Digital Scholarship in the Humanities, 2023, **38**, 750–765
https://doi.org/10.1093/llc/fqac061
Advance access publication 25 October 2022
Full Paper

OXFORD

Did Mary Shelley write *Frankenstein*? A stylometric analysis

Lee Suddaby[1] , Gordon J. Ross[1]*

[1]School of Mathematics, University of Edinburgh, Edinburgh, UK

*Correspondence: Gordon J. Ross, School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK. E-mail: gordon.ross@ed.ac.uk

Abstract

The novel *Frankenstein* was published anonymously in 1818 and was first credited to Mary Shelley in a French translation of 1821. Since its publication, several claims—both contemporaneous and recent—have been made suggesting that *Frankenstein* was written by Mary's husband, Percy Bysshe Shelley. We review the background of this controversy and then apply modern techniques from computational stylometry to determine who the true author is. Based on our analysis, we find extremely substantial evidence that Mary Shelley is indeed the true author of *Frankenstein*, and it is very improbable that Percy Shelley played a heavy role in composing the text. While our finding confirms mainstream scholarly opinion regarding *Frankenstein*, our analysis is the first application of stylometric techniques to this question and provides strong objective grounds for favouring Mary Shelley by freeing the question from some of the politics which have traditionally accompanied it.

1. Background

The novel '*Frankenstein, or The Modern Prometheus*' was published anonymously on 1 January 1818, and tells the story of the scientist Victor Frankenstein, who creates and animates his monster, and suffers terrible consequences as a result (Shelley, 1818). A French edition of the novel was published in 1821 and was the first to credit Mary Shelley as the author. Subsequent second and third English editions were published in 1823 and 1831 (Robinson, 1996). The 1831 edition was significantly revised from the original 1818 edition and is the edition that is most widely read today (Shelley and Groom, 2019, pp. li–lv).

The accepted story of *Frankenstein*'s authorship comes from the introduction to the 1831 edition (Shelley, 1831; Shelley and Groom, 2019) and is as follows: In June 1816, a ghost-story writing contest took place at Lord Byron's residence by Lake Geneva in Switzerland. The contest comprised Mary Shelley, her husband Percy Bysshe Shelley, Lord Byron, and his physician, John William Polidori. After several days of being unable to think of a story, Mary states the story of *Frankenstein* came to her in a 'waking dream', and from this initial story, she was encouraged by Percy Shelley to develop it into a full-fledged novel—her first novel—which she eventually published anonymously.

Publishing anonymously or pseudo-anonymously (e.g. under a pen name) was common for both male and female writers in the late 18th and early 19th centuries, and more so for novels compared to periodicals.[1] This was the process followed in the publication of novels by Jane Austen (Irvine, 2005, pp. 10–18), Frances Burney (Spencer, 2007, p. 24), and even Mary and Percy Shelley themselves—the 1817 collaboration *History of a Six Weeks' Tour* was published anonymously, as was Percy's 1811 novel *St. Irvyne* (Eilenberg, 2003, pp. 171–172).[2] So while the initial anonymous publication of *Frankenstein* is consistent with Mary Shelley's authorship, it has allowed for speculation that *Frankenstein* may have been written by a different author.

In March 1818, Sir Walter Scott claimed *Frankenstein* may have been written by Percy Shelley (Scott, 1818). Following the publication of Mary Shelley's second novel, *Valperga*, an anonymous 1824 review of the novel in *Knight's Quarterly* noted substantial differences between the quality of *Frankenstein* and its successor, claiming 'there is not the slightest trace of the same hand',

and suggested Percy Shelley wrote *Frankenstein* and Mary was only responsible for *Valperga* (*Frankenstein*, 1824).

These suppositions were all denied by Percy Shelley—from the start, he claimed to only have been the editor of *Frankenstein* rather than its author. However, it later transpired that he had written the preface of the novel as if he were Mary, as well as the poem *Mutability*, which was included in the novel uncredited and likely a factor in leading people to believe he was the author of the entire novel.

More recently, John Lauritsen's book *The Man Who Wrote Frankenstein* promotes the theory of Percy Shelley's authorship (Lauritsen, 2007), as do other writings from the same author (Lauritsen, 2018a,b). Others have made similar claims (Zimmerman, 1998; De Hart, 2013, Jones, unpublished). Although these claims were not—and still are not—taken seriously by mainstream scholars of Shelley's work and Romanticism, they did receive substantial media attention, with Lauritsen's book receiving the largest share of this attention.

Like the review in *Knight's Quarterly Review* published two centuries earlier, Lauritsen believes there is a substantial difference in quality between *Frankenstein* and other works of Mary Shelley—particularly *Valperga* and *The Last Man* (Lauritsen, 2018b).

Textual evidence and analysis is not the only avenue that may be—and has been—investigated in determining the true author of *Frankenstein*. One can also consider historical evidence. Indeed, it is the analysis of handwriting in the original manuscript which has fuelled speculation over Percy Shelley's contributions (Wu, 2015).

In 1996, Charles E. Robinson produced transcriptions of the original manuscript, allowing the novel to be seen in its earliest draft, and for the authorship of each part of the text to be identified down to the level of individual words. Contrary to previous analysis which concluded that Percy Shelley contributed a 'thousand or so' words to the rough draft (Murray, 1978), in the introduction to his book, *The Original Frankenstein*, Robinson concludes:

> [Percy] contributed at least 4,000 to 5,000 words to this 72,000 word novel. Despite the number of Percy's words, the novel was conceived and mainly written by Mary Shelley, as attested not only by others in their circle (e.g. Byron, Godwin, Claire and Charles Clairmont, Leigh Hunt) but by the nature of the manuscript evidence in the surviving pages of the Draft. (Robinson, 2008, p. 25)

This so-called 'handwriting argument' assumes that everything in the manuscript written in Mary Shelley's hand was originally composed by her. However, while there is no real evidence to suggest that the words in Percy Shelley's handwriting are likely to have been composed by someone other than himself,[3] there exist manuscripts in Mary Shelley's handwriting of works composed by Percy—for example, his 1820 lyrical drama *Prometheus Unbound* (Fraistat, 1991), and the poems *The Mask of Anarchy*, *The Witch of Atlas*, *The Cenci*, and *The Sensitive Plant* (Reiman and Powers, 1977, pp. 210, 236, 301, 348). Therefore, a possible theory is that the same arrangement was used in the composition of *Frankenstein*. However, the fact that Mary is known to have transcribed manuscripts for her husband is by no means indicative of the same arrangement having been used in the composition of *Frankenstein*—a view shared by mainstream scholars.

At the very least, analysis of handwriting provides us with a lower bound to the extent of Percy Shelley's contribution, but the question as to whether he could have been the intellectual force behind the novel as a whole remains somewhat open. Was his contribution to the novel merely editorial—as Robinson (1996) puts it, as 'an able midwife who helped his wife bring her monster to life'? Or was he, as Rieger (1974, p. xviii) suggests, as much as a 'minor collaborator'? Or are the fringe theories correct in claiming Percy Shelley was the real author, and *Frankenstein* his brainchild, rather than his wife's?

This is precisely the question which we seek to answer in this investigation, although we will use a different approach from the above scholars. Rather than concerning ourselves with the analysis of handwriting or literary themes or examination of the bibliographical evidence, we will instead apply modern techniques of stylometry. Doing this should yield an analysis that is not reliant on previous traditional attribution studies, which will be valuable if the authorship of *Frankenstein* continues to be contested. Of course, even if one assumes *Frankenstein* is Mary's first novel, one expects to find stylistic differences between this and the later novels as she matured and her literary career developed, but this should still make for a worthwhile analysis.

To the best of our knowledge, the only previous applications of stylometry to Mary Shelley's work are Rybicki (2016), a study on authorial gender signals, which analyses whether her early novels, including *Frankenstein*, are stylistically similar to the work of other female novelists; and O'Sullivan (2022), where *Frankenstein* is found to be closest in style to other selected works of Mary Shelley and her parents, William Godwin and Mary Wollstonecraft. However, neither of these works are explicitly concerned with the authorship of *Frankenstein*, nor do they compare it to works written by Percy Shelley.

2. Basic elements of stylometry

Stylometry involves using statistical methods to study linguistic style, often to determine the authorship of a

text of interest. For this purpose, 'style' refers to elements such as word choice, frequency, combinations, and sentence length, taken together to provide a unique profile for an author. General reviews of the subject and its methods are given in Juola (2006), Koppel *et al.* (2009), Savoy (2020), and Stamatatos (2009).

Two widely studied tasks in stylometry are 'Authorship Attribution', which seeks to determine the author of a given text from a pool of potential authors, and 'Authorship Verification', where we are given a text and a single author and wish to determine (i.e. verify) whether the text was written by the said author.

A seminal example of authorship attribution comes from Mosteller and Wallace (1963), who determined the authorship of twelve disputed *Federalist Papers* between James Madison and Alexander Hamilton. Other examples include identifying J.K. Rowling as the author of *The Cuckoo's Calling*, published under the pseudonym Robert Galbraith (Juola, 2015); an authorship study on thirty-two prose pieces and ten poems to see whether they can be attributed to Edgar Allan Poe (Schöberlein, 2017); determining whether the *Book of Mormon* was written by a single author or a collaboration (Holmes, 1992); and a recent investigation assessing the claim that *Wuthering Heights* was not written by Emily Brontë, rather her brother Branwell (McCarthy and O'Sullivan, 2020).

2.1 Function words

Since the analysis of Mosteller and Wallace (1963), function words have played a key role in authorship studies. These are considered the building blocks of language and are largely context-free, in the sense that they have essentially no meaning on their own. Function words include articles, prepositions, auxiliary verbs, conjunctions, and pronouns—for example, 'and', 'the', 'of', and so on. An overview of the suitability and usefulness of the function words in stylometry is given in Argamon and Levitan (2005), Miranda García and Calle Martín (2006), and Kestemont (2014).

While certain words should always be considered in a list of function words, such as 'the', 'a', 'and', 'it', etc., there is no single authoritative list of function words that is always used in stylometry. For example, in their analysis of the Federalist Papers, Mosteller and Wallace (1963) choose a list of seventy particular function words. Alternatively, it is common to instead take these words to be the most frequently occurring words in a literary corpus, such as the 100 or 200 most frequent words (MFWs). Once we have a list of function words (or frequent words), variations in their frequencies can be analysed to make comparisons of style between various authors and texts and perform authorship identification.

2.2 Character n-grams

An alternative method for determining the authorship of texts is to focus on character n-grams rather than function words (Kjell *et al.*, 1994; Kešelj *et al.*, 2003). A character n-gram (henceforth simply an "n-gram") is a sequence of n consecutive characters. For example, in the sentence, 'The quick brown fox jumps over the lazy dog', the 3-grams of characters are 'the', 'he ', 'e q', 'qu', etc. In stylometry, to avoid the choice of character n-gram containing too much context-specific information, a small n is suitable, for example, 2 or 3 (Stamatatos, 2009), though the actual process of modelling and analysis does not depend on the value of n. Just as for frequent words, variations in the frequencies of n-grams may be analysed to determine authorship.

One of the benefits of n-grams over function words is that, by Zipf's Law (Zipf, 1949), the number of unique words used by an author will be fairly small, so it would be difficult to scale up to using thousands of frequent words without encoding large amounts of context-specific information. As they can span word boundaries, there are significantly more common n-grams, so the number of n-grams used can be much greater than is possible with individual words.

In Grieve (2007), n-grams are found to be some of the most accurate stylistic markers of those tested in author attribution, with 2- and 3-grams being the most accurate. Other comparisons of methods have also found character n-grams to be among the most accurate style markers (Koppel *et al.*, 2009; Stamatatos, 2009), and 3-grams the most accurate type of n-gram (Sapkota *et al.*, 2015). It was found in Stamatatos (2013) that character n-grams were more reliable than those based on frequent words when comparing works from different genres and/or topics, implying the n-grams are more independent of context than word tokens.

3. Description of corpus

Our goal is to perform authorship attribution on *Frankenstein* by comparing its function word/n-gram frequencies to those of several candidate authors. The full corpus we use for this is given in Appendix A and contains works by authors such as Percy Shelley and William Godwin (Mary Shelley's father). The inclusion of Godwin is a particularly important choice—after her mother (Mary Wollstonecraft) died during childbirth, Mary's father was responsible for her upbringing and somewhat informal education. Godwin's radical views and writings had a substantial impact on his daughter's political views and literary career (Clemit, 1999), and his involvement is also felt in his editing of some of Mary's later novels such as *Valperga* (Shelley

and Rossington, 2000, pp. xiv–xvi). We now discuss the full corpus in more detail, to justify the inclusions which we have made.

In order to construct a representative authorial signature for Mary Shelley, we include the novels which she wrote during her life: *Valperga*, *The Last Man*, *The Fortunes of Perkin Warbeck*, *Lodore*, *Falkner*, and *Mathilda*. These works are particularly important since if we want to determine whether Mary Shelley wrote *Frankenstein*, we must compare its style to the other books which we know she wrote. If *Frankenstein* is stylistically similar to these undisputed works, we have found evidence that they share an author, that is, that Mary Shelley wrote *Frankenstein* too.

For our main analysis, we will use the 1818 first edition of *Frankenstein*, as substantive changes were made to the second and third editions in 1823 and 1831 (see Murray, 1981; Shelley and Groom, 2019; Appendix B).[4] In particular, many of the changes were made by Godwin; therefore, if we wish to work with the novel in its most unaltered form, it is the 1818 edition which we prefer.

For the Percy Shelley texts in our corpus, we include his two early novels, *Zastrozzi* and *St. Irvyne*. A set of his prose essays was considered for inclusion,[5] but these were found via clustering and principal component analysis (PCA) to be stylistically very different from the novels to the extent that considering them as part of his authorial profile led to a loss of accuracy when testing attribution methods. Hence the essays were omitted. This is consistent with the general belief in stylometry that prose and non-prose and fiction and non-fiction writings can have quite different authorial signatures even when produced by the same author.

Most of Percy Shelley's written work is poetry. Unfortunately, this might be unsuitable in characterizing authorial style in prose fiction, as noted for example in McCarthy and O'Sullivan (2020, note 4). Therefore, while we have included Percy Shelley's poetry in the corpus (see Appendix A for details on the precise editions), it is categorized differently, so we form two distinct authorial profiles for Percy Shelley: one for poetry and one for prose works.[6]

Finally for Mary and Percy Shelley, we included the collaborative piece *History of a Six Weeks' Tour* but considered it as a separate authorial profile independent of those for Mary and Percy as individuals.

For comparison, we include in the corpus some works of related authors: Mary Shelley's parents, William Godwin and Mary Wollstonecraft; the 'Godwinian' Charles Brockden Brown; Thomas Love Peacock, who was a close friend of Percy Shelley; Sir Walter Scott, an important figure in the literary world when the Shelleys were writing; and John Polidori, associated with the Romanic movement and author of *The Vampyre*, produced during the same ghost-story writing contest from which *Frankenstein* is said to have been conceived. It is possible to see from Mary's journals[7] that she had read some works of certain authors included here—particularly Brown, Godwin, and Scott.

Other than Godwin's *Fleetwood*, all the texts in our corpus are obtained from Project Gutenberg or Project Gutenberg Australia. Some general changes were made for clean-up: Project Gutenberg licenses, title pages, contents, and other pre-amble were removed, punctuation was removed (except apostrophes indicating contractions and possession, which were standardized), the underscores used to indicate italic text are removed, as are epigraphs, quotations of other works, introductions, prefaces (including those written by the author), and textual notes. Notably for *Frankenstein*, the text contains some poems from other authors—such as Percy Shelley's *Mutability* and Wordsworth's *Tintern Abbey*—which were removed.

The R programming language version 4.0.2 was used for our analysis (R Core Team, 2020) and the `stylo` package was used for access to certain stylometry and machine learning methods (Eder *et al.*, 2016). We implemented some other algorithms—in particular, Delta—manually. For the analysis with `stylo`, we used both individual words and character 3-grams and removed pronouns from the corpus, as these can be too particular to a genre or topic (Pennebaker, 2011).

Finally, culling is performed on the corpus when preparing the counts of words and 3-grams. Specifically, we delete any words/3-grams that do not appear in at least 70% of the texts in the corpus. This is common practice in stylometry and is intended to avoid the pathological case where a novel (or handful of novels) contains context-specific words such as a character's name which occur so often they become one of the most frequent words across the entire corpus.

3.1 Clustering

Before carrying out a more formal authorship analysis, it is helpful to visualize the relationship between the texts in this corpus. For this purpose, we use hierarchical clustering, which represents the corpus as a dendrogram by repeatedly clustering together texts which are most alike (Eder, 2017). The process works as follows: we measure the 'distance' between each pair of texts in the corpus, defined as the squared distance between the (normalized) counts of MFWs or most frequent n-grams. Then, the two texts with the smallest 'distance' are joined into a single cluster node. We then estimate the distance from this new node to all other nodes, and repeat the process, until all texts have been linked. The result is a single cluster, as shown in Figs 1 and 2. The

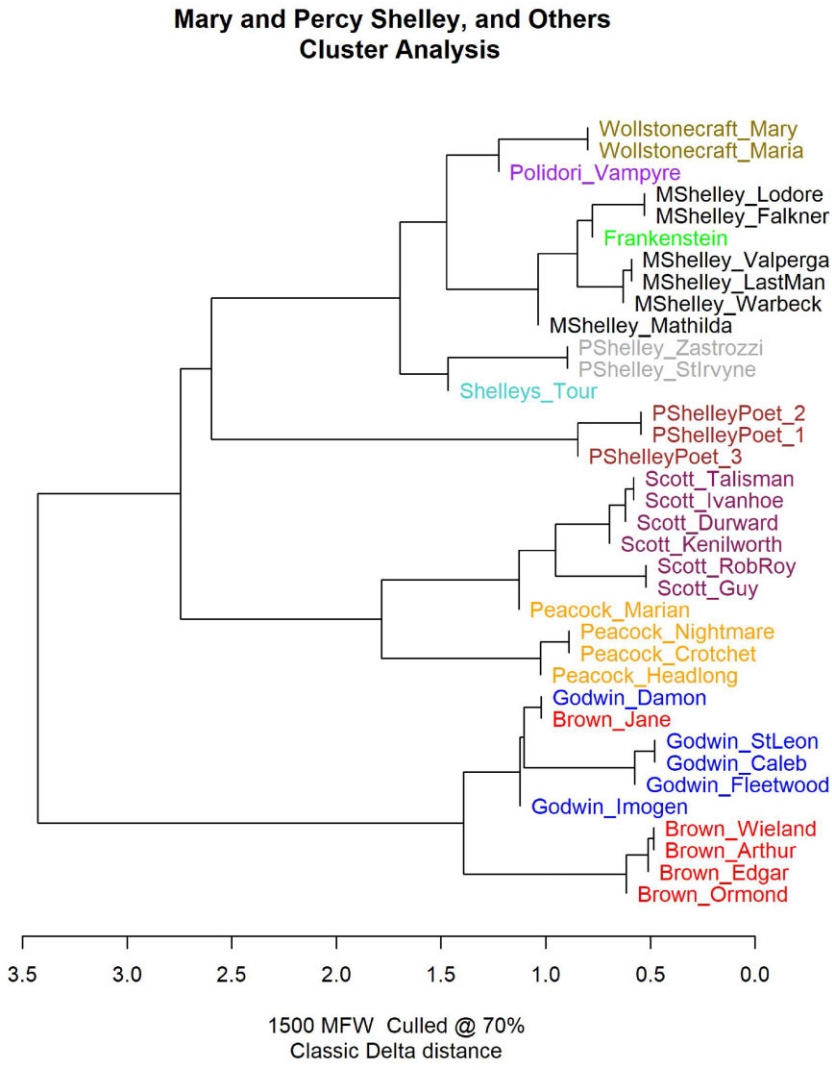

Figure 1. Hierarchical clustering of the corpus with 70% culling and using 1,500 most frequent 3-grams.

black lines link texts which are 'close' together under the above distance metric.

We see that cluster analysis generally—and marginally more so in the case of MFWs—puts the works of each author together as 'sub-clusters', suggesting that it is correctly identifying authorial signatures. There are a few exceptions: Charles Brockden Brown's *Jane Talbot* is associated with Godwin's works more than Brown's other novels, and Peacock's *Maid Marian* often occupies the same branch as Sir Walter Scott's works. It can be seen that *Frankenstein* is clustered within the cluster containing the other works of Mary Shelley, giving preliminary evidence that she is the true author.

Analysing the exact distances from the distance table produced, we find that on an individual text level, *Frankenstein* is closest to *Valperga*—Mary Shelley's novel which immediately followed *Frankenstein* in terms of publication date (*Mathilda* being composed but not published before *Valperga*). This links to the idea of stylistic drift, where works by a given author that are written closer together in time may also be closer in terms of style (Ross, 2020).

3.2 PCA

We next consider another approach to visualizing the corpus, as a prelude to a more formal study. PCA is a dimensionality reduction method, designed to transform many potentially correlated variables into a smaller set of uncorrelated variables, the *principal components* (Binongo and Smith, 1999; Jolliffe, 2002). With this process, we hope much of the variation in the original data can be accounted for in these principal

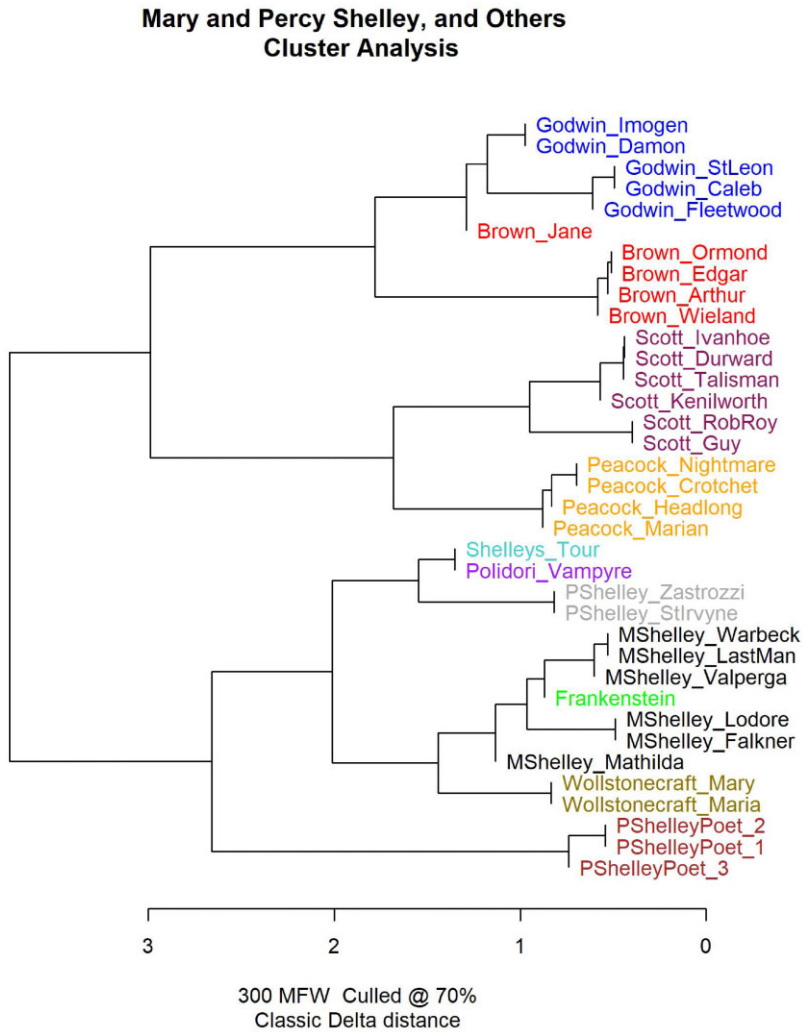

Figure 2. Hierarchical clustering of the corpus with 70% culling and 300 MFWs.

components (PCs). It is one approach to performing Multidimensional Scaling (Borg and Groenen, 2005), that is, visualizing a high-dimensional object such as a literary corpus in two dimensions, by projecting it down onto its first two PCs.

In the case of stylometry, the high-dimensional variables are the frequencies of the MFWs or 3-grams, recorded for each text in the corpus. We can use it to visualize in two dimensions some of the stylistic differences encoded in the varying frequencies of the hundreds of features considered.

To determine if PCA allows us to see significant differences between the first edition and the 1831 third edition of *Frankenstein*, we include both in this part of the analysis. For the sake of readability (in the visualization), only a subset of works from the above corpus is considered. This represents the most likely candidate authors—William Godwin, Mary Shelley, Percy Shelley, and the collaborative work between the Shelleys.

The results of the PCA can be seen in Figs 3 and 4. This plots the various novels in the corpus in two dimensions, which is a projection onto the first two PCs. The distance between various texts in this plot is based on the distance between their most frequently occurring words/3-grams, with more similar texts being shown as closer together. Here, we can see a pattern similar to the results of the clustering—the works of each author are closer together geometrically.

When using MFWs *Frankenstein* is clearly closer to Mary Shelley's other works than to Percy Shelley's or anyone else's. There is more ambiguity when 3-grams are considered, and while *Frankenstein* is not at all close to Percy Shelley's work, it is not clear whether it

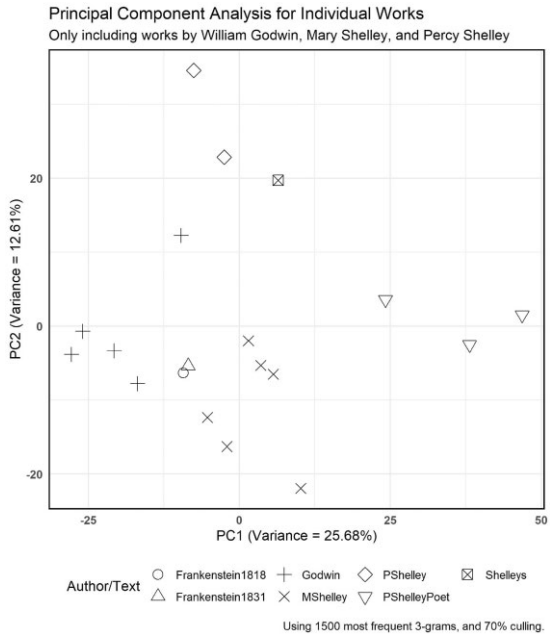

Figure 3. PCA with *Frankenstein* and other works of Mary Shelley, Percy Shelley, and William Godwin, using the 1500 most frequent 3-grams with 70% culling.

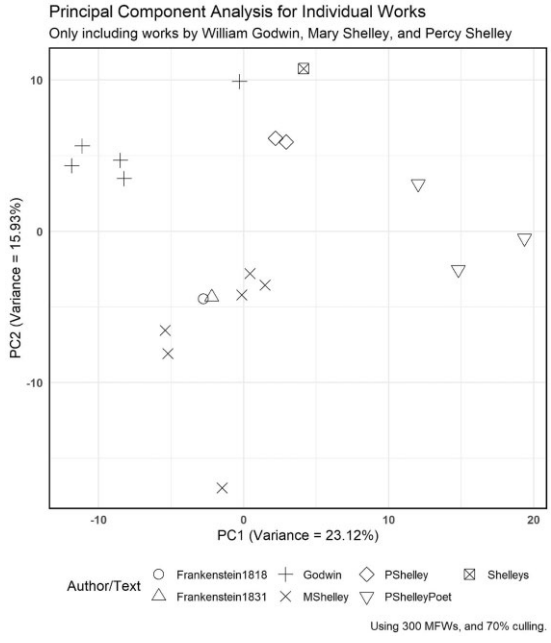

Figure 4. PCA with *Frankenstein* and other works of Mary Shelley, Percy Shelley, and William Godwin, using the 300 MFWs with 70% culling.

is closer to Mary Shelley's novels, or William Godwin's. Therefore, while this may provide some evidence against Percy Shelley's authorship, the evidence regarding Mary's authorship is not clear-cut.[8]

Furthermore, there appears to be little detectable difference between the two editions of *Frankenstein*, so even if notable differences exist between the texts, these differences are in theme and philosophy rather than quantitative style as measured by frequency of frequent words and 3-grams.

A potential concern is that since *Frankenstein* is written in epistolary form with three distinct narrators—Captain Walton, Victor Frankenstein, and The Creature—it is possible each narrator has a distinct writing style and the style associated with *Frankenstein* is the average of three individual stylistic profiles. To check this, PCA was performed in the same way as before, with the parts of each narrator being treated as individual texts. We find the three data points cluster very close to the novel as a whole, closer than any other text in the corpus, therefore this epistolary form appears to not be a concern.

4. Methodology

Having now completed the preliminary visualization of the corpus, we move on to performing a more formal authorship attribution study. Our goal is to determine which of the authors in our corpus is the most likely author of *Frankenstein*. For this purpose, we will use two standard methods of authorship attribution which are common in the stylometry literature—Burrows' Delta, and Support Vector Machines (SVMs).

4.1 Burrows' delta

The 'Delta' method introduced by Burrows (2002) is a standard authorship attribution method used for attributing texts in both English and other languages (Jockers *et al.*, 2008; Miranda García and Calle Martín, 2012; Schöberlein, 2017; Savoy, 2018), as a similarity metric in hierarchical cluster analysis (Eder, 2017), and in the 'Rolling Delta' method (Burrows, 2010; Hoover, 2012; Rybicki *et al.*, 2014). It is essentially a type of K-nearest neighbours classifier which measures the 'distance' between the text which we wish to analyse and each author from a set of candidate authors, based on their use of MFWs (or n-grams). More formally, suppose our features consist of a set of M most frequent words or n-grams. For a given candidate author a, let A_a be a length M vector which constitutes the profile of that author, which is formed by computing the (normalized) proportion of frequent words used by the author in the corpus texts (so that the first element of A_a is the proportion of times the most frequent word appeared in the set of corpus texts written by author a, etc.). Next, let C denote the text to be classified (e.g. *Frankenstein*) where again C is a length M vector of frequent word proportions. Finally,

let $d()$ be a distance function such as Manhattan, Euclidean or Cosine distance. Then, delta is given by:

$$\Delta(A_a, C) = d(A_a, C)$$

The author which produces the minimum distance score is then identified as the most likely author of the text. Certain modifications of this algorithm are introduced in Hoover (2004a); and Argamon (2008). The main modifications used in our investigation are the removal of pronouns from the set of frequent words (Hoover, 2004b) and culling, as was described earlier.[9]

4.2 SVMs

SVMs are supervised learning models commonly used in stylometry. Given an M-dimensional feature set, it aims to find a hyperplane (border) in M-dimensional space which classifies data points from two distinct classes (Joachims, 1998; Diederich *et al.*, 2003). Given that multiple possible hyperplanes are likely to exist, the standard SVM approach chooses the one which maximizes a type of distance (the 'margin') between the classes (Cortes and Vapnik, 1995). If it is not possible to perfectly separate the classes, a loss function may be used to penalize data points that lie on the wrong side of the hyperplane, and non-linear kernels can be used to project the data into a space where such a separation is possible.

In applications to stylometry, the data points used in the SVM algorithm will be the corpus texts and the unknown text to be classified, all represented in feature space as vectors of function word proportions, and the classes represent the candidate authors. With more than two candidate authors, there are two commonly used approaches to convert SVMs into a multi-class classifier. The first is one-versus-all, which pits one class/author against all others; the second is one-versus-one, where binary comparisons are made between every possible pair of classes. In both cases, the multiple binary classification results are aggregated in a kind of 'democratic process' to return a single class result. It is the one-versus-one method that is implemented in the `stylo` R package for authorship attribution, which we use in our analysis.

5. Results

5.1 Delta

We apply Delta using 100, 200, 500, and 1,000 most frequent words; 200, 500, 1,000, and 2,000 most frequent 3-grams; and culling at the 70% level. The distance functions used are the (mean) Manhattan distance,

$$\Delta_M(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_i|,$$

and the cosine distance,

$$\Delta_C(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}.$$

These results are shown in Tables 1–4.[10] In all cases except 200 3-grams and cosine distance, the algorithm attributes *Frankenstein* to Mary Shelley, often by a considerable margin when one compares the raw distance scores in the tables (e.g. 0.604 for Mary Shelley compared to 0.844 for the second-nearest author William Godwin and 1.019 for Percy Shelley's prose profile using 100 MFWs and the Manhattan distance). Therefore, in the language of Burrows' original paper, *Frankenstein* is a lot 'less unlike' Mary Shelley's other novels than any other authorial profiles we have created (Burrows, 2002).

One concern was that, despite culling being performed, some of the words that featured in the lists of most frequent words—particularly when scaling up to 500 and 1,000 MFWs—were not context free. Therefore, we applied the algorithm using a set of 200 frequent context-free function words, shown in Appendix B. This also attributes *Frankenstein* to Mary Shelley, with a score of 0.768 for her under the Manhattan distance compared to 0.850 for Godwin and 1.071 for Percy Shelley (Prose). These distances are very similar when using the cosine distance: 0.717, 0.837, and 1.053 for Mary Shelley, Godwin, and Percy Shelley, respectively.

In sum, this attribution is generally independent of which distance metric is used and the number of frequent words or 3-grams employed, which demonstrates robustness (i.e. the results are not sensitive to small variations in how we set up the analysis), as does the finding that the attribution to Mary Shelley was unchanged with the presence or absence of culling. Within the tables for the two distance metrics, we can see the general pattern does not significantly vary as the number of features used varies. For example, with most frequent words and Manhattan distances, Mary Shelley is always ranked first; Godwin, Brown, and Wollstonecraft rank near the top; Percy Shelley and Scott are central; and Peacock, Polidori, and Shelleys are towards the bottom of the rankings.

5.2 SVMs

We now perform the same authorship attribution using SVM, using the SVM implementation from the `stylo`

Table 1. Delta scores for classification of *Frankenstein* with 70% culling and Manhattan distance (Classic Delta)

	100 MFWs			200 MFWs			500 MFWs			1000 MFWs	
Rank	Author	Distance	Rank	Author	Dist.	Rank	Author	Dist.	Rank	Author	Dist.
1	Mary Shelley	0.604	1	M. Shelley	0.685	1	M. Shelley	0.733	1	M. Shelley	0.803
2	William Godwin	0.844	2	Godwin	0.871	2	Godwin	0.941	2	Godwin	0.940
3	Charles Brockden Brown	0.926	3	Wollstonecraft	0.932	3	Wollstonecraft	0.977	3	Wollstonecraft	1.013
4	Walter Scott	0.944	4	Brown	0.936	4	Brown	0.999	4	Brown	1.030
5	Mary Wollstonecraft	0.950	5	P. Shelley (Prose)	0.985	5	P. Shelley (Prose)	1.056	5	Scott	1.126
6	Percy Shelly (Prose)	1.019	6	Scott	1.087	6	Scott	1.091	6	P. Shelley (Prose)	1.137
7	Thomas Love Peacock	1.108	7	P. Shelley (Poetry)	1.102	7	Peacock	1.129	7	Peacock	1.170
8	Percy Shelley (Poetry)	1.112	8	Peacock	1.165	8	P. Shelley (Poetry)	1.173	8	P. Shelley (Poetry)	1.245
9	John Polidori	1.123	9	Polidori	1.199	9	Polidori	1.225	9	Shelleys	1.276
10	Shelleys (Collaboration)	1.374	10	Shelleys	1.256	10	Shelleys	1.247	10	Polidori	1.296

Lower numbers indicate that the author is more likely to have written the text.

Table 2. Delta Scores for classification of *Frankenstein* with 70% culling and cosine distance

	100 MFWs			200 MFWs			500 MFWs			1,000 3-grams	
Rank	Author	Distance	Rank	Author	Dist.	Rank	Author	Dist.	Rank	Author	Dist.
1	Mary Shelley	0.613	1	M. Shelley	0.686	1	M. Shelley	0.612	1	M. Shelley	0.663
2	William Godwin	0.850	2	Brown	0.885	2	Wollstonecraft	0.886	2	Wollstonecraft	0.866
3	Charles Brockden Brown	0.857	3	Godwin	0.903	3	P. Shelley (Prose)	0.944	3	Godwin	0.884
4	Mary Wollstonecraft	0.921	4	Wollstonecraft	0.904	4	Brown	0.981	4	P. Shelley (Prose)	0.967
5	Percy Shelley (Prose)	0.961	5	P. Shelley (Prose)	0.956	5	Godwin	0.990	5	Brown	0.978
6	Percy Shelley (Poetry)	1.002	6	P. Shelley (Poetry)	0.964	6	P. Shelley (Poetry)	0.999	6	Polidori	1.029
7	Walter Scott	1.065	7	Shelleys	1.042	7	Shelleys	1.001	7	Shelleys	1.033
8	John Polidori	1.083	8	Polidori	1.221	8	Polidori	1.041	8	P. Shelley (Poetry)	1.057
9	Thomas Love Peacock	1.159	9	Scott	1.223	9	Peacock	1.239	9	Scott	1.185
10	Shelleys (Collaboration)	1.260	10	Peacock	1.221	10	Scott	1.240	10	Peacock	1.235

Lower numbers indicate that the author is more likely to have written the text.

Table 3. Delta Scores for classification of *Frankenstein* with 70% culling and Manhattan distance (Classic Delta)

	200 3-grams			500 3-grams			1,000 3-grams			2,000 3-grams	
Rank	Author	Distance	Rank	Author	Dist.	Rank	Author	Dist.	Rank	Author	Dist.
1	Mary Shelley	0.648	1	M. Shelley	0.726	1	M. Shelley	0.759	1	M. Shelley	0.829
2	William Godwin	0.820	2	Godwin	0.851	2	Godwin	0.856	2	Godwin	0.914
3	Charles Brockden Brown	0.872	3	Wollstonecraft	0.938	3	Brown	0.944	3	Brown	1.004
4	Walter Scott	0.878	4	Brown	0.950	4	Wollstonecraft	0.961	4	Wollstonecraft	1.028
5	Mary Wollstonecraft	0.892	5	Scott	1.003	5	Scott	0.989	5	Scott	1.088
6	Thomas Love Peacock	1.008	6	P. Shelley (Prose)	1.090	6	P. Shelley (Prose)	1.112	6	P. Shelley (Prose)	1.165
7	Percy Shelly (Prose)	1.022	7	Peacock	1.092	7	Peacock	1.112	7	Peacock	1.174
8	John Polidori	1.051	8	Polidori	1.201	8	Polidori	1.232	8	Shelleys	1.315
9	Shelleys (Collaboration)	1.185	9	Shelleys	1.206	9	Shelleys	1.283	9	Polidori	1.327
10	Percy Shelley (Poetry)	1.364	10	P. Shelley (Poetry)	1.393	10	P. Shelley (Poetry)	1.356	10	P. Shelley (Poetry)	1.329

Lower numbers indicate that the author is more likely to have written the text.

R package. Classification of *Frankenstein* is performed in several ways to verify robustness. This is done with two different training sets—one consisting of the entire corpus, and one which reduces the classification problem to a pure binary classification by only considering the works of Mary Shelley and Percy Shelley (Prose).

Table 4. Delta Scores for classification of *Frankenstein* with 70% culling and cosine distance

200 3-grams			500 3-grams			1,000 3-grams			2,000 3-grams		
Rank	Author	Distance	Rank	Author	Dist.	Rank	Author	Dist.	Rank	Author	Dist.
1	Charles Brockden Brown	0.781	1	M. Shelley	0.703	1	M. Shelley	0.699	1	M. Shelley	0.856
2	William Godwin	0.783	2	Godwin	0.772	2	Godwin	0.756	2	Godwin	0.903
3	Mary Shelley	0.788	3	Brown	0.822	3	Brown	0.832	3	Brown	0.930
4	Mary Wollstonecraft	0.876	4	Wollstonecraft	0.882	4	Wollstonecraft	0.875	4	Wollstonecraft	0.965
5	Shelleys (Collaboration)	1.000	5	P. Shelley (Prose)	0.978	5	P. Shelley (Prose)	1.002	5	P. Shelley (Prose)	1.003
6	Thomas Love Peacock	1.105	6	Shelleys	1.029	6	Shelleys	1.077	6	P. Shelley (Poetry)	1.007
7	Percy Shelly (Prose)	1.114	7	P. Shelley (Poetry)	1.122	7	P. Shelley (Poetry)	1.109	7	Shelleys	1.017
8	Percy Shelley (Poetry)	1.114	8	Polidori	1.143	8	Scott	1.140	8	Polidori	1.069
9	Walter Scott	1.134	9	Peacock	1.181	9	Polidori	1.152	9	Scott	1.072
10	John Polidori	1.158	10	Scott	1.207	10	Peacock	1.165	10	Peacock	1.108

Lower numbers indicate that the author is more likely to have written the text.

Recall that `stylo` uses a one-versus-one approach for the multi-class problem.

To demonstrate that the attribution of *Frankenstein* is not sensitive to the parameters chosen for the model, we perform the analysis on both MFWs and 3-grams, with and without culling at the 70% level, and with a variety of MFWs. For 3-grams, we used $200, 300, \ldots, 2000$ features, and $50, 100, \ldots, 500$ for MFWs. This gives a total of 116 variations.

In all cases, *Frankenstein* is attributed to Mary Shelley, both over Percy Shelley and over any of the other authors in the corpus. Combined with the above results from Delta, this provides extremely strong evidence that Mary Shelley is indeed the author of *Frankenstein*.

6. Further robustness checks

The above analysis shows that both Delta and SVMs attributed *Frankenstein* to Mary Shelley. However, for this result to be convincing, we should verify that Delta/SVMs can correctly identify the author of the 19th-century literary texts. Although there is a large amount of existing evidence showing the general effectiveness of these algorithms in stylometry, we will now demonstrate for completeness that these are also effective on the corpus we have constructed, which increases the strength of our results.

6.1 Cross-validation

We use Leave-One-Out Cross-Validation to check Delta can correctly identify the author of literary texts. For this purpose, we attempt to classify each text in the corpus one by one, by removing it from the corpus, and treating it as if it were an unknown text, with the remaining corpus texts used to form the author profiles. In other words, our goal is to check if our methods can correctly identify the author of each text and hence demonstrate their reliability. Note two of the authors (John Polidori and Percy Shelley's poetry works) only had a single work in the corpus so could not be used for testing, leaving thirty-three works for cross-validation. The full results of cross-validation are shown in Tables 5 and 6.

Table 5. Accuracy of Delta using MFWs under cross-validation (absolute number of misclassified texts in brackets)

Num. MFWs	Manhattan Dist. (%)	Cosine Dist. (%)	Overall (%)
100	94	97	95
200	97	97	97
500	97	97	97
1,000	97	97	97
Overall	96	97	97

Table 6. Accuracy of Delta using 3-grams under cross-validation (absolute number of misclassified texts in brackets)

Num. 3-grams	Manhattan Dist. (%)	Cosine Dist. (%)	Overall (%)
200	82	91	86
500	85	91	88
1,000	94	94	94
2,000	91	94	92
Overall	88	92	90

The overall accuracy of Delta on the corpus across the two distance metrics and four counts of MFWs is 97% (a total of nine incorrect attributions across eight experiments; each option has one error except for 100 MFWs and Manhattan distances, for which there were two). The texts that were at times misclassified were Thomas Love Peacock's *Maid Marian* to Walter Scott twice (similar to the result in Fig. 1); Charles Brockden Brown's *Jane Talbot* to William Godwin three times (also not surprising given Figs 1 and 2); Godwin's *Imogen* to Percy Shelley twice; and Mary Wollstonecraft's *Maria* to Godwin twice.[11]

Using 3-grams, the overall accuracy of Delta is 88% when using the Manhattan distance and 92% using the cosine distance. Peacock's *Maid Marian* and Wollstonecraft's *Maria* were particularly troublesome, being attributed to Scott and Godwin, respectively, in every case. Wollstonecraft's *Mary* was attributed to Mary Shelley five times out of eight. The remaining misattributions were on Godwin's *Imogen* and *Damon and Delia* and Brown's *Jane Talbot*. All of Mary and Percy Shelley's works were correctly attributed under cross-validation, though they did at times pick up works from other authors.

Despite these errors, the accuracy of 97% shows that Delta is extremely reliable for attributing late 18th- and early 19th-century literary texts using most frequent words, and still has a high level of reliability when using 3-grams. This further increases the believability of the attribution of *Frankenstein* to Mary Shelley.

Next, we perform a similar analysis to verify the accuracy of SVMs. We found that using 500 MFWs, there were four misclassifications in the corpus, giving an accuracy of 88%; using 2,000 most frequent 3-grams gave an accuracy of only 79%. The accuracy was roughly the same for other counts of most frequent words/3-grams. Some incorrect classifications were Thomas Love Peacock's *Maid Marian* being attributed to Sir Walter Scott; Percy Shelley's *St. Irvyne* to Mary Shelley, and one each of Mary Wollstonecraft's novels to Mary Shelley and William Godwin. In addition, *Zastrozzi* and one volume of Percy Shelley's poetry are attributed to his wife when using 3-grams. This is potentially due to class imbalance—as seen in Figs 1 and 2, Mary and Percy Shelley and Mary Wollstonecraft had similar writing styles compared to other authors in the corpus, and with Mary Shelley having six novels in her profile, this could skew attributions towards her. The mistakes surrounding Percy Shelley's works are particularly concerning in the context of the *Frankenstein* authorship question, so the SVM trained on the corpus should not be given as much weight in concluding the novel's true author. However, Delta's near-perfect accuracy shows it is a reliable method and fortunately both Delta and SVMs agree that Mary Shelley is the most likely author of *Frankenstein*.

6.2 Impostors method

The above analysis focused on authorship attribution, that is, identifying the most likely author from a given set of candidate authors. However, the problem can also be phrased in terms of authorship verification where the question is a binary yes/no decision as to whether a particular author is likely to have written a text. We will now study *Frankenstein* as a verification problem, using the recently introduced 'Impostors' algorithm.

The Impostors method was introduced by Koppel and Winter (2014) and has since been applied to several problems in stylometry, for example in Kestemont *et al.* (2016). We provide a summary of the method, and the interested reader can consult Koppel and Winter (2014) for a full description. Suppose we have an unknown text (e.g. *Frankenstein*) represented as a vector $\mathbf{x}$ of features and we wish to determine if it was written by a particular candidate author. Let $T = \{t_1, \ldots, t_n\}$ denote a set of texts which are known to be written by this author. Next, assume we have a set of distractor documents written by other authors (i.e. the rest of the corpus) known as 'impostors', $I = \{i_1, \ldots, i_k\}$. During each iteration of the algorithm (typically 100 in total), the method selects a random subset of the impostors, I', and determines whether $\mathbf{x}$ is closer to an item in T than in I', considering a proportion (e.g. 0.5) of the total features available. The value returned by the method is a value between 0 and 1 representing the proportion of iterations for which the analysed text was closer to a work of the suspected author than one of the impostors (i.e. closer to an element of T than one of I'). If the returned score is larger than a threshold σ, then we have verified that the text was written by this author. Otherwise, verification has failed.

Similar to the discussions mentioned in the section on the Delta framework, the choice of distance/similarity metric may significantly affect the performance of the algorithm. Popular choices are the Manhattan distance, the cosine distance, and the min–max distance, as used by Koppel and Winter (2014), where for two vectors $\mathbf{x}$ and $\mathbf{y}$ of n features:

$$\text{minmax}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)}.$$

Distances between z-scores instead of relative frequencies may also be considered ('Delta' distances).

We apply the Impostors method to *Frankenstein* as a test of robustness, using the `imposters()` function from the `stylo` R package, considering both word tokens and 3-grams. We present in detail the results for using the 1,000 most frequent words and 2,000 most frequent 3-grams, though the results when using different counts are essentially the same and the conclusions identical. For simplicity, we use the default values: 100 iterations, selection of 50% of texts as impostors and 50% of features for comparisons, and the classic Delta distance.

The `R` output from running the Impostors method is given in Fig. 5. The numerical values are not fixed due to the stochastic nature of the algorithm.

In either case, the score of 1 for Mary Shelley (`MShelley`) is unambiguous—the algorithm does not find any reason to doubt that Mary Shelley wrote *Frankenstein*. In technical terms, during every iteration,

(a)

```
Brown   0
Godwin   0.06
MShelley   1
Peacock   0
Polidori   0
PShelley   0
PShelleyPoet   0
Scott   0
Shelleys   0
Wollstonecraft   0
```

(b)

```
Brown   0.01
Godwin   0.05
MShelley   1
Peacock   0
Polidori   0
PShelley   0
PShelleyPoet   0
Scott   0
Shelleys   0
Wollstonecraft   0
```

Figure 5. Results from applying the Impostors algorithm to *Frankenstein*, using the classic Delta distance. (a) 3-grams. (b) word tokens.

Frankenstein is found to be closer to one of Mary Shelley's novels than to any of the impostor works. While being different from zero, the scores for William Godwin of 0.06 and 0.05 are sufficiently small to reject him as the novel's author. With regards to the original hypothesis of *Frankenstein* being Percy Shelley's composition, we find no evidence in favour of this claim here.

We can use `imposters.optimize()` from `stylo` to tune the thresholds for a 'yes' and 'no' decision. We find the lower `p1` values to be 0.24 and 0.25 for 3-grams and tokens, respectively, indicating that a value lower than this may be taken as a rejection of that person as the author. Similarly, the `p2` values that give the cut-off for a 'yes' decision are found to be 0.75 and 0.73—these are the values of σ referenced earlier. Between the `p1` and `p2` values is the 'grey area' for the algorithm—we cannot reliably draw a conclusion as to predicted authorship. Fortunately, none of the returned scores fall within this range in our case.

To check our results for robustness, we ran the Impostors algorithm using different distance metrics and choices of MFWs and 3-grams. Attribution to Mary Shelley, combined with a notable lack of evidence in favour of Percy Shelley's authorship, occurs in every case. As before, Godwin is identified as a more likely candidate than Percy Shelley, similar to what we found when looking at Delta distances.

The Impostors method is likely to be sensitive to some authors having more works in the corpus than others since the algorithm makes instance-based comparisons. Therefore, it is possible that, for example, Scott would be put at an advantage with six works, over Polidori and Wollstonecraft with only one or two works. This is the usual problem of class imbalance which can affect the problem of some supervised learning methods.

To gain additional confidence that the attribution to Mary Shelley is not overly influenced by class imbalance, we apply the Impostors method to all other works from Mary and Percy Shelley for which authorship is undisputed, of which there are eleven in total. For example, we will confirm *Zastrozzi* is attributed to Percy Shelley, even if he has fewer works making up the authorial profile when instance-based comparisons are considered. In doing this, the work in question is removed from the corpus of works for comparison, and we work only with the data on the frequency of MFWs. In other words, we are performing leave-one-out cross-validation.

For all works, we achieve a score from the algorithm of 1 for the expected author. For Percy Shelley's works, there are non-trivial scores for Mary Shelley's potential authorship. A potential reason for this is there are only a limited number of available poetry and prose works written by Percy Shelley, and the Impostors method can suffer from the aforementioned class imbalance. To make the training data a little more balanced, we combined the prose and poetry profiles for Percy Shelley into a single profile, giving five works for Percy compared to the six for Mary. This made no significant difference to the results. In the application to Mary Shelley's novels, all receive scores of 0 or very close to 0 as far as Percy Shelley's authorship goes. Therefore, in summary, when dealing with the novels, the results from the Impostors algorithm appear consistent and reliable and further suggest that *Frankenstein* was authored solely by Mary Shelley.

7. Limitations

While we believe our above analysis is conclusive, we should mention the potential limitations of our approach.

Throughout this analysis, we have assumed that all of Mary Shelley's post-*Frankenstein* novels were actually written by her. One could object that the similarity between *Frankenstein* and her other novels is

indicative of Percy Shelley having written all these works, rather than Mary having authored *Frankenstein* (similar to theories that all of Shakespeare's works were written by the Earl of Oxford or others). However, this argument holds no weight and is impossible, since Percy Shelley died in July 1822, so could only have contributed to *Mathilda* and potentially *Valperga*.

A second potential limitation is the relatively small number of Percy Shelley prose works included in the corpus. This mostly stems from the fact that he was not primarily a novelist, and while his works of poetry and prose essays were included, both are seen to be dissimilar to his novels, leading to the poetry only being included as forming a second authorial profile for Shelley. However, our cross-validation study has confirmed the general finding in the stylometric community that Delta can accurately identify authorship even given only a relatively small number of available works, and so we do not believe this is a major issue.

8. Conclusion

In conclusion, the above analysis points strongly in favour of Mary Shelley's authorship for *Frankenstein*. This finding is robust against several stylometric methods including classic authorship attribution algorithms such as Delta and SVMs, along with the unsupervised Impostors approach for authorship verification. Our robustness studies showed these techniques have high accuracy for attributing the authorship of 19th-century literary texts, and can hence be trusted. While it is known from Robinson (2008) that Percy's contribution to *Frankenstein* was not insignificant, his style is simply not close enough to *Frankenstein* to suggest he played a heavy role in writing the novel, or was responsible for the novel as a whole, as some have suggested (Zimmerman, 1998; Lauritsen, 2007; De Hart, 2013; Jones, unpublished).

To the (small) extent that a second person is implicated as a potential author of *Frankenstein*, it is William Godwin who prevails over Percy Shelley, for example in certain cases of Delta applied to 3-grams, and in our Impostors study. This relative closeness of Godwin's style to that adopted by Mary Shelley is not altogether surprising given his previously described influence on her upbringing, education, and career as a writer. However, we do not wish to suggest or imply Godwinian authorship of *Frankenstein* given the complete absence of any literary or external evidence for this, and our results point strongly toward Mary Shelley's sole authorship.

Appendix A

Composition of Corpus

In the following table, PG = Project Gutenberg; PGA = Project Gutenberg Australia; and AL = The Anarchist Library. While *Mathilda* was only published posthumously in 1959, it was written in 1819, thus the year is recorded as 1819 (Shelley and Nitchie, 1959; Fisch et al., 1993, p. 5). *Maria, or The Wrongs of Woman* was left incomplete after Mary Wollstonecraft's death in 1797, and published by William Godwin in 1798.

Number	Author	Title	Year	Source
1	To be Determined	Frankenstein	1818, 1831	PG, PG
2	Charles Brockden Brown	Arthur Mervyn	1799	PG
3	Charles Brockden Brown	Edgar Huntly	1799	PG
4	Charles Brockden Brown	Jane Talbot	1801	PG
5	Charles Brockden Brown	Ormond	1799	PGA
6	Charles Brockden Brown	Wieland	1798	PG
7	William Godwin	Caleb Williams	1794	PG
8	William Godwin	Damon and Delia	1784	PG
9	William Godwin	Fleetwood	1805	AL
10	William Godwin	Imogen	1784	PG
11	William Godwin	St. Leon	1799	PG
12	Thomas Love Peacock	Crotchet Castle	1831	PG
13	Thomas Love Peacock	Headlong Hall	1815	PG
14	Thomas Love Peacock	Maid Marian	1822	PG
15	Thomas Love Peacock	Nightmare Abbey	1818	PG
16	John Polidori	The Vampyre	1819	PG
17	Mary Shelley	Falkner	1837	PGA
18	Mary Shelley	Lodore	1835	PGA
19	Mary Shelley	Mathilda	1819	PG
20	Mary Shelley	The Fortunes of Perkin Warbeck	1830	PGA
21	Mary Shelley	The Last Man	1826	PG
22	Mary Shelley	Valperga	1823	PGA
23	Percy Shelley	St. Irvyne	1811	PGA
24	Percy Shelley	Zastrozzi	1810	PGA
25–27	Percy Shelley (Poetry)	Complete Poetical Works (3 Vols)	Various	PG
28	Sir Walter Scott	Guy Mannering	1815	PG
29	Sir Walter Scott	Ivanhoe	1819	PG
30	Sir Walter Scott	Kenilworth	1821	PG
31	Sir Walter Scott	Quentin Durward	1823	PG
32	Sir Walter Scott	Rob Roy	1817	PG
33	Sir Walter Scott	The Talisman	1825	PG

(continued)

(continued)

Number	Author	Title	Year	Source
34	Mary and Percy Shelley	History of a Six Weeks' Tour	1817	PG
35	Mary Wollstonecraft	Maria, or The Wrongs of Woman	1798	PG
36	Mary Wollstonecraft	Mary: A Fiction	1788	PG

Appendix B

List of the 200 Most Frequent Words used in the analysis

the, i, to, and, a, of, in, that, it, my, is, you, for, was, on, me, but, so, this, with, have, be, we, at, not, all, im, as, like, are, just, its, out, up, about, they, what, or, one, if, do, from, had, get, when, will, there, dont, time, know, now, can, some, then, by, really, no, well, an, your, go, more, were, am, think, would, who, people, good, been, how, has, got, them, going, because, back, day, see, much, our, only, which, want, their, love, even, other, too, after, today, went, over, way, here, last, into, ive, still, say, could, very, things, new, did, life, work, off, something, right, make, than, us, first, thats, said, didnt, little, cant, night, down, never, thing, also, why, again, around, being, feel, ill, home, where, any, should, need, take, two, before, most, while, those, oh, come, these, made, great, though, long, always, better, ever, friends, myself, another, many, since, next, maybe, thought, look, through, fun, few, bad, actually, find, world, week, lot, sure, days, year, someone, man, pretty, years, away, getting, every, tell, may, best, god, came, anything, same, nothing, let, stuff, doing, read, old, everyone, guess, place, put, nice, left, own, told

Notes

1. As Raven (2003, p. 143) states 'the overwhelming majority of the English novels of the eighteenth and early nineteenth centuries were published without attribution of authorship either on the title page or within the preface or elsewhere in the text.' Over 80% of new novels between 1750 and 1790 were published with no clear ascription of authorship (Raven, 2003, pp. 143–145). While this percentage decreased during the 1790s and 1800s, it was still the case that, for example, in the year of *Frankenstein*'s publication, 1818, 66%, of the 62 published novels were anonymous to some degree (Raven, 2003, p. 164).
2. The act of leaving one's name off the title page was well-known to members of Mary Shelley's family (Eilenberg, 2003, pp. 171–173). Percy Shelley published his first novel *Zastrozzi* (1810) with only his initials 'P.B.S'. and his second novel *St. Irvyne* as 'by a Gentleman of the University of Oxford'. Several of his essays and poems (e.g. *A Vindication of the Natural Diet* (1813) and *Epipsychidion* (1821), respectively) were also published anonymously (Morton, 2006, p. xv). Mary's father, William Godwin, published most of his major works under his own name, though some such as the novels *Damon and Delia* (1784) and *Imogen* (1784) appeared anonymously. Her mother, Mary Wollstonecraft published the novel *Mary, a Fiction* (1788) anonymously, but did put her name to most other works. After *Frankenstein*, Mary Shelley's later novels (excepting later editions of *Frankenstein*) were published under the name of 'the Author of 'Frankenstein', 'the Author of 'The Last Man', and variations on these (Schor, 2003, pp. 114, 136). Though it should be noted that Mary's publishing anonymously was a result of Percy's father, Sir Timothy Shelley, basing his financial support on Mary not publishing under the Shelley name, rather than her personal preference (Eilenberg, 2003, pp. 173–174).
3. That is, while Mary is known to have done copy work for her husband, Lord Byron, and Thomas Love Peacock, there is no evidence of Percy Shelley doing this sort of work for anyone other than himself.
4. It is possible to view the 1818 and 1831 editions side by side here: http://knarf.english.upenn.edu/Text/text.html (accessed 8 January 2022).
5. The collection is *A Defence of Poetry and Other Essays*, available at https://www.gutenberg.org/ebooks/5428 (accessed 8 January 2022).
6. In fact, in preliminary analysis where poems were included, these were found via cluster analysis and principal component analysis to appear significantly different from the novels and essays, giving us evidence for this choice.
7. See http://knarf.english.upenn.edu/MShelley/bydates.html (accessed 8 January 2022).
8. Admittedly, using only the two principal components as we have done here is not ideal for authorship attribution, as these two PCs represent under 40% of the total variation in the underlying data (using two PCs in such cases has been criticized, for example, in Rizvi (2021)). There is no single rule for the percentage of the underlying variation the selected PCs should account for, though cut-offs tend to be in the range of 70–90% (Jolliffe, 2002, p. 113). Therefore, if we wished to claim our use of PCA constitutes robust authorship attribution, many more PCs would be necessary, but since this is only part of an initial description and visualization of our corpus, we feel using two PCs is suitable and practical for this purpose.
9. Another issue that can crop up is how to deal with contractions—and whether it is necessary to remove them or replace them with uncontracted forms. This did not appear to be a concern in our corpus as no contractions appeared within the 1,000 most frequent words.
10. The results without culling are essentially identical and thus are not given in detail for brevity. Though the choice for the number of most frequent words and 3-grams used is somewhat arbitrary, the attribution to Mary Shelley was found to be the same for any choice between 40 and 1,000 MFWs for either distance metric. Attribution using 3-grams was less consistent—when using 140–260 or 580–720 3-grams and the cosine distance, *Frankenstein* was attributed to William

Godwin or occasionally Charles Brockden Brown, though usually by a very small margin; in all other cases the attribution was to Mary Shelley.

11. This may have something to do with Godwin's influence in the editing and posthumous publication of the novel (Johnson, 2002, pp. xvii–xix), though his influence is not thought to be significant enough to affect the style of the novel as a whole (Myers, 1980; Mellor, 1996).

References

Argamon, S. (2008). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC Conference*, University Of Victoria, Canada.

Binongo, J. N. G. and Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, **14**(4): 445–66.

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. 2nd edn. New York, NY: Springer.

Burrows, J. (2002). Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. (2010). Never say always again: reflections on the numbers game. In McCarty, W. (ed), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 13–36.

Clemit, P. (1999). Mary Shelley and William Godwin: a literary–political partnership, 1823–36. *Women's Writing*, **6**(3): 285–95.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**: 273–97.

De Hart, S. D. (2013). *Shelley Unbound: Discovering Frankenstein's True Creator*. Port Townsend, WA: Feral House.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, **19**: 109–23.

Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, **32**(1): 50–64.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, **8**(1): 107–21.

Eilenberg, S. (2003). Nothing's namelessness: Mary Shelley's Frankenstein. In Griffin, R. J. (ed.), *The Faces of Anonymity: Anonymous and Pseudonymous Publication from the Sixteenth to the Twentieth Century*. New York, NY: Palgrave Macmillan, pp. 167–92.

Fisch, A. A., Mellor, A. K., and Schor, E. H. (1993). *The Other Mary Shelley: Beyond Frankenstein*. Oxford: Oxford University Press.

Fraistat, N. (1991). 'Introduction to Prometheus Unbound'. http://shelleygodwinarchive.org/contents/prometheus_unbound/prometheus-unbound-introduction/, (accessed 30 December 2021).

Frankenstein. (1824). The Anniversary - Anonymous. *Knight's Quarterly Review* **3**, 195–99. http://knarf.english.upenn.edu/Reviews/knights.html (accessed 30 December 2021).

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, **22**(3): 251–70.

Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **155**(1): 91–120.

Hoover, D. L. (2004a). Delta prime? *Literary and Linguistic Computing*, **19**(4): 477–95.

Hoover, D. L. (2004b). Testing burrows's delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2012). The tutor's story: a case study of mixed authorship. *English Studies*, **93**(3): 324–39.

Irvine, R. P. (2005). *Jane Austen*. London: Routledge.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In Nédellec, C. and Rouveirol, C. (eds), *Machine Learning: ECML-98*. Berlin, Heidelberg, Germany: Springer, pp. 137–42.

Jockers, M. L., Witten, D. M., and Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing* **23**(4): 465–91.

Johnson, C. L. (ed.) (2002). *The Cambridge Companion to Mary Wollstonecraft*. Cambridge: Cambridge University Press.

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd edn. New York, NY: Springer.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Juola, P. (2015). The rowling case: a proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, **30(Suppl_1)**: i100–13.

Kešelj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pp. 255–64.

Kestemont, M. (2014). Function words in authorship attribution from black magic to theory? *Proceedings of the Third Computational Linguistics for Literature Workshop (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 59–66.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., and Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, **63**: 86–96.

Kjell, B., Woods, W. A., and Frieder, O. (1994). Discrimination of authorship using visualization, *Information Processing & Management*, **30**(1): 141–50.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, **65**(1): 178–87.

Lauritsen, J. (2007). *The Man Who Wrote Frankenstein*. Dorchester, MA: Pagan Press.

Lauritsen, J. (2018a). 'The Real Frankenstein and Its Author'. https://www.us.mensa.org/read/bulletin/features/the-real-frankenstein-and-its-author/ (accessed 30 December 2021).
Lauritsen, J. (2018b). The true author of Frankenstein. *Academic Questions*, **31**(4): 450–7.
McCarthy, R. and O'Sullivan, J. (2020). Who wrote wuthering heights? *Digital Scholarship in the Humanities*, **36**(2): 383–91.
Mellor, A. K. (1996). Righting the wrongs of woman: Mary Wollstonecraft's Maria. *Nineteenth-Century Contexts*, **19**(4): 413–24.
Miranda García, A. and Calle Martín, J. (2006). Function words in authorship attribution studies. *Literary and Linguistic Computing*, **22**(1): 49–66.
Miranda García, A. and Calle Martín, J. (2012). The authorship of the disputed federalist papers with an annotated corpus. *English Studies*, **93**(3): 371–90.
Morton, T. (ed.) (2006). *The Cambridge Companion to Shelley*. Cambridge: Cambridge University Press.
Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, **58**(302): 275–309.
Murray, E. B. (1978). Shelley's contribution to Mary's Frankenstein. *Keats-Shelley Memorial Bulletin*, **29**: 50–68.
Murray, E. B. (1981). Changes in the 1823 Edition of Frankenstein. *The Library*, **s6-III**(4): 320–27.
Myers, M. (1980). Unfinished business: Wollstonecraft's Maria. *The Wordsworth Circle*, **11**(2): 107–14.
O'Sullivan, J. (2022). The sociology of style: writing and influence within literary families. *Poetics*, **92**: 101620.
Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About us*. New York, NY: Bloomsburg Press.
Raven, J. (2003). The anonymous novel in Britain and Ireland, 1750–1830. In Griffin, R. J. (ed.), *The Faces of Anonymity: Anonymous and Pseudonymous Publication from the Sixteenth to the Twentieth Century*. New York, NY: Palgrave Macmillan, pp. 141–66.
R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/, (accessed 30 December 2021).
Reiman, D. H. and Powers, S. B. (eds) (1977). *Shelley's Poetry and Prose: A Norton Critical Edition*. New York, NY: W. W. Norton & Company.
Rieger, J. (1974). *Frankenstein or the Modern Prometheus: the 1818 Text*. Chicago, IL: University of Chicago Press.
Rizvi, P. (2021). Shakespeare and principal components analysis. *Digital Scholarship in the Humanities*, **36**(4): 1030–41.
Robinson, C. E. (1996). *The Frankenstein Notebooks: A Facsimile Edition*. New York, NY: Garland Publishing. http://shelleygodwinarchive.org/contents/frankenstein/the-frankenstein-notebooks-introduction/ (accessed 30 December 2021).
Robinson, C. E. (2008). *The Original Frankenstein*. Oxford: The Bodleian Library.
Ross, G. J. (2020). Tracking the evolution of literary style via Dirichlet–multinomial change point regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **183**(1): 149–67.
Rybicki, J. (2016). Vive la différence: tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities*, **31**(4): 746–61.
Rybicki, J., Hoover, D., and Kestemont, M. (2014). Collaborative authorship: conrad, ford and rolling delta. *Literary and Linguistic Computing*, **29**(3): 422–31.
Sapkota, U., Bethard, S., Montes, M., and Solorio, T. (2015). Not all character N-grams are created equal: a study in authorship attribution. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, CO: Association for Computational Linguistics, pp. 93–102.
Savoy, J. (2018). Is Starnone really the author behind Ferrante? *Digital Scholarship in the Humanities*, **33**(4): 902–18.
Savoy, J. (2020). *Machine Learning Methods for Stylometry*. Cham, Switzerland: Springer.
Schöberlein, S. (2017). Poe or not Poe? A stylometric analysis of Edgar Allan Poe's disputed writings. *Digital Scholarship in the Humanities*, **32**(3): 643–659.
Schor, E. (ed.) (2003). *The Cambridge Companion to Mary Shelley*. Cambridge: Cambridge University Press.
Scott, W. (1818). Remarks on Frankenstein, or the modern prometheus; a novel. *Blackwood's Edinburgh Magazine*, **2**(12): 613–20.
Shelley, M. (1818). *Frankenstein; or, the Modern Prometheus*. 1st edn. London: Lackington, Hughes, Harding, Mavor, & Jones.
Shelley, M. (1831). *Frankenstein; or, the Modern Prometheus*. 3rd edn. London: Henry Colburn & Richard Bentley.
Shelley, M. and Groom, N. (2019). *Frankenstein: or 'The Modern Prometheus': The 1818 Text*. 3rd edn, Oxford: Oxford University Press.
Shelley, M. and Nitchie, E. (1959). *Mathilda*. Chapel Hill, NC: University of North Carolina Press.
Shelley, M. and Rossington, M. (2000). *Valperga; or, the Life and Adventures of Castruccio, Prince of Lucca*. Oxford: Oxford University Press.
Spencer, J. (2007). Evelina and Cecilia. In Sabor, P. (ed.), *The Cambridge Companion to Frances Burney*. Cambridge: Cambridge University Press, pp. 23–37.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.
Stamatatos, E. (2013). On the robustness of authorship attribution based on character N-gram features. *Journal of Law and Policy*, **21**(2): 421–39.
Wu, D. (2015). Percy bysshe Shelley wrote Frankenstein. *Myth 25: 30 Great Myths about the Romantics*. Chichester: John Wiley & Sons, Ltd, pp. 212–19.
Zimmerman, P. (1998). *Shelley's Fiction*. Los Angeles, CA: Darami Press.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. New York, NY: Addison-Wesley.