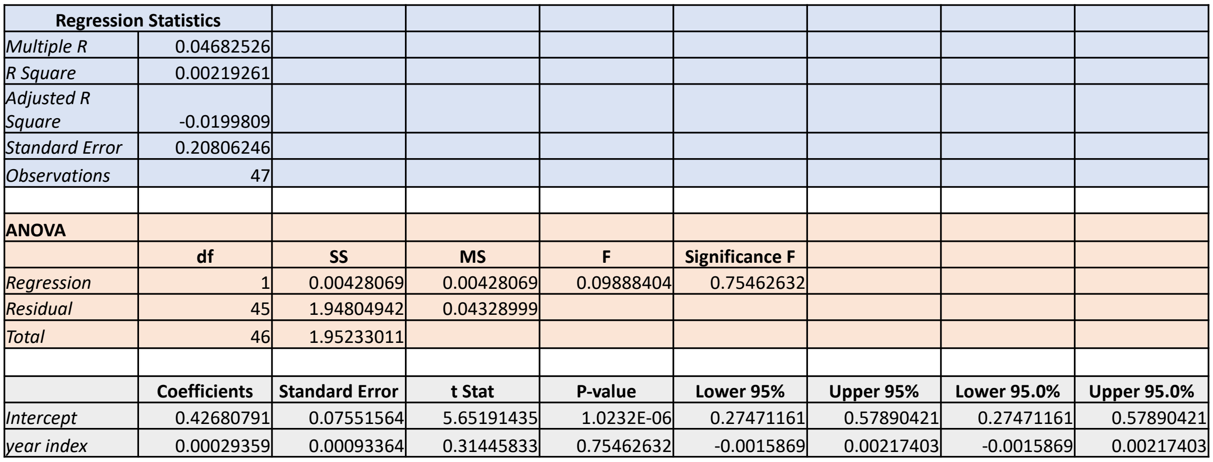
Regression Statistics								
Multiple R	0.03389946							
R Square	0.00114917							
Adjusted R Square	-0.0210475							
Standard Error	0.15134929							
Observations	47							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.00118593	0.00118593	0.05177229	0.82103821			
Residual	45	1.03079734	0.02290661					
Total	46	1.03198327						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.33731006	0.05493176	6.14052833	1.9308E-07	0.2266718	0.44794831	0.2266718	0.44794831
year index	0.00015453	0.00067915	0.22753524	0.82103821	-0.0012133	0.00152241	-0.0012133	0.00152241

Figure 4: Detailed linear regression and ANOVA results for proportion of female-specific character counts versus year index (female-authored texts only).



Regression Statistics								
Multiple R	0.0655149							
R Square	0.0042922							
Adjusted R Square	-0.0178346							
Standard Error	0.21628223							
Observations	47							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.00907408	0.00907408	0.19398171	0.66173163			
Residual	45	2.1050102	0.046778					
Total	46	2.11408428						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.3361308	0.07849898	4.28197666	9.5935E-05	0.17802574	0.49423586	0.17802574	0.49423586
year index	0.00042745	0.00097052	0.44043355	0.66173163	-0.0015273	0.00238218	-0.0015273	0.00238218

Figure 5: Detailed linear regression and ANOVA results for proportion of female-specific character occurrence counts versus year index (female-authored texts only).



Regression Statistics								
Multiple R	0.04682526							
R Square	0.00219261							
Adjusted R Square	-0.0199809							
Standard Error	0.20806246							
Observations	47							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	0.00428069	0.00428069	0.09888404	0.75462632			
Residual	45	1.94804942	0.04328999					
Total	46	1.95233011						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.42680791	0.07551564	5.65191435	1.0232E-06	0.27471161	0.57890421	0.27471161	0.57890421
year index	0.00029359	0.00093364	0.31445833	0.75462632	-0.0015869	0.00217403	-0.0015869	0.00217403

Figure 6: Detailed linear regression and ANOVA results for proportion of female-specific pronoun counts versus year index (female-authored texts only).

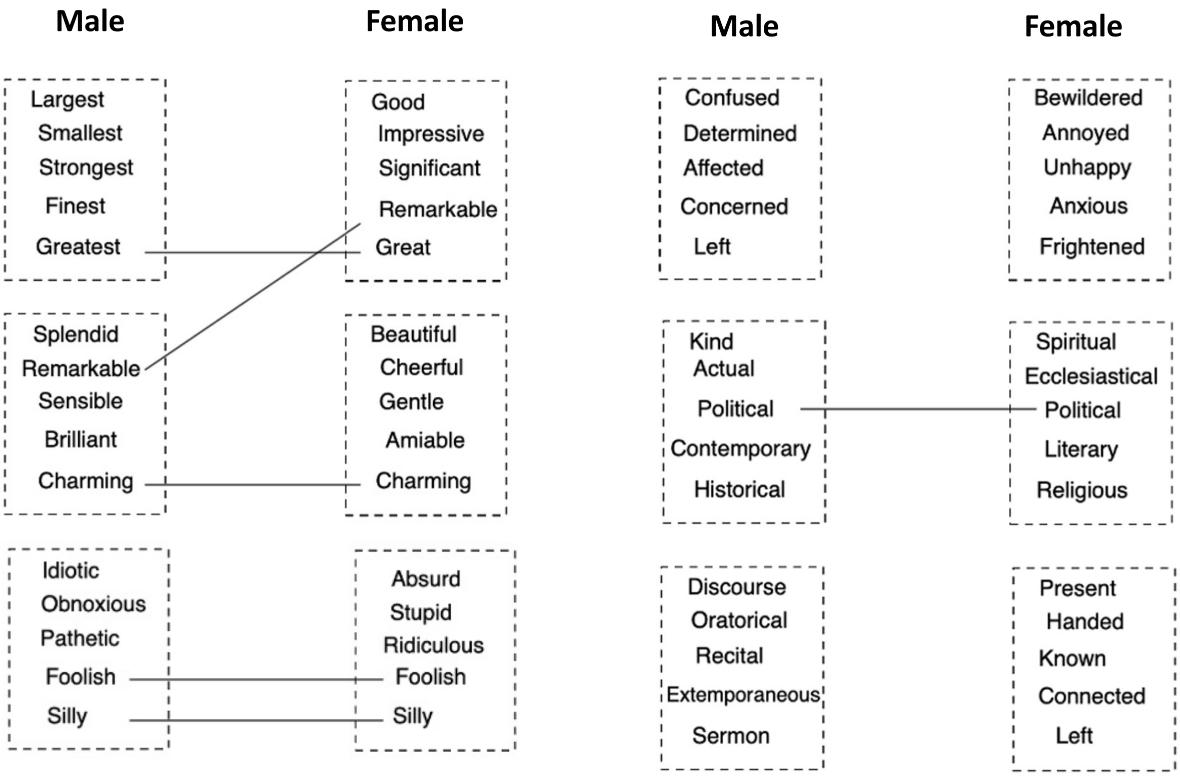


Figure 7: A comparison of the most representative words in six word-embedding clusters for male and female character occurrences in the corpus under study.