



**AALBORG UNIVERSITY**  
DENMARK

**Aalborg Universitet**

Towards Authorship Attribution in the Trykkefrihedsskrifter: A Stylistic Analysis of the Danish Freedom of the Press Writings' Main Writers

Meier, Florian Maximilian

*Published in:*  
pre-print

*Publication date:*  
2022

[Link to publication from Aalborg University](#)

*Citation for published version (APA):*

Meier, F. M. (2022). Towards Authorship Attribution in the Trykkefrihedsskrifter: A Stylistic Analysis of the Danish Freedom of the Press Writings' Main Writers. Manuscript in preparation. In *pre-print*

#### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

#### **Take down policy**

If you believe that this document breaches copyright please contact us at [vbn@aub.aau.dk](mailto:vbn@aub.aau.dk) providing details, and we will remove access to the work immediately and investigate your claim.

# Towards Authorship Attribution in the Trykkefrihedsskrifter: A Stylistic Analysis of the Danish Freedom of the Press Writings' Main Writers

Florian Meier<sup>1</sup>

<sup>1</sup>Department of Communication and Psychology, Aalborg University Copenhagen, Copenhagen Denmark

## Abstract

Authorship attribution is performed on a comparative basis where a text's style is compared to a possible authors candidate's style. Picking the right candidates, however, is not always a straightforward task. By performing a stylistic characterisation of authors, we present three different steps that can support this candidate selection process.

## Keywords

authorship attribution, trykkefrihedsskrifter, stylistic analysis, text mining

## 1. Introduction

Authorship attribution (AA), i.e. finding the true author of a text for which authorship is disputed or unknown, is usually performed on a comparative basis [1]. This means that an authorship model compares a text's style representation with the style of possible author candidates. Finding possible candidates can be done in various ways. One option is that domain experts with knowledge about a collection are able to select possible candidates based on experience from close-reading. However, if a collection is of considerable size, close-reading all texts is extremely time-consuming. Moreover, if the number of texts of unknown authorship is vast a pre-selection of texts with similar style is needed. In both situations, computational approaches for creating stylistic profiles of authors and texts can be helpful to narrow down the number of potential author candidates and texts to be considered in machine learning-based (ML-based) AA experiments. In this poster, we present three steps that can be considered when selecting appropriate candidates for ML-based AA experiments and apply them in the context of the Danish Freedom of the Press Writings.

## 2. The Danish Freedom of the Press Writings

The Danish Freedom of the Press writings ( *Trykkefrihedsskrifter* ), is a collection of pamphlets published and collected during the 1770s in the kingdom of Denmark-Norway. The publication of these short texts was made possible through the abolition of censorship by Johan Friedrich


---

 fmeier@ikp.aau.dk (F. Meier)

 0000-0001-9408-0686 (F. Meier)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

**Table 1**

Lexical concepts and measures for the ten most prominent authors in the collection.

Author	Number of Books	Avg. Book Word Count	Avg. Token Length	Avg. Type Token Ratio	Avg. Herdan C
MartinBrun	54	2323.98	4.53	0.50	0.91
J.C.Bie	16	4213.31	4.69	0.47	0.91
J.L.Bynch	16	6113.00	4.68	0.45	0.90
P.F.Suhm	14	7794.50	4.68	0.43	0.90
SarenRosenlund	14	6202.64	4.48	0.39	0.89
ChristianBagge	11	2596.64	4.95	0.48	0.90
F.C.Scheffer	9	3620.22	4.67	0.46	0.90
Chr.Thura	6	11856.83	4.76	0.38	0.89
L.Jæger	6	12752.00	4.71	0.31	0.88
O.D.Lütken	6	10584.33	5.05	0.36	0.89

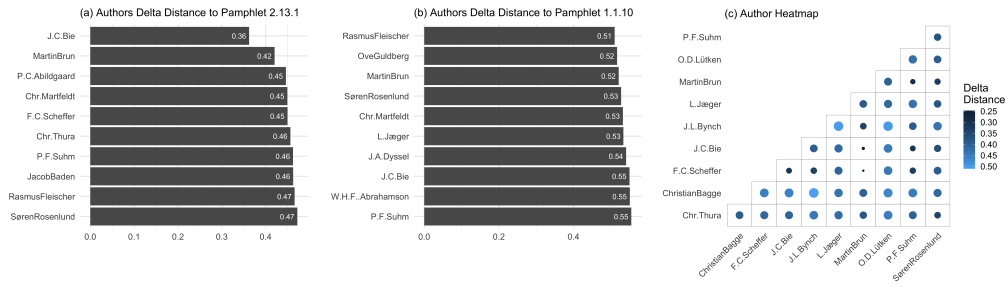
Struensee, the de-facto regent at that time. In these pamphlets, authors could for the first time freely and without restrictions discuss recent events and topics like religion, economy and trade, societal conditions, patriotism or sex. To better understand the collection and study idea generation and knowledge diffusion in that period, knowing who wrote the texts is of high importance. AA in the *Trykkefrihedsskrifter* is, however, a very challenging task, due to the long-tailed distribution of authors in the collection; 116 authors only wrote a single pamphlet and in total around 354 books (almost 50%) are of unknown authorship [2]. We build on a dataset used in [3], which is a digitized and machine-readable version of the Freedom of the Press Writings.<sup>1</sup> In this work, we further filter out eight books that during our analysis process could be identified as non-prosaic. In total, we study 717 pamphlets which are between 117 (min) to 81416 (max) tokens long (Mean=4730 tokens, Median=2904 tokens).

### 3. Finding Candidates for Authorship Attribution Experiments

To find appropriate candidates for our experiments we performed three steps. First, we perform some initial stylistic characterization by looking at lexical measures. Second, we use a distance-based measure to calculate the stylistic distance between author profiles and pamphlets. Third, we compare different dimensionality reduction techniques to study which one follows the intuition about stylistic similarity built-in step two. We have to note that our steps build on the assumption that already known authors with many texts are considered to be the authors of other pamphlets with greater likelihood.

**Measures of vocabulary richness:** In a first step, we investigate whether we can identify idiosyncrasies in writing style by looking at lexical concepts and measures of vocabulary richness. For this purpose, we created Table 1 which shows that Martin Brun wrote 54 pamphlets and is the most represented author in the collection. The *avg. word count per book* indicates that his books are rather short. Other writers, especially L.Jæger, write much longer pamphlets. With the feature *average token length* we try to cover the aspect of who might use longer words than others. However, when comparing the top ten authors no striking differences become evident. Type token ratio (TTR) is a measure of vocabulary richness. The higher this ratio the more unique words an author is using which hints to more skilled and varied word use. However, care needs to be taken as this value is not normalized with respect to text length which means

<sup>1</sup>The digitalization was performed by the Royal Danish Library and is accessible here: <https://www.kb.dk/inspiration/trykkefrihedens-skrifter>.

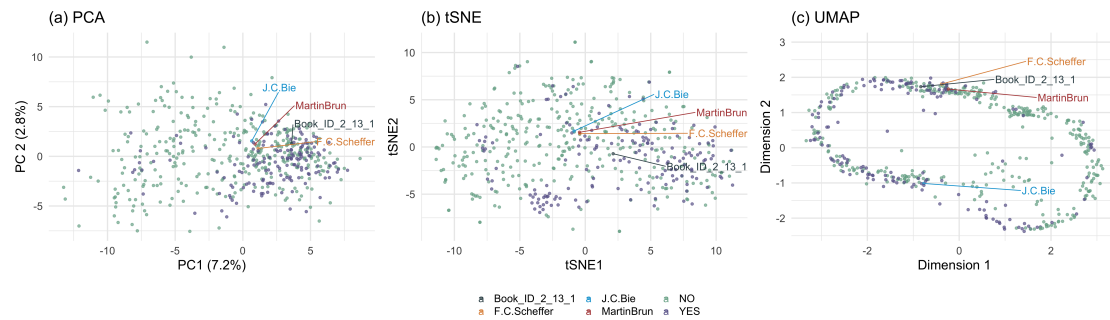


**Figure 1:** Distance between two pamphlets of the collection and between the most prominent authors.

that longer texts automatically have a lower TTR. In our case, Martin Brun seems to use a richer vocabulary compared to Thura, Lütken and Jaeger, but this is probably due to their texts being longer on average. A measure of vocabulary richness normalizing for text length is Herdan's C [1, p.28]. When looking at these values basically no differences become evident. To sum up, the lexical measures did not unearth any noticeable differences which implies a high similarity in style on that level.

**Delta distance:** A commonly used distance measure in AA is Burrow's delta [4]. Burrow suggests using the most frequent word types (MFW) as these words mainly correspond to function words which are often used unconsciously by authors, thus inherently reflecting his or her writing style. In our case, every author is represented by an author profile, which is created by concatenating all texts by this author to one entity. For all author profiles and pamphlets of unknown authorship, i.e. single texts, we create feature vectors by taking the 300 MFWs of uni-, bi- and trigrams. The values in the vectors are not raw term frequencies (TF) but get z-normalized relative to text length and with respect to the occurrence frequency of that feature in the corpus. The feature vectors can then be used to calculate the pairwise delta distance between (a) authors and their author profiles, (b) author profiles and pamphlets and finally (c) among pamphlets. These pairwise comparisons can help us investigate different aspects of our AA problem. First, the delta distance between single pamphlets and author profiles can give insights into which authors might be likely author candidates and should be considered as such in additional ML experiments. For example, Figure 1(a) indicates that Burrows delta between pamphlet 2.13.1 and J.C. Bie is considerably smaller compared to other authors in the collection, which makes him a likely author candidate. However, Figure 1(b) also highlights a problem of this approach as no real differences in distance between pamphlet 1.1.10 and authors becomes evident. This could indicate that even in a more advanced ML-based AA experiment no real conclusion will be able to be drawn. Finally, the pairwise distance between author profiles gives us a general picture about which authors are similar in style and might be difficult to distinguish. Figure 1(c) shows a heatmap of the top ten authors and the delta distance between their profiles. We can see that the distance between Brun and Scheffer, Brun and Bie, as well as Brun and Suhm is very small indicating similar writing styles. Authors like Bynch and Bagge or Bynch and Jæger seem to be stylistically far apart and are thus more easy to distinguish from each other in ML-based AA experiments.

**Dimensionality reduction:** In a final step, we investigated how dimensionality reduction



**Figure 2:** Three dimensionality reduction techniques applied to the author profiles and texts.

techniques and their possibilities of creating 2D visualisations of  $n$ -dimensional feature vectors support the intuition gained via the delta distance measures and give further evidence when picking appropriate candidates in ML-based experiments. Figure 2 compares (1) principal component analysis (PCA), (2)  $t$ -distributed stochastic neighbor embedding ( $t$ -SNE) and (3) uniform manifold approximation and projection for dimension reduction (UMAP). We note that Figure 2(c) UMAP, unlike (a) PCA and (b)  $t$ -SNE, does not support the intuition built in Figure 1(c), that J.C. Bie, F.C. Scheffer and Martin Brun are stylistically similar. However, the percentage of variance explained by the two dimensions in PCA is too low to consider this as a viable solution, leaving  $t$ -SNE as the most appropriate technique.

## 4. Conclusion and Future Work

To solve the problem of unknown authorship in the *Trykkefrihedsskrifter* ML-based AA experiments need to be performed. In the future, we want to use a combination of the delta distance and  $t$ -SNE visualizations to select appropriate candidates for these experiments. However, as apparent in the case of Brun, Bie and pamphlet 2.13.1 the stylistic closeness of some authors could make this yet again a challenging task.

## References

- [1] J. Savoy, *Machine Learning Methods for Stylometry. Authorship Attribution and Author Profiling*, 1st. ed., Springer Nature, Cham, Switzerland, 2020.
- [2] H. Horstbøll, Luxdorps samling af trykkefrihedens skrifter 1770-1773, *Fund og Forskning* 44 (2005) 397–440.
- [3] F. Meier, B. Larsen, F. Stjernfelt, Exploring the potential of bootstrap consensus networks for large-scale authorship attribution in luxdorps’s freedom of the press writings, in: *Proceedings of DHN 5th Conference, 2020*, pp. 110–124. URL: <http://ceur-ws.org/Vol-2612/paper8.pdf>.
- [4] J. Burrows, Delta: a measure of stylistic difference and a guide to likely authorship, *Digital Scholarship in the Humanities* 17 (2002) 267–287.