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

April 11, 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 35 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 10 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 subject 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 10 lines of Shakespeare's 35 plays. The text of The 35 plays in The First Folio was 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 2 discusses the raw and cleaned dataset and all variables of interest, along with the visual presentations of relationships between the variables. 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. **?@sec-result** presents model summary statistics from the last 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

2.1 Introduction to Dataset

The dataset utilized in this paper is Shakespeare's First Folio by William Shakespeare (Shakespeare 2000), downloaded from Project Gutenberg, an online library that allows free access to a vast collection of public domain books. Data was acquired, cleaned, and analyzed using the statistical programming software R (R Core Team 2023), with additional support from R packages `arrow` (Richardson et al. 2023), `dataverse` (Kuriwaki, Beasley, and Leeper 2023), `gutenbergr` (Johnston and Robinson 2023), `here` (Müller 2020), `janitor` (Firke 2023), `knitr` (Xie 2014), `marginaleffects` (Arel-Bundock 2024), `modelsummary` (Arel-Bundock 2022), `readr` (Wickham, Hester, and Bryan 2024), `rstanarm` (Brilleman et al. 2018), `tidybayes` (Kay 2023), and `tidyverse` (Wickham et al. 2019).

Choosing Shakespeare's First Folio over other literary compositions holds significance due to its importance and authority in Early Modern English. In particular, William Shakespeare, one of the greatest playwrights of that time, had also been known as a masterful wordsmith by contributing thousands of words and phrases to the vocabulary and enriching the newly developed form of the English language. By analyzing a diverse and representative collection of his plays, we are able to conclude certain language usage and linguistic patterns during the Renaissance.

2.2 Distribution and Properties of Individual Variables

To begin with, after sketching the expected dataset and our model and simulating a dataset of how the number of vowels could be distributed following the Poission distribution, we are now on our way to acquire and clean our data. We download the full text of Shakespeare's First Folio from the Project Gutenberg using `gutenberg_download()` from `gutenbergr` (Johnston and Robinson 2023). Then to avoid overly imposing on the Project Gutenberg servers, we will proceed on with our local copy.

After that, note that we are interested in only the lines that have content and the lines that are part of a play. In this case, we remove all empty lines that are provided just for spacing and remove all lines before the first play "The Tempest". Next, remove the title of each play and the concluding lines that start with "FINIS", and we are able to create counts of the number of vowels, A/E/I/O/U/a/e/i/o/u, in that line and for the first ten lines of each play.

Table 1 provides a breakdown of the counts and frequency of individual vowel usage (A/a, E/e, I/i, O/o, U/u) by each play. From the table, we are able to observe that the total count of "E/e"s is generally the highest among the vowels, indicating that this vowel is more likely to be the most frequently used vowel in Shakespeare's First Folio. On the other hand, the total count of "U/u"s is the lowest among the five vowels.

Table 1: Counts of Aa/Ee/Ii/Oo/Uu's, by play, in Shakespeare's First Folio

play	total_count_a	total_count_e	total_count_i	total_count_o	total_count_u
A Midsommer Nights Dreame	22	41	28	25	16
All's Well, that Ends Well	22	29	23	21	10
As you Like it	33	52	16	30	8
Loues Labour's lost	26	44	20	16	15
Measvre, For Measure	16	39	16	21	19
Much adoe about Nothing	20	45	23	21	8
The Comedie of Errors	29	31	14	26	15
The Famous History of the Life of King Henry the Eight	23	52	16	28	12
The First Part of Henry the Fourth	24	40	19	23	2
The Life of Henry the Fift	26	40	17	24	11

Table 2: Counts of vowels, by play, in Shakespeare's First Folio

play	total_word_count	total_count_vowel
A Midsommer Nights Dreame	72	132
All's Well, that Ends Well	60	105
As you Like it	82	139
Loues Labour's lost	66	121
Measvre, For Measure	65	111
Much adoe about Nothing	65	117
The Comedie of Errors	65	115
The Famous History of the Life of King Henry the Eight	82	131
The First Part of Henry the Fourth	69	108
The Life of Henry the Fift	71	118
The Life of Timon of Athens	57	96
The Merchant of Venice	76	120
The Merry Wiues of Windsor	63	115
The Second Part of Henry the Fourth	65	109
The Taming of the Shrew	74	133

Table 2 then provides summary statistics for the total word count and the total count of vowels for each play. From this data, the distribution of the total word count and total count of vowels across plays seem to follow a clear pattern and a direct correlation, as plays with a higher word count generally have a larger number of vowels and vice versa.

2.3 Relationships between Variables

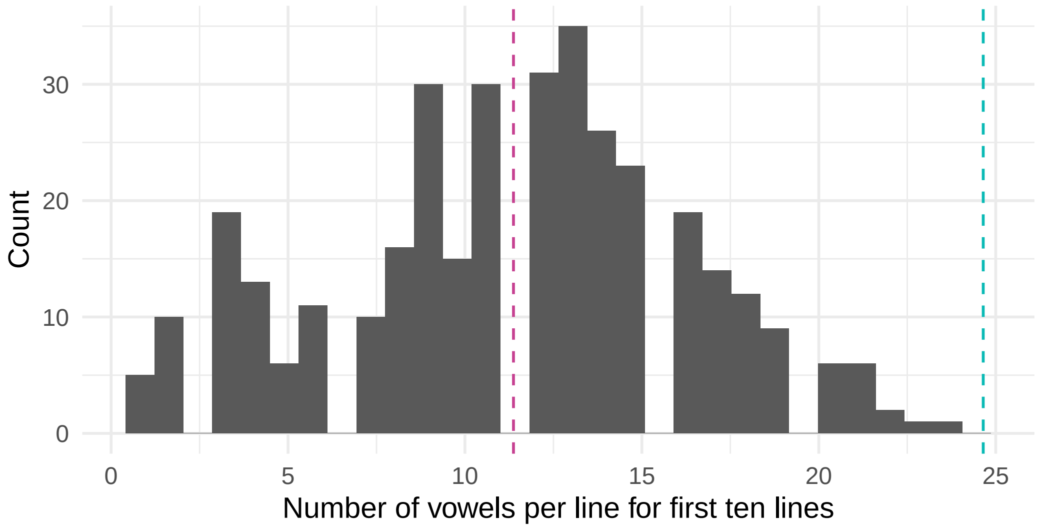

Figure 1: Distribution of the number of vowels

Figure 1 displays a histogram that represents the distribution of the number of vowels per line. The red dashed line represents the average number of the count of vowels (11.4), and the blue dashed line represents the spread of the count of vowels (24.6). A large distance between the mean and the variance suggests that the data points are not mainly clustered around the average value. This implies that certain lines have a much higher or lower count of vowels compared to the average, and these outliers contribute to the greater variability in this dataset. As we can tell from Figure 1, the tallest bars appear to be in the range of 10 to 15 vowels, suggesting that most lines contain a count of vowels within this range.

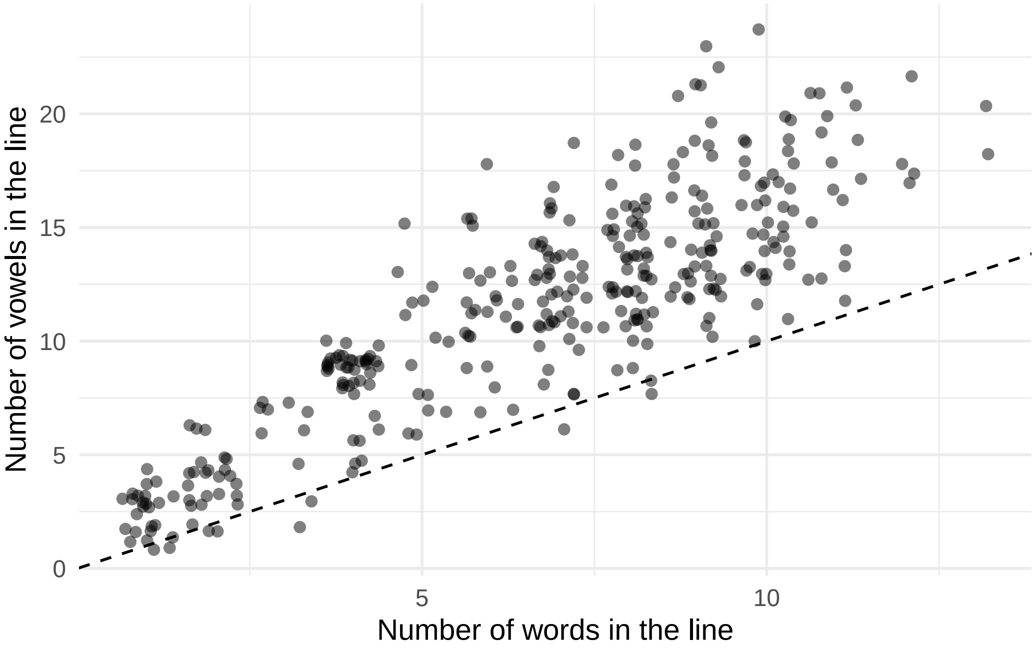

Figure 2: Comparison of the number of vowels in the line and the number of words in the line

Figure 2 displays a scatter plot, which contains numerous data points corresponding to the number of vowels and the number of words in a line. We specifically include the diagonal line to help with our understanding about the data. Suppose that the data were on the $y = x$ line, then on average there would be one vowel per word. Given the mass of point above this diagonal line expect that on average there is more than one vowel per word. However, the density of the points seems to be higher at the lower end of both axes, indicating that there are more lines with fewer words and fewer vowels. While the number of words increases, the spread of vowels also increases, implying a greater variability in the count of vowels for lines with a higher word count. This aligns with our observations in Figure 1.

3 Model

The model that we are interested in is:

$$y_i|\lambda_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \beta_0 + \beta_1 \times Number\,of\,words_i$$

$$\beta_0 \sim Normal(0, 2.5)$$

$$\beta_1 \sim Normal(0, 2.5)$$

where y_i is the number of vowels in the line and the explanatory variable is the number of words in the line. We can build this poisson regression model using `stan_glm()` from `rstanarm` (Brilleman et al. 2018).

4 Results

We now use `plot_cap()` from `marginaleffects` (Arel-Bundock 2024) to show the number of vowels predicted by the model, for each line, based on the number of words in that line.

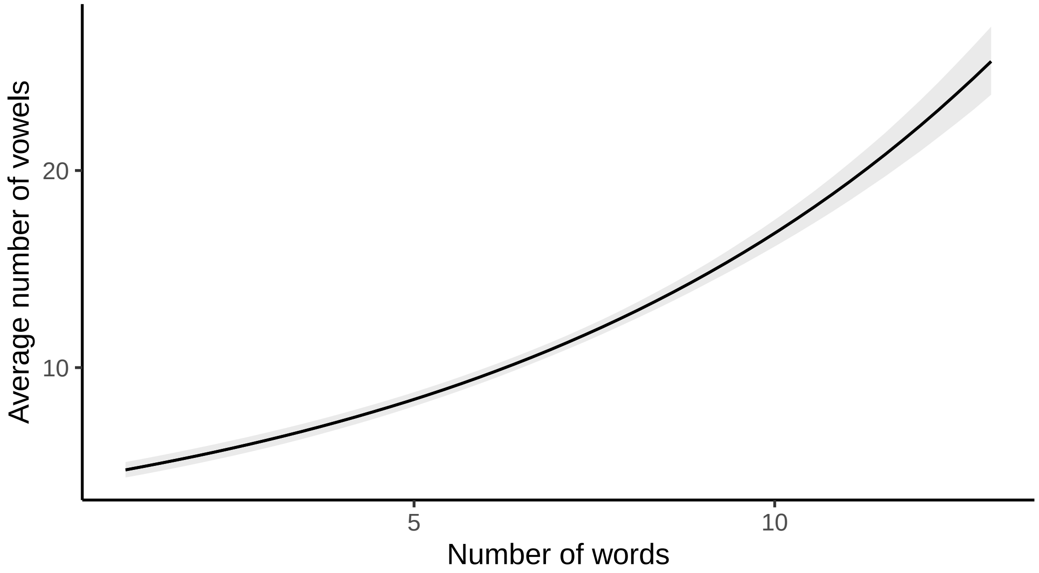

Figure 3: The predicted number of vowels in each line based on the number of words

Table 3: Whether the number of vowels is positively related with word count in Shakespeare's First Folio

	Number of vowels
(Intercept)	1.432
word_count	0.139
Num.Obs.	350
Log.Lik.	−855.126
ELPD	−856.7
ELPD s.e.	9.3
LOOIC	1713.5
LOOIC s.e.	18.6
WAIC	1713.5
RMSE	2.81

Figure 3 suggests that as the number of words in a line increases, the average number of vowels also increases, and the rate of this increase accelerates as the number of words grows. Therefore, it clearly demonstrates that our expected relationship is positive.

Finally, Table 3, the model summary, summarizes the results of this Poisson regression analysis, which is used to model count of vowels. The positive coefficient of estimates (0.139) for our predictor variable "word_count" further suggests that there is a positive relationship between the number of vowels and the word count in the selected lines.

5 Discussion

5.1 An Overview of Linguistic Patterns in Early Modern English

In this paper, we thoroughly conducted an exploratory analysis of linguistic patterns and letter frequency by exploring the distribution of vowels in Shakespeare's First Folio, marking the first collection of his legendary 35 plays. By employing the methodology of the Poisson regression model, we investigated the relationship between our response variable, the number of vowels, and our explanatory variable, the word count in the first 10 lines of each play. Our analysis revealed several important findings in English language usage during the Elizabethan and Jacobean eras.

5.2 Vowels Are More Commonly Used Than Consonants: Consistency from Early Modern English to Modern English

One notable takeaway from our analysis is the confirmation of the prevalence of vowels within Early Modern English literary compositions. Our research revealed a consistent predominance of vowels, particularly the letter "E/e," in Shakespeare's First Folio. This discovery aligns with the findings from a previous study conducted by WordsRated, which emphasized the significance of the letter "E/e" as the most frequently used letter in Modern English. By establishing this connection, we have contributed additional evidence supporting the dominance of vowels in literary works during Shakespeare's era. This finding necessarily enhanced our understanding of linguistic trends and highlights the unshakeable importance of vowels in English written literature regardless of different historical periods.

5.3 Correlation between Letter Frequency and Word Count in Shakespeare's Works

Furthermore, another key takeaway in this paper highlights the deliberate intention of Shakespeare's use of vowels in his writing, as evident in the increase in vowel usage frequency with the rise in word count. This further emphasizes the importance of vowels not just as a fundamental element of language, but also as a tool for creating depth and complexity in linguistic expression. Moreover, Figure 3 demonstrates how this increasing rise in vowel usage may be linked to the syntactic dynamics of Shakespeare's plays, suggesting a strategic use of both consonants and vowels to enhance the overall fluenceness and effectiveness of Shakespeare's writing style. In light of this connection between vowels and words, it can be argued that a better understanding of vowel distribution patterns can greatly enhance our appreciation of Shakespearean literature.

5.4 Limitations and Future Directions in Research on Early Modern English Literature

Despite the valuable insights gained from this paper, it is important to acknowledge its limitations. One such limitation is that our research was confined to analyzing Shakespeare's First Folio, which may not fully represent the linguistic trends of the Early Modern English period. Thus, further investigation could enhance the depth of our findings by incorporating a wider range of literary works from this era. Additionally, the examination of other linguistic features such as consonant distribution and word length could provide a more comprehensive understanding of language usage in Shakespeare's time. There are a number of directions that future research may pursue. For instance, researchers may choose to explore the evolution of English language over time by comparing linguistic patterns in Shakespeare's plays with those found in contemporary literature.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

———. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Kay, Matthew. 2023. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Shakespeare, William. 2000. *Shakespeare's First Folio. Project Gutenberg*. https://www.gutenberg.org/ebooks/2270.

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.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Im-*

plementing Reproducible Computational Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.