# Stylometry: An Investigation into Frankenstein's Authorship

Jacob Alban Russan-Pritchett (s2152592), Willow Lilith Fossil (s2279407), & Izabel Samaeva (s2277508)

October 2024


## *Executive Summary*

Published in the early 19$^{\text{th}}$ century, *Frankenstein* has historically had disputed authorship perhaps due to its initial anonymous writer. While the novel is typically attributed to Mary Shelley, it has been argued by literary critics and analysts that it is instead a work by her husband, Percy Shelley, or even a joint work by both. To determine the true author, we have used stylometric methods such as K-nearest neighbours (KNN), and discriminant analysis to compare a multitude of different authors, including Mary and Percy Shelley's separate and joint novel data. To make this comparison, we statistically analysed the frequency of the 69 most commonly used words for multiple authors' novels. Our data analysis using KNN showed significant similarities between the writing style of *Frankenstein* and that of author Mary Shelley, irrespective of Percy Shelley's prose and poetry styles being combined together or left separate. Our less precise and accurate method, discriminant analysis, did point towards Mary Shelley's father, William Godwin, authoring the novel. Still, we determined this result to be due to the method performing worse. This lead us to conclude that the true author of the novel is most likely to be Mary Shelley. While our conclusion was reached through rigorously tested methods, we recommend comparing to additional authors using more methods in future analysis to gain greater certainty of *Frankenstein*'s authorship.


## 1 Introduction

Anonymously published in 1918, *Frankenstein* is typically attributed to the novelist Mary Shelley. However, it has been suggested that the author of this novel is instead her husband, Percy Shelley, or perhaps it was a joint text written by both. Stylometry is the statistical study of literary styles and techniques to determine a text's author or to understand the change in a writer's style. Although these suggestions are typically considered outdated and perhaps even controversial today, we will use stylometric analysis comparing the styles of a multitude of authors, including the Shelley couple and even Mary Shelley's parents in an attempt to pinpoint the novel's true author. Our stylometric analysis will be based on the frequency of the 69 most common words in the authors' different texts. Using leave-one-out cross-validation we will find the precisions and accuracies of our two chosen stylometric methods, K-nearest neighbours and discriminant analysis, subsequently testing the outcomes of our case study by finding the most likely authors when excluding the Shelley couple from the data set. Additionally, we will use MDS to investigate Percy Shelley's different styles in prose and poetry, and whether they should be treated as one style or two entirely separate ones.

## 2 Information on our Datasets

To perform the stylometry, we compared datasets of the frequency of the 69 chosen common words, equivalently referred to as our function words, of novels from different authors and that of the *Frankenstein* novel. Our data consists of the function word frequency from our corpus from Mary Shelley, Percy Shelley, their joint-novel authorship, Mary's parents, and reference authors, used to ensure the quality of our testing, from a similar time period being: Bram Stoker, Charles Brockden Brown, Thomas Peacock, Walter Scott, William Polidori. Additionally, as Percy Shelley had written poetry, we test whether his styles should be evaluated individually or jointly in 4.1..

Our 69 function words were chosen on the premise of being the most relevant common words in the English language, excluding pronouns as those are heavily affected by a novel's specific characters rather than an author's writing style. We performed feature extraction by taking our 69 function words, categorising each text in the corpus by the quantity

of each such function word. For each text, we created a 70 entry long vector with each of the 69 word counts and the $70^{\text{th}}$-entry being the total remaining words. Each author's vectors were compared to those of the other authors and to the similarly structured vector of the novel *Frankenstein* to find our expected match.

# 3 Discussion of Statistical Methods

As outlined in the introduction, Stylometry is a statistical tool we can use to analyse text authorship based on sufficiently large corpuses. Here we will delve into the machinery of how this works and how it is employed in this report.

To mathematically set up our method, we designate a pairing $(x_i, y_i)$, where $x_i$ is our vector of function words for the $i^{\text{th}}$ author's bibliography and $y_i$ is the associated $i^{\text{th}}$ author. Our purpose is to categorise new observations, i.e. new texts, by allocating them to the corpus's most probable author. To do this, we have chosen the following two following common stylometric methods:

- Discriminant Analysis (DA):
  Simply put, our goal is to predict the author of a new given text using its function word vector. So the approach here is that we can calculate the probability that any given author in our corpus is the author of this new text, and we choose the author with the highest probability.

  This is done using Bayes Theorem: $p(y_i = c \mid x_i) \propto p(x_i \mid y_i = c)p(y_i = c)$ where each $c$ is a possible author candidate and each $x_i$ is a text written by an unknown author containing $N_i$ words ($x_i$ is the 70-element function word vector). Our class conditional densities are given by the multinomial $p(x_i \mid y_i = c) = \text{Multinomial}(x_i; N_i, \theta_c)$.

  We can estimate the multinomial parameter vector $\theta_c = (\theta_{1,c}, \theta_{2,c}, \cdots, \theta_{70,c})$, where $\theta_{i,c}$ is the probability of a random word from author $c$ being the word(s) corresponding to the $i^{\text{th}}$ entry in the function word vector, by using the MLE of $\hat{\theta_{i,c}} = \frac{k_{i,c}}{N_c}$ where $k_{i,c}$ is the total count of function word $i$, or number of non-function words for $i = 70$, for author $c$ and $N_c$ is the total words written by author $c$.

  With all of that, we take our prior $p(y_i = c) = \frac{1}{\#\text{ of Prospective Authors}}$ $\forall c$ i.e we take all authors to be equally likely to have written the unknown text, and calculate the class conditional density for each author and multiply by the prior. Rescaling these outputs to sum to 1, we pick the author corresponding to the largest probability output to have written the unknown text.

- K-Nearest Neighbours (KNN):
  This classification technique is arguably much more straightforward in the sense that we need only concern ourselves with the closest distances to our vector of unknown authorship $\tilde{x}$ to all other vectors in the $(x_i, y_i)$ pairs training set for each author and their designated function word vector. In general, this is the $K$ closest vectors, however, we will be using $K = 1$ as each vector is a separate classification so it would not make sense to use $K > 1$ in this application.

  Using $K = 1$, with these distances, we classify our author of the unknown text as follows: $r = \arg\min_i (d(x_i, \tilde{x}))$ is the index of the minimal distance given by the distance function $d$, we then classify the author of the unknown text as $\tilde{y} = y_r$. The distance function we will be using is the $L_1$ norm.

# 4 Methodical Precision & Accuracy of Known Authorship

## 4.1 Justification regarding Combination of Percy Shelley Novels & Poetry

As stated in 2, the corpus includes function word data sets of Percy Shelley's novels and poetry separately. Here we will outline the justification for leaving separate or combining these data sets. Regardless of this outcome, we will classify *Frankenstein* in section 5.1 using both the combined and separate data sets to determine if the results are

consistent across these two approaches in consideration of statistical completeness.

To quantify differences in writing style, we need to compare the datasets of different authors. A challenge arises in visualising these high dimensional function vectors: they need to be shown in an analogous lower dimensional visualisation. This is where multidimensional scaling (MDS) becomes useful, explained and further used in 5.2. We apply MDS to represent our function vectors in 2-D space, which can be visualized on a graph.

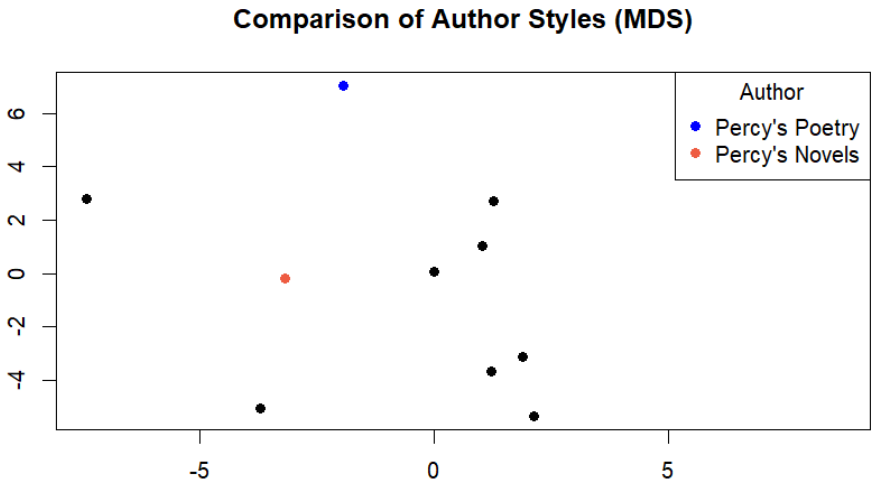


Figure 1: Comparison of Author Styles (MDS)

Figure 1 shows that the writing styles of Percy Shelley's novels and poetry differ significantly, more so than the differences between the other authors. To quantify this, we calculate the difference between Percy Shelley's poetry and novels as 11.64 (2 d.p), while the average difference between all pairs of authors is 10.56 (2 d.p). If we were to combine Percy Shelley's novels and poetry into one dataset, we would expect their difference to be far lower than the average distance between all authors. However, our results show that the difference between Shelley's novels and poetry is greater than the average difference, providing clear justification for keeping the two datasets separate.

## 4.2 Discriminant Analysis vs KNN as Classification Methods

Before applying our methods in 3 to classify *Frankenstein* we wish to test the validity of the methods when classifying know authorship. This will provide some metrics as to the reliability of our classification of *Frankenstein*.

We used leave-one-out cross-validation (LOO-CV) to assess the validity of our methods. Compared to other methods such as k-fold validation, LOO-CV can provide more reliable results, though it requires more computational resources. However, due to our small dataset size, this was not a concern.

In LOO-CV, we iterate through each author and, for each text, we classify it using either K-Nearest Neighbors (KNN) or discriminant analysis. The training data consists of the entire dataset except for the text being classified. After classifying each text, we compare the predicted classification to the true classification. Once all texts and authors have been processed, we can evaluate our model's performance based on the number of correctly and incorrectly classified texts.

Note that for our dataset specifically, some authors have only one text attributed to them. We have skipped these authors' single text in the cross-validation. If we where to perform one, the removed text would never be predicted correctly, as the training dataset would not include any information on the correct author.

We performed LOO-CV using both methods in 3 on datasets with Percy Shelley's novels and poetry both combined and separate, mentioned in 4.1. And extracted three metrics from our results:

- *Accuracy*: This is simply the proportion of texts that have been correctly classified.
- *Precision*: This is the proportion of positive classifications that are correct. This can be used to test how good classification of a specific author is, Mary Shelley and Percy Shelley are of particular interest here.
- *Mean Precision*: This is the average of precision across all authors. Unlike accuracy, this treats each class equally, regardless of quantity of texts in the class.

Table 1: Classification Metrics of the Separate Dataset as Percentages (1 d.p)

| **Method** | **Accuracy** | **Precision** | | | **Mean Precision** |
|---|---|---|---|---|---|
| | | Mary Shelley | Percy's Novels | Percy's Poetry | |
| DA | 81.6% | 55.6% | 100.0% | 100.0% | 87.8% |
| KNN | 92.1% | 100.0% | 100.0% | 100.0% | 94.6% |

Table 2: Classification Metrics of the Combined Dataset as Percentages (1 d.p)

| **Method** | **Accuracy** | **Precision** | | **Mean Precision** |
|---|---|---|---|---|
| | | Mary Shelley | Percy Shelley | |
| DA | 76.3% | 45.4% | 100.0% | 84.6% |
| KNN | 84.2% | 66.7% | 100.0% | 89.2% |

Upon reviewing the performance metrics from both Table 1 and Table 2, it is clear that KNN consistently outperforms discriminant analysis (DA) across all metrics. One area of particular concern is the precision for Mary Shelley's texts under discriminant analysis, which is notably low. This suggests that DA tends to over-classify texts as Mary Shelley's, raising doubts about the reliability of DA in correctly identifying her works. Given that accurate classification of Mary Shelley is crucial for this model, this over-classification significantly undermines the model's validity when using DA.

Additionally, the metrics for the combined dataset are consistently lower than those for the separate dataset, reinforcing the justification to keep Percy Shelley's bibliography distinct. This separation improves model performance for both DA and KNN. The most striking evidence for this is the precision of Mary Shelley's classification. Both DA and KNN show very low precision when using the combined dataset, indicating frequent misclassification of texts as Mary Shelley's. This seriously compromises the credibility of any classification of her works using the combined data.

Despite these concerns, both accuracy and mean precision remain relatively high across the board, especially mean precision. This suggests that the methods used to represent author writing styles and classify based on those representations are generally effective. This is consistent with results typically seen in stylometric analysis, where function words often provide strong indicators for author classification.

# 5 Stylometric Analysis of Frankenstein

## 5.1 Classification of Frankenstein

**Using the entire Dataset:**

As we have thoroughly weighed up each stylometric technique in the previous section, we can proceed with our authorship classification of the novel *Frankenstein*. Initially, as given in 4.2, KNN appears to be a better classification technique for this problem; for completeness we will use both in our classification and analyse them appropriately.

Using KNN, we see that the author most likely to have written *Frankenstein* is Mary Shelley and alternatively from discriminant analysis, the most probable author is given to be William Godwin; it is important to note that we get the same outcome irrespective of combination of Percy Shelley's novels and poems into one vector, so from now on we

continue without combining them as assessed prior.

Mary Shelley is the author with the most historical claim to the novel, so it makes sense that the stylometric method with the higher average precision assigns the text to her. It is interesting to note however, that William Godwin is the father of Mary Shelley (maiden name Godwin). So perhaps discriminant analysis, with its lower average precision, misattributed the text due to similar familial writing characteristics, i.e the distance between these authors function word vectors is 8.17 (2 d.p) which is less than the average difference between all pairs of authors 10.56 (2 d.p). On the other hand, it could also potentially be true that William Godwin wrote *Frankenstein* for his daughter to publish under her name to gain a footing as a novelist, given it was published as her first book, although a more concrete foundation in the historical context is needed to make further and more sound analyses of discriminant analysis classification of William Godwin as the author.

**Partial Dataset Utilisation:**

One proposition that we are keen to investigate is whether Percy Shelley may have meaningfully contributed to the writing of our the novel. From the holistic stylometric analysis of the dataset, it appears unlikely for this to be the case, but it would be wise for us to see who our techniques would classify if Mary Shelley was not an option in order to see if further analysis is warranted into this avenue.

We have three separate function vectors influenced or fully characterised by Percy Shelley: his novels, his poems, & novels he wrote together with Mary Shelley. When removing Mary Shelley's function vector from the dataset, if we produce any of the above "authors", it could, as above, indicate the need for further or more detailed research into the relationship between Percy Shelley and the novel *Frankenstein*.

When removing Mary from the dataset, we find that by using KNN and discriminant analysis, the "second most likely" to have written *Frankenstein* is Walter Scott[1] and William Godwin respectively for each technique. Therefore the prediction for KNN changes while discriminant analysis remains identical, neither of these producing any of Percy Shelley's anthology. As a conclusion to our proposition, this heavily indicates a lack of substantive involvement from Percy Shelley in the writing of *Frankenstein*.

**Summary:**

From the previous two parts of this section, there is strong evidence to suggest that Mary Shelley is the author of the novel *Frankenstein*, given the lack of Percy Shelley's involvement and also the greater average precision of KNN which classified Mary as the author.

## 5.2 Data Visualisation

Now we allocate a section of this report to help illustrate our results and data. Firstly, we will be using Multi-Dimensional Scaling, MDS, to represent our 70 dimensional function vectors in 2-D for easier visualisation. The actual distances between our authors do not scale down appropriately to 2-D, so MDS minimises the corresponding error allowing us to most correctly plot them in 2-D.

Using MDS, we can graph our function vectors for each author to gain an intuition towards the grouping and classifications that have been employed in this study so far, in Figure 2. Notably in our visualisation, all authors that our stylometric techniques predicted as the author for *Frankenstein* are some of the closest ones to it, which is logical in the KNN sense, but it is interesting that the discriminant analysis result was also surprisingly close. Specifically, we can see that Mary Shelley and Walter Scott are the first and second closest to *Frankenstein*, as found by removing Mary from the Dataset and redoing the stylometry.

Up until now, we have been grouping all novels by each author together into one cumulative function vector of their entire bibliography. Instead, separating them into individual texts, year by year, for each author of interest (Mary Shelley and William Godwin), we further use MDS to show the change in each author's style through time with every novel, displayed in Figure 3. Based on our prior intuition that the classification of Mary Shelley is by KNN and William

[1]It is perhaps amusing that Walter Scott is classified as the author of *Frankenstein* in the absence of Mary Shelley as he was one of the first authors at the time to dispute the authorship of the novel and postulate Percy Shelley as the true author.

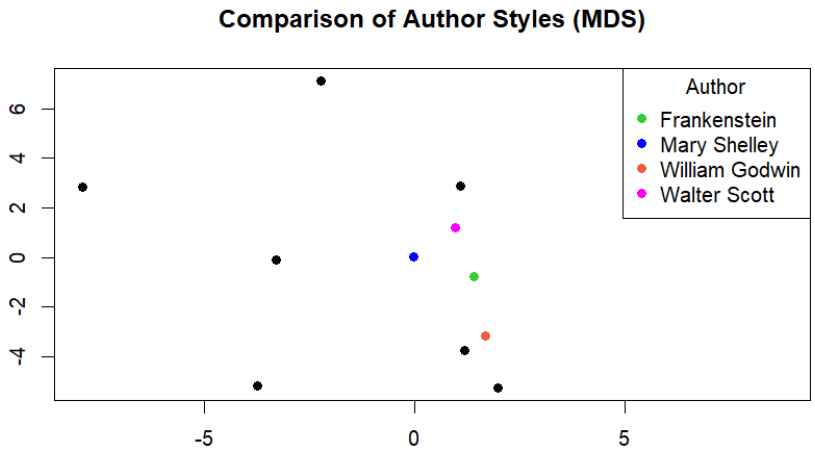


Figure 2: Comparison of Author Styles including Frankenstein (MDS)

Godwin is by DA and each methods corresponding mean precision, this visualisation makes slightly more sense to us because *Frankenstein* is at least close to one of Mary Shelley's novels, where as William Godwin has none close by, making sense as he was classified by DA not KNN.

# 6 Conclusions and Further Research

In conclusion, the above stylometric analysis done using KNN, and LOO-CV supports Mary Shelley as *Frankenstein*'s writer. When we used KNN which had 94.6% mean precision and 92.1% accuracy, Mary Shelley was determined as the author. When analysis was performed using discriminant analysis, which comparatively only had 84.6% mean precision and 81.6% accuracy, the authorship of the novel was attributed to William Godwin, her father. This result may have been caused by him being a large influence on her writing style rather than him being the writer, but as this method had significantly lower mean precision and accuracy, it is reasonable for us to disregard this result.

Additionally, using characteristic distance metrics we had solid evidence to suggest that Percy Shelley's prose and poetry writing styles were different enough to be treated as different authors entirely, however for the sake of statistical completeness we chose to test his styles both jointly as well and got the same results as for both methods, just with lower mean precisions and accuracy. Even when Mary Shelley was taken out of the data set, Percy Shelley was not found to be the novel's writer implying he did not significantly contribute to the novel's writing style, further reinforcing our determining Mary Shelley as the author.

Future research may benefit from more granular stylometric techniques used for singular texts, as we found the singular shared work between Mary and Percy Shelley very limiting. Similar issues were also found in the quantity of Percy Shelley's prose. Additional suggestions would be to add more authors to the analysis, perhaps from a more varied time range, and to consider using more and better stylometric methods tailored specifically to the qualities of our dataset.

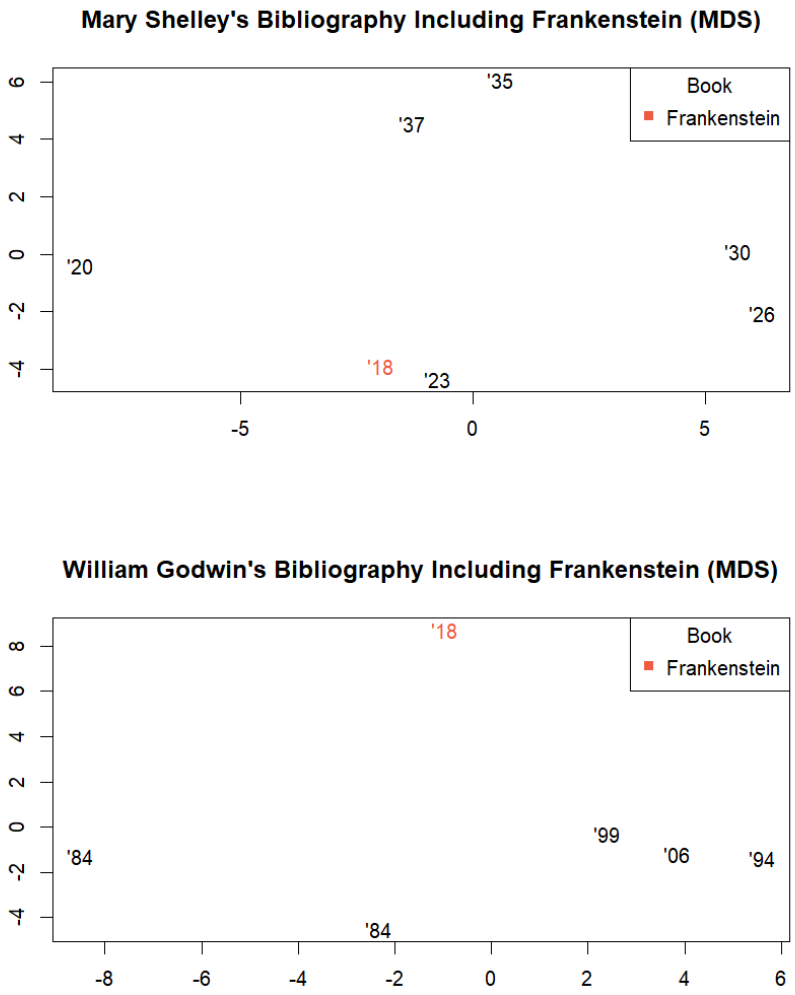


Figure 3: Mary Shelley and William Godwin Bibliography through time, including Frankenstein (MDS)