"*Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis*"

Paul Meyer's work on the composition of French legendaries in prose (Meyer, 1906) led him to discover that some of these were formed from a macro structural point of view of successive compilations, showing that some legendaries were more of a compilation of various elements juxtaposed rather than classified thematically. He observed, therefore, that the legendaries of the C_2 family were the result of an agglomeration among the legendaries of the B family, including themselves the legendary A, and new lives. It also detects on a smaller scale that certain sequences of Saint's Lives escape the thematic classification and that sub-series stand out, such as the hagiographic collection of Wauchier de Denain's *Seint Confessor* of the legendary C or the consecutive and recurrent series in the legendaries family B and C of the following three lives: Sainte Sixte, Saint Laurent and Saint Hippolyte. This serial composition of the Lives of Saints is a composition datum also noted by other specialists such as J.P. Perrot (Perrot, 1992) and G. Philippart (Philippart, 1977) who even points out that these hagiographic series, which he supposes written by the same author, must then be studied in their entirety in the same way as a literary work. However, it is still very rare today that hagiographic text editing concerns a complete author's legendary, mostly because of a lacking certainty about these grouping within corpora that were transmitted to us by anonymous people.

From there, in order to try to detect hagiographic series that could come from the same hand, we decided to confront the hypotheses of Paul Meyer which are based on a structural analysis of an exploratory stylometric analysis, on one of the family C legendary, namely the manuscript fr. 412 of the Bibliothèque Nationale de France. However, the task is complex, because the abundance of graphic variants of medieval languages that are attributable to the copyist and not the author can affect the stylometric analysis. Textual variants added to the thread of transmission and the absence of a standardized spelling add to the difficulty (Kestemont, Moens, & Deploige, 2015). While the lemmatization of the texts could have minimized the complexity of the graphic variation, in our case, the preparation of the corpus would have been extremely time-consuming. This is why we decided to work from witness written by a single hand in order to limit the biases that could have been induced by graphic variations linked to the copyist and not to the author.

Corpus	OCR Training	OCR Test	Manual Segmentation	Kraken Segmentation
Lines	8890	987	43,380	45125

Table 1: Line count per corpus.

The analysis of the complete legendary is thus made possible by the use of the software of OCR *kraken* (Kiessling, 2018) that we have trained on about 8890 transcribed lines and a test corpus of 897 lines which results in up to 95.2% success in reading the manuscript (transcribed content were the complete *Vie de Saint Brice*, *Dialogues sur les vertus de saint Martin*, *Vie de saint Martin*, *Vie de saint Nicolas*, *Vie de Saint Giles*, *Vie de saint Marcel de Limoges*). Such results allow us to hope that the stylometric analysis is not corrupted by the margin of error (Franzini et al., 2018).

Word Segmentation	Kraken Character Recognition Score
With	0.9186
Without	0.952

Table 2: Character recognition rate over 987 unknown lines with Kraken

Error count	82	63	59	58	45	43	43	42	41	41	38	37	31	29	26	23	23	22	21	19
Correct	i	u	r	s	o	l	t	n	a	e	.	∅	r	c	∅	m	u	t	n	m
Generated	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	i	i	∅	n	n	i	r	u	i

Table 3: Confusion matrix of the 20 most frequent mistaken tokens by the OCR. ∅ stands for nothing : the character was either not transcribed (Correct row) or inserted by the ocr (Generated row).

We work from the text thus obtained. Because most of the texts of the legendary are anonymous, we follow an unsupervised approach to the analysis of the texts (Camps & Cafiero, 2012), using agglomerative hierarchical clustering with Ward's criterion (Ward, 1963). Because of the errors regarding spaces and segmentation in the OCR, we chose to work with n-grams, and particularly 6-grams, a choice of length guided by the idea of getting a good proxy for the words themselves.

The texts are, in average, quite short, a known difficulty for stylometry (Eder, 2015), with a median value of 16863 characters (space excluded), but with extreme values of 1,364 and 85,378. Texts that are too short create a problem of reliability, as the observed frequencies may not represent accurately the actual probability of a given variable's appearance (Moisl, 2011). To limit this issue, we restricted the analysis to the first 400 most frequent variables, and removed texts below 5,000 characters. The metric and choices of normalization are also an important parameter, one to which much attention has been devoted (Evert et al., 2017; Jannidis, Pielström, Schöch, & Vitt, 2015)

Following the benchmark by Evert et al. 2017, we chose to use Manhattan distance with z-transformation (Burrows' Delta) and vector-length Euclidean normalization. The results are partially presented in illustrations 1 & 2.

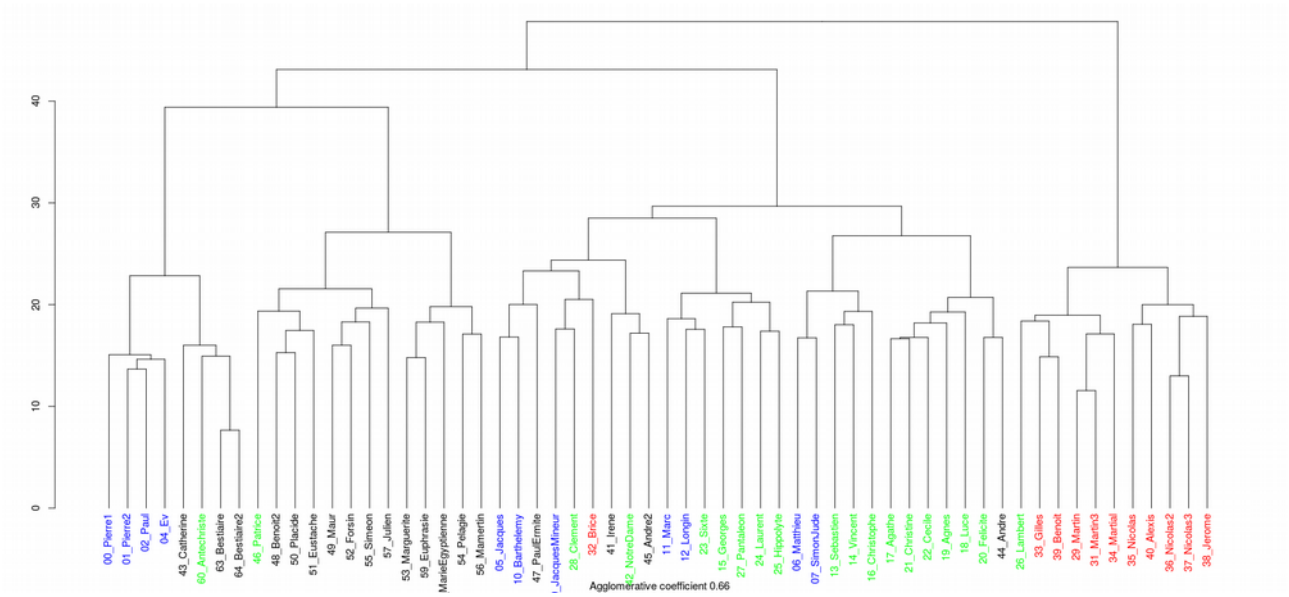

Illustration 1: Dendogram of agglomerative hierarchical clustering using manhattan distance, z-transformation and vector length normalization over 400 most frequent 6-grams

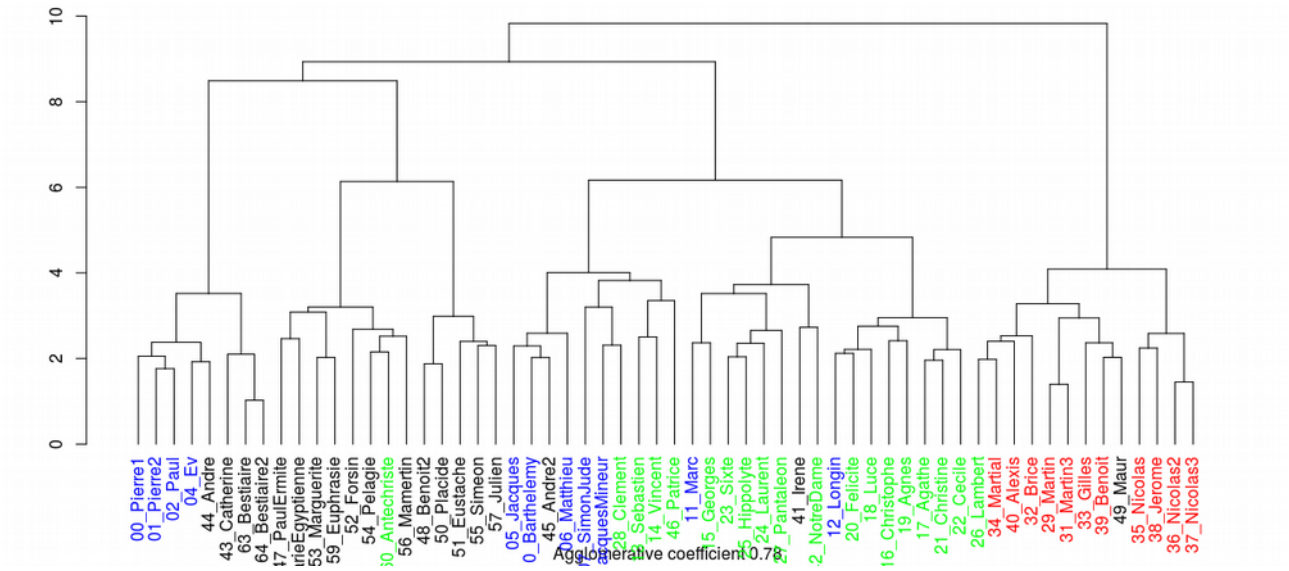

Illustration 2: Dendogram of agglomerative hierarchical clustering using Kohonen SOM coordinates over 400 most frequent 6-grams

Because, at the same time, the corpus is homogeneous, the texts can be quite short, and the data is noisy, separating them in stable clusters can prove quite hard. We try to improve the quality of the signal by applying, first, a Kohonen self-organizing map (Kohonen, 1988, p. 59-69), and then using the coordinates of the points in the SOM for hierarchical clustering (Camps & Cafiero, 2012).

In addition, the specificity of composition of the legendary C by successive additions (lives of the legendary A, then lives of the legendary B, and finally addition of new lives) allows us to ensure a quick control of the likelihood of some proposed groupings. The presence of the hagiographic collection of *Seint Confessors* of Wauchier de Denain where the author identifies itself twice (both in *Dialogues de Sulpice Sévère* and in *Vie de saint Martial de Limoges*) also serves as an indicator of validity.

The study has already shown interesting connections between the legendary of Wauchier de Denain and the *Vie de Saint Lambert de Lièges* and some collections have been revealed. Two of them are quite certain, one of the first five texts of C, all about saint apostles and hypothetically from legendary A, and another one of six virgin saints' lives, all from legendary B including the Lives of saint Agathe, Lucie, Agnès, Felicité, Christine and Cécile. Three others need more analysis to be confirmed.

To conclude, our analysis attempts to evaluate the best parameters for our study and to overcome certain difficulties inherent to our corpus. Indeed, two major obstacles have to be overcome: the lack of graphic standardization and the lack of homogeneity in the separation of words. At the end of this prospective study, we hope to be able to confirm some of Paul Meyer's work hypotheses, but also perhaps to reveal new hagiographic series prior to the composition of the legendaries that were transmitted to us and that could have escaped us so far. Such results would allow us to better understand the mode and the mediums of diffusion of these Saint's Lives who seem, in this serial form, to escape a pure use of liturgy.

Bibliography

Adam, R. (2005). La Vita Landiberti Leodiensis (ca 1144-1145) du chanoine Nicolas de Liège. *Le Moyen Age*, *Tome CXI*(3), 503-528.

Camps, J.-B., & Cafiero, F. (2012). Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature. *Revue Des Nouvelles Technologies de l'Information*, *SHS-1*(MASHS 2011/2012. Modèles et Apprentissages en Sciences Humaines et Sociales Rédacteurs invités : Mar), 55-84.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, *32*(suppl_2), ii4-ii16. https://doi.org/10.1093/llc/fqx023

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., … Rybicki, J. (2018). Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018.00004

Jannidis, F., Pielström, S., Schöch, C., & Vitt, T. (2015). Improving Burrows' Delta – An empirical evaluation of text distance measures. *Digital Humanities Conference*, 11.

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)* (p. 59–66). Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0908

Kestemont, M., Moens, S., & Deploige, J. (2015). Collaborative authorship in the twelfth century: a stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digtial Scholarship in the Humanities*, *30*(2), 199-224. http://dx.doi.org/10.1093/llc/fqt063

Kiessling, B. (2018). *OCR engine for all the languages. Contribute to mittagessen/kraken development by creating an account on GitHub*. Python. Consulté à l'adresse https://github.com/mittagessen/kraken (Original work published 2015)

Kohonen, T. (1988). Neurocomputing: Foundations of Research. In J. A. Anderson & E. Rosenfeld (Éd.) (p. 509–521). Cambridge, MA, USA: MIT Press. Consulté à l'adresse http://dl.acm.org/citation.cfm?id=65669.104428

Meyer, P. (1906). Légendes hagiographiques en français. In *Histoire littéraire de la France* (Imprimerie nationale, Vol. 33, p. 328-458). Paris. Consulté à l'adresse http://archive.org/details/histoirelittra33riveuoft

Moisl, H. (2011). Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora. *Journal of Quantitative Linguistics*, *18*(1), 23-52. https://doi.org/10.1080/09296174.2011.533588

Perrot, J.-P. (1992). *Le passionnaire français au Moyen âge*. Genève, Suisse: Droz.

Philippart, G. (1977). *Les Légendiers latins et autres manuscrits hagiographiques*. Turnhout (Belgique), Belgique: Brépols.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538-556. https://doi.org/10.1002/asi.21001

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845