Authorship Attribution for Swedish Novels

Martina Nyberg
Uppsala University
martina.nyberg.7261@student.uu.se

Abstract

This paper examines the writing style of Swedish authors from the perspective of authorship attribution. Using data consisting of Swedish novels written by 72 authors, several variants of feature sets are employed in an SVM and a Naive Bayes classifier for the task of identifying authors based on stylistic properties. The results show that features based on token frequencies and character n-grams both yield high performance. However, an examination of the most informative features confirm the importance of choosing features that are unrelated to the topic of the text, while also exposing differences between the classifiers regarding which features are the most discriminative. Building on the results of the authorship attribution task, a style representation consisting of the 120 most frequent words across the corpus is used to compute similarity between authors/novels based on the Jensen-Shannon divergence metric.

1 Introduction

The notion of writing style plays an essential role in characterizing the qualities associated with a particular text and its author. Apart from topic and genre, readers of novels may consider the style to be an important aspect of what makes a book appealing. Naturally, the type of text partly determines the style of the writing, but also sociological and psychological factors related e.g. to the author's personality, gender, age, level of education, and native language (Daelemans, 2013).

The research area of stylometry studies literary style through statistical analysis (Zheng et al., 2006). The basis for this approach relies on the assumption that an author's unique writing style is measurable in a quantifiable manner and can be defined by characteristics of the text (Neal et al., 2017). Such characteristics may include e.g. word use and syntactic constructions (Karlgren, 2004). According to Daelemans (2013), style is unique to each individual author and can be viewed as their fingerprint.

The subject of style from a computational perspective is of particular relevance in the field of literary analysis, where it can assist the process of analysis and interpretation by the use of quantitative data, which adds a level of objectivity and systematicity that cannot be acquired solely by close reading (Boukhaled, 2016).

Stylometry is closely related to the task of authorship attribution. This process aims to identify the author of a text; the text whose author is unknown is assigned to a most likely author based on previously seen training examples of texts with known authors (Jockers and Witten, 2010). Further, the classification is based on the style rather than the content of the text. In addition to the field of literary analysis, authorship attribution can be applied in areas such as sociolinguistic research, forensics, or medical diagnosis (Daelemans, 2013).

There are numerous works drawing on the approaches of stylometry and authorship attribution. The aim of these studies may in some cases be a deeper insight into, or confirmation of, some known phenomenon within the literary subject of study (Storey and Mimno, 2020; Jautze et al., 2013), while others seek to expand the methods for authorship analysis (Hay et al., 2020; Boukhaled, 2016), or apply them to certain domains such as scientific articles (Bergsma et al., 2012), or social media (Jasper et al., 2018). In addition, some works have focused on applying the approaches to more specific purposes and external tasks, such as predicting the literariness of novels (van Cranenburgh et al., 2019) or the degree of success that a book is likely to achieve, (Khalifa and Islam, 2020; Ashok et al., 2013) or the development of a book recommendation system based on writing

style (Alharthi et al., 2018).

Although there are works that consider computational methods for literary analysis of Swedish texts, few studies have addressed Swedish literature with regards to writing style on a large-scale, the subject to which this work aims to make a contribution. The papers by Zechner (2021) and Berglund and Dahllöf (2021) are the only works on writing style for Swedish literature known to the author, both with a somewhat different aim and approach compared to the present study.

A typical part of a stylometry task is the selection and extraction of features for text representation (Neal et al., 2017). The choice of the most suitable features is a prevalent issue, giving rise to many variations. According to Brocardo et al. (2013), more than a thousand different stylistic features and numerous analysis methods have been employed in the literature of authorship analysis. The types of features include token or character n-grams, punctuation, complexity and vocabulary richness measures, and syntactic, semantic or pragmatic features (Daelemans, 2013). There is no agreement on an optimal set of features that performs well across datasets and domains (Iqbal et al., 2010). However, Neal et al. (2017) argue that function words and n-grams have shown to be the most efficient types of features for authorship attribution, despite their simplicity.

In the light of this, this study aims to explore different token and n-gram feature sets and two classifiers, and assess their suitability and performance in authorship attribution on Swedish novels. More specifically, the study aims to answer the following questions:

- How well can stylistic differences between authors be measured automatically?
- What method in terms of the choice of features and classifier performs best at this task?

Additionally, based on the outcome of the authorship attribution task, a selected style representation will be used for measuring author/novel similarity, in order to analyze and discuss some noticeable phenomena regarding the authors' styles. The results provide insights about the degree of stylistic similarity of Swedish authors from a computational, more large-scale perspective compared to knowledge acquired through close reading. One potential practical use is a book recommendation system based on writing style.

The rest of this paper is organized as follows. Section 2 describes previous works related to stylometric analysis and feature selection. Section 3 presents a theoretical background on the representation of style, which is the basis for feature selection in the present study. Section 4 presents the preparation of the data. Section 5 describes the experiments, the results of which are presented in section 6. In section 7, the results are discussed and further applied to analyze stylistic similarity more closely. A conclusion is presented in section 8.

2 Related Work

The use of stylometry for literary analysis can be considered a form of *distant reading*, a term proposed by Moretti (2000) which involves the notion of focusing on parts that are smaller or larger than the text. Consequently, the study of some features of texts in a large collection can provide insights about the collection as a whole in a way that would be difficult to achieve through close reading.

For example, Storey and Mimno (2020) study the stylistic similarity across a large number of Ancient Greek authors, demonstrating that their writing style is very similar even centuries apart. To represent writing style, they employ features consisting of the 250 most frequent words used by those authors. The features are used in classification tasks to predict categories such as authorship, genre, dialect and time period.

Hay et al. (2020) propose an approach where documents are embedded in a stylometric space. Documents from the same author have similar representations in the space, thereby reflecting a distinctive style of that author. Using data from English news and blog articles, they employ deep neural network models to perform authorship clustering and attribution tasks.

Jautze et al. (2013) analyze the sentence types, and morphological and syntactic features in chic lit compared to high literature. Their results show that certain syntactic features differentiate the two genres. Principally, high literature contains more sentences with a higher syntactic complexity than chic lit.

Berglund and Dahllöf (2021) compare the stylistics of printed Swedish bestseller novels and the most popular audiobooks during the period 2015-2019. Their results show that there are clear differences between the formats across sev-

eral stylistic properties such as document length, word use and syntactic complexity.

3 The Representation of Style

A basis for many stylometric studies is the commonly made distinction between the topic and the style of a text. Topic concerns the actual content of the text, referring to things such as objects and events, while the style denotes how the content is expressed (Argamon et al., 2005). As there are many possible ways of expression, multiple texts can refer to the same topic but still display differences in terms of style. Under the assumption that style and topic are independent of each other, there is an importance in choosing features that exclusively represent stylistic properties.

The use of a set of the most common words to distinguish writing style has been shown successful by Burrows (2002). The high-frequency features of a language is often function words, and these have been widely used with success in stylometry and authorship attribution (Kestemont, 2014). Such words typically consist of articles, pronouns, adpositions, conjunctions, etc. Kestemont (2014) describes the notion of these words as units that do not carry much meaning by themselves, and they are used by all authors regardless of topic. Differences in use of function words between texts of different authors can thus be regarded as independent of topic and a consequence of different writing styles. Even so, Sundararajan and Woodard (2018) argue that not all non-function words are content-specific, and some of the most common types of these words may also be suitable in style representations. Additionally, Kestemont (2014) also discusses the high performance of character n-grams in capturing author style, arguing that they maintain the benefits of function words but also encompass morphological structures and whitespace, which makes them superior in capturing functional information.

Nevertheless, it may be difficult to completely separate style from topic, as some topics tend to require a certain style, e.g. legal matters (Karlgren, 2004). In fact, using a dataset of Swedish novels, Zechner (2021) showed that the results of authorship attribution by means of a small feature set containing the frequencies of just the 10 most common words were highly sensitive to differences in the topic of the texts. More specifically, the performance decreased when removing the influence of the topic by using reference and target samples from different books by the same author. Here each book was taken to signify a separate topic.

Contrasting results are seen in Menon and Choi (2011), where stop words were shown to be highly robust to changes in topic when employed in authorship attribution. However, one potentially significant difference to the study by Zechner (2021) is the use of 659 stopwords as compared to ten.

4 Data and Preprocessing

In this study, the data consists of Swedish novels written during the 19th and 20th centuries. The novels were collected from the Swedish Literature Bank,[1] and filtered to only include proofread novels written in Swedish. Duplicates in the form of texts in alternative editions were excluded. This resulted in a total of 249 remaining texts written by 78 different authors. 148 texts were written by a male author and 101 by a female author.

The actual novel texts were then extracted from their xml files using the Beautiful Soup python library, and further cleaned from remaining xml tags and special characters with regular expressions. Three of the texts were translations from French, out of which two contained the original French version as well as the Swedish translation. The French versions were removed from these texts. In addition, all files were manually inspected for prefaces, comments or other content written by someone other than the author. This type of information was found in 63 of the 249 texts, and since removed.

5 Experimental Setup

5.1 Feature Extraction and Encoding

Two different forms of features were used in this study: the relative frequency counts of the k most common tokens, and the k most common character n-grams with size ranging from 4 to 7. N-grams within this range yielded the most successful results for authorship attribution in a study by Kešelj et al. (2003) where ranges from 1 to 10 were tested.

The features were extracted by computing the token/n-gram counts of all tokens/n-grams across all texts. The k most frequent tokens/n-grams were then chosen as features. Following the procedure

[1] litteraturbanken.se

used in Storey and Mimno (2020), a condition was set to only include tokens/n-grams that appeared in at least 50% of the texts. This was done to avoid bias from long texts with high frequencies of certain words, such as proper nouns, that were not as common over the entire corpus. Additionally, this removed the influence of some of the spelling variations seen in the texts. One example includes the words *över* and *öfver* (over), which have the same meaning, but only one of the variations occurs for a large amount of the texts.

In the following experiments, features belonging to the category of token frequencies come in three variants with regards to the number of features used: 22, 120, and 250. The use of 250 most frequent tokens to represent style was adapted from Storey and Mimno (2020). However, upon inspecting the extracted tokens, they include several words that cannot be considered unrelated to topic, such as *ögon* (eyes), *barn* (child), and *gud* (god). These tokens occur relatively far down the list, but it is unknown how significant they may be in identifying the authors, hence this feature set is included nonetheless. The shorter feature sets are employed since they do not include such distinctly topic-related tokens. Thus, they may be considered superior representations of style, although with the drawback of encoding less information.

The character n-grams consist of n-grams with size in the set of {4, 5, 6, 7} and number of features in the set of {22, 120, 250, 500, 1000, 1500, 2000}. The higher numbers of features are employed due to character n-grams requiring more instances compared to tokens in order to encompass the same text amount. For comparison, the same sizes as used for the token features are included as well.

The collected texts of each author were split into nine segments of equal size, removing any trailing tokens. This resulted in 702 segments. Each segment was then encoded as a vector containing the token/n-gram frequency in the segment for each token/n-gram in the predefined feature set, normalized by the total number of tokens/n-grams for that segment. It is thus possible to compare segments of different lengths.

5.2 Classification Models

The different variants of encoded segments were used as input in classification tasks with a Support Vector Machine classifier (SVM) and a Multinomial Naive Bayes classifier, implemented through the scikit-klearn library (Pedregosa et al., 2011). The Naive Bayes classifier was the highest performing classifier in the study by Storey and Mimno (2020) on Ancient Greek authors. SVM is arguably the most popular model for text classification (van Cranenburgh et al., 2019), and has also been shown to outperform Naive Bayes when n-grams are used as features (Sapkota et al., 2015). In the SVM hyperparameters, kernel was set to linear, and C to 0.01. For the Multinomial Naive Bayes, alpha was set to 0.0001. All classifications were performed and evaluated with 9-fold cross validation. Hence, each model was trained on 8 folds and evaluated on the 9th, which was then repeated with each of the folds as the test fold. The average of the 9 scores were then taken as the performance metric.

6 Results

6.1 Classifier Performance

The results of classifying each text segment to one of the 172 authors with an SVM and a Naive Bayes classifier are presented in table 1 and 2 respectively. The reported scores are the mean accuracy of author prediction over the nine folds for each classification setup. Specifically, the first four rows denote the accuracy of classifications using each of the four n-gram sizes paired with each of the values for the number of features used. The last row denotes the accuracy acquired from complete tokens, using either 22, 120 or 250 features. Overall, the performances of the different feature sets are relatively high, with the exception of the smallest set of 22 features. This was not unexpected, since the encoded information is more scarce than for the other sizes.

It can be noted that SVM performs better than the Naive Bayes classifier, with the highest scores seen for 4-grams with 1000-2000 features. Generally, more features lead to better performance, which is true for all the n-grams. The highest results are closely followed by the 120 and 250 token features, which also perform well.

Considering the gender of the authors, the SVM classifier performed fairly well at predicting this variable, with accuracy scores ranging from 74% to 84% for the different feature configurations. For the Naive Bayes classifier, the acquired score is 63%, regardless of feature set. For comparison,

		Number of features						
		22	**120**	**250**	**500**	**1000**	**1500**	**2000**
n-gram size	**4**	0.69	0.95	0.98	0.98	0.99	0.99	0.99
	5	0.74	0.94	0.97	0.97	0.98	0.98	0.98
	6	0.67	0.94	0.97	0.97	0.97	0.98	0.98
	7	0.49	0.91	0.95	0.96	0.97	0.97	0.97
	Tokens	0.81	0.96	0.98	-	-	-	-

Table 1: SVM classification results. The scores indicate the accuracy of predicting the author for each text segment.

		Number of features						
		22	**120**	**250**	**500**	**1000**	**1500**	**2000**
n-gram size	**4**	0.63	0.87	0.90	0.93	0.95	0.96	0.97
	5	0.74	0.89	0.91	0.93	0.94	0.95	0.96
	6	0.77	0.91	0.92	0.94	0.95	0.95	0.96
	7	0.51	0.87	0.93	0.95	0.95	0.96	0.96
	Tokens	0.84	0.92	0.95	-	-	-	-

Table 2: Naive Bayes classification results. The scores indicate the accuracy of predicting the author for each text segment.

the chance of a correct classification by random guessing is 52%.

As noted by Neal et al. (2017), comparing the performance across different stylometric studies is difficult due to the lack of a gold standard benchmark dataset. The numerous ways to utilize features, and differences in data also complicate comparisons. The study on Ancient Greek Authors by Storey and Mimno (2020) is similar to the present study in that they also employed the 250 most common tokens in a Naive Bayes classifier. In their authorship attribution task, they acquired an accuracy of 88%, while the present study with the same classifier and type of feature yielded a 95% accuracy. These results are, of course, not directly comparable due to the different datasets. Moreover, their findings indicated that the authors in their data write in a style that is more similar to each other compared to the similarity seen in English or Icelandic authors. Thus, they were likely more difficult to distinguish for a classifier than was the case for the Swedish data in this study, which could further explain the difference in performance.

6.2 Feature Importances

As noted in section 5, the feature sets of 250 features include some topic-related tokens. Such words are identifiable among the n-grams as well, and become more common the more features are included. By examining the most contributing features per author, it is found that for the SVM classifier, the topic-related words were in fact noticeably discriminating for the task of author identification. As an example, the most important features in identifying the author Selma Lagerlöf through some of the feature sets are displayed in Table 3.

Most of these topic words are nouns, e.g. *ögonen* (the eyes), *leende* (smile), and *ansikte* (face). Several features are adverbs, which are not considered among the most typical part-of-speech categories for function words. However, they are not notably topic-related in the examples. Similar patterns are seen for all authors, indicating that when topic information is available, it is highly useful for identifying the author of the text through the SVM classifier. The reliance on topic words are not seen for feature lengths smaller than 250, in which topic words are absent, nor for any of the Naive Bayes classifications. Generally, this classifier relies more on the most common words, hence it is arguably making use of information that is more relevant to the style of the authors.

In summary of the results, a large part of the different feature sets were highly successful in authorship attribution. However, the larger feature sets which contain instances of topic-related words may be considered less suitable for representing style, under the assumption that style is separate from topic. This appears to be especially relevant when the style representation is applied in

Feature set			
250 tokens		7-gram, size 2000	
par	pair	’som om ’	as if
vad	what	’varandr’	each other
tillbaka	back	’ händer’	happens
igen	again	’ leende’	smile
mig	me	’händer ’	happens
kanske	maybe	’den där’	that
framför	in front of	’ den dä’	that
hennes	her	’nsikte ’	face
.	.	’en där ’	that
ett	a	’lötslig’	suddenly
mot	against	’ plötsl’	suddenly
upp	up	’plötsli’	suddenly
såg	saw	’ ansikt’	face
litet	small	’. henne’	her
ögonen	the eyes	’ansikte’	face
ned	down	’tsligt ’	suddenly
omkring	around	’ötsligt’	suddenly
liten	small	’ibland ’	sometimes
en	a	’ ibland’	sometimes
ögon	eyes	’ varand’	eachother

Table 3: The 20 most important featues for author Selma Lagerlöf when classified with SVM, using either 250 tokens or 2000 7-grams as features. The 7-grams which do not consist of complete words has been translated into the corresponding complete English word for clarity.

an SVM classifier, where the less common, topic-related words were exploited to a noticeable degree.

7 Analysis

On the basis of the results in the authorship attribution task, one feature set from the preceding experiments is selected to represent the writing style in pairwise comparisons of stylistic similarity, presented in this section.

The Jensen-Shannon divergence (JSD) (Lin, 1991) is used to measure similarity, which quantifies the difference between two probability distributions. The calculation is done through the Scipy library (Virtanen et al., 2020), which provides a distance metric in the form of the square root of the Jensen-Shannon divergence. To obtain the divergence, this metric is simply squared. Because a similarity measure rather than a divergence is desired, the similarity is calculated by subtracting the divergence from 1, following the procedure in Storey and Mimno (2020). Jensen-Shannon Divergence is not dominated by the most common words and gives equal importance to all words (Gerlach and Font-Clos, 2020). Therefore, a style representation that is narrow enough to exclude any topic-related words is favourable in order to avoid influence from such words. Hence, the feature set consisting of 120 tokens is chosen, which performed relatively well for both classifiers. The relative frequency vectors of each novel/author are converted into probability distributions and used as input to the similarity measure.

7.1 Writing Style in Translated Texts

One topic of interest is the three texts among the novels that were translated from French. One hypothesis is that the style of translations differs from the style in original works by the same author. These texts (*The Defence of a Fool*, *Legends*, *Inferno*), originally written by the author August Strindberg, are compared to each of the other novels in the corpus through the Jensen-Shannon similarity measure. The result is a list of works in descending similarity in relation to the chosen reference novel. These lists show that the three novels are distinctly close in similarity. Inferno and Legends both have each other as the most similar novel, while The Defence of a Fool diverges slightly more from the other two. The similarity of Inferno and Legends becomes particularly plausible given that they were both translated by the same person, while The Defence of a Fool had another translator. Irrespective of which of the three novels is used as the reference, the distance to Strindberg’s other novels which were originally written in Swedish is prominent. One thing to note is that these other novels do not always occur together in the lists, instead his works appear to be quite diverse in terms of style. This confirms findings by Zechner (2021), where the distance in similarity was shown to be higher for Strindberg’s novels when compared to other authors.

7.2 Consistency of Author Style

In spite of this, it can be hypothesised that books by the same author are generally close in style. By examining works by several authors in the data through the similarity measure, this appears to be true in many cases. As an example, the 10 novels most similar to *Den siste athenaren* (The Last Athenian) by Viktor Rydberg are listed in table 4, where the reference novel is followed by all the other novels by the same author included in the

Viktor Rydberg	Den siste athenaren
Viktor Rydberg	Fribytaren på Östersjön
Viktor Rydberg	Singoalla
Magnus Jacob Crusenstolpe	Morianen, eller Holstein-Gottorpiska huset i Sverige, part 1
Magnus Jacob Crusenstolpe	Morianen, eller Holstein-Gottorpiska huset i Sverige, part 2
Johan Vilhelm Snellman	Fyra giftemål
Emilie Flygare-Carlén	Skuggspel, part 1
Emilie Flygare-Carlén	Skuggspel, part 2
Axel Lundegård	Den stora dagen
Carl Jonas Love Almqvist	Smaragdbruden: följderna av ett rikt nordiskt arv
Otto Witt	Skapelsen

Table 4: The 10 novels most similar to Viktor Rydberg's *Den siste athenaren*.

Viktor Rydberg	1.0
Magnus Jacob Crusenstolpe	0.89
Emilie Flygare-Carlén	0.87
Johan Vilhelm Snellman	0.85
Aurora Ljungstedt	0.84
Wendela Hebbe	0.84
Fredrika Bremer	0.84
Axel Lundegård	0.84
Otto Witt	0.82
Carl Jonas Love Almqvist	0.82
Axel Gabriel Ingelius	0.81

Table 5: The 10 authors most similar to Viktor Rydberg. The right column displays similarity. This value has been scaled between 0 and 1, and calculated for all 172 authors.

corpus. Further, this list indicates that some of the works by the author Magnus Jacob Crusenstolpe are written in a style similar to the reference novel. By encoding the complete works of each author in separate vector representations, the similarity measure can be used to measure similarity between the authors. In agreement with the results on novel similarities, table 5 indicates that the author most similar to Viktor Rydberg is Magnus Jacob Crusenstolpe.

Nevertheless, some of the examined novels were not closely related to other novels by the same author. As argued by Daelemans (2013), individual style may change over time. This could possibly explain the distance in similarity for some of the novels belonging to the same author. Another possible explanation is related to the findings of Zechner (2021), who noted that the topic of the book influenced the successfulness of identifying its author, even though a very small number of function words were used as features. Thus, even when topic words were not included as features, the distribution of function words may depend on the topic words that they interact with in the full text. Consequently, books that cover highly different topics may require different forms of expression. In the present study, this notion was not thoroughly controlled for in the classification tasks, which were based on segments rather than complete books. Hence, in many cases, the training samples could be either from the same book as the test sample, or from another book by the same author. A closer look at the reasons behind stylistic dissimilarity within authors is a question for future research.

8 Conclusion

In this paper, two different classifiers and numerous feature sets were employed for the task of authorship attribution for Swedish novels. The results show that high accuracy was achievable for many of the feature sets, regardless of classifier. However, the Naive Bayes classifier may be considered more reliable in this context, since it gave more significance to features that are more typically associated with the notion of style. Choosing a set of features that model solely stylistic properties seems important to make sure that the classification is based on what is considered style and not topic.

A style representation that performed well in both classifiers and which does not include topic-related words was utilized in a closer analysis on the similarity of novels. This confirmed the ideas that translated texts differ stylistically from other non-translated novels by the same author, and that novels by the same author are generally close in style, though exceptions exist. These are just some

of the subjects that could be examined with this method. Hence, the selected examples mainly show that the style representation selected through the classification tasks could be successfully used in a distant reading related to some specific phenomena.

Given that numerous classifications were carried out in this study, it is difficult to provide an informative error analysis. The incorrectly predicted authors varied for different classifiers and feature sets. An extension of this study could be to analyze the reason behind misclassifications. The classifications are based on equally sized segments of the combined books by each author, resulting in that some segments may consist of only parts of a book in some cases, and contain several books in other cases. It is thus not trivial to investigate which particular works that tended to be misclassified. Thus, one variation of the method could be to build the data samples from separate works. Since the similarity measure has indicated that Strindberg's translated works differ from his other works originally written in Swedish, one hypothesis that could be tested under the approach of single work samples is whether the translated works also tend to be incorrectly classified by the classifiers.

It could also be interesting to consider in what way the stylistic dissimilarities between authors present themselves. This study has shown that there are discernible differences regarding their use of common words, but with 72 authors it was outside the scope to further investigate how this was expressed. For example, some authors may use pronouns or commas more than others, etc. The method of distant reading does not imply that close reading and qualitative analysis is obsolete, and more thorough explanations only possible with domain knowledge within literary analysis could further extend the results. For instance, by being familiar with the styles of these books through reading, the quality of the similarity measure could be further assessed.

References

Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. 2018. Authorship identification for literary book recommendations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 390–400.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pages 1–16.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764.

Karl Berglund and Mats Dahllöf. 2021. Audiobook stylistics: Comparing print and audio in the bestselling segment. *Journal of Cultural Analytics*, page 29802.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337.

Mohamed Amine Boukhaled. 2016. *On computational stylistics: mining literary texts for the extraction of characterizing stylistic patterns*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6.

John Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17:267–287.

Andreas van Cranenburgh, Karina van Dalen-Oskam, and Joris van Zundert. 2019. Vector space explorations of literary language. *Language Resources and Evaluation*, 53:625–650.

Walter Daelemans. 2013. Explanation in computational stylometry. In *International conference on intelligent text processing and computational linguistics*, pages 451–462.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22:126.

Julien Hay, Bich-Liên Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. 2010. Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7:56–64.

Johannes Jasper, Philipp Berger, Patrick Hennig, and Christoph Meinel. 2018. Authorship verification on short text samples using stylometric embeddings. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 64–75.

Kim Jautze, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong. 2013. From high heels to weed attics: a syntactic investigation of chick lit and literature. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 72–81, Atlanta, Georgia.

Matthew L Jockers and Daniela M Witten. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25:215–223.

Jussi Karlgren. 2004. The wheres and whyfores for studying text genre computationally. In *Workshop on Style and Meaning in Language, Art, Music, and Design*.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264.

Mike Kestemont. 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.

Muhammad Khalifa and Aminul Islam. 2020. Will your forthcoming book be successful? predicting book success with cnn and readability scores. *arXiv preprint arXiv:2007.11073*.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37:145–151.

Rohith Menon and Yejin Choi. 2011. Domain independent authorship attribution without domain adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 309–315.

Franco Moretti. 2000. Conjectures on world literature. *New left review*, 1:54.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50:1–36.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 93–102.

Grant Storey and David Mimno. 2020. Like two pis in a pod: Author similarity across time in the ancient greek corpus. *Journal of Cultural Analytics*.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van

Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Niklas Zechner. 2021. Cross-topic author identification–a case study on swedish literature. In *Swedish Language Technology Conference*, pages 72–78.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, 57:378–393.