Pavel Ramirez, Carol Zhou

Dr. Roberto Franzosi

SOC 446W Big/Small Data & Visualization

9 February 2025

SOC 446W Homework 3

Introduction

The intersection of Natural Language Processing (NLP) and literary analysis has opened up new possibilities for understanding text on a deeper level. By leveraging computational tools, scholars can now analyze vast corpora efficiently to identify linguistic trends, stylistic patterns, and thematic shifts that were previously difficult. Building upon previous homework, this assignment starts with an overview of the corpus statistics for two literary works — *Harry Potter and the Order of the Phoenix* and *The Scarlet Letter* — on which all analyses will be based. The paper then delves deeper into the use of the NLP Suite to examine N-grams, co-occurrences, and their broader implications in culturomics. Through analyses of text readability, sentence complexity, vocabulary richness, and nominalization, it examines how different measures influence the style of a corpus. Additionally, this paper discusses the role of gender and vocabulary in shaping writing style. By engaging with scholarly literature, this assignment aims to contextualize these computational approaches within the broader landscape of digital humanities.

Methods

The assigned corpora are *Harry Potter and the Order of the Phoenix* (hereafter referred to as *Harry Potter Book 5*) by J.K. Rowling and *The Scarlet Letter* by Nathaniel Hawthorne. *Harry Potter Book 5*, published in 2003, is the fifth book in the Harry Potter series and contains 38

chapters. Among young and modern readers, it is a fantasy saga known for Rowling's clear and accessible writing style and vivid descriptions ("The Writing Style of J. K. Rowling"). In contrast, published in 1850, *The Scarlet Letter* has 24 chapters. Hawthorne's writing style is "ornate and subtle", usually marked by long, complex sentences and symbolism (SparkNotes). The distinct literary genres and historical periods of these two texts make them excellent subjects for comparative analysis, particularly in terms of thematic exploration, sentence complexity, and stylistic nuances.

To gain a general understanding of each corpus, document statistics were generated for both texts using CORPUS/DOCUMENT Analysis Tools with the "Compute document statistics (*)" option. From this output, total sentence, word, and syllable count were calculated and visualized. Boxplots were then created to display word count and sentence length per chapter using data_visualization_2_main. The resulting figures for each corpus were analyzed individually and in comparison to each other.

One useful tool to identify differences and similarities in style, tone, and language structure is N-grams. An N-gram is "a sequence of N words" (Kumar). A unigram consists of a single word (e.g., "machine"), a bigram consists of two words (e.g., "machine learning"), a trigram consists of three words (e.g., "advanced machine learning"), and so on. Some N-grams appear more frequently than others. For example, "machine learning" probably appears more than "advanced machine learning". Thus, probabilities can be assigned to N-grams to indicate how likely they occur or are to follow a given sequence of words. This is useful for several applications. First, it helps identify which N-grams should be grouped as single entities. Secondly, it can assist in making predictions about the next word in a sequence, such as in autocomplete features. For instance, when composing an email, after the greeting, it is more

common to write “I hope” followed by “you are doing well” rather than “you have a good day.” In addition, N-grams are helpful in spelling correction, as words like “archive” are more frequently used than misspellings like “arhieve.”

In literary analysis, N-grams have been applied in the field of culturomics — “the application of massive scale data collection and analysis to the field of human culture” (TEDx Talks 00:13:25-00:11:38) — to study temporal shifts in language and how word frequency reflects cultural trends over time. They have also been used to examine authorial style since some authors use certain combinations of words and characters more than other authors. Furthermore, with tools like the NLP Suite, it is possible to analyze whether particular N-grams appear more frequently at the beginning, middle, or end of a document, or if they are distributed across the text. Such patterns can offer valuable insights into an author’s distinctive style (Franzosi).

Mazzocchi’s article discusses the idea that the rise of Big Data is changing the nature of scientific inquiry, leading to a debate on the role of theory in science. It explores the arguments for and against data-driven research, highlighting the importance of both inductive and deductive approaches in scientific discovery. The article also touches on the limitations of purely automated text analysis and emphasizes the need to combine quantitative and qualitative approaches to gain a more comprehensive understanding of language use in texts.

In the context of this assignment, Mazzocchi’s insights are relevant because they highlight the potential and limitations of using computational tools like the NLP Suite for literary analysis. While the NLP Suite can efficiently analyze large amounts of text and identify patterns that might be missed with traditional methods, it is important to remember that it cannot fully capture the nuances of language, such as context, irony, and authorial intent. Therefore, it is

important to combine quantitative analysis with qualitative interpretation to gain a more complete understanding of the text (Mazzocchi 1254).

Furthermore, Mazzocchi's exploration of culturomics sheds light on its potential and limitations in literary analysis. While culturomics enables researchers to track shifts in language and culture over vast periods and offer insights that would be impossible with traditional methods, it can overlook the nuance of individual texts and the complexity of language. It often relies on frequency counts rather than qualitative interpretation, and a focus on quantitative analysis may miss the subtleties of meaning, context, and authorial intent that are essential for a deeper understanding of literature.

To compute N-grams, the "N-grams & Co-occurrences" option was selected under CORPUS/DOCUMENT Analysis Tools. The "Compute N-grams" option was ticked with the "word" setting to generate word-based N-grams rather than character-based ones. Two runs were performed. In the first run, only unigrams were computed with the following settings: Case insensitive, Exclude punctuation (word N-grams only), Exclude articles (word N-grams only), Exclude determiners (word N-grams only), Exclude ALL stopwords (word N-grams only), and Lemmatize. These settings were used unigrams are more likely to include punctuation or stopwords, which would not contribute meaningfully to the analysis. In the second run, N-grams up to six-grams were computed with only the "Case insensitive" and "Exclude determiners (word N-grams only)" settings. This is because, starting with bigram, the possibility of stopwords appearing in N-grams increases, and removing them would limit the analysis. N-grams were computed for both *Harry Potter Book 5* and *The Scarlet Letter*, and the results were analyzed individually and then compared. However, in *The Scarlet Letter*, no prominent five-grams were identified, so the N-gram analysis was limited to four-grams.

The most significant words identified in the N-gram analysis were further examined for their relationships with surrounding words using the search function, and the results were visualized using the Co-Occurrences VIEWER. For *Harry Potter Book 5*, "Harry" and "Hermione" were searched, while for *The Scarlet Letter*, "Hester" and "Pearl" were searched. Two runs were performed for each corpus: the first to compute co-occurrence frequencies in sentences, and the second to compute co-occurrences in documents. Only the Co-Occurrences VIEWER was used because the N-grams VIEWER requires input files with dates embedded in the file names. Since the assigned corpora files only include chapter numbers in the file names, N-grams VIEWER could not be run. While Google N-gram Viewer is a powerful tool for analyzing N-grams across millions of books, it cannot be used for visualizing N-grams in custom corpora like those in this assignment. This is why the NLP Suite VIEWER was chosen for this analysis, as it allows for visualizing N-grams specific to our own corpus (Franzosi).

Then, these words were also searched using the "Search text file(s) for words/collocations" option under CORPUS/DOCUMENT Analysis Tools. The search returned the words immediately preceding and following the searched words, along with one sentence before and after each occurrence of the searched words in the results. A Sankey flowchart was generated for each corpus to understand the words preceding and following the target words using data_visualization_2_main. The settings applied were: a maximum of 10 words for variable 1, 5 for variable 2, and 30 for variable 3. The results reveal the surrounding words, helping to identify frequent collocations and word pairings. The returned sentences before and after the searched word could provide context, showing whether the target word carries a positive or negative connotation, its relationship to other words, and how it contributes to the overall meaning. Moreover, sentences without the searched word can show contrasts or gaps in

understanding. If the word appears in a discussion but then is not used in other areas, it could indicate ideological or conceptual differences.

To evaluate how text readability, sentence complexity, vocabulary richness, nominalization, and various other vocabulary measures affect the style, the two corpora were analyzed using the "Style analysis (ALL options GUI)" option under CORPUS/DOCUMENT Analysis Tools. Text readability refers to how easily a text can be understood, while sentence complexity is measured by the number of subordinate clauses or clauses within a sentence. Vocabulary richness is assessed through Yule's K. Unlike readability and complexity, the larger the Yule's K value, the less diverse the vocabulary (Franzosi). Given the multitude of vocabulary measures, only a select few were chosen for this analysis, as they offer insights into how vocabulary affects the style of the assigned corpora. These include "Vocabulary (via Hapax Legomena, once-occurring words)", "Abstract/Concrete Vocabulary", and "Punctuation as Figures of Pathos (? !)". Both the entire corpus was copied and pasted into the GenderGuesser to determine an author's gender based on the words used.

Nominalization is the process of changing a verb or adjective into a noun (Franzosi 26). For instance, "interfere" becomes "interference", "decide" becomes "decision", and "argue" becomes "argument". It can influence style by making writing sound more formal and help convey complex ideas more concisely and objectively. However, nominalization can also shift focus from the action (the verb) to the result or concept (the noun) and leads to more abstract and less engaging sentences (SchoolTube Community). For example, as discussed in Dr. Franzosi's paper, the sentence "The mob lynched Sam Hose" can be nominalized as "The lynching of Sam Hose". In the nominalized sentence, the mob, who is the agent of the action, is no longer

explicitly mentioned. This can make it seem as if the lynching was an event that happened without any specific cause or perpetrator.

The Franzosi paper also discusses verb voice and agency. Verb voice refers to the relationship between the verb and the subject of the sentence. In active voice, the subject acts as the verb, while agency refers to the capacity of individuals or groups to act independently and make their own choices. All three of these grammatical choices have a collective effect on the text, and the authors of the paper argue that the use of nominalization and passive verb voice in newspaper articles about lynchings served to obscure the agency of the white perpetrators and to reinforce the existing racial power structure in the Jim Crow South. These choices can influence several aspects of corpora, a lot of which can be subjects of analysis, which include analysis of implied power structures, ideological stances, aspects of character relationships in fictional texts, and evaluating the reliability of information.

Results and Discussions

Harry Potter Book 5 and the Scarlet Letter Corpus Statistics

The top right panel of Figure 1 reveals that the total word count for *Harry Potter Book 5* (n = 257,111) is significantly higher than that of *The Scarlet Letter* (n = 82,784), which results in a greater total number of sentences and syllables in *Harry Potter Book 5* as well. Figure 2 also shows a notable variation in word count between chapters: the chapter with the most words in *Harry Potter Book 5* contains 8,924 words, while the chapter with the fewest only has only 3,819. Similarly, *The Scarlet Letter* exhibits a wide range in chapter word counts, with its longest chapters containing 14,774 words and shortest having only 487, as shown in Figure 3. Despite this, most chapters in *Harry Potter Book 5* actually contain more words per chapter than those in *The Scarlet Letter*. This is indicated by the first quartile of *Harry Potter Book 5* (5,823 words)

being greater than the third quartile of *The Scarlet Letter* (3,443.5 words), and it suggests that *Harry Potter Book 5* has longer chapters. Moreover, while the sentence length distributions for the two corpora overlap, *The Scarlet Letter* has longer sentences on average, with a median sentence length of 26, compared to 17 in *Harry Potter Book 5*, as presented in Figures 4 and 5. This means that sentences in *The Scarlet Letter* are likely to be more complex, which aligns with Hawthorne's reputation for using lengthy, intricate sentence structures. Given these findings, differences in thematic depth, sentence complexity, and overall style are expected between the two corpora in the subsequent analysis.

Figure 1. Total sentence (top left), word (top right), and syllable (bottom) count in Harry Potter Book 5 and The Scarlet Letter, calculated from the NLP_corpus_stats_Dir_Harry Potter Book 5.csv and NLP_corpus_stats_Dir_The Scarlet Letter.csv files.

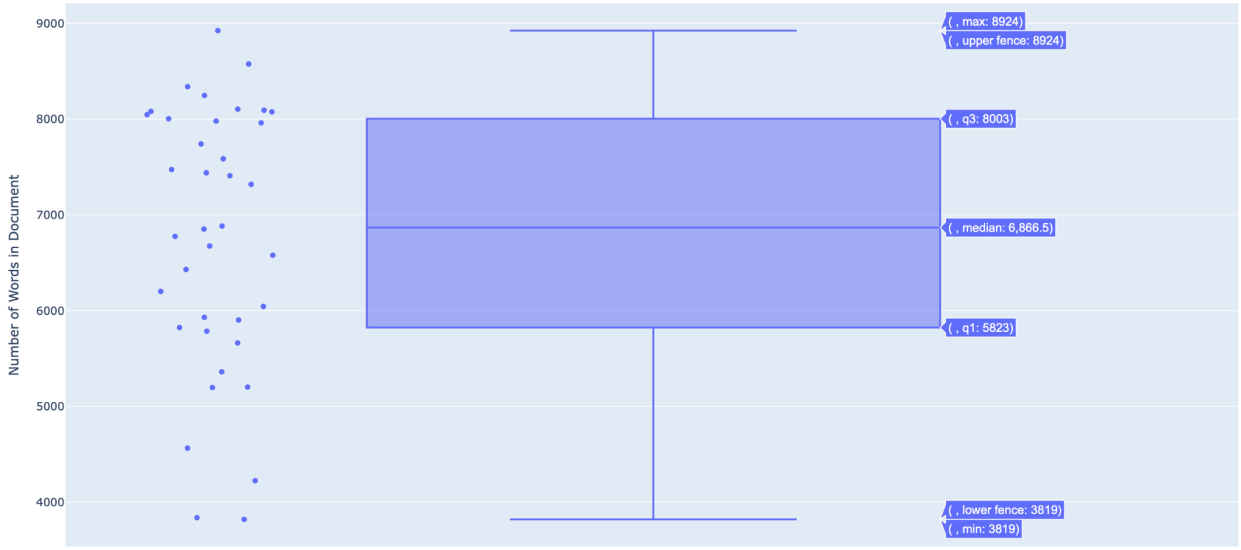

Figure 2. Boxplot of word count per chapter in Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_corpus_stats_Dir_Harry Potter Book 5.csv as input.

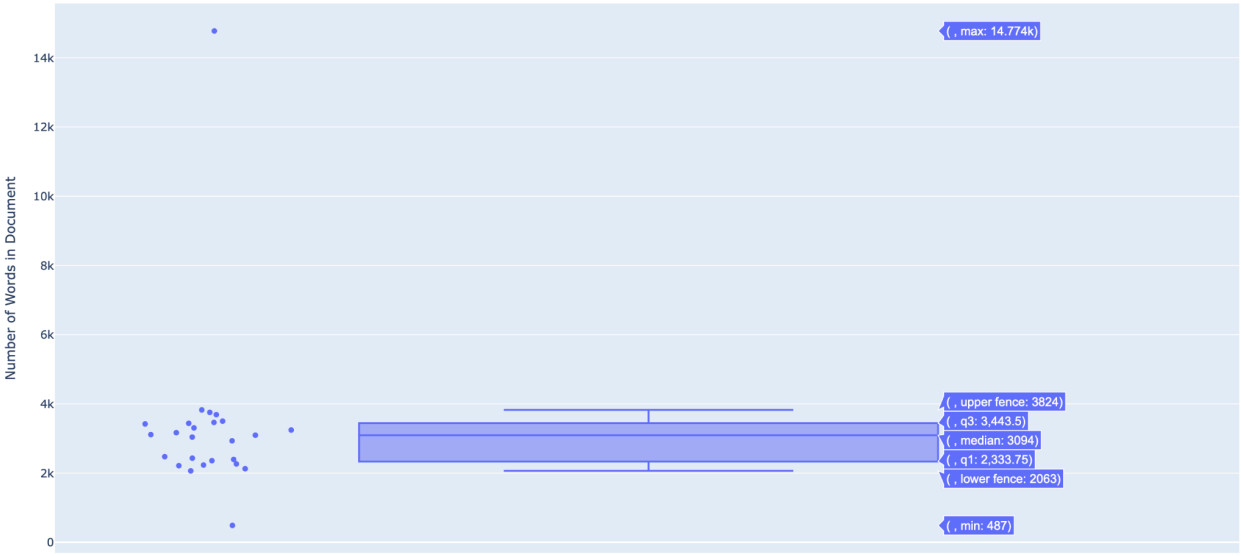

Figure 3. Boxplot of word count per chapter in The Scarlet Letter, generated using data_visualization_2_main with the file NLP_corpus_stats_Dir_The Scarlet Letter.csv as input.

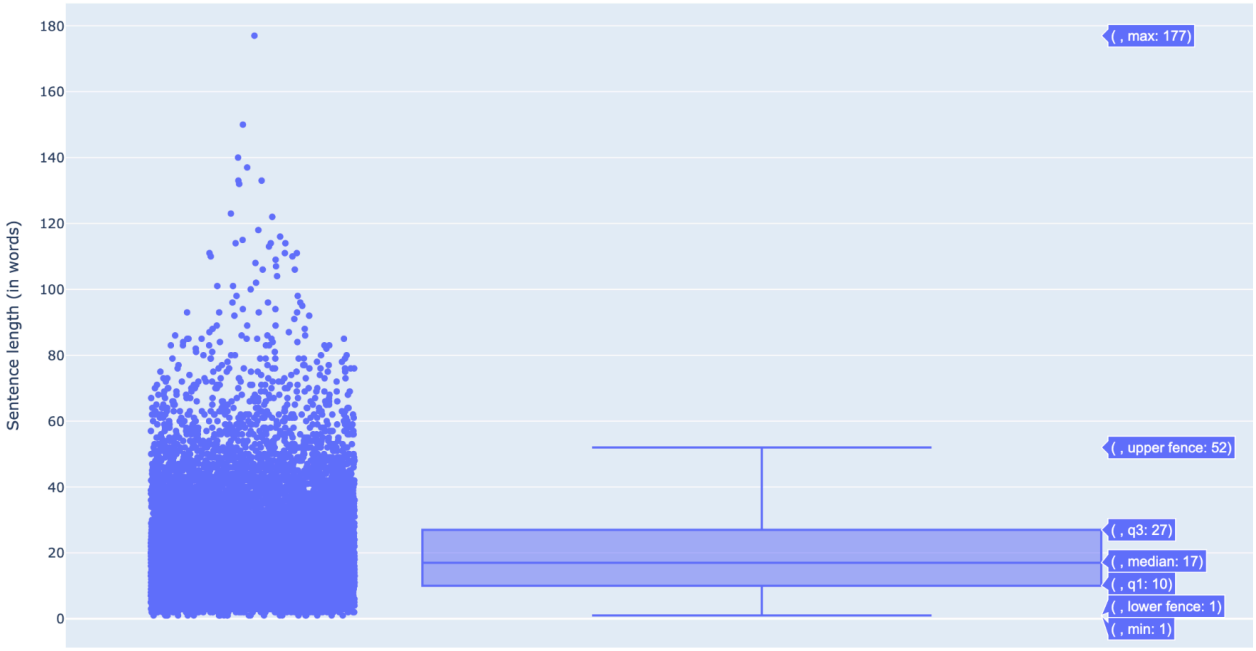

Figure 4. Boxplot of sentence length in Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_sentence_length_Dir_Harry Potter Book 5.csv as input.

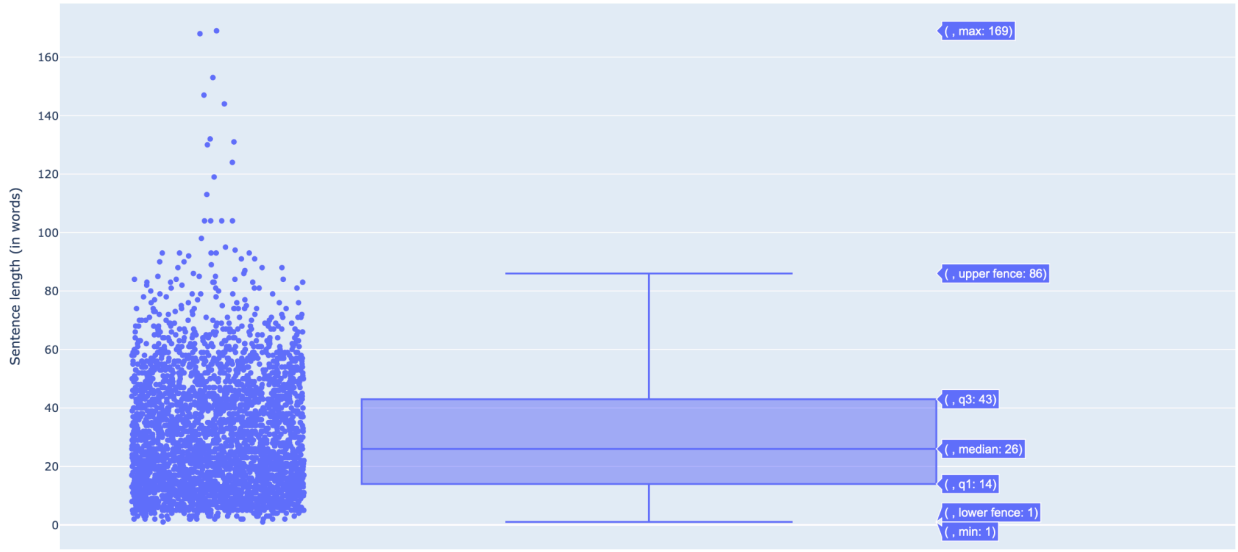

Figure 5. Boxplot of sentence length in The Scarlet Letter, generated using data_visualization_2_main with the file NLP_sentence_length_Dir_The Scarlet Letter.csv as input.

N-grams and Co-occurrences

According to Figure 6, verbs such as “say”, “look”, and “know” appear frequently in *Harry Potter Book 5*, which is mirrored in *The Scarlet Letter*, as shown in Figure 7. This suggests a strong focus on dialogue and action in both texts. Prominent character names such as “Harry”, “Ron”, and “Hermione” are prevalent in *Harry Potter Book 5*, while *The Scarlet Letter* features names such as “Hester”, “Pearl”, and “Dimmesdale” (referred to as “minister”). This highlights the character-driven narratives at the heart of both works. Both figures also reveal content-specific words. For instance, *Harry Potter Book 5* includes terms like “Dumbledore”, “Sirius”, and “professor”, which are tied to the magical setting, characters, and plot elements specific to Hogwarts. Conversely, *The Scarlet Letter* features words like “mother”, “letter”, and “heart”, reflecting the themes of sin, shame, motherhood, and its Puritan context. Moreover, the frequent use of “thou” — an archaic form of “you” — in *The Scarlet Letter* highlights its historical and linguistic context, reinforcing the formal, religious tone of the period.

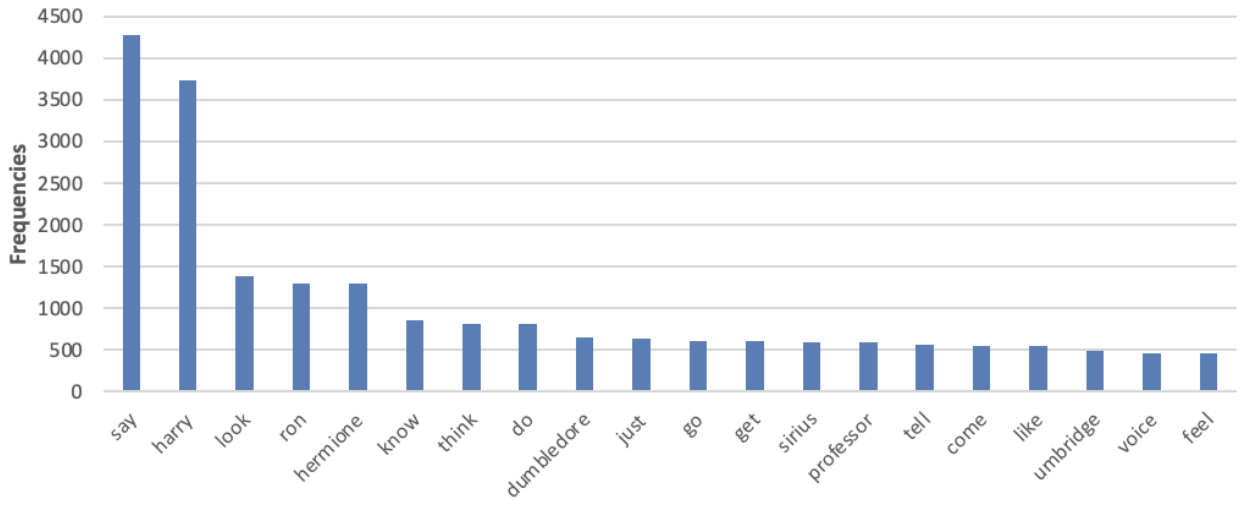

Figure 6. Frequency distribution of unigrams in Harry Potter Book 5, showing the top 20 most frequent unigrams with a cutoff frequency of 464. This figure was created and annotated using the NLP_n-grams4_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

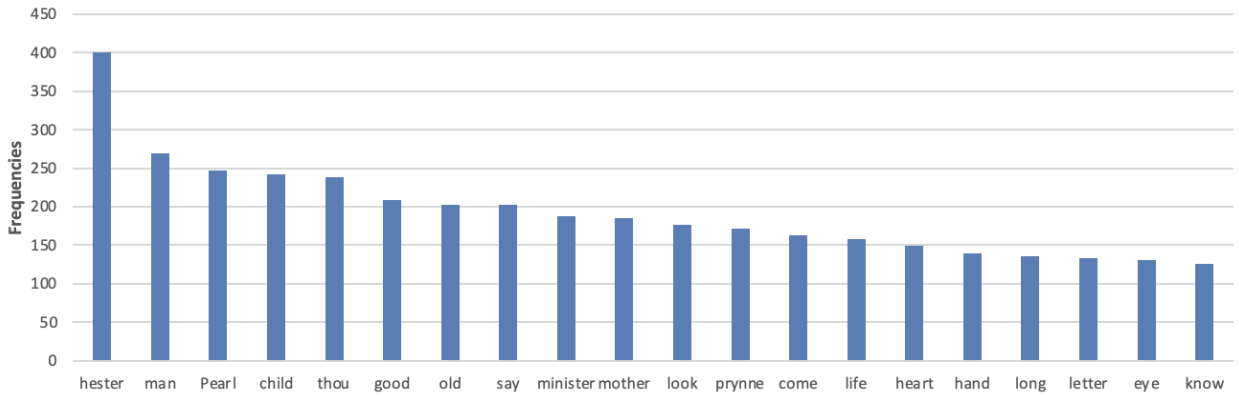

Figure 7. Frequency distribution of unigrams in The Scarlet Letter, showing the top 20 most frequent unigrams with a cutoff frequency of 126. This figure was created and annotated using the NLP_n-grams1_Word_Dir_The Scarlet Letter_stats_Bar_chart.xlsx file.

Figures 8 and 9 reveal similar patterns observed in the unigrams as in the bigrams for both corpora. Both texts emphasize dialogue. In *Harry Potter Book 5*, bigrams such as "said Harry", "said Hermione", "said Ron", "Harry and" and "Ron and" are prevalent, focusing on the trio's interactions and relationships. In *The Scarlet Letter*, the bigram "said Hester" similarly points to key character dialogues. Character names also feature prominently in both texts. For example, in *Harry Potter Book 5*, in addition to "Harry", "Hermione", and "Ron", names like "Mrs. Weasley" and "Professor Umbridge" appear a lot. In contrast, *The Scarlet Letter* includes bigrams such as "Hester Prynne", "Mr. Dimmesdale", "Roger Chillingworth", and "Governor Bellingham", emphasizing key characters and their formal relationships. Additionally, in *The Scarlet Letter*, "New England" is a direct reference to the setting and its historical context, while "scarlet letter" alludes to the central theme of the book. Differences in narrative styles are also present. "Thou hast" and "wilt thou" in *The Scarlet Letter* reflect the language of the 19th century, while *Harry Potter Book 5*'s bigrams use more contemporary and informal language (e.g., "he was").

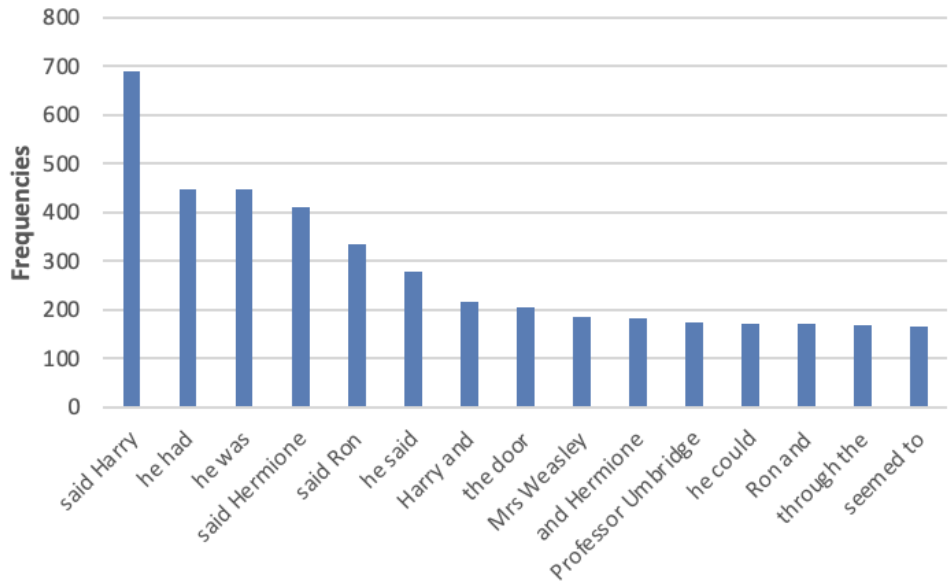

Figure 8. Frequency distribution of bigrams in Harry Potter Book 5, showing the top 15 most frequent bigrams with a cutoff frequency of 165. This figure was created and annotated using the NLP_n-grams2_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

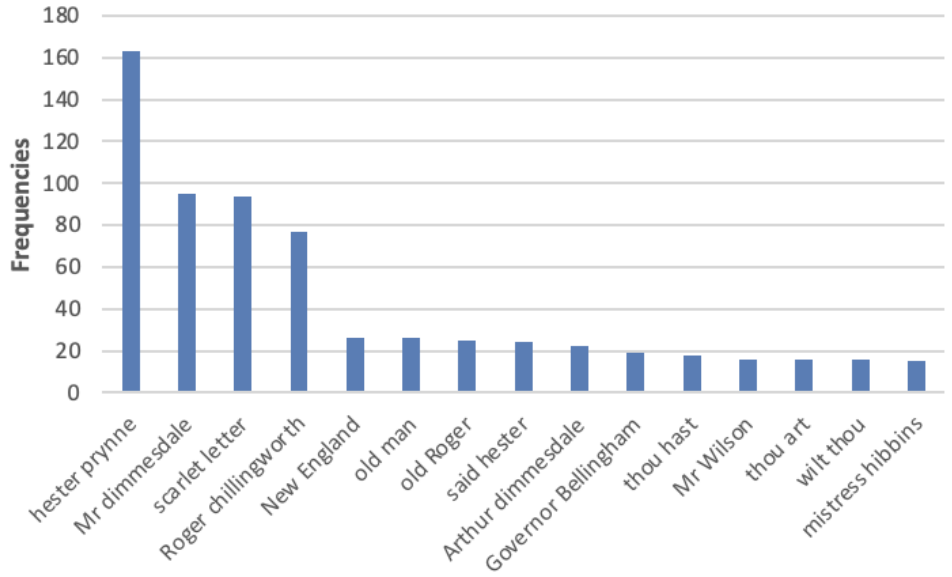

Figure 9. Frequency distribution of bigrams in The Scarlet Letter, showing the top 15 most frequent bigrams with a cutoff frequency of 15. This figure was created and annotated using the NLP_n-grams2_Word_Dir_The Scarlet Letter_stats_Bar_chart.xlsx file.

In Figure 10, the trigrams in *Harry Potter Book 5* heavily emphasize the trio's frequent exchanges, as evident in phrases like "Ron and Hermione" and "Harry and Ron". In comparison, *The Scarlet Letter*'s trigrams, such as "Reverend Mr. Dimmesdale", "old Roger Chillingworth", and "the old man", are more focused on character names paired with descriptions, suggesting a more formal and descriptive narrative style, as shown in Figure 11. Furthermore, trigrams like "in front of", "as though he", and "for a moment" in *Harry Potter Book 5* reflect a contemporary style of phrasing.

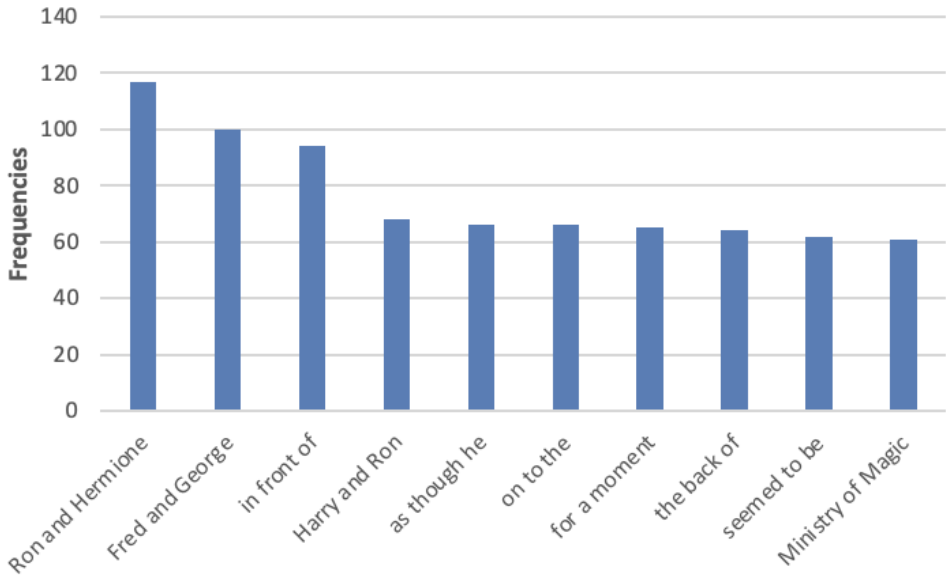

Figure 10. Frequency distribution of trigrams in Harry Potter Book 5, showing the top 10 most frequent trigrams with a cutoff frequency of 61. This figure was created and annotated using the NLP_n-grams3_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

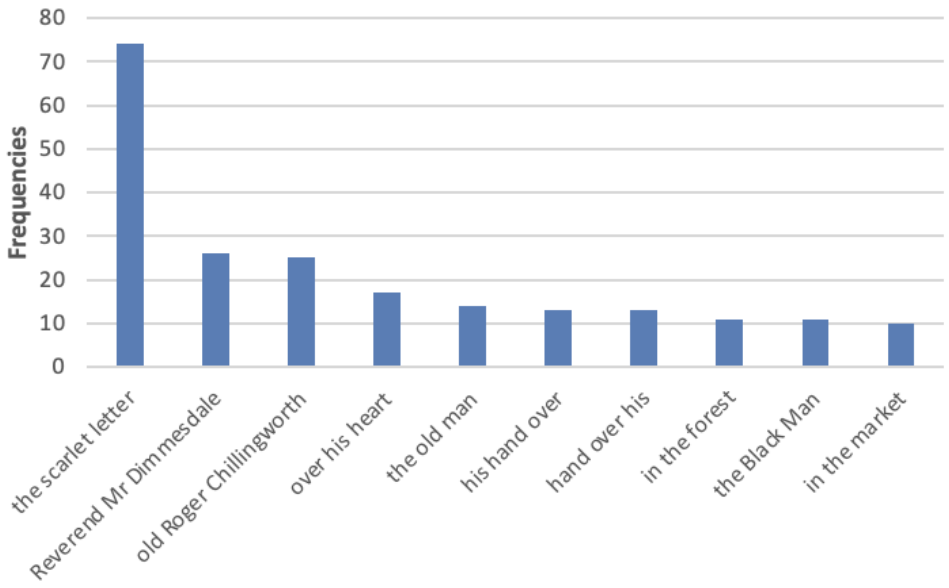

Figure 11. Frequency distribution of trigrams in The Scarlet Letter, showing the top 10 most frequent trigrams with a cutoff frequency of 11. This figure was created and annotated using the NLP_n-grams3_Word_Dir_The Scarlet Letter_stats_Bar_chart.xlsx file.

Figures 12, 14, and 15, which display the frequency distributions of four-to-six-grams in *Harry Potter Book 5*, show more context-specific phrases such as “Defense Against the Dark Arts”, “the Order of the Phoenix”, and “The Department of Mysteries”, all of which are central to the plot. On the other hand, in Figure 13, *The Scarlet Letter* highlights four-grams that reflect the novel’s themes of sin and guilt, with phrases like “hand over his heart” and “letter on her breast” prominently appearing.

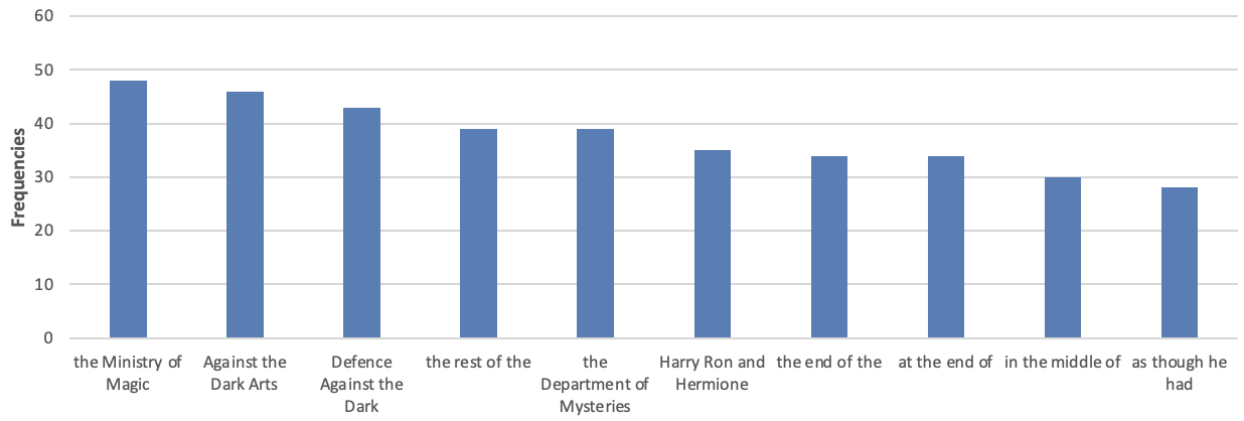

Figure 12. Frequency distribution of four-grams in Harry Potter Book 5, showing the top 10 most frequent four-grams with a cutoff frequency of 28. This figure was created and annotated using the NLP_n-grams4_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

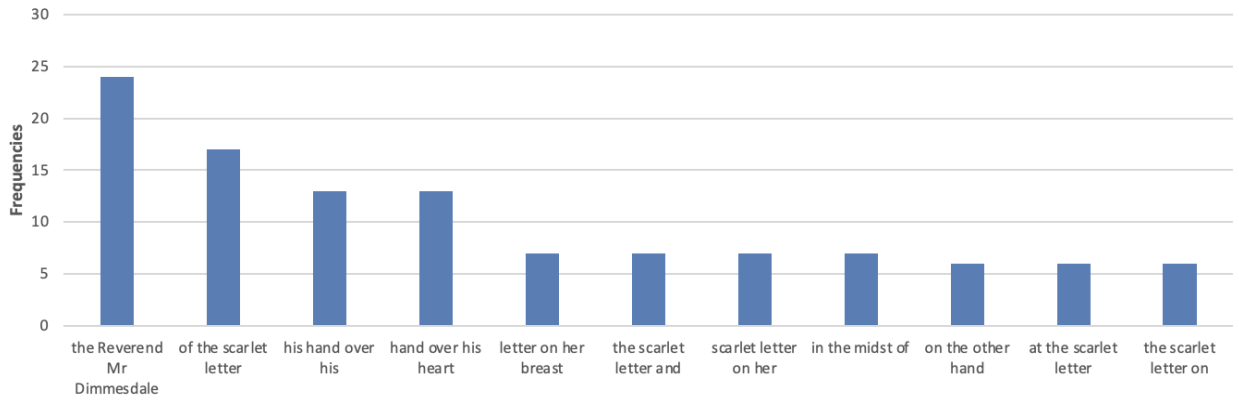

Figure 13. Frequency distribution of four-grams in The Scarlet Letter, showing the top 11 most frequent four-grams with a cutoff frequency of 6. This figure was created and annotated using the NLP_n-grams4_Word_Dir_The Scarlet Letter_stats_Bar_chart.xlsx file.

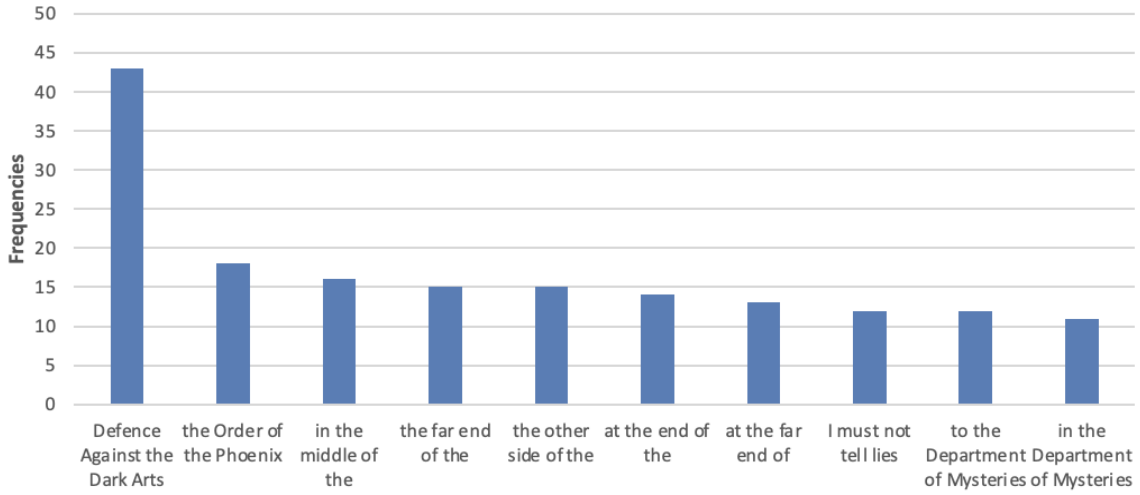

Figure 14. Frequency distribution of five-grams in Harry Potter Book 5, showing the top 10 most frequent five-grams with a cutoff frequency of 11. This figure was created and annotated using the NLP_n-grams5_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

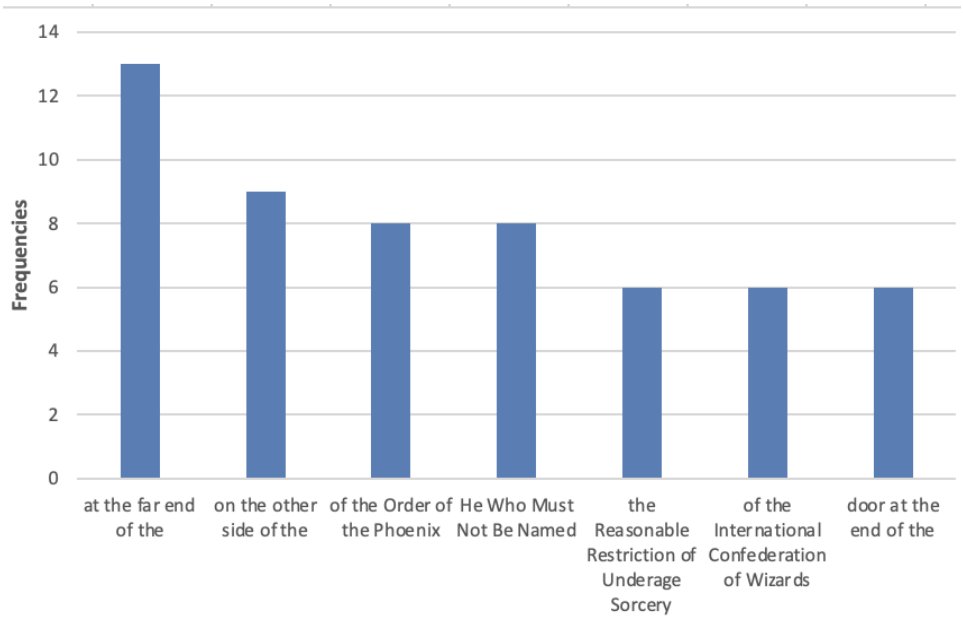

Figure 15. Frequency distribution of six-grams in Harry Potter Book 5, showing the top 7 most frequent six-grams with a cutoff frequency of 6. This figure was created and annotated using the NLP_n-grams6_Word_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

The analysis of n-grams in *Harry Potter Book 5* and *The Scarlet Letter* reveals notable similarities and differences in their narrative styles. Both texts emphasize dialogue and character-driven storytelling, with frequent use of character names and verbs like “say”, “look”, and “know”, reflecting their focus on interaction. *Harry Potter Book 5* leans on contemporary language and phrases tied to its magical setting, such as “Dumbledore” and “the Order of the Phoenix”, while *The Scarlet Letter* features archaic language and content-specific terms like “thou” and “the scarlet letter”, highlighting its historical and moral themes. The bigrams in both works focus on key character exchanges, but *The Scarlet Letter* uses more formal, descriptive language, while *Harry Potter Book 5* showcases casual, conversational phrases. Additionally, the trigrams in both novels reflect their respective narrative styles: *Harry Potter Book 5* emphasizes the trio’s relationships, while *The Scarlet Letter* often pairs character names with descriptive phrases, reinforcing the novel's formal tone.

Co-occurrences VIEWER

From previous N-grams analysis, “Harry” and “Hermione” did not appear together in any N-grams. Therefore, it is worth examining whether these characters co-occur in any sentences. Figure 16 shows that “Harry” and “Hermione” only co-occur in 2.46% of all the sentences in *Harry Potter Book 5*. This suggests that, while they are central characters, they are not frequently mentioned together within the same sentence. However, the figure also reveals that 92.11% of chapters contain both “Harry” and “Hermione”, indicating that they are almost always present in the same broader sections of the book. On the other hand, 1.54% of sentences in *The Scarlet Letter* contain both “Hester” and “Pearl”, suggesting that they are mentioned together in the same sentence a bit less often than “Harry” and “Hermione”. Similar to “Harry” and “Hermione”, “Hester” and “Pearl” also appear in 72% of the documents. This again confirms

that they are typically found within the same broader sections of the book but with slightly less consistency than Harry and Hermione.

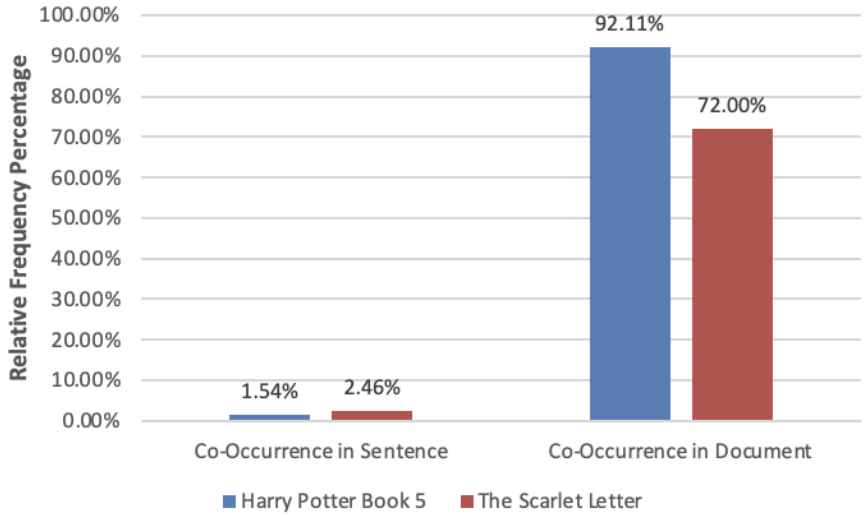

Figure 16. Relative frequency percentage of co-occurring words "Harry" and "Hermione" for Harry Potter Book 5 and "Hester" and "Pearl" for The Scarlet Letter in sentences and documents. Relative frequency percentage is the proportion of sentences/documents that contain a co-occurrence ("YES") relative to the total number of sentences/documents (both with and without a co-occurrence).

Words/Collocations Searches

Examining the word clouds in Figure 17, the most prominent words in each cloud are "Harry" and "Hermione" (left) and "Hester" and "Pearl" (right), which is expected as these were the search terms. In *Harry Potter Book 5*, words like "said", "Dumbledore", and "Ron" highlight the importance of relationships, while terms like "looked", "looking", and "thought" suggest a focus on observation and reactions. On the other hand, In *The Scarlet Letter*, words like "mother", "child", and "minister" emphasize the core relationships of the novel. Words like "bosom", "scarlet", and "sin" suggest the exploration of sin and the societal implications of Hester's actions. While *Harry Potter Book 5* seems to focus more on character interactions and

reactions between them, *The Scarlet Letter* leans more towards moral themes and character relationships within a stricter societal context.

Figure 17. Word clouds generated for sentences containing the searched words "Harry" and "Hermione" from Harry Potter and the Order of the Phoenix (left) and "Hester" and "Pearl" from The Scarlet Letter (right), created using the "Search text file(s) for words/collocations" option under the CORPUS/DOCUMENT Analysis Tools.

In Figure 18, in *Harry Potter Book 5*, “said” is the most prominent word preceding both “Harry” and “Hermione”. This reinforces that N–grams results and highlights the importance of dialogue in the narrative. In addition, punctuation marks such as commas and periods are notably frequent both before and after the searched words. In The Scarlet Letter, Figure 19 similarly shows that commas frequently precede and follow both “Hester” and “Pearl” prominently. Furthermore, “little” is commonly found preceding “Pearl”, while “Prynne” follows “Hester”. These patterns are consistent with the N-gram analysis, further supporting the observed language trends.

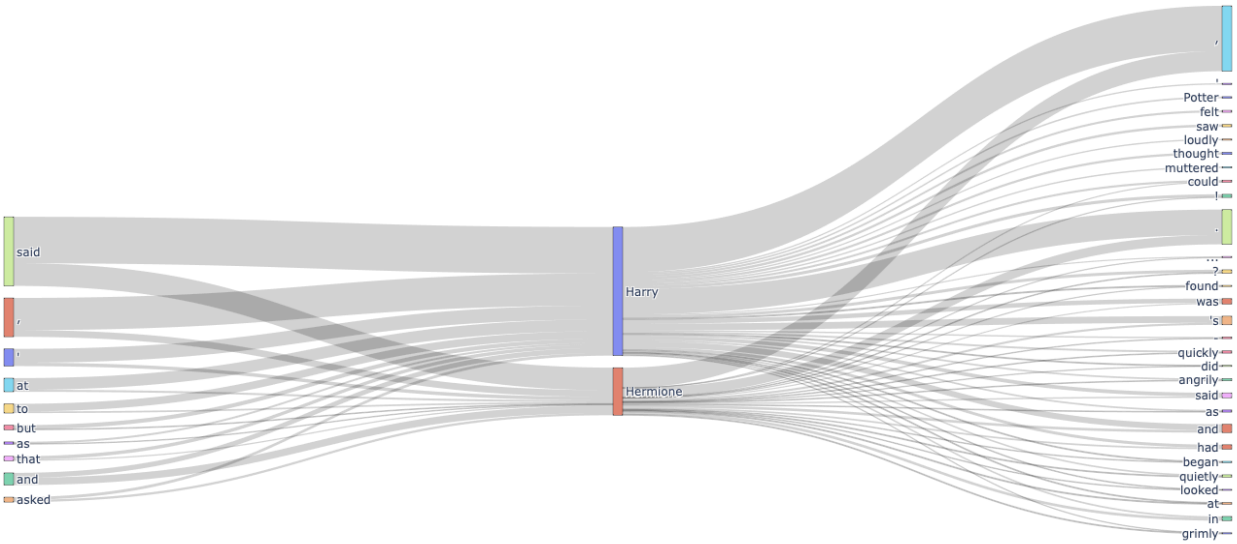

*Figure 18. Sankey chart depicting the preceding and following words of "Harry" and "*Hermione*" in Harry Potter Book 5, generated using the NLP_search_word_sent_Dir_Harry Potter Book 5.csv file.*

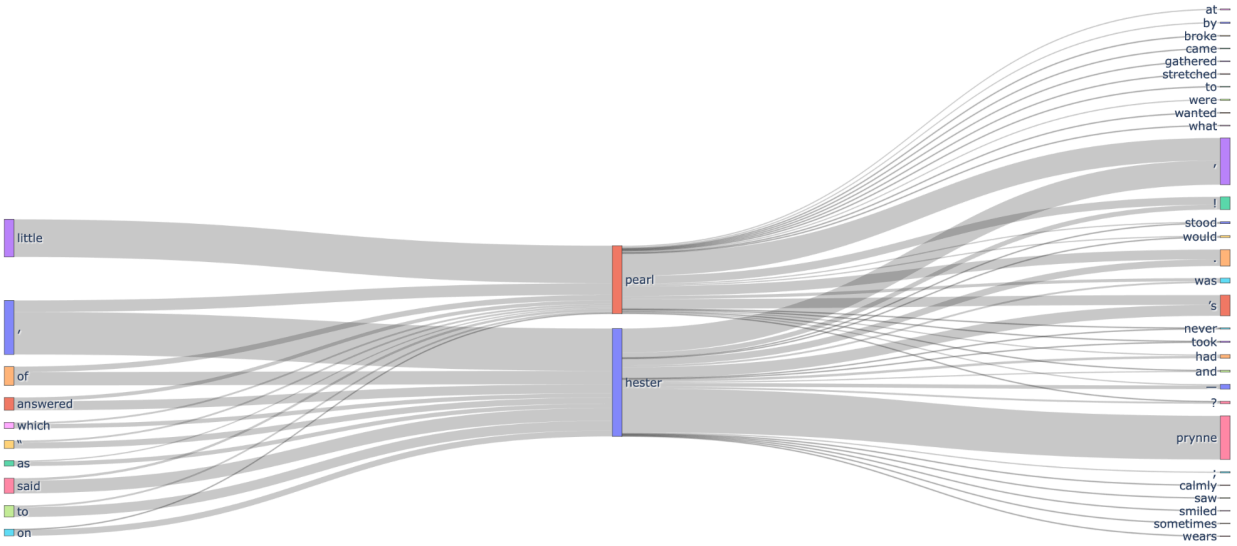

Figure 19. Sankey chart depicting the preceding and following words of "Hester" and "Pearl" in The Scarlet Letter, generated using the NLP_search_word_sent_Dir_The Scarlet Letter.csv file.

Style

Figure 20 shows that the highest frequent consensus readability grade level is 9th and 10th grade for *Harry Potter Book 5*. This indicates that this text is accessible to a broad audience, likely targeting middle school readers. In contrast, *The Scarlet Letter*'s highest consensus readability grade level is 12th and 13th Grade, as shown in Figure 21. This suggests that this corpus employs more complex language and themes, making it more suitable for high-school or college-level readers. The text-readability results align with the previous findings that *The Scarlet Letter* contains longer and more complex sentences because of the use of archaic language, dense prose, and intricate themes.

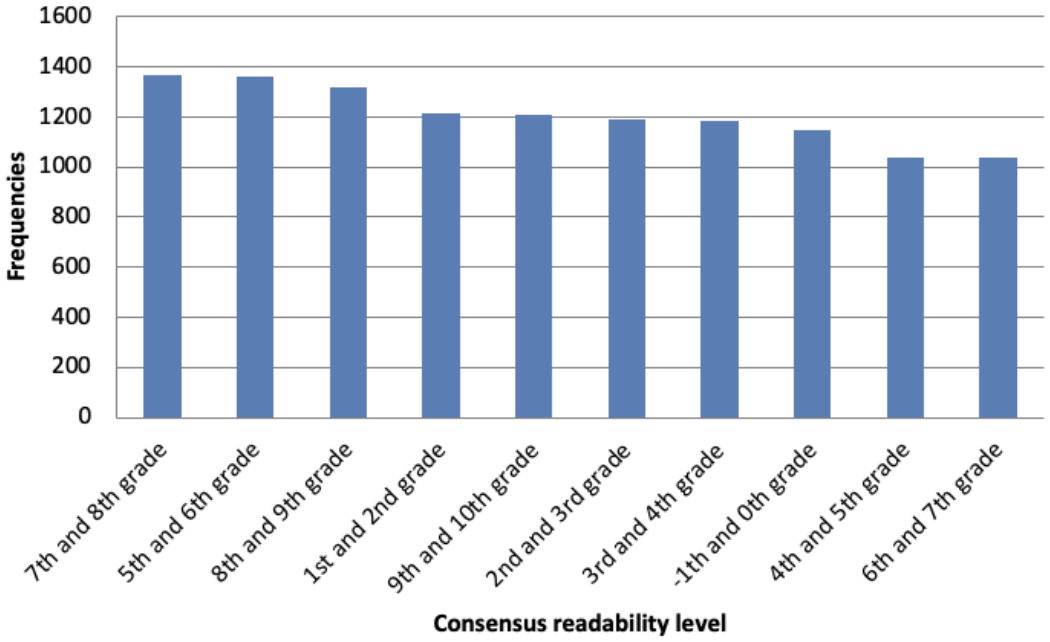

Figure 20. Text readability frequencies of overall readability consensus in Harry Potter Book 5, showing the top 10 most frequent consensus readability levels, which account for 75.63% of the sentences. This is created and annotated using the NLP_cons_READ_Dir_Harry Potter Book 5_stats_Bar_chart.xlsx file.

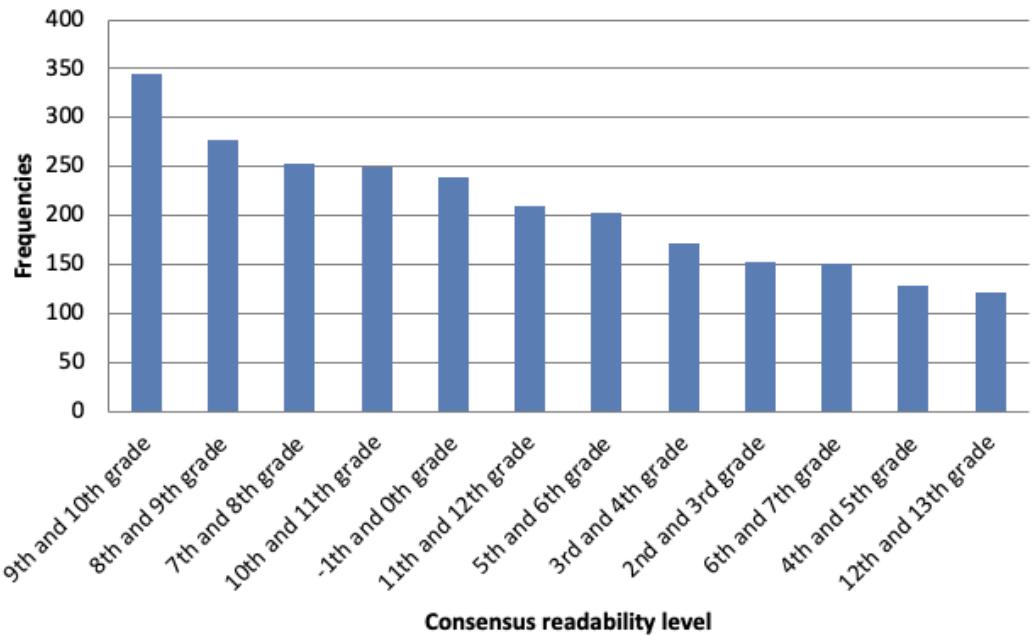

Figure 21. Text readability frequencies of overall readability consensus in The Scarlet Letter, showing the top 12 most frequent consensus readability levels, which account for 75.41% of the sentences. This is created and annotated using the NLP_cons_READ_Dir_The Scarlet Letter_stats_Bar_chart.xlsx file.

According to Figures 22 and 23, although the sentence complexity score distributions of the two corpora overlap, *The Scarlet Letter* has a higher median Yngve score of 93 compared to *Harry Potter Book 5*'s median score of 59. A higher Yngve score indicates more complex sentences, typically with more embedded clauses, longer phrases, and a greater number of dependencies between words. Thus, a "typical" sentence in The Scarlet Letter tends to be more complex on average than that in *Harry Potter Book 5*. Interestingly, the spread of the sentence complexity scores in *The Scarlet Letter* is wider than that of *Harry Potter Book 5*. This means that *The Scarlet Letter* contains sentences that vary significantly in complexity, from quite simple to very intricate. These figures again support the earlier observation that *The Scarlet Letter* contains more complex sentence structures overall.

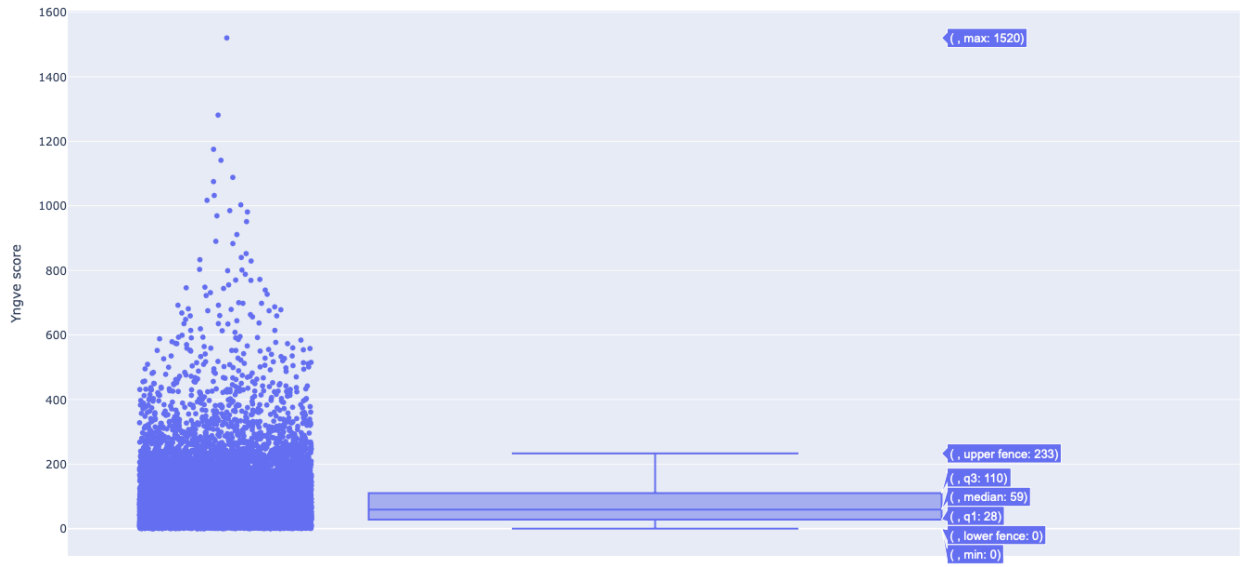

Figure 22. Boxplot showing the distribution of Yngve sentence complexity scores for all sentences in Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_SentenceComplexity_Dir_Harry Potter Book 5.csv as input.

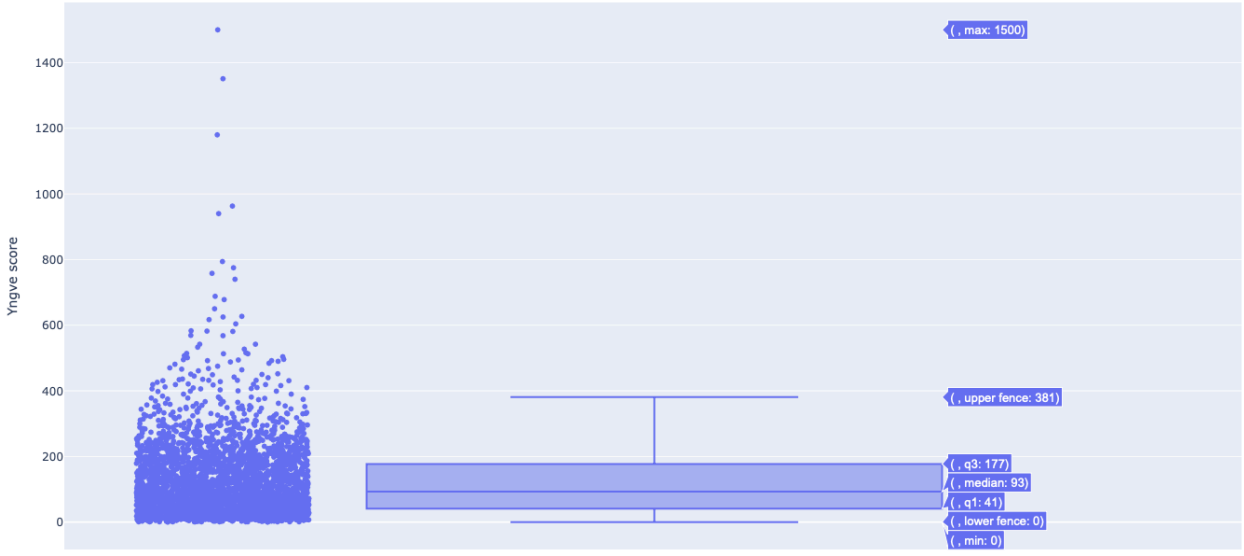

Figure 23. Boxplot showing the distribution of Yngve sentence complexity scores for all sentences in The Scarlet Letter, generated using data_visualization_2_main with the file NLP_SentenceComplexity_Dir_The Scarlet Letter.csv as input.

Though the Yule's K score distributions for the two corpora in Figures 24 and 25 overlap, the median Yule's K for *Harry Potter Book 5* is 9.445, which is lower than the median for *The Scarlet Letter*, which is 19.23. This suggests that a "typical" chapter in Harry Potter Book 5 tends to have a greater vocabulary richness on average. Moreover, the distribution for *Harry Potter Book 5* is narrower, indicating less variation in vocabulary richness across chapters. In contrast, *The Scarlet Letter* shows a wider range, indicating more fluctuation in vocabulary richness between chapters. The greater vocabulary diversity in *Harry Potter Book 5* could be attributed to its fantasy genre, which often introduces a broader array of topics and neologisms (such as magic spells), whereas classic literature such as The Scarlet Letter may focus on more specific themes with a more established vocabulary.

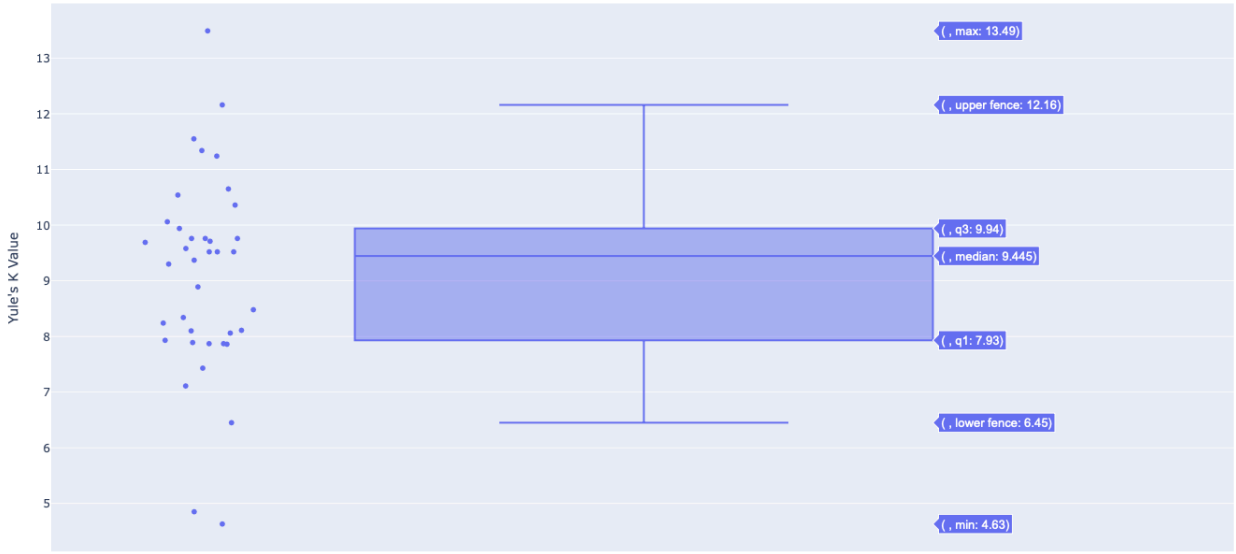

Figure 24. Boxplot showing the distribution of Yule's K vocabulary richness scores across chapters in Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_Yule K_Dir_Harry Potter Book 5.csv as input.

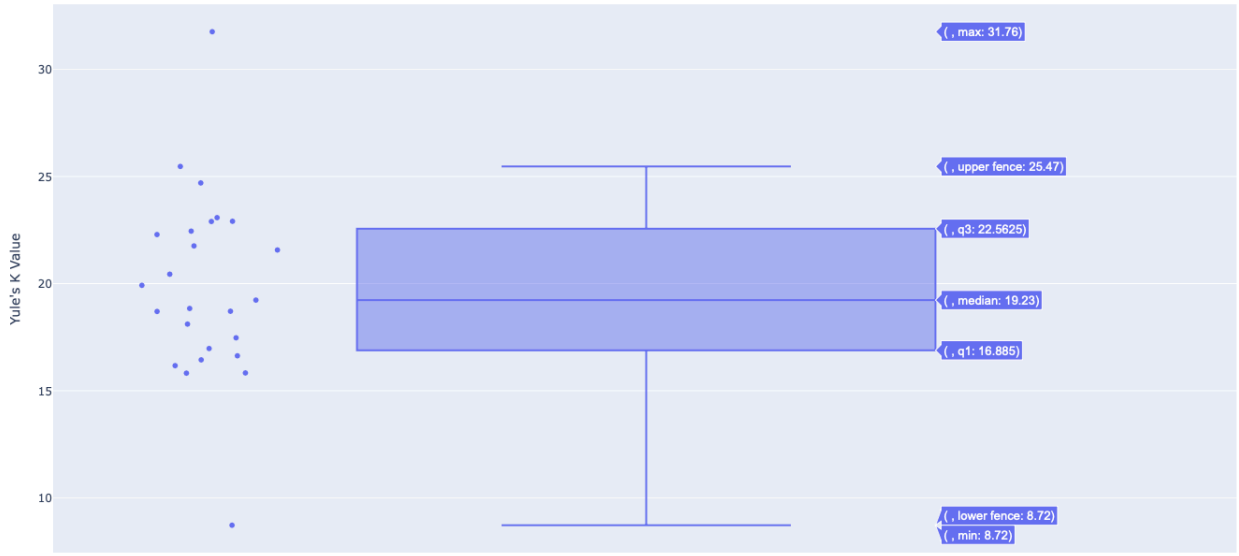

Figure 25. Boxplot showing the distribution of Yule's K vocabulary richness scores across chapters in The Scarlet Letter, generated using data_visualization_2_main with the file NLP_Yule K_Dir_The Scarlet Letter.csv as input.

Examining the frequency distributions of nominalized verbs in Figure 26, the most frequent nominalized verb in *Harry Potter Book 5* is "feeling", highlighting a focus on emotions and internal states. In comparison, *The Scarlet Letter* features "expression" as the most frequent nominalized verb, emphasizing communication, as shown in Figure 27. Interestingly, "feeling" and "expression" both appear in both figures. The specific nominalized verbs that appear most frequently suggest differences in thematic focus. Harry Potter Book 5 leans towards emotions and conflict, while The Scarlet Letter seems to focus more on communication, identity, and societal roles. In addition, when comparing the distribution of nominalization percentages in sentences (Figures 28 and 29), although the two distributions overlap, the median for *Harry Potter Book 5* is slightly higher than that of *The Scarlet Letter*. This suggests that, on average, *Harry Potter Book 5* contains slightly more sentences with nominalized verbs. While both

authors use nominalization to abstractly express actions and qualities, Rowling employs it more frequently.

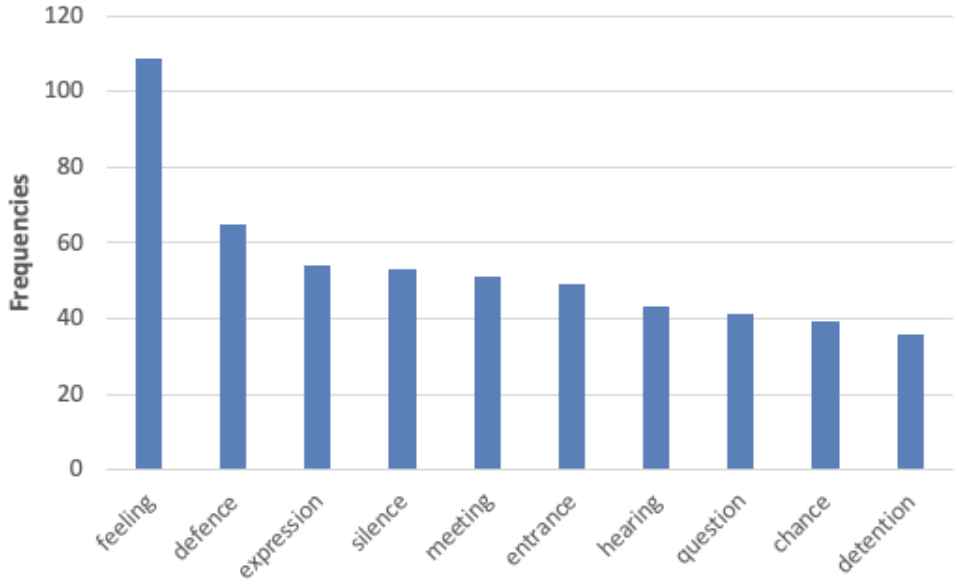

Figure 26. Frequency distribution of nominalized verbs in Harry Potter Book 5, showing the top 10 most frequent verbs with a cutoff frequency of 36. This figure was created and annotated using the NLP_NOM_verb_NOM_Dir_Harry Potter Book 5_nom_verb_freq_Bar_chart.xlsx file.

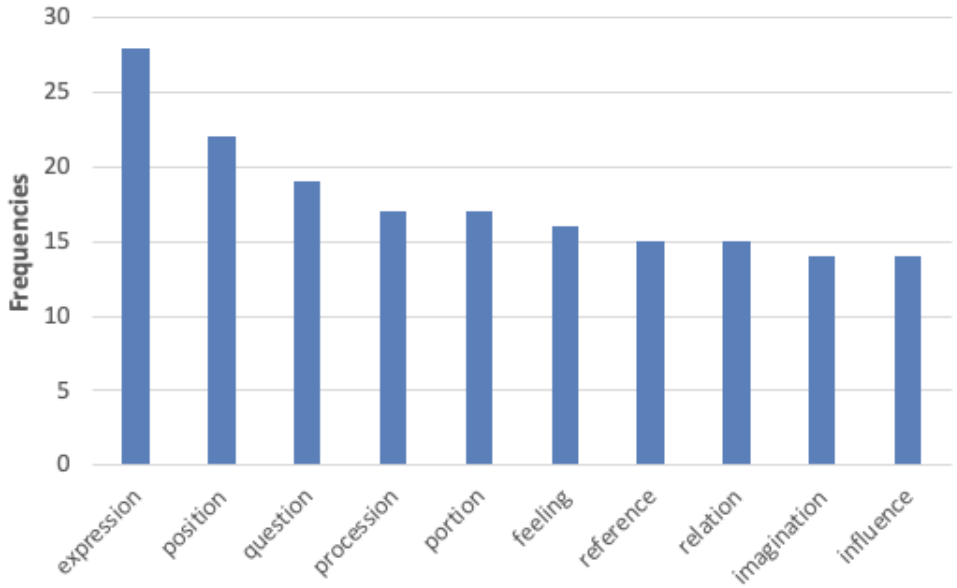

Figure 27. Frequency distribution of nominalized verbs in The Scarlet Letter, showing the top 10 most frequent verbs with a cutoff frequency of 14. This figure was created and annotated using the NLP_NOM_verb_NOM_Dir_The Scarlet Letter_nom_verb_freq_Bar_chart.xlsx file.

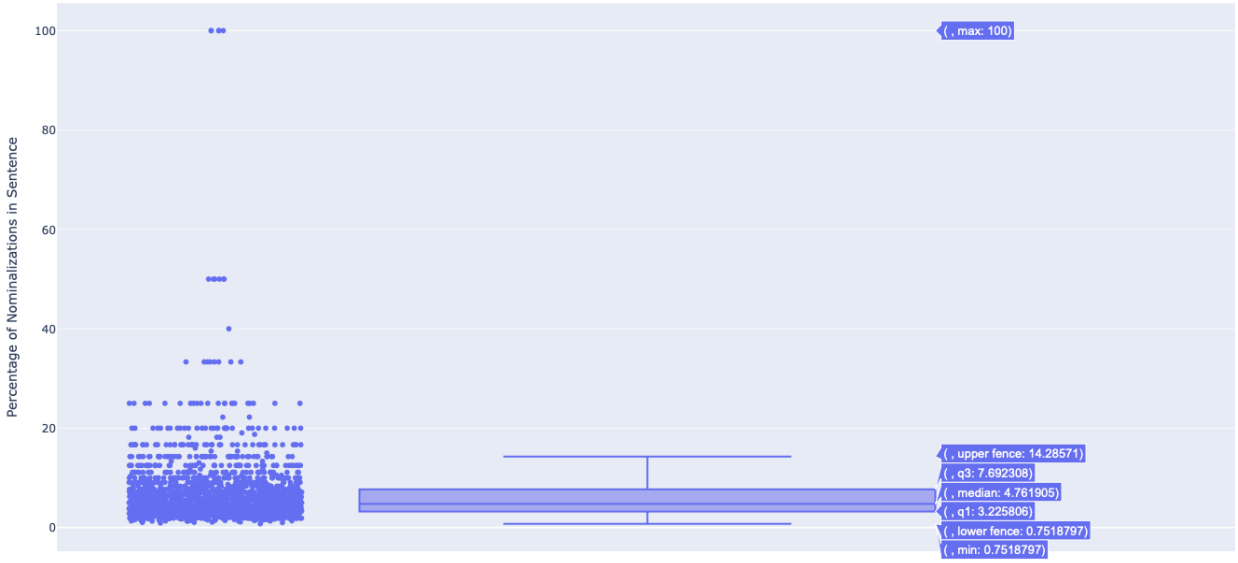

Figure 28. Boxplot showing the distribution of the percentage of nominalization in sentences from Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_NOM_Dir_Harry Potter Book 5_freq_bySent.csv as input.

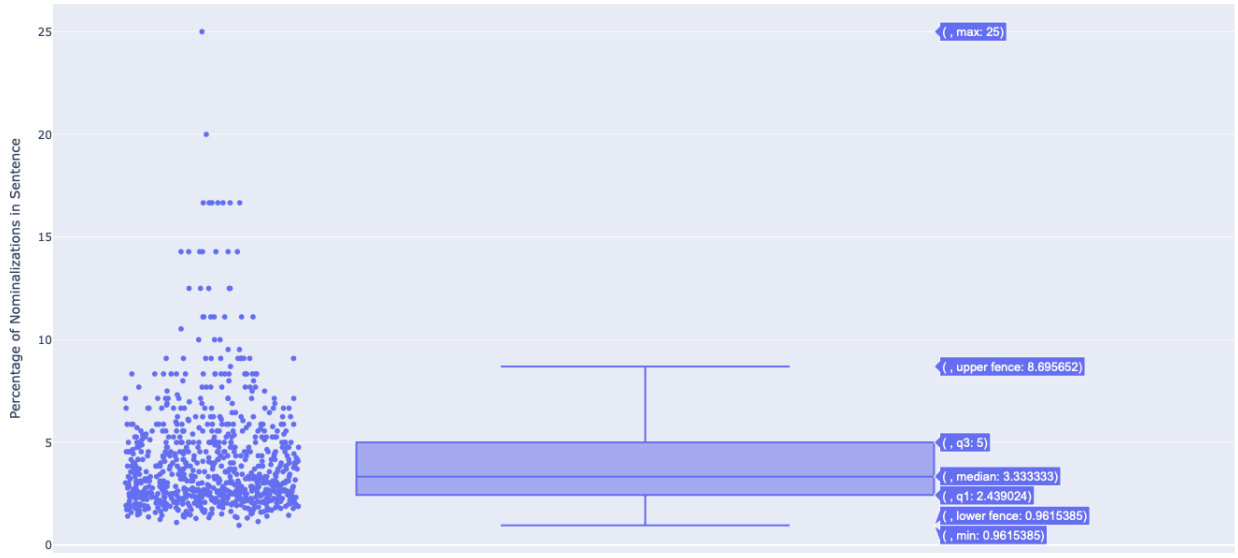

Figure 29. Boxplot showing the distribution of the percentage of nominalization in sentences from The Scarlet Letter, generated using data_visualization_2_main with the file NLP_NOM_Dir_The Scarlet Letter_freq_bySent.csv as input.

The NLP Suite offers several options for analyzing a corpus's vocabulary besides Yule's K vocabulary richness. Therefore, "Vocabulary (via Hapax Legomena, once-occurring words)", "Abstract/Concrete Vocabulary", and "Punctuation as Figures of Pathos (? !)" measures were selected because they highlight aspects of vocabulary richness, word usage patterns, and stylistic traits. Starting with hapax legomena, which are words that only appear once in the entire corpus, their frequency can indicate the diversity or uniqueness of a text's vocabulary and provide an additional measure to cross-check with Yule's K values. As shown in Figure, in *Harry Potter Book 5*, there are 4,530 hapax legomena, compared to 3,301 in *The Scarlet Letter*. After calculating the relative frequency of hapax legomena (the number of hapax legomena divided by the total number of words), only 1.76% of words in *Harry Potter Book 5* occur once, while 3.99% of words in *The Scarlet Letter* are hapax legomena, suggesting a slightly broader or more specialized vocabulary in Hawthorne's work. Even though this result contradicts the previous finding of higher vocabulary richness in *Harry Potter Book 5* based on Yule's K, it makes sense given Hawthorne's reputation for using more complex and specialized language. Meanwhile, *Harry Potter Book 5's* lower Hapax Legomena frequency can be explained by its fantasy genre, where magical terms and spells, though infrequent, are repeated throughout the book, leading to a lower percentage of words that only appear once.

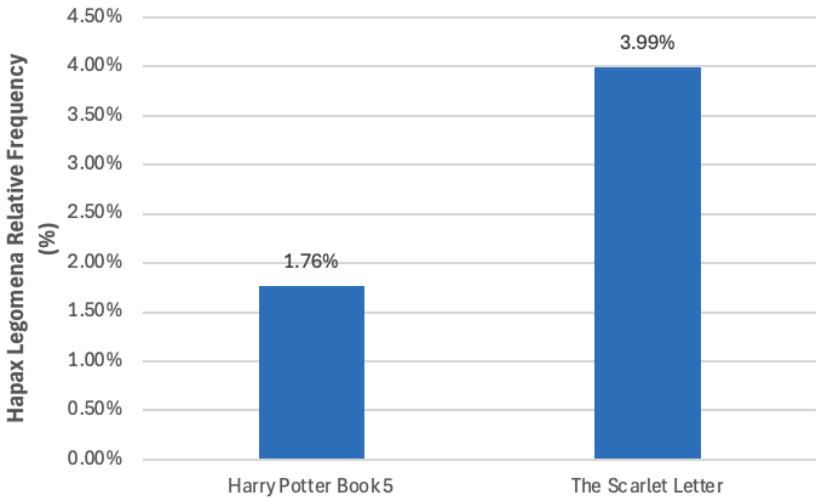

Figure 30. Hapax Legomena relative frequency distribution in Harry Potter Book 5 and The Scarlet Letter, generated using data_visualization_2_main with the files NLP_n-grams1_Word_Hapax_Dir_Harry Potter Book 5_stats.csv and NLP_n-grams1_Word_Hapax_Dir_The Scarlet Letter_stats.csv.

Next, looking at the abstract (emotions, concepts) versus concrete (tangible, physical) vocabulary in a corpus can reveal differences in narrative tone and focus. In Figures 31 and 32, the spread of both median concreteness (on a scale of 1 to 5, where 1 is abstract and 5 is concrete) distributions is quite similar, though the median concreteness value for *Harry Potter Book 5* (median = 3.12) is slightly higher than that of *The Scarlet Letter* (median = 2.92). This suggests that, in general, *Harry Potter Book 5* uses a more concrete vocabulary and has a more grounded, descriptive style. This aligns with Rowling's reputation for clear, accessible writing and vivid descriptions, making it natural for her to employ more concrete words to convey her story.

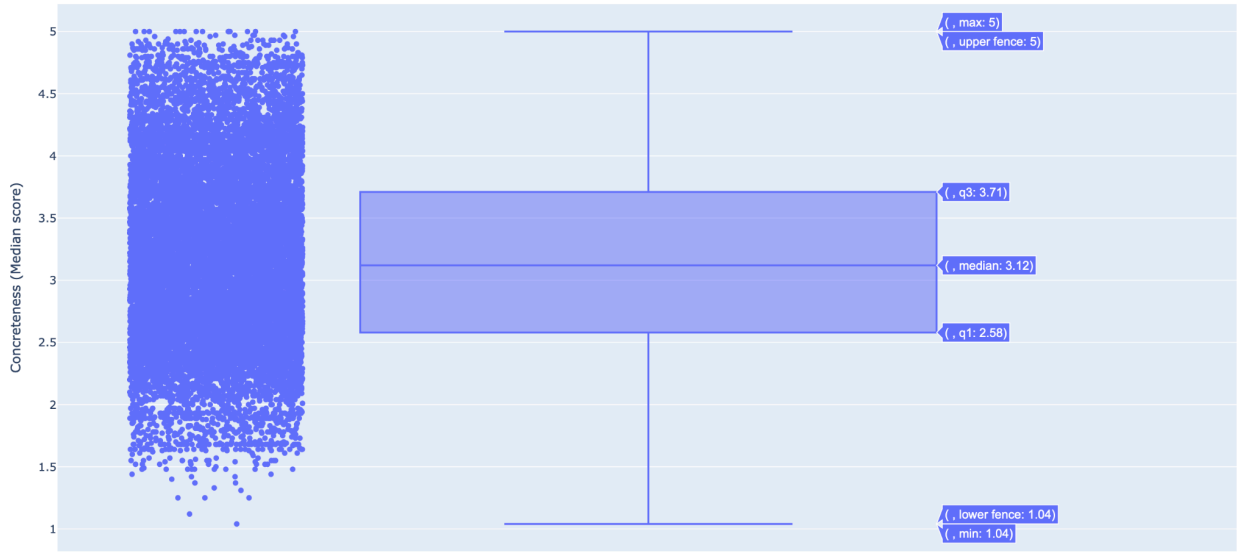

Figure 31. Boxplot showing the distribution of the median concreteness values for each sentence in Harry Potter Book 5, generated using data_visualization_2_main with the file NLP_abstr-concrete-vocab_Dir_Harry Potter Book 5_stats.csv as input.

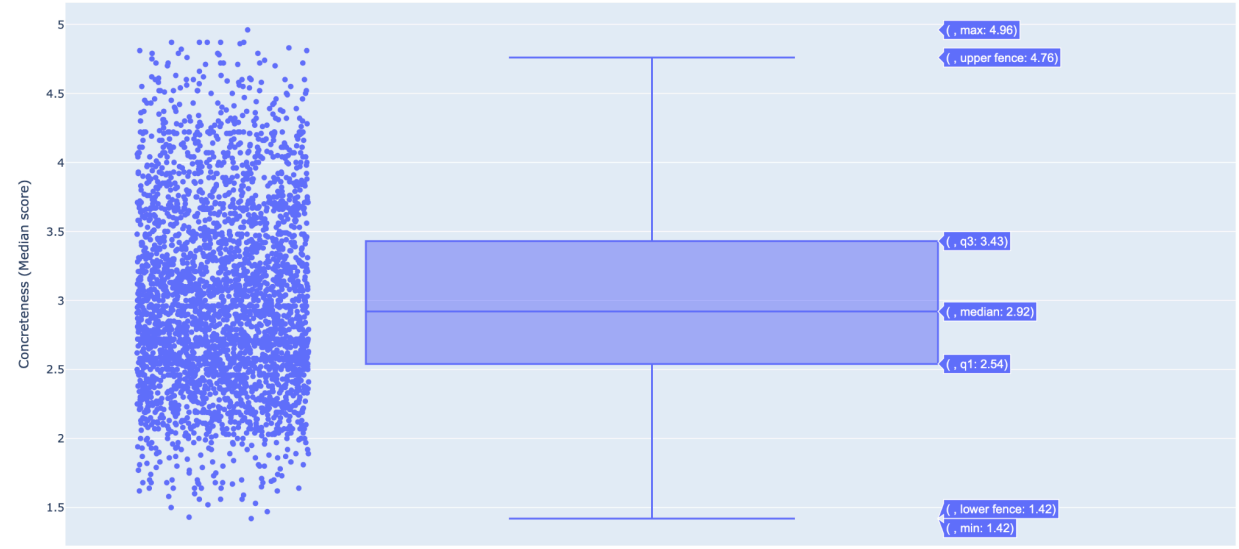

Figure 32. Boxplot showing the distribution of the median concreteness values for each sentence in The Scarlet Letter, generated using data_visualization_2_main with the file NLP_abstr-concret-vocab_Dir_The Scarlet Letter_stats.csv as input.

Thirdly, the use of punctuation, such as question marks and exclamation points, can signal emotional intensity or urgency. Analyzing how these are used could shed light on the emotional tone and stylistic choices made by the authors, potentially reflecting differences in narrative voice and style between the two works. In Figure 33, the relative frequency of question and exclamation marks is calculated by dividing the punctuation frequency by the number of sentences in each corpus, assuming that each sentence contains only one punctuation mark. While this assumption may not be entirely accurate, given the difference in word count between the corpora, this normalization helps to better understand the use of pathos in each text. According to Figure 33, 18.14% of the punctuation marks in *The Scarlet Letter* are question marks, which is higher than in *Harry Potter Book 5*. This suggests that Hawthorne may use more questions to convey a sense of uncertainty or to engage the reader more directly with the narrative. For exclamation marks, the percentages are similar between the two works, indicating that both authors use exclamation points with a comparable frequency to express emotional intensity or emphasis. To sum up, vocabulary does affect style. This is evident with *Harry Potter Book 5* showcasing a more concrete vocabulary and less uncertainty, while *The Scarlet Letter* employs more specialized language and greater emotional intensity through punctuation.

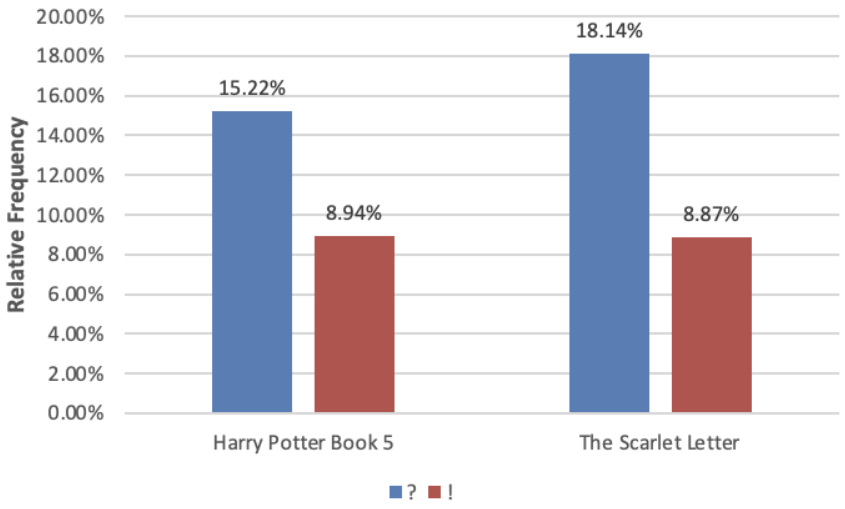

Figure 33. Distribution of relative frequencies for question and exclamation marks in Harry Potter Book 5 and The Scarlet Letter. This figure was created and annotated using the files NLP_punctuation_Dir_Harry Potter Book 5_Bar_chart.xlsx and NLP_punctuation_Dir_The Scarlet Letter_Bar_chart.xlsx.

You can use GenderGuesser to approach issues of style on a corpus. GenderGuesser can predict the gender of an author based on their writing style. This information can then be used to explore whether there are any gender differences in the writing style of a corpus. For example, you could look at the frequency of different words, the length of sentences, or the use of different grammatical structures. If you find that there are significant gender differences in the writing style of your corpus, you can then discuss the implications of these findings. For example, you could discuss whether these differences reflect gender stereotypes or whether they reflect different communicative goals. Here is the output from the browser version of GenderGuesser by Hacker Factor for random sections of our corpus:

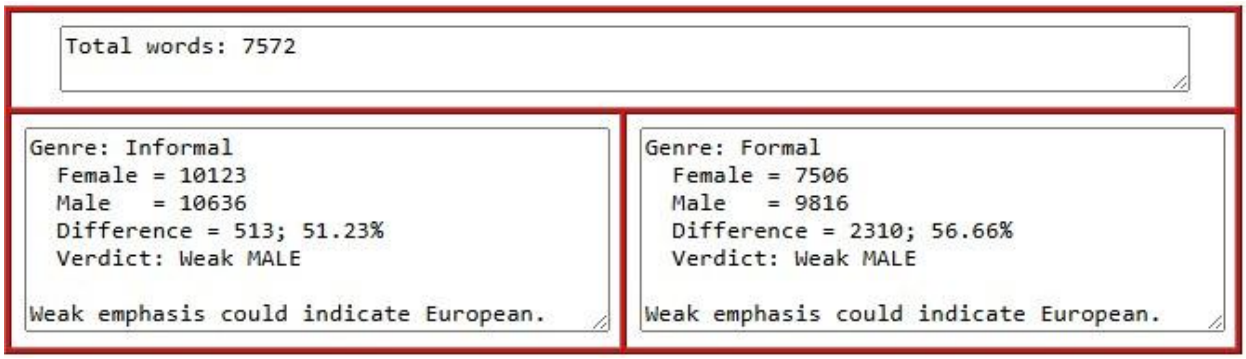

Figure 34. GenderGuessr output for Harry Potter Book 5.

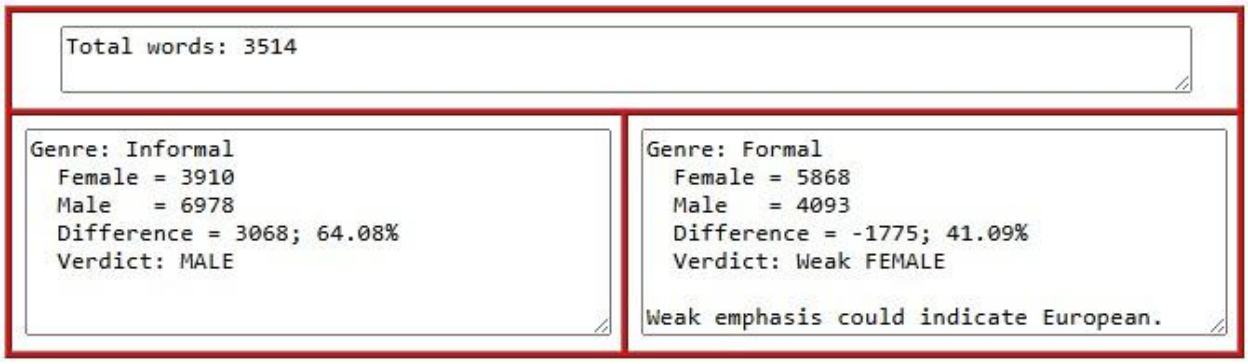

Figure 35. GenderGuessr output for The Scarlet Letter.

There are some important things to keep in mind when reading the output of this algorithm. Gender Guesser uses a neural network to predict the gender of an author based on their writing style. The neural network was trained on a large corpus of text that was labeled with the gender of the author. It learned to identify patterns in the writing style of male and female authors. These patterns then result in a model that can predict the gender of corpora input into the model.

Formal text is typically written for professional or academic purposes. It is characterized by the use of formal language, complex sentence structures, and a lack of personal pronouns while informal text is typically written for personal or social purposes. It is characterized by the use of informal language, simple sentence structures, and the use of personal pronouns. Formal text tends to have fewer markers of gender than informal text, so as a result the more formal style of European writing results in weak emphasis on outputs from the model. The gender estimate "Weak emphasis could indicate European" means that the neural network is not very confident in its prediction of the gender of the author.

The algorithms used by GenderGuesser are not perfect. They have an accuracy of about 60-70%. This means that they will incorrectly predict the gender of an author about 30-40% of

the time. The accuracy of the algorithms is limited by the fact that there is a lot of variation in the writing styles of male and female authors.

Literature Review

The rise of Big Data has ignited a debate on the role of theory in scientific inquiry. Some argue that the massive amount of data now available makes the traditional scientific method obsolete, like in the Anderson paper. However, others argue that theory is still essential for guiding scientific research and for interpreting the results of data analysis, such as in the Mazzocchi paper. This debate has led to an interest in the relative merits of data-driven vs. hypothesis-driven research. Data-driven research is an inductive approach that starts with data and then looks for patterns. Hypothesis-driven research is a deductive approach that starts with a hypothesis and then designs experiments to test the merits of that hypothesis. Although we have for centuries considered the latter approach as the gold standard for generating robust knowledge, the former approach has some advantages over the traditional one.

Quantitative Narrative Analysis (QNA) is a recently developed method for analyzing narrative texts. QNA focuses on identifying agencies within the text. This analysis can then be used to better understand the social and historical context of the text being analyzed. According to the Pennebaker paper, linguistic styles are stable individual differences in language use. These styles can be measured using tools like LIWC, which uses metrics such as the frequency of different words and word categories in texts. Linguistic styles have been shown to be linked to many different personality traits and other individual differences.

Based on Jautze's paper, syntactic structures can be used to characterize different genres of prose. For example, chick lit tends to use simpler sentence structures than literary novels. This difference in syntactic structures may be related to the different purposes of these two genres.

Automated text analysis has several limitations, including difficulty in automating methods to capture nuanced aspects of language use like context, irony, and sarcasm. This is why, according to the Franzosi paper, it is very important to combine quantitative and qualitative approaches to text analysis.

Conclusion

This analysis of *Harry Potter Book 5* and *The Scarlet Letter* using various tools in the NLP Suite reveals differences and similarities in their narrative style, thematic focus, and linguistic structures. The corpus statistics highlight that *Harry Potter Book 5* has a higher word count and shorter sentences, making it more accessible to a younger audience. Conversely, *The Scarlet Letter* features longer, and more complex sentences, which is confirmed by the sentence complexity analysis later. The readability analysis also indicates that *Harry Potter Book 5* is suitable for middle school readers, whereas *The Scarlet Letter* is more appropriate for high school or college-level readers. The N-grams analysis underscores the character-driven nature of both texts, with frequent use of character names and dialogue-related verbs. The co-occurrence analysis further supports this. It shows that even though key characters in both texts often appear in the same chapters, they are less frequently mentioned together in the same sentences, implying their roles within the broader narrative. However, *Harry Potter Book 5* employs a more contemporary and informal language, while *The Scarlet Letter* uses archaic language and formal phrasing. Style analysis reveals that *Harry Potter Book 5* has a richer and more concrete vocabulary, likely due to its fantasy genre. Meanwhile, *The Scarlet Letter* employs more specialized and abstract language toward conceptional symbolism and emotional intensity, which is seen in its frequent use of question marks. Nominalization is more prevalent in *Harry Potter Book 5* and shows a focus on emotions and conflicts, while *The Scarlet Letter* emphasizes

communication and societal roles. Finally, GenderGuesser predicts a male/weak female emphasis for *The Scarlet Letter* and a weak male emphasis for *Harry Potter Book 5*, which could be attributed to their formal genre and European-influenced writing style.

Work Cited

Franzosi, Roberto. NLP TIPS files.

Franzosi, Roberto, et al. “Ways of Measuring Agency.” Sociological Methodology, vol. 42, no. 1, Aug. 2012, pp. 1–42, https://doi.org/10.1177/0081175012462370.

Jautze, Kim, et al. “From High Heels to Weed Attics: A Syntactic Investigation of Chick Lit and Literature.” *ACL Anthology*, June 2013, pp. 72–81, aclanthology.org/W13-1410/.

Kumar, Prachi. “An Introduction to N-Grams: What Are They and Why Do We Need Them?” *XRDS*, 21 Oct. 2017, https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/.

Mazzocchi, Fulvio. “Could Big Data Be the End of Theory in Science?” *EMBO Reports*, vol. 16, no. 10, 10 Sept. 2015, pp. 1250–1255, https://doi.org/10.15252/embr.201541001.

Pennebaker, James W., and Laura A. King. “Linguistic styles: language use as an individual difference.” *Journal of Personality and Social Psychology*, vol, 77, no. 6, 1999, pp. 1296-1312.

SchoolTube Community. “Modality and Nominalization: Mastering Persuasive Language | SchoolTube.” *Schooltube.com*, 18 June 2024, www.schooltube.com/modality-and-nominalization-mastering-persuasive-language/.

“SparkNotes: The Scarlet Letter: Style.” *www.sparknotes.com*, www.sparknotes.com/lit/scarlet/style/.

TEDx Talks. “A Picture Is Worth 500 Billion Words: Erez Lieberman Aiden and Jean-Baptiste Michel at TEDxBoston.” *YouTube*, 14 July 2011, www.youtube.com/watch?v=WtJ50v7qByE.

“The Writing Style of J. K. Rowling.” *I Write Like*, 2024, iwl.me/writer/J._K._Rowling.