



HAL
open science

Transcribing Medieval Manuscripts for Machine Learning

Estelle Gueville, David Joseph Wrisley

► **To cite this version:**

Estelle Gueville, David Joseph Wrisley. Transcribing Medieval Manuscripts for Machine Learning. 2022. halshs-03725166v1

HAL Id: halshs-03725166

<https://shs.hal.science/halshs-03725166v1>

Preprint submitted on 15 Jul 2022 (v1), last revised 20 Sep 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Transcribing Medieval Manuscripts for Machine Learning

Keywords

Paris Bible; Latin Bible; handwritten text recognition (HTR); Thirteenth Century Europe; bias; transcription norms; computational text analysis

Authors:

- Estelle Guéville, Yale University, USA: estelle.gueville@yale.edu
- David Joseph Wisley, New York University Abu Dhabi, UAE: djw12@nyu.edu

Introduction

In the early twentieth century, many scholars focused on the preparation of editions and translations of texts previously available only to the few specialists able to read archaic hands and privileged enough to travel to work in person with them in manuscript. Valuable scholarship in its own right, the preparation of these editions and translations for particular texts deemed important enough to justify the effort and time, laid the foundation for generations of scholarship in medieval studies. On the other hand, for many materials in historical archival collections—including already digitised collections—medievalists have only had the time to create partial transcriptions, if any at all. Access to textual material from the medieval period has increased greatly in recent years with digitisation, and we are able to imagine many new research projects in decades to come. What challenges do new frontiers of automation in the archives raise with respect to medieval studies and in particular to the ways we transcribe? In this article, we argue that if medievalists hope to pursue the kinds of analysis that goes on in advanced computational research, we will need new kinds of transcriptions, intentionally theorized not only for human reading, but also for machine processing. We already have mature methods for remediating generations of editions of medieval works such as Optical Character Recognition (OCR), but we can ask ourselves if these are the kinds of text we want to use for future computational analysis. We suggest instead that one way forward is by going back to the scriptorium.

Practices of Transcribing Medieval Manuscripts: a Very Short Historiography

In this section, we give a brief overview of different ways that editors and publishers of medieval texts have treated the question of the difference between the writing systems that we typically use today and those that are found in manuscripts. It is not meant to be a full assessment of historical trends, but rather a way of situating our discussion of transcription. We frame that discussion by referring to work done specifically with thirteenth-century Latin Bibles, but we trust that our contextualized discussion of transcription will benefit other use cases for communities who may be considering automatic forms of text creation from manuscript.

a. Historicizing Normalisation

A transcription of an old text is both a theoretical and a practical endeavour. We all have inherited multiple methodologies for transcribing, but machine learning systems for handwritten text recognition (HTR) such as Transkribus, eScriptorium and others that will no doubt emerge in coming years foreground three main issues related to transcription. First, their emergence emphasises the question of normalisation as a historically contingent and changing category. Second, their rise in popularity foregrounds the necessity for anticipating how target transcriptions will be used in research before the transcription begins. Third, we are confronted by the question of how "general" HTR models emerging in the digital GLAM sector, that is, general-purpose models for large amounts of text, will change modes of analysis and interpretation in the humanities (Hodel *et al.*, 2021).

So, what are some of the ways that we implicitly or explicitly normalise texts when we transcribe them? Scholars working on a particular source base might have a given set of transcription norms inherited from a publisher or a philological education that encourage the normalisation of letter forms such as i/j or u/v or impose specific rules on capitalization or spacing. We should not forget that normalisation has itself become a norm; editors of the first editions and incunabula in the fifteenth century often proposed versions of a text much closer to the original manuscript than later scholars have. As they strove to reproduce medieval manuscripts the way they were, they maintained many of their features, including not only columns, running titles or rubrics but also special letter forms and abbreviations, as seen in Figure 1. Special letter types that included these abbreviations were created to print the incunabula, amongst them, what we now call the macron (represented in Unicode as 0304) but also several others (; ; etc.). These incunabula also maintained the distinction between normal and long s (s/l), normal and insular d (d/), normal and rotund r (r/), a differentiation that was usually lost in modern editions of manuscripts. It is safe to say the transition to print culture eliminated the need for abbreviations and different letter forms is a false assumption. In later centuries, normalisation also followed the period norms including the use of the long s throughout the eighteenth century (Attar, 2010).

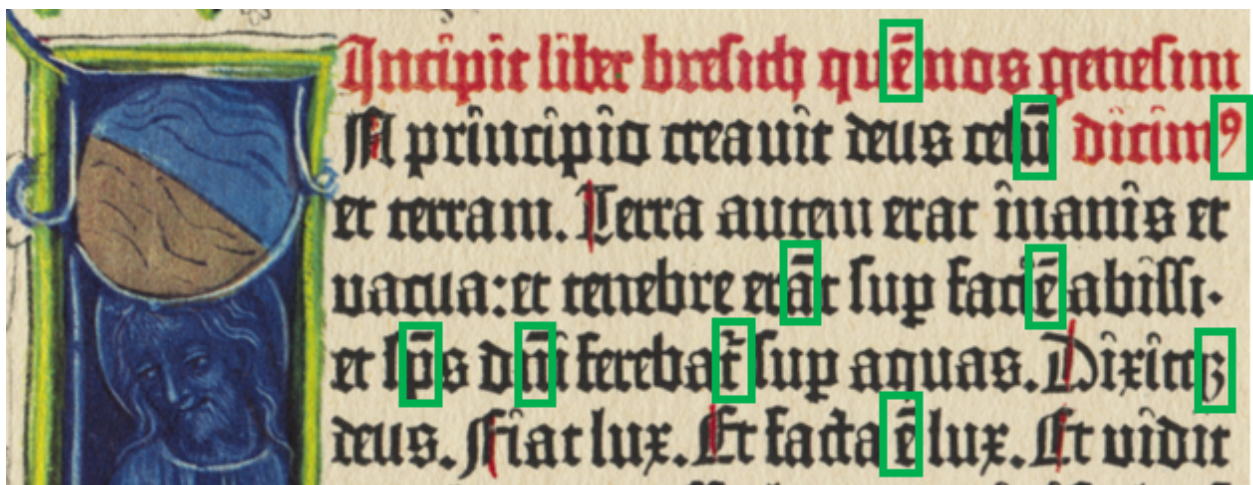

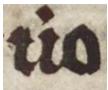





Figure 1: An image from an incunabulum Gutenberg Bible illustrating printed versions of special letter forms and abbreviations (some of which are highlighted by the green boxes). Source: Beginning of book Genesis in 42-line Gutenberg bible, fol. 5r, vol. 1, Staatsbibliothek Berlin 259,1454/55.

But what kinds of transcriptions do we find in circulation today? Scholars distinguish between normalised transcriptions, semi-diplomatic, and diplomatic ones, although often each transcriber defines their own set of rules, leading to a large variety in transcription norms. In Table 1, we give examples of transcriptions of each type, using a corpus of Latin Bibles we are studying. The first column (normalised) chooses to normalise all letter forms, capitalisation and spacing, silently expanding abbreviations, correcting the text where the transcriber feels like it is required, and replacing unfamiliar letters with rough equivalents from the Roman alphabet. The text it produces is very easy to read for anyone who is not familiar with palaeography, since it conforms to modern-day literacies and it allows some research questions such as the literary style, occurrences of words, or comparisons of texts to be made with limited intervention.

If, on the other hand, as medievalists, we want to use language resources (LR) and natural language processing (NLP) techniques, especially with the new capacities of text creation facilitated by HTR, an inevitable conflict between normalised versions of texts (the norm for corpora) and document-level transcriptions lies on the horizon (Piotrowski, 2012). We would like to suggest that the notion of study corpora is intertwined with the idea of normalization, a set of transcription norms that has been chosen and perfected in the majority of cases since the 19th century. At the same time, as researchers, we also see assembling these corpora as built based on a significant loss of information.

| In the manuscript Louvre Abu Dhabi, LAD 2013.051 | Normalised | Semi-diplomatic | Diplomatic |
|---|--------------|---------------------|------------|
|  | super faciem | <i>fuper faciem</i> | fū faciē |
|  | vero | <i>uero</i> | úo |
|  | respondit | <i>respondit</i> | ꝛndit |
|  | congregentur | <i>Congregentur</i> | Conggent̃ |
|  | Deus | euf | euf |

| | | | |
|---|-------------|--------------------------|------------------------|
|  | producerunt | p odu ^x erunt | p odu ^x unt |
|---|-------------|--------------------------|------------------------|

Table 1: A table illustrating sample words from a Parisian Bible at the Louvre Abu Dhabi, LAD 2013.051 and sample transcriptions: normalised, semi-diplomatic and diplomatic.

The second column (semi-diplomatic) usually keeps special letter forms as they are written in the text, making the difference between u/v or s/f, preserving the original capitalisation or spacing as much as possible. It expands the abbreviations but usually indicates their purposeful expansion by the use of italics (or other methods such as underlining). It is a hybrid model from which we can understand editorial intervention in human reading, but is not amenable to plain text processing approaches lacking italics to encode such interventions (Widner, 2017). The last one - and the least commonly used - is the diplomatic transcription that seeks to preserve as much information as possible from the original manuscript. Similar to practices in epigraphic transcription, this understanding of diplomatic transcription identifies written characters, linking them to Unicode "without spaces, punctuation or diacritics (unless these are in the source document), and without restoring lacunae or expanding abbreviations" (Bodard, 2021).

While there is a tradition of editing diplomatically, most editions of medieval texts are not diplomatic. For dealing with the many letter forms in medieval manuscripts that differ from modern alphabets (long s, insular d, rotund r, etc.), editorial traditions differ about what to do with them as well. For the most part, there has been a rough convergence on an ASCII or ASCII-like character set for representing necessary characters. In the case of medieval manuscripts, transcribers employ abbreviations that have been the object of a considerable amount of study in palaeography, which initiatives such as the Medieval Unicode Font Initiative (MUFI) have rallied to describe (MUFI, 2015). In our opinion, one of the key weaknesses of the MUFI, however, is the inability of all of these Unicode characters to display correctly across multiple devices, impeding both plain text workflows and "what you see is what you get" (WYSIWYG) editors, so common in academic life.

Transcription norms have been designed as a way of ensuring the consistency of a critical text to be set down in print technologies which we associate with rigor, orthography, etc. They ensure the accessibility of the critical text: for modern literacies and expectations of scholars and students alike. They give scholars access to a deeper understanding of the textual tradition, while ensuring ease of professional reading (Siemens *et al.*, 2009). They also influence the kinds of research we might do or even limit what we understand as possible. We suggest that normalization of textual features in the print-centered mentality of transcription entails a loss of information that could very well be useful, even central, to future digital research in medieval studies.

It goes without saying that the digital turn in the historical humanities has multiplied the different ways that we can access, create and use text. For some years, it has been possible to read and compare digitized medieval manuscripts at a distance and on a screen, a process which the International Image Interoperability Framework (IIIF) has sped up significantly. Access is not, however, only a question of the delivery of archival materials to far-flung parts of the world. The availability of digitized images of manuscript materials also allows us to access the text within them by creating what used to be called "machine-readable" text, now sometimes called digital text, machine processable text, automatic transcription, or simply transcription (where automation is assumed). By accessing text, we mean more

than just having a digital facsimile on the screen for human reading, but creating a transcription and putting it to some specific use.

The history of normalization of transcriptions is heavily linked to the evolving research landscape. While the first editions tended to preserve the micro features of manuscripts, the tendency to normalize transcriptions of medieval texts happened with the development of literary research. Because scholars were interested in the text "itself," for its meaning or for comparison of literary traditions from one manuscript to another, it made sense to normalize the spelling and make editions of medieval manuscripts easier to read and study by a larger community. This is what prevailed for centuries, up until the recent development of new technologies. Scholars are now interested not only on the "substantives" as defined by Greg (1950-51), that is the words themselves and their meanings, but also on the "accidentals", including the spelling, punctuation, or word-division. This change in research paradigm which had an impact on the transcribing schemes was also allowed by the evolving technologies available to the medievalists.

b. Changing Technologies and Levels of Transcription

Let us return to the question of a diplomatic transcription and ways it has been used in the field as technologies for medieval studies have developed. Scholars of pre-modern cultures often speak of a diplomatic transcription or a diplomatic edition, in which characters are recorded as they appear with minimal to no editorial intervention or interpretation. Debates around transcription have focused on varieties of transcription, allowing for different amounts of scribal information to be captured. We see these debates about diplomatic transcription as connected to late 20th century critical practices of understanding documents within the context of their production and recopying that emerged in the debates around New Philology and before, especially with the possibility of the delivery of digital versions of manuscripts in the form of images on the web (Rigg, 1983).

It might be argued that there are as many forms of diplomatic transcription practices as there are textual traditions. Critics assign different terminology to the "levels of transcription" (Robinson & Solopova, 1993), which sometimes overlap but also create subtle distinctions between transcription styles based on different editorial practices. Our purpose in raising this point is not to decide once and for all how diplomatic transcriptions should be done, but rather to suggest that one's choices for encoding must arise not only from the specific textual scenarios at hand, but also from the ways that one wants to use such text downstream. If the use for such text is a screen-based, documentary digital edition for scholarly reading, perhaps the transcription can be as specific to the textual tradition as desired and as specialized as the audience intended for the work. On the other hand, if the goal is to work with contemporary computational approaches to text, the consistency, as well as the concision of transcription norms, becomes all the more important as decisions made in both ground truth creation and subsequent model retraining embed bias within machine learning.

Let us consider some of the implicit assumptions for diplomatically transcribed texts in some projects. Many, if not most, scholars make the distinction between semi-diplomatic and diplomatic transcription though, the former usually expanding the abbreviations and the latter transcribing the text exactly as it appears on the page, but others create more granular distinctions. In "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue" for the Canterbury Tales project (1993), Peter Robinson and Elizabeth Solopova defined four different levels of transcription: **regularized** ("all manuscript spellings

are regularized to a particular norm, perhaps the spelling of a manuscript considered authoritative”); **graphemic** (“every manuscript spelling is preserved (as: ‘she’, ‘sche’) without distinction of separate letter forms as in a graphetic transcription”); **graphetic** (“every distinct letter-type is distinguished (as: r ‘short’ is transcribed apart from r ‘round’ and r ‘long descender’, etc.”); and **graphic** (“every mark in the manuscript, every space, is represented in the transcription, even to the point of decomposition of letter forms into discrete marks”). It is useful to note that Robinson and Solopova were limited by the technology they had access to which in turn impacted the kind of research questions they were able to tackle. If given the kind of resources we have nowadays, namely artificial intelligence and HTR, they would have probably chosen another transcription level, graphetic instead of graphemic, perhaps with characteristics from the graphic level.

In the Paris Bible Project, the transcriptions we have made of Latin Bibles could be categorized as a transcription using characteristics from both the graphic and the graphetic levels: we make the distinction between every letter form, we represent abbreviations, capitals, spaces, punctuation as much as we can but we do not represent every single difference in the letter forms (longer or smaller vertical strokes, length, breadth or weight).¹

In the “Menota” (Medieval Nordic Text Archive) project for machine-readable editions of medieval Nordic texts, they distinguish three levels of transcription: normalized, diplomatic, and facsimile. They describe them as follows:

- “A **facsimile** level (<me:fac>): A letter-by-letter transcription with a selection of palaeographic characteristics and the retention of abbreviations as in the manuscript.
- A **diplomatic** level (<me:dipl>): A letter-by-letter transcription with a small selection of palaeographic features and the expansion and identification of abbreviations.
- A **normalised** level (<me:norm>): A transcription in normalised orthography.”

According to this classification, the transcriptions of Paris Bibles we have made would be described as “facsimile,” although we do not encode all the properties they describe, e.g. unclear readings, erased and/or corrected text, initials, and *litterae notabiliores*, or headings. In the case of the Menota project, the diplomatic level is described as an accurate transcription letter by letter in which the abbreviations are expanded and the number of paleographic features reduced. As they explain, “a diplomatic transcription, as defined here, thus requires more editorial intervention than the facsimile transcription in the form of the interpretation of abbreviations and the normalisation of allographic variation.” The resulting transcription will thus be more readable than the facsimile transcription, the latter which, as suggested above, is perhaps more intended for computational analysis than for scholarly reading.

The two examples described above are representative of different ways digital projects in medieval studies have approached complex transcription. At one end of the spectrum, the Canterbury Tales project uses transcriptions which may be called strictly diplomatic, in which every feature which may reasonably be reproduced in print is retained, not only spelling and punctuation, but also capitalization, word division, and variant letter forms. The layout of the page is also retained. Any abbreviations in the text will not be expanded, and, in the strictest diplomatic transcriptions, apparent slips of the pen will remain

¹ For more information about the Paris Bible Project, see the project site: <https://parisbible.github.io/>. One of the project’s goals is to study countable micro-features in extant Latin Bibles in the world as a way of performing predictive analysis about the copying and scribal habits.

uncorrected. Such editions are often so close to the originals as to be too complex for the nonspecialist reader, or in any case no easier to read than the originals. At the opposite end of the spectrum in the Menota project, there are fully modernized transcriptions, where the substantives (Greg, 1950-51) are retained but everything else is brought up to date, in some cases to such an extent as to make it questionable whether they are to be regarded as transcriptions at all. In between these two extremes, a number of levels may be distinguished — ‘semi-diplomatic’, ‘semi-normalised’, etc. — depending on how the accidents of the original are dealt with.

While it is true that development and use of TEI-XML for the purpose of scholarly editions most definitely had an impact on the way scholars transcribe multiple layers of texts (Driscoll, 2006), we believe that a near future of medieval studies will include significantly more automatic transcription than at present. It is not so much that the latter will replace the former, but automation will make the latter more scalable and customizable and will open up intermediary forms of textual exploration between the browsing the digitised manuscript library and preparing a full edition. In our opinion, when the question of automation is combined with transcription, it is only logical that dealing with different layers of transcription will be facilitated, opening many new questions and challenges in the world of text creation and analysis.

The most common use for HTR-generated text at present in the archival community seems to be searchability of archival documents (Stutzmann *et al.*, 2018), although the critical literature describing the use of such technology is expanding quickly (Nockels *et al.*, 2022). There are, of course, other possibilities when we move closer to traditions of complex analysis and interpretation of texts well known in medieval studies. For example, one might want to have an automatic transcription as a draft baseline for creating a new documentary edition, or creating a diplomatic layer for a new critical edition. Likewise, an unedited, but searchable transcribed text could be used for semantic annotation or for the purposes of genetic critical analysis. In a complex textual tradition, any number of transcribed witnesses might be made for comparison, alignment or higher-level analysis (Jänicke and Wrisley, 2017). Each of these end goals is facilitated by transcription, but the nature and norms of that transcription impact the extent to which we can accomplish our "scholarly primitives" with ease. In fact, what is called "diplomatic transcription" adds new kinds of information to corpora specific to the scholarly questions underlying them. In an age of rapidly developing computer vision technologies it may be time to re-theorize the diplomatic transcription, or replace the term altogether. It is quite possible that in a global community of medievalists in only a few years that we end up with a mass of automatically transcribed text that is actually very difficult to compare computationally.

In the creation of ground truth data for training HTR models, there are many choices to be made and advice is generally quite vague: *transcribe as closely as possible to what we see*. It is important to qualify the expression "what we see," because sometimes a transcriber is confronted with a passage in which our knowledge of the ways that the platform tends to react changes the way that we transcribe. The socio-technical elements of ground truth creation have been described by others, so we will not rehearse them here (Alpert-Abrams, 2016; Cordell and Smith, 2018). How can we be sure that complex modes of transcription–encodings in their own right—are machine processable and that they do not interfere with basic downstream processes, such as tokenization and word counting? Our own approach to diplomatic transcription is linked to purpose, and yet we are aware that our purpose of transcribing different copies of what are essentially the same textual tradition will necessarily not be shared by others. How can we

mitigate these problems from the beginning in the design of a HTR-enabled text creation project, especially if our HTR models are made publicly available and will be used and adapted by others?

Designing Corpora for Transcription

a. Transcription as Encoding

Transcribing for modern readers so they can read the text without difficulty and transcribing for a machine are two very different tasks. Computational linguists have been calling for the "representation of manuscript reality" in medieval corpora for some time (Honkapohja *et al.*, 2009) by encoding linguistic, palaeographic, and codicological features in digital editions. This approach, of course, sees value in editing pre-modern texts, but wishes for them to be available in "unadulterated form" so that their non-normalised complexity can be also used for research purposes. Creating an automated transcription of a manuscript is not the same as digitizing it, rather it is creating an imperfect representation of it--with all the limitations of any computational model--in order to be able to analyze that text through specific lenses. Transcribing for a machine, however, does not preclude the eventual editing of works, but let us speak first about creating actionable transcriptions for computational reading.

The abbreviation in medieval studies is typically framed as a skill of medieval literacy (and a medievalist's literacy) in order to be able to read in manuscript. Although there have been some quantitative studies about abbreviations in Latin and vernacular languages (Bozzolo *et al.*, 1990; Hasenohr, 2002; Römer, 1997), they are usually something to learn, understand, decode and then encode or uncollapse, usually when making a transcription or an edition (Honkapohja, 2013; Honkapohja, 2018). Research tools exist for understanding them, such as Capelli's *Lexicon Abbreviatarum*, the famous white reference book on many medievalists' shelves also available in digitized form (Ad fontes; Abbreviationes Online). In thinking through our research process of transcribing medieval manuscripts using Transkribus, it occurred to us that most researchers do not use abbreviations as features that can be useful in and of themselves but as a "problem" to resolve to understand the meaning of the text, even to mark as an expansion in a critical edition to indicate how the editor has interpreted the abbreviation.

We believe that theoretical potential on working with diplomatic transcriptions is different, not as a hermeneutical act--understanding the text as a way of eliminating reading text as it is written in manuscript--but instead situating textual analysis in a context of what appears in manuscripts and with the help of computational tools. In the Paris Bible Project, we are interested in countable micro-features in copies of the many Paris Bibles extant in the world (Herrmann *et al.*, 2015) as a way of performing predictive analysis about the copying and scribal habits of such Bibles. For designing a transliteration scheme, turning to Capelli for examples of the most common abbreviations and letter forms in particular manuscripts was not particularly helpful. It is important to remember that Capelli is an all-purpose reference work that aspires to spatiotemporal breadth as well as examples of abbreviations stemming from many medieval genres. In our case, we are working with one domain and a relatively constrained space and time from which they emerge. We encountered similar issues working with MUF1, including the fact that most of its character set does not appear in our manuscript, most of the Unicode symbols do not visualize in text editors and in the transcription window of Transkribus. In sum, if we consider our

transcription system as a form of encoding medieval scribal data, we do not need a totalizing description of medieval writing systems, but rather a pragmatic one corresponding to the realia of our corpus of manuscripts.

To train an HTR model for the hands of Paris Bibles, we first needed to create a transcription of a handful of folios. That transcription had to represent the "manuscript reality" in a one-to-one relationship. It is quite possible to use an expanded abbreviation in an HTR transcription as the machine learning process is language-independent. If *faciē* were transcribed as *faciem* the model could eventually be trained, but it would limit our ability to distinguish between *facī*, *faciē*, and *faciem*--if the three forms happen to be present. To do so, we identified about 40 special characters used as abbreviations: superscript characters which are placed on top of letters, the so-called "combining letters" of Unicode (̄; ́; ̂; ̃; ̄; etc.), some special characters (; ; ; 7; etc.) and special letter forms (f; ;) to distinguish from their common form (s, d, and r); and finally; superscript letters (ⁱ; ^c; ^m; ^s; etc.). The first and the third groups can be found on many letters and we discover new possibilities on every page we transcribe. Had we used the MUFI, it would have given hundreds of possibilities whereas we work with about only forty. We opted for a simple, adaptable Unicode solution that is pragmatic given the HTR system, but also that outputs text that is easily visualized in plain text format. Of course, there were slight paleographic variations between specific letter forms and in the placement of punctuation, a fact that became particularly apparent as we moved across different manuscripts and different contexts for manuscript production. This fact did not lead to the multiplication of new forms of encoding, but rather sets of problems which we resolved in the project guidelines.

b. Basic Principles of Transcription for HTR

Each project, each manuscript, each hand being different, the list of possible Unicode characters to be used for transcription can evolve and change accordingly. Our suggestion is that there is not one definitive list that can be used and applied to any project, and especially to all projects, rather a quite large number of possibilities, which must be project-specific and aligned with project objectives. To establish a specific Unicode list, as for any transcription process, setting principles to follow is fundamental: what do we transcribe and how? How do we prioritize the principles? How do we encode variance, exceptions, and aesthetic scribal habits? In what follows, we outline five basic principles for transcribing for machine learning which have emerged at this stage of our research.

1. Although transcribing for machine learning is fundamentally an interpretative activity, the **first principle** to abide by should be: the transcription must be as close as possible as what you see in the manuscript, even if this is not enough to render all the variety and inconsistencies. If there is a basic character in Unicode which corresponds to what you see, and that letter exhibits no significant variance across your document which matches your research problem, there is no reason to opt for a more complex character encoding.
2. Since there is always the possibility of variance, the **second principle** is that it is useful to have a preliminary "scan" of the document you want to transcribe, or through samples of the corpus you will be working with, before beginning transcription. A first pass of transcription allows you to create a working list of Unicode characters.

3. Since there is inevitable variety in hands, the **third principle** is to take care when attempting to encode in maximal granularity the "aesthetic" quality to some graphemes that we don't want to reproduce (in our case, this meant spacing, the letters v and p), or a variance in the placement certain abbreviations (for example, the macron) which create too great a variety of rare encoding choices or difficulty in ordering the characters in the transcription.
4. Since we are creating machine-readable transcriptions for the purpose of computational analysis, the **fourth principle** is not to choose Unicode characters that will not display in regular text editors (i.e., MUFI green letters) or other platforms.
5. The **fifth principle** is not to make design decisions that will be undone by common NLP practices (lower casing, tokenization, removing punctuation) for working with unstructured text.
6. In the case of contradictory decisions, we add a coda to our five principles: there is a need to prioritize the principles.
 - a. For abbreviations (such as the macron), we made an arbitrary decision to consider what letter the abbreviation replaces, rather than where it is placed. Let's take the example of the word "bien". In the manuscript BnF français 24428, the scribe wrote it in ways that can be transcribed either biē or bīe, the macron being often written on top of both letters, right in the middle. That is to say that transcription is never really divorced from an interpretation of what was meant by the scribe, even though we do not normalise.
 - b. In the case of spaces and special letter forms, a larger sample should be consulted. A single occurrence is an exception not a principle and doesn't reflect particular features indicating scribal practices. How you decide on that larger sample is a function of the scope of any particular project. Do not create a new Unicode or a spacing decision based on one example, but rather on a larger sample.

These principles are not universal ones, and as we have found they are enriched by work across different domains, periods of time and languages. The logical conclusion of these principles lies in the fact that a general model for transcribing medieval Latin, or French or Arabic is unlikely, but rather a variety of sub-models is a desirable goal for a language-specific community of scholars in digital manuscript studies.

c. Building a Dataset of Manuscripts for Transcription

In this section we would like to offer some practical suggestions for thinking about building such a dataset of transcriptions from one or more manuscripts. The most obvious reason for working with manuscript-level transcriptions is that there is something about the scribal behavior in the manuscript(s) which is worthy of critical attention and that can be detected from the text. Purely paleographic variation would not be served well by HTR, as its goal is to convert a text (with paleographic variation) into machine readable text and only very different letter forms would be thus translated. Another key point when considering automatic transcription of manuscripts is that our ability to build a functioning HTR model is usually based on the manual transcription of a few dozen pages of text (a few thousand words). This step of the process can be a very time-consuming one in medieval languages and any project should budget enough resources to get started. If existing transcriptions or an edition exist, it is important to remember that documentary style editions are better than composite critical editions, and they will have to be un-normalized to be used as ground truth for model training.

One can most certainly work with transcription in a single manuscript, but the majority of the codex, or the part of the codex of interest, should be in the same hand and should be long enough to justify the startup time of preparing the training data. If you do not need a long portion of the manuscript, or multiple manuscripts, transcribed, it is probably best simply to transcribe by hand. It is important to remember that when applying one HTR model trained on one hand to a different hand, the retraining process can be significant. The same can be said for adapting a model from one scholar's project to another. An example might illustrate this last point best. In Table 2, at far left we illustrate a normalized Vulgate passage from the beginning of 3 Kings, the medieval designation for what is known as 1 Kings today. Using a public model in Transkribus, we can transcribe the second column with minimal effort after layout analysis. In the third and fourth columns are visualized transcriptions after correcting a set of sample pages and retraining the model. With time and effort, there is a gradual, although not perfect, convergence from one project's model with specific purposes to another's.

| Normalized Vulgate | Gothic_Book_Scripts_XIII-XV_M4 (Hodel) | LAD 1.0 (Paris Bible Project) | LAD 1.3 (Paris Bible Project) |
|---|---|---|--|
| Et rex David senuerat, habebatque aetatis plurimos dies: cumque operiretur vestibus, non calefiebat. | e rex dauid senuerat habebatg letatis plit umos dies Cumaz opt rireturu stibz no ca letiebat • | Croc aui scnuerat habebatq ctatis plurimos ics Cumq opí ructur u- stib nō ca- sciēbāt. | Cux aui sciucrat habebatq letatis plurimos ies Cumq opi riretur u- stib nō ca- lელიebat. |
| ² Dixerunt ergo ei servi sui: Quæramus domino nostro regi adolescentulam virginem, et stet coram rege, et foueat eum, dormiatque in sinu suo, et calefaciat dominum nostrum regem. | D veerunt ergo ei sern sii • Queramus dno noo rege adolescentulam iurgine • e stet coram rege • et foueat eu • dormiatqz in sinu suo et calefa ciat dum nrm regem • | Dfǣgtunt cigo q sip sui. Qucramus ño npo rigi a olescentulam uuginē; 7 stcī co am rege; ct soucat eū. o miatq in sinū suo ct calcta cat nm nrm regem. | Oixesunt ergo ci seciu- sui. Queramus ño nño regi- a olescentulam uirginē; 7 stet co am rege; et foueat eū. o miatq in sum suo et calefa- ciat ñm ntñ rogem. |
| ³ Quæsierunt igitur adolescentulam speciosam in omnibus finibus Israel | Quest erunt ergo adolescentulam speciosam • i omib fuubz isrl | Ouesi crunt ergo a olqscntulam spocosam. ī onmib siaib isrl | Quesi erunt ergo a olescentulam speciosam. ī oñib fñib isrl |

Table 2: Sample transcriptions of the beginning of 3 Kings (modern, 1 Kings) from Free Library of Philadelphia, Rare Book Department, Lewis E M 063:01-31, fol 63r, using three different models.

Since medievalists working on an edition or on one particular text are often working with a number of different witnesses which exist in different countries and which were copied during different periods, the question can arise whether HTR is an appropriate method for their transcription. A team needs to ask itself if the different witnesses are found in different hands, does the time allotted permit training a model from scratch or adapting a separate model to each one of the various hands? Additional question about

access to resources are also worth asking. Is there access to a downloadable scan of the document or a IIIF manifest from a digital manuscript library? Is the document scanned in its entirety or is the scan partial? Is the quality of the digitization satisfactory for your purposes? What are the laws around the reuse of images across institutions and countries? Finally, since the images will sit on external servers, does the holding library allow you to upload their images for such research purposes?

Defining the scope of the project, the type and number of manuscripts constitutes the first step of the process, but it is one which needs to be carried out with a good understanding of what HTR is able to achieve. In the Paris Bible Project, gathering digitized manuscripts, assembling and labelling data was our first struggle, largely because of the way manuscripts are preserved and made discoverable. The history of collections and the way institutions describe objects, their approach to digitization, and their policies towards accessibility and reusability were all significant hurdles.

Not all medieval manuscripts, especially Paris Bibles, are digitized, or even discoverable. Today, Paris Bibles are found throughout the world and are a symbolic object of historical collections. Paris Bibles were originally objects intended for individual use—for studying, teaching or preaching purposes—but they have not traditionally made up a single medieval collection. To our knowledge, there have not been collections of fully digitized ones either. However, today, most cultural institutions with a medieval collection, whether museums or libraries and even private collectors possess at least one Paris Bible. Because they represent an important moment in the history of book production and in the history of devotion, preaching and teaching (Light, 1994; Light, 2012; Ruzzier, 2010), they have become a “must-have” for most collections, prestige objects that are, in some cases, among the very first numbered manuscripts. In other collections, they can be notoriously undiscoverable, considered as a “common” object. Even when digitization exists, the quality is unequal from copy to copy, from one institution to another. Some institutions also choose to make the digitisations public and downloadable (or not). We also found ourselves constrained by competing or contradictory library catalogue descriptions. The term “Paris Bible” is not universally accepted, and such documents can only be found under their specific terms in various European languages: *Biblia sacra*, *Pariser Normbibel*, *Bibbia dell’ università*, *Universitetsbibel*, or even under the general denomination “Bible”. Given this diversity of nomenclature, we have relied on manual and visual identification, using discipline-specific rich metadata about these objects, primarily in European and North American collections which adopted early integral digitization.

The challenges described above in the constitution of one's corpus within the constraints of what HTR does very well are important to consider when designing a research project involving transcription of medieval manuscripts. What makes the Paris Bible a strong candidate for HTR transcriptions is the fact that they are usually written in a somewhat standard Gothic hand. A project which wanted to use HTR to transcribe the manuscripts of a large textual tradition such as the *Roman de la rose* or the *Speculum historiale* would not enjoy the same success, due to the inevitability of difference in hand across the various witnesses.

d. Problems of Metadata

Metadata are important when it comes to transcriptions made from manuscript. They are what allow us to group together different artefacts. Working with medieval manuscripts is difficult, however, as we mentioned before, since cultural institutions describe their objects in vastly different ways. The material evidence which can be gleaned from well described manuscripts can be enormously helpful for contextualizing the transcriptions we are making. A specialized library with a significant manuscripts

collection has more chances to have a specialized curator or cataloguer who would provide many codicological and artistic details, including, for example, the justification size, the number of lines, the ruling or presence of catchwords. On the other hand, a museum with a variety of objects and a normalized description process used for ceramics, paintings or manuscripts would be much less specific. Moreover, countries and institutions use different norms for writing dates or locations as well as different languages. Overall, we had to treat the metadata with caution.

The quality of legacy metadata is fundamental to our project: our process is recursive, which means that we carry out transcription, correction, retraining and analysis of different datasets. We needed to start with well-known documented manuscripts which could serve as reference points for understanding the textual tradition. In our specific case, the more that we know about the time and place of the copying of a manuscript and the different hands found in it, the more we are able to "predict" about others. There was an inevitable degree of normalization across different libraries and national traditions which allowed for us to compare like objects. It is very possible that large research projects working with many manuscripts encounter a similar phenomenon. Lastly, but connected to the question of metadata, we needed to make choices regarding the manuscripts used for training purposes. Our model was first trained on one single manuscript, the Bible (LAD 2013.051) from the Louvre Abu Dhabi collection (Guéville, 2021), and not even in its entirety: we used the text of Genesis, part of the Exodus and the books of Matthew and Mark. Prologues, marginal corrections, and other Biblical books have been completely ignored in the first phases of the project, thus creating a potential bias toward specific words contained in these books. We then added a handful of manuscripts using the same texts. Our latest "composite" model is currently based on a dozen manuscripts, which are not representative of all locations, dates and traditions.

Designing HTR Models for Research Questions

a. Training a Model for Transcribing Medieval Manuscripts

The steps to actualising an HTR model for transcribing medieval manuscripts extend beyond understanding what kind of transcription scheme is most appropriate and the availability of digitized manuscripts. The process of training a HTR model involves a non-trivial layout analysis step in which baselines and polygons are identified and with which a given ground truth transcription can be aligned. Sometimes a manuscript is available with a IIIF manifest, but the quality of the image does not allow for automated layout analysis. This situation occurred for us with a well described and well localized Bible produced in Bologna (BnF nouvelles acquisitions latines, 3189) which we were particularly excited about using as a reference manuscript due to its origin in the Italian city and its known copyist. Unfortunately, we believe that the contrast between the color of the ink and parchment was not sufficient to perform adequate layout analysis on all folios.

In some cases, given the research questions of a specific project, full transcriptions are desirable, but not truly required. That is to say that a successful project with HTR transcription of medieval manuscripts depends on a large number of factors and one is almost inevitably required to make a compromise between the availability of specific texts, the specificity of scribal hands, the quality of digitization of the manuscripts and the tasks one would like to carry out with the resulting transcriptions. In modern or contemporary archival transcription projects, often an emphasis is placed on a model being able to transcribe a variety of hands with precision. With medieval transcription projects, the added dimensions of medieval textuality (multi-scribal composition, compilation, rebinding, marginalia, etc.) have to be

taken into account for effective project design. In other words, the more specific your research questions are, or the more complex your sources are, the more problematic the idea of a general model becomes.

In the Paris Bible Project, in order to create an HTR model, we had completed a number of steps: identification of our corpus of folios, normalisation of our metadata, and design of our transcription scheme. We trained our first model in Transkribus, LAD 1.0, based on the public model in Transkribus Gothic_Book_Scripts_XIII-XV_M4 (Hodel) to which we added data from the manuscript kept in the Louvre Abu Dhabi collection (LAD 2013.051). We created eight pages of ground truth, amounting to 588 lines and 3547 words. It cannot be understated that to arrive at the first model of a project, especially when this requires careful transcription from manuscript, is an arduous process. The human effort in the process grows lighter after these initial steps, with requirements of time and energy shifting to correction of the model and assembly of new manuscript samples. We trained two subsequent models, LAD 1.1 and LAD 1.2 with 16 more pages of ground truth (1592 lines and 9632 words in total) from the same manuscript. The characteristics and performance of the various models are summarized in Table 3. In sum, HTR model design is a time-consuming endeavour which connects specific research questions to a general understanding of the performance and characteristics of the HTR system, but also for which the results are somewhat difficult to predict in advance.

| Model | Date | Base_model | Number_of_chapters | Number_of_pages | Number_of_lines | Number_of_words | Number_of_characters | Number_of_symbols | Training CER (%) | Validation CER (%) | Comments | Bias |
|---------|------------|-------------------------|--------------------|-----------------|-----------------|-----------------|----------------------|-------------------|------------------|--------------------|---|---|
| LAD 1.0 | 15/08/2020 | Gothic Books (Hodel) | 50 | 8 | 7 | 1 | 588 | 3547 | 0.27% | 4.52% | Based only on the LAD manuscript (LAD 2013.051). | Bias towards Northern France, second half of mid-13th century. Bias towards Genesis, Exodus, Mark and Matthew |
| LAD 1.1 | 22/08/2020 | LAD bible 1.0 | 50 | 24 | 19 | 5 | 1592 | 9632 | 11.89% | 7.20% | Based only on the LAD manuscript (LAD 2013.051). Same GT as 1.2 with a different base model | <i>Ibid.</i> |
| LAD 1.2 | 22/08/2020 | Charter Scripts (Hodel) | 50 | 24 | 19 | 5 | 1592 | 9632 | 0.62% | 4.14% | Based only on the LAD manuscript (LAD 2013.051). Same GT as 1.1 | <i>Ibid.</i> |

| | | | | | | | | | | | | | |
|---------|------------|----------------------|-----|----|----|---|------|------|---|-------|--------|--|--|
| | | | | | | | | | | | | with a different base model | |
| LAD 1.3 | 26/10/20 | Gothic Books (Hodel) | 100 | 39 | 30 | 9 | 2516 | 1525 | 8 | 0.51% | 3.01% | Based only on the LAD manuscript (LAD 2013.051). | Bias towards Northern France, second half of mid-13th century. |
| PBP 1.0 | 29/06/2021 | LAD bible 1.3 | 50 | 25 | 16 | 9 | 152 | 8840 | | 2.04% | 12.76% | Composite model based on Paris Bibles from around Europe in the 13 th and 14 th centuries. | Bias of books, localisation and dates mitigated. |
| | | | | | | | | | | | | | |

Table 3: Statistics on the HTR models trained in the Paris Bible Project.

b. Creating a Composite Model

Sometimes it proves valuable to combine different sources with different hands into a single model in the hope that the general result will prove more successful across a variety of textual situations. This approach can be useful for limiting the bias of a particular set of manuscripts when it comes to abbreviations or spelling and to avoid the problem known as overfitting, where assumptions about a known dataset are erroneously imposed onto a new unknown dataset. Since our first models LAD 1.0, LAD 1.1, LAD 1.2 and LAD 1.3 were developed using a very limited corpus², we soon realized their limits: they were heavily biased and often we found examples of their common abbreviations reproduced in transcriptions of new manuscripts where they were not present. In order to reflect the diversity of Paris Bibles and limit some of the bias introduced, we decided to increase the size of the dataset and train a new model based on multiple manuscripts, reflecting the multiplicity of traditions, locations and dates of production. What we were aiming for is some kind of "general model" with enough data to be able to extrapolate from one Biblical manuscript to another without overfitting.

To this extent, we constructed a dataset of approximately 450 folios from 24 manuscripts³. These manuscripts were produced mainly in France and England, but we had examples from Italy, Germany and

² LAD 2013.051, to which were added three manuscripts from the BnF collection, Latin 40, Latin 10421 and Smith-Lesouëf 19.

³ The composite dataset were made up of excerpts from Berkeley UCB 12; British Library Additional 50003; British Library Royal 1 D I; BnF Latin 40; BnF Latin 10421; BnF Latin 10428; BnF Latin 14232; BnF latin 14238; BnF SL 19; Trinity Hall Library, ms.22; Cleveland Art Museum 2008.2; New York, Columbia University, Burke Library at Union Theological Seminary, UTS MS 072; Harvard Typ 446; Louvre Abu Dhabi, LAD 2013.051; Free Library of Philadelphia Lewis E31; Free Library of Philadelphia Lewis E32; Free Library of Philadelphia Lewis E37; Free Library of Philadelphia, Lewis EM 063; Library Company of Philadelphia 9; Madrid 12802; Arc Priv 3 Montecassino 2018; Schaffhausen, Stadtbibliothek, Ministerialbibliothek, Min. 6; Swarthmore College BS 75 (partial); University of Pennsylvania Ms 1065.

Spain (see Table 4) and the folios used were taken from as many books of the Bible as possible: 74 books out of 81.

| Origin | Number of mss |
|-------------------|---------------|
| England | 7 |
| France | 6 |
| England or France | 1 |
| Germany | 2 |
| Italy | 3 + 1? |
| Spain | 2 |
| Spain or Italy | 1 |
| Unknown | 1 |
| Total mss | 24 |

Table 4: A list of the number of manuscripts used to construct our "composite" Paris Bible dataset with digitized copies from around Europe.

This composite dataset is a kind of "artificial" Paris Bible. We collected samples of folios from full and partially digitized Paris Bibles with a maximum of variables in mind (book of the Bible, dating and localization of the codex, codex size, etc.). We then trained a new HTR model (PBP 1.0) based on 3 corrected folios from each Bible manuscript. Had the Paris Bibles of the world been mostly IIF compliant, the gesture would be even much easier, as with the HORAE project carried out on Books of Hours (Stutzmann and Chevalier, 2021). Instead, for us aiming to capture a diversity of Latin Bible codices has meant turning not only to copies which are already available in better resourced European and American libraries, but to smaller, under resourced or isolated libraries, and even to legacy databases (such as Mandragore and Digital Scriptorium) which never intended to publish fully digitized codices. As one might expect, the results of this composite model were less than optimal when used to transcribe new unknown texts. What we learned in the process of creating this model is that while there are a number of well dated and localized Paris Bibles in the world, the hands represented by those codices are quite diverse and the three folios we transcribed from each most likely did not provide a large enough sample. In addition the quality of digitization and the ability to carry out layout analysis was a significant hindrance to including the data from southern European libraries.

We can extrapolate from this that each time a new HTR model is developed with a new question in mind, a considerable amount of new data needs to be created, but also that the effort to create a diverse dataset is significant. The step between our initial model based on one manuscript and the composite one was quite large indeed. In summary, we are confronted with a paradoxical situation. On the one hand, we do not see a general model being a strong possibility in the near future, even for a manuscript corpora as "uniform" as Paris Bibles; this means that a number of models for subsets of the corpus is perhaps a better strategy.

On the other hand, the number of digitized manuscripts from all the collections in which we find Paris Bibles is uneven, seemingly precluding the possibility of an extensive number of models.

c. How Much is Enough?

The field of machine learning is one which can be alienating to those with traditional humanistic training, since it introduces notions such as "ground truth," "training data," or even "gold standard" data which have traditionally belonged to the sciences. One other key idea in machine learning is that of "predictive analysis", again not one that is usually integrated into a study of medieval manuscripts. Predictive analysis can be tricky because we know that inaccuracies in ground truth, or biases built into the ground truth, have the potential of creating false conclusions. Two questions which loom in the background of our work, and will be important for a generation of medieval studies interested in transcription from manuscript, are (1) do we have a diversity of digitized data to mitigate against what might be called "collections bias" described above, and (2) how much of this transcription, analysis and interpretation need to be done in order to understand fully enough what we can about any given corpus. In data science, it is recognized that predictive analytics can mean that we do not need to be concerned with getting all the data, but rather the right kinds of data. Will this be possible in medieval studies at present? Or will we need to do targeted digitization to be able to advance? It is also important to know when enough is enough. Incrementally adding more and more data will not necessarily produce better results, but sometimes results will simply plateau. The question we are asking ourselves at present is to what extent this will be the case with Latin Bibles.

Working with Transcriptions

Our paper addresses the larger question of how medievalists have approached the question of transcription when dealing with manuscripts and how those approaches are evolving with the automation afforded by computer vision methods such as handwritten text recognition. A number of important issues about building corpora, creating balanced or even general models for medieval handwriting, even new challenges of error, bias and overfitting of machine learning models come to the fore. It is unlikely that we will be able to correct all the errors of these automated transcriptions with human labor, so we will need to find ways of looking for larger patterns while mitigating imprecision. In this section we will look very briefly at some of the scholarly reasons for having transcriptions, as well as kinds of new questions that become possible when we do. Detailed descriptions about each of these projects will not be given, as it would take us beyond the scope of this article and as research is published elsewhere.

As was mentioned above in section 1b, one of the simplest reasons that transcriptions of a manuscript are desirable outcomes in medieval studies is the question of searchability and keyword indexing. If we think about the ways we are able to search within modern texts such as pdfs or in online databases, "full text" searchability has become a basic expectation of researchers. Funded projects such as HIMANIS (HIstorical MANuscript Indexing for user-controlled Search, <https://himanis.hypotheses.org>) have focused on the voluminous archives such as chancery records and leveraging computer vision techniques to render these sources more friendly to the searching demands of researchers. A similar capability

provided by the READ COOP's Search and Read Interface makes a first step in opening up handwritten archival documents to string and fuzzy searching.

Another important way that transcriptions from manuscript can enable analysis of medieval texts well known for their instability is in the domain of intertextuality. Transcriptions from textual traditions such as the *Roman de la Rose* could offer quite a window onto the *mouvance* of that text and the ways in which scribal culture modified it. In the context of our own work, we have found the equivalent of these intertextual elements in Latin Bibles where the text differs from the Vulgate version. When sequences of strings do not match, cross checking with the Vetus Latin Database (Brepols) has revealed many examples where the language in our corpus of Paris Bibles echoes examples of pre-Vulgate text of the Vetus Latina. These examples which we have only begun to study suggest an intertextuality based on orality and sermoning culture which influences in turn copying practices of Bibles. Of course, a less conservative and more normalising approach to HTR would probably facilitate the computational identification of these variants more efficiently, but the principle of revealing intertextuality remains the same.

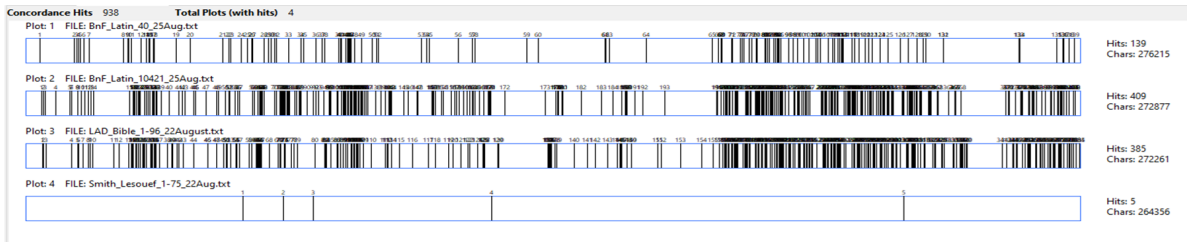
| Vulgate | LAD 2013.051 (Genesis) | VLB (Brepols) |
|--|---|---|
| et recordabor foederis mei vobiscum, et cum omni anima vivente quae carnem vegetat: et non erunt ultra aquae diluvii ad delendum universam carnem. | Et recordabor foederis mei q̄ pepiq̄i uob̄cum ⁊ cum om̄i anima uiuente que carnem uegetat ⁊ n̄ erunt ultra aque diluuii; ad delendam uniuersam carnē. | et recordabor foederis mei, quod pepiq̄i vobiscum (PS-EUCH) |
| Dixitque ei angelus Domini: Revertere ad dominam tuam, et humiliare sub manu illius. | Dixitq̄ ei angl̄s d̄omini. Reūte ad d̄ominam tuam ⁊ humiliare sub manu ipsius | Dixitque angelus Domini Revertere ad dominam tuam, et humiliare sub manu ipsius . (PS-EUCH) |
| Vimque faciebant Lot vehementissime: jamque prope erat ut effringerent fores. | Uimq̄ faciebant loth ueher̄t̄if̄ firme iam p̄pe erat ut ilfrigent fores. | Vimque faciebant Lot vehementissime: jamque prope erat ut infringerent fores. (PS-EUCH) |
| duodecim duces generabit, et faciam illum in gentem magnam. | xii. d̄uces generabit ⁊ faciā illū crefcere in gentē magnam. | duodecim duces generabit et faciam illum crescere in gentem magnum (BREV GOTH) |
| Leua oculos tuos et uide a loco, in quo nunc es, ad aquilonem et meridiem, ad orientem et occidentem. | Leua oculos tuos i oirectū ⁊ uide a loco ī quo nunc es ad aquilonē ⁊ m̄idiem ad orientem ⁊ occidentem. | Gen. XXIII, ubi dicit Dominus ad Abraham: Leua oculos tuos in directum et uide a loco in quo nunc es, ad aquilonem et meridiem et orientem et occidentem. (Nicolas de Aquavilla) |

Table 5: Some examples of where departures from the Vulgate text in Genesis in the Louvre Abu Dhabi Bible (LAD 2013.051) match possible citations in the Vetus Latin database (Brepols).

Another example of the possibility of searching specific words in the transcriptions made from manuscript comes from word counting. Using a concordancing tool, we can visually compare the frequencies of two versions of the Latin lemma *domin-* (meaning "Lord"), namely " n*" and " omin*". In Figure 2, it can be seen that the variant n* is almost never detected in manuscript Smith-Lesouëf 19, whereas it is quite prevalent in the other three manuscripts compared here: BnF Latin 40 and BnF Latin 10421. On the other hand, omin* is quite rare in LAD 2013.051 and very common in the New Testament in Smith-Lesouëf 19.

CONCORDANCE DENSITY PLOTS (ðn* vs ðomin*)

ðn* (Latin 40, Latin 10421, LAD 2013.051, SL19)



ðomin* (Latin 40, Latin 10421, LAD 2013.051, SL19)

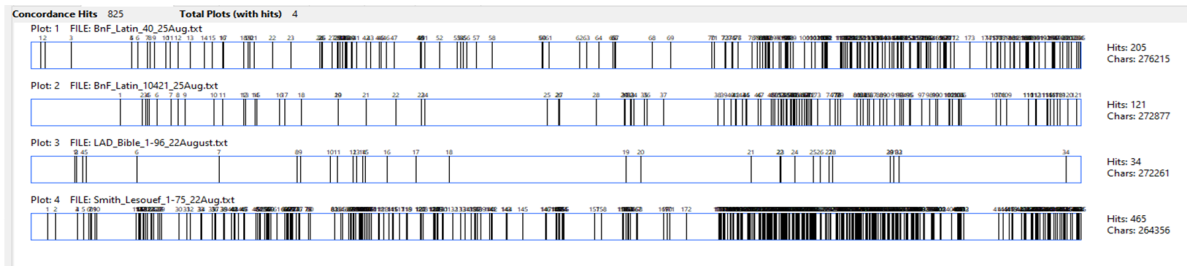


Figure 2: Concordance Density Plots made with AntConc for two strings representing the same lemma (ðn* and ðomin*) for four manuscripts of Latin Bibles (BnF Smith-Lesouëf 19, BnF Latin 40, BnF Latin 10421, LAD 2013.051). Visualized with AntConc.

These differences allow us to look at variance in individual groups of manuscripts for a single feature indicative of a scribal profile, but also contribute to computational methods we can use across many manuscripts and features to perform classification experiments. In this respect, the title of this article "Transcription of Medieval Manuscripts for Machine Learning" not only points to modes of transcription which allow for HTR transcriptions to be produced, but those to the study of those transcriptions using machine learning to identify larger scale patterns. Of course, not all corpora of interest to the medievalist might qualify in size or quality for this kind of analysis, but it does open the door to forms of predictive analysis for some.

For example, in the case of the manuscript Cambridge, Corpus Christi College 49, known to have three different hands, we are able to use the technique known as rolling stylometry to predict computationally the identity of those hands, which we have previously confirmed with our human eye. Significantly, using this method of sequential analysis, we are able to predict the identity of the copyist at any given point of the manuscript with a quite small sample of language, down to less than 1500 words, and this systematically using each and every one of the three hands in all of the segments of the manuscript. As Kestemont (2015) wrote: “superficial textual variations also present important scholarly opportunities, for instance for the identification of scribes or the dialectological analysis of texts.” Several scholars have also discussed the issues caused by data loss linked to the normalisation of transcriptions and the discarding variations that are considered random or trivial. (Driscoll 2006; Kytö et al. 2011; Kopaczky 2011; Rogos 2011; 2012; Stutzman 2014; Lass 2004). This criticism is supported by the fact that there are numerous approaches which have demonstrated the value of scribal and other accidental variation. The method of predictive analysis of the hand of a known scribe demonstrated here for the case of one

manuscript confirms what we know from close visual analysis, but also has some remarkable possibilities when working with other transcriptions, in combination with material philological evidence for example, for gaining a better understanding about the way such such manuscripts were collectively made and transcribed.

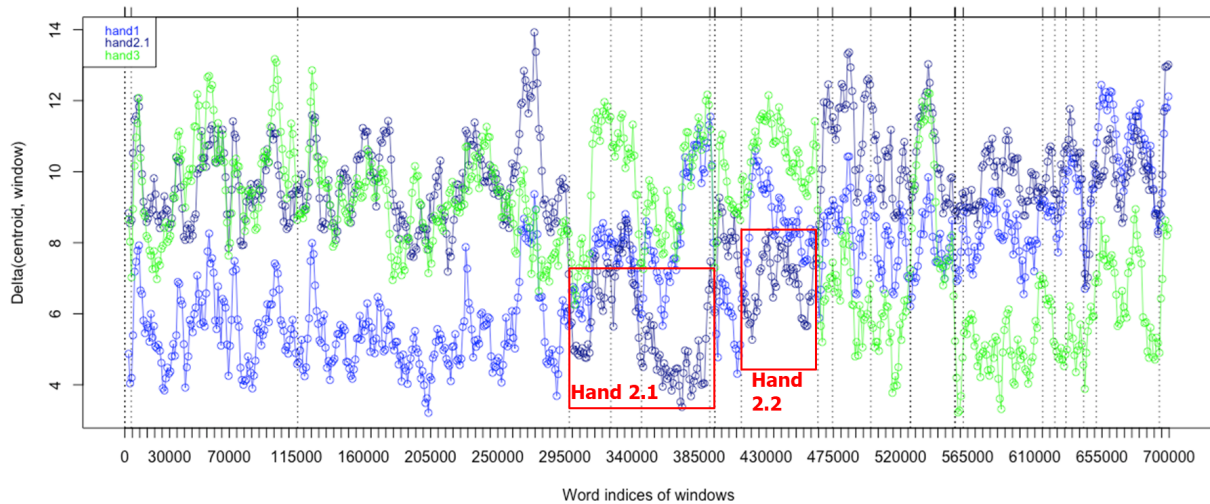


Figure 3: A graphic output of an experiment using the rolling delta method for predictive classification of the scribal hands of a Latin Bible, Cambridge, Corpus Christi College 49. Visualized in R with the Stylo package.

One last example is a very distant analysis. Using the method known as TF-IDF (or Term Frequency - Inverse Document Frequency), we can look at a number of non-normalised transcriptions of segments of manuscripts from the Paris Bible tradition in order to predict dating or localization based on what we know from legacy metadata. TF-IDF, as a classification method, helps us to approximate common words in each transcription, while balancing its importance in other documents. Clusters of transcriptions which share similar common words (and necessarily common abbreviations and words spelled with special letter forms) cluster together, but not with the rest of the corpus. Using a custom word dictionary, what we have found using this method is that there are certain high frequency words with specific spellings (in / ī and elt / ē from top right to bottom left) and (et / 7) from top left to bottom right which are indicative of regions for which we are quite sure of the localization of manuscripts. This tends to suggest that English scribes are more likely to use the Tirolian ampersand, whereas Italian or some French scribes preferred the two letter "et". A similar distinction is made between Catalanian and English scribes' preference for the preposition "in" and French scribes "ī". Such results suggest how we can use non-normalised transcription as an additional layer of data for scribal attribution, dating and localization. This added layer is promising in as much as it will help us build a handlist of high-level features distinctive of specific regions, a useful criterion for a "uniform" manuscript tradition.

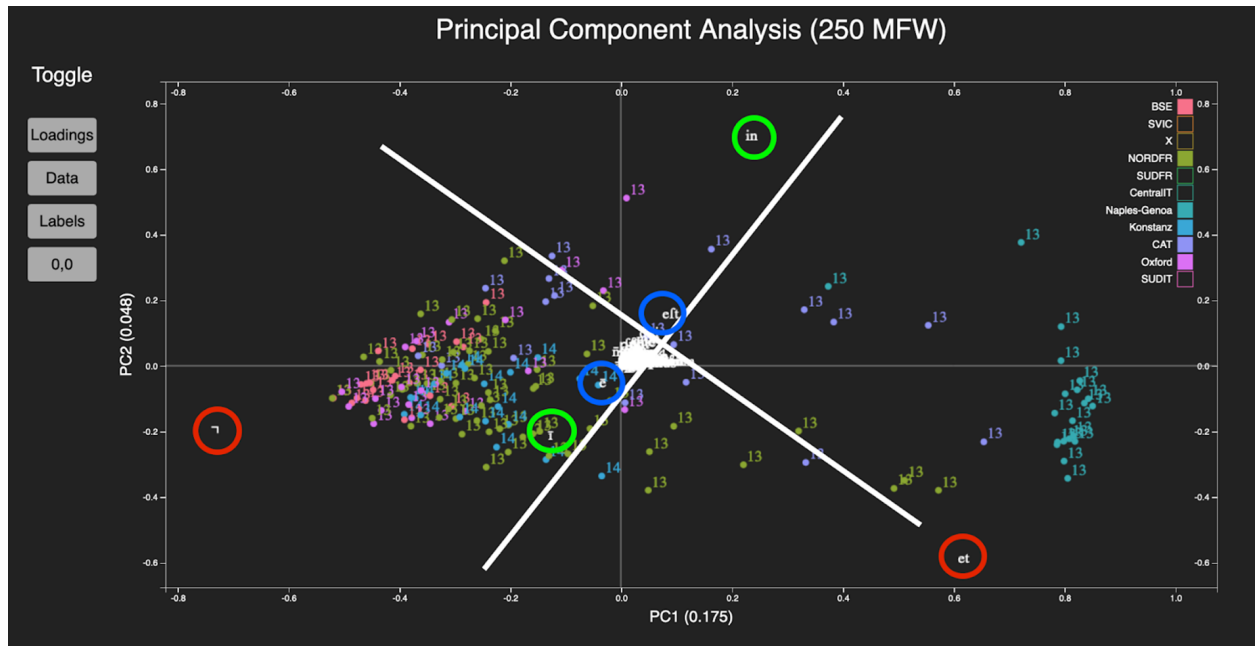


Figure 4 : Principal Component Analysis (PCA) with loadings resulting from TF-IDF analysis of 24 transcriptions of Latin Bibles from many different regions, carried out using scikit. Code adapted from Paul Vierthaler.

A Continuum of Bias

In our paper we have argued for the importance of an explicit approach to transcription norms which encode specific forms of evidence about our research documents linking them to specific research questions, by also taking into consideration both human and machine readers with an emphasis on reproducibility and plain text environments. We also have explained how, along with principles of transcription, the material quality of digitized collections and their basic availability via download or IIIF impacts the way the datasets we are able to create text transcribed from manuscript for computational experiments. In closing we have presented sample evidence which can be drawn from computational analysis of the non-normalised transcription methods we propose.

In the not so distant future, we predict that mainstream medieval studies will be using much more text which has been transcribed directly from medieval manuscripts, and the use of semi-automated methods for text extraction will be more widespread than they are now. As we have argued in our paper, automated methods make possible the inclusion of micro-features in transcription schemes for capturing different kinds of data about scribal practice in manuscripts. Conversely, HTR models can also facilitate certain normalization gestures in the process of manuscript transcription. Whereas this categorization might seem to cast transcription in terms of two extremes—one conservative and one normalizing—a more likely future is one in which there are many different customized methods for accessing certain kinds of text in the pursuit of specific research questions. Even more likely is that researchers of the future will encounter versions of texts across the spectrum of normalization, and they will want to be able to work with all of them, analogous to the way that edition-centered scholarship traditionally used critical editions of texts which did not share the same degree of editorial intervention alongside each other.

It is likely that future text processing methods will emerge to handle these discrepancies in and between texts. In the era of growing popularity of HTR and of data sharing, there is, in our opinion, a new responsibility for the medievalist to participate in thoughtful and explicit text creation. It is unlikely that we will eliminate bias in HTR models, since after all, the transcription norms that we employ based on specific textual traditions are themselves forms of bias, along with datasets combining very different domains and textual artefacts. Whereas the process of remediating medieval texts has been described in terms of the so-called "unedition" (Dark Archives, 2021), one of the pieces of the traditional edition we strongly believe should be preserved moving forward is the editorial statement. Since the relationship between transcription, manuscript work and technology has changed, and will continue to change, our scholarly practices, let us call this new version a "transcription statement." In such a statement, the need to outline one's theoretical and practical principles for transcription norms, the anticipated research questions that a particular layer of data might uncover, samples of training and output data, as well as the principles of model training and correction ought to be provided. In our paper we have attempted to model these kinds of observations using details from our own experiments with transcription. Akin to the calls for multiple forms of transparency about the process of data creation as well as a deep understanding of the context in which data are created (D'Ignazio and Klein, 2020), we imagine such "transcription statements" providing a point of access for understanding what design decisions were made in the creation of the model, unpacking in detail what kinds of information have been privileged and by whom, what kinds of bias are embedded in libraries, in codices, in particular digitized versions of codices, as well as how those various levels of bias impede our field's deeper understanding of our new objects of study on the way to a future of digital manuscript studies.

References

- Alpert-Abrams, Hannah. (2016). "Machine Reading the Primeros Libros." In *Digital Humanities Quarterly*, <http://dx.doi.org/10.17613/M6SC9G>
- Attar, Karen. (2010). "S and Long S". In Michael Felix Suarez; H. R. Woudhuysen (eds.). *Oxford Companion to the Book*. Vol. II. p. 1116.
- Bodard, Gabriel. (2021). "Diplomatic Transcriptions" Epidoc version 9.2 <https://epidoc.stoa.org/gl/latest/trans-diplomatic.html>
- Bozzolo, Carla, Dominique Coq, Denis Muzerelle and Ezio Ornato. (1990). "Les Abréviations dans les Livres Liturgiques du XV^e Siècle: Pratique et Théorie." In *Actas del VIII coloquio del Comité internatiocional de paleografía Latina*, Madrid-Toledo, Sept.–Oct. 1987, 17–27. Madrid: Joyas Bibliográficas.
- Cappelli, Adriano. (1902). *Lexicon Abbreviaturarum*. J.J. Weber Leipzig.
- Cordell, Ryan, and Smith David A. (2018). *Report: A Research Agenda for Historical and Multilingual Optical Character Recognition*. <http://hdl.handle.net/2047/D20297452>
- Dark Archives. (2021). "Birth of the Unedition" Panel. Dark Archives Conference.
- D'Ignazio, Catherine and Klein, Lauren. (2020). *Data Feminism*. MIT Press.

Driscoll, Matthew James. (2006). "Levels of Transcription." In *Electronic Textual Editing*, edited by Lou Burnard, Katherine O'Brien O'Keefe, and John Unsworth, 254–261. New York: Modern Language Association

Greg, W. W. (1950-51). "The Rationale of Copy-Text." In *Studies in Bibliography*, vol. 3, p. 19-36.

Guéville, Estelle. (2021). "Les manuscrits médiévaux occidentaux dans la collection du Louvre Abu Dhabi. 2009-2017." In *Le manuscrit médiéval: texte, objet et outil de transmission*. Volume I. Brepols: Pezia. Le livre et l'écrit. N°22. P. 105-153.

Hasenohr, Geneviève. (2002). "Écrire en latin, écrire en roman: réflexions sur la pratique des abréviations dans les manuscrits français des XII^e et XIII^e siècles." In *Langages et peuples d'Europe: cristallisation des identités romanes et germaniques (VII^e-XI^e siècle)*, edited by Michel Banniard, 79–110. Toulouse: CNRS Université de Toulouse-Le Mirail.

Herrmann, J. Berenike, van Dalen-Oskam, Karina, Schöch, Christof. (2015). "Revisiting Style, a Key Concept in Literary Studies." In *Journal of Literary Theory* (JLT); 9(1): 25–52 <http://www.degruyter.com/view/j/jlt.2015.9.issue-1/issue-files/jlt.2015.9.issue-1.xml>

Hodel, T., Schoch, D., Schneider, C., & Purcell, J. (2021). General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, 7, 13. DOI: <http://doi.org/10.5334/johd.46>

Honkapohja, Alpo, Samuli Kaislaniemi and Ville Marttila. (2009). "Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora." In *Corpora: Pragmatics and Discourse*. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). Ascona, Switzerland, May 14–18, 2008, edited by Andreas Jucker, Daniel Schreier and Marianne Hundt, 451–474. Amsterdam: Rodopi.

Honkapohja, Alpo. (2013). "Manuscript Abbreviations in Latin and English: History, Typologies and How to Tackle Them in Encoding." In *Studies in Variation, Contacts, and Change in English Volume 14: Principles and Practices for the Digital Editing and Annotation of Diachronic Data*, edited by Anneli Meurman-Solin and Jukka Tyrkkö. Accessed September 30, 2020. <http://www.helsinki.fi/varieng/series/volumes/14/honkapohja/>.

Honkapohja, Alpo. (2018). "'Latin in Recipes?' A Corpus Approach to Scribal Abbreviations in 15th-Century Medical Manuscripts." In *Multilingual Practices in Language History: English and beyond*, edited by Päivi Pahta, Janne Skaffari and Laura Wright, 243–271. Berlin: De Gruyter. DOI: <https://doi.org/10.1515/9781501504945-012>

Honkapohja, Alpo. (2021). "Digital Approaches to Manuscript Abbreviations: Where Are We at the Beginning of the 2020s?." *Digital Medievalist*, 14(1), p.None. DOI: <http://doi.org/10.16995/dm.88>

Jänicke, Stefann and Wrisley, David J.. (2017). "Interactive Visual Alignment of Medieval Text Versions," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, p. 127-138, doi: 10.1109/VAST.2017.8585505.

- Kestemont, Mike. (2015). "A Computational Analysis of the Scribal Profiles in Two of the Oldest Manuscripts of Hadewijch's Letters." *Scriptorium* 69: 159–75 <https://www.persee.fr/collection/scrip>
- Kytö, Merja, Grund, Peter, and Walker, Terry. (2011). *Testifying to Language and Life in Early Modern England: Including CD-ROM: An Electronic Text Edition of Depositions 1560–1760 (ETED)*. Amsterdam: Benjamins. DOI: <http://doi.org/10.1075/z.162>
- Lass, Roger. (2004). "Ut Custodiant Litteras: Editions, Corpora and Witnesshood." In *Methods and Data in English Historical Dialectology*, edited by Marina Dossena and Roger Lass, 21–48. Bern: Peter Lang.
- Light, Laura. (1994). "French Bibles c. 1200-30: a new look at the origin of the Paris Bible." In R. Gameson (Ed.). *The Early Medieval Bible: Its Production, Decoration and Use*. Cambridge: Cambridge University Press.
- Light, Laura. (2012). "The Thirteenth-Century Bible: The Paris Bible and Beyond." In R. Marsden and E. A. Matter (Eds). *The New Cambridge History of the Bible*. Volume two, c. 600-1450. Cambridge: Cambridge University Press.
- Piotrowski, Michael. (2012). "Spelling in Historical Texts." In *Natural Language Processing for Historical Texts*. Morgan & Claypool, ch. 3.
- Rigg, Arthur George. (1983). "The editing of medieval Latin texts: a response." in *Studi medievali Ser. 3*, vol. 24, p. 385-388.
- Robinson, Peter and Elizabeth Solopova. (1993). "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue." In *The Canterbury Project Occasional Papers volume I*, edited by Norman F. Blake and Peter Robinson, 19–52. Oxford: Office for Humanities Communication.
- Rogos, Justyna. (2011). "On the Pitfalls of Interpretation: Latin Abbreviations in MSS of the Man of Law's Tale." In *Foreign Influences on Medieval English*, edited by Jacek Fisiak and Magdalena Bator, 47–54. Bern: Peter Lang.
- Rogos, Justyna. (2012). "Isles of Systematicity in the Sea of Prodigality? Non-alphabetic Elements in Manuscripts of Chaucer's 'Man of Law's Tale.'" Accessed September 30, 2020. <https://docplayer.net/41316899-Isles-of-systematicity-in-the-sea-of-prodigality-non-alphabetic-elements-in-manuscripts-of-chaucer-s-man-of-law-s-tale-justyna-rogos.html>.
- Römer, Jürgen. (1997). *Geschichte der Kürzungen: Abbrüviaturen in deutschsprachigen Texten des Mittelalters und der Frühen Neuzeit*. Göttingen: Kümmerle Verlag <https://external.dandelon.com/download/attachments/dandelon/ids/AT003438C7F6F4A7087F7C1257EA400335047.pdf>
- Ruzzier, Chiara. (2010). *Entre universités et ordres mendiants. La Miniaturisation de la Bible au XIII^e siècle*. PhD Thesis. Université Paris 1 Panthéon-Sorbonne, Paris.
- Ruzzier, Chiara. (2016). 'Des armaria aux besaces: La mutation de la Bible au XIII^e siècle', in *Le Moyen Âge dans le texte. Cinq ans d'histoire textuelle au Laboratoire de médiévistique occidentale de Paris*. Grevin, B. & Mairey, A. (eds.). Paris: Publications de la Sorbonne, p. 73-111.

Siemens, Ray, Leitch, Cara, Blake, Analisa, Armstrong, Karin, and Willinsky, John. (2009). ““It May Change My Understanding of the Field”: Understanding Reading Tools for Scholars and Professional Readers.” In *Digital Humanities Quarterly*, volume 3, issue 4.

Stutzmann, Dominique. (2014). “Conjuguer Diplomatique, Paléographie et Édition Électronique : les Mutations du XIIe Siècle et la Datation des Écritures par le Profil Scribal Collectif.” In *Digital Diplomatics. The Computer as a Tool for the Diplomatist?*, edited by Antonela Ambrosio, Sébastien Barret and Georg Vogeler. *Archiv für Diplomatik. Beiheft 14*, 271–90. Vienna, Cologne, Weimar: Böhlau Verlag. DOI: <http://doi.org/10.7788/boehlau.9783412217020.271>

Stutzmann, Dominique, Kermorvant, Christopher, Vidal, Enrique, Chanda, Sukalpa, Hamel, Sébastien, Puigcerver Pérez, Joan, Schomaker, Lambert, and Toselli, Alejandro H. (2018). “Handwritten Text Recognition, Keyword Indexing, and Plain Text Search in Medieval Manuscripts.” Conference paper presented at Digital Humanities 2018 Conference, Mexico City, June 26–29. Accessed September 30, 2020. <https://dh2018.adho.org/handwritten-text-recognition-keyword-indexing-and-plain-text-search-in-medieval-manuscripts>.

Stutzmann, Dominique, Chevalier, Louis. (2021). “Hours: Recognition, Analysis, Editions – HORAE. Numérique et patrimoine. Enjeux et questions actuels.” In *Digital technology and heritage. Challenges and issues*, Mar 2021, Paris, France. 31 p.

Widner, Michael. (2017). Toward Text-Mining the Middle Ages. In J. E. Boyle and H. J. Burgess (Eds). *The Routledge Research Companion to Digital Medieval Literature*. United Kingdom: Taylor & Francis Group, 131-144.