

Estelle Guéville and David Joseph Wrisley*

Everyone Leaves a Trace: Exploring Transcriptions of Medieval Manuscripts with Computational Methods

<https://doi.org/10.1515/dsl-2024-0012>

Received August 30, 2024; accepted October 11, 2024; published online November 29, 2024

Abstract: The topic of this paper is a thirteenth-century manuscript from the French National Library (Paris, BnF français 24428) containing three popular texts: an encyclopedic work, a bestiary and a collection of animal fables. We have automatically transcribed the manuscript using a custom handwritten text recognition (HTR) model for old French. Rather than a content-based analysis of the manuscript's transcription, we adapt quantitative methods normally used for authorship attribution and clustering to the analysis of scribal contribution in the manuscript. Furthermore, we explore the traces that are left when texts are copied, transcribed and/or edited, and the importance of that trace for computational textual analysis with orthographically unstable historical languages. We argue that the method of transcription is fundamental for being able to think about complex modes of authorship which are so important for understanding medieval textual transmission. The paper is inspired by trends in digital scholarship in the mid-2020s, such as public transcribe-a-thons in the GLAM (Galleries, Libraries, Archives and Museums) sector, the opening up of digitized archival collections with methods such as HTR, and computational textual analysis of the transcriptions.

Keywords: scribal behavior; medieval textual creation; computational textual analysis; handwritten text recognition; transcription; Paris, BnF français 24428

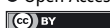
The manuscript Paris, BnF français 24428 stands out amongst many medieval compiled codices in that a scribe's name is mentioned at the end of the first text, Gautier de Metz's *L'Image du Monde*, and this named scribe, Omons, also claims to be the translator of one of the other texts in the codex. Additional texts in the same codex copied around 1265 include *Li Bestiaires Divin* (ff. 53r–78v), *Li Laipidaires*

Estelle Guéville and David Joseph Wrisley contributed equally to this work and share first authorship.

***Corresponding author: David Joseph Wrisley**, Arts and Humanities, NYU Abu Dhabi, P.O. Box 129188 Abu Dhabi, United Arab Emirates, E-mail: djw12@nyu.edu. <https://orcid.org/0000-0002-0355-1487>

Estelle Guéville, Medieval Studies, Yale University, New Haven, CT 06520-8287, USA, E-mail: estelle.gueville@yale.edu. <https://orcid.org/0000-0003-2603-1051>

 Open Access. © 2024 the author(s), published by De Gruyter on behalf of Chongqing University, China.

 This work is licensed under the Creative Commons Attribution 4.0 International License.

(ff. 79r–88v), both translated by Guillaume Le Clerc, as well as the *Fables d'Ésope* authored by Marie de France (ff. 89r–114v).¹ Textual objects from the Middle Ages were the product of many anonymous identities: scribes, correctors, illuminators, binders and such a codex provides an excellent opportunity for an in depth study of the cooperation of authors and scribes (even translators) in a specific medieval document. Due to the popularity and availability of digitized copies of other witnesses of the texts in question, it is also an ideal test case for using computational methods applied to medieval texts to examine overlapping and intertwined linguistic data found in them and to assess the kinds of claims made by its medieval creators with the conclusions one might draw from computational analysis, such as whether or not a named author also acted as a scribe.

In this article, we explore questions of collaboration in the creation of textual compilations such as a codex, by focusing on the precise orthography found in different witnesses of the texts found in this manuscript. We obtained digital texts in the form of computer-created plain text transcriptions to carry out our research using mixed methods. Where possible, when print editions of the texts in question exist, we have used Tesseract, an open-source optical character recognition (OCR) engine on scans of them to obtain computer-readable text.² We have also used diplomatic versions of texts in manuscript created in two different ways: crowd-transcribed by a group of specialist volunteers, and machine transcribed using state of the art handwritten text recognition (HTR) models we trained for old French in Transkribus and a publicly available digitized copy of the manuscript from the Bibliothèque nationale de France (BnF) (Kahle et al. 2017). Then, we use research methods well established in literary studies for authorship attribution, and to a lesser extent used for scribal detection (Haverals and Kestemont 2023; Kestemont and Van Dalen-Oskam 2009), to examine what influence the scribe(s) found in our complex manuscript might have had on the texts we find in it. In particular, we compare this evidence with textual and paratextual statements made within the manuscript to hypothesize about the validity of these medieval claims.

Finally, and more broadly, we argue in this article that all philological practices with medieval sources, both in the past and the present (editing, transcribing, HTR, scribal copying), leave a detectable trace of those carrying out these interventions by virtue of the specific choices made in working with the documents. This trace, we also

1 Most of the texts found in the manuscript Paris, BnF français 24428 have been edited, but some of them are very popular texts at the time. According to the online Jonas database (IRHT), the verse *Image du Monde* is extant in over 75 copies, *Li Bestiaires Divin* in 25 and the *Fables d'Ésope* in more than 30.

2 Tesseract is not a perfect engine for text creation from print of all eras and languages, but it performed satisfactorily for our purposes of creating digital text from print editions. See Smith (2007, 2013).

argue, must be understood when using computational tools in a pre-modern context as it has a significant impact on the outcomes of “reading machines” (Ramsey 2011). Computational textual analysis offers a number of interesting possibilities when studying the literary archival object: identifying relevant patterns in language usage specific to regional differences in orthography or phonology, exploring the interwoven linguistic traces found within the textual elements of a codex and generating hypotheses about collaborative literary creation in the past. The consensus amongst practitioners is that computational forms of analysis must be used carefully, and this is especially true for medievalists, since we do not always have enough contextual information about our archives. Nonetheless, mixed methods approaches for the study of literary texts which draw on the rapidly changing landscape of machine learning and artificial intelligence provide medieval studies, and literary studies more generally, with innovative possibilities to think more holistically about literary creation.

1 A Method for Detecting Changes in Scribal Hands

Traditional approaches to studying an entire manuscript might include examining material features, paleography, or collation to understand the conditions of its production better. Scholars might study the iconographic program of the codex or search for clues to help to establish provenance history. For decades now literary scholars have been interested in how complex the collaboration between scribes and authors has been. We turn to methods in digital authorship attribution, a set of computational practices commonly known by the umbrella term “stylometry” (Tempestt et al. 2018) in order to explore possible contributions to manuscripts. A stylometric approach assumes that linguistic features of texts are measurable and quantifiable (Herrmann, van Dalen-Oskam, and Schöch 2015), and the statistics of such measurement allow us to suggest relationships between texts. We address how such computational techniques might be used in the case of premodern handwritten codices, not only to make evidence-based claims about the singular author of a text, but also to include the scribe, the translator and the modern scholar in a more holistic study of the complex authorial regimes present in the ways that medieval texts have been transmitted across time.

This study uses digital text from a variety of sources, including transcriptions from manuscripts and digitized critical editions. One of the core assumptions in digital authorship attribution studies is that texts known to be written by the same author can be shown to be similar when subjected to statistical analysis of the words,

even if they were published anonymously or under pseudonyms (Skorinkin and Orekhov 2023). What then of the medieval author function where the orthography in texts in manuscript is highly unstable, the production of manuscripts is known to be complex, and the various actors involved in that process are unknown or debated? Moreover, when the fabrication of a codex takes place at the hand of multiple actors, including someone who claims to be more than just a scribe but also an author of a text, as Omons does in this case study, what kinds of complexities lie in waiting when we try to do statistical analysis of texts based on the words contained in it? How much of the linguistic data from medieval texts is reflective of the author, the scribe, the translator or even the contemporary editor? Given the complexity of medieval textuality, interpreting the results of computational analysis of texts requires great care.

Some research has been carried out on medieval vernacular literary collective authorship in single texts. As far as we can tell from published studies, in some cases the computational analysis has been carried out on normalized versions of the text based on editions. One such study is that of Chretien de Troyes' old French *Lancelot*, arguing that different lexical usage patterns appearing around line 6,150 suggest a shift in authorship from Chrétien to Godefroy of Lagny, responding to a long-standing debate about that text (Reilly and Dillon 2013). Other studies used TEI XML transcriptions with specific systems encoding markers of scribal signs, as in the case of the middle Dutch letters of Hadewijch (Kestemont 2015) or the middle Dutch *Roman van Walewein* and the two identified authors, Penninc and Vostaert (van Dalen-Oskam and van Zundert 2007). In the second, the research question is focused on the possibility of computationally detecting dual authorship using critical editions as data, and in the third, in exploring distinctions between the scribe and author in a single manuscript. Hand-transcribing an entire manuscript – or multiple texts found in multiple manuscripts – is technically possible, but not scalable, given the human labor required to do so. Possibilities for more scalable text creation include academic crowd-sourcing (in fact, our own work was prompted by a collective transcription event known as the Image du Monde Challenge), or AI-based transcription using HTR platforms. Here we use a combination of both.

In the first part of this article, we use a similar methodology applied to vernacular European literature, following van Dalen-Oskam and van Zundert, but in order to explore a cross-genre compilation of seven texts in one manuscript. Rather than a hand-transcribed TEI XML edition, we use an automated transcription of the entire manuscript in plain text format. Indeed, access to mature AI technologies such as HTR has provided the archival sector in general, and the medievalist in particular, with unprecedented ways of creating digital texts for analysis (Nockels et al. 2022). In our research we take advantage of this technique to transcribe texts directly from a manuscript, capturing new layers of data by using a special transcription system in

ways that would have not been practical in the past.³ These kinds of automated transcriptions that preserve scribal idiosyncrasy provide us with insight into the textured language of manuscripts, a kind of data from medieval compilations produced within complex creative regimes. Our study leverages a few trends in contemporary digital scholarship; it combines public transcribe-a-thons in the GLAM (Galleries, Libraries, Archives and Museums) sector, well-established modes of creating digital text with OCR, a newer form of opening up of digitized archival collections with methods such as HTR, and computational textual analysis of the transcriptions in both Python and R. It combines a prevalent method in medieval studies, namely studying texts in their material context, with computational methods for text creation and textual analysis (Deploige and De Gussem 2021).

In our analysis, we also make an important assumption: that the act of copying a manuscript in the middle ages leaves traces of the scribe(s) in the language, specifically with respect to orthographic and abbreviation habits. Given our interest in the scribal role in copying texts (and the “transcribal” role of modern editors in creating editions), it follows that we must be very intentional about the norms of transcription to carry out our research. This applies as much to the way we train our HTR system to maintain medieval orthography and letter forms, as to our awareness of the norms followed by modern editors.⁴

To work with the manuscript BnF français 24428, we created transcriptions from scratch, not because we lack editions of these works, but rather because we require non-normalized versions of the texts that give us access to the ways that they were transmitted in the medieval period so that we can understand more about the act of copying. In other words, non-normalized transcriptions give us access to what has been called the “closed archive,” by which is meant all the data that are hidden in manuscripts and do not find their way into an edition (Deploige and De Gussem 2021). Using a custom-designed transcription schema, the text we create with HTR includes unprecedented direct access to linguistic information, capturing the text as it was spelled in the manuscript. Strictly speaking, the transcriptions are non-normalized from the perspective of a number of micro-features such as abbreviations and special letter forms, rather than capturing paleographic diversity across the manuscript. Figure 1 gives a sample transcription of the sort we create. For example, the two columns of the transcription reproduce the exact number of words in a line, the layout of the manuscript, the special letters of medieval scribes and any abbreviations used. Lines 6–7 of the left-hand column in a traditional edition would be transcribed and normalized as follows: “La première partie contient sept chapitres et

³ For more details about the transcription method we used see Guéville and Wrisley (2024).

⁴ Many guides for editing vernacular texts exist offering criteria for normalization of spelling while editing a text. An exemplary one is Foulet and Speer (1979).

contient par tout
 ·lv· chapifres· ʒxx·
 ʒ ·viiij· figures sanzcoi
 li liures ne porroit eft
 legiereṁt qui eft deufeiz en iij· parties·
 La p̄emere partie contient· uij· cha
 pifres ʒ ix· figures fans le p̄ologne·
 Li p̄emiers chapifres eft de la poiffa
 ce n̄re fegnoꝝ· Li fecons por coi dex fift
 le monde· Li tiers porcoi dex fift lome
 a fa famblance· Li q̄rs porcoi dex ne
 fift lōme teil quil ne peut pechier· Li
 cm̄qm̄ef por coi ʒ com̄t lef ·vii· arf fuṽt
 trouees ʒ de loꝝ oꝝde· Li fiffimes def ui·
 menieres de gens que li philofophe po
 ferent au monde ʒ om̄t clergie uint en

france Li fetimes de la meniere des
 ·vij· ars· Li oomumes de nature q̄m̄t
 ele oeure ʒ que ce eft· Li noeuimes
 de la foꝝme dou firmaṁt Li difime·
 om̄t li ·iij· eleṁt font affis· Li ou
 fim̄ef q̄m̄t la terre fe tient enmi le ciel
 l i doufimes ḡm̄t ʒ queile la reondef
 ce de la terre eft· Li trefimes por coi
 dex fift le monde reont· Li q̄toꝝfimes
 def ifueleteiz dou cours dou firmaṁt
 & def ·vii· planetes·
 a feconde partie ḡtient etx cha
 Apifres ʒ ·iex· figures· L i pmier
 chapifres eft q̄m̄t la terre eft deufeiee
 en diufes parties ce quelpart ele eft
 habitee ʒ i fecons eft la maremōde

Figure 1: A sample of the automated transcription made from the manuscript Paris, BnF français 24428 using our custom transcription scheme preserving special letter forms and abbreviations typical of medieval manuscripts.

neuf figures sans le prologue” [*The first part contains seven chapters and nine images without the prologue*]. Important to note is that contemporary accents have been added in this normalized version, abbreviations have been expanded, contemporary letters have been swapped for modern ones, errors in the transcription have been corrected and punctuation would have been fixed.

2 Scribal Identification in a Single Manuscript

The computational methods we carry out depend on the existence of open digital copies so that they can be manipulated; luckily, the manuscript has been digitized and is available in Gallica, the digital library of the BnF, published with an explicit open license for data reuse. Without such openness, it would be much more difficult to carry out this work. The manuscript dates from around 1265 and Table 1 details the contents of this compilation, made up of seven texts by different authors, copied by three or four different hands. For example, on ff. 35r to 78v we find a text entitled *Li Bestiaires Divin*, a French verse translation by Guillaume Le Clerc of a Latin text entitled the *Physiologus*. We suspect based on a visual paleographic analysis that

Table 1: A detailed table of the contents of the manuscript Paris, BnF français 24428, including foliation, titles, authors, translators and scribes.

Foliation	Title/content	Author	Translator	Scribe
1r-46r	L'Image du Monde	Gautier de Metz	–	Omons
47r-48v	La Recapitulations des choses devant dites	Omons	–	
49r-52r	Li Volucraires	Hugues de Fouilloy	Omons	Probably Omons despite few changes in compactness
53r-78v	Li Bestiaires Divin	Unknown	Guillaume Le Clerc	
79r-88v	Li Laipidaires	Unknown	Guillaume Le Clerc	
89r-114v	Fables d'Ésopes	Marie de France	–	Scribe 2
115r-118r	Instruction pour la confession	Unknown	–	Scribe 3
118v	Ex-libris			

the scribe is the same as the two first texts, although there are slight changes in the hand.

In a colophon on f. 48r at the end of *L'Image du Monde*, a scribe named Omons is mentioned: “Omons a non, qui fist ceste weure” [*The one who made this work is named Omons*]. A few folios later is also written: “Dou latin a trait ceste rime / Omons, li clers, par soi meïsmes. / Proiez por lui, si ferez bien, / Qu'il ne vous a menti de rien. Explicit” (f. 52r) [*Omons the clerk made this poem from Latin all by himself. Pray for him, if you will, that he has not lied about anything. Explicit.*] According to these two inscriptions, we assume for our research purposes that Omons copied at least these two texts (*L'Image du Monde* and *Li Volucraires*), situated between ff. 1r and 52r. These give us a good sample of the features of Omons' way of copying. A paleographic analysis of the manuscript distinguished several hands: despite a few changes in compactness across the manuscript, it is probable that Omons wrote ff. 1 to 88, until the end of *Li Laipidaires*. The following text, the *Fables d'Ésopes* of Marie de France, starts on f. 89r and seems to have been written by another hand, although the paleographical difference is not always clear.

What if we turn to computational analysis to explore the claims made in the manuscript and our cursory visual assessment of the paleography? Can we use digital textual methods usually applied to authorship attribution to identify a change of scribes within this single manuscript containing several texts of different genres? Is it possible to understand who between the author, the translator and the scribe has more impact on the way a text is written and how might we quantify this impact? Since we believe that the way words are spelled tells us something about the people

who were copying them, we are then in a position to apply some higher-level forms of digital analysis to dive into the codex.

Using the method of rolling stylometry, we wanted to see if it can help to identify the shifts in word usage in the manuscript BnF français 24428, shifts that could corroborate (or disprove) the claims about scribes, authors and translators related to the texts found in this manuscript. Rolling stylometry allows us to detect a change in word usage in texts (Eder 2016). It is a kind of sequential analysis, in layman’s terms, that moves across the text in slices to predict which of the hands for which we have a transcription is most likely the creator of a given portion of text. It inspired us to explore if, by measuring the common words of a transcribed manuscript, including the same words with variant orthography, it is also possible to recognize a change in overall usage indicative of a shift of scribal hand.

This first analysis looks at whether Omons, the scribe of the first text, the author and scribe of the second and the translator of the third, can also be the scribe, as we suspected, of the majority of the manuscript, especially in the center of the manuscript where the hand is more compact. In Figure 2, the x-axis is labeled with markers of the 5,000-word windows chosen for *rolling.delta*, and shows the progression of the words in the codex. Along the top of the graph in green are abbreviated names of the texts found in the codex, situated at the beginning point of those texts. There are three connected line graphs, corresponding to (1) Omons, the scribe of the first text

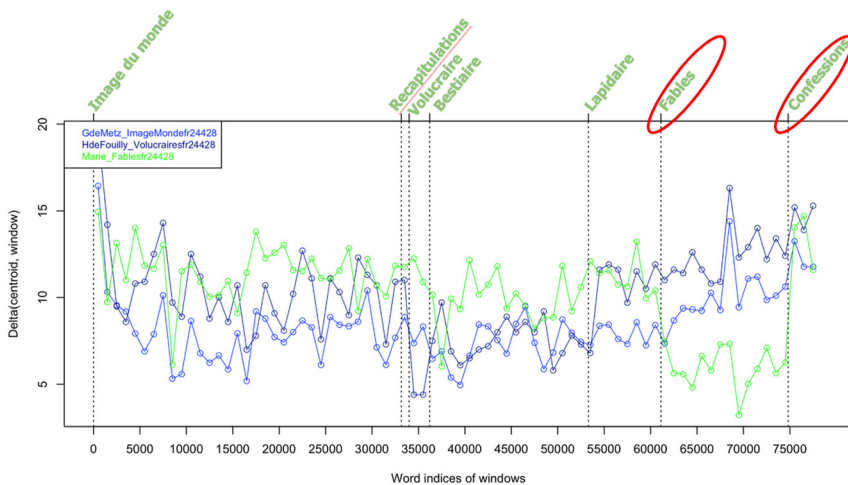


Figure 2: A plot of Delta scores for three possible reference texts (Gautier de Metz, Hugues de Fouilly and Marie de France) resulting from an experiment using the *rolling.delta* method for predictive classification of the scribal hands in the manuscript Paris, BnF français 24428. Visualized in R with the Stylo package.

(light blue); (2) Omons the translator of the third (dark blue); and (3) the unknown scribe of Marie de France's *Fables d'Ésopes* (green). Any given point along the x -axis provides the distance metric for that window of text; the point that is lowest on the y -axis across the plot indicates the best candidate for the copyist for that window of the text. The connected points visually represent the trend of the best candidate. For example, in the space between the two vertical lines on top of which we find the two red circles, the plot indicates more or less where there is a significant change in hand. What this experiment suggests is that for most of the manuscript, Omons was indeed the scribe of all the long texts except the *Fables d'Ésopes* written by Marie de France (starting around the 62,000-word mark on the x -axis). These results may be said to confirm what was written in the text and what we are able to posit with the naked eye, but the empirical confirmation using stylometry is nonetheless useful. Another layer of influence potentially here is that of the language of the translator – *Li Volucraires*, *Li Bestiaires Divin*, and *Li Laipidaires* all having been translated from Latin – although in the case of the first text the scribe and the translator are most likely the same. In the case of the second and third text, a more sophisticated approach would be required to distinguish between the author and the translator, potentially using normalized transcription (Rybicki and Heydel 2013). Our reader can appreciate the value of the rolling stylometry approach also when there is ambiguity on the page, either the material element of the page has made copying difficult or when a very professional scriptorium has made it difficult to tell when the hands change.

3 The Various Signals of a Larger Corpus

If digital methods can help identify different scribes in a single manuscript compilation of several texts as we just demonstrated, what about a larger framework in which we compare these individual texts copied here with witnesses of those same texts found in other manuscripts (and ostensibly copied by other scribes)? In these witnesses, would the author's signal override the scribe's signal? Or vice versa?

To test this hypothesis, we assembled a corpus of a larger number of transcriptions, by identifying digitized copies of manuscripts in which the texts in BnF français 24428 are also included, or other texts by the same authors as those found in our original manuscript. We worked with about sixteen different texts from seventeen different manuscripts, resulting in the HTR creation of about forty text files.⁵ The process of creating so many transcriptions is a laborious one, including the

⁵ In compiling this second corpus of witnesses of the same texts found in BnF français 24428, while there are indeed many extant copies, not all of these copies are digitized and openly available. This is

identification, preprocessing, layout analysis, and baseline detection identification of text blocks corresponding to the text, followed by the automatic transcription of the texts. This process presupposes an HTR model capable of transcribing the texts in this group of manuscripts with adequate precision.

Once our texts were processed, we had data that could be used to compare and contrast the ways that the various texts were written in different witnesses and different scribal contexts. To the abovementioned texts by known authors from other manuscripts which we transcribed using the non-normalizing method described in Section 1, we added all the transcriptions of the texts in BnF français 24428 transcribed using the same method as well as some additional texts. These additional texts were taken from both edited print versions of some of the works and transcriptions of the *L'Image du Monde* from several of the manuscripts resulting from a crowdsourcing event: the Image du Monde Transcription Challenge (Keane et al. 2021). This brought the total number of texts in our analysis to forty-seven. Combining these texts is an unconventional move on our part. Traditionally, scholars adopting computational textual analysis methods would use texts that adhere to the same, if not similar, transcription norms. Nonetheless, we combined them to assess the extent of the influence of such norms in the clustering methods we employ.

Instead of the rolling stylometry method used above, we adopted another common method for digital textual analysis known as Term Frequency Inverse Document Frequency (TF-IDF), in which our textual transcriptions are used in order to discover their relative similarity and distinctiveness in the corpus. We use an algorithm to approximate the importance of the most frequent words in each text, while also considering their importance in other texts within the corpus. Each one of the forty-seven documents is assigned a score that increases with more appearances of a word, but decreases if it is found throughout the corpus.⁶ In a Principal Component Analysis (PCA), we look for clusters of documents with similar scores, that is, which share similar abbreviated words with each other, but not with the rest of the corpus.⁷

One of the most important points we can underscore here again is that the mode of transcription is very important for the results we obtain through digital analysis. In Figure 3, we see an obvious distinction between editions (marked as 'ED' found

regrettable for the case of the many exemplars of *L'Image du Monde* found in libraries in the UK, but ultimately beyond our control. We used the following manuscripts for this corpus: Arsenal 3516; BnF français 14964; BnF français 14969; BnF français 14970; BnF français 19525; BnF français 20046; BnF français 2173; BnF français 24428; BnF français 24870; BnF français 25405; BnF français 25406; BnF français 25408; BnF français 3142; BnF latin 14470; BnF NAF 1104; Bibliothèque Sainte Geneviève 2200; Chantilly 0477.

6 For a detailed explanation of the method, see Ramsay (2011).

7 A similar attempt at looking at documents, albeit focusing on dialect, is given by Mäkinen (2020).

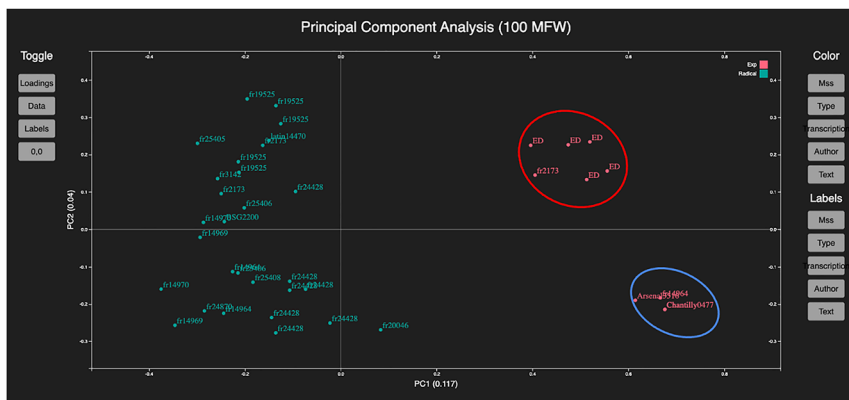


Figure 3: A Principal Component Analysis (PCA) resulting from TF-IDF analysis of forty-seven transcriptions of texts in French, including editions, normalized transcriptions and non-normalized transcriptions of *L'Image du Monde* by Gautier de Metz, the *Fables d'Ésope* by Marie de France, *Li Bestiaires Divin* by Guillaume Le Clerc and others. Carried out using the scikit library in Python. Code adapted from Paul Vierthaler.

inside the red circle) and normalized transcriptions on the one hand (clustering on the right side of the graph in the blue circle), and the non-normalized abbreviation and letter-form conservative transcription on the other hand (clustering on the left side of the graph). Normalized texts cluster together (lower right), editions cluster together (upper right), and non-normalized transcriptions cluster together (left). Working with this mixed corpus where the mode of transcription varies, neither the author, scribe nor translator has the strongest signal. Instead, the trace the modern transcriber (or editor) leaves on the text is the most significant. This fact points to the necessity of having a well-designed and consistent method for transcription in order to study scribal activity and to mitigate against the influence we have on texts when we transcribe them, changing them for modern literacies.

Looking more closely at Figure 3, there seems to be something special about the manuscript BnF français 2173 which had been produced as part of the crowd-transcription exercise mentioned earlier. It is found in the red circle, right next to the editions, which is not that surprising after all. It makes sense to expect that editions and normalized texts would appear similar: the set of rules used to transcribe texts for the purposes of an edition resembles that of the normalization process that can occur during transcription. This is particularly the case with BnF français 2173, since the group that created it used a similar set of rules as the ones usually used for editions. In fact, in the published transcription guidelines used for their transcription, not surprisingly, they refer to themselves as “editors” (Altunbas et al. 2021). This

is to say that when there are no specific guidelines for the preservation of scribal features, modern transcribers tend to normalize – consciously or unconsciously – adapting what they read in the manuscript to what they are used to reading in an edition.

In this section, we established the role that is played by transcription in text creation through editing or crowd-transcription and highlighted how a transcriber or editor leaves an inevitable trace on the text that can override any other historical signals. To preserve those signals, we recommend not combining texts transcribed using multiple methods to mitigate the risk of adding confusion; instead, a fully and consistently normalized corpus is most useful to identify authorship. When looking at scribal (regional or dialectal) influence, non-normalized transcriptions will be most useful if they follow the same set of rules and are not analyzed together with editions or normalized transcriptions.

4 Towards the Centrality of Scribal Language

The example above underscores the problem with normalization which is essentially one of data loss; once transcribed using normalization principles, it is impossible to recover the orthographic instability of documents without redoing the work. For this reason, for projects which intend to interrogate copying as part of the textual tradition, we highly recommend an unnormalized method, and if desired, devising a subsequent, semi-automatic normalization process. In order to reverse the effect of the normalizations of the crowd transcriptions, we created four new transcriptions of the segments of the manuscripts containing the same *Image du monde* (BnF français 14964, BnF français 2173, Chantilly, Bibliothèque du château, 477, and BnF Arsenal 3516), again using our HTR methodology that preserves scribal features. We would expect to obtain completely different results, and in fact, the newly transcribed texts do not cluster with their crowd-normalized counterparts (circle, bottom right), but rather they appear with the other manuscripts which have been diplomatically transcribed (oval, bottom left) (Figure 4).

For the remainder of this section, we reconstitute the corpus by removing the editions and normalized texts from the corpus altogether, reducing our corpus size to thirty-eight texts, allowing us to focus only on the texts transcribed with our feature-preserving method. When we do so, removing the transcribers' and editors' influence from the corpus, another pattern emerges.

Two clusters clearly marked by the manuscripts BnF français 24428 (red circle, lower left) and BnF français 19525 (blue circle, lower right) appear on the left and right sides of the graph. What this visualization illustrates is that when texts are transcribed including the features found in the manuscript, the scribal signal is

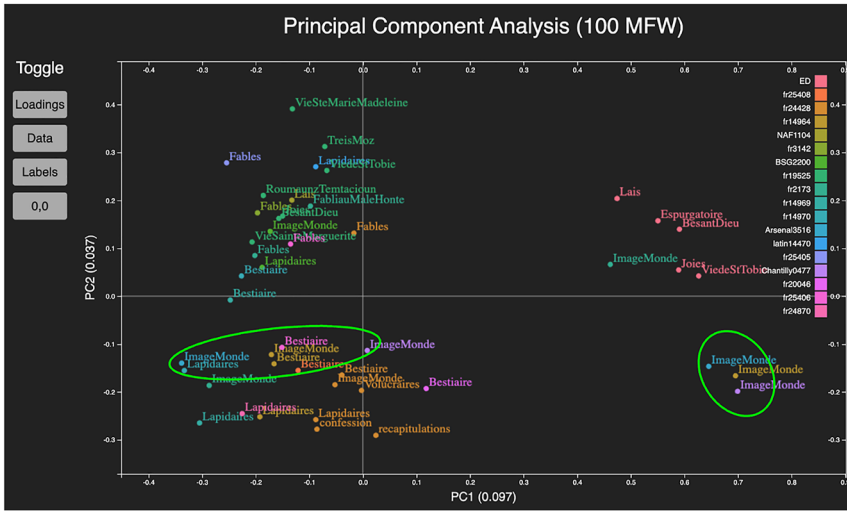


Figure 4: Principal Component Analysis (PCA) resulting from TF-IDF analysis of forty-seven transcriptions of texts in French, including four HTR-created, feature-preserving transcriptions of *L’Image du Monde* from the manuscripts BnF français 14964, BnF français 2173, Chantilly, Bibliothèque du château, 0477, and BnF Arsenal 3516. Carried out using the scikit library in Python. Code adapted from Paul Vierthaler.

stronger; instead of clusters of similar authors appearing in groups, we have groups of texts from the same manuscript. Whereas for the case of BnF français 24428, we mentioned above that it is possible that a signal emerges from the translator, again here in comparison with BnF français 19525 we are not yet able to investigate this question, lacking the appropriate study corpus. In summary, for this manuscript and for other texts in our research, we have found that orthographic and abbreviation patterns from a given manuscript are useful features that allow us to differentiate between manuscripts due to the signal of the scribe.

This technique also confirms the results we observed above about the scribes of the manuscript BnF français 24428 using rolling stylometry, again with the exception of the *Fables d’Ésope* of Marie de France (green oval, center right) which sit at a distance from the other texts from the same manuscript pointing to a distinct scribe, our second hand. This result would seem to reconfirm our hypothesis articulated above that the scribe named Omons copied all the texts from that manuscript except the *Fables d’Ésope*.

In this kind of exploratory analysis, using the technique of PCA, complex data in multiple dimensions is reduced into the two most important dimensions, and it is said these two components generally correspond to the two axes. According to our

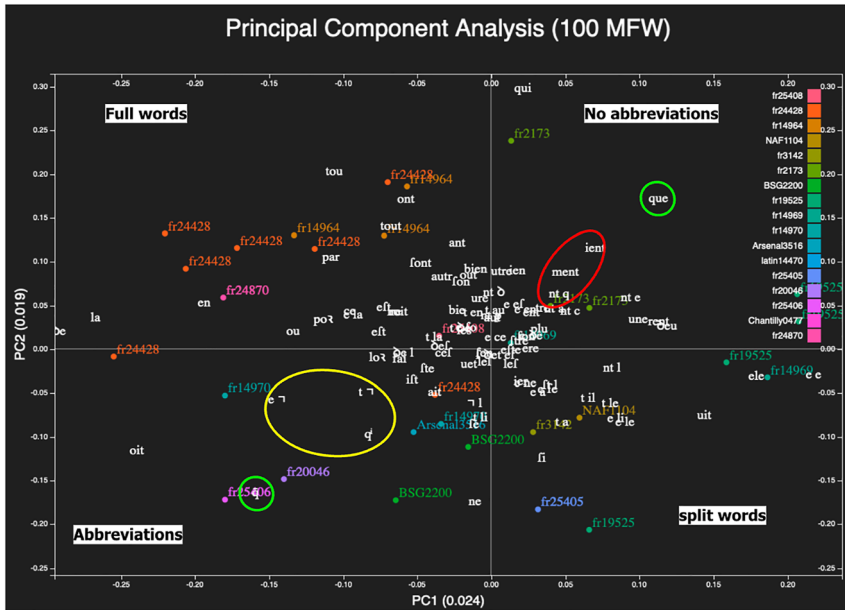


Figure 6: Principal Component Analysis (PCA) with word loadings resulting from TF-IDF analysis of thirty-eight non-normalized transcriptions of texts in French, using character 4-grams. Analysis carried out using the scikit library in Python. Code adapted from Paul Vierthaler.

syntax. It appears that it is not only abbreviations or special letter forms in words that are useful features for distinguishing between specific scribes, but also their position in words (especially at the end of words) as well as cases where scribes abstain from possible abbreviations.

In Figure 6, the graph visualizes a PCA plot of our corpus, but we have also included what are called loadings; these loadings included on top of the plot in white are the most significant 4-gram tokens that are distinguishing features for the various groupings of texts. These 4-gram tokens are helpful since they indicate important syntactical patterns in the features. We see different clusters of texts using distinctive ways of spelling (for example, “que” vs “q̄”), as well as different tendencies to abbreviate (or not) the ends of words. In the upper right quadrant we find word-endings which might be typically abbreviated, but are here instead written out, such as “ment,” “ient,” and “nt.” Those are typical of the manuscript BnF français 2173. In the lower left quadrant, we find many examples of the tironian ampersand and the abbreviation “q̄.” Perhaps most striking are the cross-word tokens we find in the lower two quadrants (but especially at right). We find both beginnings and endings of words and sometimes conjunctions between words. It seems that it is not only words,

but parts of words (character n-grams) that are important linguistic indicators of scribal practice. Put another way, what has emerged in the character 4-gram analysis represented in Figure 6 is a way to distinguish between scribes, given the ways they sequence the words in a copied text. In the loadings from a 4-gram analysis there are many words which are two or three letters. This indicates that the space, as a character, is itself also significant. It is not only special letter forms and abbreviations that factor into the ways we study such transcriptions but their position in a word – beginning, middle, end, or standalone – and this position is somewhat indicated by the spaces. Much more research needs to be done into this kind of scribal syntax, perhaps including not only the sequence of words or the position in the line, but also their position in a column or on a page of a codex, ultimately integrating scribal behavior on the physical page into the analysis.

5 Conclusions

The main conclusion we can draw from this paper is that, faced with the possibilities of digital analysis, *the mode of transcription is fundamental*. When asking the question about the impact on the writing and transmission of texts, how we can begin to answer largely depends on what texts are available to us, whether we are able to retranscribe them efficiently, and with what other texts we are able to compare them. The mode of transcription has a huge impact on the cluster analysis and rolling stylometry methods that we used above. This point is an argument in favor of being very intentional about how one transcribes: normalization processes erase the scribe's signal, whereas a non-normalized transcription will preserve it. Combining multiple modes of transcriptions (editions, normalized texts, non-normalized texts) will simply erase historical signals in favor of the modern scholar. Additionally, it is possible to locate a change of scribe in a manuscript if there is enough data and texts. It is said that *research follows record*, but in our case, we must revise this phrase to say that *research follows digitized record*. On the one hand, if some of the manuscripts discussed here were not available in digitized format we would not be able to carry out our work. On the other hand, if the numerous manuscripts of the *L'Image du Monde* located in libraries were digitized and available, we might have been able to present much more nuanced findings.⁸

Looking at the results outlined here, we argue that scribes are shaping a literary tradition as much as authors are, and this is not only in terms of content-based changes or textual interpolations. The shaping of texts can be found in many micro-

⁸ See the two manuscript lists at <https://arlima.net/no/1272> and https://jonas.irht.cnrs.fr/consulter/oeuvre/detail_oeuvre.php?oeuvre=3739.

features of language which are undone by the process of creating modern critical editions. These conclusions raise larger questions. How does the promise of automated transcription of medieval manuscripts change the way that we think about how we can use them? What will richer data-driven descriptions of manuscripts allow scholars to know about them, as individual objects, or as groups? What are the implications for incunabula and the instability of spelling and/or abbreviation?

One angle that we would like to explore in future work is how other compiled codices are constructed, moving beyond exclusively close reading and the lenses of genre or theme (Busby 2002). Perhaps most challenging of all, in our opinion, is how the future of digital medieval studies will look when we have a whole spectrum of texts available, each of which may have been created according to very different transcription criteria as we have seen in the corpora used for this article. And finally this question, which has been asked by Hodel in a recent publication: what sophisticated models of analysis will become necessary in that uneven landscape in order to compare such differently transcribed corpora (Hodel 2022)? A near future of computational medieval studies will be one in which many different versions of texts can be found – not only of the most famous and re-edited ones – or can be generated with AI-based HTR and medievalists will have to come up with the means of dealing with such heterogeneity.

Research ethics: Not applicable.

Informed consent: Not applicable.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Use of Large Language Models, AI and Machine Learning Tools: We have used the AI platform Transkribus to transcribe the medieval manuscripts used in the study. We have used the Stylo Package in R and scripts for TF-IDF in Python.

Conflict of interest: The authors state no conflict of interest.

Research funding: None declared.

Data availability: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

- Altunbas, Ahmet Deniz, Holly Barbaccia, Mary Flannery, Kyle Ann Huskin, Karl Kinsella, Anna-Amicia Litwinska, Aylin Malcolm, et al. 2021. "Team 2 Transcription Conventions." Image du Monde Challenge. <https://web.archive.org/web/20230320231734/https://imagedumonde.wordpress.com/team-2-phase-ii/>.

- Busby, Keith. 2002. *Codex and Context: Reading Old French Verse Narrative in Manuscript*. Amsterdam: Rodopi.
- Deploige, Jeroen, and Jeroen De Gussem. 2021. "Medieval Authorship and Canonicity in the Digital Age: An Introduction." *Interfaces: A Journal of Medieval European Literatures* 8: 113–24.
- Eder, Maciej. 2016. "Rolling Stylometry." *Digital Scholarship in the Humanities* 31 (3): 457–69.
- Foulet, Alfred, and Mary Speer. 1979. *On Editing Old French Texts*. Lawrence, KS: Regents Press of Kansas.
- Guéville, Estelle, and David Joseph Wisley. 2024. "Transcribing Medieval Manuscripts for Machine Learning." *Journal of Data Mining & Digital Humanities*. On the Way to the Future of Digital Manuscript Studies: 9805. <https://doi.org/10.46298/jdmdh.9805>.
- Haverals, Wouter, and Mike Kestemont. 2023. "From Exemplar to Copy: The Scribal Appropriation of a Hadewijch Manuscript Computationally Explored." *Journal of Data Mining & Digital Humanities*, On the Way to the Future of Digital Manuscript Studies. Experiences and Challenges: 1–21, <https://doi.org/10.46298/jdmdh.10206>.
- Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. "Revisiting Style, A Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52.
- Hodel, Tobias. 2022. "Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities." In *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*, 157–77. Bielefeld: Bielefeld University Press.
- Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. "Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR, Kyoto, Japan)*, 19–24.
- Keane, Monica, Laura Morreale, Christine Kralik, and Ben L. Albritton. 2021. "Image du Monde Transcription Challenge Project Archiving Dossier Narrative." BodoArXiv, <https://doi.org/10.34055/osf.io/ep3b6>.
- Kestemont, Mike. 2015. "A Computational Analysis of the Scribal Profiles in Two of the Oldest Manuscripts of Hadewijch's Letters." *Scriptorium* 69: 159–77.
- Kestemont, Mike, and Karina Van Dalen-Oskam. 2009. "Predicting the Past: Memory Based Copyist and Author Discrimination in Medieval Epics." In *BNAIC 2009 Proceedings of the Twenty-First Benelux Conference on Artificial Intelligence*, 121–8.
- Mäkinen, Martti. 2020. "Stylo Visualisations of Middle English Documents." *Journal of Data Mining & Digital Humanities*: 5614. <https://doi.org/10.46298/jdmdh.5614>.
- Nockels, Joe, Paul Gooding, Sarah Ames, and Melissa Terras. 2022. "Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research." *Archival Science* 22: 367–92.
- Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Champaign: University of Illinois Press. <https://muse.jhu.edu/pub/34/monograph/book/18394>.
- Reilly, Brian J., and Moira Rose Dillon. 2013. "Virtuous Circles of Authorship Attribution through Quantitative Analysis: Chrétien de Troyes's Lancelot." *Digital Philology: A Journal of Medieval Cultures* 2: 60–85.
- Rybicki, Jan, and Magda Heydel. 2013. "The Stylistics and Stylometry of Collaborative Translation: Woolf's Night and Day in Polish." *Literary and Linguistic Computing* 28 (4): 708–17.
- Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007, Curitiba, Brazil)*, 629–33.
- Smith, Ray. 2013. "History of the Tesseract OCR Engine: What Worked and What Didn't." *Proceedings SPIE, Document Recognition and Retrieval*, Vol. 8658, 1–12.
- Skorinkin, Daniil, and Boris Orekhov. 2023. "Hacking Stylometry with Multiple Voices: Imaginary Writers Can Override Authorial Signal in Delta." *Digital Scholarship in the Humanities* 38 (3): 1247–66.

- Tempestt, Neal, Sundararajan Kalaivani, Fatima Aneez, Yan Yiming, Xiang Yingfei, and Damon Woodard. 2018. “Surveying Stylometry Techniques and Applications.” *ACM Computing Surveys* 50 (6): 36.
- van Dalen-Oskam, Karina, and Joris van Zundert. 2007. “Delta for Middle Dutch – Author and Copyist Distinction in Walewein.” *Literary and Linguistic Computing* 22 (3): 345–62.