Linguistic Patterns in Early Modern English: An Exploratory Analysis of Letter Frequency in Shakespeare's First Folio*

Vowels Are More Commonly Used Than Consonants

Daisy Huo

March 30, 2024

This paper conducted linguistic frequency analysis in Early Modern English by investigating the distribution of vowels and consonants in First Folio, the first collection of William Shakespeare's 36 plays. Downloading from the Project Gutenberg collection and applying the methodology of the Poisson regression model, we would count the number of words along with the number of times vowels appear in the first 30 lines of each play. This analysis revealed a consistent distribution of vowels, highlighting the discovery of trends in Early Modern English language use and a vital impact on linguistic acquisition. These findings matter as word frequency counts are utilized in multidisciplinary research to study the origins and evolution of English literature.

1 Introduction

William Shakespeare is widely regarded as one of the most renowned English playwrights and poets in the history of English literature. His masterpieces hold profound significance in the context of the transition from Old English to Early Modern English, which began to develop in the 16th century. The First Folio, published in 1623, is the first collected and printed edition of Shakespeare's comedies, histories, and tragedies. Moreover, Oxford professor Emma Smith (2023) spoke highly of The First Folio as a "trophy" and that half of Shakespeare's well-known plays would have been lost without the preservation of this book. In the realm of linguistics, such analysis of historical texts would enhance our understanding of letter frequency and English phonetics across the Elizabethan and Jacobean eras during the Renaissance and Reformation (Wheeler 2018).

*Code and data are available at: https://github.com/dai929/Linguistics_First_Folio.git

Linguists process and predict human language, relying on the frequency with which a letter of the alphabet occurs on average in a written literature piece. Among the 5 vowels and 21 consonants, WordsRated investigated over 172000 English words and concluded that the most five common letters by frequency are “E”, “S”, “I”, “A”, and “R” respectively on a descending order (Talbot 2023). That is to say, vowels are more commonly used in words than consonants in Modern English. Indeed, while a substantial number of research has been conducted on how main themes and key elements including love, power, politics and free will greatly influenced theatres in Shakespeare’s time, the study of his writing style remains notably absent.

To address this gap, we analyze the letter frequency patterns during the Early Modern English period in The First Folio, aiming to visualize and model the distribution of vowels and compare the results with a similar analysis in Modern English. The estimand of our interest is the number of vowels in the first 30 lines of Shakespeare’s 36 plays. The text of The 36 plays in The First Folio was separately acquired and downloaded from the Project Gutenberg collection (Hart 2019), a public library containing over 70000 freely accessible eBooks, and further converted into a dataset. Then through the measurement of our estimand, we found that the number of vowels in Shakespeare’s words increases as the number of words totalled increases, suggesting a positive correlation between vowels and word counts. The findings further emphasized the essence of vowels acting as a keystone and the majority volume in written literature even during the Elizabethan and Jacobean periods. With the absence of vowels, consonants on their own are incapable of forming a syllable or a word. By recognizing the importance of vowels, linguists, phoneticians and educators can proceed to study the letter frequency under different authors, literary periods and genres.

The remainder of this paper is structured as follows. Section Section 2 discusses the raw and cleaned dataset and all variables of interest, along with the visual presentations of relationships between the variables. Section Section 3 includes the Poisson regression model used to analyze the correlation between the number of vowels in the line and the number of words in the line. Section Section 4 presents model summary statistics from the last section. Section Section 5 contains the main findings, a few takeaways that we learnt about the world, shortcomings and some possible future research directions.

2 Data

3 Model

4 Results

5 Discussion

References

Hart, Michael. 2019. "Project Gutenberg." *Project Gutenberg*. https://www.gutenberg.org/.

Smith, Emma. 2023. "Follow the Money: The Story of Slavery and Shakespeare's First Folio." *The Guardian*. https://www.theguardian.com/books/2023/apr/21/slavery-and-shakespeares-first-folio?CMP=share_btn_url.

Talbot, Dean. 2023. "Letter Frequency in English – WordsRated." *WordsRated*. https://wordsrated.com/letter-frequency-in-english/.

Wheeler, Kip. 2018. "LibGuides: English Literature: Literary Periods & Genres." *Libguides.com*. https://mc.libguides.com/eng/literaryperiods.