A Computational Investigation of Middle English Orthography

Nicholas Turk

April 26, 2024

1 Introduction

We do not speak the same language as 1300s English folk. Their English was structured differently, pronounced differently, and most relevant to this paper, it was written differently. Middle English is typically classified as the English dialects developed following the 1066 Norman Conquest of England, but before the 1500s, when printing became widespread. While we do have a good understanding of linguistic and orthographic trends during this time period, previous work has not leveraged computational methods for the purpose of investigating granular changes. This paper presents a thorough analysis of basic Middle English textual data, specifically observing the change in spelling and diction from 1150 to 1500.

2 Historical Context

The Norman Conquest of 1066 saw a massive change to the aristocracy within England. The Anglo-Saxon rulers who had been so influential to the development of Old English were ousted, and the Norman rulers took their place at the top of society. The Normans mostly spoke French, which became a language of prestige and culture in England. While the majority of people spoke English, the elite, and sometimes even the middle class, spoke French and Latin. The distinction between Old English and Middle English came not as a single massive shift, but as a slow drip-feed of French and Latin influence following the conquest [1].

Another major turning point in the history of the English language was the Black Death between 1348 and 1350. English had become more common in prestigious environments throughout the late 13th and early 14th centuries, but the big break came in part due to urbanization and economic upheaval caused by the massive demographic shifts that come with losing half the population. As living standards rose within England due to a lack of demand for resources leading to massive surplus that allowed the lower classes greater purchasing power. Because English was still mostly the language of the lower classes, the empowerment of these lower classes led to the wider adoption of English, and especially English writing, which became more standardized due to this institutionalization [2].

This is not to say that Middle English spelling was ever truly standardized. Far from it. Middle English spelling tended to capture very well the phonetics of its words, but there were often multiple valid ways to represent those pronounciations, with only scribal training and personal preference to guide the decision between those

variants [1]. These preferences do undergo interesting shifts, though, and it's worth remembering that scribes were often familiar with Latin and French, and thus would pull in spelling trends and etymologies from those languages to inform their English writing as well [3].

3 Methods

This study counted up the words, letters, and letter n-grams in a corpus of Middle English. Using the Helsinki Corpus of English Texts [4], we first isolated all texts in the appropriate time period for Middle English writing, that being from 1150-1500. There were six works in the corpus dating in the range of 1050-1150, but they were all in Old English, despite being in the correct time period for Middle English to have been in development. This left us with 141 remaining texts, distributed as shown in Table 1.

Table 1: Representation of the time periods our data covers

Period	1150-1250	1251-1350	1351-1420	1421-1500
# Texts	37	23	48	33

Having extracted the text from the Helsinki Corpus, we now performed simple operations to count the number of times a given word appeared, the number of times a given letter appeared, and the number of times a given n-gram of letters appeared. Words and letters are self-explanatory, but basically n-grams (in this context) are a slice of n letters from a given word. For instance, the 2-grams of the word "kynge"

are "ky", "yn", "ng", and "ge". We counted 2-grams, 3-grams, and 4-grams along with letters and words for each time period in our data.

We then graphed the 30 most frequent words, 2-grams, 3-grams, and 4-grams in bar plots representing their frequency in the time periods those graphs represent. We also created graphs for letter frequencies, but there are so few unique letters that there was no need to limit our inspection to only the thirty most used. By examining the frequency of these various items over time, we were able to determine trends in the spelling of Middle English across centuries.

4 Findings

Our findings start with an interesting, but unsurprising observation. Middle English word frequencies appear to conform to Zipf's Power Law, a statistical law which states that for some sequence of frequency data where frequencies $x_1 > x_2 > ... > x_n$, we have a rough formula where $x_r \propto \frac{C}{r^\alpha}$ where C and α are constants. In words, we can say that the frequency of a given word is inversely proportional to the ranking of its frequency in the sequence. It's well established that word frequencies tend toward Zipf's Law [5], but it's interesting that this holds even in systems where spelling isn't consistent. We can see Zipf's law in action within Figure 1.

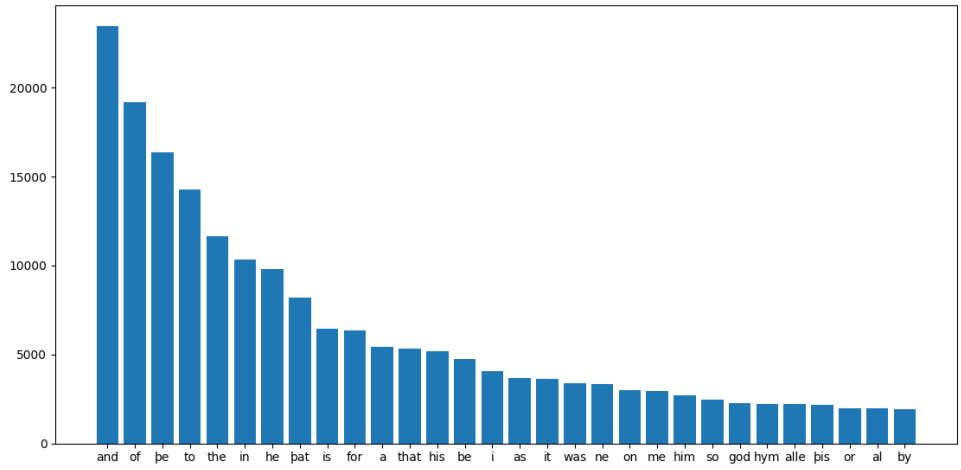

Figure 1: Zipf's Power Law causes curves that look like this one in the graph of all word frequencies across the studied time periods

A bit more unexpectedly, though, we see that while Zipf's Law applies consistently on word frequency, for letter and n-gram frequency, it seems to vary a bit more, with us seeing a major variance in the α value for the distribution, and even whether Zipf's Law seems to apply.

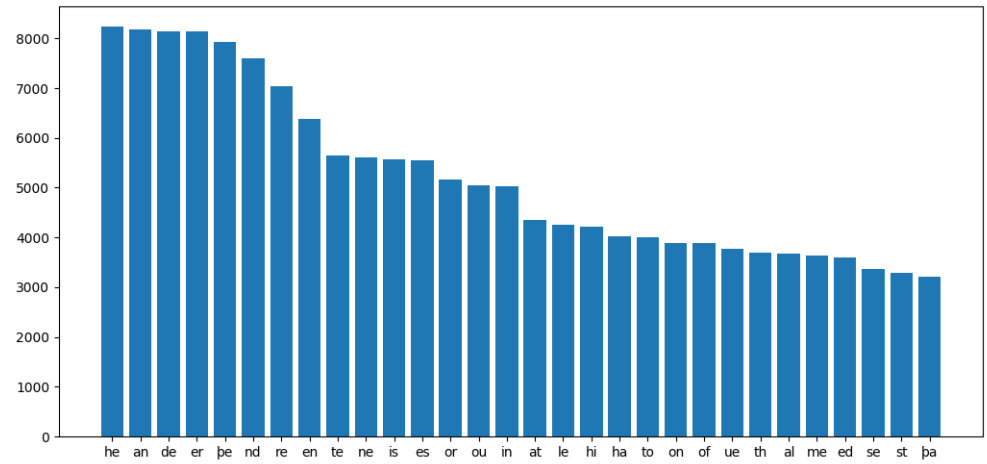

Figure 2: 2-gram frequency graph for texts from 1250-1350

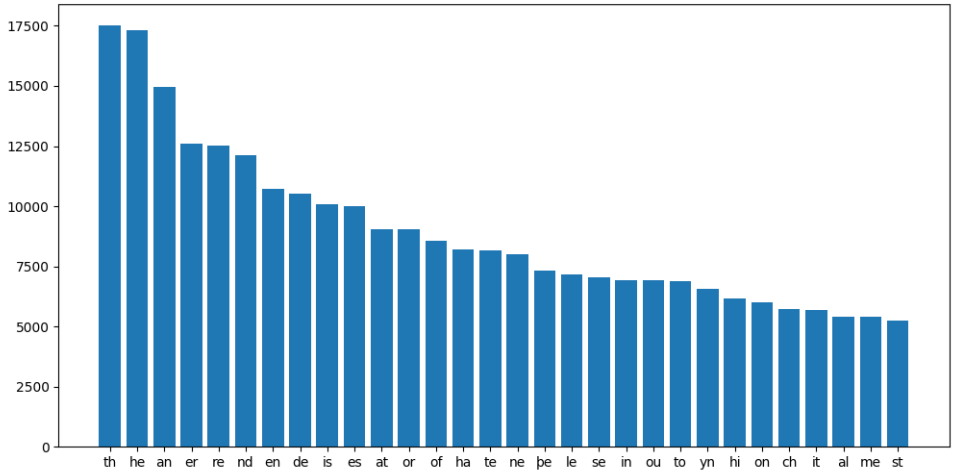

Figure 3: 2-gram frequency graph for texts from 1350-1420

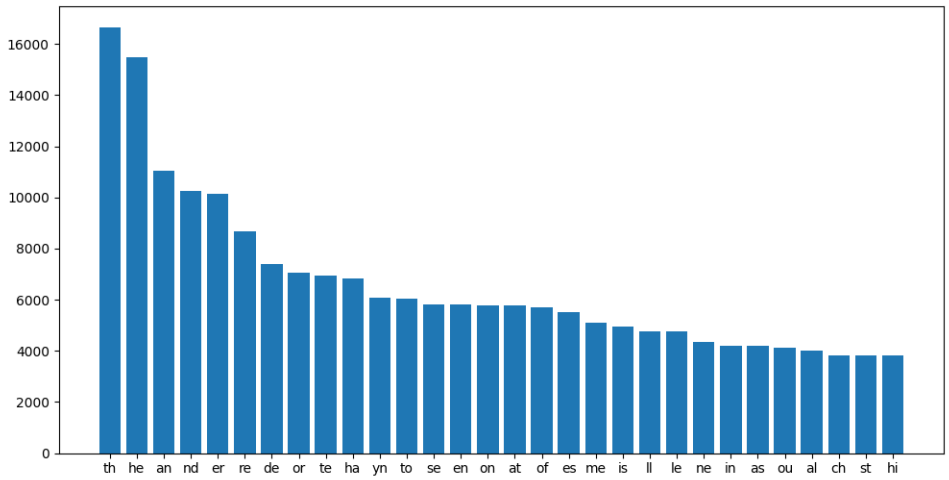

Figure 4: 2-gram frequency graph for texts from 1420-1500

Notice how the 2-gram graphs begin to take on a form more similar to Figure 1 over time. These kinds of changes correspond to major upheavals in Middle English spelling, where the Power Law emerges more strongly in texts following the Black Death, when orthography was more standardized. Whether this spelling standardization was the cause of the change in 2-gram frequency is a question better left to actual linguists, but this is a worthwhile and interesting line of inquiry.

One more observation from our data which reproduces known historical understanding comes from the analysis of word frequency, where we see a shift in the spelling of "þe" to "the," a major change that corresponds to the general trend towards the adoption of the Latin-influenced "th" replacing the native English characters "þ" (thorn) and "ð" (eth). Looking at Figures 5 and 6, we can see the rise of "the" in the aftermath of the Black Death.

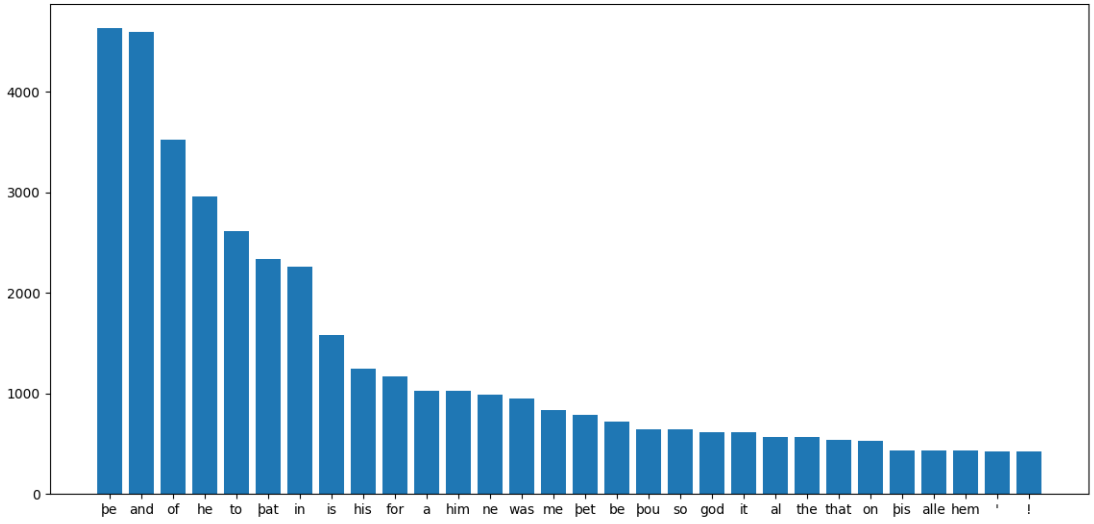

Figure 5: Word frequency graph for texts from 1250-1350

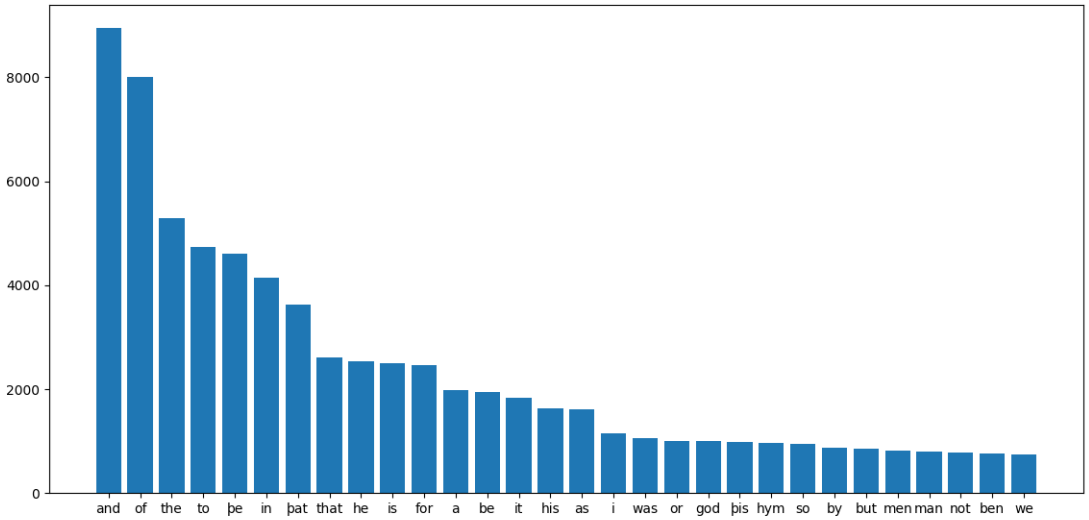

Figure 6: Word frequency graph for texts from 1350-1420

We can see a similar trend in our letter frequency graphs, where the thorn and eth characters decrease massively in frequency, while “t” and “h” increase. The effect is so great that eth goes entirely missing after 1420.

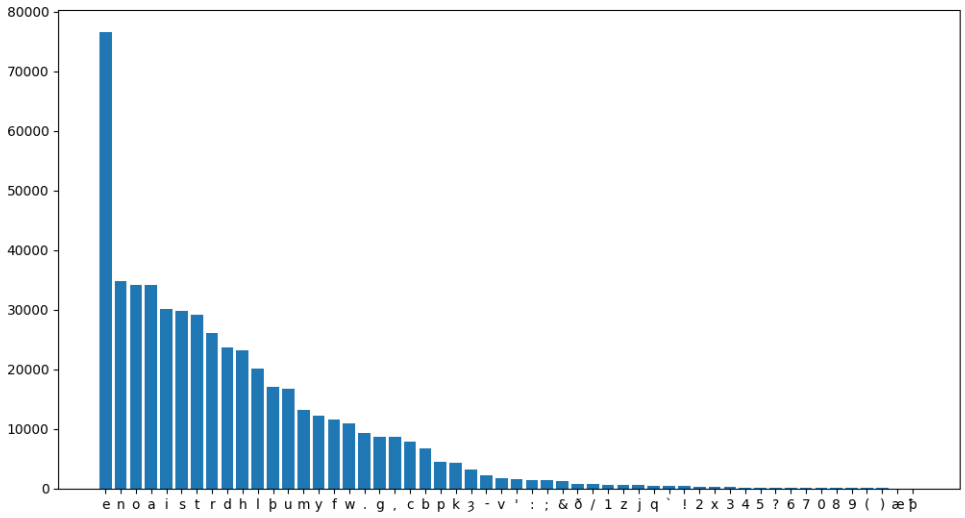

Figure 7: Letter frequency graph for texts from 1250-1350

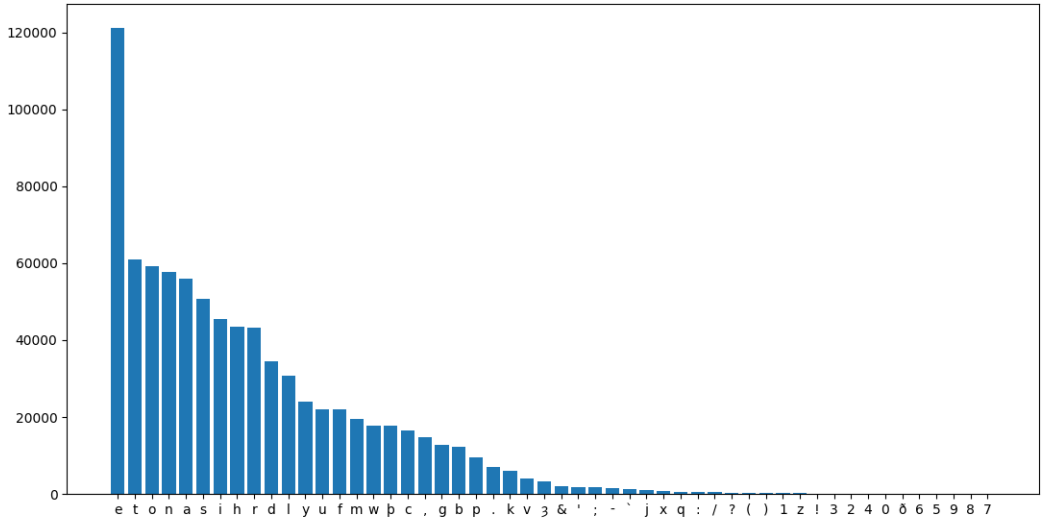

Figure 8: Letter frequency graph for texts from 1350-1420

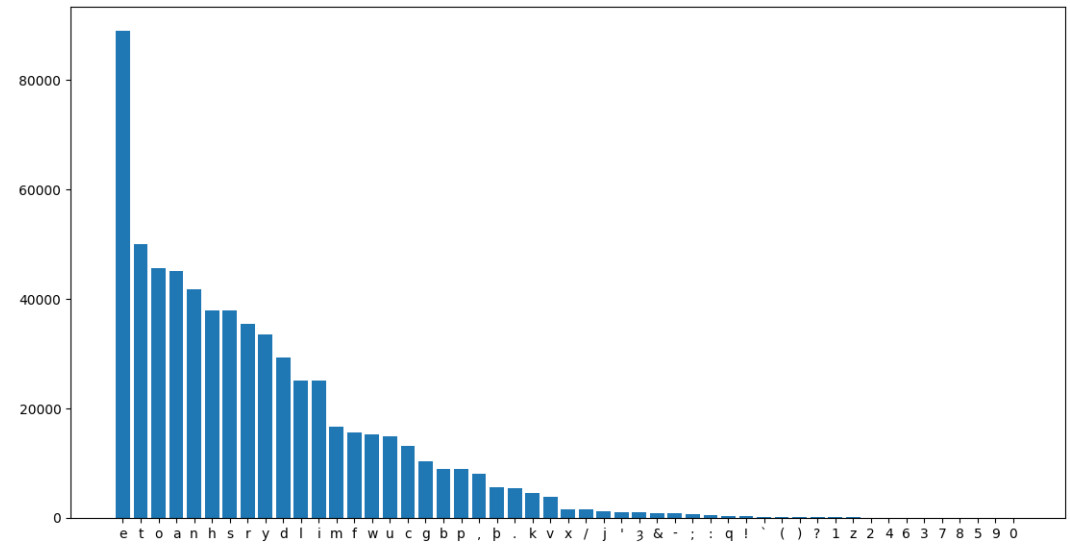

Figure 9: Letter frequency graph for texts from 1420-1500

5 Conclusion

We see that simple computational methods can reflect and build upon our understanding of historical linguistic trends. Middle English was a unique and remarkable language that saw a lot of change over the course of its lifetime. Our data reflect how the Black Death brought about massive linguistic shift, specifically through the lens of orthography. We see continuing French and Latin influence in England bringing about slow but steady changes in the general trends of spelling, making some letters and word variants more popular as time went on. The fall of thorn and eth and the rise of "th" are well documented [1], but the observations that Zipf's Power Law may only apply to letter n-grams under specific circumstances may require further study.

In general, there are plenty of unique places a follow-up analysis could go. Studying specific dialects for similar data may lead to more nuanced conclusions. It would also be interesting to see if other languages would have similar or completely different trends. No matter where future investigations lead, this one is a reminder that simple methods can still offer remarkable findings.

References

[1] C. Upward and G. Davidson, *The History of English Spelling*. Oxford: Wiley-Blackwell, 2011.

[2] L. Wright, "Rising living standards, the demise of anglo-norman and mixed-language writing, and standard english," in *The Multilingual Origins of Standard English*, 2020, pp. 1–16.

[3] R. Hotta, "Etymological respellings on the eve of spelling standardisation," *Studies in Medieval English Language and Literature*, vol. 30, pp. 41–58, 2015.

[4] U. o. H. Department of Modern Languages, *Helsinki corpus of english texts*, https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/index.html.

[5] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, 2014. DOI: 10.3758/s13423-014-0585-6. [Online]. Available: https://doi.org/10.3758/s13423-014-0585-6.