Evaluating Hierarchical Clustering Methods for Corpora with Chronological Order

Philippe Gambette, Olga Seminck, Dominique Legallois, Thierry Poibeau

Evaluating Hierarchical Clustering Methods for Corpora with Chronological Order

Philippe Gambette[1,2], Olga Seminck[2], Dominique Legallois[2], Thierry Poibeau[2]
philippe.gambette@univ-eiffel.fr, olga.seminck@cri-paris.org, dominique.legallois@sorbonne-nouvelle.fr, thierry.poibeau@ens.psl.eu

The style and language of an author evolves over time, but how and to what extent? Is evolution linear or is it more erratic? In stylometry, those questions are often addressed with hierarchical clustering. Hierarchical clustering is popular in Digital Humanities to classify texts by degree of similarity. When texts can be ordered chronologically, it is often expected that texts which are closer in the chronology will also be more similar, therefore the tree obtained from hierarchical clustering is also expected to be consistent with the chronological order of texts. This hypothesis appears not only in stylometry, when studying the evolution of the style of an author, but also in Historical Linguistics, when analysing for example a collection of Old English, Middle English, and Early Modern English texts (Moisl 2020), or in discourse analysis, for example with New Year's greetings by presidents of the French Fifth Republic (Leblanc 2016: 63, 67, 86, 87).

Hierarchical clustering can traditionally be represented through a dendrogram: a rooted tree whose leaves are documents, the length of the path between two leaves representing the stylistic/linguistic distance between the documents (see Figure 1). Clusters correspond to branching nodes: the shorter the distance between two nodes, the more they are expected to share stylistic and linguistic features. Hierarchical clustering is a method that is easily accessible thanks to freely accessible implementations, with the R package Stylo (Eder et al., 2016) for example.

We wonder how much the resulting dendrogram is consistent with the chronological order of writing. Indeed, this would provide us with a method of evaluating the result of the clustering. More precisely, the question we want to answer is: can the branching nodes of the dendrogram be re-ordered so that its leaves follow a chronological order as best as possible, while of course preserving the structure of the dendrogram?

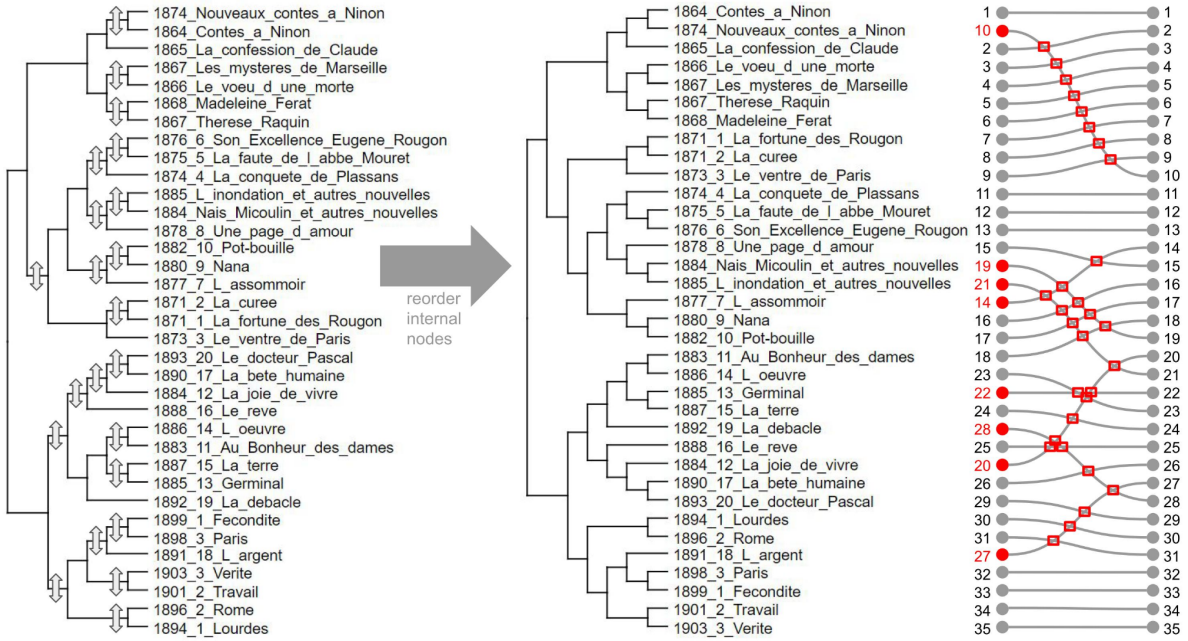
One needs to keep in mind that hierarchical clustering does not provide a fixed order on the leaves of the output dendrogram: given a binary partition of a corpus in part A and B, A can be represented before or after B. "Seriation" approaches have been introduced in the literature to find an optimal order of the leaves, based on various criteria (Bar-Joseph et al, 2001; Chae & Chen, 2011). Here, we introduce a more straightforward approach to evaluate, directly on the dendrogram obtained by hierarchical clustering, how much its topology is consistent with an ordered list of items (novels ordered chronologically in our case). In the ideal case, if the clusters below each node of the dendrogram can be reordered so that the order of the leaves corresponds exactly with the chronology of the novels, it means that the clusters displayed by the dendrogram are consistent with a chronological evolution of the style of an author.

[1] Université Gustave Eiffel
[2] Lattice (Langues, Textes, Traitements informatiques, Cognition) - CNRS & ENS/PSL & Université Sorbonne nouvelle

We thus developed methods based on two criteria to reorder the branching nodes of the dendrogram, so that the obtained order on the leaves is as close as possible to the actual chronological order.

The first criterion is the minimum number of conflicts between the chronological order and the order on the leaves of the dendrogram. More precisely, we want to minimize the number of pairs of leaves which are ordered differently in the dendrogram and in the chronological ordering. The second criterion is the minimum number of leaves which have to be deleted before the dendrogram respects the chronological order. This criterion is particularly relevant if the input chronological order may contain errors.

Both criteria are illustrated in Figure 1, where the dendrogram is built from a corpus of novels by Émile Zola extracted from corpus CIDRE (Seminck et al., 2021), classified using *motifs* (Legallois et al., 2018) and analysed with Stylo.



(a) (b)

Figure 1. A dendrogram built from motifs of novels by Émile Zola gathered in corpus CIDRE, where the file name starts with the first publication year (a): the vertical arrows are located next to each branching node whose left and right child should be exchanged in order to get a leaf ordering with 8 leaves to delete (10, 14, 19, 20, 21, 22, 27, 28) to get the remaining leaves in the chronological order (b), which minimizes the number of conflicts with the chronological order (each of the 30 red rectangles corresponds to a conflict between the two orders)

It is also possible to evaluate in both cases whether the obtained value for each criterion is lower than what we would expect by chance (for example for small dendrograms). To this aim, we can compare with the results obtained if the input order is not the chronological order but a random order. This would correspond to a situation where no chronological signal could be captured from the dendrogram built from the clustering algorithm. Therefore, for each criterion, we can estimate a

probability that the obtained value is lower than what would be expected on random data, by comparing this value with those computed on the same dendrogram and 10 000 orders picked uniformly at random.

This estimation relies on fast algorithms, implemented in Python (available at https://github.com/oseminck/tree_order_evaluation), which run in polynomial time for both criteria on dendrograms where each branching node has a fixed maximum number of children.

Thanks to these two evaluation algorithms, we can test whether our models of the evolution of the individual style of different authors (generated by the R package Stylo), which will be described in more details in upcoming publications, are able to capture the signal of a chronological evolution, and to quantify how improbable it is that this signal could be caused by chance. Future work will also contain a stage of interpretation: determining what is linguistically significant in the evolution of the style of an author and what may explain conflicts between the expected order and the observed order.

Acknowledgements

This work was funded in part by the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Bar-Joseph Z, Gifford DK, Jaakkola TS (2001). "Fast optimal leaf ordering for hierarchical clustering". *Bioinformatics*, **17**(supp. 1), S22–S29.

Chae M, Chen J (2011). "Reordering hierarchical tree based on bilateral symmetric distance". *PLOS One*, **6**(8), e22546.

Eder M, Rybicki J, Kestemont M (2016). "Stylometry with R: a package for computational text analysis". *R Journal*, **8**(1), 107–121.

Leblanc J-M (2016). *Analyses lexicométriques des vœux présidentiels*. ISTE editions, p.63, 67, 86, 87.

Legallois D, Charnois T, Larjavaara M (2018). "The Balance Between Quantitative and Qualitative Literary Stylistics: How the Method of 'Motifs' Can Help". In *The Grammar of Genres and Styles: From Discrete to Non-discrete Units*, pp.164–193.

Moisl H (2020). "How to visualize high-dimensional data: a roadmap". *Journal of Data Mining and Digital Humanities*, Special Issue on Visualisations in Historical Linguistics, pp.1–19

Seminck O, Gambette P, Legallois D, Poibeau T (2021). "The Corpus for Idiolectal Research (CIDRE)". *The Journal of Open Humanities Data (JOHD)*, **7**, 15.