Analysing Villette by Charlotte Brontë*

Using the Frequency of A's

Aryaman Suri

March 17, 2024

This paper examines the distribution of the letter A in Villette by Charlotte Brontë to determine if its frequency correlates with the number of words in the text. We discovered that when the text's word count rises, the letter A appears more frequently. Linguists may find this analysis useful since it may be used to recognize patterns and traits in the English language, gain a deeper comprehension of Charlotte Brontë's writing style, and track the development of the English language over time.

Table of contents

*Code and data are available at: https://github.com/aryamansuri/Letters-used-in-Villette.git

1 Introduction

Letter frequency analysis is an important approach in literary analysis for identifying underlying patterns and nuances in texts. Notably, as the quantity of words increases, individual letter frequencies become more important, providing insights on the authors writing style, linguistic preferences, and thematic patterns. Given that a is used so frequently in the English language and can express nuanced subtleties of tone and meaning, its frequency stands out among these letters as a very interesting focal point.

Comprehending the correlation between the frequency of the letter a and the total word count is highly relevant to computational text analysis and literary studies. This analysis was performed on Charlotte Brontës Villette, a literary classic known for its intricate plot and deep characterizations. Variations in the frequency of 'a' within longer paragraphs can indicate changes in the narrative focus, stylistic flourishes, or thematic emphasis of Villette , which can enhance our comprehension of Brontës goals and the literary landscape. We use data from the first ten lines of each chapter and see if there is a relationship between the number of 'A' s and the number of words using a Poisson regression in R (R Core Team 2022) and directions from (Alexander 2023).

Furthermore, by using these analyses, researchers can identify underlying patterns and trends that might not be immediately noticeable when using only standard close reading techniques. Scholars can reveal stylistic decisions, thematic threads woven throughout the text, and hidden layers of meaning by examining the subtleties of a frequency across different word counts in Villette. This expanded comprehension not only helps us appreciate Villette more fully but also adds to a larger literary conversation by providing new insights and directions for future research in the area. Consequently, the examination of a frequency analysis is a useful instrument in the toolbox of the literary analyst, enabling more complex readings and encouraging a more in-depth interaction with texts from various genres and eras.

2 Data

In this study, we drew upon data from Villette, written by Charlotte Brontë, sourced from the Gutenberg Project repository (Johnston and Robinson 2023), to uncover potential insights into the relationship between the frequency of the letter A and the number of words in each line using R(R Core Team 2022). In doing so, we shed light on linguistic patterns within the text. The Gutenberg Project(Johnston and Robinson 2023) is a digital library that provides unrestricted access to public domain literary works like books and plays. More details about the specific R packages and processes used are included below.

2.1 Data Collection

As mentioned above all of our raw data was aqcuired from the book Villette by Charlotte Brontë using the Gutenberg library(Johnston and Robinson 2023) in R(R Core Team 2022). Our initial dataset had over 21000 rows of text with columns of row number that corresponded to the line in the book and text on that line. This included chapter names, blanks lines and data that was not usable.

2.2 Data Cleaning

Next, the raw data was then cleaned for our analysis. This included getting rid of blank lines, adding columns for Chapter numbers and line numbers and the number of As(not case-sensitive) using data manipulation from Dplyr (Wickham et al. 2023), stringr (Wickham 2022) in the tidyverse (Wickham et al. 2019) package, in each line in addition with the number of total words in each line. This step is vital for our study since its main purpose is to analyse the relationship between the number of As and the number of words.

2.3 Data Analysis

After we have our filtered dataset, it is ready for use for analysis. We analysed this dataset.

```
                                                                    Text
1                                                                BRETTON.
2 My godmother lived in a handsome house in the clean and ancient town of
3 Bretton. Her husband's family had been residents there for generations,
4      and bore, indeed, the name of their birthplace-Bretton of Bretton:
5      whether by coincidence, or because some remote ancestor had been a
6             personage of sufficient importance to leave his name to his
  Chapter Count of A Word Count
1       1          0          1
2       1          5         14
3       1          4         11
4       1          3         11
5       1          4         11
6       1          4         10
```

We seperated each chapter by identifying the text and the roman numerals at the start of each chapter using Stringr (Wickham 2022) and filtered our data into categories for each chapter and the first 10 lines of each chapter. We now had the filtered dataset for the first ten lines of the 42 chapters in the book Villette.

3 Model

The goal of our model is to predict if the the number of As increases as the number of words increases. We will use the Bayesian Poisson model for this. The Bayesian Poisson model is a statistical tool that helps us make sense of count data, like the number of occurrences of an event within a specific time frame.

3.1 Model set-up

Define y_i as the number of As in the line and the explanatory variable is the number of words in the line.

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{1}$$
$$\log(\lambda_i) = \beta_0 + \beta_1 + \text{Number of words}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

We run the model in R (R Core Team 2022) using the rstanarm package of Goodrich et al. (2022). We use the default priors from rstanarm.

3.1.1 Model justification

The use of the Poisson model is justified for a multiple number of reasons. First, the Poisson distribution is commonly employed in analyzing count data, making it suitable for the analysis of the occurrences of As within the text. Second, the model assumes independence between events, which aligns with the assumption that the occurrences of As in different parts of the text are independent of each other. Overall, the Poisson model provides a robust framework for investigating the relationship between the A frequency and word count in our analysis.

4 Results

First we can visualize our cleaned dataset in Figure 1 and Figure 2 . We can already see that as the number of As in the line increases, the number of words in the line also increase. This is a good sign that our model is well fitting.

Our results are summarized in Table 1 using Model summaries in R(Arel-Bundock 2022). As seen in the table, we see a positive correlation between the frequency of the As and the word count. This means that the frequency of As increase by 10% or 0.1 for an additional word

Table 1: Explanatory models of Number of A's based on number of words

	Number of A's
(Intercept)	0.22
	(0.11)
word_count	0.10
	(0.01)
Num.Obs.	420
Log.Lik.	−815.984
ELPD	−817.8
ELPD s.e.	11.8
LOOIC	1635.6
LOOIC s.e.	23.7
WAIC	1635.6
RMSE	1.76

in the line. We are also provided information on the log-likelihood, which is -815.984. The negative log-likelihood in Poisson regression serves as a measure of how well the model fits the observed data. As it decreases, it indicates that the model is better at predicting the observed counts.

5 Discussion

In this paper, we undertook an exploratory analysis of the frequency of the letter 'A' in Charlotte Brontë's Villette, aiming to understand its relationship with the number of words per line. Through systematic examination and quantitative modeling, we sought to uncover potential patterns and correlations that shed light on the text's linguistic structure and stylistic tendencies.

Our analysis revealed intriguing insights into the relationship between 'A' frequency and word count in Villette. Specifically, we found a positive correlation between the frequency of 'A's and the number of words per line, suggesting that longer lines tend to contain a higher occurrence of the letter 'A'.

Furthermore, our study uncovered variations in this relationship across different sections of the text, indicating potential shifts in writing style, thematic emphasis, or narrative tone throughout Villette. These findings also offer insights into broader trends in classic literature and textual analysis methodologies.

However, it's important to acknowledge the limitations of our approach. Our analysis focuses solely on the frequency of the letter 'A' and its relationship with word count, overlooking

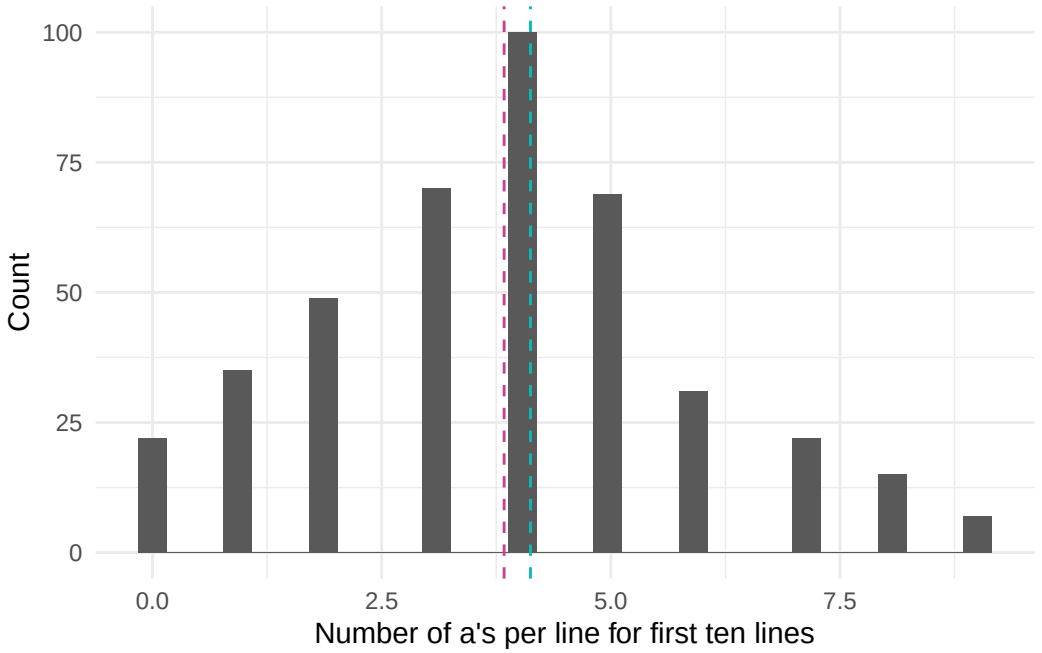

Figure 1: Distribution of the number of As

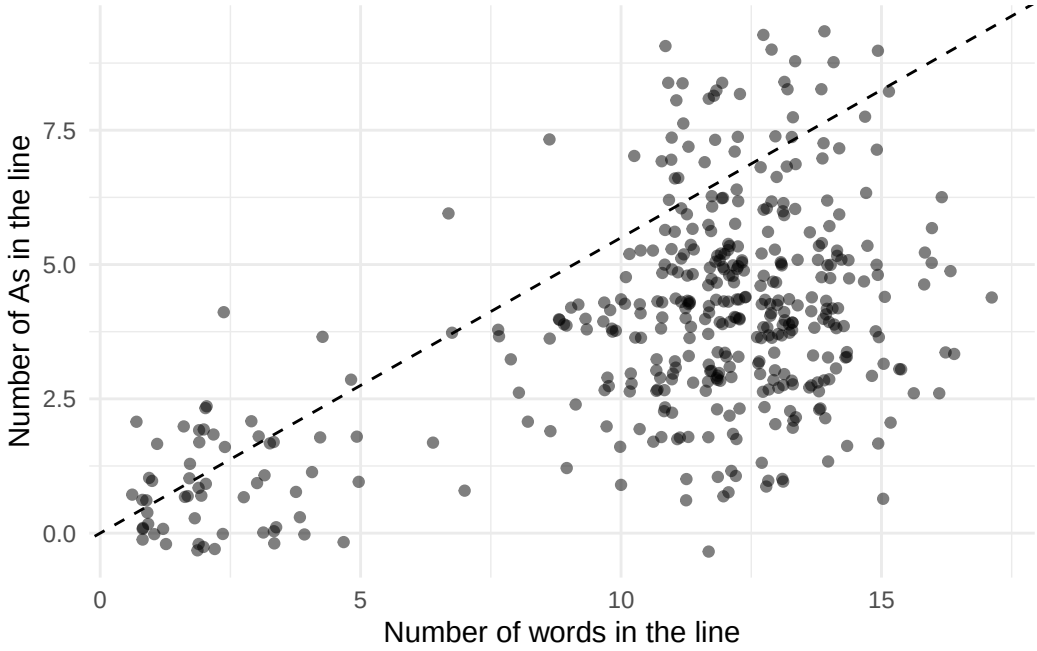

Figure 2: Compare the number of As per line and the number of words per line

potential interactions with other linguistic features or contextual factors. Additionally, our study does not delve into authorial intent or the broader socio-historical context in which Villette was written, which could offer deeper insights into the observed patterns.

Moving forward, future research could adopt a more comprehensive approach, integrating qualitative analyses, such as close reading or authorial intent, with quantitative modeling techniques. This interdisciplinary approach would provide a richer understanding of the stylistic choices and thematic motifs present in Villette, offering new avenues for exploring the complexities of Brontë's narrative craft. Furthermore, extending this analysis to other literary works and genres could facilitate comparative studies and illuminate broader trends in textual dynamics across different authors and periods.

6 Conclusion

In conclusion, our analysis of the frequency of the letter 'A' in Charlotte Brontë's Villette has provided valuable insights into the text's linguistic structure and stylistic tendencies. Through quantitative modeling and systematic examination, we have uncovered a positive correlation between 'A' frequency and word count, suggesting nuanced patterns in Brontë's prose. These findings not only contribute to our understanding of Villette's narrative dynamics but also offer broader implications for literary analysis methodologies and the study of Victorian-era literature. Moving forward, further interdisciplinary research integrating qualitative and quantitative approaches will continue to enrich our understanding of textual dynamics and deepen our appreciation of literary craftsmanship.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r and Python*. Chapman; Hall/CRC.

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.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Johnston, Myfanwy, and David Robinson. 2023. "Gutenbergr: Download and Process Public Domain Works from Project Gutenberg." https://docs.ropensci.org/gutenbergr/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2022. "Stringr: Simple, Consistent Wrappers for Common String Operations." https://stringr.tidyverse.org.

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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://dplyr.tidyverse.org.