# What Say You, Beowulf: Investigating Different Epochs of the English Language through Text Mining of Renowned Fictional Works

Claire Gellner
American University
Text Mining Final Paper

## Abstract

*English has changed significantly since its arrival in modern-day Britain in the 400s CE. By using text mining practices, such as key word and phrase analysis, this study explores how English changed between each distinct epoch and creates a snap-shot of the English language lexicon at each time period. The analysis reveals the beginning of a schism between Old and Middle contrasted with Modern English in the 1500s. Earlier Modern English retains some characteristics of the Middle English lexicon that have been shed by the 1800s.*

**Keywords:** English, historical linguistics, text mining

## 1. Introduction

The English language is an Indo-European language in the West Germanic language family, originating in what is now Germany that is closely related to Frisian, German, and Dutch (Crystal & Potter, 2018). English arrived in modern-day Britain in the year 410 C.E. with the invasion of the Goths (Nordquist, 2019). English has taken on a three distinct forms over the past 1,500 years. Much of the earlier forms of English are no longer recognizable to modern speakers.

The earliest distinctive epoch in English's history spans from the 5th century until the 11th Century and is known as the Old English period (Oxford International English Schools, 2018). At this time, English most resembled its Germanic cousins. Separated from its original milieu, English underwent a unique set of linguistic changes and was open to a different outside influences than its linguistic cousins.

English changed radically after the Norman Invasion of Britain brought French and Latin influences. This period from the 12th until the 16th century (Oxford International English Schools, 2018) is known as the Middle English period (Oxford International English Schools, 2018). Middle English adopted new words, sounds, and grammar from French and Latin (Crystal & Potter, 2018), not found in Old English.

The English Renaissance ushered in the beginning of Modern English. This period begins in the 17th century, and continues today (Oxford International English Schools, 2018). Inventions such as the printing press led to language standardization and wider literacy. This coupled with the industrial revolution, and increased contact other nations and languages through colonialism, war, and diplomacy, as well as trade shaped English into the language spoken today (Oxford International English Schools, 2018).

A question of interest here is the distinctions between these different epochs in the history of English; what distinguishes them? Text mining offers a way to highlight the lexical differences between the epochs.

## 2. Literature Review

"Stylometry" is the practice of using statistical analysis to highlight the differences between written works of different genres or authors (Laramée, 2018). If the statistical analysis of text consists of counting the word frequencies, stylometry is using those statistical features to make a claim. Many stylometry studies have used statistical and text mining analyses to analyze famous works and identify authorship. A quick way to find high-frequency features and words is to present the data in a matrix with each word, letter, type, or feature as a column and each instance as a row (Jockers & Underwood, 2016). Then, topic modeling, or similar methods, is commonly used (Jockers & Underwood, 2016) to group similar texts. Ultimately, this can then be fed into a model, such as PCR or K Nearest Neighbors to predict and identify authorship.

Similar techniques have been used to analyze text for distinctive features and attribute authorship. Çerkezi et al., 2013 begin with a statistically based method. This analysis searched for the frequency of words and n-grams in the works of Oscar Wilde and

Dino Buzzati (Çerkezi et al., 2013). This was then used on to show the similarities and differences in the styles of the two authors and compared to the translated versions of their works.

Similarly, Sallis and S. Shanmuganathan, 2008 analyzed the letters of Mary Queen of Scots; their goal was to group letters together letters from different points in the queen's life. The first step was to find the word and letter frequencies and length, followed by sentence complexity (Sallis and S. Shanmuganathan). These frequencies were then used for clustering similar letters for similar style and topic through k-means clustering, allowing stylometric and statistical features to be analyzed together (Sallis and S. Shanmuganathan, 2008). Sallis and S. Shanmuganathan, 2008 also use the statistical and clustering analysis for further study on letters of contested authorship.

For a stylometry study, the statistical analysis is a first step in a stylometric analysis. Miranda and Martín, 2012 were able to demonstrate the authorship of the *Federalist Papers* for works without a known author. Miranda and Martín, 2012 first created a list of high frequency functional words used in the *Federalist Papers*, both of known and unknown authorship. For each word, they then calculate the delta as developed by Maciej Eder (Eder, 2020). Delta is defined as the measure of stylistic difference, therefore authorship, that is attributed to words frequently used by a given author (Eder, 2020). This delta score shows that a given word was used by a specific author more frequently and thus can be used to identify the most likely author of the *Federalist Papers* with unknown authorship (Miranda García & Calle Martín, 2012).

This method is shared with Feinerer, 2008. In this analysis, a term document matrix from a corpus of the *Wizard of Oz* series was created. Then, a Principal Component Analysis model was attuned to the features and word frequencies in of L. Frank Baum's and Ruth Plumly Thompson's writing using books in the series with uncontested authorship Feinerer, 2008). This model was able to predict the authorship of later works in the series as not having been written by Frank L. Baum by the writing style features (Feinerer, 2008).

Reviewing historical works in the history of English to find the most frequently used words and phrases will give a kind of snap-shot view of English at each epoch, and how it has changed over time

## 2. Purpose

This study is a comparative statistical analysis of the lexicon at each distinct epoch (Old, Middle, Modern) and establishes the representative features of English at the time period using literary works: *Beowulf* for Old English, Chaucer's *The Canterbury Tales* for Middle English, Shakespeare's *Hamlet* for Early Modern English, and Jane Austen's works for Modern English. This study will compare word choice (word frequency) as well as phrases (n-grams) often used in each work or by each author, therefore during each epoch and how these may have changed over time.

## 3. Data Collection

The data for this analysis came from two sources: Project Gutenberg and the janeastenr package. Project Gutenberg is a repository of digitized books in the public domain. Similarly, the janeaustenr package is composed of digitized versions of all of Jane Austen's novels taken from Project Gutenberg and formatted as a package for RStudio.

One edition of each text was selected. For *Beowulf*, the version translated by Lesslie Hall was used (Anonymous, 1892). The version of *The Canterbury Tales* used was edited by David Laing Purves (Chaucer & Purves, 2000). For the Shakespeare data, only *Hamlet* was used, as it is one of Shakespeare's longer stand-alone plays (Shakespeare, 2022). The version used in this analysis was edited by Dianne Bean (Shakespeare, 1998).

The digitized books from Project Gutenberg were loaded into R as tibbles using a mirror. These books required extra data cleaning to isolate as much of the text in question as possible by removing any introductions, preface, translator notes, table of content or any other front-matter or end-matter. These steps ensured the data set was primarily comprised of the original text without much modern text added in. Empty rows were also removed.

The data imported from Project Gutenberg consisted of two variables, the text itself and the Gutenberg ID. The ID was removed, and columns were inserted to indicate the author and the epoch.

The Jane Austen data set originally contained variables for the specific novel, line number, and chapter. To match format of the Gutenberg books, these columns were removed, and the author and epoch columns were added.

All the books were then combined into one large dataset. The final data set consisted of 91179 rows with three columns: text, author, and epoch. This organization makes it possible to see which words have been commonly used in English throughout its history, as well as within each era or by each author.

The final data preparation step was to remove stop words. For the analysis custom stop words were added to the stop_words dictionary available in the dplyr package for R. The custom stop words included the character names as well as stage-related words from *Hamlet*. Had these not been removed, they would have inaccurately been labeled the most common words due to their overuse in the specific literature chosen.

## 5. Methodology

For this analysis, each of the questions was broken down for easy analysis. The question of the most commonly used words is split in two: What are the most commonly used words by each author, and what are the most commonly used words in English across all epochs. Both questions were investigated through a frequency analysis. Two versions of the compiled dataset were used, one with the author and epoch columns, and one without. The version without author or epoch was used for the "all time" analysis. In both cases, the analysis was conducted through a simple count of distinct words using tf, where the word count is the number of times the word appears out of the total.

This brought up an additional question: Would tf_idf yield different results. Because tf_idf calculates the word frequency with regards to the context and decreases the weight of importance for more common and less distinct words, it was possible that this would yield different frequencies. Thus, the analysis of most common words was repeated using the bind_tf_idf function which automatically calculates tf_idf. There was variation between tf and tf_idf; different words were counted amongst the 10 most common when using one metric or the other. There was overlap, with some words appearing in both sets. For consistency, tf_idf was used for this analysis.

The question of which phrases were most common was similarly split out into multiple questions that could be analyzed through n-grams. The first: What are the most common two-word phrases (bigrams) used. The second: What are the most common three-word-phrases (trigrams). Just as with the word-frequency analysis, this analysis was done once with n-grams grouped by author, and then repeated without author or epoch indicated to analyze the most common phrases across time. To conduct the n-gram analysis, the text was tokenized in pairs for bigrams and by triplets for trigrams.

From the word and phrase frequency analyses, it was evident that there existed a schism between Old and Modern English that was not as salient as the barrier between Old and Middle, or Middle and Modern English. An analysis to determine the nature of the schism was conducted to answer the question of if words could be attributed to a specific period.

The schism analysis used a version of the dataset with only epoch indicated, and author removed. The analysis had two distinct parts. The first was calculating the ratio of high-use words in each epoch to assess the likelihood that a given word was used in a specific epoch. This showed the correlation between the word and the epoch by indicating which specific words had higher odds of being used in one epoch over another. Three separate plots were made so that each epoch combination (Old + Middle, Middle + Modern, Modern + Old) could be compared. One epoch was the x-axis, and the other the y-axis. This created a visual representation of words that were more likely to be shared between epochs, vs fitting squarely into usage during one era. The second component of this analysis was a correlation analysis between all three epochs.

## 7. Limitations

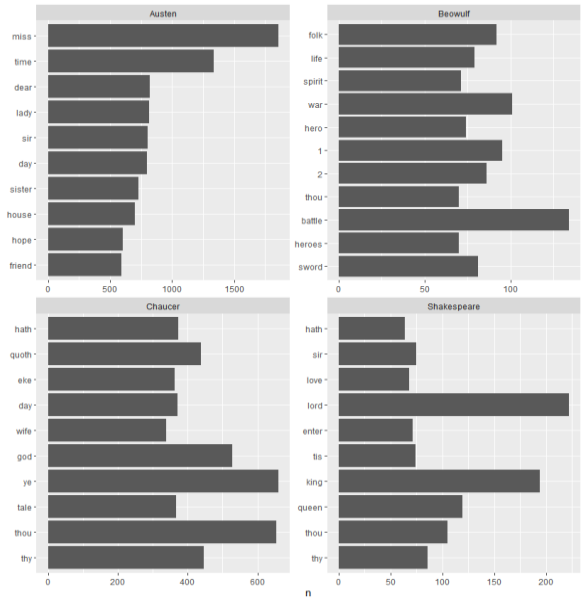
The primary limitation in this study is the availability of seminal texts for each era. Few works as pristine as *Beowulf* exist from the earlier time periods in the history of English. This lead to a disproportionate representation of Modern English when compared with the earlier epochs.

In a similar vein, only one author from each epoch was selected for this study. This causes the unintentional obfuscation of common stylistic and lexical features of the era because a given author simply does not use them. Sampling works of multiple authors from each epoch would alleviate both unequal representation as well as providing a less biased view of the epoch.

## 8. Findings

The first question investigated was which words were most used by each author, as well as in English overall. The results of the word frequency analysis provide a valuable glimpse of English at each stage in its development from Old English to Modern English, revealing many archaic and no longer used words being quite common parlance in the time of *Beowulf* and *The Canterbury Tale*s.

**Figure 1a. Most Common Word by Author – tf**

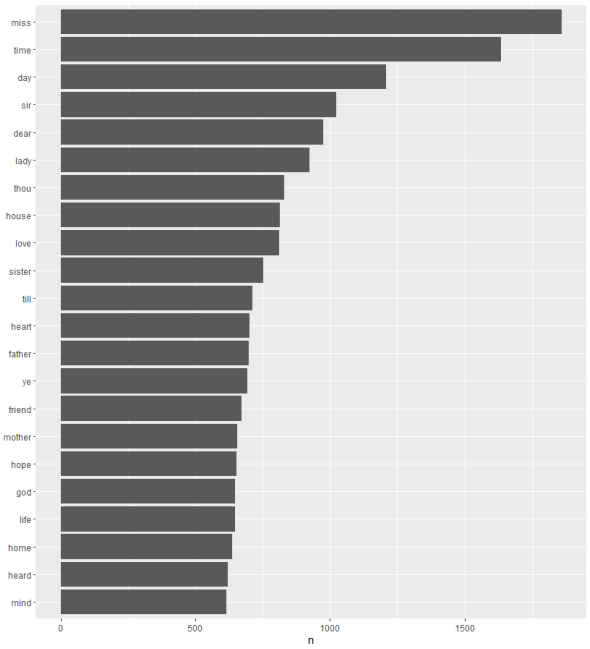



The most common word in *Beowulf* is the surprisingly modern "folk", while the most frequently used word in by both Shakespeare and Chauther is "hath"[1]. Unexpectedly, when using tf, the most frequently used words in Beowulf are modern with only "thou" being an archaic form. Chaucer has many more unfamiliar high frequency words; "eke", a word meaning "also" (Common Chaucerian Words, n.d.), both "thou" and "thy", and the no longer used conjugation "quoth".

Shakespeare's most frequent words fall somewhere in the middle between the archaic forms found in Chaucer, and the purely modern English of Jane Austen. By far, the most frequent word in Austen's combined works is "miss", with over 2000 instances, followed by "time" and "dear". *Hamlet* boasts a mix of modern and archaic words with "tis", "thou", "thy" and "hath" appearing beside "love", "enter" and "sir", a word that also ranks highly in usage by Austen.

The analysis was then repeated without distinguishing the author.

**Figure 1b. Most Common Word No Author – tf**




The analysis of word frequency across all time using tf is dominated by recognizable words: "miss" and "time", for example, with few words being archaic forms; only "thou" and "ye". One possible explanation for this may be the amount of modern English text dwarfing the amount of text from other epochs.

The word frequency analysis was repeated using tf_idf. There was found to be variation between the two metrics. This was expected as tf_idf weighs the word frequency in comparison to the words to reflect the relative importance of a word in the document, and not simply over the total number of words as with tf.

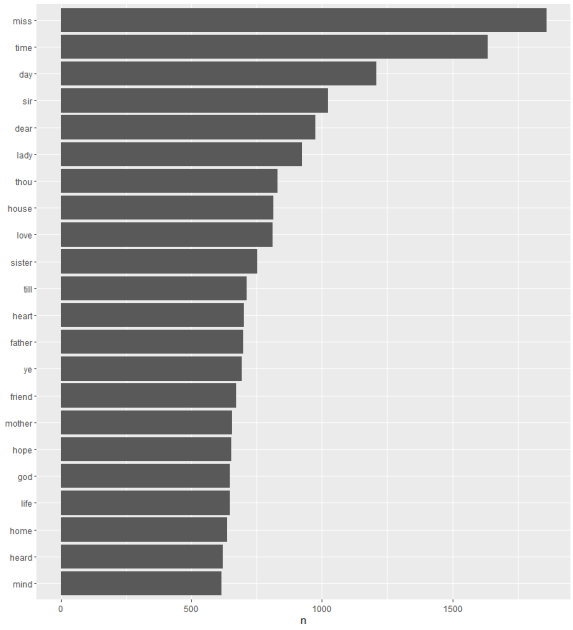
[1] The archaic present tense third-person singular of verb to have (Merriam-Webster, *hath*).

**Figure 2a. Most Common Word by Author – tf_idf**


In this version of the most frequently used words, *Beowulf* seems much more archaic with terms such as "neath"[2], and "atheling"[3]. Shakespeare, however, seems slightly more modern with frequent words including "I'll", "clown", "that's" and "ghost".

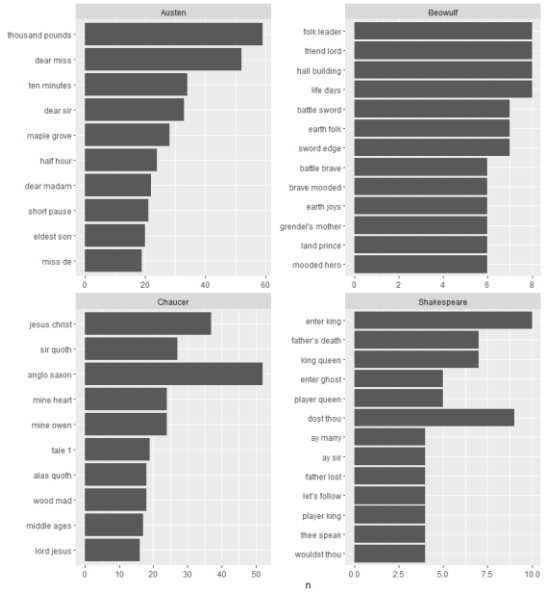
However, when the most frequent words without author was repeated using tf_idf, there was no difference with the list made using tf.

**Figure 2b. Most Common Word No Author – tf_idf**




The second question concerned the most common two- and three-word phrases used by each author and in English overall. The result of the bigram analysis is interesting and expected.

**Figure 3a. Most Common Bigrams By Author**



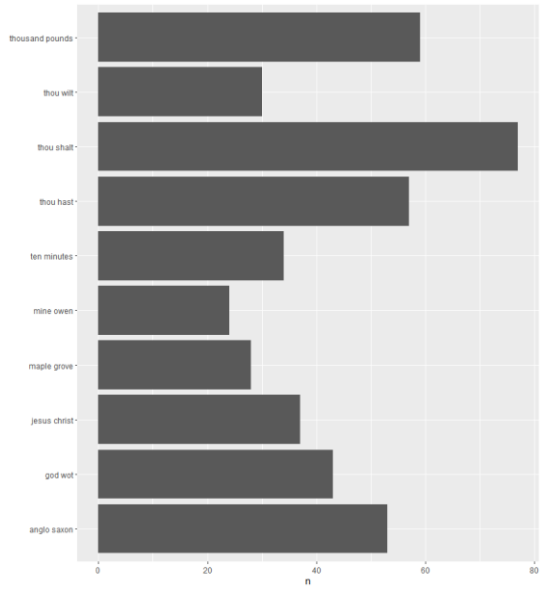

Here again, the word pairings most common for each author reflect the time with the frequent word pairs in *Beowulf* being odd collocations of modernly understood words that would not be used today such as "friend lord", "battle sword", "earth folk", or, "brave mooded". The most common word pairs used by Chaucer were more archaic, including "sir quoth", "mine heart", "mine owen", and "wood mad". These two works also reflect the attitudes of the time periods. Word pairs showcase the language of battle in *Beowulf* through "sword edge" and "battle brave", as well as the deep piety of Chaucer's time with common pairs referring to Jesus.

Surprisingly, the common word pairs in Shakespeare were not as modern as the most common words, reminiscent of the findings of the tf analysis. Instead, Shakespearian bigrams echo the archaic language of earlier time periods. The most frequent word pairs in *Hamlet* were found to include "dost thou", "ay marry", "thee speak", and "wouldst thou". By contrast, the most common pairs of words in Austen's works were found to be very modern and recognizable including: "thousand pounds", "dear miss", "ten minutes", "short time", and "half hour". The bigrams found in Austen speak to the propriety of the period and the amount of time Austen's characters spend writing and reading letters.

[2] Short for beneath. (Merriam-Webster. (n.d.). Neath)

[3] An anglo-saxon term for a prince or nobleman (Merriam-Webster. (n.d.). Atheling)

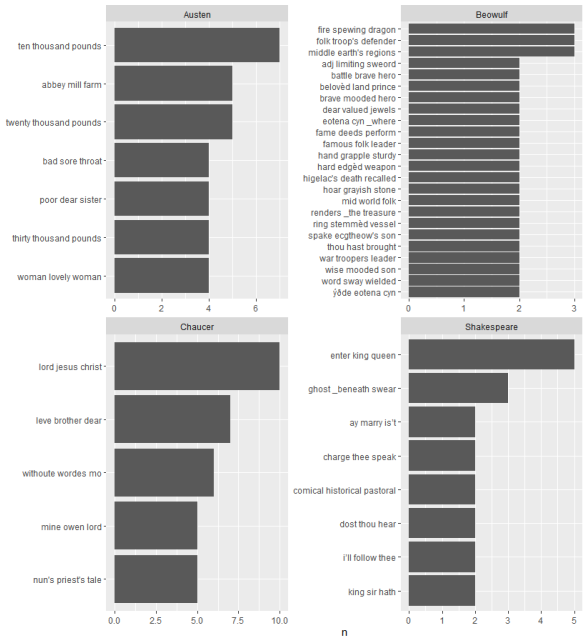
**Figure 3b. Most Common Bigrams No Author**




When the analysis was repeated without denoting author or epoch, the most common word pairings in English were found to be a mix of words mostly from Austen and Chaucer. This indicates that Austen is not, in fact, dominating the analysis despite comprising most of the source text, a result of using tf_idf for bigram analysis.

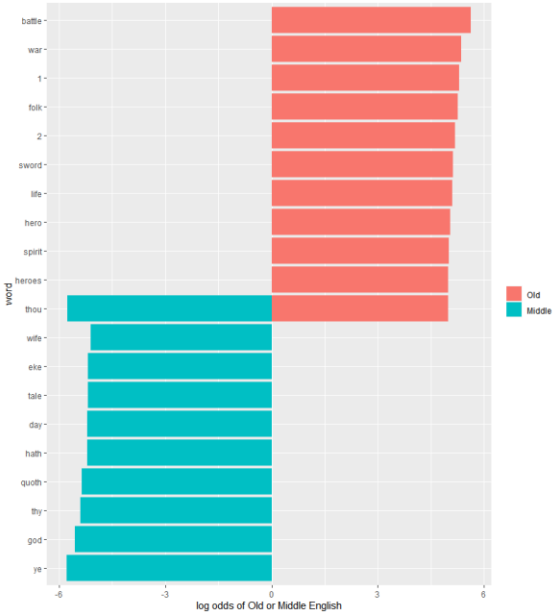
This analysis was repeated with three-word phrases, known as trigrams. The results of the trigram analysis are consistent with those of the bigram analysis. However, there are very few trigrams, and most had a frequency under 10 total uses, and thus were not as indicative features of the epoch. However, for Beowulf, the trigrams showcase more archaic language than any of the other analysis.

**Figure 4. Trigrams by Author**




The third question posed by this analysis looked to investigate the schism and porousness of the distinction between the three epochs in the history of English. The result of the schism analysis highlighted the links between Old and Middle English, and Middle and Modern English, and the distance between Old and Modern English.

**Figure 5. Log Odds of Words by Epoch**



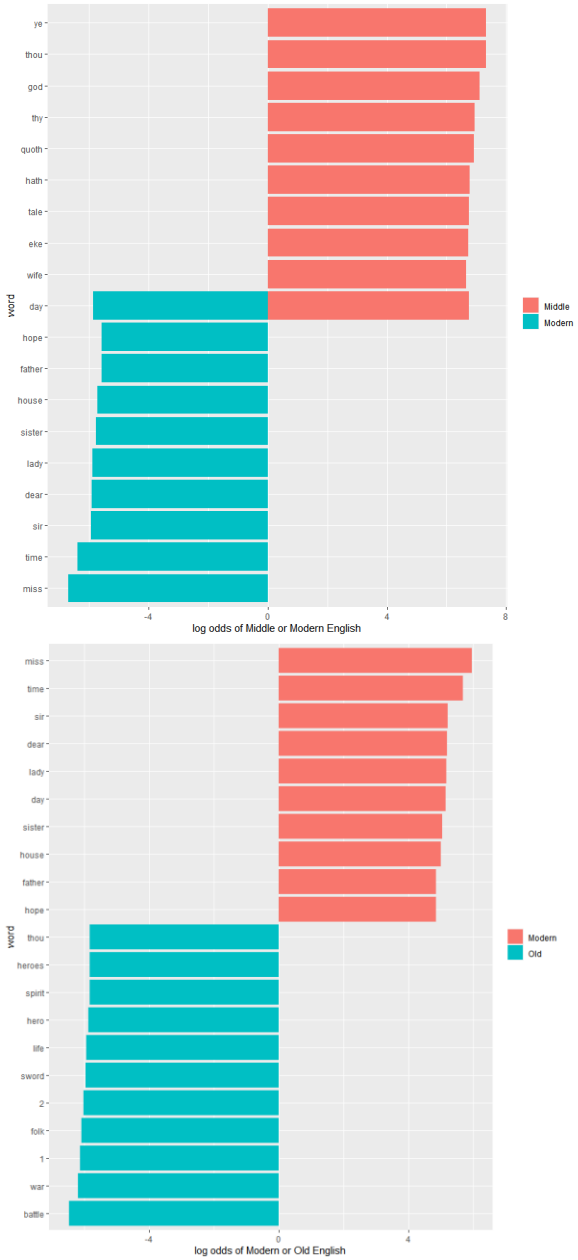
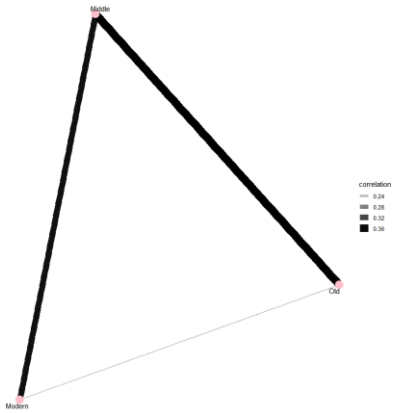



*Figure 5* shows the log odds of a word appearing in the writing of a specific epoch. In the combination of Old and Middle English, "hath" is a word that could appear in both epochs. This was previously seen in the word frequency analysis as well. Similarly, between Middle and Modern English, "day" is indicated as being of high frequency in both epochs. Unsurprisingly, Old and Modern English do not have any overlapping words. One note of interest is that the log odds of the Modern English words appearing in Modern English specifically when compared with Old English are smaller than expected. This could indicate the division between the epochs is less salient going back in time. Modern words are more likely to have previously appeared in earlier texts, while words that have died out are unlikely to appear in Modern texts.

The final step in the schism analysis was to look at the correlation between the three epochs.

**Figure 6. Correlation between Epochs**




Here it is evident that there is a strong link between Old and Middle English, and a slightly weaker link between Middle and Modern English, with the link between Old and Modern English being very weak.

## 9. Discussion

This analysis analyzed the lexicons of the three major epochs of English through exemplary works from each period. It was established that Old English and Modern English are most distinct from each other, with Middle English having features of both. Middle English, and to a lesser extent Old English, can be characterized by archaic words and constructions. Some of these are still used at the very beginning of the Modern English period but fall out of use as the epoch continues. Modern English and Middle English tend to see an overlap in words, however the words in Middle English can be opaque due to non-standard spellings.

The purpose of this study is a stylistic analysis and not authorship attribution. Further study could be done to determine if a given work can be correctly attributed to an epoch by using the features and frequent words identified here as the markers of the epoch.

Similar analyses of historical documents have also done a letter frequency analysis, though this was not done here. An analysis to track the evolution of spelling in specific words across each epoch of English may be an interesting topic for further analysis as well.

## 10. Conclusion

Reviewing the different analyses presented here brings a blurry picture of the English language and its history into view. One immediate observation is the distinct archaic lexicon found in *Beowulf* and *The Canterbury Tales,* and occasionally *Hamlet*, especially when contrasted with the easily recognizable wordage of Austen. This split between Middle and Modern English may be more salient than that between Old and Middle English, as evidence by the schism analysis where Old and Middle English were most correlated. However, it is important to note that while many of Chaucer's words are not immediately recognizable due to their spelling, a number of the words Chaucer uses are still in use, such as "woulde" and "saide" for "would" and "said". Chaucerian English may be closer to Modern English even if it does not appear so at first glance.

The distinction between Old, Middle, and Modern English becomes more complex when considering the Shakespearian bigrams and the most frequently used words in *Beowulf*. Despite being considered the beginning of Modern English, Shakespearean bigrams are peppered with archaic forms such as in "dost thou", "thee speak", and "wouldst thou", while Austenian bigrams are not. This clear difference is unexpected. The approximately 250 years between Shakespeare and Austen marks the distinct shift in the English language where it takes on the more modern form still spoken today. This provides evidence for a pattern of linguistic change whereby words that collocate together fall out of use together; meaning language change happens in chunks. This claim is further bolstered by the trigrams which showcased the most archaic language in *Beowulf*. It is possible that these words and constructions fell out of usage around the same time while other words continued to be used in more modern times.

What is more perplexing is that the most frequently used words in Beowulf from the tf analysis, as well as the bigrams are recognizable and, in some cases, indistinguishable from modern English. Phrases like "sword edge" and "folk leader" would be right at home in any modern fantasy novel. One possible explanation is that words with high enough frequency to remain in use will be quite old. These words have remained high frequency throughout the 1500-year span and are not in danger of falling out of use. The evidence for this possibility is strengthened by the results of the word frequency and n-gram analysis without author.

Another surprising finding is the quotidian nature of the high-frequency words. The most common words are those used to address the family and friends around us. Prominently on the list of most common words across all epochs are "sister", "mother", "father", and "friend. To these important people, we tell the little joys, fears, and sorrows of everyday life. Both "love" and "hope" make the list, as did "heart", "home", and "mind". It is unlikely these words were contributed solely by Austen given that the bigrams with the highest counts also showcase much of Chaucer, indicating that when the author is unknown, Austen does not dominate. Perhaps, it is that we have not changed so much since the time of *Beowulf*, after all. The greatest joy in life remains sharing news, hopes, fears, and dreams with those we love.

## 11. References


Anonymous. (1892). *Beowulf: An Anglo-Saxon Epic Poem*. In L. Hall (Trans.), *Gutenberg.org*. D.C. HEATH & CO., PUBLISHERS. https://www.gutenberg.org/files/16328/16328-h/16328-h.htm

Austen, J. (1994). Emma. In *Project Gutenberg*. https://www.gutenberg.org/ebooks/158

Austen, J. (1994). Mansfield Park. In *Project Gutenberg*. https://www.gutenberg.org/ebooks/141

Austen, J. (1994). Northanger Abbey. In *Project Gutenberg*. https://www.gutenberg.org/ebooks/121

Austen, J. (1994). Persuasion. In *Project Gutenberg*. https://www.gutenberg.org/ebooks/105

Austen, J. (1998). Pride and Prejudice. In *Project Gutenberg*. Project Gutenberg. https://www.gutenberg.org/ebooks/1342

Austen, J. (1994). Sense and Sensibility. In *Project Gutenberg*. https://www.gutenberg.org/ebooks/161

Çerkezi, E., Muka, Q., Çelo, E., & Dumi, A. (2013). Impact of New Entry Strategic Technology on Frequency Words Analysis in Translation, Literature and Inter-cultural Communication. *Procedia - Social and Behavioral Sciences*, 99, 196–205. https://doi.org/10.1016/j.sbspro.2013.10.486

Chaucer, G., & Purves, D. L. (2000). *The Project Gutenberg eBook of The Canterbury Tales, and Other Poems, by Geoffrey Chaucer and David Laing Purves* (D. O'Danachair, Ed.). *Project Gutenberg*. https://www.gutenberg.org/cache/epub/2383/pg2383.html

*Common Chaucerian Words*. (n.d.). Chaucer.fas.harvard.edu. Retrieved October 15, 2022, from https://chaucer.fas.harvard.edu/common-chaucerian-words

Crystal, D., & Potter, S. (2018). English language | Origin, History, & Characteristics. In *Encyclopædia Britannica*. https://www.britannica.com/topic/English-language

Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." *Digital Scholarship in*

*the Humanities*, vol. 30, no. 2, 14 Nov. 2013, pp. 167–182, 10.1093/llc/fqt066. Accessed 7 Sept. 2020.

Feinerer, I. (2008). An Introduction to Text Mining in R [Review of *An Introduction to Text Mining in R*]. *R News*, *8*(2), 19–22. ISSN 1609-3631.

Jockers, M. L., & Underwood, T. (2016). Text-Mining the Humanities [Review of *Text-Mining the Humanities*]. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities* (pp. 350–368). John Wiley & Sons, Incorporated. https://ebookcentral.proquest.com/lib/aul/reader.action?docID=4093339&ppg=350

Laramée, F. D. (2018). Introduction to stylometry with Python. *Programming Historian*. https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python

Merriam-Webster. (n.d.). Atheling. In *Merriam-Webster.com dictionary*. Retrieved October 14, 2022, from https://www.merriam-webster.com/dictionary/atheling

Merriam-Webster. (n.d.). Hath. In *Merriam-Webster.com dictionary*. Retrieved October 15, 2022, from https://www.merriam-webster.com/dictionary/hath

Merriam-Webster. (n.d.). Neath. In *Merriam-Webster.com dictionary*. Retrieved November 20, 2022, from https://www.merriam-webster.com/dictionary/neath

Miranda García, A., & Calle Martín, J. (2012). Testing Delta on the "Disputed Federalist Papers." International Journal of English Studies, 12(2), 133. https://doi.org/10.6018/ijes/2012/2/161791

Nordquist, R. (2019, July 19). *Key Events in the History of the English Language* [Review of *Key Events in the History of the English Language*]. ThoughtCo. https://www.thoughtco.com/events-history-of-the-english-language-1692746

Oxford International English Schools. (2018, March 4). *A Brief History of the English Language - Oxford International English Schools*. Oxford International English Schools. https://www.oxfordinternationalenglish.com/a-brief-history-of-the-english-language/

Shakespeare, W. (2022.). *Play Lengths*. Www.playshakespeare.com. https://www.playshakespeare.com/study/play-lengths

P. Sallis and S. Shanmuganathan, "A Blended Text Mining Method for Authorship Authentication Analysis," 2008 *Second Asia International Conference on Modelling & Simulation* (AMS), 2008, pp. 451-456, doi: 10.1109/AMS.2008.99.

Shakespeare, W. (1998). *The Project Gutenberg eBook of Hamlet, by William Shakespeare* (D. Bean, Ed.). Project Gutenberg. https://www.gutenberg.org/files/1524/1524-h/1524-h.htm