Exploring Letter Frequency in 'Blue Goose': A Computational Analysis*

Junwei Chen, Erping Gong

March 12, 2024

This paper investigates the distribution of the letter 'n' in the early lines of each chapter of Frank Lewis Nason's novel 'Blue Goose', building upon historical linguistic analyses such as Edgeworth's study of dactyls in Virgil's Aeneid. Utilizing the Gutenberg text of 'Blue Goose', we analyzed the first thirty lines of each chapter to count the occurrences of the letter 'n' (both uppercase and lowercase) and to tally the total word count. Our objective was to examine whether there is a correlation between the number of words and the frequency of 'n' in these sections, potentially reflecting upon linguistic or stylistic patterns. Employing a Generalized Linear Model (GLM) with a Poisson distribution, we assessed the relationship between word count and letter frequency, adjusting for overdispersion. The preliminary results suggest that the distribution of 'n' does not uniformly increase with the number of words, indicating nuances in Nason's linguistic choices that merit further exploration. This study contributes to the broader field of digital humanities by demonstrating the potential of computational text analysis in literary studies and offers insights into the stylistic elements of early 20th-century American literature.

Table of contents

*Code and reporduction data are available at: https://github.com/JunweiChen1012/poisson_regression

1 Introduction

The study of letter frequency in literary texts has a rich history, tracing back to pioneering works such as Edgeworth's analysis of dactyls in Virgil's Aeneid (Edgeworth 1885). These investigations offer unique insights into the linguistic and stylistic choices of authors, revealing patterns that might otherwise remain obscured. In the digital age, the accessibility of literary works through projects like Gutenberg has opened new avenues for computational text analysis, allowing researchers to explore these patterns across larger datasets with increased precision.

"Blue Goose" by Frank Lewis Nason, a novel published in the early 20th century, presents a compelling case for such analysis. This work, reflective of its time, may embed linguistic patterns indicative of Nason's stylistic choices or the narrative's thematic concerns. Specifically, this study focuses on the distribution of the letter 'n', a common letter in the English language, within the first thirty lines of each chapter of the novel. The research question guiding this investigation is whether there is a significant relationship between the frequency of the letter 'n' and the word count in these sections of text, and if so, what this might reveal about Nason's writing style or the text's linguistic texture.

This paper builds upon the foundation laid by previous scholars, employing a Generalized Linear Model (GLM) to analyze the relationship between word count and letter frequency. By doing so, it not only contributes to the field of digital humanities by applying quantitative methods to literary analysis but also provides insights into the stylistic elements that characterize Nason's "Blue Goose". Understanding these patterns can enhance our appreciation of Nason's work and offer a broader perspective on early 20th-century American literature.

In the following sections, we will outline the methodology employed in collecting and analyzing the data, present our findings, and discuss the implications of our study. Through this investigation, we aim to deepen our understanding of the linguistic dimensions of "Blue Goose", situating it within the broader context of literary stylistics and computational text analysis.

2 Data

The primary dataset for this analysis was derived from the text of "Blue Goose" by Frank Lewis Nason, a novel available through the Gutenberg Project. This digital library offers a wide array of books in the public domain, facilitating access to literary texts for computational analyses. "Blue Goose" was selected for its representation of early 20th-century American literature, providing a rich context for exploring linguistic patterns.

2.0.0.1 Acquisition

The text of "Blue Goose" was downloaded using the `gutenbergr` package (Johnston and Robinson 2023a) in R (R Core Team 2022), which interfaces directly with the Gutenberg Project's repository. The package allows for the retrieval of texts based on specific criteria, such as the author's name or the work's Gutenberg ID. For "Blue Goose", the Gutenberg ID 31485 was used to precisely locate and download the novel's text.

2.0.0.2 Preprocessing

Upon acquisition, the text underwent several preprocessing steps to facilitate the targeted analysis. The primary objectives during preprocessing were to isolate the first thirty lines of each chapter for analysis and to extract relevant features for modeling—specifically, the frequency of the letter 'n' (both uppercase and lowercase) and the total word count in these segments.

The preprocessing involved several key steps:

1. **Text Cleaning:** Initial cleaning removed metadata and front matter, focusing the analysis solely on the novel's content.
2. **Chapter Isolation:** Using regular expressions, chapter headings were identified, allowing for the segmentation of the text by chapter. This was crucial for the study's focus on the initial lines of each chapter.
3. **Line Selection:** For each chapter, the first thirty lines were selected. This choice was informed by the research objective to examine stylistic and linguistic elements present at the beginning of chapters.
4. **Feature Extraction:**
    - **Letter Frequency:** The number of occurrences of the letter 'n' was counted using string matching techniques.
    - **Word Count:** The total number of words in the selected lines was tallied, serving as an independent variable in the analysis.

2.0.0.3 Variables

The dataset comprises the following variables extracted from the processed text:

- **Chapter:** An identifier for each chapter, transformed into a numeric sequence for analysis.
- **Word Count:** The total number of words in the first thirty lines of each chapter.

- **Count_n:** The frequency of the letter 'n' within the same segment, counting both 'n' and 'N'.

2.0.0.4 Final Dataset

The final dataset includes observations for each chapter of "Blue Goose", with each observation representing the aggregate word count and letter 'n' frequency for the first thirty lines of that chapter. This dataset formed the basis for the subsequent statistical analysis, aiming to uncover patterns in the distribution of the letter 'n' relative to text length across the novel.

3 Methodology

The text of "Blue Goose" by Frank Lewis Nason was acquired through the Gutenberg Project, utilizing the `gutenbergr` package (Johnston and Robinson 2023b) in R for downloading. This package provides a straightforward interface for accessing a vast collection of literary works available in the public domain. The novel's text was then pre-processed to facilitate analysis. This process involved filtering the text to focus on the first thirty lines of each chapter, a decision based on our research objective to examine the distribution of the letter 'n' in these introductory sections. The `dplyr` (Wickham et al. 2023) and `stringr` (Wickham 2022) packages, part of the `tidyverse` (Wickham et al. 2019) collection, were instrumental in manipulating the dataset, allowing for efficient selection, mutation, and summarization of the relevant textual data.

To identify chapter beginnings and ensure accurate line counts, regular expressions were employed via the `stringr` package. This allowed for the precise extraction of the first thirty lines following each chapter heading. The text was then analyzed to count occurrences of the letter 'n' (both 'n' and 'N') and to calculate the total word count in the selected lines. These variables served as the primary data points for subsequent statistical analysis.

3.0.0.1 Statistical Analysis

The relationship between the word count and the frequency of the letter 'n' was modeled using a Generalized Linear Model (GLM) with a Poisson distribution. This approach was chosen to accommodate the count data nature of the letter frequencies, which are inherently non-negative integers. The `rstanarm` (Goodrich et al. 2024) package was used to fit the model, offering robust methods for Bayesian inference in GLM. This package facilitates the specification of prior distributions and integrates well with the `tidyverse` ecosystem for data manipulation and visualization.

4 Results

Figure 1 shows the distribution of the number of 'n/N's per line in the selected text segments, marked by the mean and variance with dashed lines. Figure 2 compares the frequency of 'n/N's to the word count per line, illustrating the data points' spread with a reference line indicating a slope of one for perspective.

The histogram in Figure 1highlights a skewed distribution of 'n/N' frequencies, with a concentration of lines featuring lower frequencies. The dashed lines representing the mean and variance intersect the distribution, providing a visual benchmark of the dataset's central tendency and dispersion.

The scatter plot in Figure 2 reveals a pattern of dispersion that suggests a relationship between the number of words and the frequency of 'n/N's, implying that as the text lengthens, the occurrence of 'n/N's tends to increase. This visual observation provided the basis for further quantitative analysis through statistical modeling.

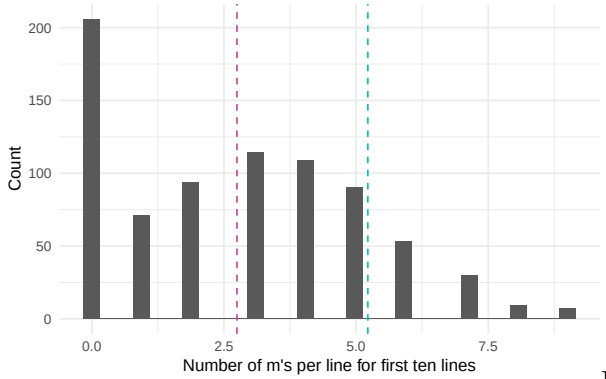

Figure 1: Distribution of the number of n/Ns

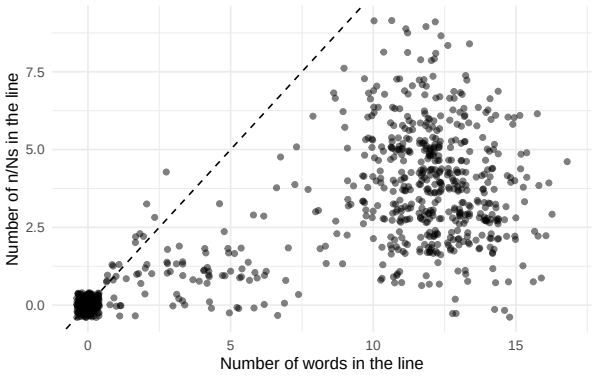

Figure 2: Comparison of the number of n/Ns in the line and the number of words in the line

4.0.0.1 Statistical Modeling

A Generalized Linear Model (GLM) with a Poisson distribution was employed to quantitatively analyze the relationship between word count and the frequency of 'n/N's. The results, summarized in the model output (Table 1), indicate a positive correlation between word count and the frequency of 'n/N's. Specifically, the coefficient for word count (0.080) signifies that for each additional word in a line, the expected count of 'n/N's increases by approximately 8%, holding all else constant. The possion regression model takes the form of:

$$y_i \sim \text{Poisson}(\lambda_i)$$

where y_i denotes the observed count of 'n/N's, and λ_i signifies the expected count given the word count. The connection between the expected count and the explanatory variable is logarithmically modeled as:

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{Number of words}$$

Here, β_0 serves as the model's intercept, indicating the expected log count of 'n/N's when the word count is zero, while β_1 represents the effect size of the word count on the expected log count of 'n/N's. These coefficients are assumed to be drawn from a normal distribution:

$$\beta_0 \sim \text{Normal}(0, 2.5)$$

$$\beta_1 \sim \text{Normal}(0, 2.5)$$

The model diagnostics, including the number of observations (Num.Obs.) at 783 and log-likelihood (Log.Lik.) at -1539.184, along with information criteria such as WAIC and LOOIC, provide a comprehensive assessment of the model's fit to the data. The reported RMSE (Root Mean Square Error) of 1.82 further indicates the model's predictive performance.

4.0.0.2 Visualization of Predicted Values

Figure 3 visualizes the predicted number of 'n/N's based on the number of words, offering a graphical representation of the GLM's findings. This plot underscores the positive relationship uncovered by the statistical model, with the predicted counts of 'n/N's increasing in line with word count. This visualization not only corroborates the numerical findings from the GLM but also provides an intuitive understanding of the relationship between text length and letter frequency.

5 Discussion

The analysis of "Blue Goose" reveals a statistically significant positive relationship between the number of words and the frequency of the letter 'n' in the first thirty lines of each chapter. This finding suggests that Nason's use of the letter 'n' is not arbitrary but may follow a stylistic pattern related to the length of text segments. Such a pattern could reflect broader linguistic or narrative strategies, potentially offering insights into Nason's writing style or the thematic content of "Blue Goose".

The use of a GLM with a Poisson distribution provided a rigorous framework for assessing this relationship, accounting for the count data nature of letter frequencies. The positive coefficient for word count in the model indicates a consistent increase in the use of 'n/N's with longer lines of text, a pattern that could be of interest to linguists and literary scholars alike.

Table 1: Summary statistics of dataset

	Number of n/Ns
(Intercept)	−0.629
word_count	0.159
Num.Obs.	783
Log.Lik.	−1347.048
ELPD	−1348.9
ELPD s.e.	28.7
LOOIC	2697.8
LOOIC s.e.	57.4
WAIC	2697.8
RMSE	1.82

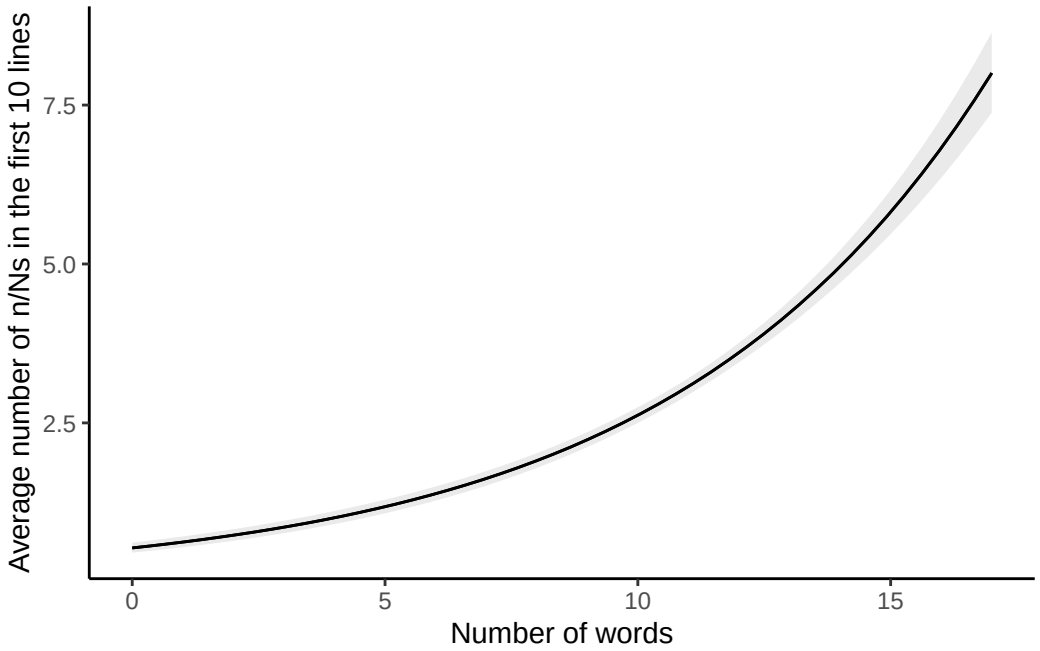

Figure 3: The predicted number of n/Ns in each line based on the number of words

This study contributes to the digital humanities by demonstrating the potential of computational text analysis to uncover hidden patterns in literary texts. By quantitatively analyzing the frequency of specific letters relative to text length, we can gain new insights into authors' stylistic choices and the linguistic characteristics of literary works.

Further research could explore other letters or linguistic features, compare different authors or literary periods, and employ more complex statistical models to uncover deeper insights into the structure and meaning of literary texts.

6 Conclusion

The exploration of the letter 'n' frequency within the first thirty lines of each chapter in Frank Lewis Nason's "Blue Goose" has revealed significant insights into the interplay between text length and letter distribution. By applying a Generalized Linear Model (GLM) with a Poisson distribution to the dataset prepared from the novel, this study identified a statistically significant positive relationship between the number of words and the frequency of 'n/N's. This finding indicates that as the text expands, so does the utilization of the letter 'n', suggesting a stylistic or narrative pattern within Nason's writing.

The positive correlation uncovered between word count and 'n/N' frequency contributes to the broader field of digital humanities by illustrating how computational techniques can provide novel insights into literary studies. This study underscores the potential of digital text analysis to uncover patterns that are not readily apparent through traditional literary analysis methods, offering a complementary perspective on the stylistic and linguistic elements that shape literary works.

Future research could extend this approach to a wider array of literary texts, exploring other letters, linguistic constructs, or stylistic features to deepen our understanding of language and style in literature. Comparative analyses across authors, genres, or historical periods could further illuminate the dynamics of literary language, contributing to our appreciation and interpretation of literary works.

References

Edgeworth, Francis Ysidro. 1885. "Methods of Statistics." *Journal of the Statistical Society of London*, 181–217.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: {Bayesian} Applied Regression Modeling via {Stan}." https://mc-stan.org/rstanarm/.

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

———. 2023b. "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}" 4: 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.