Carol Zhou

Dr. Roberto Franzosi

SOC 446W Big/Small Data & Visualization

9 March 2025

SOC 446W Homework 5

Introduction

In the realm of natural language processing (NLP), tools such as Stanford CoreNLP, spaCy, and Stanza are essential for analyzing textual corpora. They offer insights into syntactic structures, named entities, and linguistic patterns, and enable researchers to uncover the underlying grammatical and semantic features of texts. This paper employs these tools to examine two literary works — *Harry Potter and the Order of the Phoenix* and *The Scarlet Letter*. We will evaluate the parsers' and annotators' performance and explore specialized annotators like gender detection, normalized NER dates, and quote/dialogue. Additionally, this study delves into using the CoNLL table analyzer to explore the frequency distribution of "junk words," verb modalities, and noun-verb relationships. Through a combination of quantitative analysis and visualizations, this paper aims to reveal emerging themes and consolidate findings from previous weeks, supported by a review of relevant literature and visual evidence.

Methods

The assigned corpora for analysis are *Harry Potter and the Order of the Phoenix* (2003) by J.K. Rowling and *The Scarlet Letter* (1850) by Nathaniel Hawthorne. These texts belong to very different genres. *Harry Potter Book 5* is a magical fantasy, while *The Scarlet Letter* is set in the Puritan Massachusetts Bay Colony during the 17th century. Previous analyses show that *Harry Potter Book 5* has more words but shorter sentences on average, whereas *The Scarlet*

Letter has longer, more complex sentences. In terms of readability and complexity, *The Scarlet Letter* consistently demonstrates a higher level of complexity. Vocabulary analysis reveals that *The Scarlet Letter* uses more abstract and rarer words, while *Harry Potter Book 5* has a broader vocabulary likely due to its fantasy elements and its invention of new terms such as magic spells. Wordcloud, N-grams, and Co-Occurrences VIEWER analyses indicate that both texts are character- and dialogue-driven with frequent use of character names and dialogue-related verbs such as "said". Additionally, topic modeling and word embeddings reveal that *Harry Potter Book 5* focuses on themes of internal struggles, friendships, and battles, while *The Scarlet Letter* explores themes of sin, redemption, and societal judgment.

Parsers

In the context of NLP, parsers are tools used to analyze and determine the syntactic structure of sentences in a text. They help break down a sentence into its components such as words, phrases, and clauses, and identify the grammatical relationships between those components (Franzosi). Parsers help understand how words relate to each other by identifying the grammatical structure of sentences For example, in the sentence "Tom ate an apple", a parser can figure out that "Tom" is the subject, "ate" is the verb, and "an apple" is the object. Moreover, some parsers such as Stanford CoreNLP's neural-network dependency parser focus on dependency parsing. This type of parsing analyzes how words are connected, with each word in the sentence depending on a "head" word. For instance, in the sentence "Tom ate an apple", "ate" would be the head, and "Tom" and "apple" would depend on it. Parsers can also help with tasks like named entity recognition (NER) and part-of-speech tagging (POS), which are crucial for extracting information such as dates, locations, or people's names.

According to Manning et al., Stanford CoreNLP is a widely used, extensible pipeline for NLP that provides core linguistic analysis steps from tokenization to coreference resolution. The toolkit is successful for its simple design, straightforward interfaces, robust analysis components, and minimal associated baggage, making it accessible to both research and commercial users (Manning et al. 1). In the paper, the authors detail the system's architecture, usage patterns, provided annotators, and the ease of adding custom annotators, emphasizing its lightweight, Java-based framework. Stanford CoreNLP's design avoids over-engineering, focusing on ease of use and integration into larger systems, which has contributed to its broad adoption (Manning et al. 5). The paper concludes by highlighting the toolkit's accessibility, quality of components, and the importance of user-friendly design in the success of NLP tools.

For this assignment, we ran Stanford CoreNLP with the probabilistic context grammar (PCFG) parser and spaCy and Stanza with the dependency parser on our corpora to compare their performance. This is based on accuracy, speed, and multilingual support, which are essential to a good parser. The output file from the parser is the CoNLL table. This is the basis for various statistical analyses and visualizations on the text corpus later. The table contains various fields. FORM is the word form or punctuation symbol found in the input document, while LEMMA is the lemma or stem of word form. POSTAG is the language-specific part-of-speech tag. If the parser cannot determine the part of speech, an underscore is used. DEPREL refers to the universal Stanford dependency relation to the HEAD (root if and only if HEAD = 0) or a defined language-specific subtype of one. If the value of HEAD is 0, the DEPREL value is set to ROOT, with only one ROOT for each sentence. Named Entity Recognition (NER) is an NLP technique used to identify and classify entities in a text into predefined categories such as the names of persons, organizations, locations, dates, times,

quantities, etc. By analyzing a corpus using NER, we can extract and categorize specific, meaningful entities, and gain insights into the key elements present in the text (Franzosi).

Annotators

Annotating a text traditionally involves highlighting or underlining key pieces of text and making notes in the margins to emphasize the most important points, aiding in quick re-reading and recall. In NLP, annotation is done automatically by an annotator rather than by hand with a highlighter or a pencil. Stanford CoreNLP offers several specialized annotators: gender, normalized NER date, and quote/dialogue. The gender annotator identifies the gender of proper names in the text using two approaches within the NLP Suite. One method utilizes the name list file from Stanford CoreNLP, while the other relies on dictionaries stored in the lib\namesGender folder under the NLP Suite directory. These dictionaries, including those from the US Census, Carnegie Mellon, NLTK, and the US Social Security Administration, contain first names (e.g., John) and their corresponding genders (e.g., male). Some of these files also provide probabilistic measures for gender attribution, indicating the likelihood that a name (e.g., John or Jamie) is associated with a particular gender (Franzosi). For faster analysis, we only ran these on *The Scarlet Letter* since this corpus is shorter. With the first three dictionaries, the annotator only generated HTML files highlighting names in red for females and blue for males. When using the US Social Security Administration dictionary, the “by State” option was selected, and the name “Hester” was searched. However, this approach does not offer much insight into the corpus itself, as it simply shows the frequency of the name “Hester” in different US states.

The normalized NER date annotator extracts temporal expressions from the text, standardizing them into “normalized dates” that can be plotted to show their frequency over time. This data can reveal a text’s narrative structure and temporal strategy (Franzosi). The

quote/dialogue annotator extracts direct quotes and dialogue from the text and identifies the speakers. In this paper, we ran the above annotators on *Harry Potter Book 5* and *The Scarlet Letter* separately to compare gender emphasis, temporal references, and dialogue frequencies between the two corpora.

CoNLL Table Analyzer

We can use the CoNLL table analyzer to perform clause, noun, and verb analyses and provide more insights into our corpora. The clausal analysis involves dividing sentences into their constituent clauses. For example, in the sentence "When the bell rang, all the children ran out of the classroom," the main clause is "all the children ran out of the classroom," while the subordinate adverbial clause of time is "When the bell rang." Clausal measures and plots in the NLP Suite help reveal an author's style. Nouns can serve various functions, such as subject or object, active or passive, direct or indirect objects, singular or plural, proper or improper. Analyzing nouns provides important clues about language use and sentence structure.

Verbs have four different and important aspects: tense, mood, voice, and modality. Tense refers to the time of action (past, present, and future), although not all actions fit precisely into those three categories. For example, a gerundive verb may describe an action that started in the past and continues into the present or future. Verb voice indicates whether the subject of a sentence is the actor (agent) or the recipient of the action (patient). In active voice, agency (who does something) is always clear, but in passive voice, agency can be denied. This occurs when the agent is either unmarked or omitted, making the syntactic subject and semantic role of the agent separate. (Franzosi et al. 5). Modality refers to verbs that express ability, possibility, permission, or obligation. Modal verbs include can/could, may/might, must/need, will/would, and shall/should/ought. Verbs also have several moods. The indicative mood is used for factual

statements and questions. The subjunctive mood is used for statements contrary to fact, such as "If I were twenty, I would definitely go to college" (implying the speaker is not twenty). The imperative mood expresses a request or command. The conditional mood is used for statements that are dependent on another, where one statement is true only if the other is true.

Stopwords are words that are filtered out during natural language processing, typically including pronouns (e.g., I, you, we, mine), prepositions (e.g., after, in, to, on, with), articles (e.g., a, the), conjunctions (e.g., and, or), and auxiliary verbs (e.g., can, would). There is no universal list of stopwords, as the selection can vary based on the specific purpose of the analysis. Sometimes, these words are also called "junk" or function words, in constructing words that provide the bulk of meaning in a text. In addition to analyzing clauses, nouns, and verbs, we also examined stopwords to determine if our corpora align with typical frequency distributions of stopwords. Pronouns and auxiliary verbs are key markers of gender-based language style, with males using more nouns and determiners, while females use more pronouns, negations, prepositions, and conjunctions, including "I" (Franzosi). These analyses allow us to gain a deeper understanding of the language and style in our texts.

Results and Discussions

Parsers

We evaluate the performance of different parsers based on accuracy, speed, and multilingual support. In terms of accuracy, Stanford CoreNLP performs exceptionally well on well-formed texts, particularly for POS tagging (Jørgensen et al.). SpaCy and Stanza, while still accurate, are not quite as precise as Stanford CoreNLP. However, Stanford CoreNLP is the slowest out of the three since its high accuracy comes with an efficiency cost. SpaCy is the fastest, which makes it ideal for applications that prioritize real-time performance or need to

process large volumes of text quickly. Stanza is slower than spaCy but faster than Stanford CoreNLP. When we ran the parsers on *The Scarlet Letter* (24 chapters and 86,878 words) on a Macbook with an M2 chip, Stanford CoreNLP took about 9 minutes and cost high memory, while SpaCy and Stanza took around 5 minutes. However, when processing *Harry Potter Book 5* (38 chapters and 257,045 words), the runtime significantly increased. It took 30 minutes for Stanford CoreNLP, 27 minutes for SpaCy, and 14 minutes for Stanza. Lastly, regarding the number of languages that each parser supports, Stanford CoreNLP is only limited to 7 languages. However, both SpaCy and Stanza support over 60 languages. To sum up, each of these NLP packages has its strengths and trade-offs, and the best choice depends on the specific needs of your NLP tasks, whether you prioritize speed, accuracy, or multilingual support.

To visualize the distribution of NER categories in the two corpora, the percent frequency is computed for the remaining categories. Undefined NER values have been excluded due to their large quantity, which would hinder effective visualization. Figures 1 and 2 show that the most frequent entity type is PERSON for both corpora. This aligns with previous findings and indicates that both texts are character-driven. Notably, the percent frequency of PERSON in *Harry Potter Book 5* is significantly higher than in *The Scarlet Letter*, likely due to the larger number of characters in the former. Moreover, both books have a significant presence of DATE and TIME entities, a potential previously unrecognized theme, suggesting the importance of time in both narratives. In *Harry Potter Book 5*, the high frequency of ORGANIZATION might reflect the significance of institutions like the Order of the Phoenix, Dumbledore's Army, and the Ministry of Magic. In contrast, *The Scarlet Letter* demonstrates a prominence of TITLE, such as Reverend, Governor, and Mistress, likely referencing positions of authority and echoing the book's central themes.

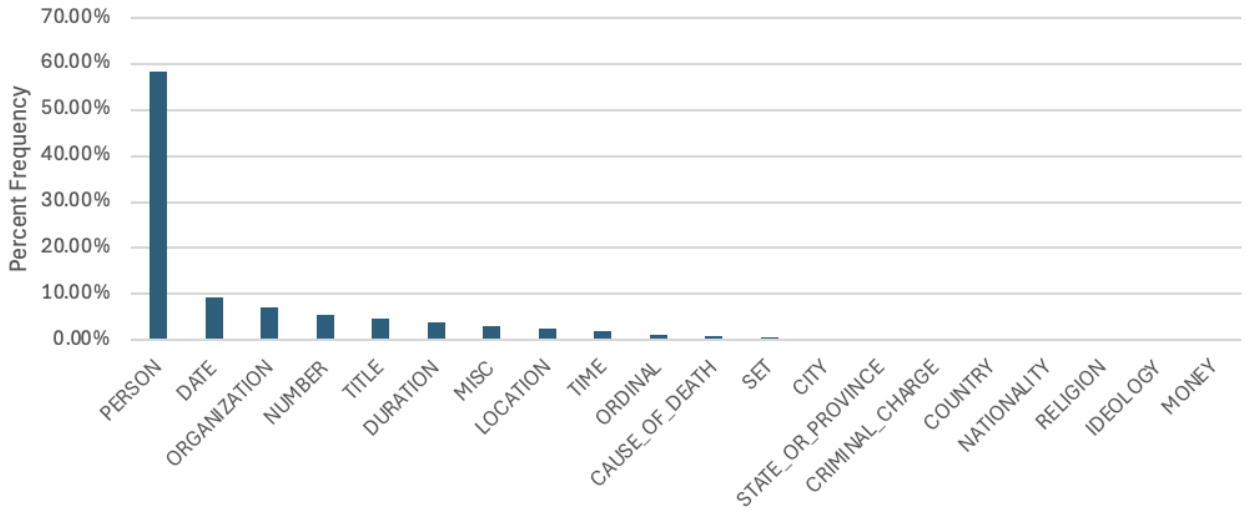


Figure 1. Percent frequency of NER categories (excluding O/undefined values) in the CoNLL Table for Harry Potter Book 5.

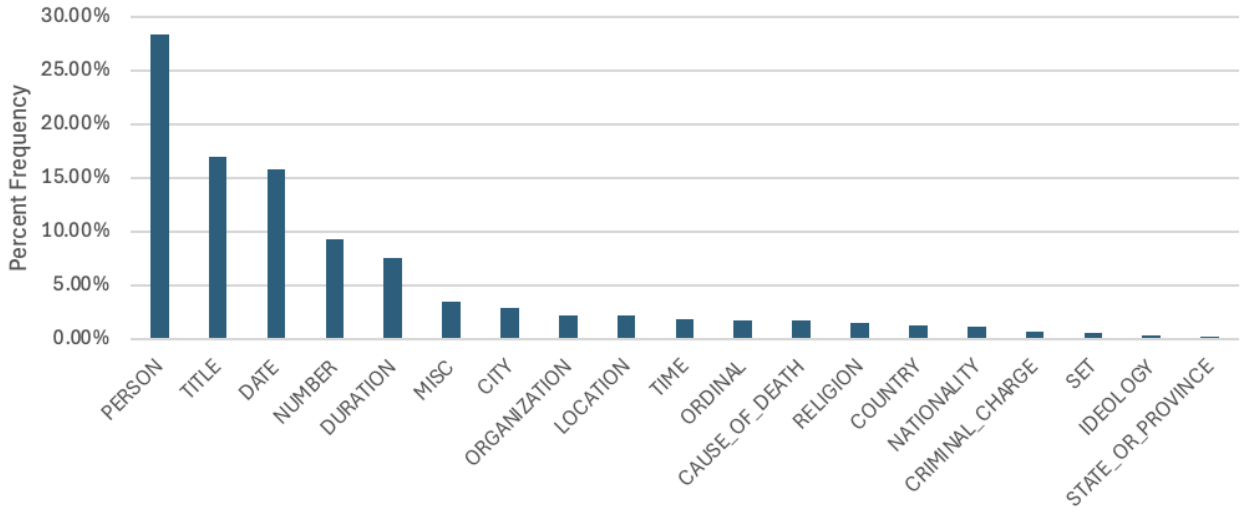


Figure 2. Percent frequency of NER categories (excluding O/undefined values) in the CoNLL Table for The Scarlet Letter.

Gender Annotator

We first ran the gender annotator with Stanford CoreNLP's dictionary. Figure 3 displays the percent frequency of gender values for identified proper names in *Harry Potter Book 5* and *The Scarlet Letter.* The figure suggests a clear dominance of male proper names in both corpora,

indicating male characters are more frequently mentioned or play a more prominent role in both narratives. This is expected for *Harry Potter Book 5* since the story centers around Harry and only a few female characters, such as Hermoine, are mentioned frequently (Figure 4). However, it is interesting that more male characters are mentioned in *The Scarlet Letter*, despite Hester being the main character and one of the book's themes being motherhood. Upon further examination, this may be explained by the fact that, besides Hester, Pearl, and Mistress Hibbins, most of the characters in the book are male, such as Roger Chillingworth, Reverend Arthur Dimmesdale, Governor Bellingham, and Reverend John Wilson. Another possibility is that the annotator mistakenly identified Hester as a male character, as shown in Figure 5. However, this seems to account for only a small percentage, which likely did not significantly impact the overall results. Furthermore, Harry Potter Book 5 demonstrates a higher percentage of male proper names (78.69%) compared to The Scarlet Letter (60.43%), but a lower percentage of female proper names (21.31% versus 39.57%). This aligns with the fact that *Harry Potter Book 5* features more male characters than *The Scarlet Letter*, while the latter contains a higher proportion of female characters.

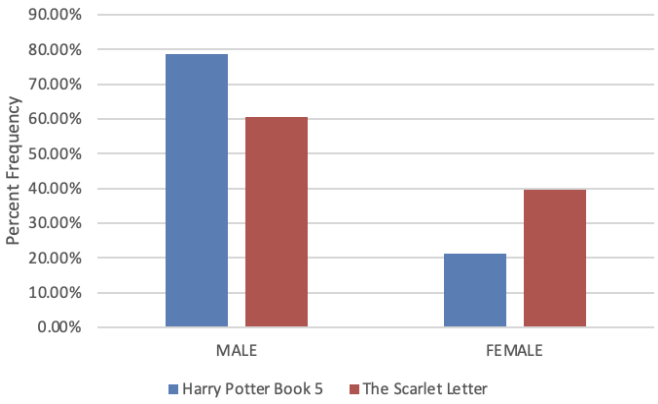


Figure 3. Percent frequency of gender values for identified proper names in Harry Potter Book 5 and The Scarlet Letter using Stanford CoreNLP's dictionary.

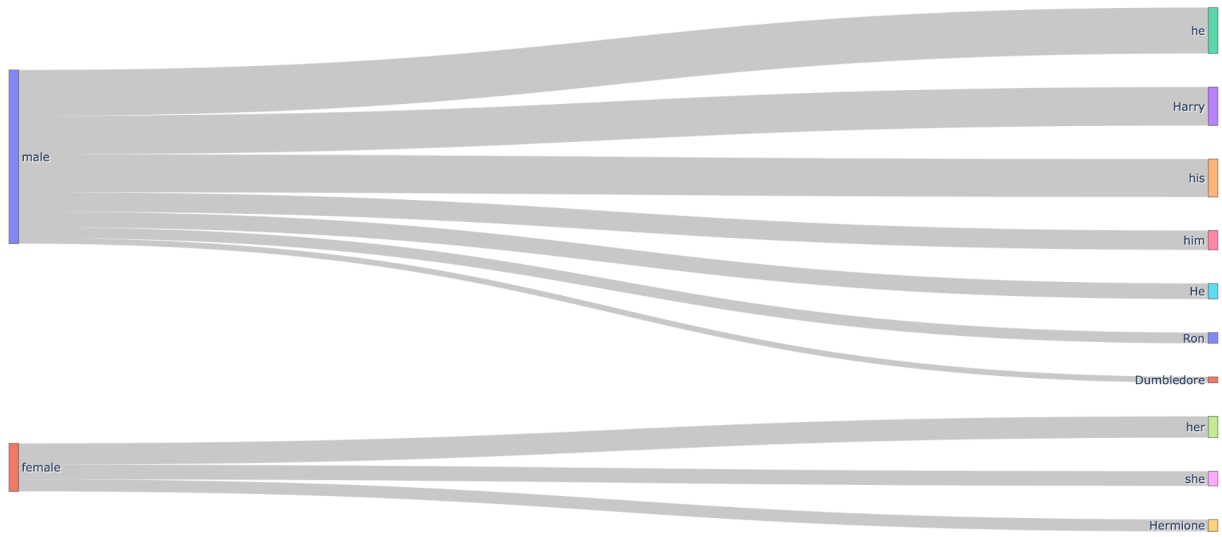


Figure 4. Sankey chart showing some proper names classified as "MALE" or "FEFMALE" in Harry Potter Book 5 using Stanford CoreNLP's dictionary.

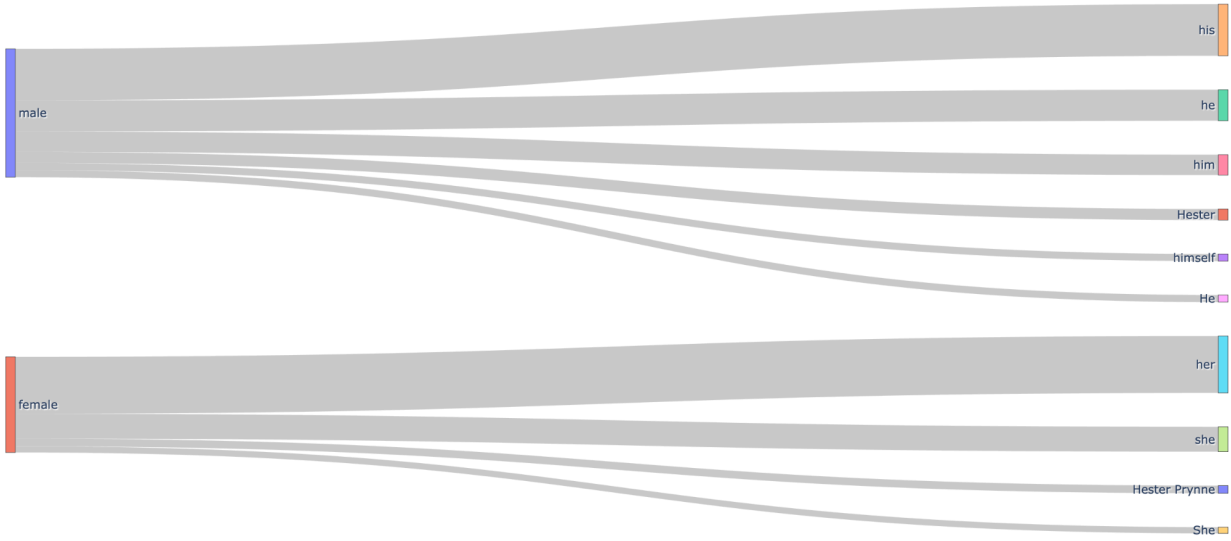
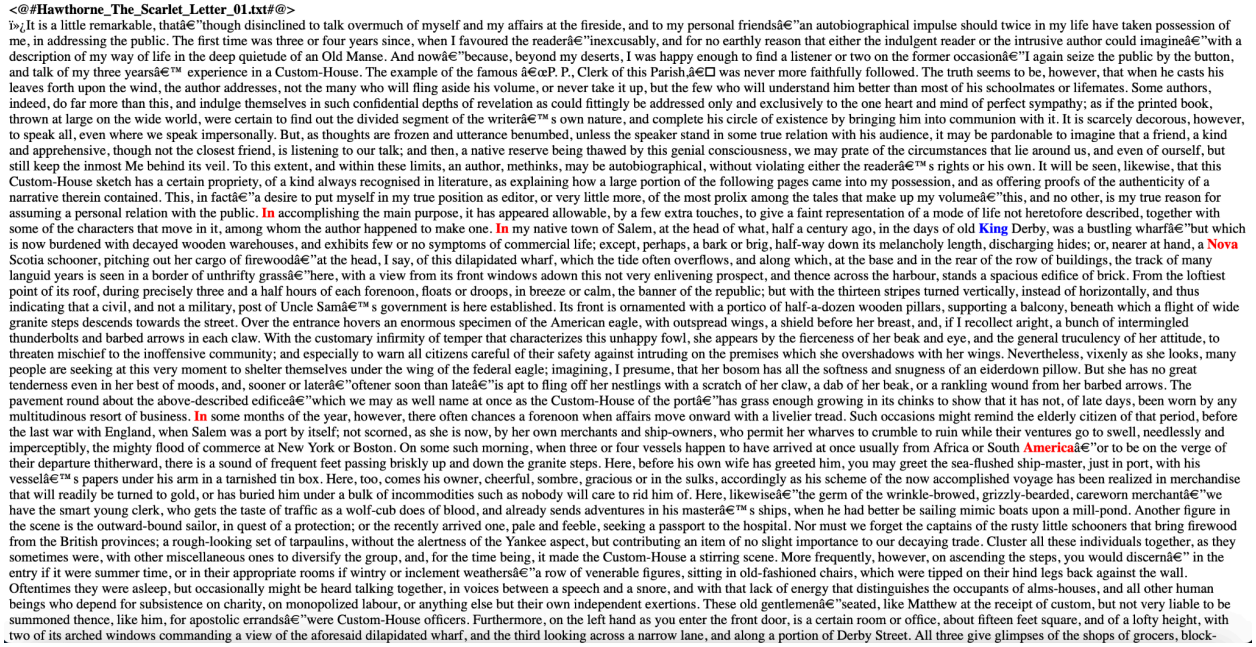


Figure 5. Sankey chart showing some proper names classified as "MALE" or "FEMALE" in The Scarlet Letter using Stanford CoreNLP's dictionary.

We also ran the gender annotator on *The Scarlet Letter* using the US census, Carnegie Mellon, NLTK, and US Social Security dictionaries to see if various dictionaries give us different results. As shown in Figures 6, 7, and 8, it appears that these annotators assign gender to any

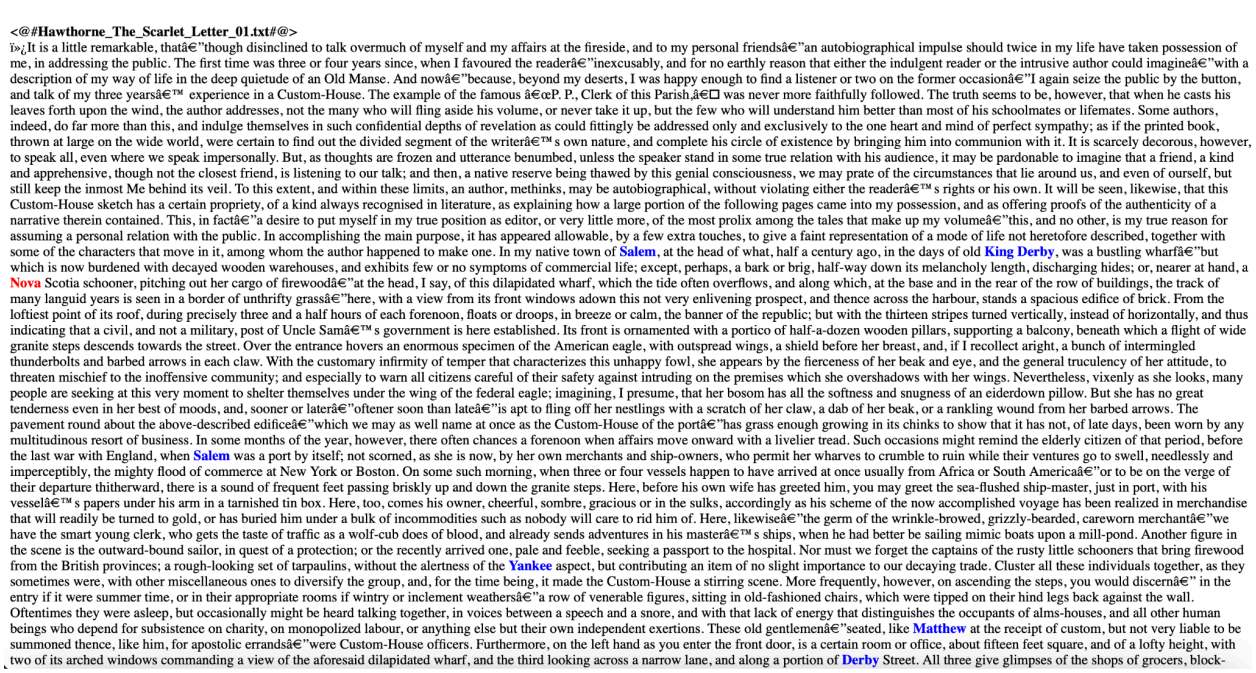
noun without first verifying whether the noun refers to a specific person. Some of the words highlighted are place names. Moreover, though the results from the Carnegie Mellon and NLTK dictionaries are consistent, the results from the US census dictionary differ significantly. For example, the US Census dictionary mistakenly identified “In” and “America” as female, likely because these words appear in the CSV file as names associated with the female gender. However, in this case, they are not proper names. On the other hand, based on Figure 9, the name “Hester” is more commonly associated with females than males across different US states. This reinforces the problem of Stanford CoreNLP identifying it as male in Figure 5. In conclusion, overall, using dictionaries other than Stanford CoreNLP’s results in inaccurate gender assignments.



<@#Hawthorne_The_Scarlet_Letter_01.txt#@>

ï»¿It is a little remarkable, thatâ€”though disinclined to talk overmuch of myself and my affairs at the fireside, and to my personal friendsâ€”an autobiographical impulse should twice in my life have taken possession of me, in addressing the public. The first time was three or four years since, when I favoured the readerâ€”inexcusably, and for no earthly reason that either the indulgent reader or the intrusive author could imagineâ€”with a description of my way of life in the deep quietude of an Old Manse. And nowâ€”because, beyond my deserts, I was happy enough to find a listener or two on the former occasionâ€”I again seize the public by the button, and talk of my three yearsâ€™ experience in a Custom-House. The example of the famous â€œP. P., Clerk of this Parish,â€ was never more faithfully followed. The truth seems to be, however, that when he casts his leaves forth upon the wind, the author addresses, not the many who will fling aside his volume, or never take it up, but the few who will understand him better than most of his schoolmates or lifemates. Some authors, indeed, do far more than this, and indulge themselves in such confidential depths of revelation as could fittingly be addressed only and exclusively to the one heart and mind of perfect sympathy; as if the printed book, thrown at large on the wide world, were certain to find out the divided segment of the writerâ€™s own nature, and complete his circle of existence by bringing him into communion with it. It is scarcely decorous, however, to speak all, even where we speak impersonally. But, as thoughts are frozen and utterance benumbed, unless the speaker stand in some true relation with his audience, it may be pardonable to imagine that a friend, a kind and apprehensive, though not the closest friend, is listening to our talk; and then, a native reserve being thawed by this genial consciousness, we may prate of the circumstances that lie around us, and even of ourself, but still keep the inmost Me behind its veil. To this extent, and within these limits, an author, methinks, may be autobiographical, without violating either the readerâ€™s rights or his own. It will be seen, likewise, that this Custom-House sketch has a certain propriety, of a kind always recognised in literature, as explaining how a large portion of the following pages came into my possession, and as offering proofs of the authenticity of a narrative therein contained. This, in factâ€”a desire to put myself in my true position as editor, or very little more, of the most prolix among the tales that make up my volumeâ€”this, and no other, is my true reason for assuming a personal relation with the public. **In** accomplishing the main purpose, it has appeared allowable, by a few extra touches, to give a faint representation of a mode of life not heretofore described, together with some of the characters that move in it, among whom the author happened to make one. **In** my native town of Salem, at the head of what, half a century ago, in the days of old **King** Derby, was a bustling wharfâ€”but which is now burdened with decayed wooden warehouses, and exhibits few or no symptoms of commercial life; except, perhaps, a bark or brig, half-way down its melancholy length, discharging hides; or, nearer at hand, a **Nova** Scotia schooner, pitching out her cargo of firewoodâ€”at the head, I say, of this dilapidated wharf, which the tide often overflows, and along which, at the base and in the rear of the row of buildings, the track of many languid years is seen in a border of unthrifty grassâ€”here, with a view from its front windows adown this not very enlivening prospect, and thence across the harbour, stands a spacious edifice of brick. From the loftiest point of its roof, during precisely three and a half hours of each forenoon, floats or droops, in breeze or calm, the banner of the republic; but with the thirteen stripes turned vertically, instead of horizontally, and thus indicating that a civil, and not a military, post of Uncle Samâ€™s government is here established. Its front is ornamented with a portico of half-a-dozen wooden pillars, supporting a balcony, beneath which a flight of wide granite steps descends towards the street. Over the entrance hovers an enormous specimen of the American eagle, with outspread wings, a shield before her breast, and, if I recollect aright, a bunch of intermingled thunderbolts and barbed arrows in each claw. With the customary infirmity of temper that characterizes this unhappy fowl, she appears by the fierceness of her beak and eye, and the general truculency of her attitude, to threaten mischief to the inoffensive community; and especially to warn all citizens careful of their safety against intruding on the premises which she overshadows with her wings. Nevertheless, vixenly as she looks, many people are seeking at this very moment to shelter themselves under the wing of the federal eagle; imagining, I presume, that her bosom has all the softness and snugness of an eiderdown pillow. But she has no great tenderness even in her best of moods, and, sooner or laterâ€”oftener soon than lateâ€”is apt to fling off her nestlings with a scratch of her claw, a dab of her beak, or a rankling wound from her barbed arrows. The pavement round about the above-described edificeâ€”which we may as well name at once as the Custom-House of the portâ€”has grass enough growing in its chinks to show that it has not, of late days, been worn by any multitudinous resort of business. **In** some months of the year, however, there often chances a forenoon when affairs move onward with a livelier tread. Such occasions might remind the elderly citizen of that period, before the last war with England, when Salem was a port by itself; not scorned, as she is now, by her own merchants and ship-owners, who permit her wharves to crumble to ruin while their ventures go to swell, needlessly and imperceptibly, the mighty flood of commerce at New York or Boston. On some such morning, when three or four vessels happen to have arrived at once usually from Africa or South **America**â€”or to be on the verge of their departure thitherward, there is a sound of frequent feet passing briskly up and down the granite steps. Here, before his own wife has greeted him, you may greet the sea-flushed ship-master, just in port, with his vesselâ€™s papers under his arm in a tarnished tin box. Here, too, comes his owner, cheerful, sombre, gracious or in the sulks, accordingly as his scheme of the now accomplished voyage has been realized in merchandise that will readily be turned to gold, or has buried him under a bulk of incommodities such as nobody will care to rid him of. Here, likewiseâ€”the germ of the wrinkle-browed, grizzly-bearded, careworn merchantâ€”we have the smart young clerk, who gets the taste of traffic as a wolf-cub does of blood, and already sends adventures in his masterâ€™s ships, when he had better be sailing mimic boats upon a mill-pond. Another figure in the scene is the outward-bound sailor, in quest of a protection; or the recently arrived one, pale and feeble, seeking a passport to the hospital. Nor must we forget the captains of the rusty little schooners that bring firewood from the British provinces; a rough-looking set of tarpaulins, without the alertness of the Yankee aspect, but contributing an item of no slight importance to our decaying trade. Cluster all these individuals together, as they sometimes were, with other miscellaneous ones to diversify the group, and, for the time being, it made the Custom-House a stirring scene. More frequently, however, on ascending the steps, you would discernâ€” in the entry if it were summer time, or in their appropriate rooms if wintry or inclement weathersâ€”a row of venerable figures, sitting in old-fashioned chairs, which were tipped on their hind legs back against the wall. Oftentimes they were asleep, but occasionally might be heard talking together, in voices between a speech and a snore, and with that lack of energy that distinguishes the occupants of alms-houses, and all other human beings who depend for subsistence on charity, on monopolized labour, or anything else but their own independent exertions. These old gentlemenâ€”seated, like Matthew at the receipt of custom, but not very liable to be summoned thence, like him, for apostolic errandsâ€”were Custom-House officers. Furthermore, on the left hand as you enter the front door, is a certain room or office, about fifteen feet square, and of a lofty height, with two of its arched windows commanding a view of the aforesaid dilapidated wharf, and the third looking across a narrow lane, and along a portion of Derby Street. All three give glimpses of the shops of grocers, block-

Figure 6. Screenshot of a partial HTML output from the gender annotator on The Scarlet Letter, using the US census dictionary. Red-highlighted words indicate female, while blue-highlighted words indicate male.



<@#Hawthorne_The_Scarlet_Letter_01.txt#@>
ï»¿It is a little remarkable, thatâ€"though disinclined to talk overmuch of myself and my affairs at the fireside, and to my personal friendsâ€"an autobiographical impulse should twice in my life have taken possession of me, in addressing the public. The first time was three or four years since, when I favoured the readerâ€"inexcusably, and for no earthly reason that either the indulgent reader or the intrusive author could imagineâ€"with a description of my way of life in the deep quietude of an Old Manse. And nowâ€"because, beyond my deserts, I was happy enough to find a listener or two on the former occasionâ€"I again seize the public by the button, and talk of my three yearsâ€™ experience in a Custom-House. The example of the famous â€œP. P., Clerk of this Parish,â€□ was never more faithfully followed. The truth seems to be, however, that when he casts his leaves forth upon the wind, the author addresses, not the many who will fling aside his volume, or never take it up, but the few who will understand him better than most of his schoolmates or lifemates. Some authors, indeed, do far more than this, and indulge themselves in such confidential depths of revelation as could fittingly be addressed only and exclusively to the one heart and mind of perfect sympathy; as if the printed book, thrown at large on the wide world, were certain to find out the divided segment of the writerâ€™s own nature, and complete his circle of existence by bringing him into communion with it. It is scarcely decorous, however, to speak all, even where we speak impersonally. But, as thoughts are frozen and utterance benumbed, unless the speaker stand in some true relation with his audience, it may be pardonable to imagine that a friend, a kind and apprehensive, though not the closest friend, is listening to our talk; and then, a native reserve being thawed by this genial consciousness, we may prate of the circumstances that lie around us, and even of ourself, but still keep the inmost Me behind its veil. To this extent, and within these limits, an author, methinks, may be autobiographical, without violating either the readerâ€™s rights or his own. It will be seen, likewise, that this Custom-House sketch has a certain propriety, of a kind always recognised in literature, as explaining how a large portion of the following pages came into my possession, and as offering proofs of the authenticity of a narrative therein contained. This, in factâ€"a desire to put myself in my true position as editor, or very little more, of the most prolix among the tales that make up my volumeâ€"this, and no other, is my true reason for assuming a personal relation with the public. In accomplishing the main purpose, it has appeared allowable, by a few extra touches, to give a faint representation of a mode of life not heretofore described, together with some of the characters that move in it, among whom the author happened to make one. In my native town of **Salem**, at the head of what, half a century ago, in the days of old **King Derby**, was a bustling wharfâ€"but which is now burdened with decayed wooden warehouses, and exhibits few or no symptoms of commercial life; except, perhaps, a bark or brig, half-way down its melancholy length, discharging hides; or, nearer at hand, a **Nova** Scotia schooner, pitching out her cargo of firewoodâ€"at the head, I say, of this dilapidated wharf, which the tide often overflows, and along which, at the base and in the rear of the row of buildings, the track of many languid years is seen in a border of unthrifty grassâ€"here, with a view from its front windows adown this not very enlivening prospect, and thence across the harbour, stands a spacious edifice of brick. From the loftiest point of its roof, during precisely three and a half hours of each forenoon, floats or droops, in breeze or calm, the banner of the republic; but with the thirteen stripes turned vertically, instead of horizontally, and thus indicating that a civil, and not a military, post of Uncle Samâ€™s government is here established. Its front is ornamented with a portico of half-a-dozen wooden pillars, supporting a balcony, beneath which a flight of wide granite steps descends towards the street. Over the entrance hovers an enormous specimen of the American eagle, with outspread wings, a shield before her breast, and, if I recollect aright, a bunch of interminglded thunderbolts and barbed arrows in each claw. With the customary infirmity of temper that characterizes this unhappy fowl, she appears by the fierceness of her beak and eye, and the general truculency of her attitude, to threaten mischief to the inoffensive community; and especially to warn all citizens careful of their safety against intruding on the premises which she overshadows with her wings. Nevertheless, vixenly as she looks, many people are seeking at this very moment to shelter themselves under the wing of the federal eagle; imagining, I presume, that her bosom has all the softness and snugness of an eiderdown pillow. But she has no great tenderness even in her best of moods, and, sooner or laterâ€"oftener soon than lateâ€"is apt to fling off her nestlings with a scratch of her claw, a dab of her beak, or a rankling wound from her barbed arrows. The pavement round about the above-described edificeâ€"which we may as well name at once as the Custom-House of the portâ€"has grass enough growing in its chinks to show that it has not, of late days, been worn by any multitudinous resort of business. In some months of the year, however, there often chances a forenoon when affairs move onward with a livelier tread. Such occasions might remind the elderly citizen of that period, before the last war with England, when **Salem** was a port by itself; not scorned, as she is now, by her own merchants and ship-owners, who permit her wharves to crumble to ruin while their ventures go to swell, needlessly and imperceptibly, the mighty flood of commerce at New York or Boston. On some such morning, when three or four vessels happen to have arrived at once usually from Africa or South Americaâ€"or to be on the verge of their departure thitherward, there is a sound of frequent feet passing briskly up and down the granite steps. Here, before his own wife has greeted him, you may greet the sea-flushed ship-master, just in port, with his vesselâ€™s papers under his arm in a tarnished tin box. Here, too, comes his owner, cheerful, sombre, gracious or in the sulks, accordingly as his scheme of the now accomplished voyage has been realized in merchandise that will readily be turned to gold, or has buried him under a bulk of incommodities such as nobody will care to rid him of. Here, likewiseâ€"the germ of the wrinkle-browed, grizzly-bearded, careworn merchantâ€"we have the smart young clerk, who gets the taste of traffic as a wolf-cub does of blood, and already sends adventures in his masterâ€™s ships, when he had better be sailing mimic boats upon a mill-pond. Another figure in the scene is the outward-bound sailor, in quest of a protection; or the recently arrived one, pale and feeble, seeking a passport to the hospital. Nor must we forget the captains of the rusty little schooners that bring firewood from the British provinces; a rough-looking set of tarpaulins, without the alertness of the **Yankee** aspect, but contributing an item of no slight importance to our decaying trade. Cluster all these individuals together, as they sometimes were, with other miscellaneous ones to diversify the group, and, for the time being, it made the Custom-House a stirring scene. More frequently, however, on ascending the steps, you would discernâ€" in the entry if it were summer time, or in their appropriate rooms if wintry or inclement weathersâ€"a row of venerable figures, sitting in old-fashioned chairs, which were tipped on their hind legs back against the wall. Oftentimes they were asleep, but occasionally might be heard talking together, in voices between a speech and a snore, and with that lack of energy that distinguishes the occupants of alms-houses, and all other human beings who depend for subsistence on charity, on monopolized labour, or anything else but their own independent exertions. These old gentlemenâ€"seated, like **Matthew** at the receipt of custom, but not very liable to be summoned thence, like him, for apostolic errandsâ€"were Custom-House officers. Furthermore, on the left hand as you enter the front door, is a certain room or office, about fifteen feet square, and of a lofty height, with two of its arched windows commanding a view of the aforesaid dilapidated wharf, and the third looking across a narrow lane, and along a portion of **Derby** Street. All three give glimpses of the shops of grocers, block-

Figure 7. Screenshot of a partial HTML output from the gender annotator on The Scarlet Letter, using the Carnegie Mellon dictionary. Red-highlighted words indicate female, while blue-highlighted words indicate male.

<@#Hawthorne_The_Scarlet_Letter_01.txt#@>
ï»¿It is a little remarkable, thatâ€"though disinclined to talk overmuch of myself and my affairs at the fireside, and to my personal friendsâ€"an autobiographical impulse should twice in my life have taken possession of me, in addressing the public. The first time was three or four years since, when I favoured the readerâ€"inexcusably, and for no earthly reason that either the indulgent reader or the intrusive author could imagineâ€"with a description of my way of life in the deep quietude of an Old Manse. And nowâ€"because, beyond my deserts, I was happy enough to find a listener or two on the former occasionâ€"I again seize the public by the button, and talk of my three yearsâ€™ experience in a Custom-House. The example of the famous â€œP. P., Clerk of this Parish,â€□ was never more faithfully followed. The truth seems to be, however, that when he casts his leaves forth upon the wind, the author addresses, not the many who will fling aside his volume, or never take it up, but the few who will understand him better than most of his schoolmates or lifemates. Some authors, indeed, do far more than this, and indulge themselves in such confidential depths of revelation as could fittingly be addressed only and exclusively to the one heart and mind of perfect sympathy; as if the printed book, thrown at large on the wide world, were certain to find out the divided segment of the writerâ€™s own nature, and complete his circle of existence by bringing him into communion with it. It is scarcely decorous, however, to speak all, even where we speak impersonally. But, as thoughts are frozen and utterance benumbed, unless the speaker stand in some true relation with his audience, it may be pardonable to imagine that a friend, a kind and apprehensive, though not the closest friend, is listening to our talk; and then, a native reserve being thawed by this genial consciousness, we may prate of the circumstances that lie around us, and even of ourself, but still keep the inmost Me behind its veil. To this extent, and within these limits, an author, methinks, may be autobiographical, without violating either the readerâ€™s rights or his own. It will be seen, likewise, that this Custom-House sketch has a certain propriety, of a kind always recognised in literature, as explaining how a large portion of the following pages came into my possession, and as offering proofs of the authenticity of a narrative therein contained. This, in factâ€"a desire to put myself in my true position as editor, or very little more, of the most prolix among the tales that make up my volumeâ€"this, and no other, is my true reason for assuming a personal relation with the public. In accomplishing the main purpose, it has appeared allowable, by a few extra touches, to give a faint representation of a mode of life not heretofore described, together with some of the characters that move in it, among whom the author happened to make one. In my native town of **Salem**, at the head of what, half a century ago, in the days of old **King Derby**, was a bustling wharfâ€"but which is now burdened with decayed wooden warehouses, and exhibits few or no symptoms of commercial life; except, perhaps, a bark or brig, half-way down its melancholy length, discharging hides; or, nearer at hand, a **Nova** Scotia schooner, pitching out her cargo of firewoodâ€"at the head, I say, of this dilapidated wharf, which the tide often overflows, and along which, at the base and in the rear of the row of buildings, the track of many languid years is seen in a border of unthrifty grassâ€"here, with a view from its front windows adown this not very enlivening prospect, and thence across the harbour, stands a spacious edifice of brick. From the loftiest point of its roof, during precisely three and a half hours of each forenoon, floats or droops, in breeze or calm, the banner of the republic; but with the thirteen stripes turned vertically, instead of horizontally, and thus indicating that a civil, and not a military, post of Uncle Samâ€™s government is here established. Its front is ornamented with a portico of half-a-dozen wooden pillars, supporting a balcony, beneath which a flight of wide granite steps descends towards the street. Over the entrance hovers an enormous specimen of the American eagle, with outspread wings, a shield before her breast, and, if I recollect aright, a bunch of interminglded thunderbolts and barbed arrows in each claw. With the customary infirmity of temper that characterizes this unhappy fowl, she appears by the fierceness of her beak and eye, and the general truculency of her attitude, to threaten mischief to the inoffensive community; and especially to warn all citizens careful of their safety against intruding on the premises which she overshadows with her wings. Nevertheless, vixenly as she looks, many people are seeking at this very moment to shelter themselves under the wing of the federal eagle; imagining, I presume, that her bosom has all the softness and snugness of an eiderdown pillow. But she has no great tenderness even in her best of moods, and, sooner or laterâ€"oftener soon than lateâ€"is apt to fling off her nestlings with a scratch of her claw, a dab of her beak, or a rankling wound from her barbed arrows. The pavement round about the above-described edificeâ€"which we may as well name at once as the Custom-House of the portâ€"has grass enough growing in its chinks to show that it has not, of late days, been worn by any multitudinous resort of business. In some months of the year, however, there often chances a forenoon when affairs move onward with a livelier tread. Such occasions might remind the elderly citizen of that period, before the last war with England, when **Salem** was a port by itself; not scorned, as she is now, by her own merchants and ship-owners, who permit her wharves to crumble to ruin while their ventures go to swell, needlessly and imperceptibly, the mighty flood of commerce at New York or Boston. On some such morning, when three or four vessels happen to have arrived at once usually from Africa or South Americaâ€"or to be on the verge of their departure thitherward, there is a sound of frequent feet passing briskly up and down the granite steps. Here, before his own wife has greeted him, you may greet the sea-flushed ship-master, just in port, with his vesselâ€™s papers under his arm in a tarnished tin box. Here, too, comes his owner, cheerful, sombre, gracious or in the sulks, accordingly as his scheme of the now accomplished voyage has been realized in merchandise that will readily be turned to gold, or has buried him under a bulk of incommodities such as nobody will care to rid him of. Here, likewiseâ€"the germ of the wrinkle-browed, grizzly-bearded, careworn merchantâ€"we have the smart young clerk, who gets the taste of traffic as a wolf-cub does of blood, and already sends adventures in his masterâ€™s ships, when he had better be sailing mimic boats upon a mill-pond. Another figure in the scene is the outward-bound sailor, in quest of a protection; or the recently arrived one, pale and feeble, seeking a passport to the hospital. Nor must we forget the captains of the rusty little schooners that bring firewood from the British provinces; a rough-looking set of tarpaulins, without the alertness of the **Yankee** aspect, but contributing an item of no slight importance to our decaying trade. Cluster all these individuals together, as they sometimes were, with other miscellaneous ones to diversify the group, and, for the time being, it made the Custom-House a stirring scene. More frequently, however, on ascending the steps, you would discernâ€" in the entry if it were summer time, or in their appropriate rooms if wintry or inclement weathersâ€"a row of venerable figures, sitting in old-fashioned chairs, which were tipped on their hind legs back against the wall. Oftentimes they were asleep, but occasionally might be heard talking together, in voices between a speech and a snore, and with that lack of energy that distinguishes the occupants of alms-houses, and all other human beings who depend for subsistence on charity, on monopolized labour, or anything else but their own independent exertions. These old gentlemenâ€"seated, like **Matthew** at the receipt of custom, but not very liable to be summoned thence, like him, for apostolic errandsâ€"were Custom-House officers. Furthermore, on the left hand as you enter the front door, is a certain room or office, about fifteen feet square, and of a lofty height, with two of its arched windows commanding a view of the aforesaid dilapidated wharf, and the third looking across a narrow lane, and along a portion of **Derby** Street. All three give glimpses of the shops of grocers, block-

Figure 8. Screenshot of a partial HTML output from the gender annotator on The Scarlet Letter, using the NLTK dictionary. Red-highlighted words indicate female, while blue-highlighted words indicate male.

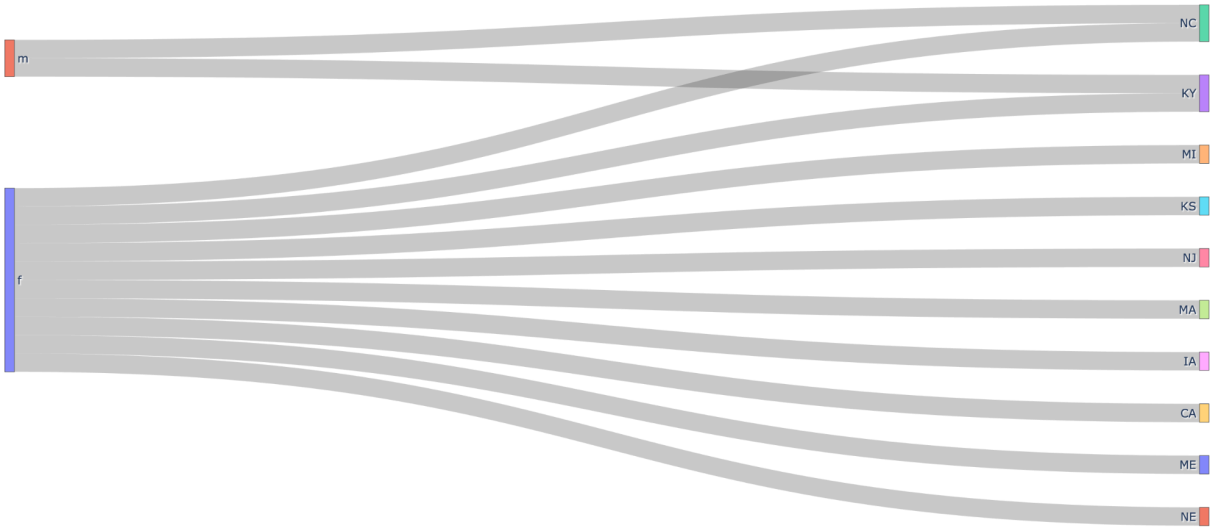


Figure 9. Sankey chart showing whether the name "Hester" is male or female across different US states.

Normalized NER Date Annotator

The normalized NER date annotator reveals a clear dominance of "PRESENT" references for both books (Figure 10). This suggests that both narratives are largely grounded in a sense of "now" or the immediate time frame of the story's unfolding. Both books also feature a significant number of "PAST" references, indicating a strong connection to events or periods that have already occurred. Some examples could be Harry's past encounters with Lord Voldemort and his struggle with his identity in *Harry Potter Book 5* and Hester's past sin of adultery and the social consequences she faces in *The Scarlet Letter*. However, *Harry Potter Book 5* shows a more noticeable presence of "FUTURE" references compared to The Scarlet Letter. This may reflect the speculative nature of fantasy, where prophecies, predictions, and future events play a

significant role. In contrast, *The Scarlet Letter* has a very low frequency of "FUTURE" references, which is consistent with its thematic focus on the consequences of past actions and sins.

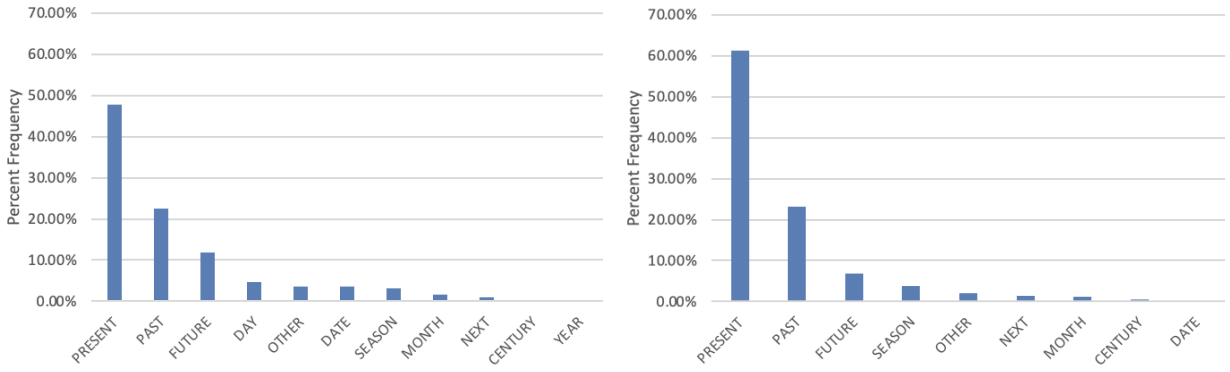


Figure 10. Percent frequency of normalized NER date categories in Harry Potter Book 5 (left) and The Scarlet Letter (right).

Quote/Dialogue Annotator

Figure 11 shows the frequency distribution of the top 10 speakers with the most dialogue in each corpus. We were initially surprised to find that Harry does not have the most dialogue, as shown in the left panel of Figure 11. However, after reviewing the output file, we discovered that many of the quotes labeled as "Unknown" speakers are actually Harry's. While we are unsure why the parser failed to correctly identify Harry's name, if most of the "Unknown" quotes are indeed Harry's, it aligns with our expectations. We are also surprised to find out that Fred, who is not part of the trio, has the second-most dialogue. Additionally, the top 10 speakers with the most dialogue are male, except for Hermione, further reinforcing the male-centered nature of the text.

On the other hand, the right panel of Figure 11 also shows that "Unknown" speakers have the most dialogue in *The Scarlet Letter*. This makes sense since many dialogues in the book are spoken by the unnamed narrator or by public crowds. Hester is the second character with the most dialogue, which is expected given her central role in the story. The right panel also

indicates a balance between female and male characters having dialogue. This suggests that, despite previous findings of the higher frequency of male proper names mentioned in the book, male characters may not necessarily be engaging in dialogues as much as female characters.

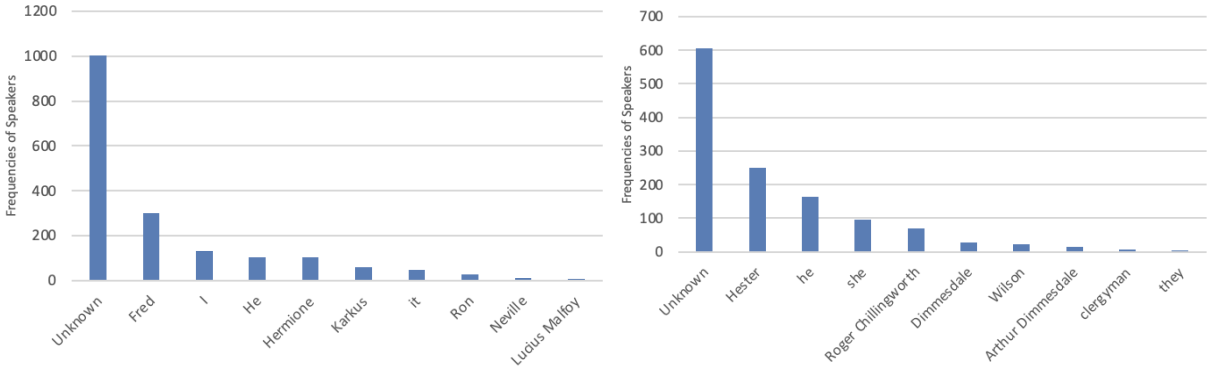


Figure 11. Frequency distribution of the top 10 speakers with the most dialogue in Harry Potter Book 5 (left) and The Scarlet Letter (right).

Clauses Analysis

According to the left panel of Figure 12, *Harry Potter Book 5* has significantly more noun phrases (NP) and verb phrases (VP) than any other clause type. This indicates a strong reliance on basic subject-verb structures in the narrative. Prepositional phrases (PP) also appear frequently, suggesting their use to provide context and detail. Similarly, as shown in the right panel of Figure 13, *The Scarlet Letter* also exhibits a high frequency of NP and VP. However, it is interesting to note that the percent frequency of VP is lower than that of *Harry Potter Book 5*, indicating a greater use of subject structures. Moreover, *The Scarlet Letter* features a more prominent presence of PP, indicating a more descriptive style than *Harry Potter Book 5*. This aligns with previous findings that *The Scarlet Letter* has a more descriptive style with frequent use of imagery and metaphors.

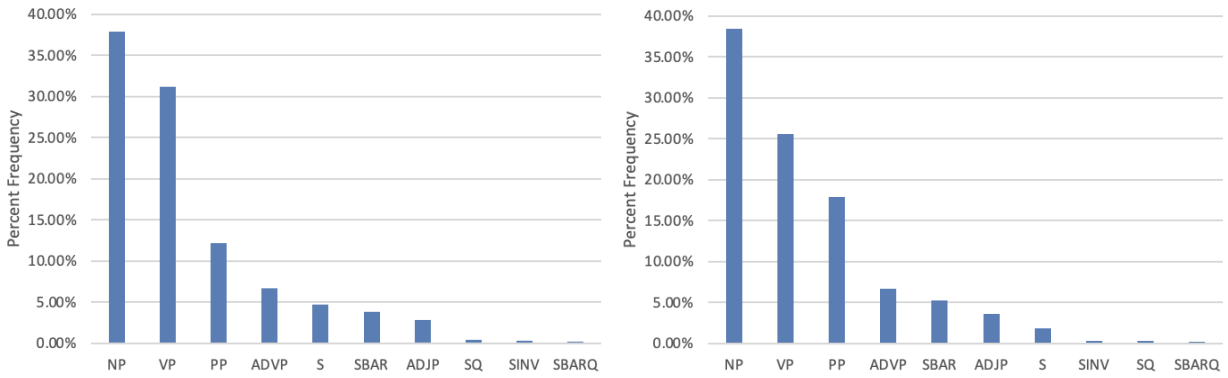


Figure 12. Percent frequency distribution of clause types in Harry Potter Book 5 (left) and The Scarlet Letter (right). NP denotes noun phrase, VP denotes verb phrase, PP denotes prepositional phrase, ADVP denotes adverb phrase, S denotes sentence, SBAR denotes subordinate clause, ADJP denotes adjective phrase, SQ refers to an inverted yes/no question or main clause of a wh-question following the wh-phrase in SBARQ, SINV refers to an inverted declarative sentence where the subject follows the tensed verb or modal, and SBARQ denotes a direct question introduced by a wh-word or wh-phrase.

Nouns Analysis

Figure 13 shows that both corpora have a high frequency of singular/mass nouns (NN). However, this is more pronounced in *The Scarlet Letter* (73.30%) compared to *Harry Potter Book 5* (49.30%). This indicates that *The Scarlet Letter* focuses heavily on objects, concepts, and general descriptions, aligning with the novel's descriptive and symbolic style typical of 19th-century literature. In contrast, *Harry Potter Book 5* has a higher percentage of singular proper nouns (NNP) at 36.44% compared to 12.13% in *The Scarlet Letter*. This emphasizes the importance of characters and their interactions in the story of Harry Potter. Both books contain a similar percentage of plural nouns (NNS) — 13.50% in *Harry Potter Book 5* and 14.38% in *The Scarlet Letter* — and a very low frequency of plural proper nouns (NNPS), which is typical as these are less common in general text.

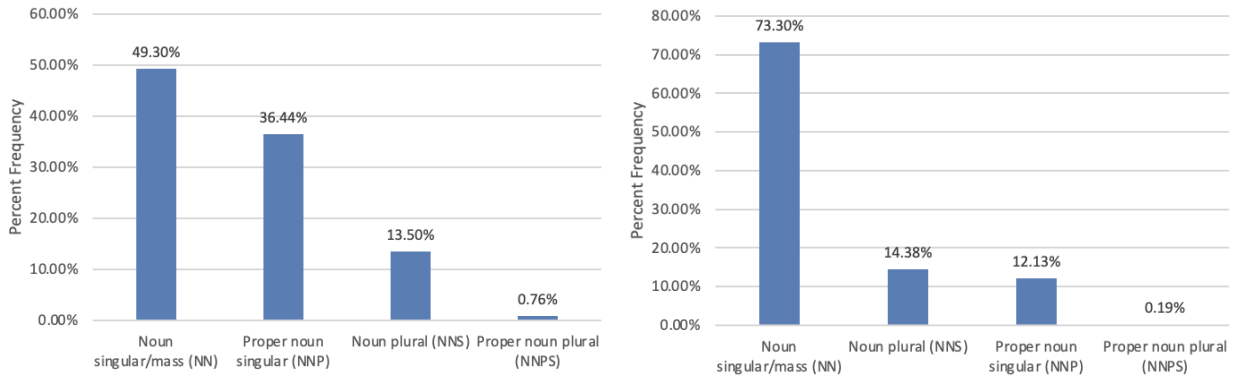


Figure 13. Percent frequency distribution of noun POS tags in Harry Potter Book 5 (left) and The Scarlet Letter (right).

As shown in Figure 14, both texts have a strong dominance of nouns functioning as nominal subjects, which means nouns frequently act as the subject of verbs. However, this dominance is more pronounced in *Harry Potter Book 5* (62.65%) compared to *The Scarlet Letter* (55.35%). Both corpora also show a significant presence of nouns functioning as direct objects. But this time *The Scarlet Letter* (38.15%) has a slightly higher percent frequency than *Harry Potter Book 5* (33.48%). The high frequency of nominal subjects and direct objects in both corpora suggests that both narratives are largely written in the active voice, where subjects perform actions. Overall, the higher percentage of nominal subjects in *Harry Potter Book 5* suggests a more straightforward and action-oriented narrative style, while the more balanced distribution in *The Scarlet Letter* may indicate a slightly more varied and nuanced approach to sentence structure.

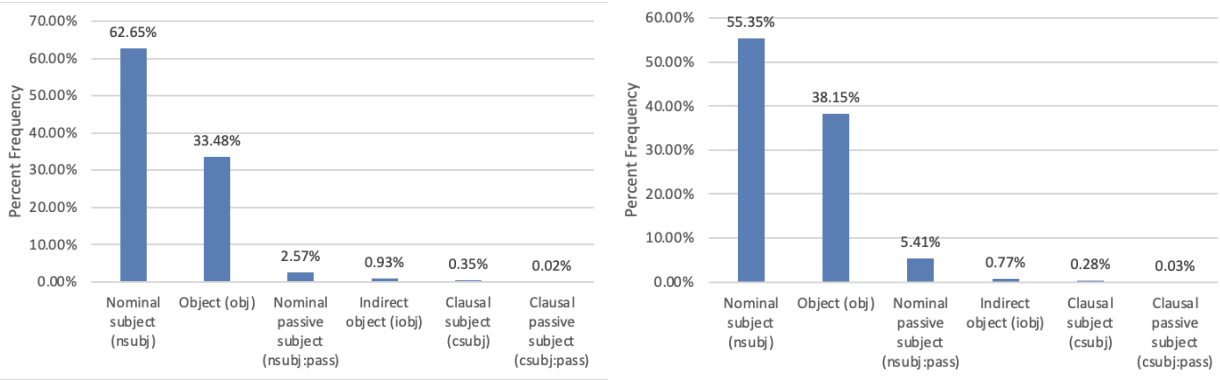


Figure 14. Percent frequency distribution of noun DEPREL tags in Harry Potter Book 5 (left) and The Scarlet Letter (right).

Verbs Analysis

According to Figure 15, *Harry Potter Book 5* has a high frequency of low-value modals (e.g., "can", "could", "may", "might"). This indicates a focus on expressing possibility, ability, and permission. Median-value modals (e.g., "will", "shall", "should") are present but less frequent, while high-value modals are rare. The prevalence of low-value modals suggests a narrative that emphasizes possibility, choice, and the character's abilities. This aligns with the fantasy genre, where characters often face challenges that require them to utilize their skills and make decisions. In contrast, *The Scarlet Letter* shows similar percent frequencies for median-value and low-vale modals, with a low occurrence of high-value modals. The dominance of median-value modals suggests a narrative that emphasizes obligation, duty, and fate. This reflects the novel's themes of morality, societal expectations, and the consequences of actions. In sum, Harry Potter Book 5's low-value modals emphasize character agency and the ability to choose one's path, while The Scarlet Letter's median-value modals suggest a stronger sense of determinism or external forces shaping the characters' actions.

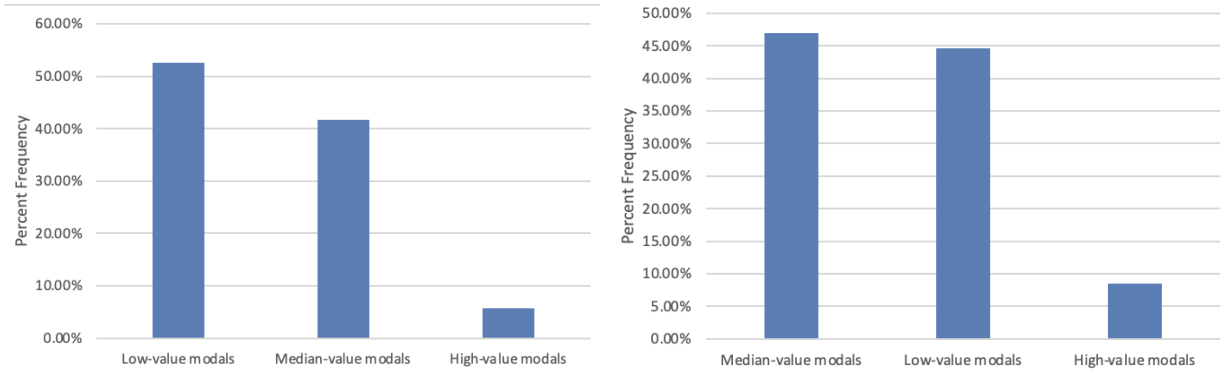


Figure 15. Percent frequency distribution of verb modality values in Harry Potter Book 5 (left) and The Scarlet Letter (right). High-value modal verbs are "must", "ought to", and "have to;" median-value modal verbs are "will", "shall", and "should;" low-value modal verbs are "can", "could", "may", and "might".

Figure 16 shows that Both corpora exhibit a strong dominance of the past tense. This is expected in narrative texts, as they primarily recount events that have already occurred. However, this finding contrasts with the results of the normalized NER date analysis, where more present date words were observed. The discrepancy could be explained by the fact that although the narratives often focus on past events, they may include significant references to the "present" in terms of the characters' immediate actions or emotions. Other than that, both corpora also show a notable presence of infinitives, which are often used to express purpose, possibility, or future actions. Both corpora contain gerunds, which are verb forms ending in "-ing" that function as nouns.

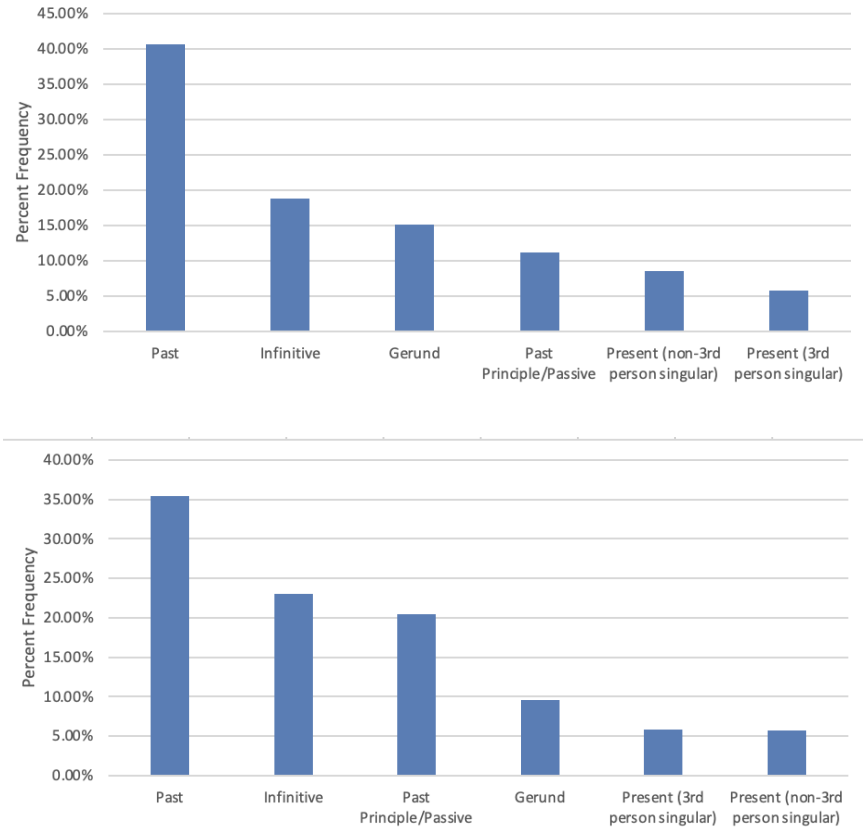


Figure 16. Percent frequency distribution of verb tense values in Harry Potter Book 5 (top) and The Scarlet Letter (bottom).

Based on Figure 17, both Harry Potter Book 5 and The Scarlet Letter utilize active verbs more than passive verbs. Both corpora show a very low frequency of the passive voice, where the subject is acted upon. This indicates that the subjects of sentences are typically performing the actions. This also suggests a narrative style that emphasizes direct action and agency, which aligns with the previous finding that most nouns function as nominal subjects. The focus on active voice enhances the immediacy and engagement of the storytelling for both corpora, making the characters' actions and decisions more prominent.

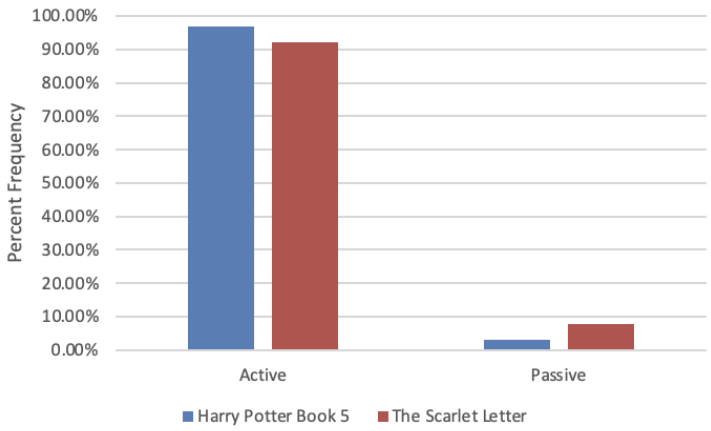


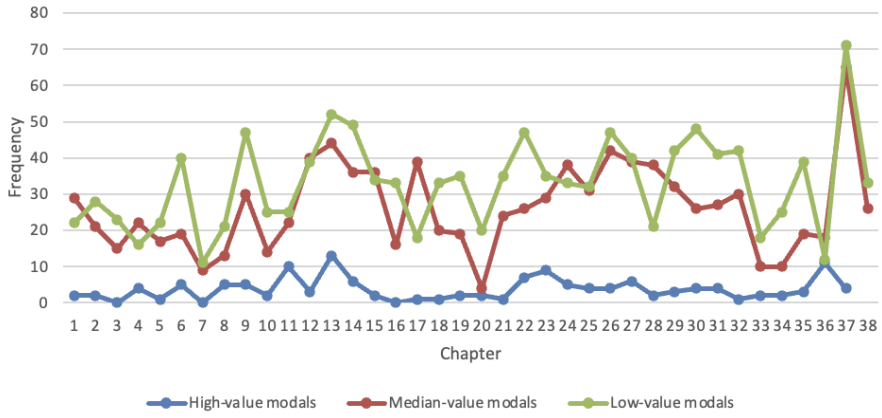
Figure 17. Percent frequency distribution of verb voice values in Harry Potter Book 5 and The Scarlet Letter.

Comparison to Moretti and Pestre

Moretti and Pestre (2015) used noun and verb statistics to uncover significant shifts in the language and priorities of the World Bank over time. They identified a transition from concrete, material-focused language (e.g., “loan”, “development”) in the early decades to more abstract, financial, and managerial language (e.g., “portfolio”, “derivative”) in later years (Moretti and Pestre 76-79). This reflected a move from industrial and infrastructural projects to financialization and governance. They also noted an increase in nominalizations (nouns derived from verbs, ending in -tion, -sion, -ment), which made the language more abstract and less concrete (Moretti and Pestre 92). This abstraction helped obscure the subjects of actions and decisions, making the language more opaque. Additionally, they observed a shift from clear delineations of past, present, and future actions to a more ambiguous use of progressive and gerund forms, suggesting a focus on ongoing processes rather than completed actions (Moretti and Pestre 99).

Noun analysis of *Harry Potter Book 5* and *The Scarlet Letter* reveals that both texts use a lot of singular/mass nouns (NN) and feature many nouns functioning as nominal subjects and

direct objects. However, it remains unclear whether most of these nouns are concrete or abstract. For example, in the sentences “The cat is sleeping” and “The dream was vivid,” both “cat” and “dream” are singular/mass nouns and nominal subjects, but “cat” is concrete while “dream” is abstract. According to Figure 18, both corpora show a low frequency of high-value modals, though *Harry Potter Book 5* exhibits an increase in median-value and low-value modals throughout the book. In contrast, *The Scarlet Letter* does not show a similar trend. This aligns with Moretti and Pestre’s observation of shifts in language reflecting different priorities, as *Harry Potter* emphasizes possibility and character agency. Figure 19 indicates a slight increase in the use of past and infinitive verbs in *Harry Potter Book 5*, while *The Scarlet Letter* shows a decrease in verb usage overall. In summary, while our results confirm some of Moretti and Pestre’s findings about the use of nouns and verbs in narrative texts, they also highlight differences due to the distinct genres and purposes of the texts analyzed. Most importantly, our findings complement Moretti and Pestre’s by showing how noun and verb statistics can reveal different aspects of language use in literature versus institutional reports.



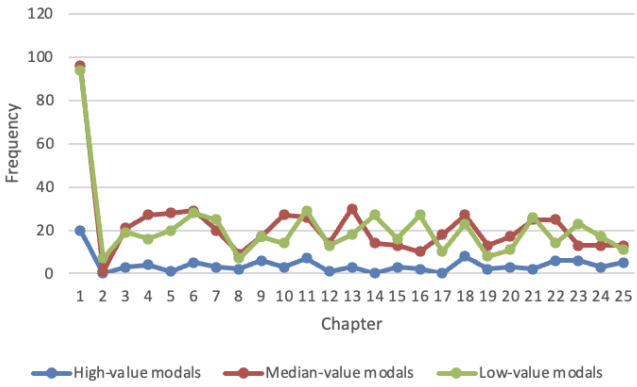
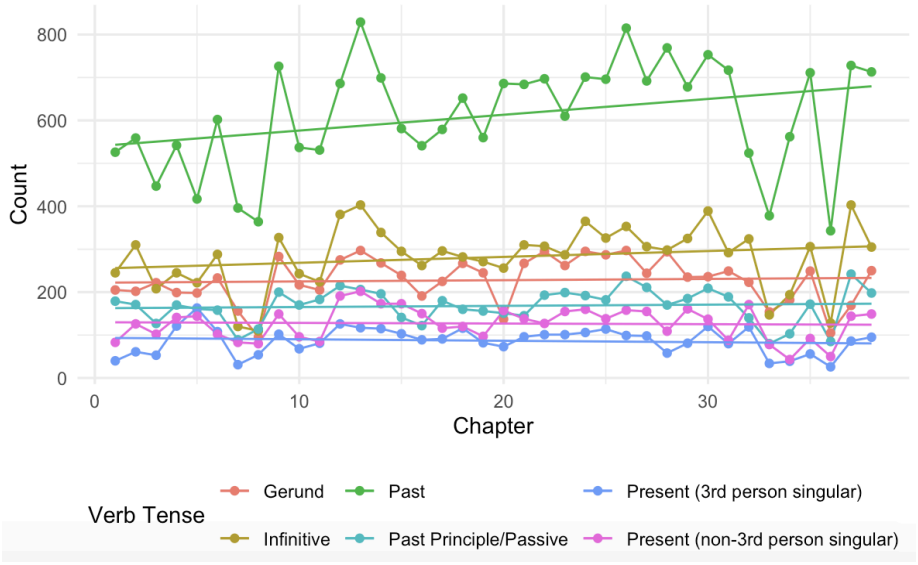


Figure 18. Frequency distribution of verb modality values in Harry Potter Book 5 (top) and The Scarlet Letter (bottom) by chapter. High-value modal verbs are "must", "ought to", and "have to;" median-value modal verbs are "will", "shall", and "should;" low-value modal verbs are "can", "could", "may", and "might".



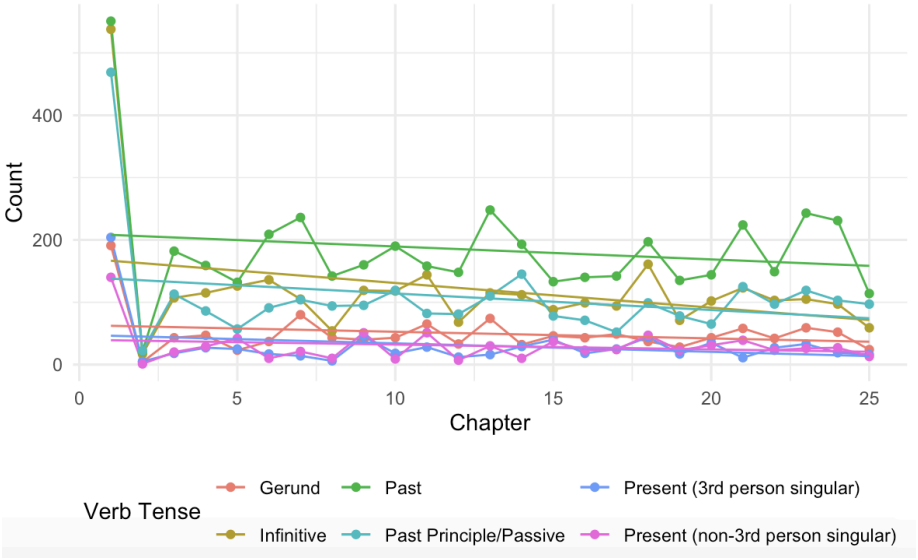


Figure 19. Frequency distribution of verb tense values in Harry Potter Book 5 (top) and The Scarlet Letter (bottom) by chapter. Linear regression lines are plotted.

Stopwords Analysis

Figure 20 shows the frequency distributions of articles, auxiliary verbs, conjunctions, prepositions, and pronouns in *Harry Potter Book 5* and *The Scarlet Letter*. In *Harry Potter Book 5*, there are 20 unique articles, 94 auxiliary verbs, 113 conjunctions and prepositions, and 45 pronouns, totaling 272 function words. In *The Scarlet Letter*, there are 17 articles, 40 auxiliary verbs, 96 conjunctions and prepositions, and 41 pronouns, totaling 194 function words. According to Chung and Pennebaker, a small set of fewer than 400 function words accounts for over half of the words used in daily speech, with around 300 being the most frequently used (Chung and Pennebaker 347). A total number of 272 function words in *Harry Potter Book 5* and 194 function words in *The Scarlet Letter* are both below the typical set of 300 most frequently used “junk” words in the English language. While these corpora do not fully comply with the typical distribution of function words in daily speech, they do reflect a substantial portion of

commonly used function words, aligning with the idea that such words form a significant portion of text in general.

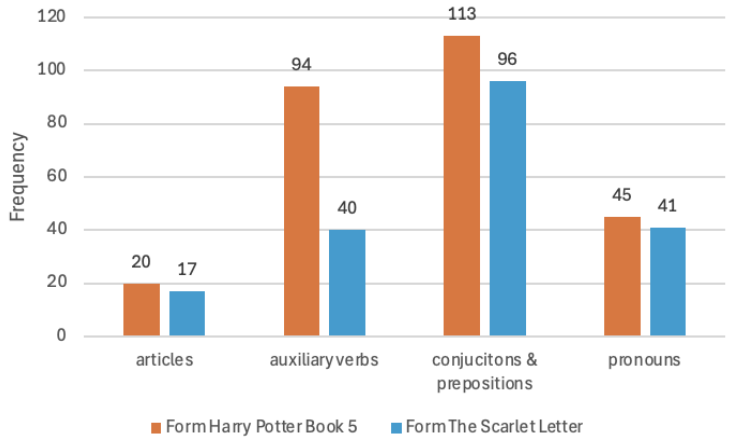


Figure 20. Frequency distribution of FORM function/stopword categories in Harry Potter Book 5 and The Scarlet Letter.

Pennebaker et al. argue that "junk" words, specifically pronouns, are far from meaningless and are actually crucial in understanding social and psychological processes. They suggest that pronouns carry significant psychological and social meaning, with first-person singular pronouns (e.g., "I", "me") indicating self-focus, while first-person plural pronouns (e.g., "we", "us") reflect group identity or social integration (Pennebaker et al. 547). Pronouns can also serve as markers of emotional state, social identity, and cognitive styles. For instance, higher use of first-person singular pronouns has been linked to depression, self-focus, and even physical health outcomes like heart disease. Conversely, shifts in pronoun use (e.g., from "I" to "we") can indicate changes in social dynamics or emotional states, such as during times of crisis or trauma (Pennebaker et al. 570). Additionally, pronouns act as referential words and require a shared understanding between the speaker and the listener. This makes them particularly important in social interactions, as they reflect the speaker's awareness of the audience's perspective (Pennebaker et al. 570). Therefore, Pennebaker et al. suggest that future research should delve

deeper into the nuances of pronoun use. Even within first-person singular pronouns, distinctions such as between the "active I" and the "passive me" have yet to be fully explored (Pennebaker et al. 571).

Search CoNLL Table Tool

Given our findings so far, we wanted to search LEMMA words "Hermione" and "Ron" since they are among the top 10 most frequent nouns in *Harry Potter Book 5* (Figure 21). Previously, we found that *Harry Potter Book 5* is male-centric, and we are wondering if there are any differences in how male and female characters are described. Similarly, we wanted to search LEMMA words "Hester" and "minister" (also known as Arthur Dimmingsdale) in *The Scarlet Letter*, given our surprise that the book mentions more male characters than female ones. We wanted to explore how Hester and Dimmingsdale are portrayed differently, despite both committing adultery, particularly in terms of how they are described by the narrator and the public. The setting we used to run the above analysis is under the "Co-occurring token POSTAG", we chose "JJ* - Any adjective". Everything else is left as default.

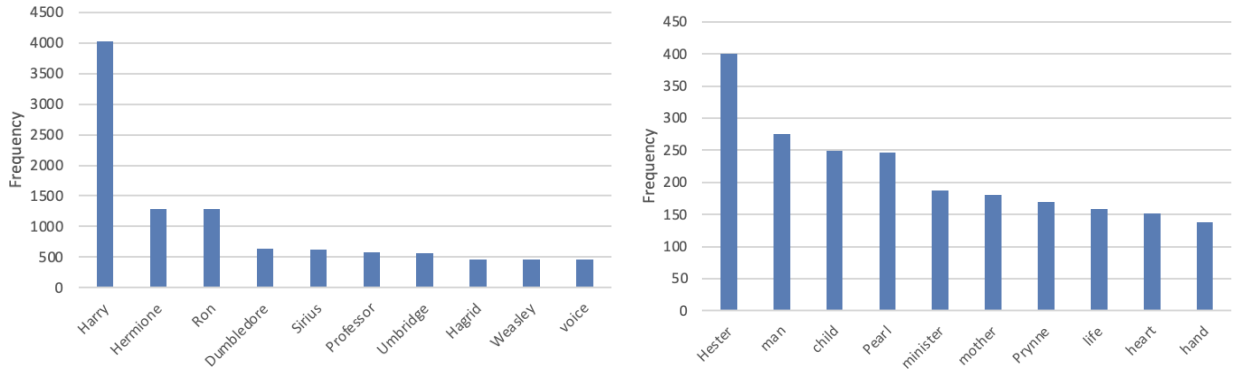


Figure 21. Frequency distribution of the top 10 most frequent nouns (lemmas) in Harry Potter Book 5 (left) and The Scarlet Letter (right).

The left panel of Figure 22 reveals that Hermione co-occurs with adjectives linked to both positive (e.g., right) and negative qualities (e.g., terrified, unconscious, white, squashed), as

well as terms related to concentration (e.g., preoccupied). These adjectives suggest that Hermione may experience fear, anxiety, and moments of incapacity in this book. In contrast, Ron is associated with adjectives reflecting positivity (e.g., good, proud), a tendency to sleep (e.g., asleep), and untidiness (e.g. disheveled). Figure 22 shows that Hermione is portrayed as intelligent, emotional, and sometimes vulnerable, while Ron is portrayed as generally good, but also flawed and prone to negative experiences or judgments. Though further analysis is needed to conclude if there are any gender stereotypes in this text, Hermione's word cloud is more emotionally charged, while Ron's word cloud shows a balance of positive and negative traits, suggesting his role as a relatable, everyman character.

Figure 22. Word clouds of co-occurring adjectives with "Hermione" (left) and "Ron" (right) in Harry Potter Book 5.

As shown in Figure 23, Hester is linked to adjectives such as "helpful", "hush", and "sight". This suggests Hester is a good person and provides assistance or support to others. She is quiet and not sturdily built. Additionally, "miserable" and "contrived" suggest a sense of

suffering, artifice, or being marked, likely due to the community's judgment. On the other hand, the minister or Dimmesdale is associated with a lot of negative adjectives such as "poor", "old", and "pale", suggesting his physical weakness. However, terms like "holy" and "pious" highlight his public persona as a religious figure. Overall, Hester's word cloud portrays her as more active and practical despite her social ostracization, while Dimmesdale's word cloud emphasizes his public image and his internal struggle driven by his guilt. It is interesting to note that though both characters suffer, Hester's suffering is more external (social rejection), while Dimmesdale's is more internal (guilt and hypocrisy).

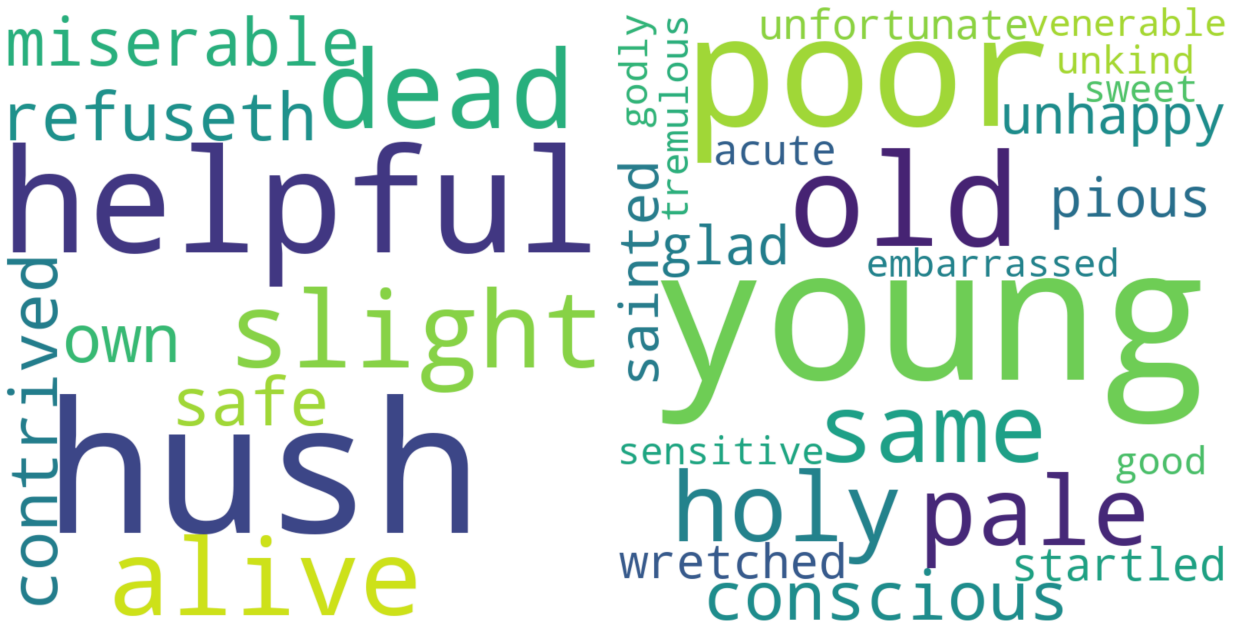


Figure 23. Word clouds of co-occurring adjectives with "Hester" (left) and "minister" (right) in The Scarlet Letter.

In sum, the CoNLL table search tool allows for more nuanced analyses by focusing on the grammatical structure and relationship between words through co-occurring tokens and parts of speech. Unlike N-gram searches which cannot identify contiguous words's part of speech, the CoNLL table search enables exploration of how specific words relate to, for example, adjectives.

This is more insight into the emotional or descriptive qualities attributed to characters. Moreover, this can address questions like whether different characters are described using positive or negative terms, revealing gendered or thematic patterns not apparent from previous analyses. Lastly, the CoNLL tool provides a more precise way to analyze semantic distance, as it allows for the comparison of characters' portrayals based on linguistic context rather than just raw word proximity or frequency generated from word sense induction or word embeddings.

Compare & Contrast This and Past Week Findings

This week's analysis reveals distinct linguistic patterns that align with previous findings. NER analysis highlights *Harry Potter Book 5* as character-driven, with a high frequency of PERSON entities. This is consistent with N-grams and topic modeling's findings since those visualizations identified character names as the most frequent words. The gender and quote/dialogue annotators show that the book is male-dominated, and verb analysis emphasizes active actions and possibility with a pronounced presence of low-value modals. This supports N-gram and topic modeling results as a lot of action verbs and the themes of prophecy and uncertainty are identified. *The Scarlet Letter*, conversely, features a more descriptive style with a focus on singular/mass nouns from the clauses and noun analysis. This aligns with previous style analysis using vocabulary richness and abstractness/concreteness. Verb analysis indicates that the novel has more median-value modals, reflecting themes of societal expectations and fate. Gender analysis reveals male dominance in *The Scarlet Letter*, which was unexpected because no previous analysis has shown that. But later dialogue analysis reveals a more balanced female and male presence in *The Scarlet Letter*. Lastly, the CoNLL table search tool adds to the total findings and offers a new understanding of character portrayals. It reveals Hermione's

intellectual strength and emotional complexity versus Ron's good personality and relatability, and Hester's practical strength versus Dimmesdale's internal struggle.

Literature Review

Muzny et al. introduce a computational metric called dialogism to analyze the dialogic nature of narration and quotations in novels. It is based on Mikhail Bakhtin's theory that all texts are fundamentally dialogic (Muzny et al. 31). Using a corpus of 1,100 English novels spanning 230 years, the authors identify grammatical features (e.g., tense, pronouns, clause structure) that distinguish spoken dialogue from narrative text. They find that novels have become increasingly dialogic over time, with dialogue playing a more central role in driving the narrative (Muzny et al. 36). In our quote/dialogue analysis, we found 1822 speakers in *Harry Potter Book 5* and 1289 speakers in *The Scarlet Letter*. To approximate the percentage of sentences that are dialogues, we could divide the number of speakers by the sentence count in each corpus. This results in 10.65% of sentences in *Harry Potter Book 5* being dialogue and 36.87% in *The Scarlet Letter*. These findings somewhat contradict Muzny et al.'s conclusion that dialogue has increased in prominence over time. However, to confirm or refine their findings, further analysis on different books would be needed.

Bonyadi's paper examines the linguistic manifestations of modality in newspaper editorials from The New York Times (NYT) and Tehran Times (TT). Specifically, it looks at how editorial writers use modality to influence readers' opinions. It identifies that both newspapers predominantly use predictive modal verbs like "will" and "would", with NYT emphasizing future predictions and TT focusing more on necessity and obligation (Bonyadi 10). In our verb analysis, instead of calling them predictive modal verbs, verbs like "will" and "would" are in the median-value modal verbs. Given that *The Scarlet Letter* actually contains more median-value

modal verbs than *Harry Potter Book 5*, this suggests that the language of *The Scarlet Letter* may emphasize obligation, duty, and necessity as well. Lastly, Bonyadi's paper concludes with pedagogical implications, suggesting that understanding modality in editorials can enhance EFL students' rhetorical awareness and critical reading skills (Bonyadi 11).

The paper by Chung and Pennebaker explores the psychological functions of function words in natural language, highlighting their role in reflecting social, cognitive, and emotional processes. Using computerized text analysis tools like LIWC, the authors show that these often overlooked function words provide insights into personality, social interactions, and psychological states, such as depression, deception, and cultural differences. For instance, the use of first-person singular pronouns is linked to self-focus and negative emotional states, while shifts in pronoun usage can indicate changes in social dynamics or stress responses (Chung and Pennebaker 352, 351). Moreover, the authors find that function words are powerful markers to study human behavior. In particular, gender and age-related linguistic differences, such as men using more concrete nouns and women more auxiliary verbs, as well as older individuals showing a preference for future tense, emphasize how language use can reflect cognitive and emotional states (Chung and Pennebaker 353-354). Given this, it would be interesting to investigate whether our assigned corpora, though both are fictional, align with Chung and Pennebaker's findings on gendered pronoun usage. Specifically, it would be interesting since *Harry Potter Book 5* is written by a female author but most of the characters are males, and *The Scarlet Letter* is written by a male author but centers on a female protagonist.

In their paper, Franzosi et al. propose Quantitative Narrative Analysis (QNA) as a method to measure agency in socio-historical research by analyzing narrative texts, particularly focusing on lynchings in Georgia. QNA organizes narrative data into structured sequences of

Subject-Verb-Object (SVO) triples, capturing actors, actions, and interactions, which can then be analyzed using tools like network analysis, GIS, and sequence analysis (Franzosi et al. 1). The authors illustrate how QNA can reveal patterns of agency in lynching events, highlighting the roles of mobs, law enforcement, and victims. They also address the limitations of newspaper sources, which often obscure agency through passive constructions and nominalization (Franzosi et al. 10). As they note, nominalization and passive voice serve to deny agency by removing or obscuring the perpetrators, thereby reinforcing power imbalances and limiting accountability (Franzosi et al. 25). Though in this assignment we did not look at nominalization, in our verb analysis, we found out that both corpora contain gerunds (verb forms ending in "-ing" that function as nouns). This indicates the potential use of nominalization in both corpora, suggesting that agency may be obscured or generalized to minimize the role of perpetrators and reinforce power dynamics.

Conclusion

This analysis explores the linguistic intricacies of *Harry Potter Book 5* and *The Scarlet Letter*, employing a range of NLP tools to uncover distinct patterns that both reinforced and expanded upon previous findings and established literary scholarship. The evaluation of parser performance reveals that Stanford CoreNLP prioritizes accuracy at the cost of speed, while SpaCy excels in speed and the number of languages supported, and Stanza offers a balanced approach. NER analysis highlights the character-driven nature of both narratives, particularly in *Harry Potter Book 5*. The book also exhibits a higher frequency of ORGANIZATION entities, reflecting its institutional themes. In contrast, *The Scarlet Letter* showcases a prominence of TITLE entities, underscoring its focus on authority.

Gender analysis reveals a dominance of male proper names in both corpora. Dialogue analysis further emphasizes the male-dominated nature of *Harry Potter Book 5*, while *The Scarlet Letter* displays a more equitable distribution of dialogue between males and females. Clause and noun analysis indicated a heavy reliance on noun phrases and verb phrases in both texts. *The Scarlet Letter* exhibits a higher frequency of singular/mass nouns and prepositional phrases, reflecting its descriptive style. Conversely, *Harry Potter Book 5* features more singular proper nouns, emphasizing character interactions.

Additionally, verb analysis reveals that *Harry Potter Book 5* prioritizes possibility and character agency through a higher frequency of low-value modals, while *The Scarlet Letter* emphasizes obligation and fate with a greater presence of median-value modals. Both corpora predominantly employed active voice and past tense. Stopwords analysis offers insights into social and psychological processes, supporting Pennebaker et al.'s research. The CoNLL table search tool facilitated a nuanced analysis of character portrayals, revealing emotional and descriptive qualities associated with specific characters and enabling the examination of relationships between words and their descriptive adjectives. Finally, the literature review demonstrated how our findings both supported and "contradicted" prior research. It also shows how they could be used to supplement existing scholarship, ultimately showcasing the power of computational linguistics in revealing hidden patterns and themes within literary texts.

Work Cited

Bonyadi, Alireza. “Linguistic Manifestations of Modality in Newspaper Editorials.” *International Journal of Linguistics*, vol. 3, no. 1, 30 Mar. 2011, https://doi.org/10.5296/ijl.v3i1.799.

Chung, Cindy, and James Pennebaker. “The Psychological Functions of Function Words.” In K. Fiedler (Ed.), *Social communication*, Psychology Press, 2007 pp. 343-359.

Franzosi, Roberto. NLP TIPS files.

Franzosi, Roberto, et al. “Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia.” *Sociological Methodology*, vol. 42, no. 1, Aug. 2012, pp. 1–42, https://doi.org/10.1177/0081175012462370.

Jørgensen, A., Hovy, D. and Søgaard, A. 2016. “Learning a POS tagger for AAVE-like language.” *In Proceedings of NAACL-HLT*. San Diego, CA, USA, pp. 1115–20.

Manning, Christopher D., et al. “The Stanford CoreNLP Natural Language Processing Toolkit.” *Www.aclweb.org*, 1 June 2014, www.aclweb.org/anthology/P14-5010/.

Moretti, Franco, and Dominique Pestre. “Bankspeak.” *New Left Review*, no. 92, 1 Apr. 2015, pp. 75–99, newleftreview.org/issues/ii92/articles/franco-moretti-dominique-pestre-bankspeak.

Muzny, Grace, et al. “Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations.” *Digital Scholarship in the Humanities*, vol. 32, no. suppl_2, 18 July 2017, pp. ii31–ii52, https://doi.org/10.1093/llc/fqx031.

Pennebaker, James W., et al. “Psychological Aspects of Natural Language Use: Our Words, Our Selves.” *Annual Review of Psychology*, 2003 vol. 54, pp. 547-77.