CBE 2

Amalia, Lauren, and Jenny

2024-11-19

Table of contents

1 Introduction

This project explores the differences in writing styles between the mid 1800s and the mid 1900s, with a focus on why older texts are often harder to understand. By examining a collection of short stories from both periods, we look at how factors like sentence length, word length, lexical density, and lexical diversity may influence the readability of older texts. This question is interesting because understanding these shifts in writing style can reveal how language and communication have changed over time, and whether these shifts have made texts more accessible or harder to read.

2 Data

The corpora used for this project are two collections of short stories from the mid 1800s and mid 1900s. The stories from the mid 1800s range from 1830s to 1850s, while the stories from the mid 1900s range from 1940s to 1950s. Given the dates chosen, almost all available texts are in the public domain and relatively easily accessible. The stories chosen come from a variety of genres and authors and are generally highly rated. Despite the small sample size, the selection of texts attempts to capture literary trends of their times. The stories were primarily sourced from Project Gutenberg, an online repository of over 70,000 eBooks. This site was chosen because it was simple to copy and paste the stories into individual text files for analysis.

Table 1: Word Count Data for Mid 1800s and Mid 1900s Texts

Time Period	Texts	Total Words	Mean Words	Word Range	Average Sentence Length	Average Word Length
Mid 1800s	8	69749	8718.625	2396 - 17574	25.44937	4.363792
Mid 1900s	13	69935	5379.615	762 - 17321	15.00980	4.062382

From Table 1, we can see that more texts from the mid 1900s were used, but the total word counts for each time period are nearly identical. This means that the average text from the mid 1800s is longer by a significant amount. More interestingly, the average sentence length for stories from the mid 1800s was much longer, and the average word length was slightly longer. Overall these corpora have similar amounts of content that can be used to draw conclusions.

3 Methods

For our initial analysis, we wanted to determine the complexity of the texts from the different time periods. In order to do so, we analyzed two different common measurements for lexical complexity - lexical density and lexical diversity (Zhou, Gao, and Lu (2023)).

First we examined the average sentence and word lengths of each of the texts from the mid 1800s and mid 1900s We did so by analyzing histograms for each period, which offered a general overview of the two comparable periods.

We then examined the lexical density of texts from the mid 1800s and mid 1900s. Lexical density is a measure of the proportion of content words to function words in a text. Content words, which include nouns, verbs, adjectives, etc, are words that carry meaning in a sentence. On the other hand, function words, such as articles, conjunctions, prepositions, etc, are used for grammatical purposes. Lexical density is higher when the proportion of content words is greater than that of function words.

We then examined lexical diversity using four different measurements - Carroll's Corrected Type Token Ratio (CTTR) (Carroll (1964)), Uber's Index (Dugast (1979), Hao et al. (2023)), and Yule's K (Choi and Jeong (2016)). We used these measurements to provide confirmation on the diversity analysis. These measurements were chosen because they all account and correct for text length sensitivities. Yule's K was chosen because it is also sensitive to frequency distributions. A low score for CTTR and Uber's Index typically indicates lower diversity when compared to a higher score. A high Yule's K score typically indicates low lexical diversity. Both signify that unique words are repeated more often within the texts. For certain measurements, we further expanded our analysis by viewing the CTTR and the Uber's Index scores through a histogram. This gave us an understanding of the trends among the individual texts within each period.

4 Results

4.1 Average Sentence and Word Length

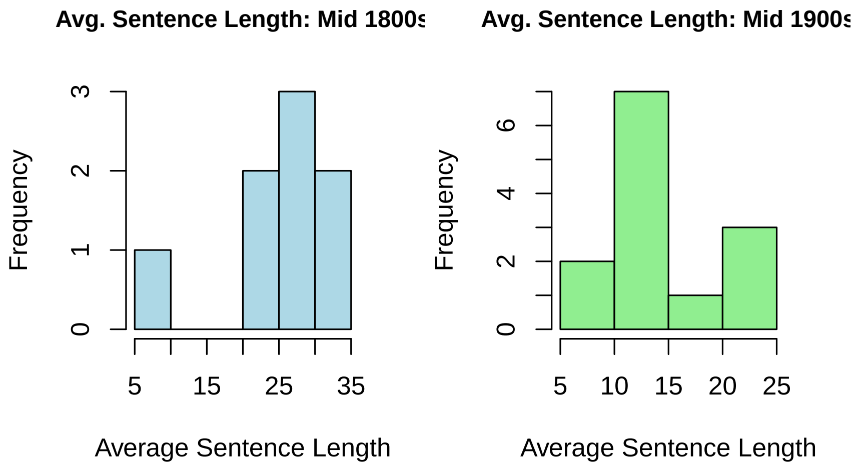

Figure 1: Distributions of Average Sentence Length

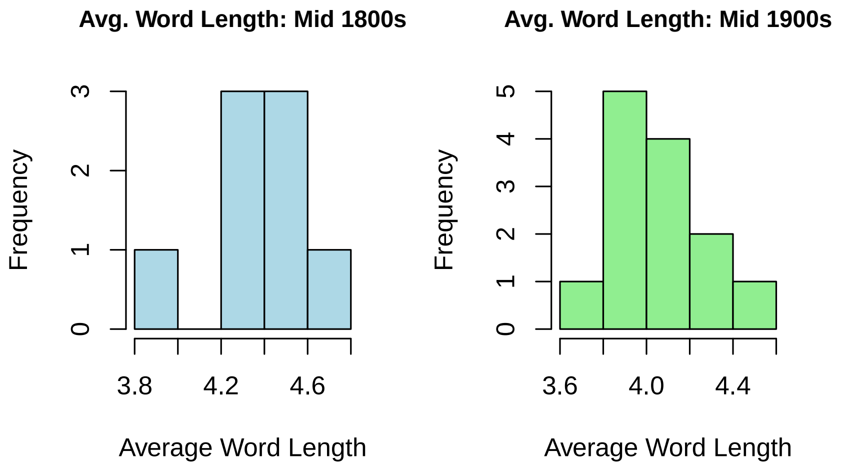

Figure 2: Distributions of Average Word Length

To begin our analysis, we looked at the distribution of average sentence length in texts from the mid 1800s and mid 1900s The histograms in Figure 1 reveal that most texts from the mid 1800s have an average sentence length between 20 and 35 words, indicating a potential preference for longer sentences during this period. In contrast, the distribution of average sentence lengths in the mid 1900s is more even, with most of the texts falling between 0 and 15 words, reflecting a possible shift towards shorter sentences.

We also examined the distribution of average word lengths. As shown in the histograms in Figure 2, the distributions for the mid 1800s and mid 1900s are similar. In both time periods, the average word length ranges from 3.6 to 4.8 characters per word.

4.2 Lexical Density

Table 2: Average Lexical Density for the Mid 1800s and Mid 1900s

Time Period	Lexical Diversity
Mid 1800s	0.4817448
Mid 1900s	0.4976068

From Table 2, we can see that the average lexical density for the texts from the mid 1800s is 0.4817, while for the 1950s, it is 0.4976. This indicates a small increase in lexical density over time, suggesting a slight rise in the proportion of content words in the mid 1900s texts. However, given the minimal difference between the two time periods, this change is unlikely to have a significant impact on the readability or complexity of the texts. A larger dataset would be needed to draw more definitive conclusions, since our sample size is relatively small.

4.3 Lexical Diversity

Table 3: Average Lexical Diversity for the Mid 1800s and Mid 1900s

Time Period	CTTR	Uber's Index	Yule's K
Mid 1800s	21.313	42.614	14.270
Mid 1900s	15.214	36.734	31.118

As seen in Table 3, for the CTTR and Uber's Index, the average lexical diversity of the texts from the mid 1800s is higher than the average lexical diversity from the 1950s. This indicates a decrease in lexical diversity between the two time periods. The Yule's K value for the mid 1800s is 14.270 where the Yule's K value for the mid 1900s is 31.118. This suggests that the texts from the mid 1900s have a more uniform distribution of word frequencies and thus less lexical diversity.

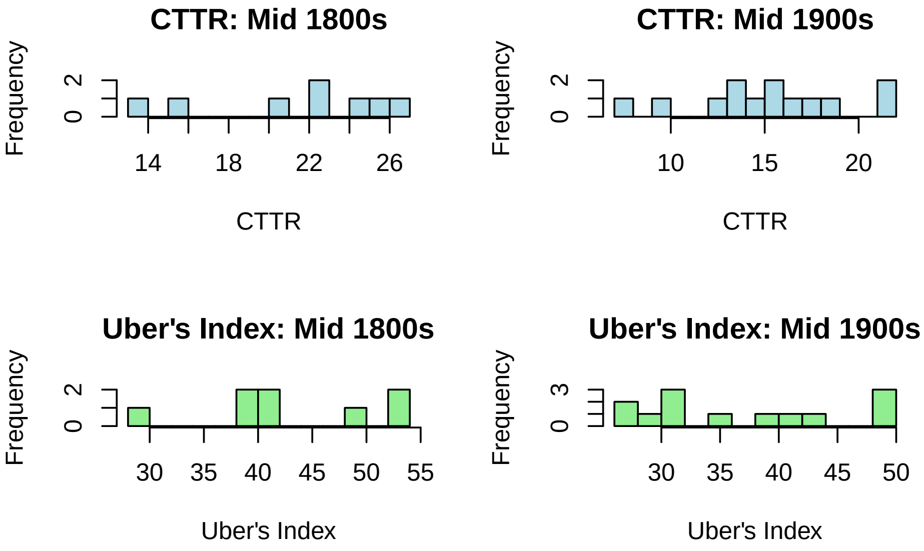

Figure 3: Distributions of CTTR and Uber's Index

For the histograms in Figure 3, we can see that, within the time periods, there is a spread of CTTR and Uber's Index values. There is no noticeable curve to the histograms, indicating mostly uniform distributions. Notably, the spread for both measures for both years is very similar. Thus, while there are specific texts within each period that greatly contribute to the previously discussed average values, many of the texts are comparably similar in lexical diversity.

5 Discussion

This exploration employed five measures for lexical complexity - four measures of lexical diversity and one measure of lexical density. Additionally, the study explored the histograms generated from mean sentence length for each of the texts. This was done to answer the primary research question concerning the change in writing styles and reader enjoyment between the 1800s and 1900s. From our exploration, we determined that the lexical density showed a small increase over time where the lexical complexity showed a decrease over time. The histograms also showed that sentences from the mid 1800s were on average longer than sentences from the mid 1900s. Thus, there is a likely decrease in lexical complexity of short stories between the two time periods, potentially lending to the preference among young adults for short stories from later time periods.

However, it is important to note that due to the small sample size, we do not have enough significance to make any conclusive statements. Thus, some limitations to consider are the relatively small size of our dataset, consisting of only about 25 texts. Expanding the dataset and incorporating a broader range of time periods beyond the mid 1800s and mid 1900s would allow us to make more meaningful statements about trends. Additionally, the texts were not randomly sampled, as we selected the top short stories suggested by ChatGPT. Ideally, a larger and more randomly selected sample would allow for more reliable conclusions about trends.

A future analysis could include more time periods and a larger variety of texts to improve the generalizability of our findings. Focusing on specific genres, such as romance or science-fiction, would allow for a better understanding of how different types of writing have evolved over time. This would help us understand whether certain genres have been more resistant to changes in style or if they have had the same characteristics over time. Additionally, analyzing how historical events or societal shifts influence language and writing styles could provide insight into how external factors influence the way people communicate.

6 Acknowledgments

ChatGPT was used to supplement our lack of knowledge of lexical analysis and finding short stories. We asked the LLM to describe some of the measures for lexical diversity and density we were considering using. We also asked it for recommendations of short stories from our two chosen time periods. The LLM was helpful in both regards as it gave us useful information on our statistical measures and pointed us towards resources we could use for building our corpus. We also utilized the texstat page for developing ideas and creating tables (Brown (2024)).

7 Works Cited

Bradbury, Ray. 1952. "A Sound of Thunder." *Collier's Weekly*. https://www.astro.sunysb.edu/fwalter/AST389/ASoundofThunder.pdf.

Brown, D. W. (n.d). 2024. "Text Analysis for Statistics & Data Science. Github.io." 2024. https://browndw.github.io/textstat_docs/.

Carroll, John B. 1964. “Language and Thought.” *Reading Improvement* 2 (1): 80.

Choi, Woonho, and HwaYoung Jeong. 2016. “Finding an Appropriate Lexical Diversity Measurement for a Small-Sized Corpus and Its Application to a Comparative Study of L2 Learners’ Writings.” *Multimedia Tools and Applications* 75: 1301513022. https://doi.org/https://doi.org/10.1007/s11042-015-2529-1.

Collins, Wilkie. 1852. “A Terribly Strange Bed.” *Household Words*. https://www.gutenberg.org/ebooks/1536.

Cummings, Ray. 1940. “The Man Who Killed the World.” *Astounding Science Fiction*. https://www.gutenberg.org/ebooks/61721.

Dugast, Daniel. 1979. *Vocabulaire Et Stylistique*. Vol. 8. Slatkine.

Eliot, George. 1859. “The Lifted Veil.” *Blackwood’s Magazine*. https://www.gutenberg.org/ebooks/2165.

Forster, E. M. 1945. “The Other Side of the Hedge.” *The Celestial Omnibus*. https://www.gutenberg.org/ebooks/34089.

Gaskell, Elizabeth. 1852. “The Old Nurse’s Story.” *Household Words*. https://www.gutenberg.org/ebooks/1404.

Greene, Graham. 1954. “The Destructors.” *The London Magazine*. https://www.ndsu.edu/pubweb/~cinichol/CreativeWriting/323/Graham%20Greene.htm.

Hao, Yuxin, Zihan Jin, Qihao Yang, Xuelin Wang, and Haitao Liu. 2023. “To Predict L2 Writing Quality Using Lexical Richness Indices: An Investigation of Learners of Chinese as a Foreign Language.” *System* 118: 103123. https://doi.org/https://doi.org/10.1016/j.system.2023.103123.

Heine, Heinrich. 1853. “Gods in Exile.” *Unpublished Work*. https://www.gutenberg.org/files/37478/37478-h/37478-h.htm.

Hemingway, Ernest. 1952. “The Old Man at the Bridge.” *Esquire*. https://biblioklept.org/2012/07/06/read-the-old-man-at-the-bridge-a-short-story-by-ernest-hemingway/.

Kornbluth, C. M. 1950. “The Mindworm.” *Worlds Beyond*. https://gutenberg.ca/ebooks/kornbluthcm-themindworm/kornbluthcm-themindworm-00-h-dir/kornbluthcm-themindworm-00-h.html.

Leiber, Fritz. 1950. “Coming Attraction.” *Galaxy Science Fiction*. https://www.gutenberg.org/files/51082/51082-h/51082-h.htm.

Melville, Herman. 1853. “Bartleby, the Scrivener.” *Putnam’s Magazine*. https://www.gutenberg.org/ebooks/11231.

O’Connor, Flannery. 1955. “Good Country People.” *A Good Man Is Hard to Find*. https://literaryfictions.com/fiction-1/good-country-people/.

O’Connor, Frank. 1950. “First Confession.” *The New Yorker*. https://www.ireland-information.com/firstconfession.htm.

Poe, Edgar Allan. 1839. “The Fall of the House of Usher.” *Burton’s Gentleman’s Magazine*. https://www.gutenberg.org/ebooks/932.

———. 1842. “The Masque of the Red Death.” *Graham’s Magazine*. https://www.gutenberg.org/ebooks/1064.

Salinger, J. D. 1948. “A Perfect Day for Bananafish.” *The New Yorker*. https://www.newyorker.com/magazine/1948/01/31/a-perfect-day-for-bananafish.

Thurber, James. 1942. "The Catbird Seat." *The New Yorker*. https://www.newyorker.com/magazine/1942/11/14/the-catbird-seat.

Turgenev, Ivan. 1852. "The District Doctor." *A Sportsman's Sketches*. https://www.gutenberg.org/ebooks/159.

Updike, John. 1951. "A & p." *The New Yorker*. https://www.newyorker.com/magazine/1961/07/22/ap.

Verne, Jules. 1954. "Master Zacharius." *Unpublished Translation*. https://www.gutenberg.org/ebooks/3538.

Zhou, Xinye, Yuan Gao, and Xiaofei Lu. 2023. "Lexical Complexity Changes in 100 Years' Academic Writing: Evidence from Nature Biology Letters." *Journal of English for Academic Purposes* 64: 101262. https://doi.org/https://doi.org/10.1016/j.jeap.2023.101262.