

**Two Parts of the Same Story: A Stylometric Analysis of
Wuthering Heights by Emily Brontë**

Jelige: claochlú

Table of Contents

Introduction	1
Methodology	5
Findings	12
Conclusion.....	19
Works Cited	21
Appendix A	23
Appendix B.....	29
Appendix C	57

Introduction¹

Given Emily Brontë's private nature and the limited documentation of her writing process, the exact details of her revisions and edits to *Wuthering Heights* are not extensively documented.² However, there seems to be a consensus that she made substantial revisions and edits during the novel's creation. In 1983, Tom Winniffrith and Edward Chitham commented on the possibility that *Wuthering Heights* was initially intended to be a one-volume novel.³ Eighteen years later, Chitham pointed out a noteworthy shift in the narrative tone between what he refers to as the first and second parts of *Wuthering Heights*, a facet which "has been shelved and worse, not even noted" (Chitham 86). He suggests that the 1846 version of the novel ". . . consisted of the chapters now numbered 1 [part], possibly a small part of 2, 3-17 and parts of 32 and 33. The expansion was effected both by writing completely new chapters [18-31] and by adding parts to the existing chapters, beginning with the 'mistakes' at the beginning and again at the start of Chapter 32" (160).⁴ This perspective also suggests that the narrative about the second generation might have been composed following the dismissal of the initial version.

Emily Brontë likely worked on *Wuthering Heights* in three periods: "January to March 1845 for the first creative effort; early autumn 1845 to June 1846 for the second; and September 1846 to May 1847 for the third" (Chitham 89). This timeline is primarily constructed based on insights from Charlotte Brontë's correspondence with friends and publishers, supplemented by Chitham's observations. Furthermore, Emily's poem "Two Children" titled by Charlotte in 1850 suggests that the main themes of *Wuthering Heights* occupied Emily's mind on May 28, 1845, as the poem bears resemblance to the novel:

I—the image of light and gladness—
Saw and pitied that mournful boy,

¹ I would like to thank Viktória Kiss for inspiring this analysis with her BA thesis on the publication history and the contemporary reception of Emily Brontë's *Wuthering Heights*. See, Kiss, Viktória. "A Strange, Inartistic Story." *Wuthering Heights: publication history and contemporary reception*. 2020. BA Thesis, University of Pécs. I would also like to thank Dr Tamás Fekete for his help and for introducing me to the programs created by Laurence Anthony.

² This study was carried out within the framework of the Kriszbacher Ildikó Tehetséggondozó Program (KITEP).

³ For more information on Winnifreth's claim, see Chitham, Edward, and Tom J. Winniffrith. *Brontë Facts and Brontë Problems*. Macmillan Press, 1983.

⁴ For a more detailed description of why Chitham divides *Wuthering Heights* as such, see, Edward Chitham. *The Birth of Wuthering Heights: Emily Brontë at Work*. Palgrave Macmillan, 2001.

And I vowed—if need were—to share his sadness,
And give to him my sunny joy. (*Poems by Currer, Ellis and Acton* 193, also qtd. in
Dunn 337)

These counterparts – “image of light” and “mournful boy” – resemble Catherine Earnshaw and Heathcliff as well as Cathy Linton and Hareton.

The first mention of the co-publication of the Brontë sisters is in a letter to London bookseller and publishers Aylott and Jones on April 6, 1846:

C. E & A Bell are now preparing for the Press a work of fiction—consisting of three distinct and unconnected tales which may be published either together as a work of 3 vols. of the ordinary novel-size, or separately as single vols—as shall be deemed most advisable. It is not their intention to publish these tales on their own account. (Smith, *Selected Letters* 72)

The reason she reached out to Aylott and Jones specifically is because, by that time, they had already agreed to publish the poems of the Brontë sisters.⁵ In a letter to publisher Henry Colburn, dated July 4, 1846, Charlotte Brontë detailed the intended three-volume publication of their collective works: “Sir, I request permission to send for your inspection the M.S. of a work of fiction in 3 vols.—it consists of 3 tales, each occupying a volume and capable of being published together or separately, as thought most advisable” (Smith, *Letters* 72). The volumes would have been *Wuthering Heights* by Emily, *Agnes Grey* by Anne Brontë and *The Master*, later revised as *The Professor*, by Charlotte Brontë. Although the sisters initially submitted their novels to Colburn in July 1846, he declined to publish them. It is worth noting that “. . . the novels considered by Colburn cannot have been the texts as known now, and when they were wholly or partly rejected by a second publisher Charlotte and Emily both rewrote their books, though it seems that Anne did not” (Chitham 87-8). Consequently, the sisters dispatched their manuscripts to alternative publishers, and ultimately, Thomas Cautley Newby accepted *Wuthering Heights* and *Agnes Grey* for publication on unfavourable terms: “. . . Anne and Emily had to pay an advance of fifty pounds, eventually to be reimbursed from sales of the works”;

⁵ Aylott and Jones published *Poems* on May 22, 1846, by Currer, Ellis, and Acton Bell, despite their preference for theological works. For more information on the process of the publication of *Poems*, see Margaret M. Smith. *Selected Letters of Charlotte Brontë*. Oxford UP, 2010.

however, after enough copies were sold, he failed to fulfil his end of the deal (Ohl 125).⁶ It is unclear whether the terms dissuaded Charlotte from signing the contract, or if it was Newby's choice to exclude her work from the volume. In either case, the remaining works had to be extended. Newby only released the works in December 1847, timing it to leverage the popularity of Currer Bell's *Jane Eyre* and create the impression that they were authored by the same person (Trapenard 17). Charlotte's disagreeable opinion of Newby likely stems from his readiness to exploit the popularity surrounding *Jane Eyre*, and blur the lines between Acton, Currer, and Ellis Bell, creating the illusion that they are the same person, and that one person behind the pseudonyms is the author of *Jane Eyre*. "Many early readers assumed that the novels had been written by one, or two, writers, not three" (Stoneman 214). Although the title page said, "*Wuthering Heights*: A novel by Ellis Bell, In three volumes"—which is untrue, as the third volume is *Agnes Grey* written by Acton Bell—in 1848, Newby promoted this book as "Mr. Bell's Successful New Novel" in the *Examiner*. By omitting the first name of the author, Newby further maintained the misconception surrounding the authorship of the book. Then, he introduced its initial American edition through Harper and Brothers under the name "By the Author of *Jane Eyre*" (Smith, *Letters*, 17-8). Despite numerous efforts to dispel this rumour, most readers continued to mix up the identities of the Bells. From Charlotte's letter to publisher W. S. Williams on July 31, 1848, it becomes evident that Emily was against public recognition: "'Ellis Bell' will not endure to be alluded to under any other appellation than the 'nom de plume'" (Smith, *Selected Letters* 113). After the death of Emily and Anne, Charlotte tried to clear the confusion in her 'Biographical Notice' in the second edition of *Wuthering Heights* and *Agnes Grey*: "It becomes, then, my duty to explain briefly the origin and authorship of the books written by Currer, Ellis, and Acton Bell" (Dunn 307). Originally, the sisters "agreed to conceal their identities because Emily refused public recognition, but the resulting confusion was repugnant to Charlotte, who soon revealed the truth 'we are three Sisters'" (Stoneman 217). Thus, the private nature of Emily not only limits scholars' ability to decipher the details of her revisions and edits to *Wuthering Heights*, but it also indirectly causes confusion among contemporary readers of the works of the Bells.

The widespread availability of circulating libraries improved the distribution of novels and eventually prompted both writers and publishers to adopt the three-volume format, also known as the three-decker format (Moretti, *History of the Novel* 7). This format allowed a single book to be simultaneously borrowed by three readers and played a role in shaping the Victorian

⁶ All translations made from the French are mine.

novel's content and structure, making it more suitable for respectable middle-class audiences and divisible into three substantial parts (MacKay 111-12). To comply with the formal requirements of the three-volume publication, Emily doubled the length of *Wuthering Heights*. It is highly plausible that the novel was not composed in a single continuous effort, and substantial alterations were made to the narrative prior to its publication. However, due to the absence of manuscripts, the exact details of these changes cannot be ascertained (Dunn 277). The assumed refinement of the novel involved three steps: Lockwood's return to Thrushcross Grange to learn the story's conclusion, the development of the new generation of characters to address critics and prepare for publication, and the decision to tighten the story's temporal framework (Chitham 165).

This study aims to investigate whether it is possible to more specifically determine than Chitham did, which parts of *Wuthering Heights* were added to the story later. Upon my recent reading of the novel, my working hypothesis centres on Chapter 18 as the assumed start of the revised novel. This hypothesis stems from the twelve-year leap between the two parts and its chapter's thematic resonance with the typical introductions of heroines in novels, which often explore the affections and character of a young female protagonist. In line with Chitham's claims and my hypothesis, this current pilot study will assume that *Wuthering Heights* consists of two parts, written approximately one year apart. Thus, the investigation will revolve around the 1846 version of the text, containing chapters 1-17 and 32-33, and the 1847 version comprising chapters 18-31 and 34. I anticipate that through Emily Brontë's use of adjectives I can determine a stylistic change in her writing, which could indicate that chapters 18-31 and 34 were written later. This is a pilot study, using only programs and software that are easily accessible on the internet.

Methodology

Stylisticians, primarily interested in text style, focus on understanding how linguistic choices in a text contribute to its overall meanings and effects (McIntyre and Walker, “Combining Corpus Linguistics and Stylistics” 16). Stylometry, as the study of linguistic style, delves into the meticulous examination of writing patterns to discern the unique characteristics of individual authors or author groups. Despite the nuanced distinction between text style and authorial style, it is important to recognize the inherent interconnectedness as acknowledged by seasoned stylisticians (16). Even in acknowledging this linkage, contemporary stylisticians tend to concentrate on the analysis of text style, emphasizing the importance of studying how linguistic choices shape the characteristics of a text. The controversy in literary studies began with Franco Moretti’s distant reading, a method where distance “is a condition knowledge,” allowing scholars to focus on units that are either much smaller or much larger than the text itself (Moretti, *Distant Reading* 48). During the process, the text itself may disappear from direct scrutiny. However, this disappearance is justified because it allows scholars to gain a broader understanding of literary systems and patterns.

The predominant methodology applied in stylometry entails scrutinizing the 100 most frequently used words in a text, database, or a chosen sample to generate a set of norms that describes the stylistic idiosyncrasies of the text and its writer (Burrows 268). An exemplary approach within this analytical framework involves cluster analysis, where statistical techniques are employed to categorize texts based on discernible linguistic or stylistic features. The process begins with the extraction of features from a text, such as word frequency, sentence length, and vocabulary richness. The data are then represented in matrices or vectors, with each row corresponding to a text and each column to a stylometric feature. The resulting clusters may represent different writing styles, genres, or authors. Equally prevalent within the stylometric analysis is the utilization of predictive modelling, commencing with the extraction of pertinent features including word frequency, sentence length, and vocabulary richness. After pre-processing the data by normalizing linguistic variations, a suitable predictive modelling algorithm, such as support vector machines or decision trees, is selected. After the training of the model on a labelled dataset, where it assimilates the intricate relationships between extracted stylometric features and authorship, evaluation metrics like accuracy and precision become pivotal in assessing the model’s efficacy on an independent dataset. Finally, the trained model can be applied to attribute authorship to unknown texts by analysing their stylometric features and making predictions based on the patterns learned during training. Carmen Klaussner and

Carl Vogel used predictive modelling for their stylochronometric analysis of the novels of Henry James and Mark Twain.⁷ Their work introduces innovative analysis and preparation techniques geared towards temporal stylometry, affirming the viability of reliably inferring the year of publication from lexical features. They conclude that, regardless of whether one is examining a specific author or a broad reference collection from the era, it is feasible to reliably infer the year of publication from lexical features.

Burrows' Delta Method is a metric for stylistic disparity between two or more bodies of text, intricately involving statistical analysis of word frequencies, specifically the ratio of the frequencies of function words (e.g. articles, prepositions, pronouns) to content words in each text or set of texts (McIntyre and Walker, "Using Corpus Linguistics" 501). Its distinctive capacity to facilitate cross-textual comparison enables the systematic grouping of documents exhibiting analogous frequency profiles (Eve 31). To account for variations in text length, a normalization process is often applied by dividing the raw frequency of each function word by the total number of words in the text (Eve 501). Burrows notes that generally, even if "authors' style change over the years, they do not change beyond recognition" (Burrows 279). However, he concedes that certain cases, exemplified by authors like Abraham Cowley, exhibit "signs of change so strong" that they impact the analytical data (Burrows 279). Eve also notes a similar point while talking about the assumed stylistic naturalism of writers, which is "the notion that authors have a set of profilable characteristics in their writing over which they can never truly have full control" (Eve 32). The flaw in the assumed stylistic naturalism is "the belief that stylometric profiles do not change even when authors try to modify their writing style. This also assumes consistency in authors' styles throughout their lives, which is a disputed claim" (32). The Brontës' worked as a collaborative force since their youth. Emily and Anne engaged in literary pursuits centered on their fictional realm, Gondal, while Charlotte and Branwell participated in a competitive literary discourse concerning their West African domain and the metropolis of Glasstown (Ohl 26-34). Drawing inspiration from Byron and Wellington, both sets of siblings evinced a shared interest, yet Charlotte and Branwell exhibited a distinct predilection for political and military themes (Bock 7-8). Charlotte's romantic preoccupation with Wellington manifested in her portrayal of an imaginary son. In contrast, Emily and Anne's early works, particularly *Wuthering Heights*, featured less politics and more realism inspired

⁷ In 2015, Klaussner and Vogel used the regression analysis applied to linguistic variables in predicting a temporal variable in the works of Henry James and Mark Twain. For more information on their findings, see Klaussner, Carmen, and Carl Vogel. *Stylochronometry: Timeline Prediction in Stylometric Analysis*. SGAI Conferences, 2015.

by their surroundings. Emily's narratives diverged from Charlotte's penchant for romanticizing upper-class characters. While writing *Wuthering Heights*, Branwell, under the pseudonym Northangerland, became the first sibling to be published with poems in northern daily papers. In 2020, Rachel McCarthy and James O'Sullivan set out to investigate the authorship of *Wuthering Heights* by analysing the writing styles of the Brontë siblings—Emily, Charlotte, Anne, and Branwell. The authors argue against alternative theories, particularly the suggestion that Branwell or Charlotte might have been the authors, presenting quantitative evidence to support their conclusion (McCarthy and O'Sullivan 6-7). Additionally, the study discusses the influence and collaboration between the siblings, emphasizing the shared style within the Brontë family.⁸ My analysis might reveal further evidence that the author of *Wuthering Heights* without doubt was Emily Brontë.

However, research of this nature might have drawbacks due to the constraints of the corpus. The analysis, anchored within the confines of a solitary text (albeit two editions), necessitates a pivot towards intrinsic features such as word choice, sentence structure, vocabulary, and patterns of language usage. Due to the lack of other novels written by Emily, I will also look at her letters and diary entries; however, there are aggravating factors here as well. Out of the four diary entries, one is co-written by Anne and only one of them was written during the composition process of *Wuthering Heights*. Of her letters, there is only one that fits the timeline. Still, even these pieces of writings are not perfectly able to provide the same stylistic features as another novel written by Brontë would. In 1999, Binongo and Smith distinguished between the style of Oscar Wilde's plays and his essays using "principal component analysis as a tool for dimensional reduction" (Binongo and Smith 787). Thus, the letters and diary entries of Emily differ in terms of style from *Wuthering Heights* and are not appropriate for this current analysis. While some stylometric techniques can still provide insights into the author's writing style based on letters and diary entries, the absence of additional works for comparative analysis unavoidably reduces the statistical robustness of the inquiry. Stylometric analysis inherently gains potency and accuracy when applied to a larger body of work from the same author, as it allows for a more comprehensive understanding of the author's consistent writing style. Although Emily did produce poems, such compositions remain excluded from this analysis because their form and style differ from that of a novel. As a result, style in the context of this essay, will refer to word preferences that are distinctive

⁸ For more information on the collaboration between the Brontë siblings, see McCarthy, Rachel, and James O'Sullivan. "Who wrote *Wuthering Heights*?" *Digital Scholarship in the Humanities*, vol. 36, no. 2, 2020, pp. 383–91, <https://doi.org/10.1093/lc/fqaa031>.

markers helping to distinguish between the proposed two parts of *Wuthering Heights*. This examination of word choice holds the promise of unveiling profound insights into the style of Brontë's text and an answer to what parts of the story were added later to the text. Central to this analysis is the essential inquiry: can these delineated sections, as assumed by Chitham, withstand scrutiny as reliably discernible products of later revision through stylistic analysis?

The construction of the corpus is underpinned by a selection of two editions of the work, both of which I have cut into chapters, for visualization. The first is the fifth edition of the Norton Critical Edition of *Wuthering Heights* edited by Alexandra Lewis in 2019. This edition assumes particular significance, as it preserves the paragraph breaks of the original text and incorporates,

the most significant of Charlotte's corrections to punctuation on the following basis: adopting the 1850 edition's removal of commas which are superfluous or would lead to misreading ("to enjoy the cheerful fire, and smoking coffee which the servant has prepared for my refreshment" [volume I, chapter III]), and the addition of commas absent in 1847 but necessary for sense ("he interrupted wincing" [volume I, chapter I]). (Brontë, Lewis 12-15)

The word use is faithful to the 1847 edition of the work. Unfortunately, I could not find the original 1850 second edition of the novel in a format that is appropriate for this research. HathiTrust holds a copy from 1851, which is a reprint of the 1850 edition; however, the full text is only downloadable in a jpg format, which would not allow me to upload the text to the software I use. In the future, I will use OCR to analyse the edition found on HathiTrust. As a result, the second work under consideration is the 1889 edition of the text which is a reprint of the 1850 edition. It includes the edits made by Charlotte Brontë, published by Smith, Elder and Company. It is a reprint of the 1850 edition, which includes the Biographical Notice of Ellis and Acton Bell written by Currer Bell, a Preface to *Wuthering Heights* by Currer Bell, as well as the two novels, *Wuthering Heights* and *Agnes Grey*. Furthermore, this is the edition that Chitham refers to in his argumentation. To determine the extent of linguistic change between the two editions, I will first analyse these editions. If there is a significant difference between the two, that would help to point out the authorial fingerprints of Emily. Then both editions will go through a thorough stylometric analysis. Such changes would include sentence structure, language use, and the frequency of the adjectives in the novel.

For the purposes of quantitative research, I rely on two programs developed by Laurence Anthony—TagAnt and AntConc. TagAnt is a freeware, multi-language segmenter and Part-Of-Speech (POS) tagger. This software lays the groundwork for preparing the chapters of *Wuthering Heights* for the subsequent use of AntConc to identify patterns, trends, and linguistic features within the data. Within the framework of this analysis, AntConc assumes a pivotal role in unearthing stylistic discrepancies between the two proposed segments of *Wuthering Heights*. With the help of this software, I will be able to do multidimensional scaling (MDS), which is often employed to analyse and visualize the relationships between words, phrases, or other linguistic elements based on their similarity or dissimilarity in certain linguistic features. It has also been used in authorship studies to analyse writing styles and linguistic features, representing texts in a space where similarities and differences can be more easily observed. To visualize my findings, I will use Voyant Tools, a free text analysis tool, built by Geoffrey Rockwell and Stéfán Sinclair. This survey will focus on where and how frequently Emily Brontë used adjectives in her novel. While the quantities of adjectives have been integrated into general n-gram methodologies, there has been limited investigation into their potential use in authorship attribution, which my study aims to do.

The preparatory steps for the text analysis involve the conversion of the two editions into a txt file format, optimized for the operational requirements of the programs TagAnt and AntConc.⁹ Then comes the use of TagAnt to tag the files, for example,

1. Coordinating conjunction: _CC
2. Cardinal number: _CD
3. Determiner: _DT
4. Existential there: _EX
5. Foreign word: _FW
6. Preposition or subordinating conjunction: _IN
7. Adjective: _JJ
8. Adjective, comparative: _JJR
9. Adjective, superlative: _JJS
10. List item marker: _LS
11. Modal: _MD

⁹ The footnotes of the novels were removed from both editions, for a clearer result. Afterwards, the works must be divided chapter by chapter into separate files. This step will be crucial when using AntConc for the analysis of word frequency.

12. Noun, singular or mass: _NN
13. Noun, plural: _NNS
14. Proper noun, singular: _NNP
15. Proper noun, plural: _NNPS
16. Predeterminer: _PDT
17. Possessive ending: _POS
18. Personal pronoun: _PRP
19. Possessive pronoun: _PRP\$
20. Adverb: _RB
21. Adverb, comparative: _RBR
22. Adverb, superlative: _RBS
23. Particle: _RP
24. Symbol: _SYM
25. to: _TO
26. Interjection: _UH
27. Verb, base form: _VB
28. Verb, past tense: _VBD
29. Verb, gerund or present participle: _VBG
30. Verb, past participle: _VBN
31. Verb, non-3rd person singular present: _VBP
32. Verb, 3rd person singular present: _VBZ
33. Wh-determiner: _WDT
34. Wh-pronoun: _WP
35. Possessive wh-pronoun: _WP\$
36. Wh-adverb: _WRB

The next step was to see if the chosen editions of *Wuthering Heights* really differ in their word frequencies. To ascertain this, the tagged versions of the texts are input into AntConc, enabling a search for “jj” (TagAnt uses “jj” to tag adjectives) in the term position right. The results are then sorted based on the frequency of appearance, with the cluster size configured to two, given the contextual insignificance of the words’ surrounding context. The outcomes of this analysis reveal discernible differences in frequency and rank between the two texts. The analysis will also use Pearson’s correlation coefficient, which can be used to measure the similarity or correlation between different linguistic features across texts or authors.

After the examination of the two source texts, I begin the analysis of *Wuthering Heights* by looking at the two editions separately. Based on the list of the 100 most used adjectives used in the novel, I look at how and where they appear in the chapters. I am looking for possible imbalances of appearance. This would show a change in the style of Brontë. Furthermore, superlatives can also be indicative of a possible shift, however, here it is also important to consider the context. From this point, based on Chitham's observations and my hypothesis, chapters 1-17, 32 and 33 will be referred to as 1846 version, while chapters 18-31 and 34 will be the 1847 version.

Findings

The comparison between the two editions shows that Charlotte made some linguistic changes to her sister's work. This aligns with Chitham's examination that "the two volumes of *Wuthering Heights* as printed by Newby contained 348 and 416 pages respectively, while *Agnes Grey* contained 363 pages. In a typical nineteenth-century edition, the Smith and Elder printing of 1889, there are 180 pages of *Agnes Grey* to a rather more tightly packed 305 of *Wuthering Heights*; this of course reflects Charlotte's revisions of 1850" (95). The two editions of the same book exhibit subtle variations in their linguistic characteristics. In terms of vocabulary density, the 1889 edition (0.105) surpasses the Lewis edition (0.096) indicating a potentially more diverse and enriched lexicon. However, when examining the average words per sentence, Lewis's edition (25.3) slightly outpaces the 1889 edition (24.8), suggesting a writing style that leans towards longer and more complex sentences. There are slight differences in the frequency of adjective use in the two editions. The adjective "beautiful" appears in the 100 most frequent words in the 1889 edition, but it is not in the Norton edition, which is based on Emily's first edition. Because the two editions have different lengths, I performed term frequency (TF) normalization. In natural language processing, term frequency is often normalized to address the length of documents and make comparisons between documents more meaningful. However, to get numbers that are easier to deal with, I multiply the sum by 100 000 based on the overall number of words in the text.

$$\text{Normalized T} = \frac{\text{Occurrence of Term}}{\text{Total Number of Terms in the Document}} * 100\,000$$

Although AntConc provides an estimated normalized frequency, I have decided to do calculations using Excel, because for my analysis I need normalized term frequency. To see the full list of the 100 most frequent words in each edition and their TF, refer to Appendix A. In the Lewis edition there are 711 words that appear only once in the novel, while in the 1889 edition there are 817. Interestingly, "accomplished" has a frequency of 2, and a normalized frequency of 1,368906654 in the Lewis edition, while in the 1889 edition it has a frequency of 1 and a normalized frequency of 0,803038698. This shows that the text edited by Charlotte holds more than "diluting the use of dialect and (possibly in conjunction with the printers) modifying paragraphing and punctuation well beyond the scope of correction of obvious errors"

(Lewis 13).¹⁰ In conclusion, the comparison between the two editions of *Wuthering Heights* reveals significant linguistic changes made by Charlotte Brontë to her sister Emily's work. This proves that the investigation should be done on an edition that follows Emily's original text, since the small changes that Charlotte introduced can produce faulty results for determining the style of Emily. Thus, the analysis will only use the fifth edition of the Norton Critical Edition of *Wuthering Heights* in hopes that it will give a clearer representation of the style of Emily Brontë.

To determine whether there are discernible stylistic changes in the text written during different periods, the focus was first on imbalances in word usage. 63,6% of the adjectives appear in the presumed first writing of *Wuthering Heights* (460 of the 6381). The analysis uses the tagged version of the text to provide a better understanding and visualization of Emily's word usage. The most frequent adjective “little” appears 173 times, and 54,9% of it appears in the proposed first writing. The three most frequently used words are “and_cc” (4745), “the_dt” (4573) and “I_prp” (4034) (see Table 1).

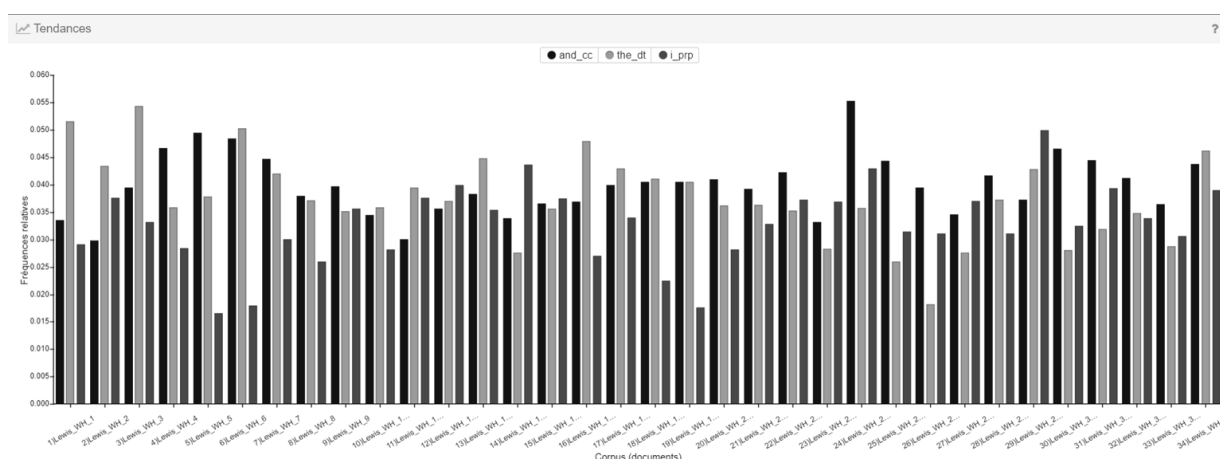


Table 1 shows the frequency of the three most used words in *Wuthering Heights*.

As the table shows, in the parts presumably written after Emily had to double the size of the book, the word “the” (black in Table 1) appears less frequently and less balanced in the 1847 text, indicating a change in the style of the affected chapters.¹¹ The word “and” appears 2850 times (2,1% of the part) in the first version, while 1895 times in the 1847 version (3,39% of the

¹⁰ Here Lewis talks about the 1850 edition of the text; however, since the edition that I am using for my analysis is a reprint of the second edition, her summary of the changes are true for my copy as well.

¹¹ From this point, based on Chitham's and my hypothesis, I will refer to chapters 1-17 and 32-33 as 1846, while chapters 18-31 and 34 will be 1847. The dates denote the period when Emily wrote the sections.

second part). However, the use of the word seems to be balanced except for chapter 24, when there is a jump.¹² Emily Brontë seemingly changed her sentence structure when she picked up the novel for the second time. Using VoyantTools Phrases function, we can see repeating sequences of words in a given text. When applied to *Wuthering Heights*, the longest sequence is 11 words long. For a better representation of style, I have added the column “where” to show in which chapter the sequence appears in (see Table 2). By using the tagged text, the results show a more faithful representation of the repeating sequences in the text. Without the tags, cases such as “and I’ll tell you all about it,” do not appear in the list.

Terms (tagged)	Terms (without tags)	Total	Length	Where
if_in you_prp do_vbp n't_rb let_vb me_prp in_rp i_prp ll_md kill_vb you_prp	if you don't let me in i'll kill you	2	11	Ch 24 and 24
and_cc i_prp ll_md tell_vb you_prp all_dt about_in it_prp	and i'll tell you all about it	2	8	Ch 6 and 32
at_in the_dt end_nn of_in the_dt house_nn and_cc	at the end of the house and	2	7	Ch 1 and 34
at_in the_dt top_nn of_in the_dt steps_nns and_cc	at the top of the steps and	2	7	Ch 7 and 24
common_jj humanity_nn and_cc a_dt sense_nn of_in duty_nn	common humanity and a sense of duty	2	7	Ch 14 and 14
have_vbp only_rb to_to do_vb with_in the_dt present_nn	have only to do with the present	2	7	Ch 9 and 9
mr_nnp and_cc mrs_nnp linton_nnp were_vbd	mr and mrs linton were	2	7	Ch 6 and 10
the_dt first_jj of_in the_dt seventy_nnp_hyph first_nnp	the first of the seventy hyph first	2	7	Ch 3 and 3
are_vbp you_prp glad_jj to_to see_vb me_prp	are you glad to see me	2	6	Ch 23 and 23
at_in mr_nnp edgar_nnp s_pos door_nn	at mr edgar s door	2	6	Ch 24 and 28

Term	Count	Length ↓
if you don't let me in i'll kill you	2	9
common humanity and a sense of duty	2	7
have only to do with the present	2	7
are you glad to see me	2	6
i love him better than myself	2	6
the first of the seventy first	4	6
to be as accomplished as linton	2	6
whet is thur tuh do nah	2	6
a dog in the manger	2	5
and i defy you to	2	5
at the end of the	2	5
because he is handsome and	2	5
i cannot live without my	2	5
i should like to be	2	5
if we can keep him	2	5

Table 2 shows repeating sequences of words, ordered from the longest to the shortest. The table only shows the first 10 results. The calculations were carried out using VoyantTools.

This approach is particularly useful for capturing the repetitive linguistic structures that contribute to the overall style of the work. Repeated phrases can be indicative of the author's thematic choices, or narrative motifs. By associating these phrases with particular chapters, the presence of recurring linguistic elements is highlighted. The sequence “at_in the_dt end_nn

¹² The frequency of these words is not normalized.

of_in the_dt house_nn and_cc” is an example, where the two occurrences appear in the proposed two parts of the novel. This is the only instance among the top ten results where the same sequence was used in both the 1846 and 1847 versions.

In my future analysis, this examination will be expanded to include the top 100 repeating sequences of words from the text. Another intriguing finding in connection with larger units within the text is the words per sentence ratios in the chapters (see Table 3).

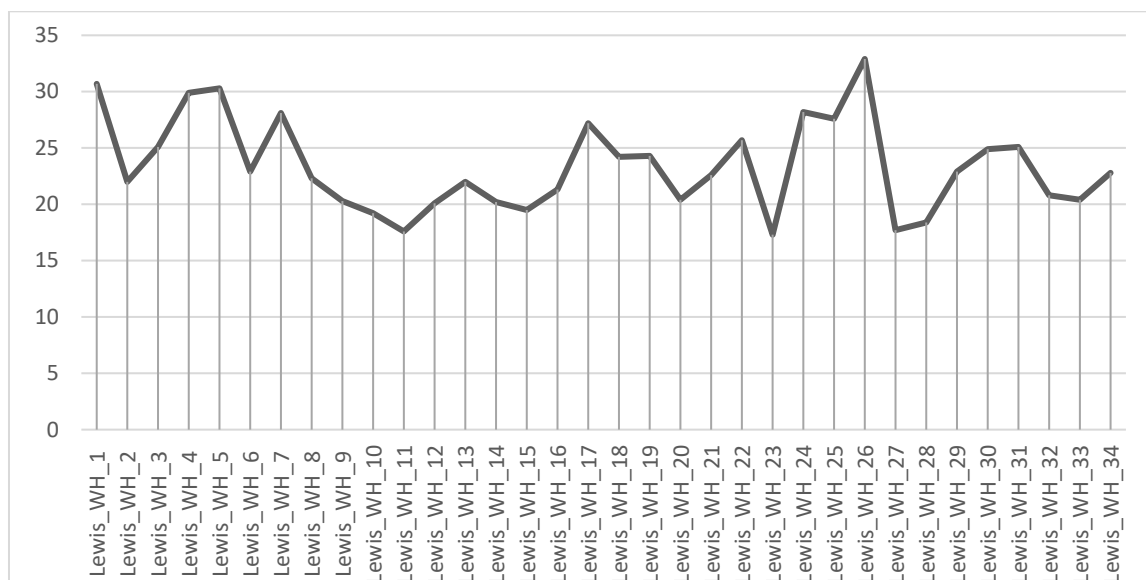


Table 3 shows that the 1847 version has more words per sentences.

However, when compared to the chapters with the highest frequency of adjectives, there does not seem to be such balanced distribution. Chapter 26 is 3rd on the list when it is sorted according to normalized term frequency (this is the 3rd shortest chapter of the entire book). At this point in the narrative, Cathy and Nelly meet with Linton, who “walked so feebly, and looked so pale” (Lewis 251). However, Chapter 24, which is 2nd on the words per sentences table, is only 15th when looking at adjective frequency. Thus, it cannot be said that chapters from the 1847 version appear to have a higher frequency of adjectives. Of the chapters from the 1846 version, chapter 17 ranks highest in sentence complexity, but it is only the 20th in the adjective use. The findings suggest that the increase in sentence complexity might not necessarily be accompanied by a proportional increase in adjective frequency. This could imply that Emily’s stylistic evolution involves more nuanced changes beyond a connection between sentence length and adjective usage. However, there appears to be a strong relationship between words per sentence and vocabulary density, as the trends appear to be similar.¹³ However, there

¹³ The vocabulary density of the chapters are determined by Voyant Tools calculations, based on the frequency of unique words in each chapter.

seems to be a close relation between words per sentences and vocabulary density. Table 4 shows this connection.

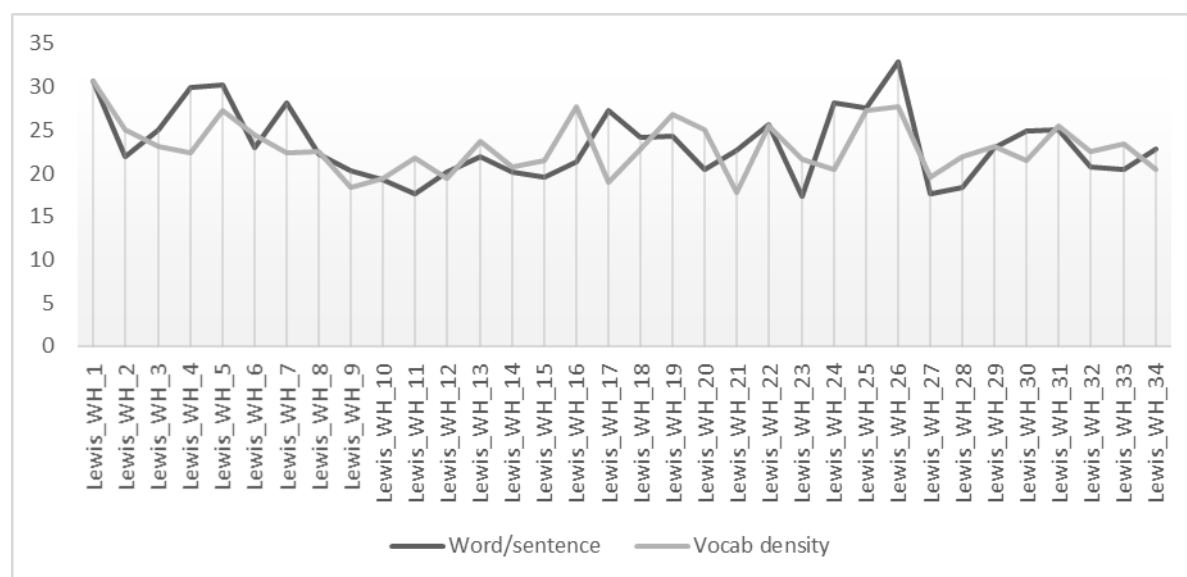


Table 4 shows the relation of vocabulary density and words per sentences in *Wuthering Heights*.¹⁴

Table 4 reveals intriguing connections and disbalances in the vocabulary density and words per sentences in the novel. The most conspicuous of these is in chapter 34, where the two graphs are moving in the opposite direction. This phenomenon is present in other chapters as well. However, there seems to be no connection between the time of writing and the differences or similarities of the vocabulary and the words per sentences.

There are 6374 adjectives in *Wuthering Heights*, consisting of 1450 different ones. 917 adjectives only appear in one part of the book. Adjectives that are exclusive to one part of the book could signal a stylistic evolution over time. If certain adjectives appear prominently in the later part of the book, it may indicate a refinement or change in the author's language choices or narrative style. For this purpose, I have looked at 1450 adjectives and identified which chapters they appear in (see Appendix B). The 1846 version has more words that only appear there than the 1847 version, for example, "dirty" appears 9 times throughout the novel, all of which are in the 1846 version. This is also the adjective with the highest frequency out of every word that only appears in only one version. Although "rich" appears in both parts, it only

¹⁴ For the visualization of the relation between the two datasets, the data of word density had to be altered. The tendency of the graph was kept. For the original data, refer to Appendix C. In the visualisation the sums were multiplied by 62 so that the two graphs start from the same point, ensuring an easier understanding of the connection between vocabulary density and words per sentences in *Wuthering Heights*.

appears once in the 1847 version, in the first sentence. After that, Emily Brontë did not use it again. This is specifically interesting because, out of the 13 occurrences, 12 are from the 1846 version. Another intriguing piece of information that was revealed by looking at where the adjectives appear is that there are 242 words that only appear in the 1847 version. The words with the highest frequency in the list are “timid” (4), “tiresome” (4), “spiteful” (4), “pony” (4), “perplexed” (4). An interesting question arises when looking at the number of adjectives used only in the 1846 version: why do they not appear in the 1847 version? In my future research, I will try to examine the context of these words to find a reason for it. My assumption is that these adjectives refer to the second generation, and such an analysis will I hope reveal that even the way Brontë described her characters have changed between the two versions.

Using Voyant Tools’ Correlation function, I found that the correlation coefficient of -0.97 indicates an extremely strong negative linear relationship between the terms “high_jj” and “open_jj.” This implies a strong stylistic preference or pattern. The absolute value of -0.97 is very close to 1, indicating an exceptionally strong negative correlation. The significance level (p-value) of 0.0000011 is very low. In statistical hypothesis testing, a low p-value suggests that the observed correlation is unlikely to have occurred by random chance. Thus, when Emily Brontë uses more adjectives related to “high,” there is a tendency to use fewer adjectives related to “open,” and vice versa (see Table 5). Because “little” is the most frequent adjective in the novel, the next word to analyse is “little.” A correlation coefficient of 0.82 was found between the terms “high_jj” and “little_jj.” This indicates a strong positive linear relationship, with a consistent tendency for the occurrences of these terms to increase together. The positive correlation suggests that as the frequency of “high_jj” increases, there is a proportional increase in the frequency of “little_jj,” and vice versa. Furthermore, the significance level associated with this correlation is 0.0035, which is below the commonly used threshold of 0.05. This low p-value suggests that the observed correlation is statistically significant, indicating that the relationship between “high_jj” and “little_jj” is unlikely to have occurred randomly (see Table 6).

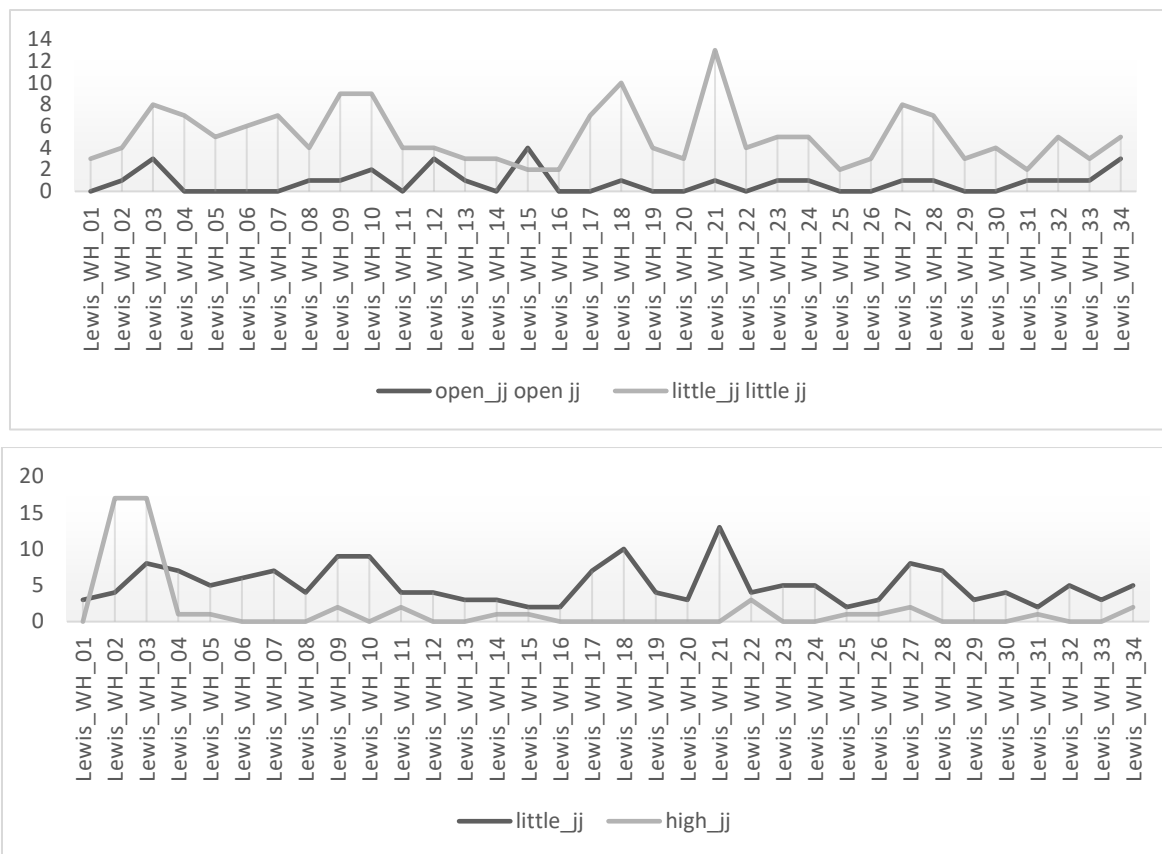


Table 5 and 6 show these adjectives appear in the novel.

This shows that on average, when “high” is used frequently in a chapter, “little” tends to do the same. The graph also shows that the two adjectives tend to decrease and increase together, with the exception, for instance, between chapters 2 and 3, or 21 and 22. It is also worth keeping in mind that “little” is the most frequent adjective in *Wuthering Heights* and it occurs 173 times while “high” only occurs 17 times.

Conclusion

This pilot study aimed to investigate whether it is possible to determine which parts of *Wuthering Heights* might have been added later. The results show that the parts that according to Chitham and my hypothesis were written after the expansion of the novel, differ from first part in its adjective use. Furthermore, imbalances in word usage were found in the text revealing changes in the frequency of certain words, especially in chapters presumed to be written after Emily expanded the novel. The analysis of repeating sequences of words provides insight into recurring linguistic structures and their association with specific chapters, shedding light on the author's stylistic choices. In the top ten results, we can see only one occurrence of repetition from different parts of the text.¹⁵ The correlation analysis between specific adjectives, such as “high” and “open,” and “high” and “little,” unveils strong relationships, suggesting consistent stylistic preferences or patterns in Emily Brontë's writing. The investigation into adjectives exclusive to particular parts of the book raises questions about the stylistic evolution of *Wuthering Heights* and Emily Brontë over time. The presence of certain adjectives exclusively in the 1846 version prompts further exploration of their contextual significance.

Throughout the analysis, I employed various analytical tools and techniques to scrutinize linguistic elements, vocabulary, sentence structure, and the frequency of adjectives in different editions of the novel. By doing so, my aim was to identify any stylistic changes that could indicate revisions made by Emily Brontë. Based on the findings of this pilot study, there are discernible stylistic differences in chapters 1-17 and 32-33, and in the later addition of chapters 18-31 and 34. In future analyses, the study plans to extend the examination to the top 100 repeating sequences of words and delve deeper into the context of words exclusive to specific versions. The presumed outcome of this investigation would show that the words exclusive to the 1847 version refer to words, people, and themes that only appear in that part of the novel. By examining linguistic patterns, narrative techniques, and stylistic changes across different chapters of the novel, such a research can offer new material for discussing Brontë's artistic growth, even in the absence of other novels written by her. Furthermore, an exploration of how her style may have changed or matured over the course of writing the book would also be possible to discern.

¹⁵ Here parts refer to what I called the 1846 and 1847 version of the text.

Works Cited

- Binongo, José Nilo, and M. W. Smith. "A bridge between statistics and literature: The graphs of Oscar Wilde's literary genres." *Journal of Applied Statistics*, vol. 26, no. 7, 1999, pp. 781–87, <https://doi.org/10.1080/02664769922025>.
- Bock, Carol. *Charlotte Brontë and the Storyteller's Audience*. University of Iowa Press, 1996.
- Brontë, Emily, and Alexandra Lewis. *Wuthering Heights: The 1847 Text, Backgrounds and Contexts, Criticism*. Norton Critical Edition, 5th ed., W. W. Norton & Company, 2019.
- Brontë, Emily, and Anne Brontë. *Wuthering Heights*. Smith, Elder, 1889.
- Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing*, vol. 17, no. 3, 2002, pp. 267–87, <https://doi.org/10.1093/llc/17.3.267>.
- Chitham, Edward, and Tom J. Winniffrith. *Brontë Acts and Brontë Problems*. Macmillan, 1983.
- Chitham, Edward. *The Birth of Wuthering Heights: Emily Brontë at Work*. Macmillan, 1998.
- Dunn, Richard J., editor. *Wuthering Heights* by Emily Brontë. Norton Critical Edition, 4th ed., W. W. Norton & Company, 2003.
- Eve, Martin Paul. *The Digital Humanities and Literary Studies*. Oxford UP, 2022.
- Klaussner, Carmen, and Carl Vogel. "Stylochronometry: Timeline Prediction in Stylometric Analysis." *SGLAI Conferences*, 2015.
- MacKay, Marina. *The Cambridge Introduction to the Novel*. Cambridge UP, 2011.
- McCarthy, Rachel, and James O'Sullivan. "Who wrote *Wuthering Heights*?" *Digital Scholarship in the Humanities*, vol. 36, no. 2, 2020, pp. 383–91, <https://doi.org/10.1093/llc/fqaa031>.
- McIntyre, Dan, and Brian Walker. "Combining Corpus Linguistics and Stylistics." *Corpus Stylistics: Theory and Practice*, Edinburgh UP., 2019, pp. 1–24.
- McIntyre, Dan, and Brian Walker. "Using Corpus Linguistics to Explore the Language of Poetry: A Stylometric Approach to Yeats' Poems." *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keefe and Michael McCarthy, Routledge, Taylor & Francis Group, Abingdon, Oxon, 2022, pp. 499–516.

- Moretti, Franco. "History of the Novel, Theory of the Novel." *NOVEL: A Forum on Fiction*, vol. 43, no. 1, 2010, pp. 1–10. *JSTOR*, <http://www.jstor.org/stable/27764362>. Accessed 2 Oct. 2023.
- Moretti, Franco. *Distant Reading*. Verso, 2015.
- Ohl, Jean-Pierre. *Les Brontë*. Gallimard, 2019.
- Sinclair, Stéfán and Geoffrey Rockwell. "Résumé." *Voyant Tools*, 2024, <https://voyant-tools.org/?view=Summary&limit=59&corpus=7a1749495cb68faea8812c945750781a>.
- Sinclair, Stéfán and Geoffrey Rockwell. "Tendances." *Voyant Tools*, 2024, https://voyant-tools.org/?view=Trends&query=i_prp&query=the_dt&query=and_cc&corpus=7a1749495cb68faea8812c945750781a.
- Smith, Margaret, editor. *The Letters of Charlotte Brontë with a Selection of Letters by Family and Friends. 1848-1851*. Oxford UP, 1995.
- Stoneman, Patsy. "The Brontë Myth." *The Cambridge Companion to the Brontës*, edited by Heather Glen, Cambridge UP, 2009, pp. 214–41.
- Trapenard, Augustin. "Authorizing Emily: The Production of an Author-Function in Charlotte Brontë's 1850 Edition of *Wuthering Heights* and *Agnes Grey*." *Études Anglaises*, vol. 60, no. 1, Jan-Mar. 2007, pp. 15-28. *ProQuest Central*, <https://search.proquest.com/docview/205859964?accountid=16746>.