CREMMA Medii Aevi: Literary manuscript text recognition in Latin

Thibault Clérice, Malamatenia Vlachou-Efstathiou, Alix Chagué

Journal of Open Humanities Data

CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin

RESEARCH PAPER

THIBAULT CLÉRICE
MALAMATENIA VLACHOU-EFSTATHIOU
ALIX CHAGUÉ
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper presents a novel segmentation and handwritten text recognition dataset for Medieval Latin from the 11th to the 16th century. It connects with Medieval French datasets, as well as earlier Latin datasets, by enforcing common guidelines, bringing 263,000 new characters and now totaling over a million characters for medieval manuscripts in both languages. We provide our own addition to Ariane Pinche's Old French guidelines to deal with specific Latin cases. 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 Old French base model on Latin datasets, improving accuracy by 5% on unknown Latin manuscripts.

CORRESPONDING AUTHOR:
Thibault Clérice
Centre Jean Mabillon, École nationale des Chartes, PSL University, Paris, France
thibault.clerice@chartes.psl.eu

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

TO CITE THIS ARTICLE:
Clérice, T., Vlachou-Efstathiou, M., & Chagué, A. (2023). CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin. *Journal of Open Humanities Data*, 9: 4, pp. 1–19. DOI: https://doi.org/10.5334/johd.97

1 CONTEXT AND MOTIVATION

Institutional and academic contexts Handwritten text recognition (HTR) and its upstream task layout segmentation (LS) 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 truths (GTs, data that will be used for training), training new models (for the transcription and, 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 or infrastructure requirements.[2] At the end of the 2020 European Union funding, Transkribus became a paid service accelerating 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. The CREMMA funding consisted of a grant for the initial cost of the infrastructure as well as an evaluation grant for providing base models for the community of CREMMA's users. 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) 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 in an HTR context and providing machine-actionable metadata for datasets. *HTR-United*, both a catalog of open source HTR ground truths and a toolkit to strengthen the control of documentation and validity of HTR data, records, as of late October 2022, 56 datasets composed of 41.5 million characters, 725,862 lines in over 13 languages and 6 scripts. *HTR-United*'s catalog provides a useful overview to build new datasets which can complement previous ones.

HTR for Latin and Old French Handwriting in the Middle Ages can, in a simplistic way, be divided into two big writing systems: cursive and calligraphy (Bischoff, 1985, pp. 58 sqq.). They reflect two complementary practices, namely cursive hands (*écritures d'usage*), which are more common to everyday and administrative documents such as accounting books and letters, and book hands.[3] While cursive represents a harder challenge due to the variability of handwriting styles, both families have the potential to be 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. This situation resulted in mainly two different kinds of strategies for creating *HTR* ground truth datasets: (1) datasets that would resolve abbreviations directly in the transcription (a practice found mostly used by historians, and quite common for cursive, specifically in France) and (2) datasets that would keep a diplomatic approach to transcription.

Our dataset builds on the experience of Ariane Pinche, 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. As the first recipient of the CREMMALab post-doctoral funding, Pinche co-organized a research seminar around the formalization of transcription guidelines for graphemic transcription of Old French (Pinche, 2022c). Based on her recommendations, a few datasets emerged around the École nationale des chartes and the CREMMA project. Notably, the Gallic(orpor)a corpora (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 (Hawk, Karaisl, & White, 2018) was realigned in ALTO XML and adapted to the guidelines as it provided some foundations for recognizing the

1 Gallery Library Archives Museums.

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

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

Caroline script specific to the first centuries of the early middle age. Vlachou-Efstathiou (2022a,b) provided a complementary dataset based on the transcriptions of two Latin manuscripts from the 9th century. When the work for *CREMMA Medii Aevi* began, we identified a lack of data for the second half of the middle age (1100–1500, see Table 1).

AUTHORS	DATASET	PROJECT	CHARACTERS	PERIOD	LANGUAGE
White, Karaisl, and Clérice (2022)	Caroline Minuscule	Rescribe	17,000	800–1200	Latin
Vlachou-Efstathiou (2022a)	Eutyches	–	87,000	850–900	Latin
Pinche (2022a)	CREMMA Medieval	CREMMAlab	**593,000**	1100–1499	French
–	**CREMMA Medii Aevi**	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' counts are rounded to the closest thousands.

2 DATASET DESCRIPTION

Object name: CREMMA-Medieval-LAT-0.1.1.zip

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 (http://dx.doi.org/10.5281/zenodo.7013436)

Publication date: 2022-10-20

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. 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. Sampling criteria and focal variables correlate. 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 represents exclusively the written Latin language while giving room to texts of multiple functions addressed to different audiences belonging to various genres (while not aiming at exhaustiveness at this stage). 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. Estimated abbreviation rate and use of specific characters, known genres and scripts were implemented to compensate for what was thought to be missing from the network of the corpus and the corpus itself in order to make it as "representative" as possible. HTR engines are language agnostic, but the same cannot be said for the resulting models, which means that it depends on the representativeness of the sample to determine whether a model will work on "similar" or "out-of-domain" documents.

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 classical texts such as Seneca's *Medea*. Script was not taken into account.
2. In a logic of complementarity, the second part of the corpus was dictated inversely by content. More specifically, given the relative absence of ligatures and abbreviations in classical texts, we chose documents that would display a higher degree of abbreviations. This both induced or led to a genre selection process, specifically for medical and scholastic data. At the same time, 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 alignment of the Eichenberger and Suwelack (2021) dataset, produced in the context of a transcribathon in Berlin and containing genres new to our corpus (Book of Hours, Psalms, etc.).

Quantitative aspects of the corpus 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 HTR model applicable to as large a range of unseen manuscripts as possible is undoubtedly the end goal. 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.

More specifically, while conducting an experiment on Caroline Minuscule OCR models, Hawk et al. (2018) conclude that "relative preponderance"[4] 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 be 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 the 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 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 training set was limited to 5 pages per script variation (depending on the density of the layout),[5] examining whether this balance can contribute to the production of generic models.[6]

4 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.

5 Aside from 3 documents coming from the *Faithful transcriptions dataset*, that were utilized rather as evaluating tools at the end of the project.

6 On GT size for OCR experiments see Ströbel, Clematide, and Volk (2020).

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. Alongside text recognition, eScriptorium allows for layout annotation using ontologies and controlled vocabularies. For this, researchers need to agree on a limited common vocabulary and share common practices to facilitate the interoperability of their ground truth.

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 small/restricted 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 Pinche and connected to the CREMMALab project such as Gabay et al. (2022), resulting in segmentation models mainly for late medieval manuscripts and early prints.[7] As per the *CREMMA Medii Aevi* 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 the first is completely standardized, 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 utilized, such as `MainZone` for columns and `MargintextZone` for *marginalia*.

Pinche's Transcription Guidelines Pinche (2022c) stressed that HTR was an answer to the need for scientific projects to acquire textual data either to undertake editions or to constitute large corpora. Her guidelines address the need to establish principles common to projects dealing with the transcription of manuscripts in order to:

- build shareable, reusable, and durable ground truth data sets;
- produce robust generic models, reusable in "out-of-domain" manuscripts;
- minimize the collective cost, including that of training people;
- build GT that seeks to optimize the learning space of HTR models.

Pinche has privileged a graphemic transcription, which reproduces graphemes, *i.e.* a canonical form for each character, instead of a graphetic one, which tries to reproduce each variation of a letter (such as ſ and s).[8] 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). 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: for example, for every manifestation of the paragraph sign, we opt for the pilcrow sign "¶" (U+00B6) on every occasion, instead of several variations such as "" (U+F1E1).[9] In the context of the guidelines, we set up a list of allowed characters and a list of common and rare cases (such as Table 2 and 3).

On the topic of abbreviations, resolving them produces specific difficulties for HTR engines, as it leads them to learn more about the language than they originally intended.[10] Abbreviations

7 More information and case studies can be found here: https://segmonto.github.io/.

8 See Pinche, Duval, and Camps (2022). On this particular topic, Gueville and Wrisley (2022) and Pinche took different paths. However, while graphemic transcription cannot be turned and reused for graphetic model training, graphetic transcriptions can be turned easily into graphemic GT, at the cost of establishing a "translation" table for each character.

9 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.

10 Most HTR engines learn directly from transcriptions and do not include a separate mechanism for abbreviation resolution or spotting. Transcriptions produced by these models are thus not showing where an abbreviation was resolved, making it difficult to distinguish HTR errors from abbreviation resolution errors. The stakes do not specifically concern the scores, which seem to be close to each other (Camps, Vidal-Gorène, & Vernet, 2021), but the long-term use of ground truth data and silver data in a sustainable way.

are not resolved in our dataset, as this constitutes rather an interpretative act linked to the specificity of each document. 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. Pinche's graphemic approach without abbreviation resolution simplifies the interpretation step of the text, and in turn, the reduction of characters diversity ultimately smooths both the human transcriber and the HTR engine's learning curves.

In order to ensure the rigorous application of these guidelines and the homogeneity of the data produced, we introduced quality control to the production and publication workflow. Each manuscript transcription was passed through ChocoMufin (Clérice & Pinche, 2021), using project-provided character translation and control 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 (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 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 MEDII AEVI

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

The project adheres to the general principles laid out by Pinche (Pinche, 2022c, Tables pp. 4–15) concerning the base cases (punctuation, word separation, functional signs, superscript letters, abbreviations, ligatures, and roman numerals). Using the project-provided character conversion table, ChocoMufin controls the transcription and corrects any anticipated error by transforming the character automatically so it conforms to the pre-defined guidelines (data should be used in their post-ChocoMufin converted state as it sometimes corrected mistranscription). 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.[12]

In general, the main differences that we isolated between the CREMMA Medieval and Medii Aevi 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);[13]
2. no normalization or distinction of `u` and `v` was provided, nor of `i` and `j`;
3. two variations of *con* are found, namely the antisigma and the 9-shaped form;
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 example.

11 In order to read the translation table, MUFI-compatible fonts are recommended, such as Junicode.

12 This includes discussion with other projects, such as Gervers, Manton, Boutreux, and Elema (2018), which led to the inclusion of the stricken-through D (see Table 3).

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

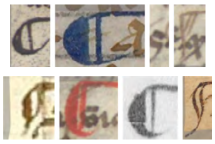
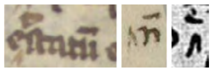
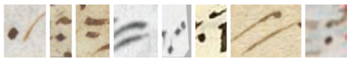
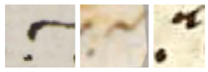
TYPE	TRANSCRIPTION	UNICODE	DESCRIPTION OR RESOLUTION	EXAMPLES
Punctuation	¶	U+00B6	Content change	
Punctuation	-	U+002D	Hyphenation	
Punctuation	/	U+2215	Diastole	
Reference mark	‸	U+2038	Omission sign 'caret'(reintroduction of content)	
Punctuation	:	U+003A	Punctus elevatus	
Punctuation	:	U+003A	Punctus 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 (2022c). Thus, macrons and generally horizontal lines diacritics over the letter such as tildes are represented by combining horizontal tildes, and any vertical zigzag and similarly shaped forms are simplified into combining 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), as long as it was a semantic feature and not a decorative one.

Abbreviation marks significant in themselves: "Standard" Abbreviations signs have been preserved as such, like *pr(a)e* -p̃ (p + combining tilde, p + U+0303), *pro* -ꝓ (U+A753), *hoc* - ħ (U+0127), ẜ (s with diagonal stroke, U+1E9C) for *secundum* or *ser-*, ꝯ for 9 shaped *con/cum* (U+A76F), Tironian sign ꝰ for the desinence *-us* (U+A770),᷑ for *(t)ur* (U+1DD1), and Ꝙ / ꝙ for *quod*. Absent from the *CREMMA Medieval* but present in *Medii Aevi*, the truncated ending *-is* is transcribed using the character ꝭ (U+A76D). The "inverted c" variation of the preposition con/cum is a good example for the difference of approach between the graphetic and graphematic approach: while using the *antistigma* (ɔ) is more faithful, it simply is an allograph of the original ꝯ. For *-rum*, the symbol ꝵ is used rather than the rotunda -rum ꝝ (U+A75D).[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-shaped ; sign (U+F1AC), avoiding the private domain ligature specific ꝙꝫ (U+E8BF) character but also avoiding confusion with the regular semi-colon.

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 ≈ (U+2248). The Division sign ÷ is used ubiquitously for the abbreviation sign of *est/id est*. Tironian *et* (U+204A,

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".

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 (see Table 4).

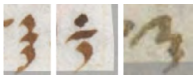
CHARACTER(S)	UNICODE	RESOLUTION	EXAMPLES
⁊	U+204A	Et	
⁊+◌̃	U+204A + U+0303	Etiam	
ſ	U+A76D	–is	
đ	U+0111	d + any desinence truncation	
ꝯ	U+A76F	con	
≈	U+2248	esse	
÷	U+00F7	est/id est	
;	U+F1AC	-que/-bus/-m/-et	
ꝝ	U+A775	-rum	

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 datasets.

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, avoiding the MUFI sign ꝗ that belongs to the private domain. More examples are provided in Table 4.[15]

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 (Pinche, 2022c, p. 11). *Ergo* and *igitur* are two of the most frequent examples of abbreviations with superscript letters. Letters without any baseline letter are simply represented with the same combining superscript character and space as the supporting baseline character (*e.g.* "ᵃᵗ": space + combining a + space + combining t *cf.* Figure 1).

Superscript letters, alongside abbreviating functions, were sometimes used to render interlinear additions. Missing content or annotations are added in the interlinear space, especially in manuscripts of scholastic and medical content. This was something that was at first a challenge for the transcription process due to segmentation constraints. It can be, at times, impossible to completely differentiate the segmentation masks of two vertically adjacent letters (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).

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".

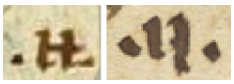
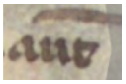
TYPE	TRANSCRIPTION	UNICODE	DESCRIPTION OR RESOLUTION	EXAMPLES
Ligature	st	–	Normally transcribed ligature	
Ligature	.n.	–	enim	
Ligature	qr	–	quia	
Monogrammatic Ligature	qd	–	quod	
Monogrammatic ligature	Et	–	Et	
Contraction	aũt̃	–	Long vertical tilde transcribed by two tildes	
Contraction	ẽẽ	–	Long vertical tilde transcribed by two tildes;	
Contraction	tp̃a	–	Two stacked tildes	

Table 4 Ligatures and special contraction cases.

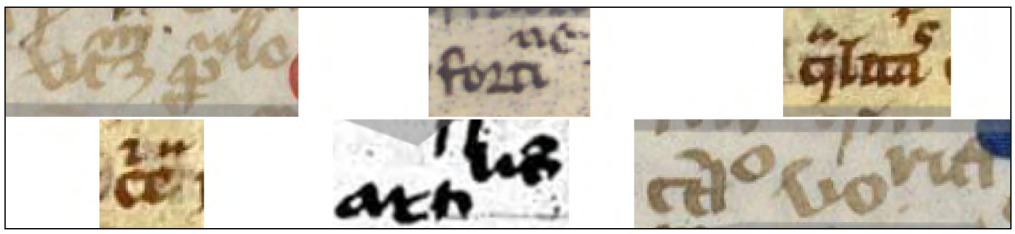


Figure 1 Examples of contraction use of superscript letters. Manuscripts in the following order: BIS 193, CML 13027, Montpelier H-318, Montpelier H-318, Vat. Pal. lat.373, BIS 193.

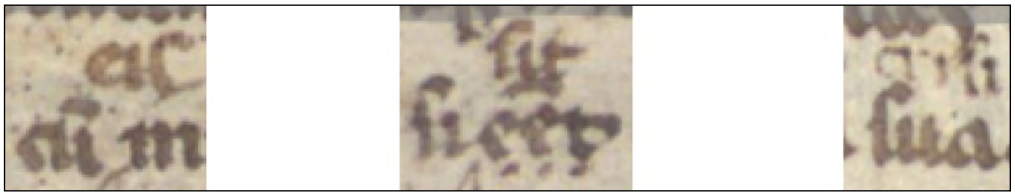


Figure 2 All examples come from the CML 13027 manuscript.

Rare characters and Numerals Referring to corpus construction practices for balanced corpora, Maniaci (1993) 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 *Medii Aevi* dataset are therefore given special attention (*cf.* Table 5). 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, ʒ (U+2125) and ℒ (U+10192) were used. "Barred O" is represented by ⌀ (U+2205) and is widely used to transcribe the word *instans* instead of ꝋ (U+A74B) that, according to MUFI documentation stands for the abbreviation of *obi(i)t* (Coulson & Babcock, 2020, p. 10).

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

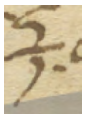
TYPE	TRANSCRIPTION	UNICODE	DESCRIPTION OR RESOLUTION	EXAMPLES
Symbols	ʒ	U+2125	Ounce	
Symbols	ℒ	U+10192	*Semi-Ounce	
Abbreviations	⌀	U+2205	instans	

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

Figure 3 Manuscripts in the following order: Latin 16195, Phi. 10 a. 135 (x3), BIS 193, CML13027, Egerton 821, Latin 6395.

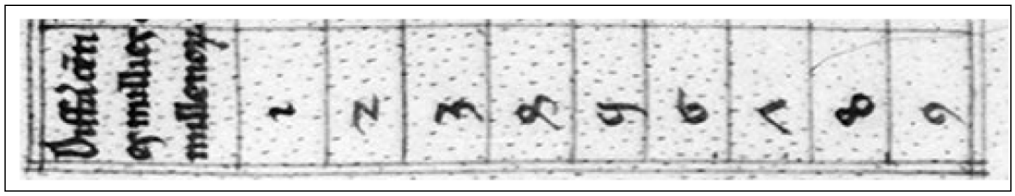


Figure 4 Snippet of Arabic numerals from BnF, lat.15461, fol.13r for comparison purposes.

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, and finally the text is transcribed. Once each sample is entirely annotated, its use of characters is controlled via the ChocoMufin software, while its conformity to the segmentation classification vocabulary is controlled by HTRVX. Finally, data are released on Github.[16] All the combining and abbreviation signs suggested for use by the present adaptation of Pinche's guidelines can be also found on a custom-made eScriptorium keyboard configuration, in order to facilitate reuse and compatibility with the guidelines.[17]

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 a wide 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 not to be 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* texts. Medical and scholastic genres, furthermore, induce the use of a range of rare characters and often underrepresented letters (such as "z", "y" and "k").

16 https://github.com/htr-united/cremma-medieval-lat.

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

Other features, such as layout and type of digitization (microfilm or original), provide 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 timespan of 5 centuries separates 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,[18] including personal hands alongside formal categories such as the one described by Rossi (2022), with different levels of execution (cursivity and formality).

Table 6 Basic features and length of the dataset in chronological order. Medic. stands for medical, Lit. for literature, 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. The complete metadata table can be found in the more detailed `data-registry.csv` of the dataset.

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	Lit.	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	Lit.	1300–1310	Color	Textualis Meridionalis	Sequential	low
Arras 861	5	Lit.	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	Lit.	1325–1399	Microfilm	Semitextualis Libraria	Sequential	low
Laur. Plut. 39.34	5	Lit.	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	Lit.	1471–1499	Microfilm	Humanistica Cursiva	Random	low
† CCCC MSS 165	5	Schol.	1500–1599	Color	Personal Cursive	Sequential	medium

18 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).

Character frequencies in the CREMMA Medieval and the Medii Aevi datasets We set up this corpus to both complement the CREMMA Medieval 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.

Unlike *CREMMA Medieval*, our approach has been feature-driven to compensate for rare characters in the dataset network. In this regard, we succeeded, as we have a higher frequency of special characters in our dataset than in Pinche's dataset, despite being smaller overall (see Table 7 and Figure 5). Only three characters are more represented in *CREMMA Medieval*: the Tironian Et, the superscript combining R (common on words such as "grand"), and "&". The character ꝯ is equally present in both datasets: resolved as con- or com-, it is often used in words such as ꝯmence (*commence*). Some very frequent diacritics, such as the horizontal lines and vertical lines transcribed as tildes, are more frequent in our dataset, by a factor of 2.51 for horizontal ones and of 3.93 for vertical ones. 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 4,500 for the vertical one, making them the first and the third most represented abbreviating characters.

LANG	TYPE	WORDS	WORDS %	UNIQUE WORDS	UNIQUE WORDS %	FREQ. OF UNIQUE WORDS > 1
Latin	abbr.	6,855	11.94%	1,460	6.24%	279
Latin	others	50,557	88.06%	21,935	93.76%	5,025
Old French	abbr.	5,755	4.15%	1,457	4.89%	286
Old French	others	132,828	95.85%	28,315	95.11%	8,726

Table 7 Comparative statistics table on abbreviations: for each dataset, we look at words that are abbreviated (abbr.) or non-abbreviated (others). It reads the following way: "11.94% of words in the Latin corpus are abbreviated."

CHARACTER	UNICODE	LATIN	OLD FRENCH	% IN LATIN	RATIO
⁊	U+204A	2228.0	4400.0	33.61	0.51
◌ᷜ	U+036C	148.0	219.0	40.33	0.68
&	U+0026	83.0	116.0	41.71	0.72
ꝯ	U+A76F	850.0	779.0	52.18	1.09
ꝑ	U+A751	1500.0	919.0	62.01	1.63
◌ᷣ	U+0365	1486.0	820.0	64.44	1.81
◌̄	U+0303	14445.0	5759.0	71.50	2.51
◌ᷓ	U+0363	2024.0	732.0	73.44	2.77
ꝰ	U+A770	1763.0	523.0	77.12	3.37
◌ᷤ	U+033E	3827.0	973.0	79.73	3.93
◌ᷛ	U+0364	518.0	120.0	81.19	4.32
ꝓ	U+A753	462.0	80.0	85.24	5.78
◌᷑	U+1DD1	1018.0	137.0	88.14	7.43
◌ᷠ	U+1DE4	978.0	55.0	94.68	17.78
◌ᷚ	U+0366	870.0	61.0	93.45	14.26

Table 8 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 Medii Aevi. Despite this difference, most abbreviated characters are more represented in the Latin dataset.

Some manuscripts have nearly no abbreviation (*cf.* Table 9). Laur. Plut. 39.34 notably so, as it only contains 3 abbreviated words which is a single character abbreviation (⁊, *et*). A little less than half of our manuscripts are less abbreviated than the most abbreviated text in the *CREMMA Medieval* dataset, while the other half can exceed it by up to ten points. However, both languages show similar maximum frequencies in terms of non-single letter abbreviations (abbreviations made up of a single Unicode codepoint such as ⁊, &, ꝑ).[19]

19 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.

Finally, despite showing a similar number of *pages*, we see a large variation in terms of word density with a limited variation in terms of unique words (*cf.* Table 8). 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 potentiality 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).

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 represents the highest abbreviation ratio in the CREMMA Medieval dataset.

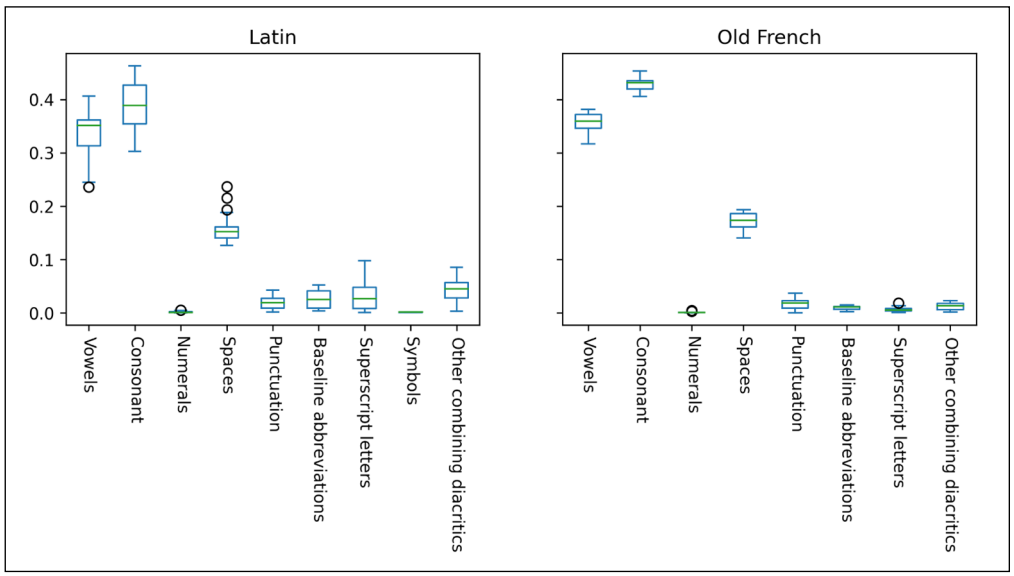


Figure 5 Frequences of character classes across manuscripts.

5 IMPLICATIONS/APPLICATIONS

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

HTR data and models have a fairly high level of reuse potential. First and foremost, while it is still relatively a rare reuse, these data, visualised correctly, can easily serve as teaching materials: e-teaching of paleography has been gaining some traction,[20] but simply moving away from printed to digital and interactive hand-outs using open data and transcription is a first step that undoubtedly some have already taken.[21] Then reuse can move to the analysis of transcription themselves: Stutzmann (2018) and Stutzmann, Mariotti, and Ceresato (2020) have shown that analysis of graphematic data can yield information about scribal practices. Finally, such data can be used for model training. Project like Possamaï, Gaiffre, Souvaye, Duval, and Ducos (2022) and Foehr-Janssens, Ventura, Carnaille, and Meylan (2021) have used automatic transcription models to speed-up the transcription process of large collection of manuscripts, using base models which were then fine-tuned on sample of data to yield better results, such as described by Pinche (2022b, 4.4). Finally, models can be used for data-mining and operating research at scale on non-manually transcribed manuscripts: Camps, Clérice, and Pinche (2021) proved the hypothesis of a 19th-century scholar by analysing a full manuscript with automatic transcription, Franzini et al. (2018) proposed also a stylometrical analysis of data obtained through automatic transcription.

As a direct output, we trained a model which would allow for transcribing or starting the transcription of Latin medieval manuscripts. In order to evaluate the gain from our data, we trained three models:[22]

1. a model containing all data from 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*.

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

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 CREMMA Medii Aevi	94.04	80.68	67.68	78.02	81.89
Only Old French	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 CREMMA Medii Aevi* contains everything but the present dataset, model *Only Old French* contains all datasets but Latin one (Eutyches, Caroline, CREMMA Medii Aevi). 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 Medii Aevi but not used for training purposes.

20 See Brookes, Stokes, Watson, and De Matos (2015), Burghart (2011) and the interactive facsimiles of http://theleme.enc.sorbonne.fr/dossiers/index.php. *Ad fontes* Hodel and Nadig (2019) has implemented training material in paleography also via interactive facsimiles: https://www.adfontes.uzh.ch/en/3001/training/einleitung.

21 We know at least of Olivier Canteaut at the Ecole nationale des chartes who has been using it for this purpose.

22 All models are trained with Kraken 4.1.2. Parameters are the base one of this version, as well as the following: NFD Unicode normalization (-u NFD, augmentation of data through *albumentations* (--augment), batch size 16 (-B 16), fixed splits (--fixed-splits), learning rate 0.0001 (--lrate 0.0001), model architecture [1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do0.1,2 Lbx200 Do] (--spec "[...]")

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 CREMMA Medieval 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.

MODEL	TEST	[SPACE] %	[SPACE]	TILDE	VERT. TILDE	7	ꝰ	ꝯ	ꝑ	ħ	ꝥ	ꝓ	ꝑ̃	ꝛ	;
All	CREMMA Medieval	1.7	803	77	46	0	10	17	15	0	0	0	4	0	0
No CREMMA Medii Aevi	CREMMA Medieval	1.7	726	89	55	0	15	12	15	0	0	0	3	0	0
Only Old French	CREMMA Medieval	1.7	733	86	50	0	12	15	20	0	0	0	2	0	0
All	CREMMA Medii Aevi	1.7	74	27	31	0	3	2	3	0	8	2	2	0	1
No CREMMA Medii Aevi	CREMMA Medii Aevi	2.8	138	78	92	0	17	8	11	1	17	15	2	15	32
Only Old French	CREMMA Medii Aevi	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 CREMMA Medii Aevi	BGO	2.3	13	3	0	1	1	0	0	0	0	0	0	1	2
Only Old French	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 CREMMA Medii Aevi	BH25	1.8	73	68	21	0	4	0	9	0	5	1	2	3	12
Only Old French	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 CREMMA Medii Aevi	UBL	6.0	482	256	76	0	38	0	69	11	37	7	16	71	71
Only Old French	UBL	5.7	484	239	77	0	28	0	59	11	37	7	15	71	71

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 *Medii Aevi* provides overall better results on out-of-domain datasets: UBL Mss. 758 and Berlin, Hdschr. 25 transcriptions improved by 4.2 points at least while Berlin, Germ. Oct. 511 (BGO), the smallest transcription set of the dataset, only improved by 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 *CREMMA Medieval*. We actually see in Table 11 that there are much fewer errors on characters that saw their frequencies reach new highs. The *All* model does only a fourth of the error of the *Only Old French* model on tildes or two-thirds on vertical tildes for the UBL manuscript. The -rum abbreviation (ꝝ) or the -et/-ed/-ibus one (;) are quite new to Medieval datasets in general, which explains the clear difference in results. Overall, this dataset helped create a model allowing for readable outputs (see Appendix Table 12 for a side-by-side comparison) on Medieval manuscripts, or at least transcriptions that can help produce new data.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Appendix Table 12.** Ground-truth (left) and prediction (right) of the new model on UBL Mss. 758, 24r. Yellow highlighting shows the differences between transcriptions. DOI: https://doi.org/10.5334/johd.97.s1

ACKNOWLEDGEMENTS

Several transcriptions are the product of alignment and adaptation of existing projects that have worked on the manuscripts in question. In the case of existing digitized transcriptions, an alignment and correction were performed. In the case of printed editions, they served as a guide for obscure passages and *dubia*.

- 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 Witt (2016).[23] In the case of *dubia*, additional corrections have been made for the faithful reproduction of the abbreviations;
- 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.

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

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 authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Thibault Clérice and Alix Chagué were responsible for the project administration. TC super- vised 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.

AUTHOR AFFILIATIONS

Thibault Clérice orcid.org/0000-0003-1852-9204
Centre Jean Mabillon, École nationale des Chartes, PSL University, Paris, France

Malamatenia Vlachou-Efstathiou orcid.org/0000-0002-9397-356X
École nationale des Chartes, PSL University, Paris, France

Alix Chagué orcid.org/0000-0002-0136-4434
INRIA, CNRS, Paris, France

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. DOI: https://doi.org/10.4135/9781849209731.n2

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: https://doi.org/10.5281/zenodo.6126376

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

Brookes, S., Stokes, P. A., Watson, M., & **De Matos, D. M.** (2015). The digipal project for european scripts and decorations. *Essays and Studies*, *68*, 25–59. DOI: https://doi.org/10.1017/9781782046073.003

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

Burghart, M. (Ed.). (2011). *Album interactif de paléographie médiévale/interactive album of mediaeval palaeography*. UMR 5648 CIHAM. Retrieved from https://paleographie.huma-num.fr/

Camps, J.-B., Clérice, T., & **Pinche, A.** (2021, 11). Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis. *Digital Scholarship in the Humanities*, *36*(Supplement 2), ii49–ii71. DOI: https://doi.org/10.1093/llc/fqab033

Camps, J.-B., Vidal-Gorène, C., & **Vernet, M.** (2021). Handling heavily abbreviated manuscripts: Htr engines vs text normalisation approaches. In *International conference on document analysis and recognition* (pp. 306–316). DOI: https://doi.org/10.1007/978-3-030-86159-9_21

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. DOI: https://doi.org/10.5209/CFCL.56186

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: https://doi.org/10.5281/zen-odo.5356154

Coulson, F., & **Babcock, R.** (2020). *The oxford handbook of latin palaeography*. Oxford University Press, USA. DOI: https://doi.org/10.1093/oxfordhb/9780195336948.001.0001

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.). DOI: https://doi.org/10.5281/zenodo.5582483

Foehr-Janssens, Y., Ventura, S., Carnaille, C., & **Meylan, A.** (2021). *« canoniser les sept sages » (c7s): livre, langues, écriture sérielle (xiiie-xve s.)*. FNRS. Retrieved from https://data.snf.ch/grants/grant/197853

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., ... Rybicki, J. (2018). Attributing authorship in the noisy digitized correspondence of jacob and wilhelm grimm. *Frontiers in Digital Humanities*, *5*, 4. DOI: https://doi.org/10.3389/fdigh.2018.00004

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

Gervers, M., Manton, A., Boutreux, A., & **Elema, A.** (2018). *Text as image, image as text* [Project]. Retrieved from https://www.utsc.utoronto.ca/research/prj/text-image-image-text-charter-integrity-and-topic-modelling (Funded by the Social Sciences and Humanities Research Council of Canada (SSHRCC))

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*.

Hodel, T., & **Nadig, M.** (2019). Grundlagen der mediävistik digital vermitteln: Ad fontes, aber wie? *Das Mittelalter*, *24*(1), 142–156. DOI: https://doi.org/10.1515/mial-2019-0010

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). DOI: https://doi.org/10.1109/ICDAR.2017.307

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). DOI: https://doi.org/10.1109/ICDARW.2019.10032

Maniaci, M. (1993). Che fare del proprio corpus? *Gazette du livre médiéval*, *22*(1), 27–37. DOI: https://doi.org/10.3406/galim.1993.1230

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

Pinche, A. (2022b, November). *Generic HTR Models for Medieval Manuscripts The CREMMALab Project*. Retrieved from https://hal.archives-ouvertes.fr/hal-03837519 (working paper or preprint)

Pinche, A. (2022c, 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., Duval, F., & **Camps, J.-B.** (2022, March). *Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français, Xe-XIVe siècles*. Retrieved from https://hal.archives-ouvertes.fr/hal-03615557 (working paper or preprint)

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

Possamaï, M., Gaiffre, B., Souvaye, G., Duval, F., & **Ducos, J.** (2022). *Liber, les décades de bersuire*. ANR. Retrieved from https://anr.fr/Projet-ANR-21-CE27-0008

Rossi, M. C. (2022). Scritture di dotti nell'università del duecento. In *Xxiie colloque de paléographie latine*. Retrieved from https://cipl.hypotheses.org/maria-cristina-rossi-univ-pisa

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.

Stutzmann, D. (2018, March). Variability as a Key Factor For Understanding Medieval Scripts: the ORIFLAMMS project (ANR-12-CORP-0010). In S. Brookes, M. Rehbein, & P. Stokes (Eds.), *Digital Palaeography*. Routledge. Retrieved from https://shs.hal.science/halshs-01778620

Stutzmann, D., Mariotti, V., & **Ceresato, F.** (2020, February). Les abréviations dans les manuscrits français du XIIIe siècle : analyses statistiques. In *L'emersione delle scritture volgari – L'émergence des écrits en langue vulgaire – The rise of vernacular writing La prospettiva paleografica – Le point*

de vue paléographique – The palaeographical perspective. XXI Convegno del Comité international de paléographie latine. Firenze, Italy. Retrieved from https://shs.hal.science/halshs-03560918

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.

TO CITE THIS ARTICLE:
Clérice, T., Vlachou-Efstathiou, M., & Chagué, A. (2023). CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin. *Journal of Open Humanities Data*, 9: 4, pp. 1–19. DOI: https://doi.org/10.5334/johd.97

Published: 12 April 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.