Report

Project: Stylometry of pre-revolutionary French pamphlets: authorship attribution and verification

Antonina Martynenko

January 2024

The goal of the project was to investigate authorship attribution and verification possibilities for non-fictional historical texts connected to the French revolution.

The objectives were twofold: first, to compile a sample corpus and test its features regarding historical language; second, to verify the authorship of specific pre-revolutionary writings.

We focused on authors associated with *Palais Royal* as important pre-revolutionary figures. We chose fragments from the *Galerie Universelle*, a little-known multi-volume encyclopaedia, which is historically proven to be connected with the *Palais Royal* circles, as our test texts. Our hypothesis was that the compiled corpus can be used to test whether *Galerie Universelle* sections were written by people close to *Palais Royal*.

Over the four-month fellowship, a text corpus by selected authors was compiled and tested for authors' styles coherence. After achieving a high level of same-author clustering, I analysed the texts in question using exploratory stylometric methods (hierarchical clustering) and General Imposters for authorship verification; for both `stylo` implementation was used (Eder et al., 2016). The analysis showed that the methods usually applied to fiction (e.g., novels) can be adjusted to non-fictional historical texts, however the problems of genre bias and text reuse should be addressed in further studies.

1 Background

Text in question choice

The subject of analysis, *Galerie universelle*, was taken as an example of a non-fictional text from the pre-revolutionary period. The *Galerie* itself is a largely understudied source with no modern scholarship, so I have done an additional background research on the sources and history of the encyclopaedia.

1.1 Historical context

The *Galerie Universelle* is a series of biographical 'portraits' of notable individuals from various times, cultures, and countries. Scarce information about the sources and authors of the portraits were found in the 19th-century and early 20th-century sources. In particular, I found an excerpt from *Mémoires de la Société académique d'agriculture, des sciences, arts et belles-lettres du département de l'Aube* (1901) that gave us an evidence of the *Galerie Universelle* project was linked to *Lycée de France* and *Palais Royal*.

Other found sources helped to identify the authors of certain fragments in the *Galerie*[1]. The *Galerie* itself references an author of introductory fragments as 'M. B***', possibly implying Jacques Pierre Brissot, whose involvement in the project is confirmed in his correspondence with Gabriel Luce de Villar.

No additional information is available neither about the editor of the *Galerie* Comte Imber de la Platière, nor about the publishing process and editions of the *Galerie*. Even though some general bibliographical sources name Imber de la Platière as the sole author of the *Galerie*[2], multiple sources confirm that each portrait may have been written by a different author, making it a multi-authored collection.

[1] Namely: l'abbé Sabatier de Castres - the author of the portrait of Marie-Thérése; Ludwik Antonie de Caraccioli - the portrait of Wacław Rzewuski; l'abbé d'Espagnac - the portrait of Catinat.

[2] See, e.g., *Nouveau dictionnaire bibliographique ... par Nicolas-Toussaint Des Essarts* (1801).

1.2 Editorial history

There are at least two editions of the *Galerie Universelle*. The first one published in 1785 as a book that includes introduction in five parts and five portraits. The second edition, which is roughly dated 1787-1788, consists of multiple separate brochures. A list of 75 brochures is available on Gallica[3], an 8-volume edition presumably compiling the same brochures is available on Google Books (Bavarian State Library's exemplar). However, there are some discrepancies between the portraits included in the two sources (7 brochures from Gallica are not found in the 8-vol. edition; 3 portraits included in the 8-vol. are not found among the Gallica's list of brochures). The bibliographical inconsistencies are complicated by the fact that only the 1785 edition has continuous pagination, in other cases the pagination in each portrait restarts from page 1.

1.3 Composition

The *Galerie* editions available for our study include five introductory fragments (1785), 70 portraits, and number of editorial documents (e.g., dedications and advertisements written by Imber de la Platière, 1787-1788). Two volumes of the 8-volume edition are dedicated solely to descriptions of female figures (volumes 2 & 3). No other organisational or character selection principles of the *Galerie* were found. Each portrait in the collection is about 50 pages long (about 10 000 words).

1.4 Texts in question

I tested the authorship of the five introductory fragments from the 1st volume of the 1785 and four additional portraits from the 8-volume edition (namely, the portraits of *Phillippe duc d'Orléans* (vol. 4), *Le Chancelier de l'Hôpital*, *Le comte Lally de Tollendal*, and *Colbert* (vol. 8)). The authorship of the introductory fragments is intriguing due to the connection of the whole *Galerie* project to *Palais Royal*: we hypothesise that 'the programme' of this multi-volume work could have been written by a prominent member of the *Palais Royal* circle. The four portraits were chosen with a similar rationale: the portrayed individuals are related with the political context of the 1780s and *Palais Royal* itself.

[3]http://ark.bnf.fr/ark:/12148/cb30634165v

2 Attribution of the *Discourse préliminaire*

2.1 Summary

Corpus: 29 texts by 12 authors, 18th-century sources, Gallica's OCR;

Text in question: *Discourse préliminaire*, 3670 words (1st vol. of *Galerie Universelle*, 1785)

Hypothesis: the text in question is written by Jacques Pierre Brissot

Result: the probability for the *Discourse préliminaire* to be written by any of the 11 authors (incl. Brissot) is very low.

Corpus & code: see the Github page

During the first month of work, a preliminary corpus of 11 authors was compiled to test the authorship of the introductory part of the *Galerie Universelle* titled *Discourse préliminaire, ou de l'influence des lettres sur les hommes en société*. As this part of the encyclopaedia gives introduction to the goals and aims of the whole multi-volume series, our hypothesis was that these parts might be written by one of the major writers connected to *Palais Royal*, specifically, by Jacques Pierre Brissot. We tested this hypothesis using clustering techniques and analysing distance distributions. Neither method provided strong evidence that *Discours préliminaire* or any other introductory parts were written by any of the 11 authors.

2.2 Corpus

- Composition. Eleven authors were selected for the corpus: Bergasse, Brissot, Condorcet, D'Alembert, Delisle de Sales, Garat, Gouges, La Harpe, Marmontel, Sieyes, Villar.
- Bibliography. For each author, we obtained a list of digitised books available on Gallica and selected from two to four books per author. The books were chosen based on genre similarity to the text in question. The texts for chosen books were retrieved automatically.
- Preprocessing. Gallica's OCRs were used; the most frequent OCR errors (based on the 500 MFW) were fixed with simple replacements, such as converting 'eft' to 'est', 'fi' to 'si', etc.

- Sampling. Due to the variation in book sizes in words, the largest texts were sampled down to 15 000 randomly taken words. This prevented an author's corpus from being biased towards the themes of a single, overly long book. The reduced samples were returned back to each author's text collection that results in samples from 45 000 (Delisle de Sales) to 11 423 (Gouges) words.
- Extension of the sample of anonymous texts. The text in question is accompanied in the volume by other introductory texts. We included these texts in our corpus to check if all introductory texts, including the *Discourse préliminaire*, were written by the same author. Extending the test set is advantageous as other introductory parts are longer than the text in question, which is only 3 670 words.

2.3 Analysis

2.3.1 Corpus heterogeneity

Despite selection of writings that should be thematically close to Galerie Universelle, a simple hierarchical clustering of individual texts reveals not authorial, but thematic or genre signal in clusters (see, for example, a cluster uniting *Éloges* by D'Alembert and Condorcet in the Figure 1).

To reduce the genre signal, I merged separate texts by each author into one text file (the details on sampling are explained above, *Corpus - Sampling*). After that independent samples were taken from each author's text file. Each text from the *Galerie Universelle* (abbr. as *G1785*) was processed separately.

Random samples of 1800 words from each author lead to perfect author-wise clustering, implying that after all manipulations we managed to capture the authorial rather than thematic signal. However, the resulting tree (Figure 2) shows no author to be associated with either of the texts in question (Stylo settings: 200 MFW, cosine distance, Ward's criterion).

2.3.2 Stylometric exploration

As the text in question seems to be highly similar in style to other texts from the same volume but it is also very short (3 670 words), I tested larger samples from other texts from *G1785*. The clustering, based on two samples of 3 000 words from each author, results in the dendrogram showing some

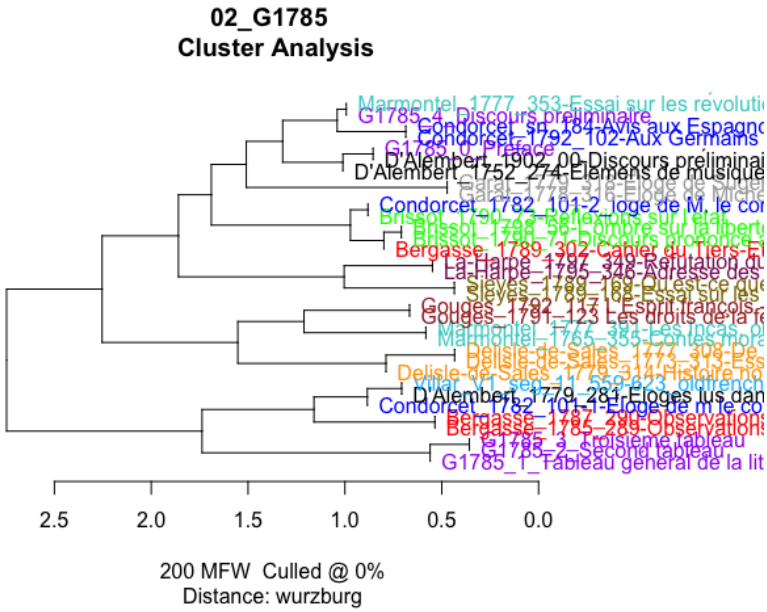


Figure 1: Initial cluster analysis for individual texts by 11 authors. Texts in question are abbreviated as *G1785*.

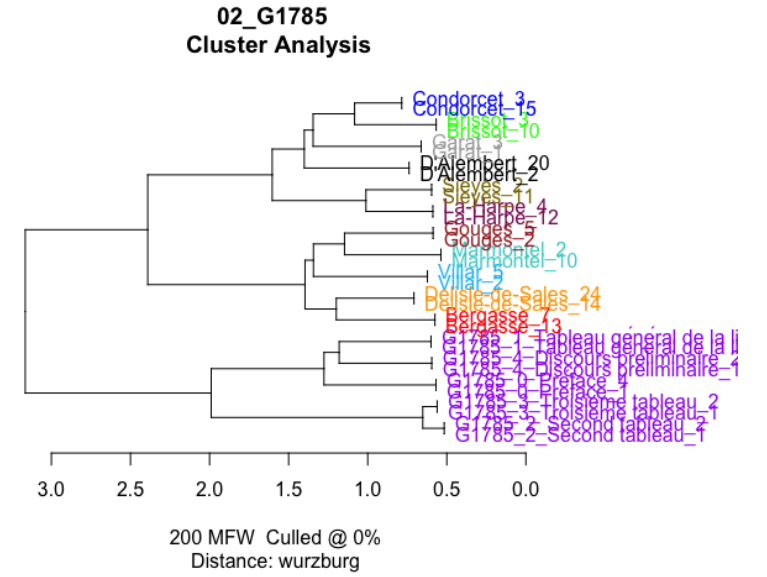


Figure 2: Clustering after merging each author's texts into one file, two independent samples of 1 800 words taken.

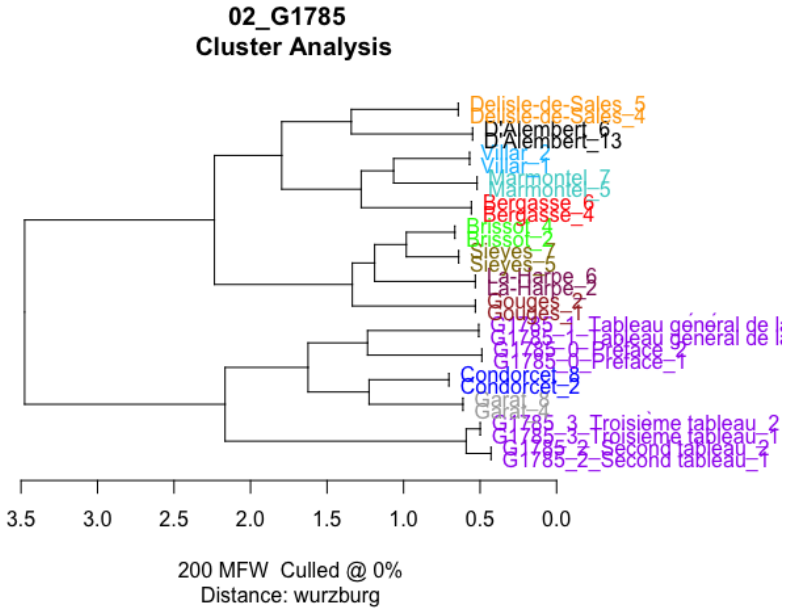


Figure 3: Cluster analysis of longer samples (3000 words).

similarities between Condorcet's and Garat's writings and the introductory parts of the *G1785* (Figure 3).

Removal of 'strong' words

Using the `seetrees` package[4] it is possible to extract the most distinctive words for the two root branches, as illustrated in Figure 4. Some of these words represent purely thematic signal, for instance, words related to the topics of language and history (*gaules, langue, lettres, parler, langage, arts, sciences, louis, france, francois*, etc.). At the same time, there are distinctive words suggesting different modes of narration, such as the 1st vs 3rd person pronouns and verb forms (see the word list associated to the *Cluster 2*, Fig. 4).

In attempt to homogenise the corpus, a list of words related to language and history was removed alongside the 1st and 2nd person pronouns. These manipulations yet had little impact on the clustering, giving perfect author clusters and a separate cluster for all texts from *G1785*. The cluster-influensing

[4]See: https://github.com/perechen/seetrees

Figure 4: 'Strong' words in clusters filtered by seetrees package.

words were removed in number of trials and different number of MFW were tested during the cluster analysis (100, 200, 300, bootstrap consensus tree from 50 to 250 MFW). In some trials, the texts in question were positioned in the neighbouring branches with samples from Condorcet, Garat, or D'Alambert, but they were never placed in the same branch with other author's samples. To sum up, the hierarchical clustering provided inconclusive results, telling us more about the thematic closeness of some samples rather than the authorial signal.

2.3.3 Distances distribution

To have a closer look into texts similarities not influenced by tree-based representation, I analysed the distance distributions between each pair of texts, a method proposed by Artjoms Šeļa (Šeļa, 2023). The general idea behind the method is that the iteratively taken distances from an author to itself should form a distribution with a smaller mean than distribution of distances to a different author. The distances are build on a random number of MFW ranked from 50 to 500, 100 iterations taken. The resulting distribution matrix (Figure 5) shows that none of the authors closely match the distance distribution of the texts in questions to themselves. This allows us to suggest that, given the corpus, there is a very small probability that any of the introductory parts were written by any of the 11 authors, including Brissot.

3 *Galerie Universelle* as an object of stylometric study

3.1 Summary

Corpus: 434 texts by 37 authors, 18th-century sources, Gallica's OCR;

Texts in question: *Galerie Universelle*'s introductory texts (1785) and four 'portraits' (1787-88);

Hypotesis: By gathering more authors who were close to *Palais Royal*, there is a chance to attribute at least some parts of the *Galerie Universelle*;

Results: Although we see issues with corpus compilation and heterogeneity, there is an evidence that stylometric methods works on this type of data; in

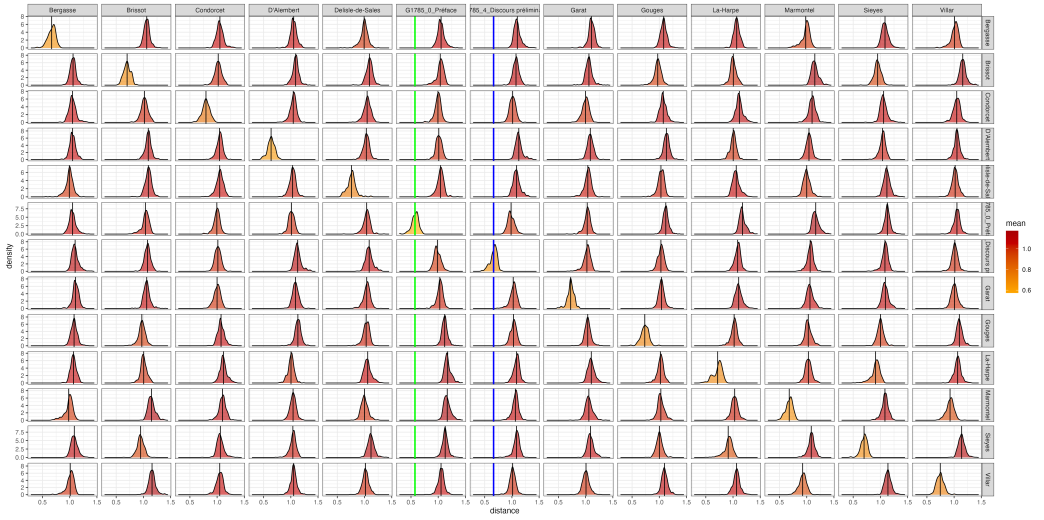


Figure 5: Distribution of distances from an author to himself and to other authors.

particular, we were able to find a suspect author for at least one of the four portraits.

Corpus & code: see the Github page

As the preliminary analysis gave no conclusive results, during the following months of the fellowship I worked on the expanded version of the corpus and of the whole experiment. The corpus was extended from 11 to 37 authors, related to *Palais Royal* circles. With this larger corpus, the authorship of two sets of texts will be addressed: firstly, the same introductory texts from the *Galerie Universelle* (vol. 1 dated 1785; abbr. as *G1785*); secondly, the four portraits from the 8-volume edition of the *Galerie* (abbr. as *G1787*). As the set of potential authors is not closed, the main method for the analysis was General Imposters method (Koppel and Winter, 2014; Kestemont et al., 2016; Eder, 2018), with hierarchical clustering used only for exploration. The aim of the experiment was to see if General Imposters method would be able to consistently identify if any of the texts in question were written by an authors present in the corpus.

3.2 Corpus

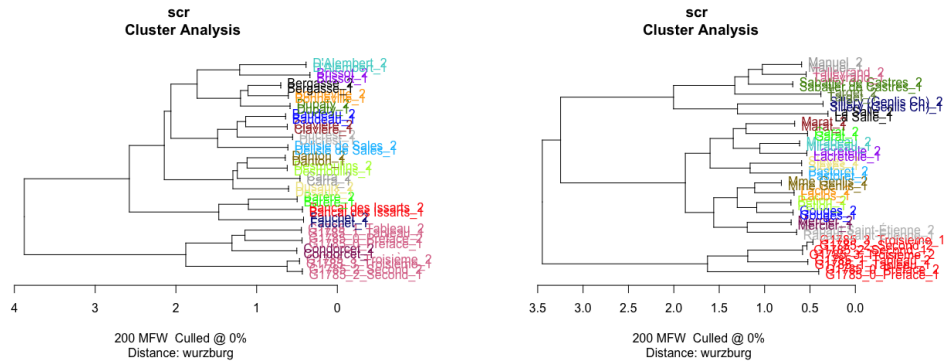
- Bibliography. 37 authors[5] associated with the *Palais Royal* were selected for the extended bibliography. For each author a list of all books printed before 1820 and digitized by Gallica was exported, then thematically similar to the *Galerie* texts were selected (without limits in the number of books). In the cases where an author's corpus was too small, all available books were included in the corpus.
- Corpus size. The resulting corpus for this experiment comprises 434 texts from 37 authors. Most of the author's samples include more than 15 000 words; for two authors it was not possible to collect enough data (Imber de la Platière's corpus contains only 902 words, all Epremesnil's writings total only 5874 words).
- Preprocessing. The preprocessing steps were the same as in the **Section 2**: the longest texts were reduced in size, a list of replacements used to improve the OCR quality. All texts by one author collapsed into one file, drawing independent samples for the analysis from this file.

3.3 Analysis

3.3.1 Authorship of the introductory fragments (G1785)

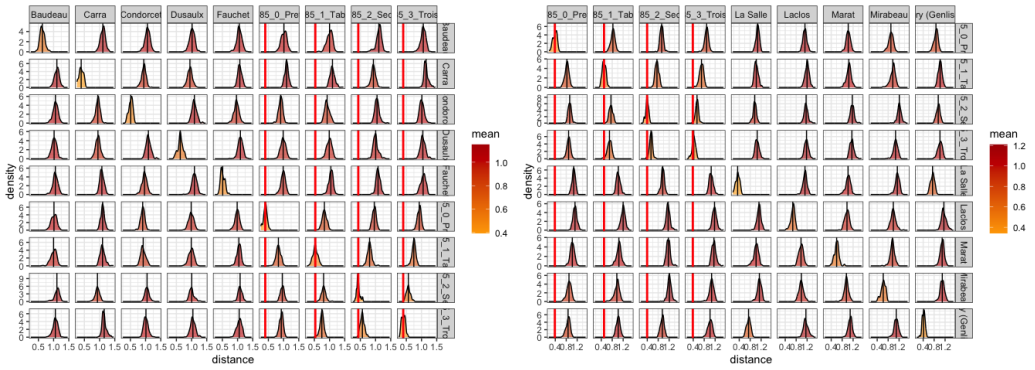
The stylometric exploration on the expanded corpus shows similar results from our previous study with the 11-authors corpus. The introductory parts of the *Galerie* typically form a separate cluster (Figure 6a); in rare cases, they cluster with Condorcet's or other authors writings (Figure 6b). However, the clustering does not provide concrete evidences on the closeness between G1785 and Condorcet, or any other author, as confirmed by the distribution of distances (Figure 7). The distribution of distances reveals a similarity between G1785-2 and G1785-3, suggesting that these two texts might have been written by the same author.

[5]Bancal des Issarts, Barère, Baudeau, Bergasse, Bonneville, Brissot, Carra, Clavière, Condorcet, D'Alembert, Danton, Delisle de Sales, Desmoulins, Ducrest, Dupaty, Dusaulx, Epremesnil, Fauchet, Garat, Gouges, La Platière, La Salle, Laclos, Lacretelle, Manuel, Marat, Mercier, Mirabeau, Mme Genlis, Pastoret, Petion, Rabaut Saint-Étienne, Sabatier de Castres, Sieyès, Sillery (Genlis), Talleyrand, Target.



(a) Authors set 1, G1785 in pink. (b) Authors set 2, G1785 in red.

Figure 6: Introductory parts' clustering (authors are divided into two sets for visibility).



(a) Authors set 1 (b) Authors set 2

Figure 7: Distances distributions (selected).

3.4 General Imposters

General Imposters method was used to determine whether an author of any introductory part is present in the corpus. From the exploratory analysis we assume that this is unlikely, although the texts G1785-2 and G1785-3 might have the same author.

3.4.1 GI implementation

The imposters method was applied via its default implementation in `stylo` (`imposters(distance = 'wurzburg')`). As a test set I used two independent samples from one of the G1785 texts, all the other texts used as reference corpus.

For each sample of 3000 words:

- 50 iterations of general imposters was performed;
- 50% of the features were tested in each trial;
- for each GI iteration a new set of independent samples was taken from the corpus;

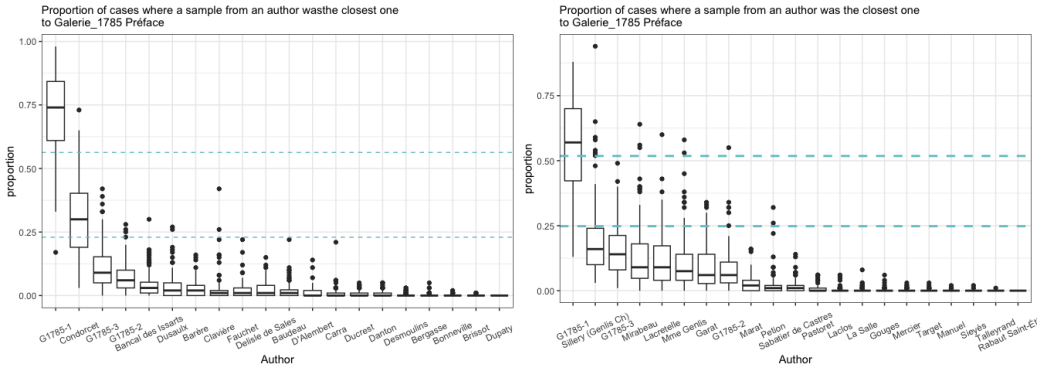
As each test text had two samples, the GI procedure was applied a hundred times to each text in question[6]. Afterwards a 'confidence interval' for the GI was calculated using the `imposters.optimise()` function, iterated over the data 20 times.

3.4.2 Authorship verification of the introductory texts

Figures 8 and 9 display the results for the test texts G1785-0 and G1785-3, respectively. Each boxplot depicts the results from 100 iterations of GI, with higher values indicating a higher proportion of cases when an author was the closest one to the test text.

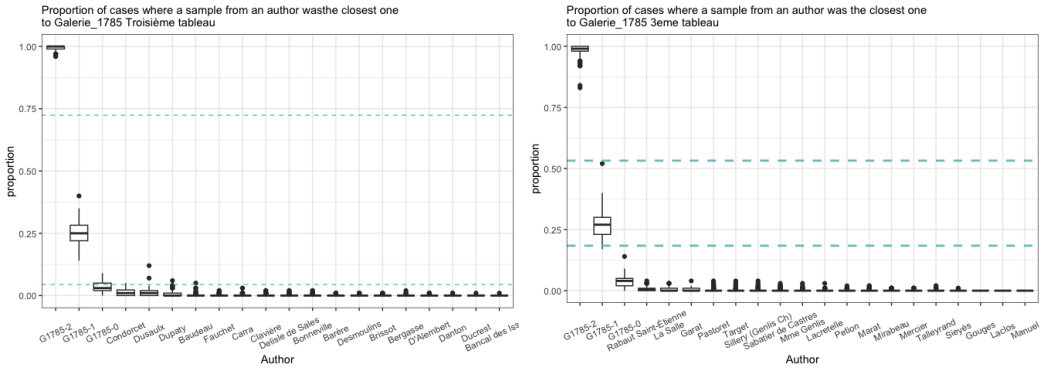
Figure 8 demonstrates that there is a high probability of G1785-0 (*Préface*) to be written by the same author as G1785-1 (*Premier tableau*). The GI results also reflect *G1785-0*'s closeness to Condorcet's writings, a relationship previously visible from the dendrograms. Nevertheless, the values indicating Condorcet's authorship are placed clearly beneath the zone of confidence

[6]The pipeline is based on the studies Eder, 2018 and Šeļa, 2023.



(a) Authors set 1 (b) Authors set 2

Figure 8: General Imposters trials results for the G1785-0 (Préface)



(a) Authors set 1 (b) Authors set 2

Figure 9: General Imposters trials results for the G1785-3 (Troisième tableau)

(dashed blue lines). Thus, given the corpus, there is not sufficient evidence to propose Condorcet as a potential author. Importantly, the GI method clearly rejected the hypothesis that any other author in the corpus could be the author of the G1785-0 (*Préface*), including the authors of the G1785-2 and G1785-3.

Figure 9 shows the case then G1785-3 (*Troisième tableau*) was set as a test text, with only a negative result obtained. None of the known authors' styles in the corpus are similar to this fragment. As we expected, there is still a high chance of G1785-2 and G1785-3 were written by the same author, who is not included in the current corpus.

3.5 Authorship of the four 'portraits' (G1787)

3.5.1 Exploration

Simple clustering techniques applied to the 4 portraits from the 8-volume edition did not yield a concrete result, see Figure 10. However, in contrast with the introductory texts *G1785*, these texts (*G1787*) do not have a tendency to form a separate cluster but tend to be close to other authors' samples in the corpus.

3.5.2 Authorship verification of the portraits

Each of the four portraits was used as a test text for iterative GI analysis, using the method described above (2 samples of 3000 words, 50 GI iterations for each sample, new independent samples taken in each iteration). The results are summarised in Figures 11 and 12.

Based on the results, it can be suggested that texts G1787-5 and G1787-7 (portraits of Phillippe-duc d'Orleans and Le comte Lally de Tollendal) were likely written by the same author, who is not included in the reference corpus. There is no clear conclusion about the authorship of the 'portrait' G1787-8 (Colbert).

The most intriguing result was obtained for the portrait of Chancelier de l'Hôpital (G1787-6, Fig. 11b). There is significant evidence to suggest that Claude Fauchet could be the author of the fragment. Further investigation into the portrait's style and thematics, with adjustments to the reference corpus as needed, is required to confirm this attribution.

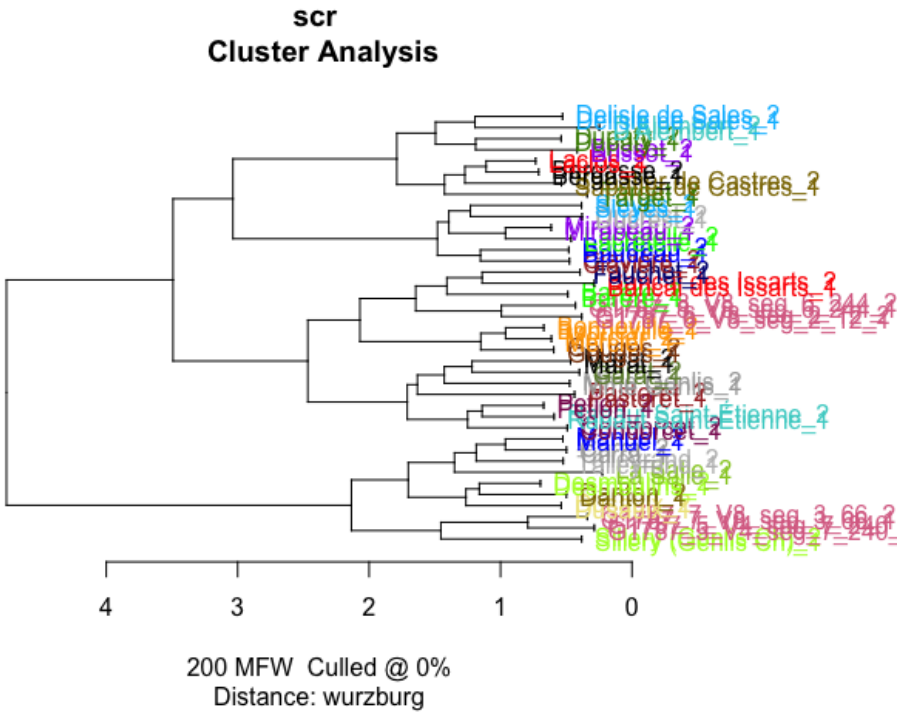
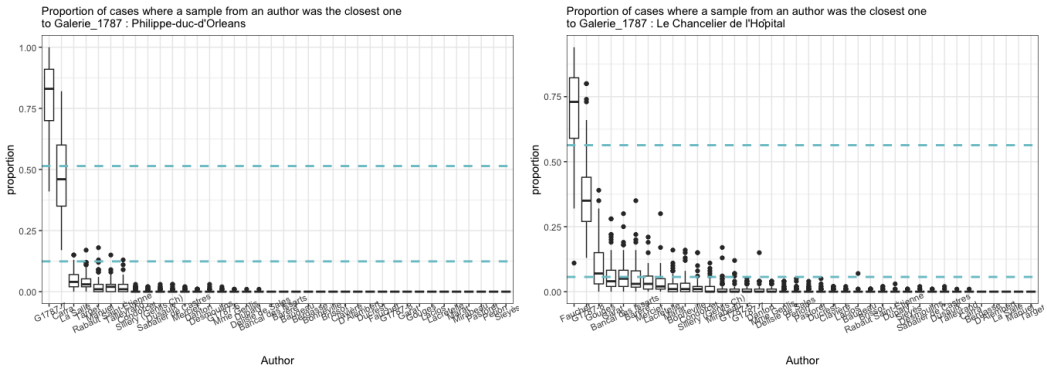
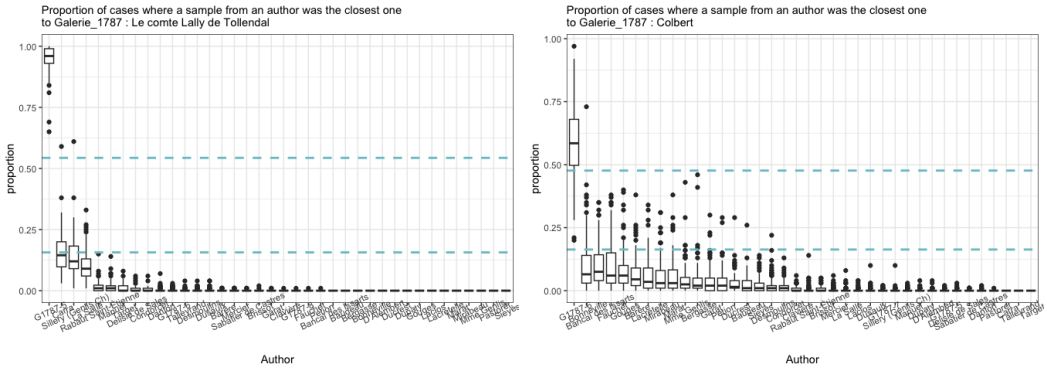


Figure 10: Basic cluster analysis of the samples from the four portraits (in pink, abbr. as G1787).



(a) Phillippe-duc d'Orleans (b) Chancelier de l'Hôpital

Figure 11: General Imposters trials results for the portraits G1787-5 (Phillippe-duc d'Orleans) and G1787-6 (Chancelier de l'Hôpital).



(a) Le comte Lally de Tollendal (b) Colbert

Figure 12: General Imposters trials results for the portraits G1787-7 (Le comte Lally de Tollendal) and G1787-8 (Colbert).

4 Conclusions and prospects

This study explores the challenges and potential of authorship attribution and verification for pre-revolutionary non-fictional texts in French. The preliminary findings display limited exploratory potential in hierarchical clustering for this type of data, but the General Imposters method predictably demonstrates better robustness and interpretability for an open authorial set problem. Thanks to the GI method, it was possible not only to find a possible suspect for one of the texts but, more importantly, to eliminate authors with low probability of connection to the texts. A closer examination of the results reveals consistency between the clustering and GI outputs.

A more advanced project in the authorship of the 18th-century political pamphlets in French will require more work in data preparation and methods application.

Additional bibliographic search and digitization will be needed. The major issue in this case is the style/genre heterogeneity of the non-fictional texts, which manifested itself in the mode of narration (1st vs 3rd person) and other very frequent vocabulary features that influence the text's overall style. One potential solution is corpus harmonisation at the bibliography level, meaning the compilation of a genre-specific bibliography and corpus with a) roughly thematically-homogenous and b) well-attributed texts (no dubious, co-authored or edited books).

Another challenge, common for the texts of this historical period, is text reuse (see, e.g., Olsen et al., 2011). Its influence on the stylometric analysis and authorship attribution should be assessed separately.

Despite current study showed no visible influence of different OCR algorithms on the stylometric analysis, there is an evident possibility of having one. Therefore, a controlled OCR pipeline, tailored to 18th-century prints, should be used for further corpus building (e.g., OCR4all infrastructure).

The stylometric methods applicable for 19-20th century narrative texts might be less efficient for earlier writings, as illustrated by the preliminary naive clustering in Figure 1. In this study, perfect author clustering was only possible after grouping texts by each author, though no author clusters appeared at the level of separate texts. On the one hand, further study should test other features than words, such character n-grams and combined features, which may be more effective for historical texts (Eder, 2013; Camps et al., 2020). Some models also take genre influence into account in stylometric analysis (e.g., Schöch and Riddell, 2014). On the other hand, it is also possible that 18th-century non-fiction genres might have less individual stylistic imprint than fictional texts. Thus the methods applied to non-fictional (e.g., political) texts should be additionally tested and probably adjusted. This kind of ground-truth testing task requires a well-built corpus that is partly described above.

References

Camps, J.-B., Clérice, T., & Pinche, A. (2020, December). Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis [arXiv:2012.03845 [cs]]. Retrieved September 9, 2023, from http://arxiv.org/abs/2012.03845

Eder, M. (2013). Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, *28*(4), 603–614. https://doi.org/10.1093/llc/fqt039

Eder, M. (2018, May). Authorship verification with the package stylo. https://computationalstylistics.github.io/docs/imposters

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *8*(1), 107. https://doi.org/10.32614/RJ-2016-007

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, *63*, 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178–187. https://doi.org/10.1002/asi.22954

Olsen, M., Horton, R., & Roe, G. (2011). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies/Le champ numérique*, *2*(1). https://doi.org/10.16995/dscn.258

Schöch, C., & Riddell, A. (2014, July). Progress Through Regression. Modeling Style across Genre in French Classical Theater [Publisher: Zenodo]. https://doi.org/10.5281/zenodo.7785295

Šeļa, A. (2023). Navalny_r. https://github.com/perechen/navalny_R/tree/main