CREMMA Medieval Latin: Literary manuscript text recognition in Latin

Thibault Clérice, Malamatenia Vlachou-Efstathiou, Alix Chagué

CREMMA Medieval Latin: Literary manuscript text recognition in Latin

Thibault Clérice[a*], Malamatenia Vlachou-Efstathiou [b], Alix Chagué [c]

[a] Centre Jean Mabillon, École nationale des Chartes, PSL University, Paris, France

[b] École nationale des Chartes, PSL University, Paris, France

[c] INRIA, CNRS, Paris, France

[*] Corresponding author: Thibault Clérice; thibault.clerice@chartes.psl.eu

Abstract

This paper present a novel segmentation and handwritten text recognition dataset for Medieval Latin, from the 11th to the 16th century. It connects with Medieval French dataset as well as ealier Latin dataset, by enforcing common guidelines. We provide our own addition to Ariane Pinche's Old French guidelines to deal with specific Latin case. We also offer an overview of how we addressed this dataset compilation through the use of pre-existing resources. With a higher abbreviation ratio and a better representation of abbreviating marks, we offer new models that outperform the base Old French model on Latin dataset, reaching readability levels on unknown manuscripts.

Keywords: Handwritten Text Recognition; Latin; manuscripts; Middle Ages; Layout Segmentation

Author roles: Thibault Clérice and Alix Chagué were responsible for the project administration. TC supervised and curated the data of the project. Malamatenia Vlachou-Efstathiou and TC produced the data, MV-E contributing more than half of the transcription work. All contributed to the writing of this paper.

1 Context and motivation

Institutional and academic contexts Handwritten text recognition (HTR) and layout segmentation (LS), its upstream task, have become two important topics in the context of Digital Humanities and digital approaches to cultural heritage collections in the GLAM[1] domain. Its growth over the past three to four years in digital projects can easily be linked to the emergence of user interfaces (UI) allowing for the annotation of ground truth (data which will be used for training), training new models (for the transcription and also, lately, for the segmentation) and for the automatic transcription of the users' own data. At first, only Transkribus (Kahle, Colutto, Hackl, & Mühlberger, 2017) provided such a service through the READ project without fees nor infrastructure requirements[2]. At the end of the 2020 European Union funding, Transkribus moved onto a paid service which accelerated the interest growth of at least one alternative, namely eScriptorium (Kiessling, Tissot, Stokes, & Ezra, 2019) at the EPHE-Scripta-PSL. Unlike the former, the latter is completely open-source, at the cost of not offering a centralized server.

In this context, the *Consortium pour la Reconnaissance d'Écritures Manuscrites des Matériaux Anciens* (CREMMA) project was created to fund a regional server. Its aims are to support students training and to provide local researchers with a free solution, in exchange for guaranteeing the release of data. The CREMMA funding consisted of a starting grant for the initial cost of the infrastructure (Graphic cards, servers, routers, etc. for around 40,000€ VAT excluded) as well as an evaluation grant for providing base models for the community of CREMMA's users (around 8000€). The latter was divided into two main languages: French and Latin, from the 9th to the 21st century. A postdoctoral position, CREMMAlab, provided the infrastructure with complementary time for building a dataset (CREMMA Medieval Old French) and expertise around transcribing medieval manuscripts.

As the CREMMA project was being drafted, Chagué and Clérice (2020) provided a solution for facilitating the FAIR principles (Findability, Accessibility, Interoperability, Reusability) in an HTR context and providing machine-actionable metadata for datasets. HTR-United, both a catalog of open source HTR ground truth and a toolkit to strengthen the control of documentation and validity of HTR data, is born from this necessity, and offers, as of early September 2022, 53 datasets, composed of 13.3 million characters, 43 100 lines in over 12 languages and 6 scripts. Designing HTR-United, we became aware of the stakes of spending our budget in the creation of new corpora and models. They are useful and can complement other existing projects.

[1] Gallery Library Archives Museums.

[2] https://readcoop.eu/our-story/

Authors	Dataset	Project	Characters	Period	Language
White et al. (2022)	Caroline Minuscule	Rescribe	17,000	800-1200	Latin
Vlachou-Efstathiou (2022a)	Eutyches	-	87,000	850-900	Latin
Pinche (2022a)	CREMMA Medieval	CREMMA(lab)	**593,000**	1100-1499	French
-	**CREMMA Medieval Lat**	CREMMA	263,000	1100-1600	Latin
Biay et al. (2022)	DecameronFR	-	20,000	1430-1455	French
Gabay et al. (2022)	Manuscrits du 15e siècle	GalliCorpora	169,000	1400-1500	French
Total			1,149,000		

Table 1: Datasets following the Pinche Guidelines, or adapted through Choco-Mufin. Characters' count are rounded to the closest thousands.

HTR for Latin and Old French Handwriting in the middle ages can generally, and in a simplistic way, be divided into two big writing techniques/systems – calligraphy and cursive (Bischoff, 1985, pp.58 sqq.) – that reflect two complementary practices, cursive hands (*écritures d'usage*) - which are more common to everyday documents such as accounting books, charters, letters, intellectual work - and book hands[3]. While cursive represents a harder challenge, both families have the particularity to be potentially highly abbreviated, depending on the expected audience of the document: literary classics such as Cicero or Vergilius might be less abbreviated than pharmaceutical recipes, scholastic works, or accounting books of an abbey. This situation resulted in mainly two different kinds of *HTR* ground truth dataset creation strategies: datasets that would resolve abbreviations directly in the transcription ground-truth (a practice found mostly used by historians, quite common for cursive, at least in France) and datasets that would keep a diplomatic approach to transcription.

Our dataset builds on the experience of A. Pinche and specifically her work on the CREMMA Medieval dataset, which treats different variations of Old French from the 13th to the 15th century, with a heavy focus on the first section of the period. It started as an adaptation of Pinche (2021) into an OCR dataset and follows our common work regarding HTR and Stylometrical analysis (WAUCHIER 2019). During her postdoctoral work, A. Pinche co-organized a research seminar around the formalization of transcription guidelines for graphemic transcription of Old French (Pinche, 2022b). It notably involved French and Swiss medievalists, including D. Stutzmann, who applied HTR to a large amount of data in the framework of the HIMANIS project, O. Canteaut, who is involved in the e-Notre-Dame de Paris Project, both opting for the resolution of abbreviations in their data.

Based on this work from A. Pinche, a few datasets emerged around the École nationale des Chartes and the CREMMA project. Notably, the GalliCorpora Project (Gabay, Pinche, Leroy, & Christensen, 2022) and the course project DecameronFR (Biay, Boby, Konstantinova, & Cappe, 2022) provided two additions for Old French and Middle French data, centered around the end of the middle ages. On the opposite, the Caroline Minuscule project (White, Karaisl, & Clérice, 2022) was adapted for eScriptorium, as it provided some foundations for recognizing the Caroline script specific to the first centuries of the early middle age. On top of this project, Vlachou-Efstathiou (2022b) provided a dataset based on the transcriptions of two Latin manuscripts from the 9th century. At the start of the present dataset production, there was a lack of data for the second half of the middle age (1100-1500) which we focused on for our dataset (see Table 1).

2 Dataset description

Object name Typically the name of the file or file set in the repository.

Format names and versions XML (ALTO), JPEG

Creation dates 2022-01-01 / 2022-09-22

Dataset creators Thibault Clérice (Organization, Curation, Transcription, Design), Malamatenia Vlachou-Efstathiou (Curation, Transcription, Design), Alix Chagué (Organization)

Language Latin

License CC0

Repository name Zenodo

Publication date 2022-09-XX

[3]The distinctive functions gradually ceased to exist/converged as they were used interchangeably depending on the context.

3 Method

3.1 General aspects of the corpus

Corpus construction theory Borrowing the terminology from the linguistic domain (Bauer & Aarts, 2000), where data construction methods have long been examined, evaluated, and reconsidered, we shall examine the following methodological aspects[4]. Contrary to the notion of "sampling" which is by definition a random selection procedure, "corpus construction" implies a systematic selection of materials that obey a specific rationale, where its efficiency depends on the research question. "Representative sampling" is where these two approaches converge. Sampling secures efficiency in research by providing a rationale for studying only parts of a population without losing information. Its key feature is "representativeness" of the system in question. The larger the range of population representation the smaller the error. Sampling criteria and focal variables correlate. Language corpora and specifically those oriented towards formal criteria (handwriting) rather than content (dialects etc.) are easier to deal with than, let's say, population or natural language corpora. In HTR for medieval manuscripts, "representativeness" was approached in terms of the medieval handwritten Latin language's characteristics - as a system comprised of abbreviations, ligatures, and punctuation signs alongside graphemes. Different genres, scripts, and their degrees of formality served as instances of this system.

Document sampling strategy From the three registers making up the construction of a qualitative corpus according to Bauer and Aarts (2000), namely channel, domain, and function, only the first parameter is constant in our case, the sample representing exclusively the written Latin language, while it gives place to texts of multiple functions addressed to different audiences belonging to various genres (while not aiming, at this stage at exhaustiveness). The corpus construction can be regarded as a cyclical process: it has not been entirely determined *a priori* but rather evolved, bearing in mind the logic of complementarity regarding the already existing datasets. Statistical features, such as abbreviation rate or use of specific characters, can only be estimated from afar and only the analysis of the transcriptions could provide feedback for selecting new documents to fill the gap or strengthen some features. Different genres and scripts were implemented to compensate for what was thought to be missing from the corpus in order to render it as "representative" as possible. HTR engines are language agnostic, but the same cannot be told for the resulting models, which means that it depends on the representativeness of the sample to determine whether a model will work on "similar" documents. The selection of variables can only take place within the framework of a dialectical process in which knowledge of the object and the historical substratum, acquired preliminarily, feeds and controls the prediction.

Three distinctive selection processes have been applied in our case:

1. The first set of documents was selected purely on their linguistic feature, their readability, and their availability as both digitized manuscripts and editions which could be found either online or in local libraries. It led to the inclusion of mostly classical texts such as Seneca's *Medea*, and the *Priapea*. The script did not dictate this selection step.
2. In a logic of complementarity, the second lot of the corpus was dictated, inversely by content. More specifically, given the relative absence of ligatures and abbreviations in classical texts, an important feature of medieval handwritten practices, documents that would display a higher degree of abbreviations were chosen. This led to a genre selection process, specifically for medical and scholastic data. At the same time, and always seeking not to repeat already existing features for the sake of saturation, script diversity was added to the consideration and came naturally as a sort of by-product.
3. Finally, as we wanted to test Kraken models, we sought a transcription project that would provide us with data that would help us evaluate our own. This led to the inclusion of Eichenberger and Suwelack (2021) dataset, produced in the context of a transcribathon in Berlin, in the CREMMA Medieval Lat corpus: it contains a new genre of documents for the corpus (Book of Hours, psalms, etc.).

Quantitative aspects of the corpus Size depends largely on the subjective criteria and resources of each project and little can be said as a general rule: one needs to consider the limitations that stem from the effort put into producing the corpus, the budget available, the number of representations one wants to characterize, and some minimal and maximal requirements (in our case the quota for the production of an efficient HTR model).

Building a turn-key OCR model applicable to as large a range of unseen manuscripts as possible is undoubtedly the end goal (cf. the work of CREMMA Medieval and CREMMA Lab mentioned above) With the production of ground truth being expensive but with increasingly more open-access models available to the public, the challenge is finding the right combination of GTs (either to create a model from scratch or to fine-tune an existing one) that yield the best results. This is where considerations of size and variety enter the discussion and affect directly the quantitative corpus construction strategy.

[4]More specialized studies concerning quantitative approaches in codicology and paleography Maniaci (1993) other than theoretical factors, stress practical factors such as availability, medium, diversity, and internal homogeneity

More specifically, while conducting an experiment on Caroline Minuscule OCR models, Hawk, Karaisl, and White (2018) conclude that "relative preponderance"[5] in small training pools was a considerably more important factor than that of size, which inversely impacts the accuracy of the models resulting from larger training pools. A careful conclusion would have it that a specific combination of manuscripts can yield exceptional results, even though the reasons behind such results or the criteria for the respective manuscripts to be combined are not entirely clear yet. This means that quantity-wise we sought to find a balance between diversity and size of the GT, always making sure that the ground truth yields an efficient model for individual manuscripts on the training set. Training and fine-tuning experiments conducted by Ariane Pinche showed that a specialized model per script isn't always necessary, but the variety of the training set increases its robustness. Therefore, the size of each GT belonging to the train set [6] was limited to 5 pages per script variation (depending on the density of the layout), examining whether this balance can contribute to the production of a generic models[7].

Segmentation vocabulary: SegmOnto With the emergence of efficient layout analyzers and easy-to-use interfaces, the need for efficient segmentation models increases, as does the need for large amounts of data, based on the aggregation of heterogeneous documents. For this, researchers need to agree on a limited common vocabulary and share common practices to facilitate the interoperability of their ground truth.

Alongside text recognition, eScriptorium allows for layout annotation using ontologies and controlled vocabularies. A controlled vocabulary is a lexicon whose purpose is to enable the organization of knowledge to optimize information retrieval and requires the use of terms predefined by the vocabulary designer. In order to identify the different areas of the document and the type of lines present on the page as well as to characterize them from a codicological point of view, we decided to implement the controlled vocabulary SegmOnto (Gabay, Camps, Pinche, & Jahan, 2021). SegmOnto was born out of the need for a limited common ontology based on existing standards, for the description and analysis of document layout, ranging from content categorization to text recognition, mainly addressing the case of manuscripts and early printed books.

SegmOnto has already been implemented in several projects led by A. Pinche and the CREMMA Lab project such as Gabay et al. (2022), resulting in segmentation models mainly for late medieval manuscripts and early prints[8]. As per the CREMMA Medieval Latin dataset, the documents present two kinds of layout: multi-columns and singular columns, for which lines are most often long, except for the psalms and book of hours. SegmOnto offers multiple levels of description, of which only one is completely standardized (the first level), as the second is intended for custom refinement and the third for local and document-based differentiation. For the purposes of the project, only the first level of SegmOnto has been used, notably, the *MainZone* and occasionally the *MargintextZone* for *marginalia*, while the *DefaultLine* tag for the characterization of the lines.

Pinche's Guidelines Pinche (2022b) stressed the fact that due to the need for scientific projects to acquire textual data in mass either to undertake editions of long texts or to constitute large corpora to be interrogated, the use of HTR is becoming necessary to process such a mass of data. The guidelines are the result of the need to establish principles common to projects dealing with the transcription of manuscripts in order to:

- accompany the creation of training seeking to optimize the machine learning of HTR models;
- build shareable and reusable ground truth data sets;
- produce robust generic models, reusable in "out of domain" manuscripts with or without customization, useful to the scientific community and
- minimize the collective cost, including of training people;
- ensure the durability and reuse of the data produced.

A graphemic transcription has been privileged instead of a graphetic one that reproduces the manuscript as truthfully as possible [9]. Pushing the imitation too far through a graphetic approach induces a risk of making the transcription harder to complete (as it requires technical skills to recognize differentiated shapes of characters), harder to make uniform (specifically as more annotators are to participate in a dataset) and potentially unusable for HTR (as it might introduce more characters and ultimately noise for HTR engine to learn). A graphemic transcription preserves the sequence of letters and reduces each form to its meaning and each letter "shape" to a standardized representation in an alphabetical system. Therefore, in cases where functional signs have more than one graphetic manifestation but essentially the same function, they could be represented by the same sign[10]: for

[5] the proportionally higher or lower representation of a manuscript or subgroup of manuscripts in the training pool and the subsequent effect on the accuracy of the respective test manuscript or subgroup.

[6] Aside from 3 documents coming from the Berlin Transcribathon, that were utilized rather as evaluating tools at the end of the project.

[7] On GT size for OCR experiments see : Ströbel, Clematide, and Volk (2020).

[8] More information and case studies can be found here: `https://segmonto.github.io/`.

[9] See Pinche, Camps, and Duval (2021). On this particular topic, we share an opposite view with Gueville and Wrisley (2022). However, while graphemic transcription can not be used for automated graphetic transcription, graphetic transcriptions can be turned easily into graphemic ones, at the cost of establishing a "translation" table for each character.

[10] While most of the special characters exist as such in MUFI a conscious effort has been made to avoid as much as possible the private domain of MUFI, so the data can be as reusable and flexible as possible.

example, for every manifestation of the paragraph sign, we opt for the pilcrow sign on every occasion, instead of several variations such as the `F1E1 PARAGRAPHUS` sign, for the sake of homogeneity (*cf.* Table 2).

On the topic of abbreviation, resolving them produces specific difficulties for HTR engines, as it leads them to learn more about the language than they are originally intended for. In our dataset, abbreviations are not resolved as this constitutes rather an interpretative act linked to the specificity of each document and it is not the same as a textual prediction and it could prove to be detrimental to the extension of an HTR model in the long term. However, as noted regarding the graphetic and graphemic distinction, diplomatic transcriptions entail their respective difficulties, specifically regarding the way in which transcribing specific letters should be handled. Graphetic and graphemic approaches present their own level of granularity, given that neither all variations of characters are available in the UTF8 standard nor do they exist in the private zone where projects such as the Medieval Unicode Font Initiative (MUFI) abound, nor do all transcribers recognize the sometimes minor differences between different instances of the same character. By setting up a list of allowed characters and a list of common and rare cases (such as Table 2 and 3), while not requiring much in terms of interpretation of the text, we allow for a simpler transcription step through reducing the characters diversity, ultimately satisfying both the human transcriber and the HTR engine in terms of learning curves.

In order to ensure the rigorous application of these guidelines and the homogeneity of the data produced, quality control softwares are introduced to the pipeline. Each corpus was passed through ChocoMufin (Clérice & Pinche, 2021), using project-provided character translation tables. This software, alongside these tables, allows for each dataset to be both controlled at the character level and adapted to guideline specifications and modifications. It also allows for project-specific transcription guidelines to be translated to a more common one such as CREMMALab's one (Pinche & Camps, 2022)[11]. This process has been largely used in the first months of the CREMMA Medieval project, as the guidelines were still being drafted: it allowed A. Pinche to produce or align datasets first, and harmonize later, as long as the harmonization was from an upper level of details (closer to graphetic) to a lower level (closer to graphemic).

3.2 Transcription Guidelines for the CREMMA Medieval Lat

The section that follows aims to guide the reader through the transcription norms followed for the Latin dataset, illustrating the process and the more common and complex cases, especially where new characters have been introduced compared to the FRO dataset.

As a member of the CREMMA initiative, the project adheres to the general principles laid out by Pinche (2022b)(*cf.* the detailed tables of pp. 4-15) concerning the base cases (punctuation, word separation, functional signs, superscript letters, abbreviations, ligatures, and roman numerals). In cases of incoherence on a character level, mostly when there was a misunderstanding between transcribers, "incorrect" characters are handled automatically by ChocoMufin. Using the project-provided character conversion table, ChocoMufin controls the transcription and corrects any anticipated error by transforming automatically the character so it conforms to the pre-defined guidelines. This example stresses the fact that data should be used in their post-chocomufin control or *chocomufin'ed* state (manually or in the releases). However, where the guidelines were not directly addressing the situation (new characters, new types of abbreviations), we positioned ourselves and interpreted the guidelines in light of the situation: each decision was discussed with the original guidelines' author.

In general, the main differences that we isolated between the FRO and LAT datasets, stemming from the language as well as the genre's own characteristics, are:

1. the dataset bears no accentuated vowels like in the Old French texts (a rare event though for the corpus)[12];
2. no normalization of `u` and `v`s was provided, nor of `i` and `j`'s[13];
3. two variations of con (antisigma) and 9-shaped are found;
4. a higher diversity of abbreviating character usage and signification;
5. Arabic numerals alongside roman, mostly in scholastic and medical treatises.

Reference marks, functional signs, and punctuation In general, complex medieval punctuation has been simplified as much as possible, with single sign punctuation being reduced to `.` and commas will be rendered as `,`. Double sign punctuation (mainly *punctus elevatus* and *punctus interrogativus*) are consistently reduced to `:`. The hyphenation for words that continue to the next line has been marked with a unique U+002D sign, following 3.1. Table 2 gives a representative specimen.

[11] In order to read the translation table (generally named table.csv in ChocoMufin using repositories), MUFI compatible fonts are recommended, such as Junicode: https://github.com/psb1558/Junicode-font.

[12] This is different from `i` pointing, which is not taken into account by either corpus.

[13] They are however undistinguished ultimately by the newer version of guidelines.

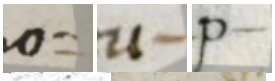
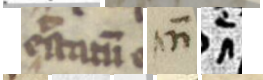
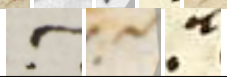
Type	Transcription	Unicode Code	Description or resolution	Examples
Punctuation	¶	00B6	Content change	
Punctuation	-	002D	Hyphenation	
Punctuation	⁄	2215	Diastole	
Reference mark	ˆ	2038	Omission sign 'caret'(reintroduction of content)	
Punctuation	:	003A	punctus elevatus	
Punctuation	:	003A	punctus interrogativus	

Table 2: Punctuation, functional signs and hyphenation.

Contractions, Abbreviations, and Ligatures Cappelli (1899) categorized abbreviations into six categories: Truncation, Contraction, Abbreviation marks significant in themselves, Abbreviation marks significant in context, Superscript letters, and Conventional signs. As Pluta (2020) stresses, the six aforementioned categories are not mutually exclusive, but the functional grouping is helpful.

Contractions: A word is abbreviated by contraction when one or more of the middle letters are missing. Such an omission is indicated by one of the general signs of abbreviation, present in both corpora, always following Pinche (2022b). Thus, macrons and generally horizontal lines diacritics over the letter such as tildes are represented by horizontal tildes, any zigzag similarly shaped forms are simplified into superscript vertical tildes. In our corpus, in cases where a macron is extended to more than one letter due to the cursivity of the script, this trait has been reproduced in the transcription, as well as in the case of stacked diacritics, usually in later medieval manuscripts (*cf.* Table 4).

Abbreviation marks significant in themselves:

Standard Abbreviations signs have been preserved as such, like *pr(a)e* -p̃ (p + combining tilde), *pro* -ꝓ, *hoc* - ħ, ſ (s with diagonal stroke) for *secundum* or *ser-*, �underscore for 9 shaped *con/cum*, Tironian sign ꝰ for the desinence *-us*, ᵓ for *tur*, and ꝙ/ Qꝺ for *quod*. Absent from the CREMMA Medieval Old French dataset but present in the LAT one, the truncated ending *-is* is transcribed using the codepoint U+A76D ꝛ. For the "inverted c" variation of the preposition con/cum, the *antistigma*, ↄ is used. For *-rum*, in a graphemic transcription, the symbol ꝝ is used rather than the specific rotunda -rum (U+A75D) of the MUFI[14].

Abbreviation marks significant in context : The abbreviation for the enclitic *-que* or simply *--bus* or vertical *-m* in later manuscripts, has been reduced to the semicolon ; sign (u+F1AC), in order to avoid the ligature specific u+E8BF (ꝯ) codepoint belonging to the private domain of MUFI and in order to avoid confusion with the regular semicolon.

Conventional signs: a category that includes all signs that stand for a frequently used word or phrase, and they are almost always isolated (*cf.* Pluta (2020)). First, a rather frequent one, the abbreviation sign for *esse* is represented by the mathematical operation Almost equal ≈ always abiding by the MUFI recommendations. In the same vein, the Division sign ÷ is used ubiquitously for the abbreviation sign of *est/id est*. Tironian *et* (all variations of it, *cf.* below) is transcribed by ⁊. *Etiam* can also be found abbreviated by a combination of the Tironian *et* and the macron symbol, for which a horizontal tilde is used (see Table 4.).

[14]The same two-shaped mark on the baseline, combined with a downward stroke, may stand as well for "-ris" as in "Aristoteles", though it is more often used at the end for "rum".

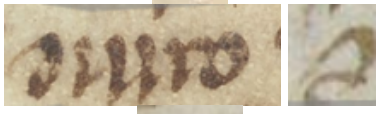
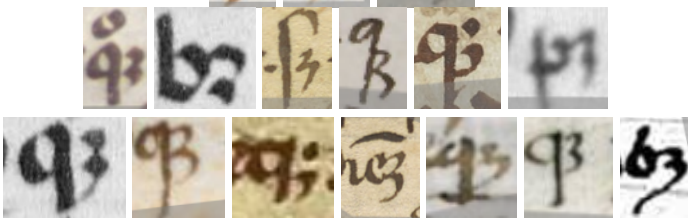
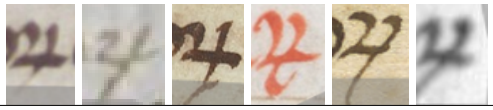
Character(s)	Resolution	Examples
⁊	Et	
⁊+◌̃	Etiam	
ꝑ	-is	
đ	d + any desinence truncation	
ↄ	con	
≈	esse	
÷	est/id est	
;	-que/-bus/-m/-et	
ꝯ	-rum	

Table 3: Freestanding, letter-combining abbreviations and their corresponding transcription signs. đ cannot be found in our dataset and is mentioned here as it might be a common case in other dataset.

Ligatures, *ie.* combinations of more than two letters in one form with the reduction of proclitic and enclitic letters or abbreviating symbols placed above or joined with letters are reduced to their original alphabetical components. Ligatures between letters in cursive scripts such as the ﬅ (U+FB05) ligature or the two ﬀ (U+FB00) ligature are resolved as *-st-* and *-ff-*. For the very frequent *quia*, the transcription *qr* has been privileged [15], avoiding the MUFI sign ꝗ that belongs to the private domain. More examples are provided in the Table 4.

[15] Other transcription guidelines privilege "q2" as a reference to the "r rotunda-shaped" abbreviation sign that lays next to q the choice of *qr* from our part being the reduction to the r rotunda-shaped abbreviation sign to the simpler r. The original insular abbreviation has a simple vertical tilde next to the letter "q".

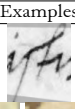
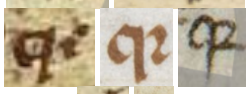
Type	Transcription	Unicode Code	Description or resolution	Examples
Ligature	st	-	Normally transcribed ligature	
Ligature	.n.	-	enim	
Ligature	qr	-	quia	
Monogrammatic Ligature	qd	-	quod	
Monogrammatic ligature	Et	-	Et	
Contraction	aũt̃	-	Long vertical tilde transcribed by two tildes	
Contraction	ẽẽ	-	Long vertical tilde transcribed by two tildes [16] ;	
Contraction	tp̃̃a	-	Two stacked tildes	

Table 4: Ligatures and special contraction cases.

Superscripts letters and interlinear additions A standard way of contracting a word is by adding a superscript letter which gives information about the abbreviated sequence. Frequent ones are open a, u, o, or the ending of a word altogether. These were all rendered with the aid of superscript characters available in MUFI (Pinche, 2022b, p. 11). *Ergo* and *igitur* are two of the most frequent example of abbreviation with superscript letters.

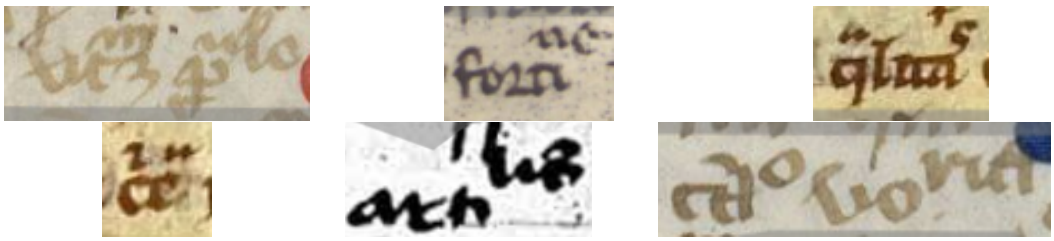


Figure 1: Examples of contraction use of superscript letters.Manuscripts in the following order: BIS193, CML13027, Montpelier H-318, Montpelier H-318, BAV Pal. lat.373, BIS193.

A special case where superscript letters were used with a non-abbreviating function in the project and merits to be mentioned was for the transcription of interlinear additions. Especially in manuscripts with scholastic and medical content, missing words/explanations are added in the interlinear space, something which was at first a challenge for the transcription process due to segmentation constraints. More specifically, it can be, at times, impossible to completely differentiate the segmentation masks of two words that are very close to each other on the vertical axe (like the interlinear additions). Therefore, provided that the corresponding combining letter exists and both words can be formulated, no new lines were carved for the interlinear additions. Where this was deemed too complex, interlinear additions were omitted (see Figure 2.

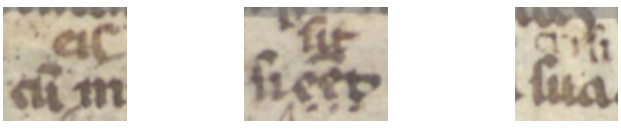


Figure 2: All examples come from the CML 13027 manuscript.

Rare characters and Numerals Maniaci (1993), as they refer to corpus construction practices for balanced corpora, stresses that "sporadically attested variables will therefore be preferred to those that appear in all - or almost all - the individuals that are part of the corpus". Rare characters, a subset of freestanding abbreviation signs, specifically occurring in the LAT dataset are therefore given special attention. In two of the manuscripts,

both of medical content, some occurrences of graphemes for the denotation of the metric values *ounce* and *semuncia* were encountered. For their transcription, the MUFI codepoints ℥ (U+2125) and ℔ (U+10192) were used. The character "barred O" is represented by the Unicode codepoint ∅ (U+2205, mathematical representation of an Empty Set) and is widely used to transcribe the word *instans* instead of the ꝋ (U+A74B) that, according to MUFI recommendation stands for the abbreviation of *obi(i)t*(Coulson & Babcock, 2020, p.10).

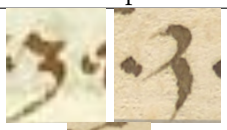
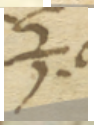
Type	Transcription	Unicode Code	Description or resolution	Examples
Symbols	℥	2125	Ounce	
Symbols	℔	10192	*Semi-Ounce	
Abbreviations	∅	2205	instans	

Table 5: Rare characters found in Montpellier H318, Phil., Col. of Phys. 10a 135 and BIS 193.

Last but not least, in addition to roman numerals, which are fairly frequent in medieval manuscripts and in the FRO dataset, often preceded and followed by dots such as ".ii.", Arabic numerals are also comprised in the dataset, mainly due to the medical treatises[17] (see Figures 3 and 4).

Figure 3: Manuscripts in the following order: Latin 16195, Phi10a135, Phi10a135, Phi10a135, BIS 193, CML13027, Egerton821, Latin6395.

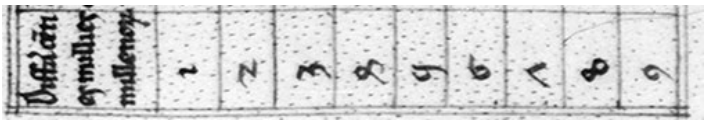


Figure 4: Snipet of Arabic numerals from BnF,lat.15461,fol.13r for comparison

Production pipeline The data was built using eScriptorium and Kraken for both segmentation of zones and lines (specifically the BLLA model). Manuscripts were annotated successively. First, the manuscript is automatically segmented, then its segmentation is manually corrected (addition, deletion, and modifications), and the text is transcribed. After each sample is completely annotated, it is controlled through the ChocoMufin software for its use of characters while the respect of the segmentation classification vocabulary is controlled by HTRVX. Finally, data are released on Github.

All the combining and abbreviation signs suggested for use by the present adaptation of Old French guidelines can be also found on the custom-made Unicode keyboard, which can be imported in the *eScriptorium* interface, conceived in order to facilitate and accelerate the transcription process and compatibility with the Unicode recommendations [18].

[17] An excellent article dealing with numerals is Burnett (2020).

[18] Available here: https://github.com/HTR-United/CREMMA-Medieval-LAT/blob/main/keyboard.json

4 Results and discussion

Properties of the resulting dataset The resulting version of the dataset (see Table 6) is built on 18 + 3 manuscripts. All alignments are original alignments, but some draw their original transcription from online projects (*cf.* Acknowledgements).

The current version of the dataset shows a wide variety of genres, and thus vocabulary. From medical and grammatical content to literary and scholastic a certain level of arbitrariness is introduced in the sequence of characters as they are not as repetitive and predictable from the machine as in a homogeneous genre or topic-driven dataset. The collection was built as to be not representative of one specific use of the Latin language and is not thematically unified - while the CREMMA Medieval dataset focuses more on literary texts, specifically hagiographic and *chanson de geste* textsMedical and scholastic genres, furthermore, induce the use of a range of rare characters and often underrepresented letters (such as zeta and upsilon, as well as some rare kappa).

Other features, such as layout and type of digitization (microfilm or original), provides different representations of texts, with more or less noise in the mask of each line given the space between them, with more or less contrast between information (colored text yields less "information" in digitized manuscripts as they tend to be a duller form of grey that black ink, while clearly departing from the manuscript "background" in color).

A time span of 5 centuries between the earliest and the oldest manuscripts, with a clear focus on the period starting in the 1200s and finishing in 1500. This leads to a good representation of a variety of Gothic scripts [19], including personal hands alongside formal categories [20], with different levels of execution (cursivity and formality). We note an intra-manuscript variation for letters such as single and two-compartment or open and closed s normal present in the same manuscript.

Shelfmark ID	Pages	Type	Date	Status	Script	Folio Sampling	degree of abbreviations
Egerton 821	4	Medic.	1100-1199	Color	Praegothica	Sequential	medium
Montpellier H318	5	Medic.	1100-1299	Color	Semitextualis Libraria	Sequential	high
CCCC MSS 236	5	Litt.	1200-1225	Color	Textualis Libraria	Sequential	medium
CLM 13027	5	Medic.	1250-1299	Color	Southern Textualis Libraria	Sequential	high
Latin 16195	4	Medic.	1250-1299	Microfilm	Semitextualis Currens	Sequential	high
† MsWettF 15	5	Schol.	1270-1280	Color	Textualis Libraria	Sequential	high
Laur. Plut. 33.31	5	Litt.	1300-1310	Color	Textualis Meridionalis	Sequential	low
Arras 861	5	Litt.	1300-1399	Color	Textualis Formata	Sequential	medium
† BIS 193	5	Schol.	1300-1399	Color	Textualis currens	Sequential	high
Phil., Col. of Phys. 10a 135	5	Medic.	1300-1399	Color	Cursiva recentior	Sequential	medium
† Mazarine Ms. 915	4	Schol.	1300-1399	Color	Textualis Meridionalis	Sequential	high
‡ UBL, Ms 758	15	Eccl.	1320-1340	Color	Textualis Libraria	Semi-Sequential	low
Latin 6395	6	Litt.	1325-1399	Microfilm	Semitextualis Libraria	Sequential	low
Laur. Plut. 39.34	5	Litt.	1400-1499	Color	Humanistica Cursiva	Sequential	low
† Vat. Pal. Lat. 373	4	Schol.	1400-1499	Microfilm	Hybrida Currens	Sequential	low
Laur. Plut. 53.08	4	Gramm.	1459	Color	Personal Humanistica	Sequential	medium
Laur. Plut. 53.09	4	Gramm.	1400-1499	Color	Humanistica Rotunda	Sequential	low
‡ Berlin, Hdschr. 25	17	Eccl.	1400-1499	Color	Textualis Formata	Semi-Sequential	low
‡ Berlin, Germ. Oct. 511	6	Eccl.	1400-1499	Color	Hybrida formata	Semi-Sequential	low
Latin 8236	5	Litt.	1471-1499	Microfilm	Humanistica Cursiva	Random	low
† CCCC MSS 165	5	Schol.	1500-1599	Color	Personal Cursive	Sequential	medium

Table 6: Basic features and length of the dataset in chronological order. Medic. stands for medical, Litt. for Litterature, Schol. for scholastic commentaries, Gramm. for grammatical commentaries, Eccl. for church literature (book of hours, psalms, etc.). Texts preceded by a ‡ are aligned and corrected using the Berlin Transcribathon dataset, by a † using the SCTA TEI editions.

Character frequencies in CREMMA Medieval datasets We set up this corpus to both complement the Old French CREMMA dataset and grow the available set of data for Latin through the Middle Ages, noting that at least two datasets for Medieval Latin existed already (Caroline Minuscule and Eutyches) in abbreviated form for pre-10th century documents.

[19] Characterisation of scripts was made by the transcriber where the information was not available on the notice of the manuscript. The criteria followed for the Gothic scripts are those of Derolez (2003).

[20] For the particular case of the "scritture di dotti", or the distinctive scripts of scholars which do not wholly conform to Delorez' classification criteria, see the contribution of Maria Christina Rossi (Univ. of Pisa) at the 22nd edition of CIPL (September 2022) https://cipl.hypotheses.org/maria-cristina-rossi-univ-pisa

Lang	type	Words	Words %	Unique words	Unique words %	Freq. of unique words > 1
Latin	abrégé	6855	11.94%	1460	6.24%	279
Latin	complets	50557	88.06%	21935	93.76%	5025
Old French	abrégé	5755	4.15%	1457	4.89%	286
Old French	complets	132828	95.85%	28315	95.11%	8726

Table 8: Comparative statistics table on abbreviations

Character	Codepoint	Latin	Old French	% in Latin	Ratio
⁊	204A	2228.0	4400.0	33.61	0.51
◌ʳ	036C	148.0	219.0	40.33	0.68
&	0026	83.0	116.0	41.71	0.72
ꝯ	A76F	850.0	779.0	52.18	1.09
ꝑ	A751	1500.0	919.0	62.01	1.63
◌ⁱ	0365	1486.0	820.0	64.44	1.81
◌̃	0303	14445.0	5759.0	71.50	2.51
◌ᵃ	0363	2024.0	732.0	73.44	2.77
ꝰ	A770	1763.0	523.0	77.12	3.37
◌̒	033E	3827.0	973.0	79.73	3.93
◌ᵉ	0364	518.0	120.0	81.19	4.32
ꝓ	A753	462.0	80.0	85.24	5.78
◌ᶻ	1DD1	1018.0	137.0	88.14	7.43
◌ˢ	1DE4	978.0	55.0	94.68	17.78
◌ᵒ	0366	870.0	61.0	93.45	14.26

Table 7: Abbreviating signs, present more than 50 times in both the Latin and the Old French CREMMA datasets. The CREMMA Medieval (Old French) dataset is comprised of 693,052 characters in total, which makes it more than twice the size of CREMMA Medieval Latin. Despite this difference, most abbreviated characters are more represented in the Latin dataset.

Unlike CREMMA Medieval, our approach has been feature-driven, as we tried, as much as possible, to find data that would ultimately allow for better recognition of special characters outside the classical A-Z range. In this regard, we succeeded, as we have a higher frequency of MUFI or special characters in our dataset than in the Medieval Old French dataset, despite being smaller overall (see Table 7). Only three characters are more represented in the other dataset: the Tironian ET, the superscript combining R (common on words such as "grand" [large, big]), and the ampersand &. The character ꝯ is equally present in both datasets: resolved as con- or com- in French, it is often used in words such as ꝯmence (*commence*, to start). Some very frequent diacritics, such as the horizontal lines and vertical lines (transcribed in vertical or horizontal tildes according to Pinche's guidelines), have seen a rise in presence: horizontal tildes are 2.51 times more seen in our Latin corpora, and the vertical tilde 3.93. This will allow better recognition of these two frequent marks, as it now totals around 19,000 occurrences in both datasets for the horizontal tilde and 4500 for the vertical one, making them the first and the third most represented characters in the CREMMA-funded datasets.

Overall, our dataset presents texts that are more varying in terms of features than the Old French dataset (see Figure 5). This is the result of our feature-driven approach. However, some manuscripts have nearly no abbreviation, Laur. Plut. 39.34 notably so as it only contains 3 abbreviated words which is a single character abbreviation (⁊, *et*). According to our analysis of the Old French data, a little less than half of our manuscripts are less abbreviated than the most abbreviated text in the Cremma Medieval FRO dataset, while the other half can beat it with up to ten points. However, both language show similar maximum frequencies in terms of non-single letter abbreviations (abbreviations made up of a single character in MUFI such as ⁊, &, ꝑ)[21].

Finally, despite showing a similar number of *pages*, we see a large variation in terms of words density with a somewhat limited variation in terms of unique words. This shows how pages as a metric are not enough to characterize a corpus for HTR and Layout segmentation purposes: the number of columns, lines, and potentially of words or characters supplements the first. To showcase this argument, the Berlin, Hdschr. 25 manuscript has the highest number of pages (17) but the third lowest amount of words (961).

[21] This definition, while useful to quantify some phenomenon, is debatable and should not be used to make a quantitative conclusion on these languages, they merely inform us about our dataset. For example, etiam (⁊ + tilde) is technically a single letter with a diacritic, but will be counted as two characters in our case.

Manuscript	Words	Un. Words	Abbr. words	Abbr. ratio	*NSCA*	*NSCA* ratio	Un. abbr.	Un. abbr. ratio
Laur. Plut. 39.34	783	571	3	0.38%	0	0.00%	1	0.18%
. Berlin, Germ. Oct. 511	**171**	134	1	0.58%	0	0.00%	1	0.75%
Berlin, Hdschr. 25	961	654	12	1.25%	3	0.31%	6	0.92%
Latin 8236	1475	1057	33	2.24%	5	0.34%	6	0.57%
Laur. Plut. 33.31	1278	858	36	2.82%	17	1.33%	21	2.45%
Laur. Plut. 53.09	1300	798	38	2.92%	10	0.77%	9	1.13%
CCCC MSS 165	1521	713	49	3.22%	28	1.84%	23	3.23%
CCCC MSS 236	1239	874	68	5.49%	44	3.55%	24	2.75%
Latin 6395	3304	2418	190	5.75%	85	2.57%	72	2.98%
Laur. Plut. 53.08	2985	1870	195	6.53%	94	3.15%	67	3.58%
UBL, Ms. 758	4468	2393	297	6.65%	72	1.61%	64	2.67%
Arras 861	2416	1601	164	6.79%	101	4.18%	80	5.00%
Egerton 821	981	677	71	7.24%	28	2.85%	31	4.58%
Phil., Col. of Phys. 10a 135	1487	1057	151	10.15%	52	3.50%	44	4.16%
Montpellier H318	4456	2316	458	10.28%	131	2.94%	109	4.71%
Vat. Pal. Lat. 373	2258	1203	234	10.36%	69	3.06%	67	5.57%
Latin 16195	4135	1676	569	13.76%	168	4.06%	107	6.38%
MsWettF 15	3574	1452	501	14.02%	172	4.81%	107	7.37%
CLM 13027	6499	3612	970	14.93%	340	5.23%	257	7.12%
BIS 193	**7370**	2731	1161	15.75%	413	5.60%	244	8.93%
Mazarine Ms. 915	4751	1873	824	17.34%	350	7.37%	195	10.41%

Table 9: Statistics per manuscript. “Un.” stands for Unique, “Abbr.” for Abbreviated or Abbreviation, “NSCA” for Non-Single Character Abbreviation. The lowest and highest values are in bold typeface. The separation between Laur. Plut. 53.08 and UBLMs. 758 symbolize the highest ratio of abbreviation in the Old French dataset.

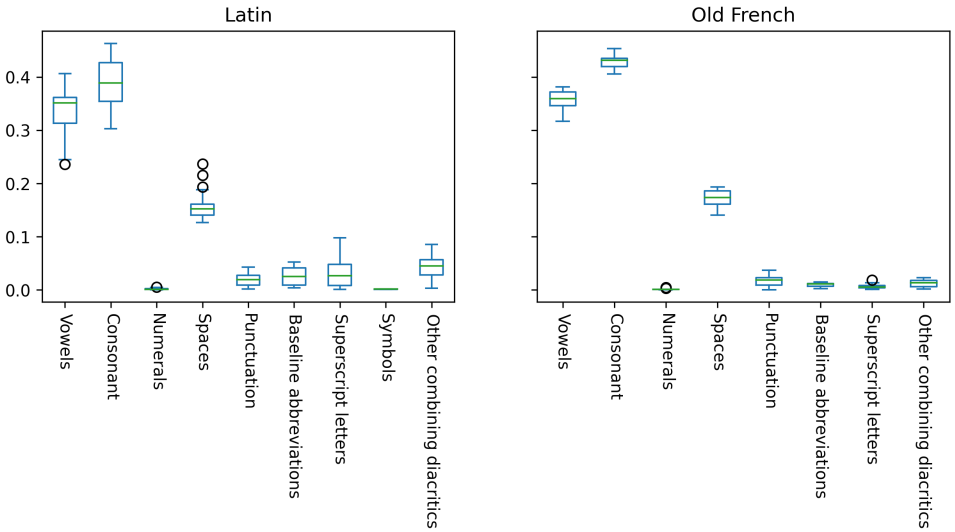


Figure 5: Frequences of character classes across manuscripts

Model	Medieval Old French (In Domain)	Medieval Latin (In Domain)	UBL	BGO	BH25
All	94.30	90.15	71.69	79.12	85.10
No Crem. Med. Lat.	94.04	80.68	67.68	78.02	81.89
Only Fro	94.01	78.10	67.49	76.81	80.74

Table 10: General accuracy results of the models. Model *All* contains all data presented in Table 1, model *No. Crem. Med. Lat.* contains everything but the present dataset, model *Only Fro* contains all datasets but Latin one (Eutyches, Caroline, Cremma Medieval Lat). Two types of test sets are present: the "In Domain" dataset are pages from the same manuscripts as the models, all others (UBL 758, BGO 511, and B.H. 25) are manuscripts from the Faithful Transcriptions Data Set aligned in CREMMA Medieval Lat but unused for training.

5 Implications/Applications

With this addition to the overall amount of datasets available, we have now 1.149 million characters for Medieval manuscripts with book scripts, ranging from the 9th to the 15th century. These data offer more than characters, as we can imagine using them in the context of linguistic studies (evolution of dialects, abbreviation usage, etc.) thanks to the common transcription norm or in codicology studies (evolution of layouts, relation between layouts) thanks to the common segmentation vocabulary, both using the original data or automatically annotated one.

As a direct output, we trained a model which would allow for transcribing or starting transcription of Latin medieval manuscripts. In order to evaluate the gain from our data, we trained three models:

1. a model containing all data from the Table 1, to help transcribe Latin and Medieval French manuscripts, which is the end goal of this paper;
2. a model containing every dataset but our own, to evaluate the impact regarding the quantity of data we add for Latin (*i.e.*, to find out if the original Carolingian datasets were enough to break the language model of the Old French datasets);
3. a model containing only Old French data, from *incunabula* of the 15th century to the main dataset CREMMA Medieval.

Each model uses at least 10% of the pages of each dataset for the development set. CREMMA Medieval FRO and Lat are split furthermore with another 10% subset for evaluation, proposing "In Domain" evaluation. From CREMMA Medieval Lat, as stated earlier, all aligned data from the *Faithful Transcription Data Set* are kept for testing, as an out-of-domain set.

The results show a massive improvement for the in-domain Latin dataset (see Table 10) and an insignificant one for Old French. The addition of the CREMMA Medieval Dataset provides overall better results on out-of-domain datasets from the three manuscripts taken into account, UBL Mss. 758 and Berlin, Hdschr. 25 display an improvement of 4.2 points at least (over around 30% of CER) while Berlin, Germ. Oct. 511 (BGO), the smallest transcription set of the dataset, shows an improvement of below 2.4%. This improvement derives equally from the simple addition of Latin into the model, as shown by the clear gap between the mixed model with Carolingian data: not only the model might benefit from Latin in general (as potentially shown by the simple addition of the Carolingian data), but it also gains in performance out of the amount of data from the same period as the generic Old French CREMMA Medieval dataset. We actually see in table 11 that there are much fewer errors on characters that saw their frequencies jump from a few thousand or hundreds of occurrences to many thousands one. The *All* model does only a fourth of the error of the *Only Fro* model on Tilde or two-thirds on vertical tildes for the UBL manuscript. The -rum abbreviation (ꝝ) or the -et/-ed/-ibus one (;) are quite new to the medieval datasets in general, which explains the clear difference in results. Overall, this dataset helped create a model allowing for readable output (see Table 12 for a side-by-side comparison) on medieval manuscripts, or at least transcription that can help produce new data.

Acknowledgements

A number of transcriptions are the product of alignment and adaptation of existing projects that have worked on the manuscripts in question. In case of already digitized transcriptions, an alignment and correction was done. In case of a printed edition, it served as a guide for difficult to read passages.

- For the manuscripts: MsWettF 15 / BIS 193 /Latin 6395 /Vat. Pal. Lat. 373 and CCCC MSS 165, the transcriptions of Sentences Commentary Text Archive (SCTA) Project by Jeffrey C. Witt (Witt, 2016) [22]. In the case of *dubia*, additional corrections have been made for the faithful reproduction of the abbreviations ;

[22] The GitHub repository of the project can be found here: https://github.com/scta-texts and their reading room here: https://scta.lombardpress.org/

model	test	[Space] %	[Space]	Tilde	Vert. Tilde	7	ꝰ	ꝯ	ꝑ	ħ	ł	ꝙ	ꝓ	ꝥ	;
All	fro	1.7	803	77	46	0	10	17	15	0	0	0	4	0	0
No Crem. Med. Lat.	fro	1.7	726	89	55	0	15	12	15	0	0	0	3	0	0
Only Fro	fro	1.7	733	86	50	0	12	15	20	0	0	0	2	0	0
All	lat	1.7	74	27	31	0	3	2	3	0	8	2	2	0	1
No Crem. Med. Lat.	lat	2.8	138	78	92	0	17	8	11	1	17	15	2	15	32
Only Fro	lat	3.1	149	91	87	0	16	10	9	1	17	20	2	15	33
All	BGO	2.9	22	1	0	1	1	0	0	0	0	0	0	1	2
No Crem. Med. Lat.	BGO	2.3	13	3	0	1	1	0	0	0	0	0	0	1	2
Only Fro	BGO	2.3	13	1	0	1	1	0	0	0	0	0	0	1	2
All	BH25	1.9	63	44	18	0	2	0	8	0	5	1	2	3	10
No Crem. Med. Lat.	BH25	1.8	73	68	21	0	4	0	9	0	5	1	2	3	12
Only Fro	BH25	2.1	100	71	21	0	5	0	5	0	5	1	2	3	12
All	UBL	4.4	274	67	48	0	12	0	54	10	30	2	14	30	43
No Crem. Med. Lat.	UBL	6.0	482	256	76	0	38	0	69	11	37	7	16	71	71
Only Fro	UBL	5.7	484	239	77	0	28	0	59	11	37	7	15	71	71

Table 11: Details on errors from the test presented in Table 10. Space % shows the portion of error points due to bad spacing, *e.g.* *All* Model has a 94.30% accuracy on Fro test set, which means a 5.7% Character Error Rate (CER): not recognized SPACES represent 1.7 points of CER, more than a quarter of the CER. Other numbers are absolute values of missed characters (deletion or substitutions) to make comparisons between models possible; insertions are not accounted for.

- for Berlin, Hdschr. 25, Faithful Transcriptions Data Set (Eichenberger & Suwelack, 2021);
- for the Donatus manuscripts: Laurentianus Pluteus 53.08 and 53.09, the edition of HyperDonat By Bruno Bureau & Christian Nicolas has been consulted (Bureau, Nicolas, & Ingarao, 2008) and (Pinche, Bureau, & Nicolas, 2016), preserving, nevertheless, the manuscript *lectiones*/errors;
- In the same vein, for Latin 16195, the critical edition of*Questiones de coitu* (Cartelle, 2017), for Montpelier H 318 and CLM 1302, the critical edition of *Liber minor de coitu* (Cartelle, 1987) and for Philadelphia, College of Physicians, 10a 135, the critical edition of the *Tractatus de sterilitate* (Cartelle, 1993) by Enrique Montero Cartelle were consulted respectively as reference.

Tools used for verification of any *dubia* in original transcriptions:

- The online version of the Capelli: https://www.adfontes.uzh.ch/fr/ressourcen/abkuerzungen/cappelli-online
- During the deliberation regarding the use of special characters, the MUFI recommendations for Latin (last version) were respected (Wills, 2016) available here: https://folk.uib.no/hnooh/mufi/specs/MUFI-CodeChart-3-0.pdf.

Funding Statement

The project CREMMA was funded by the DIM MAP (now DIM PAMIR) under the supervision of the Conseil Régional d'Île de France. Part of the alignment of data for the *Faithful Transcription Data Set* and the complete writing time for this paper was done under the funding of the second phase of CREMMALab post-doc (Thibault Clérice). The article publication fees are provided by the Centre Jean Mabillon.

Competing interests

The author(s) has/have no competing interests to declare.

References

Bauer, M. W., & Aarts, B. (2000). Corpus construction: A principle for qualitative data collection. *Qualitative researching with text, image and sound: A practical handbook*, 19–37.

Biay, S., Boby, V., Konstantinova, K., & Cappe, Z. (2022). *Tnah-2021-decameronfr*. Retrieved from https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-DecameronFR DOI: 10.5281/zenodo.6126376

Bischoff, B. (1985). *Paléographie de l'antiquité romaine et du moyen âge*. Paris: Ed. Picard.

Bureau, B., Nicolas, C., & Ingarao, M. (2008). *Hyperdonat, commentaire attribué à aelius donat aux comédies de térence.*

Burnett, C. (2020, 12). 25The Palaeography of Numerals. In *The Oxford Handbook of Latin Palaeography.* Oxford University Press.

Cappelli, A. (1899). *Dizionario di abbreviature latine ed italiane: usate nelle carte e codici specialmente nel medio-evo ripprodotte con oltre 13000 segni incisi.* Hoepli.

Cartelle, E. M. (1987). *Liber minor de coitu: Tratado menor de andrología. anónimo salernitano.* Universidad, Valladolid.

Cartelle, E. M. (1993). *Tractatus de sterilitate: Anónimo de montpellier (s. xiv).* Universidad, Valladolid.

Cartelle, E. M. (2017). Questiones de coitu. *Cuadernos de Filología Clásica. Estudios Latinos*, *37*(1), 51.

Chagué, A., & Clérice, T. (2020). *Htr-united: Ground truth resources for the htr and ocr of patrimonial documents* [dataset].

Clérice, T., & Pinche, A. (2021, 9). *Choco-mufin, a tool for controlling characters used in ocr and htr projects.* Retrieved from https://github.com/PonteIneptique/choco-mufin DOI: 10.5281/zenodo.5356154

Coulson, F., & Babcock, R. (2020). *The oxford handbook of latin palaeography.* Oxford University Press, USA.

Derolez, A. (2003). *The palaeography of gothic manuscript books: From the twelfth to the early sixteenth century.* Cambridge University Press.

Eichenberger, N., & Suwelack, H. (2021, October). *Faithful transcriptions data set: Tei/xml-encoded transcriptions of medieval theological manuscripts* (Tech. Rep.). Retrieved from https://doi.org/10.5281/zenodo.5582483 DOI: 10.5281/zenodo.5582483

Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021). Segmonto: common vocabulary and practices for analysing the layout of manuscripts (and more). In *16th international conference on document analysis and recognition (icdar 2021).*

Gabay, S., Pinche, A., Leroy, N., & Christensen, K. (2022). *Données htr manuscrits du 15e siècle.* HTR United. Retrieved from https://github.com/Gallicorpora/HTR-MSS-15e-Siecle

Gueville, E., & Wrisley, D. J. (2022, July). *Transcribing Medieval Manuscripts for Machine Learning.* Retrieved from https://halshs.archives-ouvertes.fr/halshs-03725166 (working paper or preprint)

Hawk, B., Karaisl, A., & White, N. (2018). Modelling medieval hands: practical ocr for caroline minuscule. *Digital Humanities Quaterly*, *13*.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (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)* (Vol. 4, pp. 19–24).

Kiessling, B., Tissot, R., Stokes, P., & Ezra, D. S. B. (2019). escriptorium: an open source platform for historical document analysis. In *2019 international conference on document analysis and recognition workshops (icdarw)* (Vol. 2, pp. 19–19).

Maniaci, M. (1993). Che fare del proprio corpus? *Gazette du livre médiéval*, *22*(1), 27–37.

Pinche, A. (2021). *Édition nativement numérique des oeuvres hagiographiques "li seint confessor" de wauchier de denain d'après le manuscrit 412 de la bibliothèque nationale de france.* (Doctoral dissertation). Retrieved 2017-05-09, from http://www.theses.fr/s150996

Pinche, A. (2022a, 6). *Cremma medieval.* Retrieved from https://github.com/HTR-United/cremma-medieval DOI: 10.5281/zenodo.5235185

Pinche, A. (2022b, June). *Guide de transcription pour les manuscrits du Xe au XVe siècle.* Retrieved from https://hal.archives-ouvertes.fr/hal-03697382

Pinche, A., Bureau, B., & Nicolas, C. (2016). Hyperdonat, digital edition project. In *Tei conference and members' meeting 2016.*

Pinche, A., & Camps, J.-B. (2022). Cremmalab project: Transcription guidelines and htr models for french medieval manuscripts. In *Colloque" documents anciens et reconnaissance automatique des écritures manuscrites".*

Pinche, A., Camps, J.-B., & Duval, F. (2021). *Création de modèle (s) htr pour les documents médiévaux en ancien français et moyen français entre le xe-xive siècle séance 3: L'allographie, entre besoins scientifiques et pragmatiques. comment modéliser et optimiser les données d'entraînement pour l'htr (i)?*

Pluta, O. (2020, 12). 9Abbreviations. In *The Oxford Handbook of Latin Palaeography.* Oxford University Press. DOI: 10.1093/oxfordhb/9780195336948.013.109

Ströbel, P. B., Clematide, S., & Volk, M. (2020). How much data do you need? about the creation of a ground truth for black letter and the effectiveness of neural ocr.

Vlachou-Efstathiou, M. (2022a). *Voss.lat.o.41 - eutyches "de uerbo" glossed.* HTR United.

Vlachou-Efstathiou, M. (2022b, 6). *Voss.Lat.O.41 - Eutyches "de uerbo" glossed.* Retrieved from https://github.com/malamatenia/Eutyches

White, N., Karaisl, A., & Clérice, T. (2022). *Caroline minuscule by rescribe.* HTR United. Retrieved from https://github.com/rescribe/carolineminuscule-groundtruth

Wills, T. (2016). The medieval unicode font initiative. *Medieval Unicode Font Initiative.*

Witt, J. C. (2016). Digital scholarly editions as api consuming applications: lessons and examples from the sentences commentary text archive and lombardpress. *Digital Scholarly Editions as Interfaces*, 24.

Appendix

	Ground Truth		Prediction
1	ouum faciet dñs sup terram. mulier circumdabit vi	1	ouum faciet dñs sup cerram. mulier cicumdabit uix aa-
2	gremio uteri sui .Ieremie xxxi. .Vere nouũ fecit doi᷑	2	I gremiq uteri sui Jeremie .xxx .i. Isere nouũ fecit an᷑
3	omnibus hominib; mirande multum .Et nota in annuñcciacõ	3	Ionsmbus hominib; nurande sisulcuti. Et nata ni aminẽti ac̃i
4	qd scdm triplicem terram tria fecit dñs genera marie.	4	qd scdm triplitem terram tia fecit dñs genera marie:
5	nouo Dicitur .ii. terra bt̃a uirgo maria. terra dicitur ip	5	nouor pititur .n. terra bt̃a uirgo maria. terra dicitur ip
6	sa ciuitas bethleem .terra dici romanũ imperiũ in qualib;	6	sa tiuitas bethleem. tei ra diti romanũ mperiũ inqualib;
7	istaꝫ terra fecit dñs nouũ Dic g Nouũ faciet dñs et cetera.	7	istaꝫ teri ax fecit dñs nouũ ddit g gonũ faciet dñs ceta.
8	De prima terra in qua dñs fecit magna .i. in bt̃a ugine ma	8	de prima tenia inqua dñs fecit magna .i. mnbt̃a ugine ma
9	ria dicit ꝑs. Bñdixisti dñe terram tuam. Vere bñdixit eã	9	ria ditit ꝑs. bñdixisti dñe terram tuami. Vere bñdixit eã
10	deus mũdando ab originali pc̃o. in utero materno ipsam	10	deus mũdando aborignnali pc̃o inucero materno ipsam
11	sc̃ificando ȝ eam donis celestibu; replendo | Iuxta illud Ecc̃.pt᷑	11	sc̃ificando ȝ eani douis celestib; replendo suxta iud occ̃. pc᷑
12	hec deus in terram aspexit ȝ repleuit eam donis suis De᷑	12	hec deus interram asperit ȝ repleuit eam donis suis de᷑
13	bt̃am uirgĩem donis celestib; repleuit quando in eam de	13	bc̃am uil. gĩem donis telestib; repleuit quand ineam de
14	scendit ȝ ex ea deus ȝ homo nasc̃i uoluit .vnde sc̃a fuit an	14	scendit ȝ exea deus ȝ homa nasti uolurt. unde sc̃a fuit an
15	teꝺ nata .prima eñ terra .s. eua fuit maledicta ȝ ideo tri	15	teꝺ nata. prima eñ terra .s. eua fuit maledicta ȝ ideo ni
16	bulos ȝ spinas germinauit s; hec terra est bt̃a ugo ma	16	hulos ȝ spmas perminauit s; hec terra est bt̃a ugo ma
17	ria bñdicta .Luce .Bñdicta tu in mulierib᷑ ȝc̃ Vn de iꝑa dicit	17	ria bñdicta. Lice. Bñditta tu in mulierib᷑ ȝt.vn de iꝑa dicit
18	Ecciastici .i. Generacio pterit ȝ gn̈acio aduenit terra ue	18	Eccliastici .i. Eeneracio pterit ȝ gn̈acio aduenit terra nie
19	ro in eternũ stat .Generacio angeloꝝ p̃terit fugiendo et	19	ro meternũ srati feneracio angeloꝝ p̃terst fugiend et
20	credendo ȝ generacio aduenit .s. latro ꝯfitẽs xp̃m .ra ů	20	credendo.ȝ gener atia aduenit .s. latra ꝯfitẽs xp̃u tra ů
21	.s. bt̃a uirgo in eternu stat .i. ꝑmansit ȝ ꝑmanet ĩmobi	21	s.bt̃a iurgo ineterмt stat .i. ꝑmansit ȝ ꝑmanet ĩ mabi
22	lis quia fundata erat sup nchilũ pautatis .Vñ iob fũ	22	lio quia fundata erat sup nchilũ pauptatis uñ iob fũ
23	dauit terram sup nichilum .s. hui᷑ ůgĩs pautatis Pau	23	dauit terram sup nichilum .s hur᷑ ůgĩs pauptders pdu
24	tas eius patuit | quando filiũ eius pannis inuoluit.	24	ptas eiua patuit quanto aliũ eius pannis muoluit
25	In hac terra ȝ sup hanc terram fecit dñs multa noua et	25	In hat terra ȝ sup hant terram fecit dñs multa noua et
26	p̃cipue qn̈q;.l ꝑmum nouũ qd ꝯcepit deũ ȝ hõiem Unde ꝓ	26	p̃tipue qn̈q; ꝑꝑmum nolũ qd ꝯcepit deũ ȝ hõiem unde ꝓ
27	uerbio Diues ȝ pauꝑ obuiauerũt sibi Diues ȝ pau deus	27	ner bior Diues ȝ pauꝑ obuiauerũt sibi diues ȝ pdup deus
28	ȝ homo obuiauerũt sibi in utero ugĩali Ps homo natus	28	ȝhono obulauerũt sibi in utero ugĩali ps. Honio natus
29	est in ea ȝc nõne magnũ nouũ ȝ admirandũ oĩb; quia ĩ	29	est in es ȝc nõne magnũ nouũ ȝ admirandũ oĩb; quiat̃
30	virgĩe yma sũmis sociant᷑. pater in filio qui oĩa creauit	30	virgĩt ima sũmis sociant᷑ pater infilio qui oĩa creduit
31	ex nichilo solo uerbo ꝯcludit᷑ in ugĩs utero O qn̈ta est benig	31	exnichilo salo uerbo ꝯdudit᷑ in ugĩs utero O qn̈ta est benig
32	nitas ȝ huilitas. regẽ uelle fieri seruũ. panẽ angeloꝝ lacte	32	nitas ȝ hilitas.regẽ uelle fieri sernũ panẽ augeloꝝ lacte
33	uginis modico pasci .ůbũ in utero esse incarnatũ. leticiam	33	uginis modico pasci. ůbũ mutero esse incarnatũ leticiam
34	flere. regem omniũ regũ esurire .lassari ȝ mestum ẽe. Ecce	34	fere. regem ommũ regũ esurure. lassari ȝ mestu ẽe f cce
35	ꝑmum nouũ ȝ ualde magnũ .Sc̀m nouu ÷ quia nouo modo	35	ꝑmu nouũ ȝ ualde maam cdm nonũ ÷ quia nouo modo
36	cõcepit .s. fidem ȝ amorẽ caritatis. Vñ ysaie ix. Egredieť	36	cãcepit .s. pfidem ȝ amorẽ caritatis vñ ysaie ic. Carediet᷑
37	uga de radice iesse ȝ flos de radice ei᷑ ascedet, quia xp̃c concep	37	uga de radice iesse ȝ flos deradice es᷑ ascodet quia xp̃t toncep
38	tus ÷ incendio dilecciois ȝ feruore caritatis. ideo dicit hu	38	rus icendio dilecciois ȝ feruore cariatis. ideq dicit hie
39	go de sancto Victore Nam quia ĩ corde ei᷑ amor singularis	39	go de sancto uictor e slam quia ĩ corde ei᷑ amor singularis

Table 12: Ground-truth (left) and prediction (right) of the new model on UBL Mss. 758, 24r. Yellow highlighting shows the differences between transcriptions.