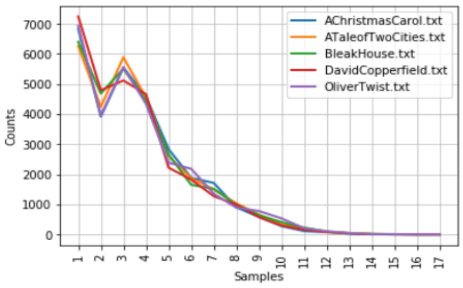
Readability of the Works of Charles Dickens and Nathaniel Hawthorne

By Travis Gubbe

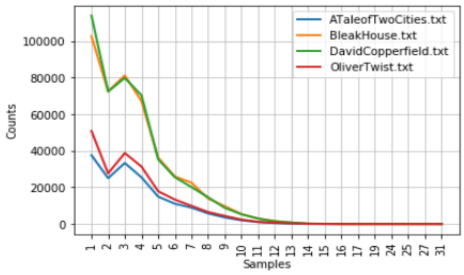
Considered two of the most well-known authors of his time, Charles Dickens and Nathaniel Hawthorne were 19th century writers, with Dickens known for his themes on social issues and satire while Hawthorne is known for his style of romanticism. Though these authors wrote during the same time period, the authors have a varying style due to their different upbringing, with Charles Dickens born in England and Nathaniel Hawthorne born in the United States. Due to these authors having a different background, this paper wants to answer the following questions: does the readability of Charles Dickens and Nathaniel Hawthorne differ from one another, despite the authors writing during the same time period? If so, what may be the cause of the difference in readability?

For this corpus, the corpus builder GutenTag was used to construct a corpus of Charles Dickens' and Nathaniel Hawthorne's various works. The corpus builder found 51 works by Charles Dickens and over 90 works for Nathaniel Hawthorne. Since this corpus wants to focus on the readability of these two authors, the corpus was narrowed to five works of Charles Dickens and six works of Nathaniel Hawthorne. For Charles Dickens, this includes *A Christmas Carol*, *A Tale of Two Cities*, *Bleak House*, *David Copperfield*, and *Oliver Twist*. For Nathaniel Hawthorne, this includes *A Wonder Book*, *Doctor Grimshawe's Secret*, *The Blithedale Romance*, *The Scarlet Letter*, and *Twice Told Tales*. A smaller amount of Dickens text was acceptable for this corpus since the Dickens texts totaled over 1 million words while the Hawthorne texts totaled a little over 600,000. Despite the slight discrepancy in the amount of words for these two authors, the analyses performed should not be affected.

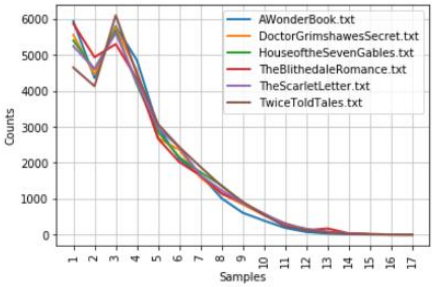
For this analysis, the first analysis done was word length frequencies for Dickens and Hawthorne. Two different word length frequency distribution plots were done for each author: one showing the word length frequency distribution for the first 30,000 words of each text (to include each text) and one showing the word length frequency distribution of the works that contained at least 100,000 words (Figure 1). These two different distributions were done to include works such as *A Christmas Carol* (29,256 words) and *The Scarlet Letter* (69,811 words), which contained much less than 100,000 words and would have vastly different word length distributions than the works that contain more than 100,000 words.



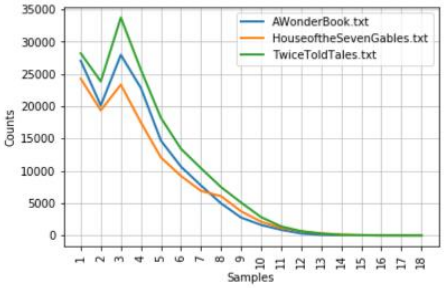
Word length frequency for all Dickens texts



Word length frequency for the Dickens texts with over 100,000 words



Word length frequency for all Hawthorne texts



Word length frequency for the Hawthorne texts with over 100,000 words

As seen from the graphs above, Dickens tended to use words with character length of 6 or less at a higher ratio than Hawthorne, with Dickens having a peak at a word length of 1 while Hawthorne having a peak at a word length of 3. Even when looking at the first 30,000 words for each author, Dickens used much more words containing one to three characters, with up to 7,000 one-character words for a few texts in the first 30,000 words. This change in word length distribution can lead to a noticeable difference in readability, which we can note with various readability metrics.

According to scholars Neil Newbold and Lee Gillam of the University of Surrey, readability can be defined as "a measure of how easy a text is to understand," which include text factors such as syntax and vocabulary (Newbold and Gillam 2). Three readability metrics were done for each text: Automated Readability Index (ARI), Coleman-Liau Index (CL), and Flesch-Kincaid Readability (FK). Both ARI and CL provide grade reading levels for written material, where ARI was originally created for technical documents in the US Air Force (Zhou, Jeong, and Green 4) while CL was created to "facilitate scoring using mechanical counting devices," (Zhou, Jeong, and Green 4). FK also provides a grade reading level of the written material and was developed for the US Navy (Zhou, Jeong, and Green 3). In their equations, ARI uses the total

characters, words, and sentences in a document while CL calculates by the average amount of letters and sentences per 100 words, whereas FK uses the total number of syllables in the document in its equation to determine its reading level in addition to total words and sentences (Zhou, Jeong, and Green 3-4). These reading metrics have the same principle: the higher the score, the more difficult the material.

The three readability indexes were performed on each text in the corpus. The following grade reading levels are given below:

Dickens Texts	**ARI**	**CL**	**FK**
A Christmas Carol	5.62	6.64	5.67
A Tale of Two Cities	7.34	7.36	7.45
Bleak House	6.85	6.80	7.41
David Copperfield	7.08	6.50	7.65
Oliver Twist	7.37	7.41	7.49
Hawthorne Texts	**ARI**	**CL**	**FK**
A Wonder Book	9.76	7.93	9.09
Doctor Grimshawes Secret	13.66	9.06	12.37
House of the Seven Gables	11.34	9.30	10.65
The Blithedale Romance	10.16	8.79	10.01
The Scarlet Letter	12.98	9.43	11.82
Twice Told Tales	11.83	9.21	10.73

For the most part, the three grade reading levels are relatively close to each other when predicting the reading level for Charles Dickens, while the three grade reading levels vary much more when predicting the reading level for Nathaniel Hawthorne, particularly the CL compared to ARI and FK. This result should not be surprising given what was seen with the word frequency distributions for both authors. Since Dickens seemed to have more words that were 5 characters or less compared to Hawthorne, it can be expected that Hawthorne will have higher readability scores than Dickens since the readability metrics looked at various sentence structuring for each text.

No matter what reading metric is preferred, however, Nathaniel Hawthorne's works typically have a higher readability score than Charles Dickens' works. This answers the first question, confirming that Dickens and Hawthorne differ in readability. The next part to answer is what may be the cause of this difference in readability. To answer this part, we will view three different factors that may help determine what is the cause of this difference in readability for these two authors.

The first factor viewed to see what may affect the readability of the texts was average sentence length for each text. As seen in the equations for ARI, CL, and FK reading metrics, sentences play a pivotal role in determining the readability score. What we hope to see is that the greater the average sentence length, then the greater the readability score for a text. For each text, the average sentence length was calculated by dividing the total amount of words in the text by the total number of sentences in the text. The average sentence length, or ASL, for each text is found below:

Dickens Texts	**ASL**
A Christmas Carol	14.99
A Tale of Two Cities	17.77
Bleak House	17.66
David Copperfield	18.77
Oliver Twist	17.73

Hawthorne Texts	**ASL**
A Wonder Book	22.24
Doctor Grimshawe's Secret	28.71
House of the Seven Gables	30.33
The Blithedale Romance	21.59
The Scarlet Letter	26.61
Twice Told Tales	24.53

Overall, Hawthorne texts have a much greater average sentence length than Dickens texts. The average sentence length seems to correlate with the increase in readability for the Dickens text. For example, when looking at Dickens' texts, *A Christmas Carol* has the shortest sentence length at 14.99 and *David Copperfield* has the longest sentence length at 18.77. When looking at the reading scores, *A Christmas Carol* has the lowest reading score among all three reading indexes, while *David Copperfield* has the highest score when looking at the FK index. Though the average sentence length seems to work well as an indicator for reading levels for Dickens, it does not seem to work so well for Hawthorne. When looking at Hawthorne's texts, *The Blithedale Romance* has the shortest average sentence length, while *House of the Seven Gables* has the longest average sentence length. When looking at the reading scores, however, *The Blithedale Romance* does not have the lowest reading score on any of the three reading metrics, while *House of the Seven Gables* does not have the highest reading score on any of the three reading metrics.

Overall, average sentence length may not be the most important factor in determining readability, though it can still be an important factor. The reason average sentence length may work as a factor for Dickens' texts compared to Hawthorne's texts is because the readability equations use words, character counts, and even syllables in determining the readability of a text.

The average sentence length may have been a more important factor with the Dickens texts among the variables measured, while a different variable, possibly words, may have been a more important factor with the Hawthorne texts. Due to this, the next factor to look at is how important word variation may be for determining the change in readability.

The next factor that may explain this variation in readability is type token ratio. Type token ratio determines the variation in vocabulary of the text by taking the total number of unique words divided by the total number of words (or tokens) in a text (University of Minnesota). A type token ratio, or TTR, close to 1 means that a text uses a high variation in language, while a type token ratio close to 0 means that a text uses a very low variation in language. Investigating the TTR of the authors' texts may show why the authors have different readability scores. The initial goal of looking at the TTR for each text was the belief that a higher ratio value would mean that the readability score would increase. Below is the TTR for each author's text:

Dickens Texts	**TTR**
A Christmas Carol	0.353
A Tale of Two Cities	0.342
Bleak House	0.335
David Copperfield	0.331
Oliver Twist	0.360

Hawthorne Texts	**TTR**
A Wonder Book	0.347
Doctor Grimshawe's Secret	0.363
House of the Seven Gables	0.400
The Blithedale Romance	0.397
The Scarlet Letter	0.389
Twice Told Tales	0.407

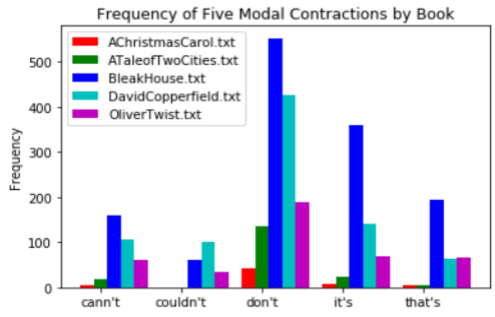
Though the TTR for Hawthorne texts tend to have larger values than the TTR for Dickens texts, this does not seem to confirm that the readability score will be higher for a higher TTR value when comparing individual texts. For example, Dickens' *A Christmas Carol* has a TTR of 0.353 while Hawthorne's *A Wonder Book* has a TTR of 0.347. If it were true that the text that has a higher TTR would have a higher readability score when comparing texts, then *A Christmas Carol* should have a higher reading score than *A Wonder Book*. If we compare reading scores, however, *A Christmas Carol* has an ARI score of 5.62, a CL score of 6.64, and an FK score of 5.67, while *A Wonder Book* has an ARI score of 9.76, a CL score of 7.93, and an FK score of 9.09. Since *A Wonder Book* has readability scores higher than *A Christmas Carol*, then the above statement is not true in this case.

The fact that there does not seem to be a clear relationship between type token ratio and readability score is not too surprising. Unlike readability, TTR does not consider sentence length

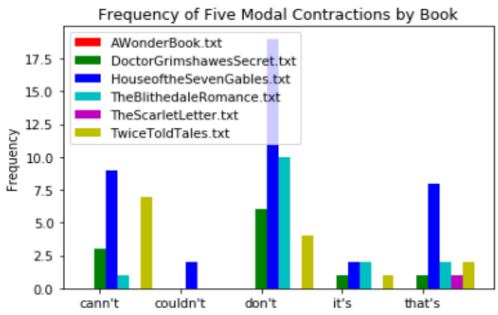
or word length. Instead, TTR looks for the total amount of unique words, no matter their length. In the case of *A Christmas Carol* and *A Wonder Book*, *A Christmas Carol* could have more unique words but have a small length, while *A Wonder Book* can have a little more of the same words but with a larger length and longer sentences. If this were the case, then *A Christmas Carol* would have a larger TTR value, but a smaller readability score compared to *A Wonder Book*. Also, the text length can influence the TTR for a given text. For example, *A Christmas Carol* contains 29,256 words, while *A Wonder Book* contains over 100,000 words. A text with a much larger text length is bound to repeat words much more often than a much shorter text, which can cause the shorter text length to have a higher TTR. Since word length may be a possible factor in the change in readability scores between Dickens and Hawthorne, we can look at the frequency of various uses of grammar, particularly contractions.

In English grammar, a contraction is the shortening (or combining) of two words to make one word. An example in grammar are the words “I’m” and “you’re”, which are contractions of “I am” and “you are” respectively. Authors can use contractions for a variety of reasons, including shortening their sentences as well as having text appear the way that people speak. Contractions can also improve the readability of a document both for the text and for the reader. Since contractions shorten words, this would affect the syntax of the text and change the readability compared to if an author chooses not to use contractions. The use of contractions can also assist the reader in reading the document, since contractions are used in every day speech and recognizable.

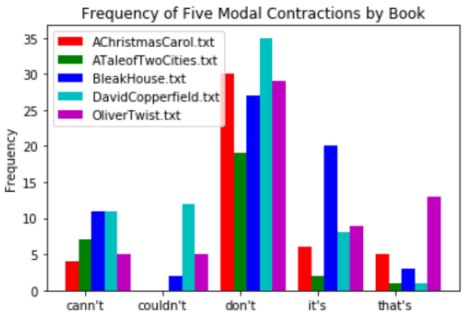
To compare the various contractions between texts, the frequencies of the five most common contractions used were plotted below:



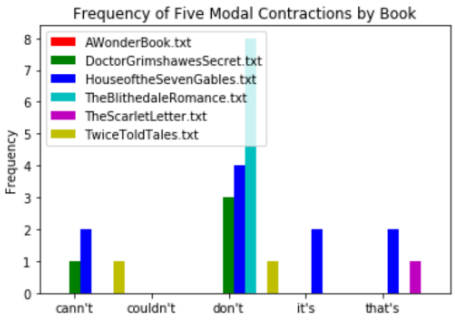
Full frequency of the contractions for each Dickens text



Full frequency of the contractions for each Hawthorne text



Frequency of the contractions of the first 30,000 words for each Dickens text



Frequency of the contractions of the first 30,000 words for each Hawthorne text

For the plots above, the five contractions used were cann't, couldn't, don't, it's, and that's. The contraction "cann't" is used over the modern spelling "can't" since cann't was the common contraction for "cannot" since "cann't" was the more common form to use in 19th century writings. From the frequency plots, Dickens tended to use contractions much more often than Hawthorne. It can also be seen that Dickens used contractions throughout his texts, noting that many words have a much higher word count in the top left graph (full text) compared to the bottom left graph (first 30,000 words to consider the texts with shorter length). Hawthorne, on the other hand, did not use contractions as much as Dickens, and barely used contractions with as much frequency later in his texts compared to Dickens.

Considering that the Hawthorne texts had higher readability scores than the Dickens texts, it's a possibility that the use (or lack) of contractions in a text can affect the readability score for a text. In order to see if contractions truly influenced the readability scores, a separate statistical analysis would need to be done. For the purpose of this paper, it can be concluded that contractions are a possible factor in affecting the readability scores of a text.

From the analysis above, the readability of Dickens and Hawthorne differ despite both authors coming from the same era, with Hawthorne having higher readability levels compared to Dickens. This paper attempted to answer what may be the causes of this change in readability among the authors, both as a comparison among each other and among their own works. After viewing three factors that may influence the readability scores for Dickens and Hawthorne texts, it seems that average sentence length and the use (or lack) of contractions may be factors in

determining a work's readability score. An analysis was also done on type token ratio, however, it was not clear what role, if any, TTR had on the change in readability. A text's total word count may be a factor in the lack of a relationship between TTR and readability, since much shorter texts may lead to more unique word counts compared to a much longer text length. From the three factors done in this paper, it appears that the average sentence length has the most influence in the change in readability, though further analysis is required to confirm. For future analysis, it would be beneficial to perform statistical analyses on the variables in the readability formulas to see which variable is the most important. It's possible that logistic regression would be able to determine if average sentence length, word count, character count, or syllables are the most influential variables to determine readability. Overall, this paper was able to answer the questions posed at the beginning, showing that Hawthorne and Dickens differ in readability and exhibiting a few factors that may explain this change.

References

- GutenTag Corpus Builder. https://gutentag.sdsu.edu/
- Newbold, Neil and Lee Gillam. "Populating a framework for Reability Analysis: Word frequency = word difficulty." *University of Surrey*, 2005, 12/10/2019. http://ucrel.lancs.ac.uk/publications/cl2009/318_FullPaper.pdf
- University of Minnesota. "Activity 4: Complexity in Oral vs. Written Language." *The Center for Advanced Research on Language Acquisition (CARLA)*, 2019, 12/11/2019. carla.umn.edu/learnerlanguage/spn/comp/activity4.html.
- Zhou, Shixiang, Heejin Jeong, and Paul A. Green. "How Consistent are the Best-Known Readability Equations in Estimating the Readability of Design Standards?" *University of Michigan. IEEE Transactions of Professional Communcation, Vol. 60, No.1.* March 2017. 12/11/2019. http://umich.edu/~driving/publications/ZhouHeejinGreenHowConsistReadability.pdf