Stories, Sonnets and Scripts: A Stylometric Analysis of Literature

Pepa Birkett, Anisah Khanom, Runlin Chen

Year 4 Project
School of Mathematics
University of Edinburgh
March 5, 2025

Abstract

Stylometry is the statistical analysis of literary style. Stylometric methods are typically applied to prose to quantitatively answer literary questions on authorship attribution. We present a collection of studies that apply stylometry to different literary forms. To illustrate the conventional uses of stylometry, we study the purported collaboration between the celebrated Alexander Dumas and the lesser known Auguste Maquet. We find evidence that Maquet may have contributed significantly to the writing in one of Dumas works. We also push the boundaries of stylometry by applying it to poetry from the Romantic period and whether stylometry predicates literary style transcending time. We find that the works of poets resist the rationalisation and reduction of stylometry. We also look at novel and screen adaptations. This results in more questions on literary style depending on other features rather than being consistent to one writer and one specific form. The implications of stylometry and literary form are discussed.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Pepa Birkett, Anisah Khanom, Runlin Chen*)

Example dedication...

Contents

Chapter 1

Introduction

1.1 What is Stylometry?

The statistical analysis of literary style is better known as stylometry. Research in stylometry has established different techniques to quantifiably measure literary texts and such analysis has been applied to different fields. Notably, stylometric analysis is the spearhead of authorship attribution and author profiling [Savoy, 2020]. Throughout the history of literature, there are many texts whose authorship has been contested, for example whether a given play or poem was really written by Shakespeare. Merriam and Matthews (1994) apply stylometry to Shakespeare to discriminate between his works and the works of his contemporary, Christopher Marlowe and later go on to study the purported works of Shakespeare and Fletcher [Merriam and Matthews, 1994]. This application of stylometry isn't just useful in the world of literary fiction; historically politics has benefited from stylometry. Notably where stylometry has been used to study the authorship of the 12 out of 85 unknown Federalist Papers [Monsteller and Wallace, 1963]. In other cases, it is not the exact name of the author that one is looking for but rather identifying certain demographics about the writer. This is known as author profiling and through this we could begin to establish the gender, age range and even social status of the author [Savoy, 2020]. This has further potential and in some cases is used in the criminal justice system - this is known as forensic linguistics where stylometry can be used to establish the criminal profile of a perpetrator.

The purpose of our foray into stylometry is motivated by sheer academic intrigue. We will push the boundaries of the statistical analysis of literature by applying stylometry to different literary forms - novels, poetry and scripts. Chapter 2 details the methodologies that have been used in each of the constituent parts of our study. In Chapter 3 we begin with a more common stylometric analysis of authorship attribution within non-English novels, where we investigate the implications of collaboration or ghost-writing within novels. The rest of the report delves into forms of literature which have been subject to little stylometric research. Chapter 4 takes on the poetic form and questions whether stylometry can be used to see if literary style can transcend time. Chapter 5 concerns a different form of literary analysis altogether in analysing the similarities and differences between novels and their film-adaptations. Through our study we wish to explore the nuances of stylometry when applied to varying literary forms.

1.2 Supervised and Unsupervised Learning

Machine learning models are pertinent tools to explore patterns and verify hypotheses in data sets. There are many different machine learning approaches that are used in stylometry. We begin by providing an overview of the two that we will focus on in this paper - supervised and unsupervised learning. In stylometry, these models aid authorship attribution and author profiling. The key difference between these two approaches is the existence of labels in the training data subset. In essence, Supervised learning involves, as Kotsiantis (2019) suggests, a predetermined output attribute that is to be classified [Allogani et al., 2019]. For example, whether Shakespeare was the real author of a given play. Here, Shakespeare is the predetermined output attribute. Research into authorship attribution comes under supervised learning. In the next chapter we will detail the supervised learning techniques that we have applied for authorship attribution of French novels.

Unsupervised learning on the other hand, is more exploratory. It uses pattern recognition where there is no target attribute. Essentially there is no learning dataset that we know for certain is written by a specific author. Thus, all the information is provided by the sample of texts we have. Clustering algorithms are a technique used in unsupervised learning which attempts to identify patterns and groupings in unlabelled data by determining an intertextual distance measure based on a predefined set of stylistic features [Savoy, 2020]. These features, which are function words in our case, and details on the methods we have used are presented in the next chapter. In this report, we will use both supervised and unsupervised learning techniques in the literary questions we explore.

Chapter 2

Methodologies

The results in this report rely on statistical analyses. In this section we look at how such analyses are implemented. First we study the importance of function words. We then look into methodologies under supervised and unsupervised learning that we will use to perform our stylometric analysis in the proceeding chapters.

2.1 Function Words

The Statistical Analysis of Literature draws together two very different fields. The quantitative study of statistics is, on the surface, in diametric opposition to the qualitative nature of literary studies. Scholars in literature look at literary features - metaphors in prose, variation of lines in poetry, stage directions in scripts to give just a few examples - to build a contextual picture of a piece of literary work. In doing so they can categorise, compare and analyse where a specific text fits in the literary landscape and postulate on authorial intent.

With this, literature, and the literary features of a text, are open to interpretation. These literary features are not, in there elementary form, conducive to a quantitative, statistical analysis. To prepare a text for a statistical dissection we must present it as a mathematical object. For this, we look to the fundamental characteristics of words themselves. Intuitively, one may think that an author's use of uncommon words would make their writing identifiable. However, such words depend heavily on context. This would become quite apparent in our paper. In Chapter 4, with the study of Romantics we will likely see words that are rooted in lofty descriptions of nature or melancholy woes. It is unlikely that such words would appear in the film scripts used in Chapter 5 (or so we'd think!).

Therefore we instead look at the occurence of common words. Monsteller and Wallace looked at so called 'function words' in their study of the Federalist papers [Monsteller and Wallace, 1963]. The set of 70 function words that they used were the most common words in the English Language, covering determiners, prepositions, conjunctions, pronouns, and auxiliary verbs. These function words include, but are not limited to:

> *the, and, of, in, to, with, for, on, by, about, as, at, but, if, or, so, than, that, when, where, which, who, will, shall, should, must, have,*

> *has, had, was, were, been, be, do, does, did, not, no, only, more, some, such*, and others.

The full list of function words used in this study is provided in the appendix 7.4. Function words are context free. It is interesting to note that the variations in how these function words are used by different people is what allows their writing to be identified - the frequency with which one uses the word 'the' or 'be' varies so much so that it can distinguish ones writing style from another. Such words are predominantly used in an unconscious manner and covers a majority of the words used in a given text therefore using function words to quantify the text should suitably reflect the writer's style [Kestemont, 2014]. As our study is an analysis of differing text forms - novels, poems and scripts - we deemed common function words the most appropriate feature choice, applicable in all three of our analyses.

Suppose that we have a list of n function words which are the most common words in the English Language. We can then represent any given text mathematically as an $(n + 1)$-length vector where, for $i \in 1, ...n + 1$, the x_i component of the vector is the frequency of the i^{th} word from the function word list. The x_{n+1} component is the frequency of non-function words. With this, we can construct mathematical objects of each literary text so that they are suitable for further analysis. Let's consider an example applied to the following sentence:

Last night, Ellie finished a book and started a new one.

Suppose our list of function words is: [*last, a, and, one*]. Clearly, in this situation $n = 4$. We can then represent this sentence mathematically as the 5-length vector: $\begin{bmatrix}1 & 2 & 1 & 1 & 6\end{bmatrix}^T$, where the first 4 elements of our vector represent the frequency of each of our chosen function words, and the 5th element is the number of non-function words in the sentence.

It must be noted that for effective comparison of texts of different length, this function word feature vector would provide more accurate and beneficial results after normalisation so that the total sum of features in each vector is 1. This means our feature vectors represent proportions of the words in our text which are function words, rather than a simple count of function words within the text. Therefore, when creating function word vectors for the texts involved in our corpuses, we redefined this vector of counts with the elements as:

$$x'_i = \frac{x_i}{\sum_{i=1}^{n+1} x_i}, \tag{2.1}$$

where x_i is the i^{th} element of our initial count feature vector, $i \in \{1, ..., n+1\}$ and n is the total number of function words we are considering, with the addition of 1 to account for a non-function word count. x'_i is the i^{th} element of the function word vector where the frequency of the corresponding i^{th} word from the function word list is represented as a proportion. Therefore, normalising our example vector from earlier gives the new vector: $\begin{bmatrix}\frac{1}{11} & \frac{2}{11} & \frac{1}{11} & \frac{1}{11} & \frac{6}{11}\end{bmatrix}^T$ where one can check that adding all elements of our vector should sum to 1.

For a set of texts, the vectors make up a matrix, the columns of which are also normalised to have a mean of 0 and standard deviation of 1. It is essential that we normalise both rows and columns of the matrix to ensure that we are looking at the patterns in the data regardless of length of texts and scale.

However, it must be noted that within our supervised learning approach in Chapter 3, we don't standardise over each individual text, but rather a collation of texts representing one author's writing style - this is a procedure more commonly known as the Delta procedure, which will discussed further in Section **??**.

Our study consists of three different function word lists used in three different analyses. Each of the most frequent function word lists chosen within our study lies between 70 and 100 words. This is supported by Burrows (2002) who suggests that a selection of function words larger than 40 gives more accurate results [Burrows, 2002].

2.2 Unsupervised Learning Techniques

2.2.1 Multi-Dimensional Scaling

Groenen and Borg (2005) describe multidimensional scaling as an exploratory technique to reveal structure when studying data that cannot be obviously interpreted [Borg and Groenen, 2005]. Literary texts are a paradigmatic example of such data as they don't follow a rigid numerical structure, instead consisting of words, themes, styles and literary devices that are often open to interpretation. The literary texts that we will look at in the following chapters of this report are converted to data sets as described above and have many features - in fact, each of the function words we have acquired is an individual feature. We often plot data on a 2-dimensional plot or perhaps even a 3-dimensional plot but since we cannot begin to visualise higher dimensional spaces we are limited given that we would ideally want to plot all the features at once, in order to compare and analyse in full.

The way in which multi-dimensional scaling combats this problem is by reducing the dimensionality of the data, in essence, projecting it down into a 2-dimensional space. We are looking to see how close in proximity the literary style of one writer is to another. Multi-dimensional scaling attempts to represent proximities literally - as distances. Suppose we have n observations in a K-dimensional space; if we have 100 function words then $K = 101$. Take two observations $\mathbf{i} = (i_1, ..., i_K)$ and $\mathbf{j} = (j_1, ..., j_K)$ and let $d(\mathbf{i}, \mathbf{j})$ be the distance function that computes the distances between the points. There are a number of ways we could compute the distance but the most frequently used and the one that we will use is the Euclidean distance. If we were in a 2-dimensional space this is simply Pythagoras' theorem for the length of the hypotenuse of a right angled-triangle. This is generalised to the $K-$dimensional case to give

$$d(\mathbf{i}, \mathbf{j}) = \sqrt{\sum_{k=1}^{K} (i_k - j_k)^2}. \tag{2.2}$$

On an MDS plot, the closer in distance two points are corresponds to similarity in literary style of those the texts or authors that these points represent, in terms of the specific features chosen, e.g. function words in our case. Texts by the same author and, therefore usually exhibiting the same literary style, will be clustered together on an MDS plot. The results of this can be used to classify the data. MDS optimises the placement of datapoints to minimise how much information is lost in the distance calculations. This ensures that the most important patterns in similarity or dissimilarity are retained. However it is important to note that this method gives a generic picture of how the data is spaced rather than a clear classification due to the fact that the dimensions are reduced so significantly.

2.2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely employed dimensionality reduction technique that transforms high-dimensional data to a lower-dimensional space while preserving most of the variance [Abdi and Williams, 2010]. PCA is particularly useful in handling datasets with many correlated features, for example function word distributions. For an n sample dataset with K variables, PCA finds a set of orthogonal axes, known as principal components, that have the maximum variance in the dataset. Then, PCA can be performed by calculating the singular value decomposition (SVD) of the centered matrix:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{U}$ contains the left singular vector and $\mathbf{V}$ contains the right singular vector, while $\mathbf{\Lambda}$ is a matrix with singular values on the diagonal. The right singular vector $\mathbf{V}$ are eigenvectors (principle directions). The eigenvectors corresponding to the highest eigenvalues in this context are the directions of maximal variance. There are several built-in functions in R that can be used to perform PCA such as `prcomp()`and `princomp()`. In section 5, function `prcomp()` is employed to reduce the function word distributions from 71-dimensions to plot the data in a two-dimension plane.

2.2.3 K-Means Clustering

While MDS plots provide a visualisation of potential clusters by reducing dimensionality, clustering is used to assign similar data points to the same cluster whilst keeping different objects in separate clusters [Sinaga and Yang, 2020]. Among various clustering methods, K-Means clustering is one of the most popular in partitioning data into meaningful groups [MacQueen, 1967].

K-Means Clustering works through an iterative algorithm that separates a dataset into k clusters by minimizing the variance within each cluster. This is also known as within-cluster variance. The number of clusters is defined depending on the data. The algorithm first chooses k initial cluster centers, $\mu_1, \mu_2, \ldots, \mu_k$, randomly from the dataset. Then each data point x_i, is assigned to the cluster with the center that it is closest to:

$$C_j = \{x_i \mid \arg\min_j d(x_i, \mu_j)\} \tag{2.3}$$

where $d(x_i, \mu_j)$ represents the Euclidean distance between data point x_i and cluster center μ_j. The new center of each cluster is then defined by taking the average of all the points that are assigned to it:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{2.4}$$

where $|C_j|$ is the number of points in cluster C_j. This is repeated until the centers are stabilised; in essense, when little or no changes occur in cluster assignments between iterations.

2.2.4 Bootstrap Methodology

The Bootstrap resampling method is a statistical technique used to test for a significant difference between two groups. It is a technique that approximates the sampling distribution of a statistic by resampling with replacement from the original dataset repeatedly [Mooney, 1996].

In section 5, we apply bootstrap resampling to two groups. We generate $B = 1000$ bootstrap samples by randomly picking observations from Group B with replacement. Let $\bar{X}_B^{(b)}$ represent the sample mean of group B and $\bar{X}_A^{(b)}$ represent the sample mean of Group A in the b^{th} bootstrap iteration. We define $\hat{\delta}^{(b)}$ as the mean difference between the two groups in a given bootstrap iteration b. Thus, for each resample, the mean difference can be computed:

$$\hat{\delta}^{(b)} = \bar{X}_B^{(b)} - \bar{X}_A^{(b)}, \quad \text{for } b = 1, 2, \ldots, B$$

The distribution of $\hat{\delta}^{(b)}$ gives an approximation of the variability in the observed difference. We can also find a bootstrap confidence interval for δ:

$$[\hat{\delta}^{(\alpha/2)}, \hat{\delta}^{(1-\alpha/2)}]$$

where $\hat{\delta}^{(\alpha/2)}$ and $\hat{\delta}^{(1-\alpha/2)}$ correspond to the lower and upper percentiles of the bootstrap distribution. This methods helps to test if one group shows higher values than another group in terms of their mean difference.

2.2.5 Hierarchical Clustering

Hierarchical clustering is another unsupervised learning technique. We have seen how implementing Multi-dimensional Scaling will create spatial representations of the data while reducing the dimensionality. K-Means Clustering, on the other hand, partitions the data into k clusters, where k is specified before running the algorithm. For large datasets this should create well-separated clusters. Comparatively, hierarchical clustering does not require a pre-defined k. For small datasets, this type of clustering identifies the hierarchical relationships through either an agglomerative approach or a divisive approach [Nielsen, 2016].

The Agglomerative or Bottom-Up Approach begins by considering each data point as its own cluster. Iteratively, it merges the two closest clusters together until only one cluster remains. The Divisive or Top-Down Approach starts with all the data points as one cluster. It then recursively splits this into smaller clusters and stops when each data point is its own cluster.

Hierarchical clustering creates tree diagrams called dendrograms which highlights the nested clustering. In Chapter 4 we use the `hclust` function in `R` to implement agglomerative hierarchical clustering. This function can use different agglomeration methods that measure the dissimilarity between two clusters of observations. We have used Ward's method (`ward.D()` in `R`) which minimises the total within-cluster variance.

2.3 Supervised Learning Techniques

Diverging from the previous unsupervised learning methodologies, Chapter 3 of our study uses supervised learning techniques for text classification.

2.3.1 K-Nearest Neighbours (KNN)

Specifically, we will be discussing the methodology behind the non-parametric K-Nearest Neighbours algorithm. KNN uses distance metrics to classify the objects in question.

The basic mathematical framework behind KNN is as follows: assume we have a dataset consisting of pairs (x_i, y_i) where $x_i \in \mathbb{R}^K$ is a K-dimensional feature vector, and y_i is a classification label, with $i \in \{1, ..., n\}$ with n being the number of observations in our training dataset. Given a new observation with feature vector $\tilde{x}$, the goal is to determine its classification label $\tilde{y}$ using KNN [Chong et al., 2021].

To find the k-nearest neighbours, we compute the distance between $\tilde{x}$ and every observation x_i in the training set using a specified distance metric. KNN then assigns $\tilde{x}$ to a label $\tilde{y}$ based on the class of its closest observation. In mathematical notation, for some distance function $d(x_i, \tilde{x})$, we let

$$r = \arg\min_i d(x_i, \tilde{x}).$$

This is then used to predict $\tilde{y} = y_r$, where $\arg\min$ denotes the index of the minimum value.

Before using KNN, to ensure all features contribute equally to the distance measure, accounting for the difference in length of our texts via normalisation of the function word count is necessary. As discussed in section 2.1, in our study we have chosen to normalise using the relative occurence frequency of function words, as given by Equation (2.1).

To address the high computational cost of computing the distance between every text in the dataset, the dataset can be adapted to Burrow's (2002) Delta context which has proven effective in authorship attribution, particularly when dealing with large corpora [Burrows, 2002]. Specifically, instead of considering individual texts, we sum the word frequencies of every text in a selected group of texts. The characterisation of these text groups can differ depending on one's research question. For example, in authorial classification, a group of texts may represent a given author, as is done in this paper. Other examples of possible groupings are genre or time period.

The main idea behind the Delta procedure, as initially proposed by Burrows (2002), is to create a composite profile of an author by summing the word frequencies across all texts by that author [Burrows, 2002]. Once the word frequencies of each author are aggregated, we then normalise these profiles, such that each row in our training set now represents one author instead of individual texts. Thus, when applying KNN, each observation x_i would represent a group of texts/author. Note that Burrows' (2002) use of the term Delta refers to the distance metric used to calculate the distance between a single text and a grouping of texts, thus in our case is the Euclidean distance [Burrows, 2002].

2.3.2 Cross-Validation

Above we have seen methods that attempt to classify the data into clusters. Cross-validation is a way to evaluate the performance of classification algorithms. Broadly it helps estimate how well a model can be generalised to unseen data. A model is a mathematical representation of relationships within the data [Wong, 2015]. First, the dataset is split into multiple subsets or folds. We then train the model that we are using with a subset of the data. The remaining data is used to test the model's performance. This process is repeated with a different fold used as the test set each time. The type of cross-validation that is used is usually chosen with regard to how much data is available. Often for large data sets, k-fold cross validation is used where the dataset is split into k equally sized subsets and the model is trained on $k-1$ folds and tested on the remaining fold.

Leave-One-Out Cross Validation (LOOCV) is a special form of k-fold cross validation where each observation in the dataset is 'left out' to be used once as a test set and the rest of the data serves as the training set. Intuitively, within the corpora used in this report, LOOCV iterates over each author, then over each of their texts. For each iteration, we assign a single text to a test set, letting all remaining texts in the corpus form the training set. The text is then classified and this process is repeated for the remaining texts in the corpus. While it certainly maximises the data usage, LOOCV is computationally expensive and thus used for smaller data sets. In a broader context, our datasets are relatively small and so this report favours the use of LOOCV to assess accuracy of the models used.

Chapter 3

Friends or Foes? A Stylometric Analysis of Maquet's Involvement in Dumas' Works

3.1 Background and Motivation

Alexandre Dumas père is a world renowned novelist, a pioneer in French romantic historical fiction and, due to his popularity, his books have been translated into a large number of languages to be sold the world over [Martone, 2020]. However, the analysis that follows considers the initial language his works were written in - French. Studies show that stylometry appears to be effective amongst all natural languages based on letters, not just English [Savoy, 2020]. In light of the 2010 French film *L'Autre Dumas*, this movie highlights the relatively unknown, amongst the general public, subject of Auguste Maquet's involvement in many of Dumas' most famous works [Nebbou, 2010]. In fact, the belief that Dumas used a ghost-writer, or a collaborator, is a topic of disagreement within the scholarly world, some diminishing Maquet's role to that of an assistant only providing the bare minimum on which Dumas could craft his chef-d'oeuvres, while others believe that without Maquet, these masterpieces frankly were not a possibility [Davies, 2010]. The extent of Maquet's input in Dumas' novels is not certain, with lots of debate over how much can be said to have been written by whom, but there is evidence to suggest that Maquet often wrote the first drafts and came up with much of the plot of these novels, before Dumas then embellished on elements of the chapters first written by Maquet [Paraschas, 2018].This debate over authorship has also been reflected in the publishing history of many of Dumas' works written during the known collaboration period throughout the 1840s, with some publishers crediting both authors whilst a large number only credit Dumas.

With these insights in mind, this chapter provides a quantitative stylometric analysis of some of Dumas' most famous works, attempting to give insight into the collaborative nature of these novels, and whether it may be possible to predict whether some chapters of *Le Vicomte de Bragelonne*, the third novel of the

D'Artagnan Romances: The Three Musketeers, were in fact predominantly written by Dumas or Maquet [Maquet, 2004]. Unlike classic authorship attribution problems in which we want to identify a single author of a previously unknown document, we want to classify the chapters of a known collaborated work to see if our findings suggest whether we can attribute a chapter to only one of the authors involved. This research aim is based on controversial literature suggesting that some of the chapters of Dumas' most famous works may have only had a small proportion of the words changed by Dumas when going from Maquet's original drafts to the final published text. We will also consider the limitations of this stylometric exploration and the challenges that arise in attributing authorship in collaborative works.

In Sarah Mombert's (2022) article, Mombert details how, due to *Le Vicomte de Bragelonne* being twice as long as origianlly planned, at times Maquet ended up replacing Dumas when providing a finalised draft for publishing within the serialisation of the novel for newspaper *Le Siècle* [Mombert, 2022]. It must be noted that most of Maquet's own work was predominantly included in later volumes of the novel. The theory that a number of chapters were written solely by Maquet in serialised novels for *Le Siècle* or *La Presse* is again supported in Callet-Bianco's (2020) article *Dumas et alii. L'écriture en collaboration* [Callet-Bianco, 2020]. In fact, Claude Schopp's (1991) edition of *Les Trois Mousquetaires/Vingt Ans Après* provides a list of chapters that can be attributed solely to Maquet [Schopp, 1991]. Unfortunately, we do not have access to this list, so to improve the reliability of the outcomes of our study, it would be beneficial to compare the chapters we predicted to be written by Maquet to those Schopp details were written by Maquet.

Using common French function words of the 19th century as extraction features, the analysis begins with unsupervised learning methods, using multi-dimensional scaling (MDS) plots to explore patterns between a variety of French historical fiction novelists from the French romantic literary period of the 19th century, with particular focus on Dumas and Maquet's placements within the plots. Secondly, based on these findings, as well as literature surrounding the collaborative nature of Dumas' works, we turn to supervised learning techniques to perform a chapter-by-chapter analysis of the third novel in the *D'Artagnan* Romances series (more commonly known as *The Three Musketeers* trilogy): *Le Vicomte de Bragelonne* . Finally, we examine how our experimental results impact authorship attribution and whether these results give further insight into whether a chapter-by-chapter analysis can determine to some extent Maquet's role within one of Dumas' works. We seek to answer the question of whether our analysis confirms the findings of previous studies or existing literary scholarship.

3.2 Data

Collection and Preprocessing

To begin our analysis on the works of Dumas, we considered 18 novels written during the Romantic Period, specifically in the mid-19th century. Now over 100 years old, the copyright terms for these first publications have expired and are therefore in the public domain. Using the online library Project Gutenberg, known to be a source of readily accessible classic literary works, we obtained texts from the following authors: Alexandre Dumas, Auguste Maquet, Victor Hugo and Stendhal [Project Gutenberg, 2025]. These authors were chosen for initial exploratory analysis due to their similarities in genre - all novels are classified as historical romanticism within the literary world. In addition, the original publication dates for all novels were within a 40 year period, between the years of 1830 and 1869, meaning we did not have to account for stylistic changes over time. A significant limitation within this study is that acquiring texts written solely by Dumas in which he didn't collaborate with other lesser-known novelists has proved to be a challenge due to the nature of his writing process [Callet-Bianco, 2020]. As such, the selection of texts representing Dumas' works pre-collaboration may skew resulting analysis due to collaboration with authors other than Maquet. See appendix 7.1 for a list of novels used for each author, including the editions used in this study as well as the publication dates.

Each novel is a separate file in plain text form, with any additional information removed, such as content and title pages, information about the source from which the text was required, chapter titles, prefaces, and notes. This was done by hand to ensure no actual text from the original novel was deleted.

In addition to having text files of each novel, we also performed a chapter-by-chapter analysis. This was favoured over dividing the text into 500 word chunks as our research aim is to classify whether Maquet may have solely written certain chapters of *Le Vicomte de Bragelonne*, as supported by the literature mentioned in Section 3.1. We did this for both Hugo's *Notre-Dame de Paris* - in English, *The Hunchback of Notre-Dame* - and Dumas' *Le Vicomte de Bragelonne.* The choice to analyse *Notre-Dame de Paris* first was to assess whether our model would correctly attribute all chapters to Victor Hugo, and thus in this scenario although the authorship of *Notre-Dame de Paris* is undisputed, we still treated the novel as having unknown authorship. This served as a preliminary test to ensure that our stylometric approach should work, not only at the novel level, but also on a chapter-by-chapter level, and could therefore be applied to the novel of actual disputed authorship *Le Vicomte de Bragelonne.* By hand we created separate plain text files of each chapter of these novels, again removing chapter titles. It should be noted that three chapters of *Notre-Dame de Paris* were omitted from the corpus as they contained less than 500 words. Based on previous stylometric studies, texts with higher word counts produce more accurate results [Mosteller and Wallace, 1963]. On reflection, seeing as there are over 50 chapters to consider for the novels chosen, we used R to create files of batches of 5 consecutive chapters and stored these in new text files. This method of text segmentation meant our initial visualisations of the data using MDS plots were easier to interpret due to less noise, as well as potentially more accurate results due to the higher word count of each chapter batch. The limitation to this is that we are now unable to classify who wrote each chapter, but instead whether it appears that Maquet may have written a selection of 5 consecutive chapters.

To preprocess the data, for each plain text file we used R to replace all punctuation with spaces. The French written language has the addition of accents on letters, so we ensured these were not deleted when removing punctuation. We also replaced multiple spaces with a single space, trimming any leading and trailing whitespaces, converted the entire text file to lowercase, and then tokenised the string into words using spaces as separators to create a vector of words for each text file of all the words in that novel/chapter.

The dataset used for our machine-learning model includes 4 authors, along with an 'Unknown' folder containing chapters of the book *Le Vicomte de Bragelonne.* Each author is a separate folder containing word count proportion vectors of the 100 most frequent French function words for each of their novels included in this study, with the function words ordered identically in each feature vector. This file setup intentionally aligns with easier application of the Delta method we are using for authorship attribution, as mentioned in Section 2.3.1.

Function Word Selection

As mentioned in Section 2.1, throughout this report we carry out our analysis based on common function words. In this chapter, we opt to use a list provided by the Eduscol, Ministère de L'Éducation National, De L'Enseignement Supérieur et de la Recherche, which details a hierarchical table of the most frequent words of the French written language from the 19th and 20th centuries, to extract our choice of function words from [Eduscol, nd]. As the table was hierarchical in terms of their frequency, we took the first 100 words from this table, ensuring not to include words that may skew our data. Within the French language, unlike English, this entails excluding gendered articles such as 'le' or 'la'. The equivalent word in English is 'the', and therefore would not introduce bias into the data based on whether the novel/chapter being considered is centered around a female or male object. We also had to exclude some nouns, such as 'homme' (man) or 'deux' (two), for the reason that it is gendered or context-specific, and so again may skew our analysis slightly. Finally, there were some words with repeated spellings in the data set, but different grammatical functions. We only included each of these words once as when using R to count our function words we are not able to account for the grammatical function of each word, only the spelling. Note that more thorough stylometric analyses often tag homonyms to distinguish between them, as Burrows (2002) does within his analysis [Burrows, 2002]. The full list of function words for this chapter of the study is provided in Appendix 7.2.

As mentioned in Chapter 2.1, all function word vectors include a 101^{st}component counting the number of non-function words in the text. This allows for the normalisation of each function word vector so that our function word vectors now represent proportions of function words within a text, rather than counts.

Burrow's Delta Context

When we implement the supervised learning technique of KNN to consider the true author of a chapter-by-chapter analysis of one of Dumas' novels, we consider very large texts consisting of novels of numerous volumes, and as such, we deemed the Delta context, mentioned in Section 2.3.1, an appropriate method to yield accurate results for our analysis due to the word count of these novels significantly surpassing $1,500$ words. In fact, most chapters of the unknown novel in question surpass 1,500 words.

Using the mathematical notation given for KNN

3.3 Results

MDS of French Historical Romance Novels

In this section of our study, we begin exploring the works of French Romantic authors through multi-dimensional scaling, as detailed in Section 2.2.1, to determine whether there is separation between Dumas' novels before his collaboration period with Maquet, Dumas' novels during his collaboration period between the years of 1841 and 1858, and Maquet's own novels [Mombert, 2022, Paraschas, 2018]. We also include novels by French historical romance authors Stendhal and Hugo to ensure the MDS plot has been created effectively, as these works have definitively known authors and thus should both cluster separately within a plot. A visualisation of the intertextual distances provides us with a basis upon which we can extend our analysis. As mentioned in Chapter 2.2.1, the limitations of dimensionality reduction must be noted.

If natural clusters appear separating authors on the plot, then it would be reasonable to assume we can continue our stylometric analysis assessing Dumas' most famous works.

In Fig. 3.1, each novel has been colour-coded based on the author of the text, with Dumas' work having been separated into work known to be written before his collaboration with Maquet and those texts under Collab, as seen in the legend, representing texts written by Dumas during his collaboration period with Maquet. From now on, for the purpose of ease within our written analysis, 'Collab' will simply be used as the author when referencing work that may have been written by both Dumas and Maquet. As briefly mentioned in Section 3.2, Hugo and Stendhal's novels are used throughout this chapter to evaluate whether the methods applied are effective and make sense.

From Fig.3.1, the majority of the novels are clustering as would be expected, emphasised by the works of Stendhal and Hugo which appear to be in separate distinct clusters away from the main authors of focus of this study. Notably, the Collab novels appear to exist in their own cluster, away from the rest of Dumas' work, suggesting that Dumas' writing style was perhaps influenced by Maquet's input. Finally, the more scattered points representing the intertextual distances between Dumas' pre-collaboration novels suggest that Dumas' novels, many a product of collaboration with other lesser-known authors, may not show similarities in writing style based on frequent function words [Callet-Bianco, 2020]. This is potentially due to the effect of Dumas' collaboration with other authors. As such, we will not consider Dumas' selection of novels within this study as a reliable reflection of works written solely by Dumas. However, their distance from the Collab novels in the plot do suggest that Dumas' novels when collaborating with Maquet are distinct from his previous works of collaboration with other authors.

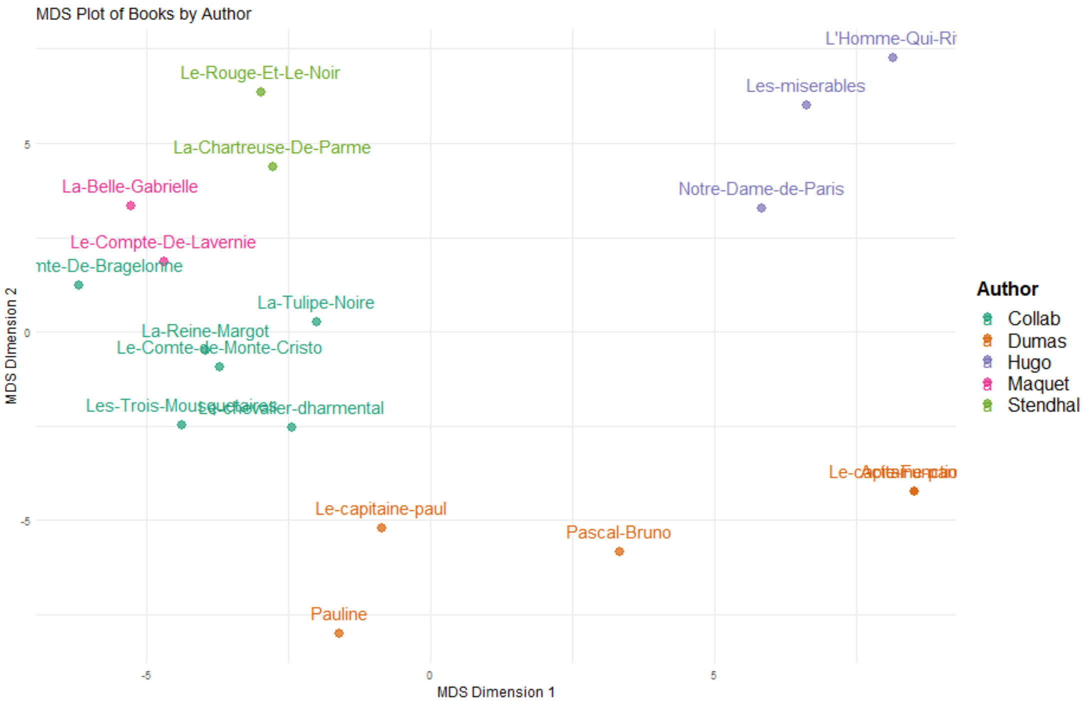

Figure 3.1: MDS plot of novels by author

KNN Model for Authorship Attribution of French Historical Romance Novels

To train a model to learn a representation for each possible author, we decided to remove those novels written by Dumas before the collaboration period. This is due to the novels not showing similarity in writing style based on Fig. 3.1. Our dataset thus consists of novels by Collab, Hugo, Maquet, and Stendhal. Each author has more than one text in their corresponding folders and, thus, our training set should never be non-empty, as is required when implementing the KNN algorithm.

To assess performance of the classifiers, we first create a test set consisting of a random single book from each of the 4 authors in our corpus, ensuring to remove these from our training set. We then applied KNN as detailed in Section 2.3.1. We do not apply discriminant analysis due to the small size of our test set.

To validate our model, Leave-One-Out Cross-Validation was performed to ensure that every text is tested individually using the remaining texts for training, to ensure a robust evaluation of our classification. Our code showed that KNN achieved 92.3% accuracy, suggesting KNN performed well on this corpus, with the following confusion matrix given as the output:

	Reference			
Prediction	Collab	Hugo	Maquet	Stendhal
Collab	5	0	0	0
Hugo	0	3	0	0
Maquet	1	0	2	0
Stendhal	0	0	0	2

Table 3.1: Confusion Matrix of LOOCV of KNN

According to this matrix, only one novel does not lie on the diagonal, and therefore only one novel has been misclassified. In fact, although statistically this is considered a misclassification, if we consider our research aim to investigate one of the Collab novels, *Le Vicomte de Bragelonne*, and Maquet's potential involvement in the solo authorship of some chapters, this prediction that one of the Collab texts was written by Maquet could support this theory.

MDS Plot For Visualisation of Chapter Similarities

Following this general exploratory analysis, the intention remains to delve further into exploring Maquet's inputs towards certain chapters of the novel *Le Vicomte de Bragelonne*. However, we first check the output of an MDS plot visualising the distance between chapters of Hugo's *Notre-Dame de Paris* and our previous novels in the corpus to ensure a chapter-by-chapter analysis is an appropriate method. Since it is known that *Notre-Dame De Paris* was written by Hugo and Hugo alone, it would be expected that when separated by chapter, each chapter should be near other Hugo novels in the plot. When originally performing this, the resulting plot was not easily read due to the large number of chapters being considered, 59, and thus the smaller word count of the text being used to create our frequent word proportion vectors. This was also the case for Dumas' novel, *Le Vicomte de Bragelonne*. Thus, we adapted our function word proportion vectors to represent batches of 5 consecutive chapters and got the MDS plot depicted in Fig. 3.2.

From the plot, it is clear that although there is variety in the proportion of frequent function words in the chapters of *Notre-Dame de Paris*, the majority of 5-chapter batches are closest to other novels by Hugo in distance, suggesting that even with dimensionality reduction, there is still visual evidence that a more in-depth chapter-by-chapter analysis of our novels will provide some interesting insight into a prediction of who wrote what chapters.

Following this, the MDS plot in Fig. 3.3 visualises the stylistic similarities and differences between the chapters of *Le Vicomte de Bragelonne*, also grouped in batches of five consecutive chapters, and other novels written during this period of French history. In fact the decision to explore this specific novel is supported within Fig. 3.1 as *Le Vicomte de Bragelonne* could be interpreted as being part of Maquet's clusters of works. This is interesting as literature suggests that this is one of the novels in which Maquet was the sole author of some of its chapters.

As expected, due to the collaborative nature of this book the resulting plot does not appear to have as distinct clustering as the previous plot of Fig. 3.2.

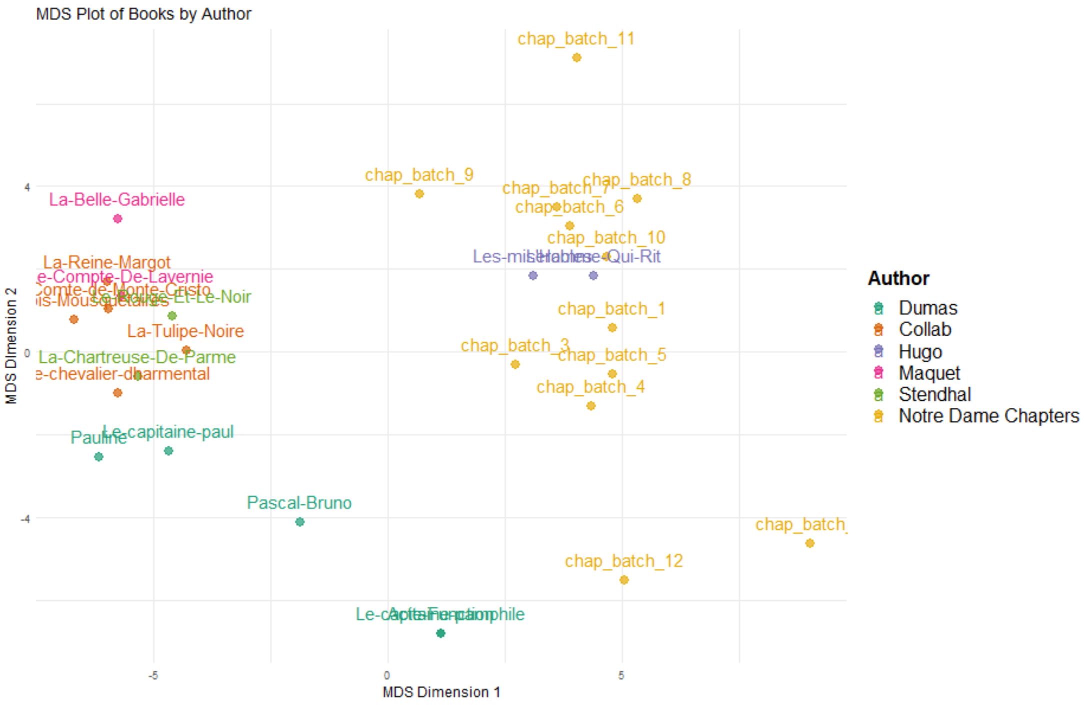

Figure 3.2: MDS of novels and chapter batches of Hugo's *Notre-Dame De Paris*

However, the chapters appear closer in similarity, based on proportion of function words in the texts, to Maquet's novels and the Collab novels, with the exception of chapter batches 14, 17, 8 and 10. These exceptions could be a consequence of the reduction in dimensionality and the effect of collaborative writing on writing style. Furthermore, we initially considered the idea that some individual chapters could have been written solely by Maquet, however our plots have grouped chapters in batches of five, another factor that could create misleading results in our MDS plot.

Finally, the sparsity of the chapter batches in our diagram could speak to the notion that a stylometric analysis by chapter of a collaborated work may not give as distinct and obvious results in comparison to works written by one author, as above in Fig. 3.2. However, if Maquet did in fact write some chapters of the novel alone, the expectation would be for these chapters to cluster together within the plot away from those chapters which were the outcome of collaborative writing. Although the plot does not provide this outcome, it is clear that some batches of chapters group closer to novels by Maquet and so in the next section we will implement supervised learning to attempt to classify the chapters of *Le Vicomte de Bragelonne* .

Supervised Learning of *Le Vicomte de Bragelonne*

In the previous section, our MDS plots revealed those novels with the greatest stylistic similarities to one another, and to the chapter groupings of the novels specified earlier. Now, we will explore our chapter-by-chapter analysis fur-

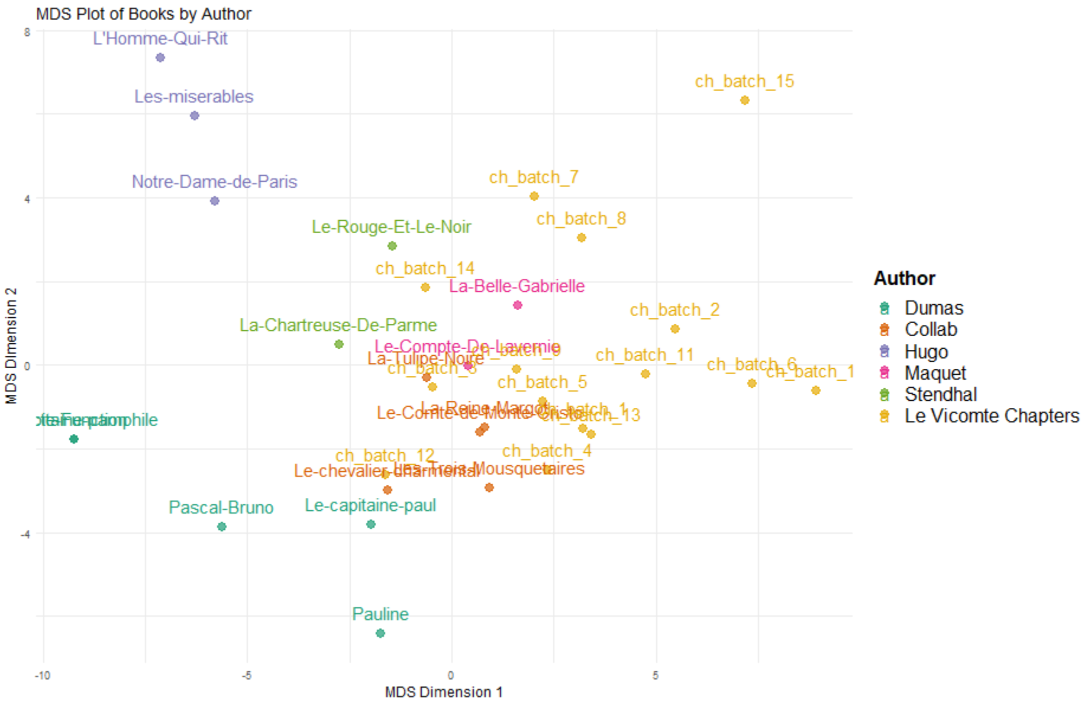

Figure 3.3: MDS of novels as well as chapters of disputed novel *Le Vicomte de Bragelonne*

ther using a supervised machine-learning model, specifically applying K-Nearest Neighbours to our corpus of books.

This leaves us to use our cross-validated trained model to predict the authorship of the unknown chapters of *Le Vicomte de Bragelonne*, one of the novels deemed a collaborative piece by Dumas and Maquet. The difference is that the test dataset now consists of chapters with uncertain authorship rather than known novels. This was briefly tested on the chapters of *Notre-Dame de Paris*, even though the authorship is known, to give brief insight into whether it could correctly predict chapters of a known text. Note that chapters are grouped in batches of 5, as previously mentioned. The resulting classification was that of the 12 chapter batches of *Notre-Dame de Paris*, 11 of them were correctly predicted as Hugo.

Following this, we move to prediction of the chapter batches of *Le Vicomte de Bragelonne* . The output from our analysis was the following:

The predictions above suggest that the chapters of this novel exhibit significant stylometric similarities with multiple authors, with multiple chapters classified to have been written solely by Maquet, whilst some chapters are predicted as the Collab author (presumed to be predominantly Dumas, but with Maquet's collaborative input as mentioned in literature). Note that chapter batch 14 is a definitive misclassification, predicting Hugo, yet we know from literature that either Maquet is the sole author or it is a collaboration of both Dumas and Maquet. To improve the above trained model, it would be beneficial to have a selection of texts deemed solely to have been written by Dumas. Unfortunately, this has not

Chapter Batch	Assigned Author
1	Maquet
2	Maquet
3	Maquet
4	Maquet
5	Collab
6	Maquet
7	Maquet
8	Collab
9	Collab
10	Maquet
11	Collab
12	Collab
13	Maquet
14	Hugo
15	Maquet

Table 3.2: KNN Author Attribution for Each Chapter

been possible in our case due to Dumas' novels not showing consistent stylometric similarities, most likely due to the collaborative nature of most of his works. Regardless, based on the high accuracy of our model, it could be suggested that Maquet has a dominant presence in the writing of *Le Vicomte de Bragelonne* as 9 out of the 15 chapter batches are predicted to have been written by Maquet, rather than classified as a collaborative authorship.

Within literary discourse, it is believed that Maquet was definitively the sole author of the last chapter, which our data seems to reflect, in fact suggesting Maquet was the author of the last 5 chapters of the novel. However, interestingly the data suggests that a siginificant proportion of the first volume of this novel may have been written by Maquet, rather than Dumas.

3.4 Discussion

In this section of our study, the main aim was to explore the collaborative nature of Alexandre Dumas' works, with a particular focus on whether certain chapters of *Le Vicomte de Bragelonne* could be attributed to Auguste Maquet, a topic of debate in the world of literature. Through the use of unsupervised learning techniques, we gained initial insight into the stylistic similarities and differences between Dumas' works, Maquet's novels, and other authors from the French Romantic period. Our MDS plots revealed potential clusters of authorship, which helped contextualise the relationship between between Dumas and Maquet in the larger literary landscape. This initial analysis asserted that, although there is some stylistic overlap, there are distinctive patterns in each authors' works when using frequent French function words as the feature on which our analysis is based.

The supervised learning analysis applied to a chapter-by-chapter breakdown of *Le Vicomte de Bragelonne* allowed us to explore the specific contributions Maquet may have made to this novel, when compared with Dumas' other known collaborated works with the author. In line with previous literature, our results suggest that certain chapter batches in *Le Vicomte de Bragelonne* exhibit stylistic markers that are more closely aligned with Maquet's writing than with the previous collaborative works of Dumas and Maquet, enforcing the theory that Maquet may have been the sole author of some of its chapters.

While our analysis contributes new perspectives to the discourse on literary collaboration in the 19th century, it is essential to recognise the limitations of our study. One such limitation is that common authorship attribution techniques are known to be less effective when used on collaboratively written text, especially in cases where the boundary between one author's contribution and another is blurred [Dauber et al., 2017]. This encouraged our decision to focus on chapters where there is evidence of sole authorship by Maquet in the novel of our focus. In addition to this, although the use of function words as features is an attractive choice for authorship attribution, as mentioned in Section 2.1, additional choices of features could provide a more holistic analysis on the stylistic elements used to define an author's style. Furthermore, the exclusion of some chapters due to word constraints and the absence of direct access to Schopp's list of Maquet's chapters limits the scope of our analysis. Future studies could expand upon this work by incorporating more sophisticated features and building upon the corpus by including more works by both authors. It would also be beneficial to consider the influence of other external collaborators used in Dumas' other works to provide a fuller picture of Dumas' written works over time.

Overall, our results indicate that Maquet's contributions, often overshadowed by Dumas' fame, may have been more significant than traditionally acknowledged, especially in certain chapters of *Le Vicomte de Bragelonne.* While there are significant limitations to our study, we hope to encourage further research on collaborative authorship attribution techniques, linking computational analysis with the world of literature.

Chapter 4

Ode to Statistical Analysis: A Stylometric Exploration of the Romantic Literary Period

“A mathematician who is not also something of a poet will never be a complete mathematician.”
Karl Weierstrass

4.1 Background

Does literary style transcend time?

Literary criticism of the Romantic period explores ideas relating to the key defining features of the time - which often includes reverence for the natural world, a focus on emotion over reason and rejection of urban civilisation. As Hahn posits, the Romantic view regards “the natural world as a living mirror to the soul, not as dead matter for scientific dissection”[Wellek, 1949]. Ironically what we are to conduct in this report is the “scientific dissection” of a corpus of works by the Romantic poets.

Romanticism can be seen as both an aesthetic and a movement. That is, when referring to Romanticism some may be speaking of the aesthetic literary style that involves ‘Romantic’ ideas or the movement influenced by the French Revolution. In this report, we are asking whether the aesthetic face of Romanticism exists outside the constraints of time. In essence, through our statistical analyses, we want to use the Romantic period as a landscape to explore whether literary style transcends time.

The Romantic Period roughly spanned the years between 1798-1837. When specifically looking at British Romanticism, the world of literature was dominated by 6 key figures and there is agreement that these poets are split into two groups - the ‘early’ Romantics, William Blake, William Wordsworth and Samuel Taylor Coleridge and the ‘later’ generation, John Keats, Percy Bysshe Shelley and Lord Byron. Our stylometric analysis is aimed at establishing groupings of

Romantic poetry - whether there is any linguistic evidence for the literary critical views concerning the constituent periods of Romanticism. Using our specifically constructed corpus of Romantic poetry, we will determine whether there is evidence of change in Romantic style over time. We might intuitively assume that the three late Romantic poets share more lexical characteristics with each other than they do with the works of the early poets, the stylometric analysis reveals whether this is actually the case.

Stylometric approach to Yeats' poems

In Ross' (2020) paper, *Tracking the evolution of literary style*, he considers a "potential limitation of most statistical approaches to stylometry is the assumption that style of an author remains constant over time" [Ross, 2020]. McIntyre and Walker's (2022) study of Yeats' poetic style, the inspiration for this chapter, over time pushes the boundaries of this potential limitation [McIntyre and Walker, 2022]. They use both stylometric techniques and corpus stylistics to establish groupings of Yeats' volumes of poetry. With this, they determine whether there is any evidence for the critical views that suggested Yeats' style changed over his writing career and that these changes fall into three distinct phases.

McIntyre and Walker [McIntyre and Walker, 2022] used different stylometric methods to paint a clearer picture; notably using Cluster Analysis and specifically Burrow's method of clustering to see which of Yeats' texts are grouped together as well as Principal Component Analysis (PCA) to reduce the complexity of the data and distinguish the texts from each other. In addition to and motivated by the results of the stylometric analysis they also used corpus stylistics, specifically Keyword Analysis to identify the unusually high or low frequencies of particular words which revealed interesting conclusions. Ultimately the multivariate analysis provided support for the literary critical views that Yeats' poetry can be divided into three stylistic periods. Yeats', particularly in his earlier writing phases, was heavily influenced by the works of the Romantics who preceded him. From the study detailed above, we extended the question on whether literary style changes over time for one poet to a group of poets within a particular literary genre.

4.2 Procedure

Constructing and cleaning the dataset

To explore whether Romantic literary style transcends time we first construct a corpus of poems for each poet. It has been over 70 years since the death of each of our poets, similar to the novels used in Chapter 3, the poems that make up our dataset are in the public domain. We, at first, obtained 10 poems for each poet from the online library, Project Gutenberg. Upon closer inspection, the texts varied dramatically in length. For instance, Blake's *Songs of Innocence and Experience* are made up of relatively very short poems. Such short poems would cause inconsistencies in our MDS plot as there would be insufficient values to calculate the distances. Thus we have altered the corpuses such that some short poems have been compiled together to create text files with a minimum of 500 words. For every poet, where we have needed to compile poems together, we have selected poems from the same works or written in the same year. For example B1, the first text for Byron in our data set is made up of 3 poems from his works *Fugitive Pieces* - written in 1806. Blake's Songs of Innocence and Experince, an anthology of 28 poems have been split into 'texts' of 5 poems, in the order in which they appear in the songs. We have encoded each poem or compilation of poems, for ease of analysis. The texts we have chosen are displayed in appendix 7.3, ordered by poet and including the publication dates. We have also included 9 texts from 3 different Modernist poets as a test set to ensure that our stylometric techniques are robust.

The process of cleaning our data was implemented both manually and using R. Unlike in Chapter 3, where it was important to leave the accented punctuation in the French words, we used an R script to remove all punctuation and other formatting such as trailing white spaces. For many of the poems, the stanzas are labelled using Roman numerals and many footnotes which also had to be removed meticulously by hand.

Function Word Selection

Often, as we see mentioned in Section 2.1, it is the frequency of these seemingly meaningless words that give way to characterising literary style. However, in our analysis we are not trying to deduce whether for example, there is a Keats poem that is contested as being written by Wordsworth. We are instead comparing the literary styles of each poet to see if we can establish a different classification to that of time. Thus, our function words were obtained by creating our own list of the most common words from canonical 18th Century texts, as shown below.

the, and, to, of, a, in, that, was, it, he, for, with, as, on, his, at, by, an, which, not, but, from, this, or, had, be, were, we, have, can, all, so, if, when, no, do, my, their, me, him, then, her, them, now, our, any, out, what, up, some, into, could, how, there, been, would, is, am, are, will, one, they, you, thy, thine, shall, should, yet, doth, did, may, must, ought, like, unto, about, such, thus, even, where, why, through, upon, against, without, over, above, below, among, between, within, beyond

Table 4.1: A list of the 100 most common words from the 18th Century.

We then input the list of function words into an R Script that counts the prevalence of these words in the texts data sets. Each text in the corpus is represented as a length 101 vector where each entry is the frequency of the corresponding function word. The poem vectors make up a normalised matrix, as described in section 2.1, which we can then use for our statistical and stylometric analysis

Stylometric and statistical techniques

In the Romantic vein of unifying supposed dichotomies, and as with the French Historical Romance Novels in Chapter 3, we again bring together both supervised and unsupervised learning techniques. Primarily, the exploratory question of whether Romantic literary style transcends time would be subjected to an unsupervised learning approach since we look to see if the clusters of early and late generation Romantic poets emerge naturally.

As Hofman posits, unsupervised learning algorithms are suitable for creating the labels in the data that are then used to implement supervised learning tasks [Allogani et al., 2019]. Thus, we use Multi-Dimensional Scaling as a dimensionality reduction technique to see if the Big Six Romantic poets fall into the natural groups of early and late. As detailed in section 2.2.1, MDS allows us to visualise the texts from each poet compared to the others as it calculates the distances between all the pairs of texts and reduces it to a two-dimensional projection of the distance matrix - thus the closer the texts corresponds to how similar they are in literary style. We also use agglomerative hierarchical clustering, explained in section 2.2.5 to see if we can identify an early and late cluster or rather if any other clustering appears.

We then use Cross-Validation, a supervised learning technique, to test the validity of the purported early and late grouping. For this, we have two classes that contain the early and late poems respectively and at each iteration we leave one poem out as our test set, retrain the remaining data and see which class the test poem falls into. We acknowledge the limitation that arises through leaving each text out at each iteration rather than each poet. Although this may intuitively seem to create an over-estimation when classifying the poems, the rationale behind this decision is in how we refrain from claiming that the literary style of each individual poet stays the same over their own works - instead we are looking at literary style changes across the genre. With the results from the Leave-One-Out Cross Validation we perform a Chi-squared test to see whether the association between true class and predicted class is statistically significant.

4.3 Results

The MDS plot is shown in Figure 4.1 with the texts of each of the Romantic poets. Following the MDS, we have also performed a hierarchical cluster analysis, based on the function word frequencies. The results of this form a dendrogram, shown in Figure 4.2.

Multi-dimensional scaling and Hierarchical Clustering

In order to illustrate the potential differences in literary style of different time periods and thus different genres, we have also included texts from the Modernist poets T.S Eliot, Ezra Pound and William Carlos Williams. The reason for choosing Modernist poets to compare with the Romantics was motivated by the idea that the Modernist literary movement existed at a distinctly separate time to Romanticism but also that Modernism was considered to be a reaction to Romanticism. This claim is verified to a certain extent in the dendrogram in Figure 4.2, where, on the left-hand side we do indeed see a clustering of 6 of the 9 Modernist texts that were used. It is then interesting to note that on

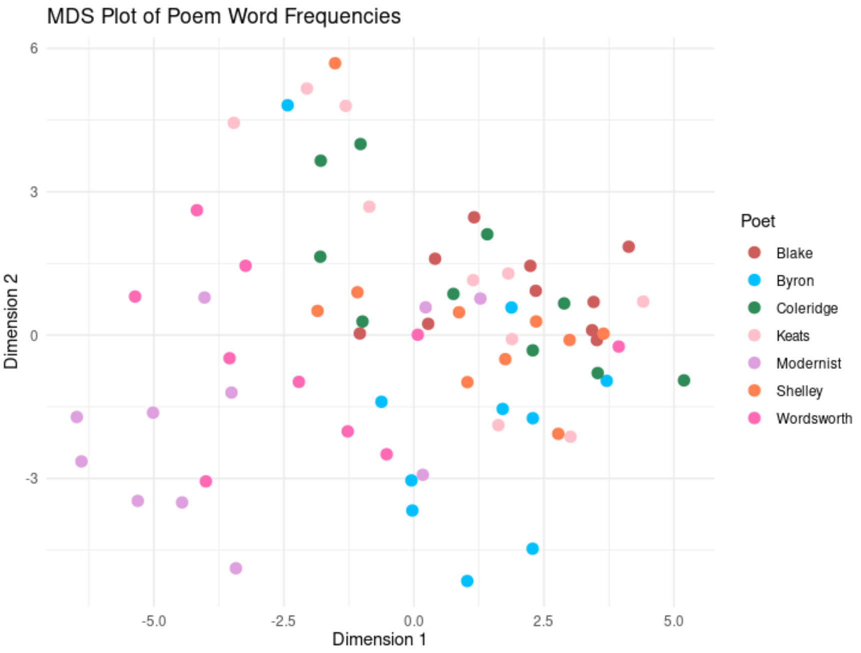

Figure 4.1: MDS plot to show groupings of Romantic poets.

the MDS plot, although the corpus of Modernist works tends toward the bottom left-hand side, some of the texts are also interspersed and close in distance to the Romantic texts. We would have to extend our study further to include the interactions between different literary periods but this initial lack of a particularly distinct cluster could perhaps align with the idea that the Romantic and Modernist movements were in conversation.

Looking closer at the purported groupings of the Early and Late Romantic poets we again fail to see clear clustering in both the MDS and the dendrogram. With the Early Romantics, we do indeed see the Wordsworth texts further from both Blake and Coleridge which lie in the cluster on the right of the MDS plot and the dendrogram highlights that the Blake texts (which form a relatively distinct cluster on one of the central branches) are rarely ever clustered with the Wordsworth texts. We could perhaps interpret this through the poets relative contextual influences. In a similar vein to stylometric author profiling (as mentioned in section 1), literary critics often consider the writer's upbringing and how this may influence in their writing style. Although the quantitative description of the texts, through function word frequencies, strips the text of context, the influences of a writer's upbringing may still be detected in their very writing style regardless of subject matter. Blake's works heavily concerned social justice which was seen to be influenced by his less affluent upbringing as opposed to that of Wordsworth. So, despite writing over the same time period Blake's emotions are projected outwards communicating political views and criticising institutions that impede on nature whereas Wordsworth is noted for his introspection communicated through nature and his surroundings. Take for example the depictions of London by each poet. Where Blake talks about the 'chartered streets' controlled by the state which in turn marks every face with 'weakness' and 'woe',

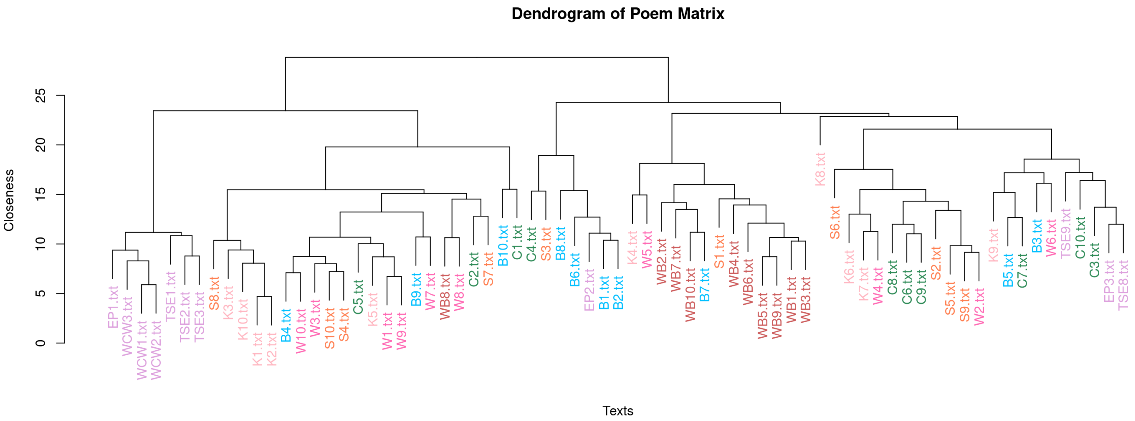

Figure 4.2: Dendrogram to show poem clustering.

Wordsworth optimistically speaks of the 'smokeless air' which highlights his hope beyond the industrialisation of the city and could indicate a certain disillusionment.

As one would imagine, the literary style of these poems and Wordsworth and Blake who have aesthetically and teleologically different approaches to the Romantic movement, would naturally fall in to different groups - which is reflected in our stylometric analyses. While our stylometric methods do not make use of close-word analysis as we have considered above, further research using word and phrase classes as features may identify and verify patterns in literary style.

When we consider the late Romantics we see that the works of Keats and Shelley seem closer together in the right cluster of the MDS plot. From a literary perspective some of the Keats and Shelley poems share similar themes; Keats' *La Belle Dame Sans Merci* expresses motifs of decay similar to Shelley's *Ozymandias*. We have also included Shelley's *Adonais* which was an elegy to Keats himself on his untimely death. Thus, the closeness of Shelley's texts and Keats' texts could give rise to a broader question on whether literary style depends on attitudes to a specific subject matter.

Once again it is not the case that all of the late Romantic poets works have clustered together. Byron's text tend to the bottom right of the MDS plot. In the dendrogram, we see that while some of the Byron poems are clustered together, they are rarely clustered with the other late Romantics, Shelley and Keats. As discussed, literary critics tend to temper their studies of texts by considering the influences of writers. We know that Blake, Wordsworth, Coleridge, Shelley, Keats were all notably influenced by John Milton - a Restoration Period poet - whereas Byron was chiefly influenced by Pope [Raupp, 2022]. While Pope respected the ideas and concepts within Milton's works - the ideas that go on to be integral to Romanticism - he certainly was not as attached to the concepts and the language. Thus for Byron, perhaps the most flamboyant of the Romantics, declaring his chief influence to be Pope is a way in which he distances himself from the other Romantic poets. This is particularly seen in his first major satirical works *En-*

glish Bards and Scotch Reviewers which was said to uphold Pope's tradition and certainly reflected the literary style of Pope through the use of heroic couplets. This fortifies the individualisation we deduce within the group of late Romantics.

Ultimately, where stylometric analysis would usually illustrate that there are different literary styles, from both the MDS plot and the dendrogram, we see that there are no distinct clustering of each poets works. We will discuss the implications of this further both in terms of Romanticism and poetry as a literary form in general in section 4.4 .

Cross Validation

To further explore the dynamics between time and literary style we performed a Leave-One-Out Cross Validation (LOOCV) which is explained in Chapter 2.

In the table below we can see the results from the LOOCV.

Poet	'True' Class	Classified as Early	Classified as Late
Blake	Early	8	2
Coleridge	Early	5	5
Wordsworth	Early	3	7
Byron	Late	2	8
Keats	Late	6	4
Shelley	Late	3	7

Table 4.2: Classification of Poems by Poet

The results here coincide with what we found from the MDS and cluster analysis in that there is no clear classification of early and late poems. For the early poets, 46.67% of the poems were misclassified as late. Whereas for the late poets, the LOOCV misclassifies 36.67% as early poems. Interestingly, looking further into the individual poets, we identify similar patterns from our earlier clustering analysis - where many of the Blake texts are classified as early, many of Wordsworth's are classified as late. The Leave-One-Out predictions do indeed place these two poets in different categories. Similarly, 8 of the 10 Byron works are classified as late, whereas the melancholic Keats sees more than half of his works classified as early. To further analyse whether these results are significant we performed a Chi-squared test with the null hypothesis that there is no association between the true class and the predicted class against the alternative hypothesis that there is an association between the true class and the predicted class. In this case, by 'true class' we mean the time-driven literary classification of early and late generations.

The Chi-squared test gave a p-value of 0.07594. Since this p-value is greater than 0.05, we fail to reject the null hypothesis, meaning that based on this test, there is no significant association between the true class and the predicted class. This suggests that the model is not reliably distinguishing between the two groups, Early vs Late Romantics.

4.4 Discussion

The Romantic poets regarded themselves and were ostensibly seen as non-conformists. To a certain extent, our stylometric analysis has verified this. We have seen through the various methods used that there is no distinct classification of the Big Six Romantic poets into Early and Late groups. Hence we cannot statistically conclude that literary style is influenced by the time at which one is writing. We also have not been able to see another such clustering occur. Had this been the case it would perhaps suggest that there may be a literary style, on a functional level at least, that categorises certain texts together. In fact, from the MDS and cluster analysis we see the Romantic works of the different generations interspersed within each other and the cross validation reported that many of the early poems would be predicted as late poems.

Through further study, given the small test we included with the Modernist poets, it would be interesting to explore this question to a higher order, between different genres. We may be able to see that there is mathematical harmony in the works of the Romantics that distinguishes them in terms of literary style, compared to poets of other literary eras.

On the surface, our results muddy the stylometric waters when it comes to poetry. As discussed in the Introduction, stylometry has high accuracy with authorship attribution and identifying patterns in literary style. With poetry, the writing style is arguably more intentional and orchestrated rather than inherent to a writer. To expand on this, the form of the poem is a key feature in itself. As evidenced by the Romantics, who quite consciously wrote in metre or in a sonnet form for example. Poetry, therefore is not like natural speech or prose and this could point to why even within individual poets works it was difficult to see clustering.

This suggests a limitation in applying stylometry to poetry. We must acknowledge that significant amounts of information is lost through the function word analysis; which is arguably to the dismay of literary critics who would place more importance on the intentional use of other features that illustrate poetic style. Gorman discusses how morphosyntactic annotation would allow us to include some of these lost features and create a clearer picture in literary stylometry [Gorman, 2024]. He argues that since the 'academic study of literary style has roots in the tradition of Poetics and Rhetoric' and these two approaches are in consensus on the most important part of the description of the style of a text being diction and word arrangement. Thus, the function word analysis strips away arguably the most important part of literary style. For a more heuristic analysis of a literary genre or works of an author, Gorman suggests that syntactic annotation encodes the choices of and relationships between words. It is perhaps a better compromise with the procedures of literary critics as it naturally expands on the traditional approach of looking at both Poetics and Rhetoric.

For now, however, it is not then that literary style transcends time but through this study it essentially resists such reduction and rationalisation. This would possibly come at the relief to both scholars in literature and the Romantic poets themselves.

Chapter 5

Behind the Scenes: Investigating Novel and Movie Adaptation Similarity

5.1 Background and Motivation

The adaptation of novels into movies has been a long-standing practice in the entertainment industry. While the market for film adaptations continues to thrive given audience interest in movies based on literary works, it also generates debates regarding the faithfulness of these adaptations. Many studies have been conducted to adapting literary works into movie, but were mostly done by using analyzing the plots of them. For example, Astiantih looked into the similarities and differences between the novel *Twelve Years as a Slave* and its adaption movies, and eventually find out that the movie is similar to its novel based on descriptive qualitative method [Astiantih et al., 2017]. We believe it would be more convincing if stylometry is applied. Some adaptations stay remarkably true to their source material, preserving narrative structure, themes, and even dialogue, while others take creative liberties to enhance visual storytelling or appeal to broader audiences. For example, the screenplay of the movie *Gone Girl* is also written by *Gillian Flynn*, the author of the novel. This unique scenario gives a high degree of dialogue similarity between the movie and the novel. In contrast, many adaptations are handled by different screenwriters and thus lead to significant linguistic difference. This chapter aims to explore how function word distributions can help us understand the linguistic similarities between novels and their movie adaptations - whether a film script stays similar to the dialogue of its source novel or diverges in it's screenplay adaptations. Function word usage is employed to quantitatively distinguish adapted novel-movie pairs from randomly paired novels and movies and assess whether function word distributions offer meaningful insights for identifying adapted works.

5.2 Data and Statistical Techniques

Data Collection and Function Word Selection

The data for this study contains 34 dialogues from movie scripts and their corresponding novels, which were obtained from publicly available online repositories and in digital formats. The list of movies can be found in Appendix 7.5. These movies are selected based on two factors: popularity and diversity. For instance, movies like *To Kill a Mockingbird* and *Pride and Prejudice* are chosen because of their popularity. Moreover, the topics of these movies vary from psychological thrillers (e.g. *Gone Girl* and *Fight Club*) to science fiction (e.g. *The Martian*, *Ender's Game*), fantasy stories (e.g. *Harry Potter* series) and historical dramas (e.g. *Litter Women*). These movie selections ensure our dataset spans a variety of genres. Given the focus on linguistic similarity, only the dialogues of both the novel and movie are used for analysis. To process the data we first extract dialogue from the 34 novels and their corresponding movie scripts into plain text files. Additionally, 34 novel-movie random pairs were generated for comparison. Each text is then represented as a 71-dimensional vector, corresponding to the standardised frequencies of function words.

The set of 70 function words that were used are the most common words in the English Language. The source and reasons for choosing these function words are described in section 2.1. Additionally, we introduce a 71st category, which accounts for all the words that are not function words. The distance between two pairs is measured by Euclidean distance as described in Equation (2.2).

Stylometric and statistical techniques

Through the use of Euclidean distance, we can measure how closely a movie script is to the novel that it is adapted from. Beyond measuring the similarity by simply looking at the distance, we applied Confidence Interval Analysis, t-test and Bootstrap Methodology to investigate whether adapted novel-movie pairs are statistically significantly different from random pairs in terms of dialogue similarity. Moreover, Principal Component Analysis (PCA) and K-Means Clustering are used to see if the adaptations follow specific linguistic patterns or if they vary.

5.3 Results

Most Similar Adapted Novel-Movie Pairs

To analyse the dialogue similarities between novels and their adapted movies, we computed the Euclidean distance, Equation (2.2), of the function words for each pair. Table 5.1 lists the top 3 adapted pairs with the smallest distances, indicating the pairs with the highest degree of similarity.

We found that *Gone Girl* has the smallest function word distance of 0.1493, meaning that this pair has the highest similarity among the 34 pairs in the dataset. This is an interesting result as Gillian Flynn, the author of the novel, also wrote the screenplay. Moreover, after careful inspection of the two scripts of dialogues,

Table 5.1: Top 3 Most Similar Adapted Novel-Movie Pairs

Rank	Novel-Movie Pair	Distance
1	Gone Girl	0.1493
2	To Kill a Mockingbird	0.1756
3	The Great Gatsby	0.1767

we found that this similarity is due to the similarity of function words, and therefore the writing style of the author of the script and novel, rather than the dialogue being copied and pasted from novel to script. This means that this direct adaptation preserved the author's unique linguistic style, making the screenplay highly similar to the novel it adapted from. The second most similar pair is *To Kill a Mockingbird* with a distance of 0.1756. This adaptation is a strong example of faithful adaption. The novel, written by Harper Lee, was adapted by Horton Foote, who is known for his careful and respectful approach to adaptation [Foote, nd]. The movie follows the novel's dialogue and themes closely, which could contribute to the high similarity. Lastly, the classic adaption of *The Great Gatsby* shows a function word distance of 0.1767. This adaptation is known to preserve the novel's original elegance - the dialogue in the movie is similar to the style of the author, F. Scott Fitzgerald, potentially leading to this low function word distance.

These results indicate that novel writers who contribute directly to their screenplay adaptations, or screenwriters who closely adhere to the novel's dialogue and style, produce movies that exhibit strong linguistic consistency with their source.

Same Author, Different Distance: The Case of Cormac McCarthy

An interesting observation in this study is the substantial function word distance (0.8580) between *No Country for Old Men* (2005), a novel written by Cormac McCarthy, and *The Counselor* (2013), a movie that is also written by McCarthy. Despite both works being written by the same writer, their linguistic patterns, as measured by function word usages, exhibit huge differences. This raises an interesting question: how does an author's writing style change when transitioning from novel writing to screenplay writing?

One of the reasons may be the structural differences between the two mediums. A novel normally allows for more narration and descriptive sentences, all of which influence function word usage. In contrast, a screenplay is primarily composed of dialogue and scene descriptions, naturally leading to a different distribution of function words. *The Counselor*, being an original screenplay, was written with a cinematic lens, whereas *No Country for Old Men* was written as a novel.

Another possible explanation for the high function word distance is the difference in the nature of these two works. Despite both work belonging to same genre (crime thriller), *No Country for Old Men* is written in a story-telling style, whereas *The Counselor* is delivered in dense monologues. Although written by the same author, their seems to be a different literary style in this case. As discussed with the poetic form in Chapter 4, this could open up the question of whether the literary style of a single writer varies with subject matter.

Confidence Interval Analysis

To compare the differences in the usage of function words between adapted and random novel-movie pairs, we computed the 95% confidence intervals for both groups. The plot is shown in the Figure 5.1.

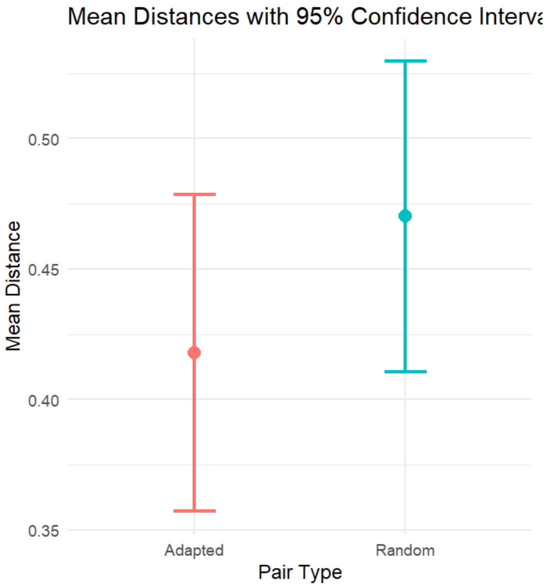

Figure 5.1: Mean function word distances with 95% confidence intervals for adapted and random pairs.

We can tell from the plot that the adapted pairs have a mean distance of 0.4178 with a confidence interval ranging from 0.3571 to 0.4785, while the random pairs have a slightly higher mean distance of 0.4701 with a confidence interval from 0.4105 to 0.5297. Despite the overlap between the confidence intervals, the random pairs tend to have a higher mean distance, suggesting that adapted pairs exhibit more similarity on average in function word usage than random pairs. This suggest that normally adaptions do follow a similar dialogue or linguistic pattern to the novels that they are adapted from. To prove this further, we use more statistical analysis to see if the distance is significantly different between the adapted pairs and random pairs.

Further Statistical Analysis

Figure 5.2 shows the bootstrap distribution of the mean differences between random and adapted pairs. The red vertical line represents the zero difference as a reference line. The majority of bootstrap samples exhibit a positive mean difference, indicating that random pairs generally have higher function word distances than adapted pairs. This result aligns with the previous Confidence Interval Analysis which shows adapted pairs have lower mean differences than that of random pairs. However, the 95% confidence interval (CI) is [-0.0004, 0.1103] which includes 0. Moreover, given that the red reference line overlaps with the distribution heavily, the evidence for a significant separation between the two groups are inconclusive.

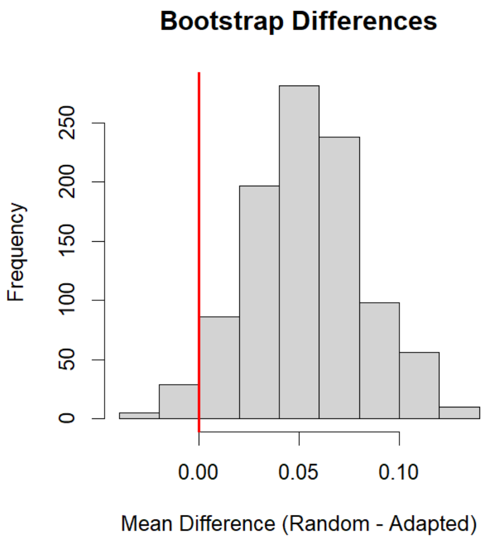

Figure 5.2: Bootstrap distribution of mean differences between random and adapted pairs. The red vertical line represents zero difference.

To further verify whether the mean function word distances differ significantly between adapted and random pairs, we performed a Welch's t-test. The results are summarized in the following table 5.3.

Statistic	**Value**
t-statistic	-1.2517
Degrees of Freedom (df)	65.978
p-value	0.2151
95% Confidence Interval	(-0.1358, 0.0311)
Mean of Adapted Pairs	0.4178
Mean of Random Pairs	0.4701

Table 5.2: Welch's t-test comparing the mean function word distances between adapted and random pairs.

The Welch's t-test results indicate that the mean function word distance between adapted and random novel-movie pairs is not significantly different,$p = 0.2151$. This implies that function word usage alone may not be sufficient to distinguish adapted pairs from random pairs based on their linguistic similarity.

K-Means Clustering Analysis

K-Means clustering was applied to investigate whether function word distributions can naturally separate adapted novel-movie pairs from random pairings.

The function word distributions were first standardized, and clustering was performed with $k = 2$, representing the two expected groups (adapted pairs and random pairs). Principal Component Analysis (PCA) was then used to reduce the 71-dimensional function word space into two principal components for visualization purposes:

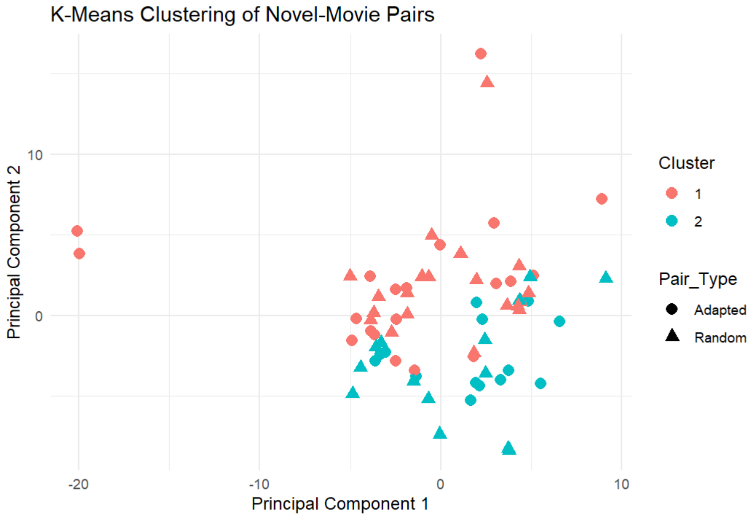

Figure 5.3: K-Means Clustering of Novel-Movie Pairs based on Function Word Usage

Figure 5.3 displays the results of K-Means clustering projected onto the first two principal components. Each point represents a novel-movie pair, with colours indicating cluster assignments and shapes distinguishing adapted pairs from random pairs.

The plot demonstrates that the clustering results do not align with the expected adapted and random pair categories. This suggests that function word distributions alone may not be enough to separate the two groups effectively. The overlapping nature of points indicates that the similarities in function word usage between adapted and random pairs might be more complex than initially hypothesized.

5.4 Discussion

In this study, we found that the similarities between novels and movies seem to vary across different pairs. The results shows that while some adaptations are similar to the novels that they are adapted from, other pairs differ greatly. Moreover, adaptations where the original author was involved in screenwriting, such as *Gone Girl (2014)*, exhibited the highest similarity, showing the idea that the style of author plays a key role in adaptation similarity. On the other hand, adaptations that changed narrative structures such as *The Counselor (2013)*, displayed a bigger difference in terms of Euclidean distance. However, the statistical tests including Welch's t-test and bootstrap analysis, shows that there is no statistically significant difference between adapted novel-movie pairs and random pairs in terms of mean difference. This suggests that the differences of usages of function words alone may not be enough to distinguish adapted pairs from random pairs.

These findings highlight the complexities of adaptation, where linguistic consistency is just one factor among many that contribute to the transformation from novel to movie. While this study provides certain insights into novel-movie similarity, we must also acknowledge some limitations. Despite function words being useful for stylometric analysis, they do not capture broader content such as structure patterns. In fact, it has been mentioned in some studies that blindly using function words sorely could be dangerous. Some function words, especially pronouns, do correlate with other factors such as author's gender and narrative perspective [Kestemont, 2014]. Future work could include additional stylistic features such as structure patterns and apply word adjacency networks (WANs) to account for the rational structure [Segarra et al., 2015]. Moreover, there is 71 function words being considered in this study, which may not be a proper number for the function words. More research should be done to determine the optimized number of function words in order to maximally capture the style of novels and movies [Segarra et al., 2015]. Furthermore, it has been suggested that the usage of function words is also depend on the topic of the text [Kestemont, 2014]. Considering the fact that the dataset in this study only consists of 34 pairs of movie and novels, expanding the dataset to include more movies covering diverse topics would also provide a more comprehensive view and convincing result of how linguistic patterns change from novels to movies.

END SCENE.

Chapter 6

Conclusion

The Plot Thickens...

In this paper we have statistically analysed literary style in forms that are relatively unexplored in stylometric literature. Where literary style can be analysed through different literary features, it is not typically analysed quantitatively. With the help of function words we have been able to mathematically and statistically represent the literary style of different writers across different forms. The beauty of literature is its subjectivity and to a certain extent our stylometric analysis has upheld this beauty. Our results varied through the different applications of stylometric methods - in French novels, stylometry worked as expected to verify authorship. The standard authorship attribution that was applied to the works of Alexander Dumas gave rise to the notion that Auguste Maquet had more involvement in these literary texts than he was given credit for. This supported literary critical views that contended the authorship of these works.

As our analysis moved towards the poetic form, the accuracy that stylometry has for novels and prose wavered such that the works of each Romantic poet did not fall into identifiable clusters. The question of literary style transcending time was left open to debate since there was no clear grouping of an early and late generation of Romantic poets. In fact there seemed to be no clustering altogether. True to the intangible nature of Romanticism, our analysis brought about more questions on whether inter-genre literary style (like that between Modernism and Romanticism for example) transcends time and whether literary style alters with different subject matters. Given how poetry is structurally different to prose, from our study, we were lead to believe that poetry resists standard stylometric analysis.

Finally, a similar result arose with our exploratory study into novel adaptations. Some screenplays written by the same author showed similarity in literary style whereas others did not. Again, this gave rise to the question of how subject matter may impact literary style unconsciously as well as the consideration of the original form the work was written in.

Ultimately, it is still quite remarkable how the smallest and most common words has allowed us to capture similarities and differences in literary style in such notable pieces of literature. Beyond the scope of this study, it would be interesting to further explore these questions in stylometry. It may be that we have to swap our function words for word classes or other literary features to be able to further clarify literary styles of a writer, genre or a specific literary form. One may even want to use stylometry to deduce which of the authors wrote each section of this very report.

Chapter 7

Appendix

7.1 French Historical Romance Corpus

Here we have listed the novels that form the corpus used in Chapter 3, detailing the original year of publication, as well as the edition used. Note that only the novel 'Le Comte de Lavernie', by Auguste Maquet, was not accessed via Project Gutenberg, instead using the Bibliothèque Nationale de France [Bibliothèque nationale de Franc

Novel (Year originally published)	**Author**	Edition Extracted From
Pascal Bruno (1837)	Alexandre Dumas	Pauline et Pascal Bruno, eBook 71510, 2023, original publication: Paris: Michel Lévy frères, 1848; Claudine Corbasson and the online Distributed Proofreaders Canada team at http://www.pgdpcanada.net (This file was produced from images generously made available by The Internet Archive/Canadian Libraries.)
Pauline (1838)	Alexandre Dumas	Pauline et Pascal Bruno, eBook 71510, 2023, original publication: Paris: Michel Lévy frères, 1848; Claudine Corbasson and the online Distributed Proofreaders Canada team at http://www.pgdpcanada.net (This file was produced from images generously made available by The Internet Archive/Canadian Libraries.)

Le Capitaine Paul (1838)	Alexandre Dumas	Le capitaine Paul, eBook 15574, 2020, Ebooks libres et gratuits at http://www.ebooksgratuits.com.
Le Capitaine Pamphile (1839)	Alexandre Dumas	Le capitaine pamphile, eBook 18697, 2006, Produced by Chuck Greif and www.ebooksgratuits.com
Acté (1839)	Alexandre Dumas	Acté, eBook 18321, 2006, Produced by Chuck Greif and www.ebooksgratuits.com
Le Chevalier d'Harmental (1842)	Alexandre Dumas, Auguste Maquet	Le chevalier d'Harmental, eBook 18028, 2006, Produced by Chuck Greif and www.ebooksgratuits.com
Les Trois Mousquetaires (1844)	Alexandre Dumas, Auguste Maquet	Les trois mousquetaires, eBook 13951, 2024
Le Comte de Monte Cristo (1845)	Alexandre Dumas, Auguste Maquet	Le comte de Monte-Cristo - Tome I, II, III, IV, eBook 17989-17992, 2024, Chuck Greif and www.ebooksgratuits.com
La Reine Margot (1845)	Alexandre Dumas, Auguste Maquet	La reine Margot - Tome I, II, eBook 13856-13857, 2020, Ebooks libres et gratuits
Le Vicomte de Bragelonne (1848)	Alexandre Dumas, Auguste Maquet	Le vicomte de Bragelonne - Tome I, II, III, IV, eBook 13947-13950, 2024, Ebooks libres et gratuits, Revised by Richard Tonsing
La Tulipe Noire (1850)	Alexandre Dumas, Auguste Maquet	La tulipe noire, eBook 26504, 2008, Produced by Chuck Greif
Le Comte de Lavernie (1852)	Auguste Maquet	Le Comte de Lavernie - Tome I, II, III, 2010, original publication: Paris: A. Bourdilliat, 1860, Bibliothèque nationale de France, département Littérature et art, Y2-50546
La Belle Gabrielle (1854)	Auguste Maquet	La belle Gabrielle - Tome I, II, III, eBook 15686, 2023, Produced by Distributed Proofreaders Europe, http://dp.rastko.net Project by Carlo Traverso and Josette Harmelin This file was produced from images generously made available by the Bibliothèque nationale de France (BnF/Gallica) at http://gallica.bnf.fr

Notre-Dame De Paris (1831)	Victor Hugo	Notre-Dame de Paris, eBook 19657, 2006, Produced by Chuck Greif and ebooksgratuits.com
Les Misérables (1862)	Victor Hugo	Les miserables - Tome I, II, III, IV, V, eBook 17489, 2006, Produced by www.ebooksgratuits.com and Chuck Greif
L'Homme Qui Rit (1869)	Victor Hugo	L'homme Qui Rit, eBook 5423, 2015, Produced by Carlo Traverso, Robert Rowe, Charles Franks and the Online Distributed Proofreading Team.
Le Rouge Et Le Noir (1830)	Stendhal	Le rouge et le noirL chronique du XIXe siècle, eBook 798, 2020, Produced by Tokuya Matsumoto HTML version produced by Chuck Greif
La Chartreuse De Parme (1839)	Stendhal	La Chartreuse De Parme, eBook 796, 2020, Produced by Tokuya Matsumoto, HTML formatting by Walter Debeuf, Project Gutenberg Volunteer.

Table 7.1: List of French Romantic Novels

7.2 List of French Function Words used in Chapter 3

de, être, et, à, avoir, ne, je, que, se, qui, ce, dans, en, du, au, pour, pas, vous, par, sur, faire, plus, dire, me, on, lui, nous, comme, mais, pouvoir, avec, tout, y, aller, voir, bien, où, sans, tu, ou, leur, si, moi, vouloir, te, venir, quand, celui, notre, devoir, l'a, prendre, même, votre, rien, encore, aussi, quelque, dont, trouver, donner, temps, ça, peu, falloir, sous, parler, alors, chose, mettre, savoir, passer, autre, après, regarder, toujours, puis, jamais, cela, aimer, non, croire, donc, fois, seul, entre, vers, chez, demander, jusque, très, moment, rester, répondre, premier, car, entendre, ni, ainsi, contre

Table 7.2: List of 100 Function Words Used in Chapter 3

7.3 The Romantic Poets Corpus

Here we have listed the poems that form the corpus for the Romantic Poetry analysis along with the year of their publication as well as the codes we have given to aid interpretation of the dendrogram in Figure 4.2. We have also included the poems that make up the Modernist poem set which we used as a test for robustness in our stylometric techniques.

Wordsworth Poems	**Text Code**	**Year**
Tintern Abbey	W1	1798
To my sister	W1	1798
Lines written in very early youth	W2	1798
Lines written in early spring	W2	1798
The Prelude Book I	W3	1798
The Prelude Book II	W4	1798
Hart-Leap Well	W5	1800
A Slumber did my Spirit Seal	W5	1800
To The Same Flower	W6	1802
It is a Beauteous Evening Calm and Free	W6	1802
I Wandered Lonely as a Cloud	W7	1807
My Heart Leaps Up	W7	1807
The World Is Too Much With Us	W8	1807
The Sun Has Long Been Set	W9	1807
She was a Phantom of Delight	W10	1807

Blake Poems	**Text Code**	**Year**
Songs of Innocence and Experience	WB1 - WB9	1794
Auguries of Innocence	WB10	1803

Coleridge Poems	**Text Code**	**Year**
Ode to the Departing Year	C1	1796
This Lime Tree Bower My Prison	C2	1797
The Rime of the Ancient Mariner	C3	1798
Fears in Solitude	C4	1798
The Nightingale	C5	1798
Ode to Tranquility	C6	1801
Dejection: An Ode	C7	1802
Hymn Before Sunrise	C8	1802
Kubla Khan	C9	1816
Christabel	C10	1816

Shelley Poems	**Text Code**	**Year**
On Death	S1	1816
A Summer Evening Churchyard	S1	1816
Mont Blanc	S2	1817

Ozymandias	S3	1818
Stanzas Written in Dejection; Near Naples	S4	1818
Ode to the West Wind	S5	1819
The Mask of Anarchy	S6	1819
A Winter's Day	S7	1819
To a Skylark	S8	1820
Adonais	S9	1821
The flower that smiles today	S10	1821
Music, when soft voices die	S10	1824
To the Moon	S10	1824

Byron Poems	**Text Code**	**Year**
On the death of a young lady - from *Fugitive pieces*	B1	1806
To D—— - from *Fugitive Pieces*	B1	1806
To Caroline- from *Fugitive pieces*	B1	1806
Childish Recollections	B2	1806
To Romance - from *Hours of Idleness*	B3	1807
I would I were a careless child - from *Hours of Idleness*	B3	1807
Soliloquy of a bard- from *Hours of Idleness*	B3	1807
Stanzas to a lady - from *Poems on Various Occasions*	B4	1807
To MSG - from *Poems on Various Occasions*	B4	1807
Love's Last Adieu	B5	1808
She Walks in Beauty	B6	1814
Stanzas for Music	B7	1815
Oh! Snatched Away in Beauty's Bloom	B7	1815
Fare Thee Well	B8	1816
Solitude	B8	1816
Darkness	B8	1816
Prometheus	B8	1816
When We Two Parted	B9	1817
So We'll Go No More a Roving	B9	1817
Apostrophe to the Ocean	B10	1818

Keats Poems	**Text Code**	**Year**
Hyperion Book I	K1	1818
Ode on a Grecian Urn	K1	1819
Lamia	K3	1820
The Eve of St Agnes	K4	1820
Isabella: The Pot of Basil	K5	1820
La Belle Dame sans Merci	K6	1820
Ode to a Nightingale	K7	1820
To Autumn	K8	1820

Lines on the Mermaid Tavern	K9	1820
Ode on Melancholy	K10	1820

Modernist texts	**Poet**	**Text Code**	**Year**
The Wasteland	T.S. Eliot	TSE1	1922
The Love Song of J Alfred Prufrock	T.S. Eliot	TSE2	1915
Portrait of a Lady	T.S. Eliot	TSE3	1920
Portrait	Ezra Pound	EP1	1918-21
Fourth Canto	Ezra Pound	EP2	1918-21
Fifth Canto	Ezra Pound	EP3	1918-21
Sour Grapes : a book of poems	William Carlos Williams	WCW 1-3	1921

7.4 List of function words used in Chapter 5

a	all	also	an	and
any	are	as	at	be
been	but	by	can	do
down	even	every	for	from
had	has	have	her	his
if	in	into	is	it
its	may	more	must	my
no	not	now	of	on
one	only	or	our	shall
should	so	some	such	than
that	the	their	then	there
things	this	to	up	upon
was	were	what	when	which
who	will	with	would	your

Table 7.10: List of Function Words Used in Chapter 5

7.5 List of the Movies & Novels in Chapter 5

The following table lists the movies and the corresponding novels used in this study, along with their publication years.

Movie Title (Year)	Novel Title (Year)
A Beautiful Mind (2001)	A Beautiful Mind (1998)
Big Fish (2003)	Big Fish: A Novel of Mythic Proportions (1998)
Brooklyn (2015)	Brooklyn (2009)
Call Me by Your Name (2017)	Call Me by Your Name (2007)
Charlie and the Chocolate Factory (2005)	Charlie and the Chocolate Factory (1964)
Doctor Zhivago (1965)	Doctor Zhivago (1957)
Ender's Game (2013)	Ender's Game (1985)
Fight Club (1999)	Fight Club (1996)
Forrest Gump (1994)	Forrest Gump (1986)
Gone Girl (2014)	Gone Girl (2012)
Harry Potter and the Sorcerer's Stone (2001)	Harry Potter and the Sorcerer's Stone (1997)
Harry Potter and the Chamber of Secrets (2002)	Harry Potter and the Chamber of Secrets (1998)
Harry Potter and the Prisoner of Azkaban (2004)	Harry Potter and the Prisoner of Azkaban (1999)
Harry Potter and the Goblet of Fire (2005)	Harry Potter and the Goblet of Fire (2000)
Jurassic Park (1993)	Jurassic Park (1990)
Life of Pi (2012)	Life of Pi (2001)
Little Women (2019)	Little Women (1868)
Love in the Time of Cholera (2007)	Love in the Time of Cholera (1985)
No Country for Old Men (2007)	No Country for Old Men (2005)
Pride and Prejudice (2005)	Pride and Prejudice (1813)
Ready Player One (2018)	Ready Player One (2011)
The Shawshank Redemption (1994)	Rita Hayworth and Shawshank Redemption (1982)
Stardust (2007)	Stardust (1999)
The Da Vinci Code (2006)	The Da Vinci Code (2003)
The Great Gatsby (2013)	The Great Gatsby (1925)
The Green Mile (1999)	The Green Mile (1996)
The Hunger Games (2012)	The Hunger Games (2008)
The Martian (2015)	The Martian (2011)
The Notebook (2004)	The Notebook (1996)
The Princess Bride (1987)	The Princess Bride (1973)
The Reader (2008)	The Reader (1995)
The Shining (1980)	The Shining (1977)
The Silence of the Lambs (1991)	The Silence of the Lambs (1988)
To Kill a Mockingbird (1962)	To Kill a Mockingbird (1960)

Table 7.11: List of Movies and Their Corresponding Novels Used in This Study

7.6 Code

For reproducibility, all R code used in our analysis is available at the following GitHub repository: `https://github.com/pepabirkett/Maths-Project-R-Code`. Cleaned raw text files are originally downloaded via Project Gutenberg [Project Gutenberg, 2025

Bibliography

[Abdi and Williams, 2010] Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

[Allogani et al., 2019] Allogani, M., Al-Jumeily, D., et al. (2019). *Supervised and Unsupervised Learning for Data Science*, chapter A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Springer.

[Astiantih et al., 2017] Astiantih, S., Rahman, F., and Makka, M. (2017). From narrative slave to movie: Adaptation theory. *Imperial Journal of Interdisciplinary Research (IJIR)*, 3(6):659–663.

[Bibliothèque nationale de France, 2025] Bibliothèque nationale de France (2025). Bibliothèque nationale de france (bnf). Accessed: 2025-02-21.

[Borg and Groenen, 2005] Borg, I. and Groenen, P. J. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer.

[Burrows, 2002] Burrows, J. (2002). 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

[Callet-Bianco, 2020] Callet-Bianco, A.-M. (2020). Dumas et alii. l'écriture en collaboration. *Cahiers de L'Herne.*

[Chong et al., 2021] Chong, B. et al. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, 4(5):37–40.

[Dauber et al., 2017] Dauber, E., Overdorf, R., and Greenstadt, R. (2017). Stylometric authorship attribution of collaborative documents. In *Cyber Security Cryptography and Machine Learning: First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings 1*, pages 115–135. Springer.

[Davies, 2010] Davies, L. (2010). Film reignites literary debate over alexandre dumas's ghostwriter. *The Guardian.*

[Eduscol, nd] Eduscol (n.d.). Mots les plus fréquents de la langue écrite française (xixe et xxe siècles). Accessed: 2025-01-21.

[Foote, nd] Foote, H. (n.d.). *Trip to Bountiful, Tender Mercies, To Kill a Mockingbird.* Grove Atlantic.

[Gorman, 2024] Gorman, R. (2024). Morphosyntactic annotation in literary stylometry. *Information.*

[Kestemont, 2014] Kestemont, M. (2014). Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press.

[Maquet, 2004] Maquet, A. (2004). *Le Vicomte de Bragelonne.* Ebooks libres et gratuits. Revised by Richard Tonsing.

[Martone, 2020] Martone, E. (2020). *Alexandre Dumas as a French Symbol since 1870: All for One and One for All in a Global France.* Cambridge Scholars Publishing.

[McIntyre and Walker, 2022] McIntyre, D. and Walker, D. (2022). Using corpus linguistics to explore the language of poetry: a stylometric approach to yeats' poems. *Routledge Handbook of Corpus Linguistics*, pages 499 –516.

[Merriam and Matthews, 1994] Merriam, T. and Matthews, R. (1994). Neural computation in stylometry ii: An application to the works of shakespeare and marlowe. *Literary and Linguistic Computing.*

[Mombert, 2022] Mombert, S. (2022). Dans les archives d'un atelier littéraire. exploration génétique de la collaboration entre alexandre dumas et auguste maquet. *Genesis. Manuscrits–Recherche–Invention*, (54):31–42.

[Monsteller and Wallace, 1963] Monsteller, F. and Wallace, D. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association 58.302.*

[Mooney, 1996] Mooney, C. Z. (1996). Bootstrap statistical inference: Examples and evaluations for political science. *American Journal of Political Science*, pages 570–602.

[Mosteller and Wallace, 1963] Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

[Nebbou, 2010] Nebbou, S. (2010). L'autre dumas.

[Nielsen, 2016] Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science*, chapter Hierarchical Clustering. Springer.

[Paraschas, 2018] Paraschas, S. (2018). *'Tous pour un, un pour tous': Alexandre Dumas, Auguste Maquet, and the Musketeers Trilogy*, pages 163–197. Springer International Publishing, Cham.

[Project Gutenberg, 2025] Project Gutenberg (2025). Project gutenberg: Free ebooks. Accessed: 2025-02-21.

[Raupp, 2022] Raupp, E. R. (2022). Caucasus journal of milton studies.

[Ross, 2020] Ross, G. (2020). Tracking the evolution of literary style via dirichlet–multinomial change point regression. *Journal of the Royal Statistical Society Series A: Statistics in Society.*

[Savoy, 2020] Savoy, J. (2020). *Machine learning methods for stylometry.* Springer.

[Schopp, 1991] Schopp, C., editor (1991). *Les Grands romans d'Alexandre Dumas: Les Trois Mousquetaires, Vingt ans après.* Laffont, Paris.

[Segarra et al., 2015] Segarra, S., Eisen, M., and Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, 63(20):5464–5478.

[Sinaga and Yang, 2020] Sinaga, K. P. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

[Wellek, 1949] Wellek, R. (1949). *Comparative Literature*, chapter The Concept of 'Romanticism' in Literary History. I. The Term 'Romantic' and Its Derivatives.

[Wong, 2015] Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48.