Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis

Jean-Baptiste Camps, Thibault Clérice, Ariane Pinche

Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis

Jean-Baptiste Camps
École Nationale des Chartes, France

Thibault Clérice
École Nationale des Chartes; Université Lyon 3, France

Ariane Pinche
École Nationale des Chartes; Université Lyon 3, France

Abstract

Stylometric analysis of medieval vernacular texts is still a significant challenge: the importance of scribal variation, be it spelling or more substantial, as well as the variants and errors introduced in the tradition, complicate the task of the would-be stylometrist, by inducing noise and perhaps even interferences in the authorship signal. Basing the analysis on the study of the copy from a single hand of several texts can partially mitigate these issues (Camps and Cafiero, 2013, Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature. Revue Des Nouvelles Technologies de l'information (RNTI), SHS-1, pp. 55–84), but the limited availability of complete diplomatic transcriptions might make this difficult. In this article, we use a workflow combining handwritten text recognition and stylometric analysis, applied to the case of the hagiographic works contained in MS BnF, fr. 412. We seek to evaluate Paul Meyer's hypothesis about the constitution of groups of hagiographic works, as well as to examine potential authorial groupings in a vastly anonymous corpus.

Correspondence:
Thibault Clérice, École Nationale des Chartes, 65, rue de Richelieu, Paris 75002, France.
E-mail:
thibault.clerice@chartes.psl.eu

1. Introduction

1.1 Understanding the French hagiographic tradition

The history of the early French saint's Lives collections in prose is still an enigma. Indeed, at the beginning of the thirteenth century, *legendiers* (i.e. manuscript containing Saint's Lives collection) were already constituted and preliminary steps are missing. In the case where those collections do not adopt the liturgical calendar, they are often built around thematic series (Perrot, 1992, pp. 11–15): apostles, martyrs, confessors, saint virgins, but the organization within those themes is not clear. One of the hypotheses about the composition of hagiographic collections in Latin with similar structure is that they are a compilation of pre-existing *libelli*, independent units about one saint or a series of saints (Philippart, 1977).

Digital Scholarship in the Humanities, Vol. 36, Supplement 2, 2021.
doi:10.1093/llc/fqab033

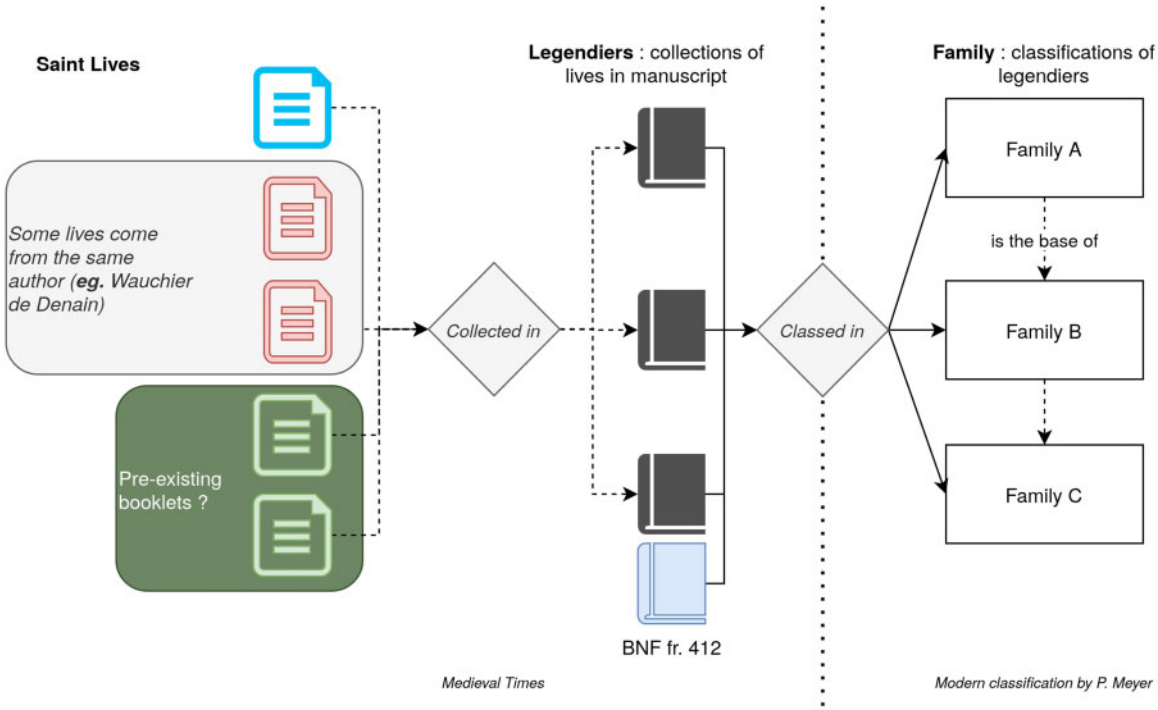

Fig. 1 Representation of the possible composition of French hagiographic collections based on Paul Meyer's hypothesis

Paul Meyer's[1] work on the composition of Old French prose legendiers (Meyer, 1906) led him to discover that some of these came from successive compilations (Fig. 1). Using their macrostructure, he tried to organize the French manuscripts into families on the basis of the thematic similarities, proximity of the groups of Lives and recurrent series in the manuscript tradition. The first three collections named A, B, and C are composed by successive additions. Thereby, collection A is a collection of saint apostles' Lives, collection B adds a collection of saint martyrs' Lives, and collection C is the aggregation of collection A, B, and 22 new texts: saint confessors' Lives, saint virgins' Lives, one text about the antichrist and another about the purgatory. New additions to the collection are not united by a thematic object and seem messier. Studying those compilations, Paul Meyer had also the intuition that the collections possess some smaller pieces. He identified a few series using authorship when he could (e.g. *Li Seint Confessor* of Wauchier de Denain in collection C) and proposed the existence of primitive series based on the repetitive grouping of selected lives in different manuscripts as for instance the series: Saint Sixte, Saint Laurent, and Saint Hippolyte.

Because most of the French saint's lives are anonymous and because the collections were rearranged by multiple editors over time, it is extremely difficult to locate what could have been the primitive series and Meyer could not go further. This serial composition of the Lives of Saints is a datum also noted by other specialists of Latin hagiography such as Perrot (1992) and Philippart (1977), who even points out that these hagiographic series must be studied in their entirety in the same way as a literary work. However, despite these academic positions, to our knowledge, there is no complete edition of hagiographic series, most probably because the identification of the series themselves is a matter of debate, except in the context of a full manuscript edition.

As such, the aim of our article is, first, to determine if Paul Meyer's intuitions and hypotheses can be infirmed, nuanced, or completed. Secondly, we would like to discover if some other links between

saint's lives can reveal series from single anonymous authors and help reconstitute some of the hypothetical pre-existing *libelli*. In order to do so, we performed a stylometric analysis on a manuscript representative of the collection C. To perform this computational approach, and because there is no complete edition of any of the manuscripts holding the collection C, we created a pipeline to acquire the text. After the presentation of the data acquisition pipeline, we will explain how we approach the inherent problems of both Old French and automatic text acquisition variability in our stylometric analysis. Finally, we'll propose an evaluation of the results in regard to the traditional knowledge we have of the manuscript transmissions.

2. Development and Evaluation of a Data Pipeline

For this work, the BnF fr. 412 manuscript, written in a single hand during the thirteenth century, seems to be a valid source for text acquisition. This manuscript is very ornamented, and juxtaposes calendaries, hagiographic collection, and bestiary. The copist dates the manuscript and identifies the illuminator.

> *Icis livres ici finist/Bone aventure ait qui l'escrist/ Henris ot non l'enlumineur/Dex le gardie de deshouneur/Si fu fais l'an* MCCIII^{XX} *et* V (fol. 227v)

The handwriting is a regular gothic *textualis* of rather large modules. The letters are rather tightly packed and the words not very detached, so that it is sometimes difficult to discern cases of agglutination. It is, however, very easy to read and has very few abbreviations. The ink is black and of good quality. It has only lightened in very few places. The manuscript was most probably written in a short time, which does not impact too much writing style (the writer's hand does not 'age'), it is in pristine condition, has been digitized by the Bibliothèque Nationale de France (BNF) and made available on Gallica.[2] To be able to analyze the text, we had to build a pipeline that would, step by step, enrich the data with more information: from pictures to text, from raw text to normalized version, from normalized version to linguistically annotated data so that multiple stylometrical approaches could be combined and evaluated.

2.1 Line detection

Handwritten Text Recognition (HTR) has evolved much over the course of the past years, with easy-to-use tools, such as Transkribus and Kraken. We distinguish two steps of text acquisition: layout detection (and particularly line detection) and the actual text recognition.

As Garz et al. (2012) put it, 'segmenting page images into text lines is a crucial pre-processing step for automated reading of historical documents': unlike printed books from modern editions, parchments present various issues from ink bleed-through (the capacity of a verso writing or picture to be seen on a recto) to inconsistent background color. On top of these traditional issues, costly manuscripts like the BnF fr. 412 accompany texts with illumination, including historiated lettrines, flourished initials and marginal ornamentation, as well as rubrics, underscoring in this way the discontinuity between texts (Fig. 2). Line detection was found to perform very poorly in Kraken compared with Transkribus, two well-known and performant HTR and Handwritten Text Recognition (OCR) engines. Kraken in its 2.0.5 release contains a traditional line segmenter based on contrast which cannot be trained on a specific layout, while Transkribus is using deep-learning models for the same work.[3] While it should be stressed that we cannot offer a methodical, re-applicable evaluation for this performance, we can definitely say that Kraken would often miss lines, create a lot of false positives in ornamentation, and—not often but enough to be seen—incorrectly sort the lines. On the other hand, Transkribus would rarely miss lines, rarely find text in illuminations (although it could happen), but had some time issues with last lines of columns (Fig. 3). We believe in subsequent results Kraken output to be much noisier than Transkribus.

2.2 HTR

Text acquisition was evaluated using both Transkribus (HTR+) and Kraken. Two datasets have been created for these specific reasons:

(1) The main dataset, Pinche Dataset below, is the combination of 271 columns transcribed by A.

Fig. 2 Examples of manuscript layout issues for line segmentation: ink bleed-through and disappearance of text (bottom left), flourished and highlighted initials (second column from the left, both top and bottom), historiated initials (third column), litterae elongatae on first and last lines (last column) from fol.9r and fol.10r

Fig. 3 Line detection in both Transkribus (first and second from the left) and Kraken. White areas are marked as lines. In these examples, we can see the illumination incorrectly identified as text (much more important in Kraken that in Transkribus), the false negative on the last line in Transkribus and various issues in Kraken, in general

Pinche, spanning from folio 103r to folio 170v (Pinche, in progress). It has the advantage of having only one transcriber and has been proofread in the context of an ongoing PhD thesis. It contains ninety-six characters (single spaces included), of which twenty-eight are found fewer than ten times, and in total makes up to around 495,000 occurrences. However, it has the downside of being both consecutive and attributed to a single author (Wauchier de Denain).

(2) The second dataset, below TNAH Dataset, is the combination of forty-three columns transcribed by S. Albouy, C. Andrieux, H. Dartois, M. Frey, O. Jacquot, M. Morillon, M.-C. Schmied, and L. Vieillon, students of A. Pinche in the context of her TEI course (Pinche *et al.*, 2019). Unlike the Pinche Dataset, it is neither consecutive nor attributed to the same legendaries or authors, in fact, all of them are anonymous. They are composed of the *Vie de Saint Philippe* (45ra–45vb), *Vie de Saint Jacques le Mineur* (45vb–46vb), part of *Vie de Saint Longin* (51rb–52vb), *Vie de Sainte Lucie* (71ra–72va), *Vie de Saint Sixte* (87ra–88va), *Vie de Sainte Marguerite* (213ra–214rb), *Vie de Sainte Pélagie* (214rb–215rb), and *Vie de Saint Eufrasie* (224vb–225vb).

The downside of this dataset is that it was mostly transcribed by nonspecialists and despite several attempts to unify it still presents differences in how the text was transcribed. It

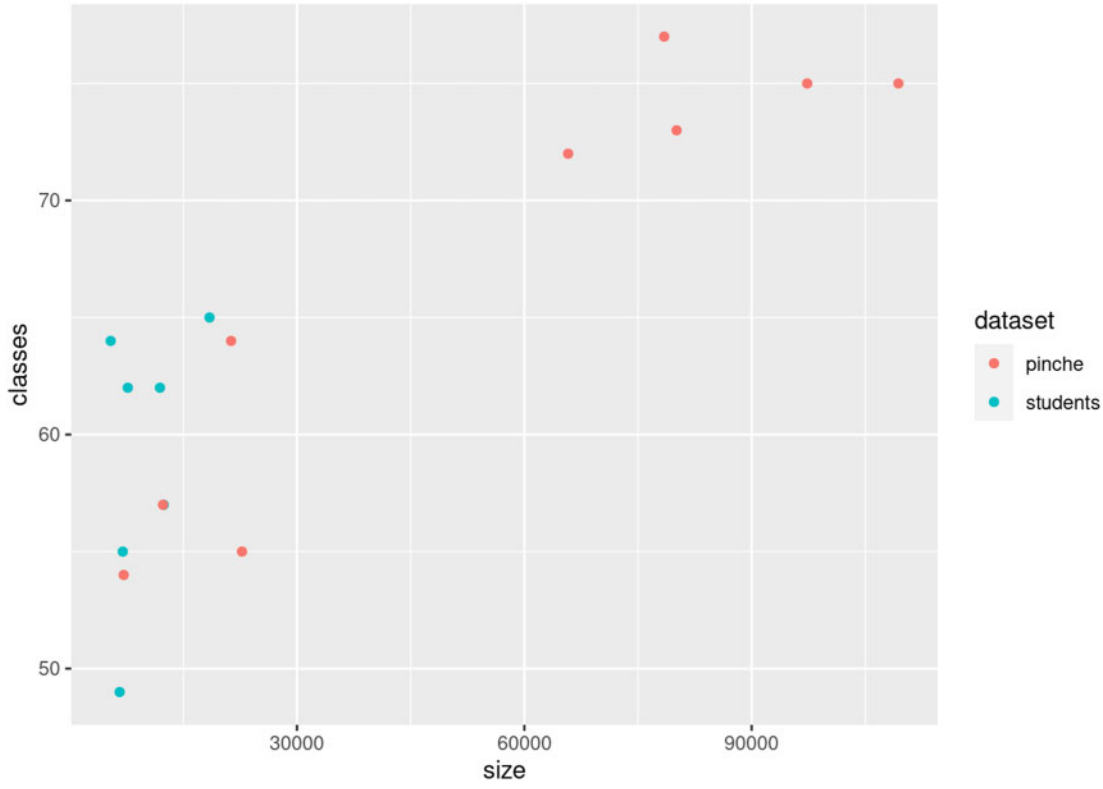

Fig. 4 Distribution of character counts

contains 102 different characters (single spaces included), of which forty-six are found less than ten times, and in total makes up to around 70,000 occurrences. Despite the differences in length, and the limited scope of any comparison for such a limited number of samples, the texts transcribed by the students seem to show a higher variability in the number of unique characters (Fig. 4). For example, for the CON letter, we found five variants in the later: U+A76F (114 occurrences, regular con letter), U + 0039 (twelve, regular nine), U+A770 (seven, modifier con), U+F1A6 (two, 'Latin Abbreviation Sign Spacing Base-line US'), and U + 2079 (two, superscript nine).

We trained three models, each of them tested on the same subset of Pinche dataset. As expected, the training set from Pinche was more efficient (most probably due to its single expert transcriber). However, we found Kraken to be quite impacted by the recognition of spaces. As such, we trained a supplementary second model that would not try to recognize spaces. The Transkribus HTR+ model performed best on the Character Error Rate (Table 1).

Folios 1r–3v were excluded from OCR, because they contain unrelated resources (mostly calendars).

2.3 Word segmentation

As it can be seen in the resulting text (Fig. 5), spaces are one of the least stable features to be correctly recognized. If spacing in handwriting is rarely really regular, Old French manuscripts are the prime examples of it.[4] Indeed, a quick look at one column from the f.10r in Fig. 5 shows that spaces are sometimes really small, sometimes nonexistent. Moreover, there are no marks for hyphenation in the manuscript, which requires us to detect and concatenate some of the tokens passing from one line to another. As an indication, the Kraken model trained with spaces had 905 errors related to

Table 1. Character error rate based on software and corpora

	Training set	Kraken	Kraken (No Space)	Transkribus (HTR +)
Test Set PINCHE	PINCHE	4.87	3.37	**2.29**
Test Set TNAH	PINCHE	31.16	27.16	**8.07**
Test Set PINCHE	TNAH	N/A	N/A	**4.97**

Training set PINCHE and test set PINCHE are split from the dataset PINCHE. In bold, the best scores.

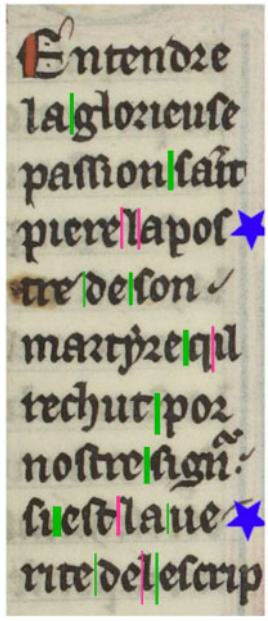

Fig. 5 Part of f.10ra. Green lines are measurable space, pink are spaces that are not written, stars represent hyphenation. Transcription with space width in square brackets (for reference, first 'a' width is 31px wide), ± 2px: 'Entendre la[4]glorieuse passion[6]saint pierre[0]l[0]apos tre[2]de[4]son martyre[6]q[0]il rechut[6]por nostre[6]sigñ si[9]est[0]la[2]ve rite[2]de[0]l[3]escrip'. Largest space in this area represents 35% of the reference a, while smallest measures 6.45% of it. A space is the area available between the furthest right pixel of the left letter and the leftmost one of the next

spaces from which 810 were deletions and insertions: it represents a 1.82-point drop of performance in CER and an impressive 37.39% of the test set errors.

An option is to treat the notion of space as a natural language processing task, where the image is not taken into account. Of course, the notion of words and grammar has evolved and what most of the other tools of the pipeline expect are words perceived as such by modern and contemporaneous medievalists. Unfortunately, due to the extreme variation in spelling of Old French, dictionary approaches do not perform well. In a previous study (Clérice, 2019), we have shown that they do not extend to new unknown domains as much as deep learning models. In this context, we used Boudams, the tool developed for the aforementioned article. It currently removes all spaces before reinserting new ones. We used the Old French model built for that study, which had a 0.99 FScore on the in-domain test set (which contained resources from the Pinche Dataset and TNAH Dataset) while having a 0.945 FScore on an out of domain dataset. Of course, the resulting output is not expected to be perfect (Table 2), and in fact, each step we pursue might introduce new errors as they were not manually transcribed or corrected (Tables 3–5). However, we did keep the output of each step for later stylometric analyses.

2.4 Abbreviation resolution, normalization, and lemmatization

With word segmentation available, there were two others forms of the dataset that were needed: one where each word would be normalized and have its abbreviations resolved, a second where each word would be tagged with both its Part-Of-Speech and its lemma. We actually treated normalization and abbreviation resolution as a lemmatization task, as they both require understanding of phenomenon such as prefix and suffix and replace them with a neutral value.

As such, we trained Pie (Manjavacas *et al.*, 2019a,b) on a corpus of Old French transcriptions available in TEI.[5] The training set was composed of around 125,000 tokens (including punctuation), the evaluation set 16,000, and the test set 15,000 taken from both Pinche and Oriflamms project (Stutzmann *et al.*,

Table 2. Example of results and ground truth after and before word segmentation

Transkribus	Entendre la glorieuse passion saït aun piere lapos stre de son. " martẏre qil rechut por nostre sign: si est la ue rite del escrip ture
Transkribus + Boudams	Entendre la glorieuse passions aït a un piere la posstre de son ." martẏe qil rechut por nostre sign: si est la uerite del escripture
Kraken (No Space) + Boudams	Entendre la glorieuse passions at piere la positre de son. mantyǨre qil rechut por nostre sign: si est la uerite del escriptire
Correct	Entendre la glorieuse passion saït piere l apostre de son martẏre q il rechut por nostre sign∼ si est la uerite de l escripture

Table 3. Output of abbreviations resolutions and normalizations in Table 2 content with ground truth

Transkribus Raw	entendre la glorieuse passion **saint** aun piere lapos stre de son. martyre q'il rechut por nostre sign**or**: si est la ve rite del escrip ture
Transkribus + Boudams	entendre la glorieuse passions **aint** a un piere la posstre de son. martyre q'il rechut por nostre sign**or**: si est la verité del scripture
Kraken (No Space) + Boudams	entendre la glorieuse passions **art** piere la positre de son. mantyere q'il rechut por nostre sign**or**: si est la verité del escriptire
Correct	Entendre la glorieuse passion **sain**t piere l apostre de son martẏre q il rechut por nostre sign**eur** si est la uerite de l escripture

In bold, the same word "saint" in the ground truth and its different versions in the various generated datasets.

Table 4. Test scores of Pie over Old French

		Accuracy	Precision	Recall	Support
Lemma	All	96.38	71.23	70.89	48,317
	Ambiguous Tokens	96.65	75.85	76.43	27,844
	Unknown lemma	72.9	26.85	26.03	1,236
	Unknown form	64.29	42.9	42.49	1,792
POS	All	96.13	78.28	75.72	48,317
	Ambiguous Tokens	95.49	78.88	75.3	32,232
	Unknown form	86.77	59.59	59.18	1,792

Table 5. Automatic tagging of the OCR text by Pie and ground truth

Transkribus Raw	entendre<VERinf> le<DETdef> glorïos<ADJqua> passïon<NOMcom> amer1<VERcjg> a3<PRE> un<DETndf> piere<NOMcom> le<DETdef> postre<NOMcom> de<PRE> son4<DETpos> .<PONfrt> martire2<NOMcom> que2<CONsub> '<PONfbl> il<PROper> recevoir<VERcjg> por2<PRE> nostre<DETpos> seignor<NOMcom> :<PONfbl> si<ADVgen> estre1<VERcjg> le<DETdef> verité<NOMcom> de+le<PRE.DETdef> escriture<NOMcom>
Correct	entendre<VERinf> le<DETdef> glorïos<ADJqua> passïon<NOMcom> saint<ADJqua> Pierre<NOMpro> le<DETdef> apostle<NOMcom> de<PRE> son4<DETpos> martire2<NOMcom> que2<PROrel> il<PROper> recevoir<VERcjg> por2<PRE> nostre<DETpos> seignor<NOMcom> si<ADVgen> estre1<VERcjg> le<DETdef> verité<NOMcom> de<PRE> le<DETdef> escriture<NOMcom>

2013). They contained abbreviation resolution, accentuation, and punctuation introduction (sen -> s'en). The results were promising with 96.86% accuracy, with 96.96% on ambiguous tokens (whose input can be normalized in different fashions), 91.42% on unknown output form, and finally 90.72% on unknown origin form.

To improve statistical calculations based on occurrence counts, we applied lemmatization. Unlike modern English, Old French is both defined by its spelling variation (not only between regional *scriptae* but also inside them), and its rich morphology. As such, the same word with different flexions can be written in

different fashions. In the Pinche Dataset, which represents 27.34% of the whole corpus to be lemmatized in Transkribus,[6] the verb *avoir* (to have) has fifty-seven different spellings, the pronoun *il* seventeen, the nouns *emperëor* eight, the adverb *tout* (all) fourteen, the adjective *saint* eleven: for example, '*compagnie*' can be found written as *compagnie*, *compaignie*, *compaignies*, *compaigniez*, *conpagnie*, *conpaignie*, and *conpaignies*.

Pie is a lemmatizer specifically designed to deal with historical languages with such traits as those found in Old French. We trained a lemmatizer on a dataset of approximately 500,000 lemmatized tokens which were taken from the Chrestien corpus (Kunstmann, 2009), the Geste corpus (Camps, 2019), the Institutes (Olivier-Martin *et al.*, 2018), the Lancelot (Ing, in progress) and the Wauchier (Pinche, in progress) dataset.[7] The overall model had 96.38% accuracy on the test corpus comprised of 48,317 tokens, punctuation included (Clérice *et al.*, 2019).

The final result is a lemmatization and pos-tagging of each document. Error accumulation through successive postprocessing steps, and noise in the source HTR dataset leads to a dataset with varying quality, although some parts of the document, if not most, are treated with satisfying results.

To evaluate the impact of all pipeline steps on lemmas and POS 3-grams frequencies, in a case where the total number of words can differ, we evaluate the differences with the ground truth in the following equation:

$$\Delta_{A,B} = \frac{\sum_{i=1}^{n} tf(A_i) - tf(B_i)}{\sum_{i=1}^{n} tf(A_i)}$$

where $tf(A_i)$ is the absolute term frequency of feature i in document A to be evaluated, and $tf(B_i)$ it's frequency in document B, the ground truth.

We also provide the ratio of lemma or POS 3-grams, which are present in A but not in B and vice versa (labeled as difference in Table 6). Difference of OCR against Gold is higher than its counterpart as a result of noise accumulation in the pipeline (it has more token).

Given the previous results, we kept only the Transkribus HTR + model output and its variations (through Boudams; through Pie for lemmatization and POS-tagging). Each figure states specifically which version of the Transkribus pipeline output it uses.

3. Stylometric Analysis

The stylometric analysis has to address several challenges, resulting both from the nature of the texts and from the data acquisition pipeline: the short length

Table 6. Comparison of the Gold corpus from Pinche dataset and the HTR results at the end of the pipeline (HTR/Boudams/Pie/Pie)

	Lemmas					POS 3-gram			
	Accuracy delta			Difference		Accuracy delta		Difference	
Corpus	All	Function	Moisl	OCR	Gold	All	Moisl	OCR	Gold
Martin (29, 30)	33.35	10.69	11.43	34.01	16.92	44.8	32.28	29.59	24.23
Dialogues (31)	29.38	9.77	9.99	33.01	19.69	48.38	35.85	29.84	23.49
Brice (32)	39.49	12.14	16.56	29.53	21.39	66.09	47.51	30.95	30.86
Gilles (33)	32.24	9.83	11.07	25.42	16.91	46.38	34.03	25.56	23.64
Martial (34)	28.26	7.92	9.68	37.36	21.63	50.29	39.09	30.09	24.83
Nicolas (35, 36, 37)	29.44	9.33	10.02	38.76	21.8	47.19	35.42	31.42	24.27
Jerome (38)	34.13	12.59	14.38	19.53	15.12	61.92	52.07	28.66	28.06
Benoit (39)	27.97	9.64	11.93	30.98	17.74	52.88	44.09	30.88	24.22
Alexis (40)	30.19	10.65	11.58	21.83	13.17	57.71	47.77	30.16	26.73
Total	27.76	9.02	9.76	51.57	29.18	43.46	34.90	32.65	21.81

We provide deltas for all values and for function lemmas as well as for values selected according to Moisl's procedure. Difference is the relative accumulation of frequencies of lemmas or POS 3-gram that are found in one version of the corpus but not the other.

and anonymity of most texts; the noise in the authorial signal caused by successive errors or innovations in the tradition of the texts (variants) as well as the amount of spelling variation; the noise (and potential biases) resulting from the data acquisition pipeline. Even though stylometric methods have shown to be relatively resilient to a—simulated or observed—moderate amount of noise (Eder, 2013; Franzini *et al.*, 2018), devising a stylometric set-up to partially eliminate or circumvent it is still likely to lead to more reliable results.

3.1 Unsupervised analysis of short anonymous texts

The texts from the manuscript are, on average, quite short, with a median value of 3,539 words, and extreme values of 298 and 18,971 (Fig. 6). Texts that are too short create a problem of reliability, as the observed frequencies may not accurately represent the actual probability of a given variable's appearance (Moisl, 2011). To limit this issue, we removed texts below 1,000 words, a relatively low limit when compared with existing benchmarks (Eder, 2015, 2017), but motivated by the necessity to not exclude too many texts.

Given the short length of the texts and the sparsity caused by noise, we implement a procedure to select for analysis only those features that satisfy a criterion of statistical reliability. In this, we follow the procedure suggested by Moisl (2011), in the implementation already used by Cafiero and Camps (2019). To summarize it, features are only retained if they match the desired confidence level and margin of error even for the smallest text in the corpus. For each feature (e.g. the function word 'et'), the minimum text size n is calculated with

$$n = \overline{p}(1-\overline{p})\left(\frac{z}{e}\right)^2$$

where $p-$ is the mean probability of the feature in our corpus; z, the confidence level, and e, the margin of error. We take $z = 1.645$ to obtain a confidence margin of 90%, and $e = 2\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the feature's standard deviation. Beforehand, to correct for normality, we generate a mirror-variable (Moisl, 2011):

$$vmirror_{ji} = (max_v + min_v) - v_{ji}$$

where v_j is the vector of the feature j, $\max_v$ and $\min_v$ are the maximum and minimum values in v_j, and v_{ji} is

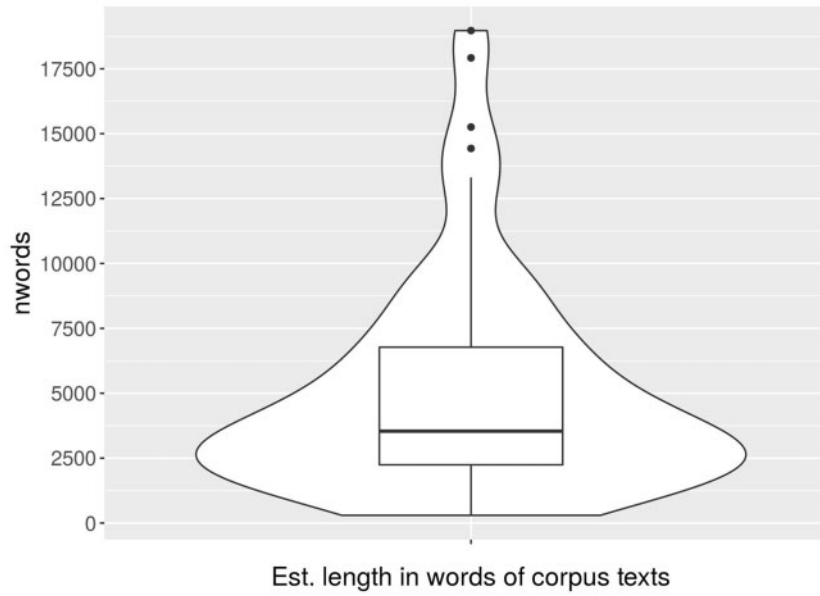

Fig. 6 Boxplot of text lengths in words, based on the Transkribus + Boudams data

the relative frequency of *j* in a sample *i*. This mirror-variable is concatenated with the original variable in order to compute *n*. If *n* is superior to the length of the smallest text in our corpus, we then exclude the feature from further analysis.

Because most of the texts of the manuscript are anonymous, we follow an unsupervised approach to their analysis (Camps and Cafiero, 2013; Cafiero and Camps, 2019), using agglomerative hierarchical clustering with Ward's criterion (Ward, 1963), guided by its ability to form coherent clusters.

The metric and choices of normalization are also an important parameter, one to which much attention has been devoted (Evert *et al.*, 2017; Jannidis *et al.*, 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.

3.2 Noise reduction and choice of features

In the form in which they have reached us, medieval texts are noisy, in respect to the authorial signal. The perturbations in the authorial signal can be inherent to the data, as is the case with the successive errors and modifications made by generations of scribes in the successive copies of the works. Such is the case for substantive variants, but it also affects the linguistic form of the texts itself: stratified, it can contain spellings and other linguistic features originating from the dialect and regional scripta of any and all of the successive scribes, creating a very important and heterogeneous spelling variation. The choice of working with the texts of a single manuscript was already guided by the aim of limiting this kind of noise, but is not, in itself sufficient. For this reason, further normalizations, such as abbreviation expansion and lemmatization, were included in the data acquisition pipeline. Yet, even though it achieves satisfying accuracy at each step, the pipeline itself, through the residual presence of errors, introduces noise as well. Moreover, as the training corpora for each algorithm were not selected by a perfectly random process, they introduce the risk of potential biases.

To handle these risks, we chose to retain raw as well as normalized data for the analyses, using three feature sets:

(1) Character: *n*-grams from raw HTR data (baseline).
(2) Functors: pseudo-affixes from expanded data, function words and POS *n*-grams.
(3) Words: word forms from expanded data and lemmas.

The aim of feature set 1 is to avoid biases resulting from the pipeline, and for this reason to use the initial raw output of the Transkribus HTR model, excluding all further normalization steps. Previous research has shown that character *n*-grams could be a way to circumvent issues due to noisy OCR output, especially when compared with most frequent words (Eder, 2013). Following existing benchmarks (Stamatatos, 2013), we choose $n = 3$ for our character *n*-grams. Because it fits our case closely, we consider this feature set to be our baseline, and complement it with two others.

Feature set 2 is built to capture *functors*, that is, grammatical morphemes (Kestemont, 2014), while circumventing the noise due to scribal variation of paleographic and graphematic nature. Functors have long been—and often still are—considered the most effective feature for authorship attribution, because they capture unconscious individual variation, while being less dependent on generic or thematic context. In this feature set, we used expanded data to extract pseudo-affixes, that is, a specific kind of *n*-gram that has been shown, along with punctuation *n*-grams, to outperform others (Sapkota *et al.*, 2015), perhaps because of its ability to capture grammatical morphemes. Since there is no authorial punctuation in our case, we extracted four kinds of pseudo-affixes *n*-grams: 'prefix' and 'suffix' (the *n* first or last characters of words of at least $n + 1$ characters), as well as 'space-prefix' and 'space-suffix' (the interword space with the *n*-1 characters preceding or following it), with $n = 3$. For instance, for '*annoncier*', we extracted '^ann', 'ier$', 'an', and 'er_'. We also included function-words. Function words are commonly recognized as one of the most effective features (if not the most) for authorship attribution (Argamon and Levitan, 2005; Koppel *et al.*, 2009; Kestemont, 2014). Finally, we added information on the morpho-syntax

of the texts, by extracting Part-of-Speech 3-grams such as 'PRE DETdef NOMcom' (preposition, definite article and noun, e.g. '*a la corone*'). POS 3-grams have sometimes shown to be a quite effective feature for cross-topic authorship attribution (Gómez-Adorno *et al.*, 2018). In this case, multiplying the measurements by concatenating three types of features in this set is done to help deal with short noisy texts and improve reliability.

Feature set 3 is constituted because—despite the broad consensus on the use of functors—some recent studies seem to advocate the use of longer word lists as a feature for authorship attribution (Evert *et al.*, 2017). Using words' forms is, in our case, both interesting, because it allows us to retain morphological information, and risky, due to the extent of spelling variation, attributable to the scribes. To account for that, we also include lemmatized words, which, in turn, are dependent upon the accuracy of the lemmatizer.

3.3 Results and cross-validation

The results on the three feature sets are included in Fig. 7, HC1. Our baseline result (Fig. 7, HC1, top) is also the one closest to Meyer's classification, often up to the ordering of the texts, though displaying a few differences (six out of fifty-nine texts, concerning mostly texts of B included with C). The results on feature sets 2 and 3, though keeping the same macrostructure, display some interesting variations with the inclusion of a mixed B/C subgroup within Meyer's A.

In order to get more insight into feature sets 2 and 3, we also give supplementary results on their components (Fig. 8, HC2). This can be useful since differences on clusterings based on separate aspects (e.g. morphosyntactic sequences versus function words or affixes) could reflect differences in groupings when alternative perspectives are taken on the language; or punctually yield useful information on some texts, as we vary the lens with which we observe it.

In order to check the robustness of our results, we give, for each analysis shown in Fig. 7, HC1 and Fig. 8, HC2, four indicators:

- The number of analyzed features and the agglomerative coefficient, that, taken together, give an indication of the quality of the clustering;
- The cluster purity of the groups (with $k = 5$) as compared with Meyer's Hypothesis on A, B, and C, and Wauchier's alleged texts;
- The cluster purity of the groups when compared with our baseline (results on feature set 1).

These figures are given in Table 7.

3.4 Classification resistant texts and volatility

Eder (2017) showed that, whatever the variation of sample length, some texts were never correctly attributed (at least when a given feature set is used) and suggested to measure the diversity of attributions of individual texts—what we call volatility—to help identify these cases when the authors are not known. Following this, one could hypothesize that the presence of classification resistant (or volatile) texts is to be expected in a sufficiently large corpus.

To measure the volatility of any individual text, in the context of unsupervised analysis, we wish to measure the stability or volatility of its neighborhood. We devise a specific metric V_i that aims to compute the volatility of the neighborhood of a specific text i in the groups of which he is a member in all the clusterings performed. Let $(G_j)_{j \in J}$, $i \in G_j$ be the family of sets of which i is a member, where J is the total number of clusterings performed. We can then construct a set X, containing all unique texts $\{x_a \dots x_n\}$ occurring in at least one set of the family (G_j), $X = \{x \mid \exists\, j \in J,\ x \in G_j\}$. For each x, the family (G_j) can be split in two subfamilies, $(A_k)_{k \in J,\ x \in A_k}$ and $(B_l)_{l \in J,\ x \notin B_l}$. We then compute a global volatility index as follows:

$$V_i = \frac{\sum_{a=1}^{n} \frac{Card\left((A_k)_{k \in J,\ x_a \in A_k}\right) - Card\left((B_l)_{l \in J,\ x_a \notin B_l}\right)}{Card\left((G_j)_{j \in J}\right)}}{Card(X)}$$

Since we normalize it by the total number of elements in X, this index is limited by $[-1; 1]$, where -1 would indicate a perfect volatility (all sets with no member in common) and 1 perfect stability (all sets with the same members).

The results of this procedure are given in Table 8. We notice that the texts attributed to Wauchier are the

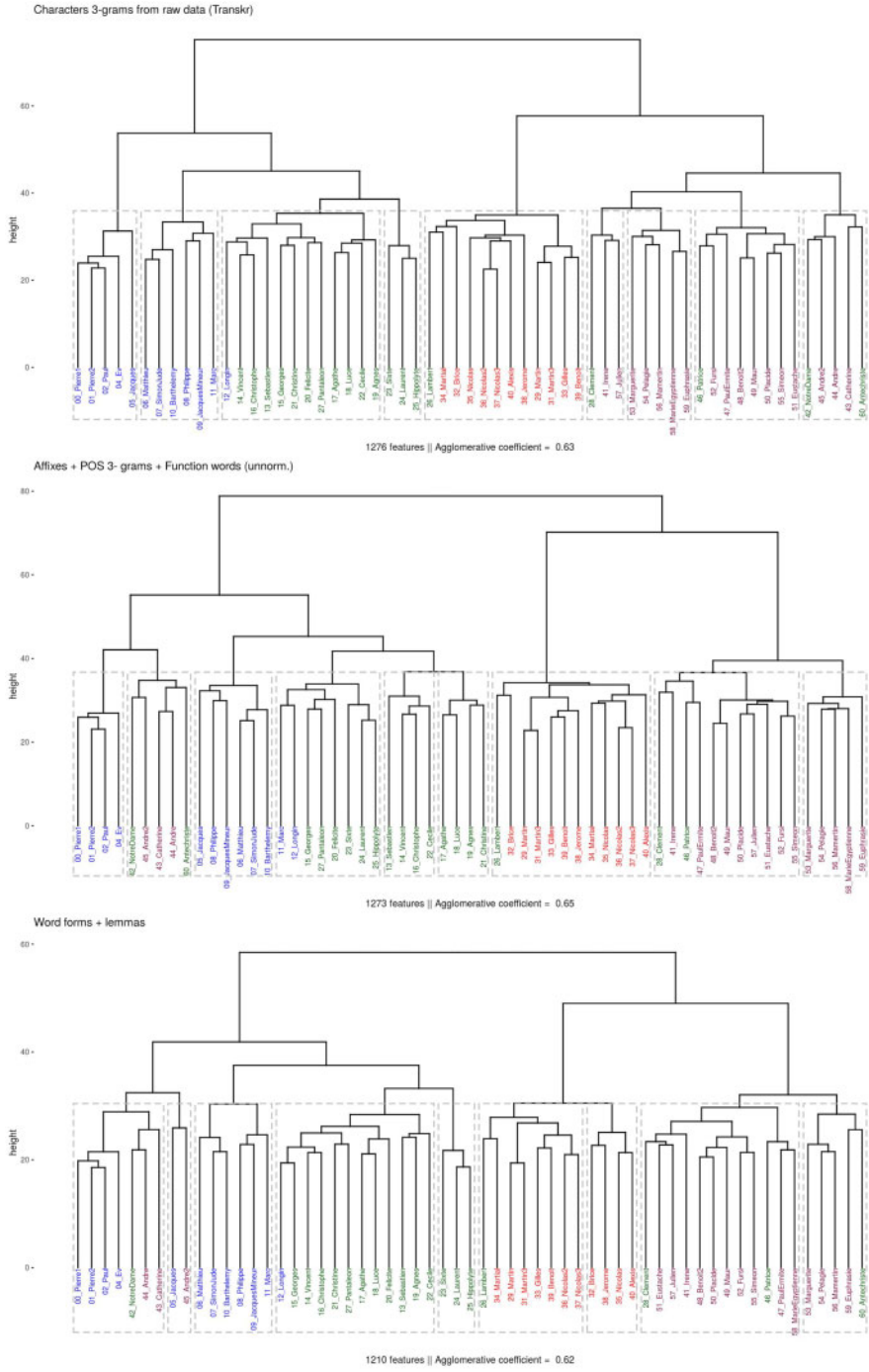

Fig. 7 HC1: Main results from agglomerative hierarchical clustering (Ward's method, Manhattan distance with *z*-scores and vector-length Euclidean normalization) on the three feature sets (top, character *n*-grams from HTR data; middle, functors; bottom, word forms and lemmas)

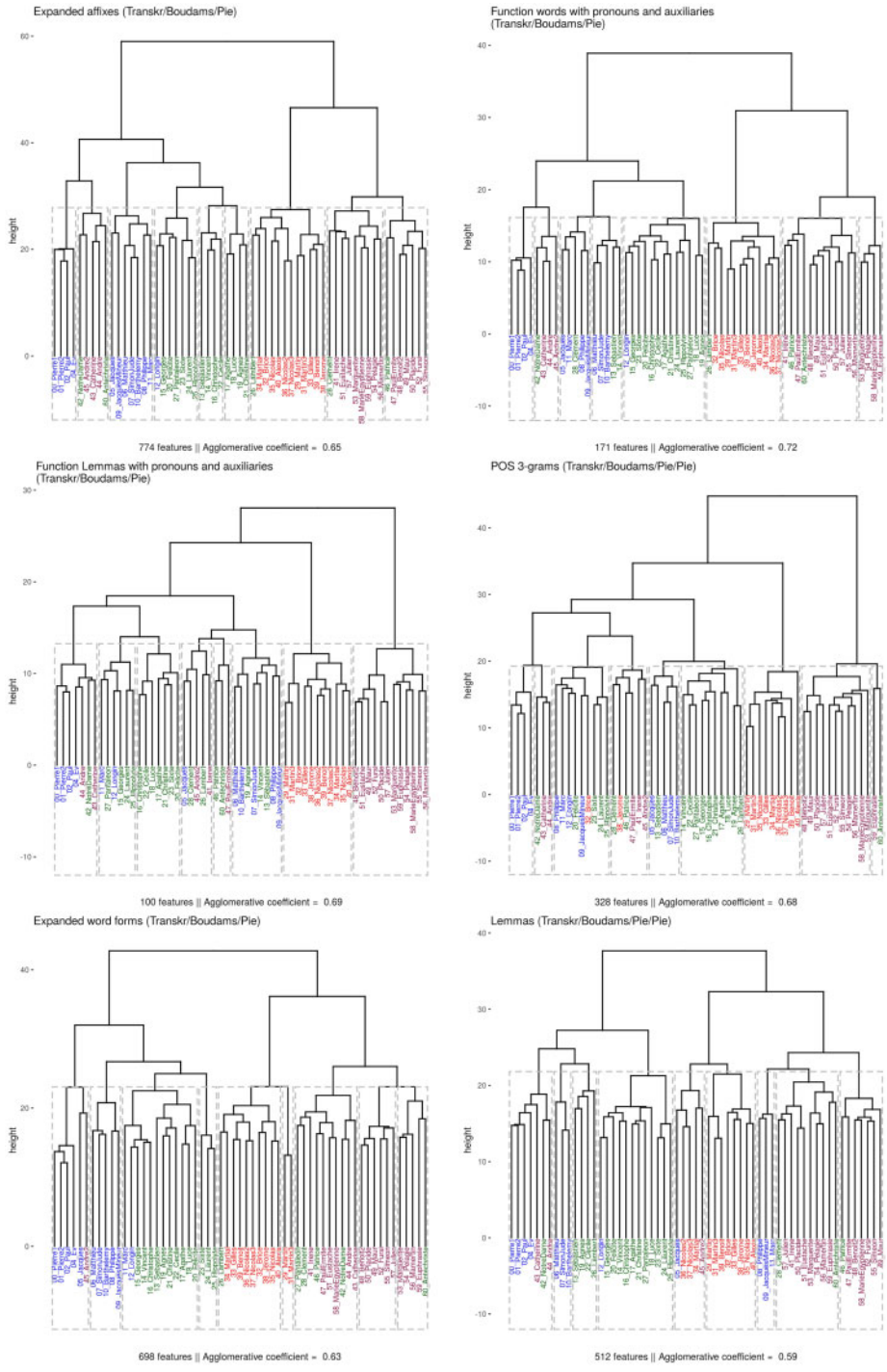

Fig. 8 HC2: Supplementary results from agglomerative hierarchical clustering (identical setup) on the components of feature set 2

Table 7. Number of analyzed features (*N*), agglomerative coefficient (*AC*), cluster purity of the clusters ($k=5$) when compared with Meyer's hypothesis and Wauchier's known texts (*CP* Meyer), cluster purity when compared with the baseline

	N	*AC*	*CP Meyer*	*CP Ref*
Feature set 1				
Raw 3-grams	1276	0.63	0.90	1
Feature set 2				
Affixes	774	0.65	0.85	0.92
Function Words	171	0.72	0.81	0.86
Function Lemmas	100	0.69	0.76	0.76
POS3gr	328	0.68	0.71	0.69
Function words + POS 3-grams + Affixes	1273	0.65	0.83	0.90
Feature set 3				
Forms	698	0.63	0.85	0.95
Lemmas	512	0.59	0.75	0.80
Words + Lemmas	1210	0.62	0.85	0.93

least volatile, while there is a small group of volatile texts achieving a score <0.5.

Another indication yielded by this index is that volatility is not (or almost not) due to variation in sample length (Fig. 9). The small relationship, on the edge of significance, between text length and volatility that we observe when looking only at the supplementary analyses disappears totally when we look at the reference analyses. This could be an indication that the strategy we have adopted, of concatenating several measurements to increase reliability on short texts, is working.

3.5 Controlling for pipeline bias

To control for the presence of bias due to the training data used by the pipeline, we perform the same set of analyses on the data obtained with models trained on alternate data (TNAH corpus) or with different tools (Kraken), and compare it with our analyses displayed above. The results are displayed in Table 9. Even though the models achieve quite different accuracies, the results are not significantly different and show a particular stability on the three main analyses (mean CP of 0.95 and 0.92, Table 9). The results based on a change of training corpus are actually closer than the one obtained with the same corpus but a different HTR software though the difference remains small.

An inspection of the features most correlated to the nine clusters of each of the three reference analyses (see Supplementary Material, Table A and Fig. A–C) shows a wide range of features, of different nature, that is, variation at the graphematical, morphological, or syntactical level. It also shows that, in the case of presumably less grammatical and more thematic features such as words or lemmas, thematic interference can come into play. The variation in the use of these features can sometimes be attributed to diachronic or diatopic variation, while in other occasions, they seem to be characteristic of a given idiolect, such as Wauchier's. For instance, the trigrams 'que' ('that') may be the mark of a more common subordination in C and the evidence of a more recent syntax in these texts, while the trigram 'qil' ('that he') in Wauchier's texts is certainly the sign of a recurrent duplication of subordination when there is an imbricated subordination, a feature of his writing style. The use of 'com' ('as') is also characteristic of Wauchier, while the use of the personal pronoun 'tu' ('you') shows a more contrasted situation. Syntactic sequences such as CONcoo ADVgen VERcjg display an evolution that could be perceived as chronological. Finally, we also have more problematic connections with thematic words, such as «apostle» in collection A (see Appendices).

4. Interpretation of the Results

The manuscript fr. 412 allows us to control the results of our approaches by checking the unity of the Wauchier de Denain collection[8] in the stylometric trees. On the three reference analyses (Fig. 7, HC1), the Wauchier group, with the adjunction of the *Life of saint Lambert*, is the most clearly distinguished group. This same

Table 8. For each text, its number of words, and volatility index (V) based on the three reference analyses (V Ref) or the supplementary analyses (V Suppl)

Texts	No. Words	V Ref	V Suppl
11_Ano_Leg-A_Ap_NA_Vie_Marc	1,820	−0.11	−0.37
05_Ano_Leg-A_Ap_NA_Vie_Jacques	17,920	−0.05	−0.43
42_Ano_Leg-B_Vi_NA_Ass_NotreDame	3,119	0	−0.24
43_Ano_Leg-C_Vi_NA_Vie_Catherine	8,877	0	−0.24
44_Ano_Leg-C_Ap_NA_Vie_Andre	3,118	0	−0.24
45_Ano_Leg-C_Ap_NA_Pas_Andre2	13,315	0	−0.44
60_Ano_Leg-B_NA_NA_NA_Antechriste	1,485	0.25	−0.14
00_Ano_Leg-A_Ap_Ev_Dis_Pierre1	6,774	0.53	0.47
01_Ano_Leg-A_Ap_NA_Vie_Pierre2	5,527	0.53	0.47
02_Ano_Leg-A_Ap_NA_Pas_Paul	4,798	0.53	0.47
04_Ano_Leg-A_Ap_NA_Vie_Jean_Ev	4,955	0.53	0.47
10_Ano_Leg-A_Ap_NA_Vie_Barthelemy	4,360	0.71	−0.17
28_Ano_Leg-B_Ma_Ho_Vie_Clement	2,544	0.71	−0.18
41_Ano_Leg-C_Vi_NA_Vie_Irene	3,145	0.71	−0.13
46_Ano_Leg-B_Co_NA_Pur_Patrice	7,872	0.71	−0.13
47_Ano_Leg-C_Co_er_Vie_PaulErmite	3,753	0.71	−0.13
48_Ano_Leg-C_Co_ev_Tra_Benoit2	3,234	0.71	0.36
49_Ano_Leg-C_NA_NA_Vie_Maur	6,310	0.71	0.36
50_Ano_Leg-C_NA_NA_Vie_Placide	2,783	0.71	0.36
51_Ano_Leg-C_Ma_ho_Vie_Eustache	3,099	0.71	0.36
52_Ano_Leg-C_Co_NA_Vie_Fursi	2,492	0.71	0.36
53_Ano_Leg-C_Vi_NA_Vie_Marguerite	1,935	0.71	0.36
54_Ano_Leg-C_Vi_NA_Vie_Pelagie	1,506	0.71	0.36
55_Ano_Leg-C_Co_NA_Vie_Simeon	2,894	0.71	0.36
56_Ano_Leg-C_Co_NA_Vie_Mamertin	2,202	0.71	0.36
57_Ano_Leg-C_Vi_NA_Vie_Julien	2,766	0.71	0.36
58_Ano_Leg-C_Vi_NA_Vie_MarieEgyptienne	5,529	0.71	0.36
59_Ano_Leg-C_Vi_NA_Vie_Euphrasie	1,293	0.71	0.36
06_Ano_Leg-A_Ap_NA_Vie_Matthieu	6,447	0.71	−0.17
07_Ano_Leg-A_Ap_NA_Vie_SimonJude	6,784	0.71	−0.17
08_Ano_Leg-A_Ap_NA_Vie_Philippe	1,014	0.71	−0.32
09_Ano_Leg-A_Ap_NA_Vie_JacquesMineur	1,356	0.71	−0.32
12_Ano_Leg-A_Ma_Ho_Vie_Longin	2,244	0.92	0.11
13_Ano_Leg-B_Ma_Ho_Vie_Sebastien	3,539	0.92	−0.1
14_Ano_Leg-B_Ma_Ho_Vie_Vincent	4,838	0.92	−0.05
15_Ano_Leg-B_Ma_Ho_Vie_Georges	4,548	0.92	0.32
16_Ano_Leg-B_Ma_Ho_Vie_Christophe	9,122	0.92	0.32
17_Ano_Leg-B_Ma_Fe_Vie_Agathe	3,109	0.92	0.32
18_Ano_Leg-B_Ma_Fe_Vie_Luce	2,366	0.92	0.32
19_Ano_Leg-B_Ma_Fe_Vie_Agnes	4,177	0.92	−0.07
20_Ano_Leg-B_Ma_Fe_Vie_Felicite	1,676	0.92	0.11
21_Ano_Leg-B_Ma_Fe_Vie_Christine	7,481	0.92	0.32
22_Ano_Leg-B_Ma_Fe_Vie_Cecile	6,782	0.92	0.24
23_Ano_Leg-B_Ma_Ho_Vie_Sixte	1,894	0.92	0.11
24_Ano_Leg-B_Ma_Ho_Vie_Laurent	3,243	0.92	0.11
25_Ano_Leg-B_Ma_Ho_Vie_Hippolyte	2,513	0.92	0.11
27_Ano_Leg-B_Ma_Ho_Vie_Pantaleon	6,565	0.92	−0.28
26_Ano_Leg-B_Ma_Ev_Vie_Lambert	5,247	1	−0.27
29_Wau_Leg-C_Co_Ev_Vie_Martin	14,432	1	0.64
31_Wau_Leg-C_Co_Ev_Dia_Martin3	18,971	1	0.64
32_Wau_Leg-C_Co_Ev_Vie_Brice	1,385	1	−0.04

(*Continued*)

Table 8 (continued)

Texts	No. Words	V Ref	V Suppl
33_Wau_Leg-C_Co_Er_Vie_Gilles	4,415	1	0.64
34_Wau_Leg-C_Co_Ev_Vie_Martial	15,255	1	0.64
35_Wau_Leg-C_Co_Ev_Vie_Nicolas	1,960	1	0.64
36_Wau_Leg-C_Co_Ev_Mir_Nicolas2	10,473	1	0.64
37_Wau_Leg-C_Co_Ev_Tra_Nicolas3	8,379	1	0.64
38_Wau_Leg-C_Co_Ev_Vie_Jerome	2,425	1	−0.04
39_Wau_Leg-C_Co_Ev_Vie_Benoit	12,792	1	0.64
40_Wau_Leg-C_Co_Er_Vie_Alexis	4,103	1	0.64

Texts are sorted on V Ref in increasing order.

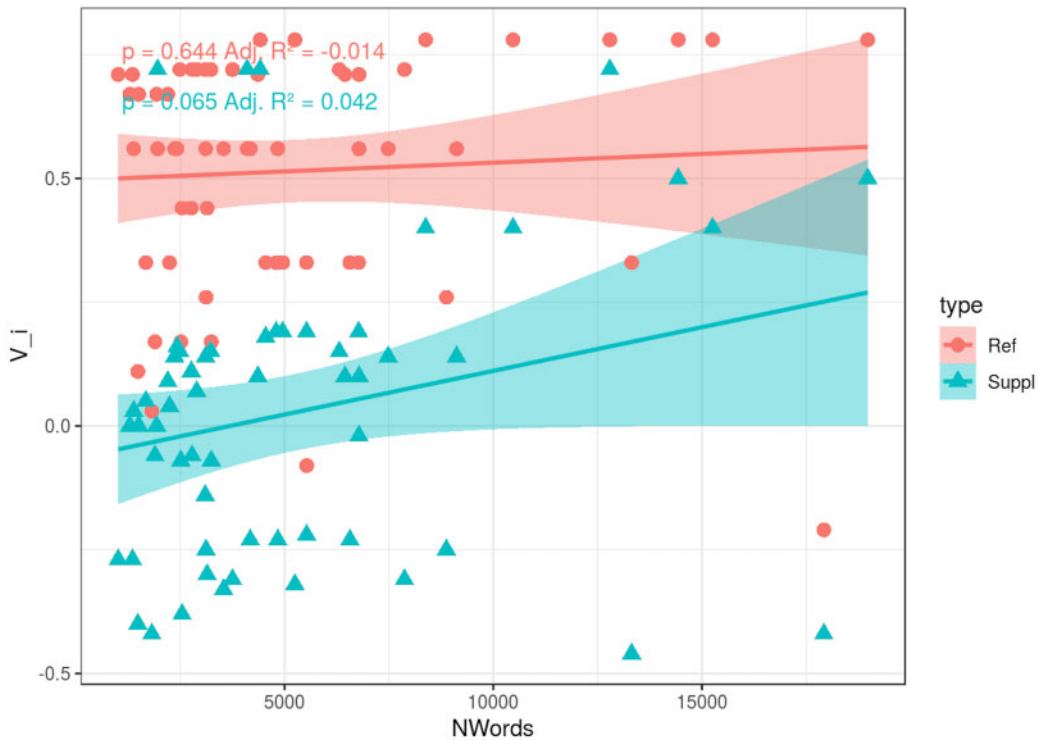

Fig. 9 Scatterplot and regression lines for V ∼ NWords, for both reference analyses and supplementary analyses

configuration, with or without *Lambert*, is also visible on all supplementary analyses, except the ones based on POS 3-grams and lemmas (Fig. 8, HC2). These two analyses also achieve low agglomerative coefficient, given their number of features, and low cluster purity, both in comparison with Meyer's classification and with our baseline; facts which advocates for considering them as outliers, with low reliability.

Moreover, in their globality, the results seem to agree with Meyer's hypothesis, with CP from 0.83 to 0.9 for the reference analysis (0.71–0.85 for the others, Table 7). This is particularly obvious for our baseline (see Fig. 7, HC1, top, reproduced here as Fig.10), that represents the manuscript fr. 412 as a successive addition of collections A, B, and C, which appear in separated branches. This can also be observed in the other trees, even if in a slightly

Table 9. Cluster purity of the analyses replicated using models trained on the TNAH corpus or with Kraken, with regard with the analyses presented in Fig.7, HC1 and Fig. 8, HC2

	TNAH Corpus	Kraken Model
Feature set 1		
Raw 3-grams	0.93	0.90
Feature set 2		
Affixes	0.90	0.92
Function Words	0.83	0.73
Function Lemmas	0.78	0.71
POS3gr	0.81	0.88
FW + POS 3-grams + Affixes	0.98	0.92
Feature set 3		
Forms	0.84	0.90
Lemmas	0.84	0.73
Words + Lemmas	0.95	0.95
Geom. mean Main analyses	0.95	0.92
Geom. mean Suppl. analyses	0.83	0.81
Geom. mean all	0.87	0.84

noisier fashion. For this reason, Paul Meyer's hypothesis seems to be confirmed by our results. Nonetheless, they can be nuanced or made more accurate in a few cases, as we will see.

4.1 Volatile texts and exceptions to Meyer's classification

4.2.1 Andrew–Catherine *(43–45) and Assumption–Antichristi (42 and 60)*

A subgroup mixing texts from Meyer's B and C can be observed in our results (Fig. 7, HC1). It contains *Saint Catherine Life*, *Saint Andrew Life and Miracles*, *Saint Andrew Passion* (n. 43–45), as well as the *Assumption of Our Lady* and, but not always, *Antichristi* (n. 42 and 60).[9] In the trees, this subgroup is sometimes included in C, sometimes in A. The A*ssumption* and the *Antichristi* were identified by Paul Meyer as texts from collection B and the second one was assumed to have certainly been published first in an autonomous way with some others texts: *Saint Patrice's Purgatory*, Julian and Brendan Lives (Meyer, 1906, p. 405).

4.2.2 *Clément and Patrice*

Clément and Patrice are integrated in collection C (as opposed to Meyer's B) in almost all analyses. A precise interpretation of this remains to be found, but one can

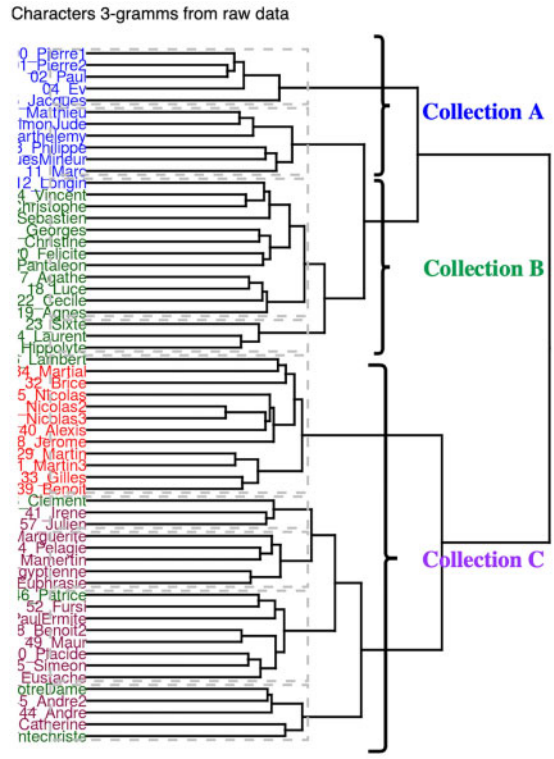

Fig.10 Results on char. 3-grams from raw HTR data, with indication of the successive collections

note that Patrice has been supposed by Paul Meyer as having been published first as part of a preexisting *libelli* and autonomously from collection B (or C) with *Antechristi*.

However, given the very small amount of texts erratically classified, and given the pre-existing difficulties faced by Paul Meyer concerning some of these, we do not expect this to contradict our former conclusion. The apparent volatility of Marc and Jacques can partially be disregarded, as they appear to only switch subgroups while remaining inside A.

4.2.2.1 Collection *A*

Paul Meyer identified collection A as a group of twelve texts. It is most apparent on the baseline analysis (Fig. 7, HC1, top). From our results, it can be refined in two subseries A1 and A2 (Table 10). Thematically, this regroupment makes sense. On one side, we find major apostles, creators of Christian Church, with two sequences of three and

Table 10. Subseries of A

Series A1	Series A2
Sequence 1	Passion, translation and miracles of Saint James the Greater (5)
Dispute of Saint Peter and Saint Paul against Simon the magician (0)	Saint Matthew Life (6)
Saint Peter Life and Passion (1)	Life of saint Simon and Jude (7)
Saint Paul Passion (2)	Saint Philip Life (8)
Sequence 2	Life of Saint James the Minor (9)
[*Martyrdom of Saint John in front of the Latin Door (3)*][a]	Saint Bartholomew Life (10)
Saint John the Evangelist Life (4)	Saint Marc Life (11)

[a]This text is amongst the one removed due to length below 1,000 words.

two texts following each other. On the other side are minor apostles which do not present sequences in theme or the likes.

To conclude about the collection A, in contrast to Meyer's hypothesis, *Saint Longin Life* is almost never grouped with any of the aforementioned series, but is nearby, being clustered as the first element of B.

4.2.2.2 COLLECTION *B* Collection B contains nineteen texts, thematically centered around martyrdoms. Of those, sixteen are sequentially found in MSS BnF fr. 412, the three others being *Our Lady, Antichristi, and Saint Patrice*. The latest are nearly never grouped with the main body of B in the trees.

We can observe a strong group composed of twelve Lives: Christophe, Agatha, Lucy, Agnes, Felicity, Christine, Cecile, Sixte, Laurent, Hippolyte, and Pantaleon Lives are clearly gathered (Fig. 7, HC1). We can add to them Georges, Vincent and Sebastien Lives, but saint Clement Life is always missing. Thus, globally, the sequence of the manuscripts is reflected in the classification.

Moreover, in all the selected trees, the *Life of Saint Longin* classified as A by Paul Meyer is gathered with texts from collection B. Furthermore, thematically, *Saint Longin's Life* isn't coherent with a series of saint apostles, given the fact that he is a martyr. In manuscript BnF fr. 412, given the order of the compilation, we can considerate *Saint Longin Life* as the last text of collection A or as the first of collection B. Looking at the manuscripts' tradition, in fact this life is mixed with the apostle lives just once in the manuscript BnF, nouv acq. fr. 23686, which happened to be the one Paul Meyer used as his prime material for studying collection A. So, regarding our results and manuscript tradition, it seems more accurate to classify this life to saint Martyrs within collection B.

In the light of this hypothetical classification, we can observe two subcollections (Table 11):

A micro-series B1a is composed by *Saint Longin, Saint Sebastien, Saint Vincent, Saint Christopher Lives*.

In addition, in B, a micro-series B1b of saint women Lives appears: Saint Agatha, Saint Lucy, Saint Agnes, Saint Christine, and Saint Cecile. Those Lives are about virgins and also close in the manuscript tradition: the first three texts of the series are often gathered together, as are the last three. There are also textual links between them. One explanation for the proximity between Saint Agatha and Saint Lucy can be the fact that the last seems to be in the continuation of the first's story.[10] Indeed, at the end of her story, Lucy defines herself as an heir of Agatha:

> *Aussi com la cites de Cathenense est secorue et aidie par seinte Agathe ma seror, aussi sera ceste citez aidie et socorue par moi, se uoz auez foi et creance en nostre Signor.*[11]
>
> *Just as the city of Catania was rescued by the help of St. Agatha, my sister, this city will be rescued by my help, if you have faith in our Lord.*

We can add that both of them come from Sicilia: Agatha from Catania and Luce from Syracuse. There are also some links between the Lives of Saint Lucy and Saint Agnes. The Life of Agnes starts at Rome where the Life of Saint Lucy stopped and they both have to face the threat of a spurned lover who wants to send them to a brothel. We can also note that Saint Christine, as Saint Agatha, is one of the four patrons of

Table 11. Subcollection B1 of men and virgin martyrs and subcollection B2

Subseries B1	
B1a—Men martyrs	B1b—Virgin martyrs
Saint Longin (*Meyer's A - 12*)	Saint Agatha + Saint Luce (*17-18*)
Saint Sebastien (13)	Saint Agnes (19)
Saint Vincent (14)	Saint Christine (21)
Saint Christophe (16)	Saint Cecile (22)
Subseries B2—Roman martyrs?	
Saint Sixte (23)	
Saint Laurent (24)	
Saint Hyppolite (25)	
Saint Pantaleon (27) ?	

Palermo in Sicilia and that the thematic of the snatched breast, iconic for Saint Agatha, can also be found in Saint Christine's Life. The reason behind the adjonction of Saint Cecile is more obscure: there might be a redundant theme around family and conversion. Finally, we can add that the Lives of this group are amongst the least volatile in our corpus after the Wauchier de Denain collection (Table 8).

However, both microseries B1a and B1b can be grouped together as a collection B1 in five of the selected trees. The rapprochement seems logical from the point of view of literary construction because it builds a collection with, on one side, five Lives of men martyrs, and on the other side, six Lives of women martyrs. A stylometric study cannot determine the order of apparition.

Finally, Paul Meyer, during his work about the different hagiographic collections, has seen that, in the collection B, the series Sixte-Laurent-Hippolyte was frequent in the manuscript tradition (Meyer, 1906, p. 495). This reunion appears in our three analyses. Furthermore, the dendrogram based on function words (Fig. 8, HC2, top-right) links Laurent and Hippolyte with Pantaleon. This addition isn't in contradiction with the tradition: collection G^{12} contains them sequentially in three of its four witnesses, and in collection C (three manuscripts) saint Pantaleon's Life is only separated by Saint Lambert's Life from the other ones. As such, it is possible that their gathering confirms a predating series.

4.2.2.3 Collection *C* Collection C contains twenty-two texts without any apparent major theme.

Table 12. Subseries of C

C1—Wauchier de Denain, *Li Seint Confessor*	
Saint Martin (29)	
[Saint Martin 2 (30)]	
Saint Martin 3 (31)	
Saint Brice (32)	
Saint Gilles (33)	
Saint Martial (34)	
Saint Nicolas (35)	
Saint Nicolas 2 (36)	
Saint Nicolas 3 (37)	
Saint Jerome (38)	
Saint Benoit (39)	
Saint Alexis (40)	
Saint Lambert (Meyer's B) to attribute to Wauchier?	
C2—Benedict and his disciples	
Translation of Saint Benoit (48)	
Saint Maur (49)	
Saint Placide (50)	
C3	
C3a	C3b
Saint Marguerite (53)	Saint Simeon (55)?
Saint Pelagie (54)	Saint Mamertin (56)
Saint Euphrasie (59)	
Saint Mary the egyptian (58)?	

[a]Wauchier's *Martin 2* is amongst the texts removed due to their length being below 1,000 words.

Collection C seems to have two major series, one constituted by Wauchier de Denain's *Seint Confessor*, and the other one containing all the others texts (Table 12).

First, we can see that the Lives of Saint Maur and Placide and *Saint Benoit's Translation*[13] form a series C2. The *translation* and the *Life of saint Maur* are grouped in our three analyses, and the *Life of Saint Placide* is also close to or part of the group. Those Lives have a thematic unity: the translation of the body of Saint Benoit, followed by the Lives of his disciples, Saint Maur and Saint Placide.

Another series C3 appears in all three analyses (Fig. 7, HC1): Saint Marguerite, Saint Pelagie, Saint Euphrasie and probably also *The Life of Saint Mary the Egyptian*. Surprisingly, we can extend this series to a subseries C3b containing one, perhaps two, men's lives: the *Life of Saint Mamertin*, always present as well, while the *Life of Saint Simeon* is more punctually associated with this group, when considering only function words or function lemmas.

Finally, this study has revealed an astonishing rapprochement between *Li seint Confessor* and the *Life of saint Lambert*. Normally, *Saint Lambert's Life* is part of collection B, but following our preceding analysis, Saint Lambert's does not fit in any group of the collection. In fact, we have to look at some of the supplementary analyses to find results where it is not associated with Wauchier's works (POS 3-grams, function lemmas, and lemmas, Fig. 8, HC2), all potentially influenced by the nature of the training corpus. However, from a close reading perspective, it is difficult to affirm the authorship given the fact that *Saint Lambert Life* does not possess any of the usual distinctive marks of Wauchier de Denain's style such as verses in prose, signatures or vernacular translation of Latin citations. Moreover, there is no reference to Philippe of Namur, Jeanne of Flanders or Roger squire of Lille, Wauchier de Denain's patrons (Douchet, 2015). There is also no evident Latin common substrate between the lives of *Li Seint Confessor* and *Saint Lambert Life*. The only common point which can be found is in the localization and the theme. Liege is close to the Namur area.[14] Saint Lambert is an important saint, a bishop linked to power, having contact with Pepin, king of the Franks. The chosen version is the one with a positive representation of royal power. So, we are in the presence of a bishop, ally of power, like Saint Martin or Saint Martial. On the other hand, one possible hypothesis regarding a potential Wauchier's authorship is that *Saint Lambert Life* is an early text, where the author erases himself and stays in the role of a simple translator. Consequently, without any easy proof of classification, further study of *Saint Lambert Life*'s relationship with Wauchier's work should be done.

5. Conclusion and Further Research

The aim of our approach was to conduct an experiment that went beyond the HTR stage and to offer a complete pipeline to acquire and analyze medieval data, using a hybrid approach of human and artificial intelligence to answer a research problem, in this case the compilation mechanisms of the legendary C. Using machine learning, we are able to acquire and process textual data from medieval manuscript images, and then submit it to a stylometric analysis setup that seems able to deal with noisy data. The application of this procedure offers perspectives for ancient or medieval Cultural heritage data, and could be extended to other material.

From a methodological and stylometric perspective, the challenge was, in a context where supervised analysis was not an option, to deal with short texts,[15] whose data were noisy in two regards: first, because of the noise generated during text recognition (HTR) and further processing steps; secondly, because of the noise to the authorial signal inherent to medieval data (spelling variation, variants, etc.). From our observations, our baseline that involved character *n*-grams with raw HTR data (already suggested as the most adapted to noisy OCR or HTR data in previous studies) can still be considered a very efficient procedure. Our attempts to suppress noise due to spelling variation by using other types of features such as lemmas or POS 3-grams, though offering alternative insights into the data, do not seem yet able to surpass it significantly, perhaps because of the cumulative error rates for each processing step. On the other hand, concerning the shortness of the texts, our results seem to agree with the notion that it is possible to analyze texts below 3,000 words; more specifically, by using less sparse features, such as characters 3-grams, or by concatenating different features, different views on the same text, we seem to achieve stable results, independent of the variation of sample length in our corpus.

From a thematic perspective, on the whole, our results confirm (or fail to disprove) Meyer's hypothesis regarding the constitution of Old French *legendiers*. They also bring to light some new facts, such as potential subseries that were not previously identified, and raise questions about *Saint Longin*'s life, that, we believe, can be considered part of collection B instead of A, and the life of *Saint Lambert*, whose possible attribution to Wauchier de Denain is deserving of further investigation. The whole process could be applied to the two other witnesses of the legendary C, but also to legendaries, such as those of the family G (3 witnesses) which are later manuscripts with a more complex tradition and whose compilation is less conservative of the original blocks in order to further strengthen our conclusions and sustain the analysis.[16]

Finally, we hope that our approach can motivate new investigations, using computational humanities,

on philological and historical holistic hypotheses formulated in the nineteenth century, that still sometimes form the basis of our understanding of the sources. By bringing together the work of the founders of our fields, such as Paul Meyer, and novel computational methods, we can hope to achieve progress in many areas, and perhaps more specifically in those that are left out of the literary canon envisioned by many close reading studies.

Supplementary Data

Supplementary data are available at DSH online.

References

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference;* Victoria, BC, Canada.

Boldsen, S. and Paggio, P. (2019) Automatic dating of medieval charters from Denmark. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, CEUR Workshop Proceedings*, **vol. 2364**, University of Copenhagen, Copenhagen, Denmark, pp. 58–72.

Cafiero, F. and Camps, J.-B. (2019). Why Molière most likely did write his plays. *Science Advances*, **5**(11).

Camps, J.-B. (ed.) (2019). Geste: Un Corpus de Chansons de Geste, 2016-. . . (Version 02). Paris, France. http://doi.org/10.5281/zenodo.2630574.

Camps, J.-B. and Cafiero, F. (2013). Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature. *Revue Des Nouvelles Technologies de l'information (RNTI)*, SHS-1, pp. 55–84.

Careri, M., Fery-Hue, F., Gasparri, F., et al. (2001). Album de manuscrits français du XIIIe siècle. Rome: Viella.

Clérice, T. (2019) Evaluating deep learning methods for word segmentation of Scripta Continua texts in old French and Latin. *Journal of Data Mining and Digital Humanities*, **2020**. https://hal.archives-ouvertes.fr/hal-02154122v2

Clérice, T., Camps, J.-B. and Pinche, A. (2019). Deucalion, Modèle Ancien Francais (0.2.0). *Zenodo*. https://doi.org/10.5281/zenodo.3237455.

Dahllöf, M. (2020). Classification of medieval documents: determining the issuer, place of issue, and decade for Old Swedish Charters. DHN 2020 Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference / [ed] Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, and Jānis Daugavietis, 2020, pp. 12–23.

Douchet, S. (ed.) (2015). Wauchier de Denain, polygraphe du XIIIe siècle, Aix-en-Provence: Presses universitaires de Provence.

Eder, M. (2017). Short samples in authorship attribution: a new approach. In DH. https://dh2017.adho.org/abstracts/341/341.pdf. (accessed 1 November 2019).

Eder, M. (2015). Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, **30**(2): 167–82. doi: 10.1093/llc/fqt066.

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, **28**(4): 603–14.

Evert, S., Proisl, T., Jannidis, F., et al. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **32**(suppl_2): ii4–ii16.

Franzini, G., Kestemont, M., Rotari, G., et al. (2018). Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm. *Frontiers in Digital Humanities*, **5**. doi: 10.3389/fdigh.2018.00.

Garz, A., Fischer, A., Sablatnig, R. and Bunke, H. (2012) Binarization-free text line segmentation for historical documents based on interest point clustering. In *2012 10th IAPR International Workshop on Document Analysis Systems*, Gold Coast, QLD, Australia. pp. 95–99. doi: 10.1109/DAS.2012.23.

Gómez-Adorno, H., et al. (2018). Document embeddings learned on various types of N-Grams for cross-topic authorship attribution. *Computing*, **100**(7): 741–756.

Ing, L. (in progress). *Disparitions lexicales en diachronie: traitements automatiques sur le Lancelot En Prose.* PhD thesis, Paris: École nationale des chartes. http://www.theses.fr/s221114 (accessed 1 November 2019).

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Improving Burrows' delta—an empirical evaluation of text distance measures. in *'Book of Abstracts of the Digital Humanities Conference 2015' Sydney, 2015.* https://www.researchgate.net/profile/Steffen_Pielstroem/publication/280086768_Improving_Burrows'_Delta_-_An_empirical_evaluation_of_text_distance_measures/links/573ad8ae08ae9f741b2d3d40.pdf (accessed 23 May 2017).

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), Gothenburg, Sweden. pp. 59–66.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science & Technology*, **60**(1): 9–26. doi: 10.1002/asi.20961.

Kunstmann, P. (ed.) (2009). Chrétien de Troyes: Cligès, Erec, Lancelot, Perceval, Yvain – Manuscrit P (BnF Fr. 794). http://www.atilf.fr/dect.

Manjavacas, E., Kádár, Á. and Kestemont, M. (2019a). Improving lemmatization of non-standard languages with joint learning. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota. Association for Computational Linguistics ArXiv:1903.06939 [Cs]. http://arxiv.org/abs/1903.06939 (accessed 23 October 2019).

Manjavacas, E., Kestemont, M. and Clérice, T. (2019b). emanjavacas/pie v0.2.3. Zenodo. https://doi.org/10.5281/zenodo.2654987

Meyer, P. (1906). Légendes hagiographiques en français. In Histoire littéraire de la France vol. 33. Paris, France, pp. 328–458.

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.

Olivier-Martin, F., Duval, F. and Ing, L. (2018). Les Institutes de Justinien en français, Paris, 1935, éd. revue par F. Duval, lemmatisée par F. Duval et L. Ing.

Perreaux, N. (2011). De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques, Digital diplomatics: the computer as a tool for the diplomatist? [International Conference "Digital Diplomatics 2011", September 29th–October 1st 2011]

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

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

Pinche, A., et al. (2019). Chartes-TNAH/digital-edition: 2019. doi: 10.5281/zenodo.3522144.

Pinche, A. (2021). *Edition nativement numérique des oeuvres hagiographiques 'Li Seint Confessor' de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France.* Ph.D. thesis, Lyon: Université Lyon 3. http://www.theses.fr/s150996 (accessed 9 May 2017).

Sapkota, U., Bethard, S., Montes, M. and Solorio, T. (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*, Denver, Colorado. pp. 93–102.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, **21**(2): 421–39.

Stutzmann, D. (2013). Ontology research, image features, letterform analysis on multilingual medieval scripts – ORIFLAMMS.

Stutzmann, D. (2019). Words as Graphic and Linguistic Structures: Word Spacing in Psalm 101 Domine Exaudi Orationem Meam (11th–15th c.). Victoria Turner; Vincent Debiais. *Les Mots au Moyen Âge – Words in the Middle Ages*, Brepols, pp. 21–59.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301): 236–44.

Wahlberg, F., Mårtensson, L. and Brun, A. (2016). Large scale continuous dating of medieval scribes using a combined image and language model. In *12th IAPR Workshop on Document Analysis Systems (DAS 2016)*, Santorini, Greece, pp. 48–53.

Notes

1 Paul Meyer (1840–1917) is one of the most famous French romance philologists. He is the author of the chapter about French hagiographic Lives in the book *Histoire littéraire de la France* (1906).

2 Manuscript BnF fr. 412 on Gallica (https://gallica.bnf.fr/ark:/12148/btv1b84259980/f1.item).

3 It should be noted that Kraken main developer, Benjamin Kiessling, is working on implementing learnable layout analysis (cf. https://github.com/mittagessen/kraken/issues/155).

4 However, it is not a specific handwriting issue, but more a question of what was space and words at the time. On this subject, see (Stutzmann, 2019). Moreover, there is a strong difficulty in appreciating the difference between

the intent of the scribe and what we perceive of it (Careri *et al.*, 2001: XXXVII).

5 We used a model on version 0.2.3 (https://github.com/emanjavacas/pie/releases/tag/v0.2.3).

6 In the transkribus dataset, Wauchier counts for 122,000 tokens over 446,900 for the whole treated part of the manuscript.

7 The lemmata are taken from Tobler, A., and Lommatzsch, E. (1952). *Altfranzösisches Wörterbuch*. E. Steiner.

8 The collection contains the nine following texts : *Saint Martin Life, Sulpicius Severus Dialog, Saint Brice, Saint Gilles, Saint Martial, Saint Nicolas, Saint Jerome, Saint Benoit*, and *Saint Alexis Lives*.

9 *Antechristi* is only 1,346 words long, which is one of the smallest size of the corpora, and below the first quartile. One could hypothesize that the *Jugement* (61) could function with *Antechristi*, but this text was excluded because it is below the 1,000 words limit.

10 In the manuscript tradition (we check eighteen manuscripts from *Li Seint Confessor* tradition), there is a really strong link between Agatha and Lucy Lives, which appear together in eleven manuscripts, against seven witnesses where they appear separated in which only two manuscripts have both Lives.

11 Manuscript BnF fr. 412, fol. 72v.

12 Bruxelles, Bibliothèque royale, 9225, Paris, BnF, fr. 183, Paris, BnF, fr. 185.

13 The *Translation of Saint Benoit* is not attributed to Wauchier neither by Paul Meyer nor by our stylometric analysis. Moreover, it is separated from the Life in MSS fr. 412 by 8 others texts.

14 Indeed Wauchier de Denain comes form the North of France and worked under the patronage of the court of Flanders.

15 There has been more substantive research on the supervised classification of medieval charters, in terms of typology, date or issuer (Perreaux, 2011; Wahlberg *et al.*, 2016; Boldsen and Paggio, 2019; Dahllöf, 2020).

16 This, however, means also creating more ground truth transcriptions, normalization and lemmatization of manuscripts. We estimate that the Pinche Dataset (271 columns) was done over the span of 3 years, on the side of other tasks a PhD students typically has to do (including teaching and other research). It is also highly dependent on the digitized version of manuscripts and the ability to use the original model to kickstart transcription.