# Lexical Richness as a Lens: Exploring the Influence of Social Class on Authors' Writings*

Qi Er (Emma) Teng  Wentao Sun  Yang Cheng

March 20, 2024

In this study, we explore the influence of social class on lexical richness in the works of authors after the Great Depression, utilizing a comprehensive analysis of lexical data from literary works and social context of authors. Our findings reveal upper-class authors exhibiting a higher Corrected Type-Token Ratio which indicates greater lexical richness, and reflecting their social milieu. This study employs a thorough lexical analysis, aiming to elucidate the complex interplay between class background and language use in literature. This research contributes to our understanding of how social class background shapes language use in literature, offering insights into the broader cultural forces that drive linguistic creativity.

## Table of contents



*Code and data are available at: https://github.com/dwz92/Analyzing-the-impact-of-social-class-on-authors-writings

# 1 Introduction

By the late 20th century, the quantitative study of vocabulary in texts reached a level of precision that spurred a series of linguistic studies. Vocabulary is a fundamental part of language and is the unit of meaning that makes up larger structures such as sentences, paragraphs, and full texts (Schmitt and Schmitt 2020). The use of vocabulary represents the author's writing style and is also influenced by family education and social class. The purpose of this paper is to gain insight into whether class differences affect the richness of vocabulary and central ideas in works by examining whether upper-class authors use richer writing vocabulary in their essays than middle-class authors.

Utilizing a combination of quantitative data analysis and authorship studies, we investigated the Type-Token Ratio of upper- and middle-class authors to reflect the richness of textual vocabulary over the period 1930-1990. This richness gives us an idea of the number of different terms used in a text and the diversity of the vocabulary(Torruella and Capsada 2013). For this study, we selected two upper-class authors, Nancy Mitford and Anthony Powell, and two working-class authors, Hugh Garners and David Herbert Lawrence. We utilized the Corrected Type-Token Ratio (CTTR) for a detailed lexical analysis of their texts. In language acquisition studies, this often entails transcriptions of spontaneous speech, such as narratives, retellings, and picture-elicited speech(Van Hout and Vermeer 2007). A higher Corrected Type-Token Ratio (CTTR) indicates that the author possesses a richer vocabulary, demonstrating a greater diversity in word usage within their texts. This phenomenon suggests that different classes of writers have different writing styles and preferred vocabularies, and that the writers' writing styles and ideas in their works are influenced by their vocabulary choices.

The paper is structured to facilitate a comprehensive understanding of the study and its implications. Following Section 1, Section 2 presents the data, detailing the data sources, analytical techniques, and the rationale behind the chosen methods. Section 4 discusses the results, elaborating on the observed trends and patterns in the authors writings. Section 5 provides an in-depth discussion of these findings, exploring potential factors influencing these trends, drawing connections to broader social-economic issues, and providing suggestions for future research in this area.

### 1.1 Estimand

This study focuses on estimating the causal effect of class origin on writers' vocabulary use, specifically how it shapes and changes in writing style. By examining the vocabulary richness of authors from upper and working classes, we aim to understand how variations in education, life experiences, and personal perceptions due to class differences influence writing.

## 2 Data

This section aims to offer an insightful understanding of the dataset utilized in our analysis. We selected 10 books written by each of two upper-class authors and two middle-class authors between 1930 and 1990. The dataset captures all the texts in the selected books. These data provide the number and type of words in each book, allowing us to analyze the richness of the vocabulary used in the writing of different class authors and to understand the relationship between class background and writing vocabulary use.

### 2.1 Source and Methodology

Data on word counts and unique word counts for the four selected authors were sourced from the original text files of the selected authors' writings. The original files of these writings were gathered through online library sources (please refer to Section B), these files were processed and transformed into workable data using R(R Core Team 2020). For key operations, please refer to the Section A.

While there were alternative text files available from other public and private sources, these works were chosen due to the completeness and accuracy of these files.

## 2.2 Variables

To better understand the data and the research process, three randomly selected paragraphs have been chosen to provide a detailed description of the research methodology used, explaining its relevance and how it contributes to our understanding of the topic. Our focus is on word count and unique word count in the novels written by the selected authors that offer a comprehensive view of how different authors use the vocabulary in their writing.

Table 1: First 3 Paragraph of Waste No tear by Hugh Garners

| Text |
|---|
| "TO LOOK AT ME, you'd never know that I'd ever been anything but a skid row bum. I haven't been one long, as years are reckoned down here, and I won't be one much longer. As soon as I get out of this hospital and get a drink or two into me, I'm going to blow the top off this town. Yeah, I know. I've been saying the same thing for months, but this time I mean it." |
| " " |
| "In our city the skid row is located ten blocks from Waltham Avenue. I was born on Waltham Avenue thirty-seven years ago, and it took me the last twenty to get to skid row. Twenty years is a third of a lifetime, and my twenty should have been the formative and pleasurable ones of my life. Instead, they were something else again, something I am going to try to put down here. It may help me, if I lay out my life in black type on white paper. It may give me the guts to do the things I've got to do." |
| " " |
| "When I was a kid Waltham Avenue had not yet become a slum street. It was situated in a poor working-class district, but its tenants were decent people who believed in work, thrift, and of some day moving to one of the new residential districts that were springing up in the suburbs of our city. Many of them moved away over the years when I was a boy, but my family were left behind. It might have been my old man's drinking that stopped the Matterson family from getting ahead, or, as he claimed, it might have been bad luck. Whatever it was, it kept us chained down to our little rented house on Waltham Avenue. We were still there when Waltham became an industrial street, with the soap factory and the planing mill and the fish wholesaler's where there had formerly been rows of little houses like ours." |
| " " |

Table 1, created with `kableExtra` (Zhu 2021), showcases the first 3 paragraphs of Waste No Tears by Hugh Garners. To examine the diversity of vocabulary utilized in the published works of various authors, approximately eight books per author were meticulously selected. The data were then systematically aggregated into two distinct categories: the word count, representing the aggregate number of conventional words, and the unique word count, denoting the tally of distinct words used. This methodical aggregation enables a concentrated comparison between

the two types of word counts across the selected works, providing insight into the authors' lexical richness and stylistic nuances.

Table 2: First Ten Rows of Writings Word Count and Uniques Word Count

| author | title | word count | unique word count |
|---|---|---|---|
| David Herbert Lawrence | england_my_england | 12885 | 3441 |
| David Herbert Lawrence | kangaroo | 149227 | 21989 |
| David Herbert Lawrence | lady_chatterleys_lover | 117257 | 17006 |
| David Herbert Lawrence | odour_of_chrysanthemums | 7516 | 2509 |
| David Herbert Lawrence | sons_and_lovers | 160136 | 19927 |
| David Herbert Lawrence | the_plumed_serpent | 170581 | 21632 |
| David Herbert Lawrence | the_rainbow | 185724 | 21378 |
| David Herbert Lawrence | the_virgin_and_the_gipsy | 30421 | 6982 |
| Hugh Garners | waste_no_tears | 44293 | 6504 |
| Hugh Garners | the_conversion_of_willie_heaps | 3592 | 1271 |

Table 2, built with `kableExtra` (Zhu 2021), displays the first ten rows of writing's word count and unique word count. This is a more concise table after the word count, which shows the authors and the books they wrote and lists the two types of word counts across the selected works. Provides a streamlined view for subsequent analysis and processing.

## 2.3 Author's average word counts distribution

Table 3: Word Count Comparison by Authors

| author | word count mean | word count std. dev. |
|---|---|---|
| David Herbert Lawrence | 104218.38 | 75127.43 |
| Hugh Garners | 7173.85 | 11240.12 |
| Nancy Mitford | 71379.60 | 14009.45 |
| Powell Anthony | 74185.55 | 7229.23 |

Initially, the length and consistency of the literary outputs were systematically evaluated. Presented in Table 3 are the mean word count and standard deviation for the works of four selected authors. David Herbert Lawrence, with a mean word count of 104,218.38, ranks highest among the authors, yet the substantial standard deviation of 75,127.43 indicates a significant variation in the lengths of his works. Conversely, Hugh Garners possesses the minimal average word count, registering at merely 7,173.85 words. On the other hand, Nancy Mitford and Anthony Powell exhibit notably higher average word counts, recorded at 71,379.60 and 74,185.55, respectively, with the lengths of their works demonstrating relative stability. Remarkably, Anthony Powell manifests the lowest standard deviation, at 7229.23, underscoring the uniformity in the length of his works. Contrasting with the brevity and conciseness characterizing the outputs of Garners, Mitford, and Powell, Lawrence's compositions are distinguished by their extensive length and variability.

## 2.4 Measurements

In our study, ten books published by each of the four authors between 1900-1990 were analyzed in detail. The txt files of these books were extracted from online libraries to ensure the authenticity and accuracy of the data obtained. We measured textual vocabulary richness using CTTR, a modified form of TTR to minimize the effect of text length on vocabulary richness measurements.

For each published book, CTTR is calculated by dividing the number of words in the text that are not repeated (number of types) by the square root of the total number of words (number of tokens) multiplied by two. The type count refers to the number of unique words in the text, i.e. words that are not repeated. The number of tokens, on the other hand, refers to the total number of words in the text, including repeated words. This method allows for a reasonable comparison by knowing the vocabulary richness of each book of the four authors chosen. This calculation method allows for a fairer comparison ignoring the length of the text to each other.The value of CTTR quantitatively responds to the diversity and richness of the text's vocabulary.

## 2.5 Average CTTR From Data

Table 4: Authors CTTR Average and Standard Deviation

| Author | Average | Standard Deviation |
|---|---|---|
| David Herbert Lawrence | 31.61 | 7.37 |
| Hugh Garners | 16.31 | 2.31 |
| Nancy Mitford | 31.97 | 1.49 |
| Powell Anthony | 34.67 | 1.64 |

Table 4 shows the average Corrected Type-Token Ratio (CTTR) and the associated standard deviation for each of the four studied authors. David Herbert Lawrence and Nancy Mitford both have high average CTTRs at 31.61 and 31.97 respectively, indicating a greater lexical richness in their works. However, Lawrence's works show more variability with a standard deviation of 7.37, as opposed to Mitford's 1.49, suggesting her use of vocabulary is more consistent. Hugh Garners has a notably lower average CTTR of 16.31, paired with a standard deviation of 2.31, which is indicative of a consistently narrower range of vocabulary. Powell Anthony tops the average CTTR at 34.67 with a standard deviation of 1.64, demonstrating not only the highest lexical richness but also a high consistency in vocabulary usage across his works. These figures collectively offer a quantitative insight into the diversity and consistency of language used by authors from different social backgrounds.

## 2.6 Word Count vs. CTTR

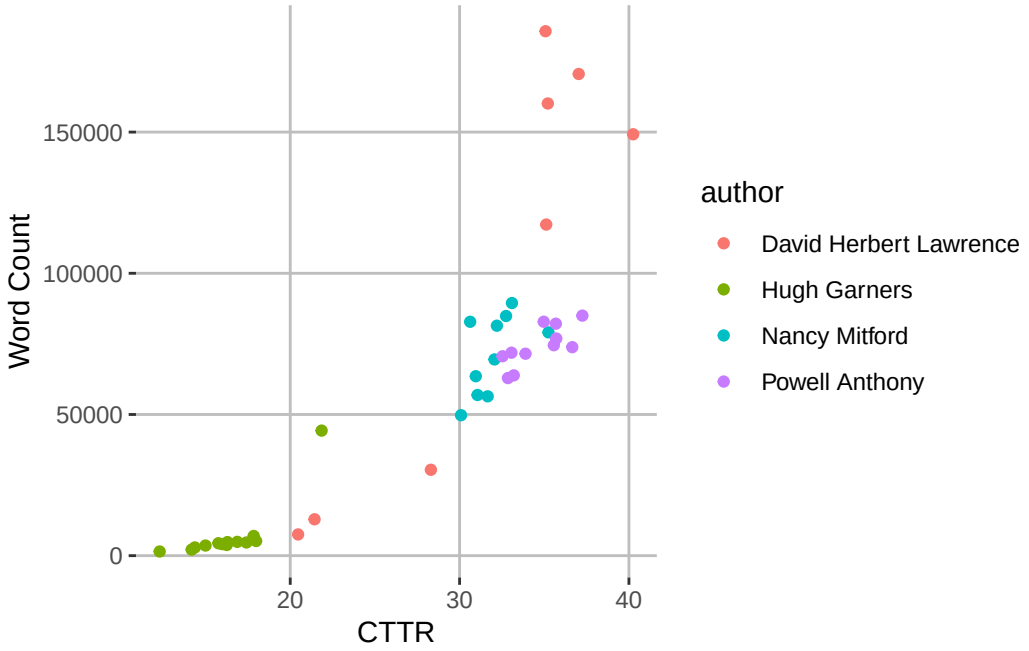


Figure 1: Comparison Between CTTR and Word Count by Authors

Figure 1 presents a scatter plot comparing the CTTR and word count for works by David Herbert Lawrence, Hugh Garners, Nancy Mitford, and Powell Anthony. The plot reveals a pattern that suggests a correlation between the lexical richness of an author's work, as measured by CTTR, and the length of the text, as indicated by word count. David Herbert Lawrence is a conspicuous outlier in this analysis. His works are scattered across a wide range of both CTTR and word counts, showing that while he has a propensity for longer works, the lexical richness within those works is quite variable. Hugh Garners' points cluster at the

lower end of both word count and CTTR scales, suggesting that his works are shorter and less lexically rich. This could reflect the impact of social and educational limitations on his writing style.

Nancy Mitford's works are positioned in the middle range of the word counts but show a higher CTTR, similar to that of Lawrence, indicating a richer vocabulary that is not necessarily dependent on the length of the text. Powell Anthony stands out with a group of works that not only have high CTTR values but also fall into a moderate range of word counts. His data points suggest a strong consistency in lexical richness that does not fluctuate widely with the length of the text. Overall, Figure 1 points to a general trend where longer works tend to have higher CTTRs, but also highlights individual variations among authors. David Herbert Lawrence's variability points to a unique stylistic approach, while the other authors demonstrate a more predictable relationship between the volume of text and lexical diversity, aligning with the anticipated influence of their social class and educational experiences on literary production.

# 3 Model

Negative binomial regression is a type of generalized linear model that is particularly useful for modeling count data, especially when equal mean and variance cannot be assumed. The negative binomial regression model extends the Poisson regression model by disregard this restriction by allowing data to exhibit overdispersion.

Moreover, as a negative binomial regression model combines with multilevel modeling (hierarchical linear modeling), the principles of multilevel analysis and the flexibility of the negative binomial distribution combines to handle overdispersed count data.

In the context of our paper, we will fit our multi-level model with the Negative Binomial regression to discover the correlation between authors lexical richness and their social class. Although Corrected Type Token Ration (CTTR) is not an exact count value, for the context of this paper we will temporarily disregard this assumption by rounding the ratio to zero decimal digit.

## 3.1 Model set-up

Define $y_i$ as the CTTR of the author's writings, then $\alpha_a$ as the author.

$$y_i|\pi_i \sim \text{Negative Binomial}(r, \pi_i) \tag{1}$$
$$logit(\pi_i) = \beta_0 + \alpha_a \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 3) \tag{3}$$
$$\alpha_a \sim \text{Normal}(0, 3) for a = 1, 2, 3, 4 \tag{4}$$
$$\tag{5}$$

In this model:

$y_i$: The dependent variable, $y_i$, represents the CTTR of the author's writings.

$\beta_0$: The intercept $\beta_0$ has a prior distribution that is normal with mean 0 and standard deviation 3.

$\alpha_a$: Each author effect $\alpha_a$ also has a normal prior with mean 0 and standard deviation 3.

We will run the model in R (R Core Team 2020) using the `rstanarm` package of (Goodrich et al. 2020).

### 3.2 Model Justification

The goal of this model is to understand how authors' lexical richness, as measured by CTTR, varies among authors. To account for the inherent variability and overdispersion in the count data, we model our data with the Negative Binomial regression. The multilevel aspect, from the background of each authors, allows the model to account for differences between authors. This aspect acknowledges that some authors may consistently use a more diverse vocabulary than others. With this setup, we expect to see a model that explores the variability in lexical richness across authors' writings, accounting for overdispersion in corrected type token ratio (CTTR) data. Moreover, since this setup is particularly useful for exploring the individual characteristics of authors, we also anticipate to see cprrelation of authors' social class with the lexical richness of their writings.

## 4 Results

Section 4 presents the core findings on the relationship between authors' word count and CTTR, specifically focusing on the influence of social class and educational background on literary vocabulary richness.

### 4.1 Model Visualization

Table 5 provides coefficients from a negative binomial regression and a multilevel negative binomial regression, indicating how each author's work differs from the intercept in terms of Corrected Type-Token Ratio (CTTR).

Hugh Garners shows a negative coefficient (-0.658 in Neg Binom and -1.308 in Multilevel Neg Binom), suggesting his CTTR is significantly lower than the baseline set by the intercept, which is consistent with the shortest average word count observed in the previous analysis.

Nancy Mitford's coefficient is slightly above zero (0.021 in Neg Binom and 0.152 in Multilevel Neg Binom), indicating her CTTR is almost at the baseline level. This aligns with her position as neither the lowest nor the highest in terms of lexical richness. Powell Anthony's positive coefficient (0.107 in Neg Binom and 0.153 in Multilevel Neg Binom) implies his CTTR is above the baseline, though not as high as David Herbert Lawrence's. David Herbert Lawrence does not have a coefficient listed here but is included in the distribution figure below.

Table 5: Model Summary of Multi-level and Negative Binomial model

| | Neg Binom | Multilevel Neg Binom |
|---|---|---|
| (Intercept) | 3.444 | 3.291 |
| | (0.119) | (0.267) |
| authorHugh Garners | −0.658 | |
| | (0.153) | |
| authorNancy Mitford | 0.021 | |
| | (0.156) | |
| authorPowell Anthony | 0.107 | |
| | (0.152) | |
| Num.Obs. | 42 | 42 |
| ICC | | 1.0 |
| Log.Lik. | −136.460 | −137.308 |
| ELPD | −137.9 | −138.8 |
| ELPD s.e. | 2.1 | 2.0 |
| LOOIC | 275.9 | 277.5 |
| LOOIC s.e. | 4.2 | 4.1 |
| WAIC | 275.8 | 277.5 |
| RMSE | 3.48 | 3.54 |

To gain a more intuitive understanding of the model, we visualized the coefficients from Table 5 in the form of a distribution figure. Figure 2, presents the distribution of CTTR estimates for each author, providing a visual representation of the variance around the mean estimate. Powell Anthony's distribution is relatively narrow and centered around the mean, suggesting his lexical richness is consistently above the baseline with less variation in CTTR scores. Nancy Mitford's distribution also clusters closely around the mean but shows a slight skew towards higher values in the multilevel negative binomial regression, indicating a modest variability with a tendency for higher CTTR scores.

Hugh Garners' distribution is broader and notably shifted to the left of the baseline in both models, implying a lower CTTR and greater variability in lexical richness compared to the others. David Herbert Lawrence, while not explicitly compared against a coefficient in the table, is represented in the figure with the broadest distribution. This suggests that his CTTR is highly variable, confirming the previous analysis which indicated his works are both lengthy and diverse in terms of lexical richness.

Overall, the Figure 2 corroborates the numerical findings of Table 5, offering a visual confirmation of the differences in lexical richness and variability among the authors. It highlights the significant spread in the lexical diversity of Lawrence's works, as well as the relative consistency observed in the writings of Powell Anthony and Nancy Mitford, with Garners' works showing both a lower mean and a higher dispersion.

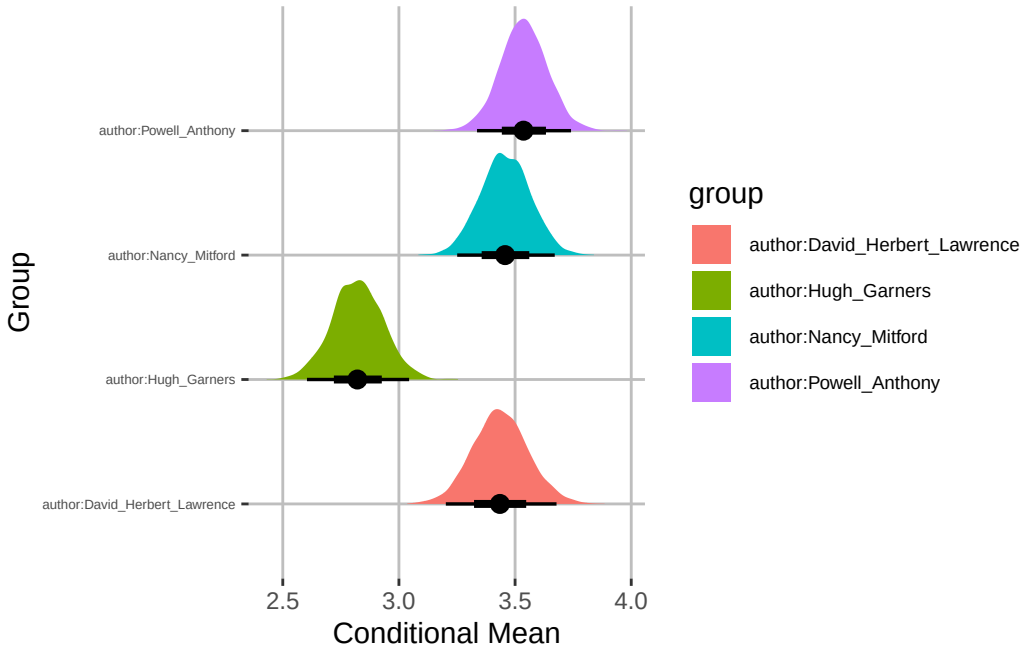


Figure 2: Distribution of Draws by Authors

# 5 Discussion

This study provides a new perspective on literary analysis, particularly through quantitative substantiating the impact of social class on the use of literary language. Traditionally, literary analysis has focused primarily on qualitative methods, emphasizing subjective interpretation and thematic exploration. By introducing a quantitative dimension, this study not only complements these traditional approaches, but also provides a new lens through which the interplay between socioeconomic and cultural factors can be understood.

## 5.1 Findings

The results of this research strongly support our initial hypothesis, underscoring a notable correlation between an author's social class, educational background, and the lexical richness of their work, despite the clear outliers in our data. The observed differences in Corrected Type-Token Ratios (CTTR) between upper-class authors like Nancy Mitford and Anthony Powell and those of Hugh Garners and David Herbert Lawrence reinforce this link. It is noteworthy that David Herbert Lawrence, while presenting a high CTTR, also has a significant work length. When normalized for a moderate length, Lawrence's CTTR settles around 20, which is consistent with our hypothesis. The study not only delineates the association between social class and lexical richness in literature but also emphasizes the importance of considering an

author's background for a more nuanced interpretation of their work. This indicates that the vocabulary an author employs is substantially shaped by their social context, rather than being solely a reflection of individual style.

## 5.2 Social and Historical Influences

The novels selected for this study were written during the mid-20th century. This period was heavily influenced by World War II as well as the Great Depression, which resulted in profound social changes and significant class divisions. World War II interrupted the schooling of many young women and men and the outbreak of the war led to an immediate and dramatic decline in high school graduation rates, which fell back to the levels of the early 1930s (Jaworski 2014). Social stratification had a significant impact on the educational and cultural experiences of individuals, which in turn was reflected in the richness of the vocabulary of their literary works. Writers from upper class backgrounds, such as Nancy Mitford and Anthony Powell, were exposed to a rich educational environment and a leisurely pursuit of literature and language that was usually unavailable to writers from the lower classes. This may have helped to enrich their vocabularies by observing a wider and more subtle choice of words in their work. Conversely, authors from lower socio-economic backgrounds, such as Hugh Garners and David Herbert Lawrence, experienced more limited educational opportunities, often confined to vocational or primary education. Their writing tended to reflect the immediate, more constrained realities of their lives and the spoken language of their surroundings. This difference in education and cultural exposure is evident in comparative analyses of the vocabulary richness of their work, which is much narrower in the works of lower-class writers.

Thus, the unique social circumstances and educational differences of the mid-twentieth century played a key role in shaping the lexical character of literary works. The unique socio-economic contexts of the period created different opportunities for the linguistic development of writers, leading to a clear stratification of vocabulary use closely related to the class background of the authors. This observation underscores the importance of considering socio-economic factors when analyzing literary language trends, especially during periods of significant social and economic contrasts. This understanding is crucial for a comprehensive statistical analysis of language patterns across social classes and historical contexts in literary studies.

## 5.3 Weaknesses and Future Research Directions

One limitation of our study is the potential influence of unobserved variables that may affect the lexical richness in authors' works, such as individual educational background or specific life experiences, which were not fully captured in our analysis. Moreover, while we used an improved TTR formula, its inherent sensitivity to text length variability and the possibility of other underlying linguistic factors were not entirely addressed. These could include stylistic choices or genre-specific language use that our study does not explore in depth.

Future research should aim to unravel these complex interactions further, perhaps by integrating more granular data on authors' personal histories or by employing a wider range of linguistic analysis tools. Additionally, exploring the application of advanced computational methods, such as natural language processing algorithms, could offer more detailed insights into how social class impacts lexical choice and style in literature. Understanding these nuanced relationships is vital for comprehensively interpreting the intersection of social factors and literary expression.

# Appendix

## A Data Manipulation and Cleaning

In the context of our raw data, we are dealing with raw text files of authors' writings. Thus, we begin by omitting unnecessary portions (such as publishers' notes) the text files. To transform these raw texts to workable data, we created three functions, `word_dist()`, `cttr_dist()`, and `cttr()` using R libraries (Wickham et al. 2019, 2023; Müller 2020; Zhu 2021). These functions aim to gather the descriptive statistics of each writings partially (average, mean, cttr), and store them into `data/analysis_data` with the author name.

Table 2 was created using the dataset imported from `data/analysis_data/cttr_all.csv`, which was created by using `cttr()`. The cleaned texts along with the titles and authors were passed into `cttr()`, which loop through each of the text to gather the word counts and unique word counts. Then, these temporary parameters were formulated into the calculation for CTTR, then stored into `data/analysis_data/cttr_all.csv`.

Table 3 was created using the dataset imported from `data/analysis_data/wordcount_dist.csv`. This dataset was created by using the function `word_dist()`, which returns the author, average word count, and std. dev. of word count of each texts. Different from `cttr()`, the stored value from `data/analysis_data/cttr_all.csv` were passed into the function to generate the values.

Table 4 was created using `data/analysis_data/cttr_dist.csv`. This dataset was created by using the function `cttr_dist()`, which returns the author, average word count, and std. dev. of CTTR of each texts. Similar to the creation of Table 3, `data/analysis_data/cttr_all.csv` were passed into the function to generate the values.

Table 5 and Figure 2 were created using the imports from `models/author_cttr_model_multi.rds` and `models/author_cttr_model.rds`. These `.rds` files were created in `scripts/03-models.R` using functions such as `stan_glm()` from R libraries (Wickham et al. 2019; Goodrich et al. 2020). Moreover, the `corrected_type_token_ratio` from `data/analysis_data/cttr_all.csv` were rounded to integers in order to use Negative Binomial regression.

Figure 1 were created using the imports from `data/analysis_data/cttr_all.csv`. Which has the same creation process as Table 2.

## B Referenced Writings

(Mitford 2011, 2015, 1970, 2010a, 2010b; Mitford and Foreman 2001; Mitford and Mansel 2012; Powell 1977, 1988, 1968, 2011c, 2010, 1994, 1957, 1955, 2011a, 1966, 2011b; Lawrence 1992, 2005, 2022, 2012, 2013, 2007, 2015; Lawrence and Steele 2002; *Hugh Garner's Best Stories* 2003; Garner 2014)

## References


Garner, H. 2014. *Waste No Tears*. Ricochet Series. Vehicule Press. https://books.google.ca/books?id=Kn26mwEACAAJ.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm.

*Hugh Garner's Best Stories*. 2003. Canadian National Institute for the Blind. https://books.google.ca/books?id=vlr7xAEACAAJ.

Jaworski, Taylor. 2014. "'You're in the Army Now:' The Impact of World War II on Women's Education, Work, and Family." *The Journal of Economic History* 74 (1): 169–95.

Lawrence, D. H. 1992. *The Virgin and the Gipsy*. Vintage International. Knopf Doubleday Publishing Group. https://books.google.ca/books?id=dEI7gTC-GI8C.

———. 2005. *The Rainbow*. Collector's Library. CRW Publishing Limited. https://books.google.ca/books?id=f6leiqFvgTsC.

———. 2007. *Lady Chatterley's Lover*. Bantam Classic. Random House Publishing Group. https://books.google.ca/books?id=K9KbYvVN3xUC.

———. 2012. *Sons and Lovers*. Dover Thrift Editions: Classic Novels. Dover Publications. https://books.google.ca/books?id=bDnCAgAAQBAJ.

———. 2013. *Odour of Chrysanthemums: Short Story*. HarperCollins Canada. https://books.google.ca/books?id=Mrn3llY-IgcC.

———. 2015. *England, My England*. Start Classics. https://books.google.ca/books?id=-DaaDAAAQBAJ.

———. 2022. *The Plumed Serpent*. Aegitas. https://books.google.ca/books?id=01ZhEAAAQBAJ.

Lawrence, D. H., and B. Steele. 2002. *Kangaroo*. Cambridge Edition of the Letters and Works of d. H. Lawrence: Works of d. H. Lawrence / General Ed. James t. Boulton. Cambridge University Press. https://books.google.ca/books?id=qd4wPBvd7FEC.

Mitford, N. 1970. *Frederick the Great*. Harper & Row. https://books.google.ca/books?id=geWAIRf1wccC.

———. 2010a. *Love in a Cold Climate*. Penguin Books Limited. https://books.google.ca/books?id=JoV9E82ecxQC.

———. 2010b. *The Pursuit of Love*. Radlett and Montdore. Knopf Doubleday Publishing Group. https://books.google.ca/books?id=m631JM1stCcC.

———. 2011. *The Penguin Complete Novels of Nancy Mitford*. Penguin Books Limited. https://books.google.ca/books?id=klMHD8saMVgC.

———. 2015. *Don't Tell Alfred: The Wickedly Funny Sequel to the Pursuit of Love*. Penguin Books Limited. https://books.google.ca/books?id=ZyanCgAAQBAJ.

Mitford, N., and A. Foreman. 2001. *Madame de Pompadour*. New York Review Books. https://books.google.ca/books?id=ke7H2dcy2jMC.

Mitford, N., and P. Mansel. 2012. *The Sun King*. New York Review Books Classics. New York Review Books. https://books.google.ca/books?id=X7ines7ITAkC.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/

package=here.
Powell, A. 1955. *The Acceptance World: A Novel*. Dance to the Music of Time. Penguin Random House. https://books.google.ca/books?id=3UEFAQAAIAAJ.
———. 1957. *At Lady Molly's: A Novel*. Dance to the Music of Time. Fontana. https://books.google.ca/books?id=DmdJAAAAMAAJ.
———. 1966. *The Soldier's Art*. Dance to the Music of Time. Little, Brown. https://books.google.ca/books?id=blwaAAAAMAAJ.
———. 1968. *The Military Philosophers: A Novel*. Dance to the Music of Time. Flamingo. https://books.google.ca/books?id=ajxbAAAAMAAJ.
———. 1977. *The Kindly Ones: A Novel*. Dance to the Music of Time. Fontana. https://books.google.ca/books?id=1Y9wzgEACAAJ.
———. 1988. *A Buyer's Market*. Flamingo. https://books.google.ca/books?id=fkZMnQEACAAJ.
———. 1994. *The Valley of Bones*. https://books.google.ca/books?id=5VbZzQEACAAJ.
———. 2010. *Hearing Secret Harmonies*. Random House. https://books.google.ca/books?id=F85ta9y6t-UC.
———. 2011a. *Books Do Furnish a Room*. Random House. https://books.google.ca/books?id=B4Oy3Klb3gYC.
———. 2011b. *Casanova's Chinese Restaurant*. Random House. https://books.google.ca/books?id=wIuVMvWuI4QC.
———. 2011c. *Temporary Kings*. Random House. https://books.google.ca/books?id=1vHrhepiZ7MC.
R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Schmitt, Norbert, and Diane Schmitt. 2020. *Vocabulary in Language Teaching*. Cambridge university press.
Torruella, Joan, and Ramón Capsada. 2013. "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness." *Procedia-Social and Behavioral Sciences* 95: 447–54.
Van Hout, Roeland, and Anne Vermeer. 2007. "Comparing Measures of Lexical Richness." *Modelling and Assessing Vocabulary Knowledge* 93: 115.
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.
Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.
Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.