Generic HTR Models for Medieval Manuscripts. The CREMMALab Project

Ariane Pinche

Journal of Data Mining and Digital Humanities

Generic HTR Models for Medieval Manuscripts
The CREMMALab Project

Ariane Pinche[1]

[1]CNRS, CIHAM (UMR 5648)

Corresponding author: Ariane Pinche , ariane.pinche@cnrs.fr

Abstract

In the Humanities, the emergence of digital methods has opened up research to quantitative analysis and/or to publication of large corpora. To produce more textual data faster, automatic text recognition technology (ATR)[1] is increasingly involved in research projects following precursors such as the *Himanis* project. However, many research teams have limited resources, either financially or in terms of their expertise in artificial intelligence. It may therefore be difficult to integrate ATR into their project pipeline if they need to train a model or to create data from scratch. The goal here is not to explain how to build or improve a new ATR engine, nor to find a way to automatically align a pre-existing corpus with an image to quickly create ground truths for training. This paper aims to help humanists develop models for medieval manuscripts, create and gather training data by knowing the issues underlying their choices. The objective is also to show the importance of data consistency as a prerequisite for building homogeneous corpora and training more accurate models. We will present an overview of our work and experiment in the *CREMMALab* project (2021-2022), showing first how we ensure the consistency of the data and then how we have developed a generic model for medieval French manuscripts from the 13th to the 15th century, ready to be shared (more than 94% accuracy) and/or fine-tuned by other projects.

Keywords

HTR model dataset medieval text transcription

I INTRODUCTION

Since the 1980s, optical character recognition (OCR) has been used to automatically acquire printed text from images [Rice et al., 1993]. However, it was not until the late 2000s that handwritten text recognition (HTR) technologies began to become usable on medieval manuscripts and exploitable in the Humanities through the use of deep learning and neural network architecture [Graves and Schmidhuber, 2008, Fischer et al., 2010]. At the time, these technologies were mostly experimental and required the support of an expert in image processing and deep learning. Only projects with significant funding and expertise in the field could afford to use this type of technology, such as *Oriflams* [Stutzmann et al., 2017]. Since the late 2010s and early 2020s, huge improvements have been made with ATR engines able to handle handwritten documents, such as *Pylaia*, *HTR+*, *Kraken* and their interfaces, such as *Transkribus* and *eScriptorium* [Kahle et al., 2017, Kiessling, 2022, Kiessling et al., 2019]. Nowadays, models can reach a character

[1]ATR is used here as a global term for automatic text recognition in both prints and manuscripts.

error rate (CER) between 8% and 2% or better for manuscripts. "From a computer science point of view, the recognition of handwriting seems to be a resolved task. The latest recognition engines allow for the successful recognition of specifically trained hands producing a text as reusable data" [Hodel et al., 2021].

These advances have made ATR accessible to any humanities research project and have significantly reduced manual work in terms of text acquisition. Cultural heritage institutions today aim to digitize large-scale collections of historical documents. To enrich digital images or make them searchable, automatic text acquisition is the next step. Evidence of this growing use can be seen in the program of the conference *Ancient documents and automatic recognition of handwriting* (June 2022) or in the number of papers presenting projects using ATR in the DH 2022 and TEI 2022 conferences.[Committee, 2022, Cummings, 2022]

But the challenge of HTR for historical documents remains because of the wide variety of handwriting across time. The production of training data is now a major challenge to build efficient HTR models adapted to a given source. During 2021-2022, within the infrastructure of the *CREMMA* project[2] supported by the DIM MAP ("Matériaux Anciens et Patrimoniaux" research funded by Île-de-France Region), the *CREMMALab* project has been designed to share open training data and HTR models for medieval manuscripts between the 13th and the 15th century. All data and models produced are available in the *Cremma Medieval* repository [Pinche, 2022a] and listed in the *HTR-united* catalogue [Chagué et al., 2021]. During the project, transcription protocols have been put in place to optimize the production of homogeneous and shareable data. Through the gathering of a corpus of medieval manuscripts, the learning process of the HTR algorithms has been examined to evaluate the impact of the training corpus on the robustness and genericness of the models.

Regarding the need for HTR data and models in the community, this paper builds on the work and experience of the *CREMMALab* project to present some recommendations for the production of consistent data for medieval manuscripts that could be shared with other projects, as it is well known that the production of training data is extremely time consuming [Fischer et al., 2010]. It also offers an overview of the constitution of two generic models for medieval manuscripts: *Bicerin* and *Cortado*. The goal here is multiple: (i) to put in place methods for making, collecting and sharing more coherent ground truths, (ii) to propose a process for producing a generic model, and (iii) to help humanists understand genericness, scoring and fine-tuning of HTR models, so that they can reuse previous models in their own projects.

The paper is organized as follows. The next section (II) presents backgrounds of HTR, training process and related works. Section III provides guidelines for data building through the example of the *Cremma Medieval* dataset. Section IV describes methods used to train HTR models with *Kraken* and their results, but also how to fine-tune them on a given corpus. The results will be analysed in Section V. Finally, Section VI summarizes the work presented and draws conclusions.

II BACKGROUND AND RELATIVE WORKS

[2]Consortium for Handwriting Recognition of Ancient Materials.

2.1 General principles of the ATR

Automatic text recognition is the ability of a machine to accept pixels from images of sources containing text as input and render each of the characters in the digitization as code points[3] readable by computers. The technical progress of recent years in artificial intelligence and neural networks has made it possible to automatically produce textual data from scanned documents, thus considerably reducing the manual transcription work required for any corpus study. Tools as *Transkribus* [Kahle et al., 2017] or *Kraken* [Kiessling, 2019] and its interface *eScriptorium* [Kiessling et al., 2019] offer graphical interfaces to facilitate the use of ATR by non-specialists, whether for applying or training models, but also for creating training data, called ground truth (GT).

Figure 1: Example of segmentation, ms. BNF fr. 412

Behind the generic name of ATR, there are in fact two distinct steps : segmentation and text recognition. The segmentation phase identifies the different zones and lines in images to isolate the written lines as a unit for the next phase. Depending on the technology used, segmentation can be down to line, word or character level. In contrast to projects of the 2000s, it is not necessary to prepare the images upstream for line segmentation before the actual phase of text recognition [Fischer et al., 2009]. Indeed, binerisation or colour processing to reduce noise in the images is no longer required [Kiessling, 2022] if the digitization is of high quality. Thanks to advances in layout analysis over the last ten years, segmentation can be fully automated. However, even if *Transkribus* or *Kraken* provide excellent line detection models, segmentation remains the part that needs to be optimized in order to be able to train fully performing zone detection models on historical materials. Zone naming in this phase[4], as proposed by the *SegmOnto* project [Gabay et al., 2021], is also a major challenge to create an enriched layout analysis (see figure 1) that could be part of a pipeline from digitized images to text pre-editorialization, as in the *Gallic(orpor)a* project [Sagot et al., 2022].

Text recognition implies applying a model that fits a given collection of historical documents in order to produce a transcription. This in turn requires training a model, and to do this GT has to be produced. This is a time-consuming task, but it can be integrated in a virtuous production cycle (figure 2). To produce a specific model, there are two possibilities: to create a new model from scratch or to fine-tune a model from a pre-existing

[3] A code point is a number assigned to represent a character in a system for representing text, such as Unicode.

[4] Using appropriate terms to describe the different components of the layout of a document.

model with new data representative of the corpus to be acquired.

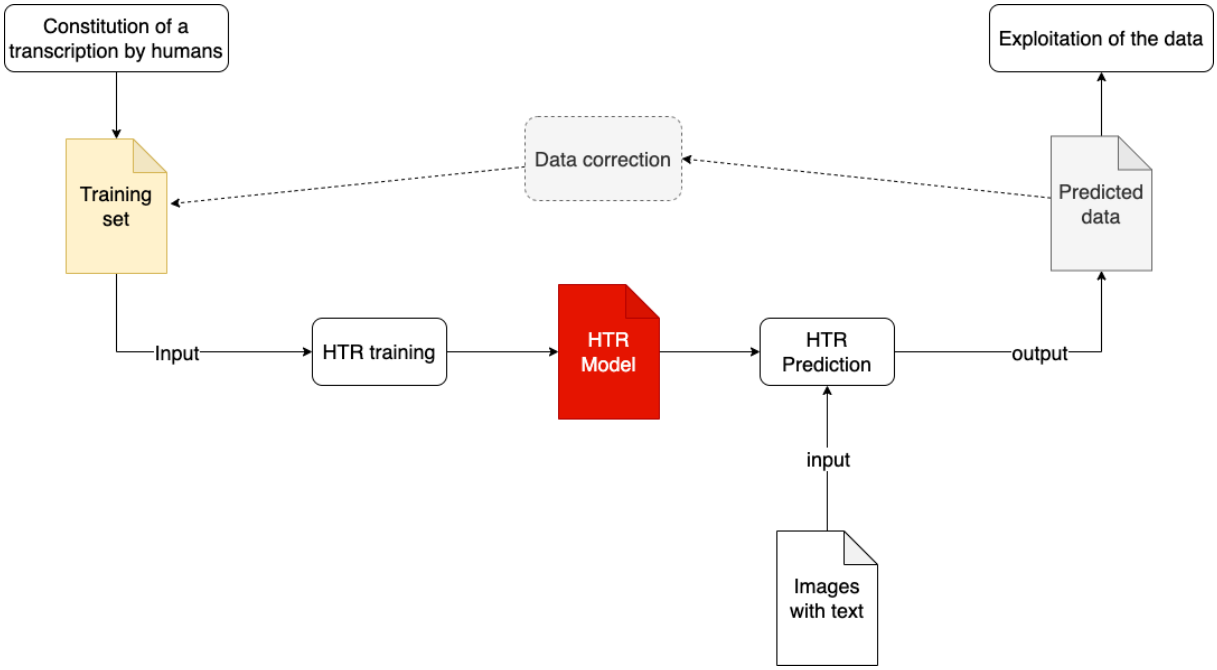

Figure 2: Example of an HTR training process

2.2 HTR and ground truth production

Making GT to share must be a concern of research projects in the coming years if the user community is to grow. There are several ways to produce those data, including (i) aligning pre-existing transcriptions with images and (ii) production of manual transcriptions to build one's own corpus.[5]

(i) Alignment: this method makes it possible to quickly produce enough data to train a model when the availability of ground truth data is still very limited. This is the path chosen by projects such as *Himanis* or *Home* [Stutzmann et al., 2021]. For the *Himanis "chancellery" corpus*, the project used "the monumental text edition provided by Paul Guérin [that] contains a (relatively small) set of transcriptions of more than 1 770 acts from this vast collection." [Bluche et al., 2017]. Then, the lines of text on the images were semi-automatically detected and aligned with the edition. This is also the case of the automatically generated GTs from the Parzival database or the IAM-HistDB dataset [Fischer et al., 2009, 2010, 2011]. However, advanced computer skills are required to align edited text with an image, whether at the level of lines, words or characters. Furthermore, these methods generally provide modernized transcriptions, even if the source corpus is highly abbreviated, thus giving a representation of the text that is already the result of a high-level interpretation. In these cases, modern editorial choices may introduce biases in the perception of the text and can weaken the performance of HTR if the corpora are not large enough for the model to learn conventions that may be project specific and that are not a strict reproduction of what appears in the document.

(ii) Production of manual transcriptions from scratch: this method allows complete control over documents and transcription rules to best suit the objectives of a project. The biggest

[5]From now on, we will only focus on HTR and its specificities in terms of GT production (due to the large variety of variations in handwritten documents) and in terms of model training.

disadvantage of this approach is the significant effort in terms of time, transcription skills and money that has to be made each time a new project is launched. One way to reduce these costs is by sharing more data and models, but also by ensuring that their compilation and sources are well-document in order to help teams find and use the most appropriate ones for their own purposes.

2.3 Ground truth and transcription

This leads us to an important question: what is the best way to transcribe for GT? Should we produce an abbreviated or normalized text? Two recent articles address this issue: "Handling Heavily Abbreviated Manuscripts: HTR Engines vs. Text Normalization Approaches" [Camps et al., 2021b], which deals with medieval Latin manuscripts, and "Modern vs. Diplomatic Transcripts for Historical Handwritten Text Recognition", which deals with the *Carabela collection* of manuscripts related to Spanish naval travel and trade during the 15th-19th centuries [Romero et al., 2019]. In these papers, it seems that global reading with normalized transcripts gives better results based on word-level performance, but the scores are close to those of abbreviated corpora associated with a pipeline to automatically develop abbreviations in order to calculate the word error rate (WER). The scores also depend on the HTR engine, and in particular on the role of the language model within it. Therefore, the choice cannot be made only on the basis of the score, it is also about the use of predictions. For raw text production or keyword spotting (KWS), it simplifies the workflow to have a normalized text for querying the corpus. The text is also easier to read for non-specialists. In the case of a language with no graphical variation, the normalization of the word does not have a huge impact on the perception of the text. But, in the case of a vernacular language such as Old French, with no graphic convention, the development of an abbreviation can be based on external criteria such as the geographical area of production. The abbreviation system also gives a lot of information about the text itself, the hand, the area of production or reception, the status of the text (formal or working text), etc. Tools, like CATTI [Romero et al., 2012], can produce both character-level diplomatic transcripts and the corresponding modernised version, but the prediction can be a bit complicated to handle because of the added information (special characters, tags, etc.).

HTR for historical documents is a valuable resource. But most of the solutions proposed today are individual solutions for a particular corpus over a given period of time, which leads to the production of specific models using data not made for re-use or long-term sustainability. The next step, which is the subject of this paper, is to build more general models to be able to handle not only different hands, but also different scripts from different periods and linguistic areas so that they can be easily shared without multiplying models to be applied even on large collections. "Well-prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently" [Hodel et al., 2021, p. 7]. This is why, we need more reusable data and this is only possible if we put strategies in place to gather and share it.

III DATA PRODUCTION AND GATHERING

3.1 The *Cremma Medieval* dataset

How is one to produce a dataset for a generic HTR model? We will try to answer this question through the experiments we have conducted on manuscripts from the 12th to the

15th century. We must specify that these experiments concern a "low complexity corpus" with relatively homogeneous material, but they are a good starting point to test this approach which can later be applied to more complex and heterogeneous corpora.

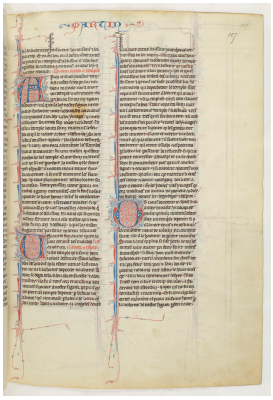

Figure 3: BnF, fr. 412, 13th c.

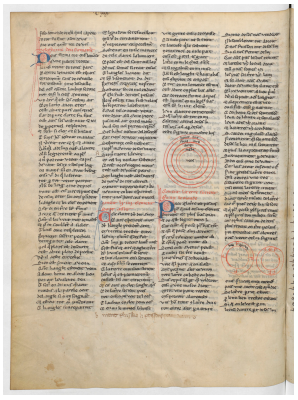

Figure 4: BnF, Arsenal, 3516, 13th c.

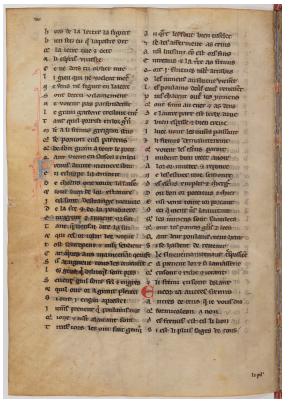

Figure 5: BnF, ms fr. 24428, 13th c.

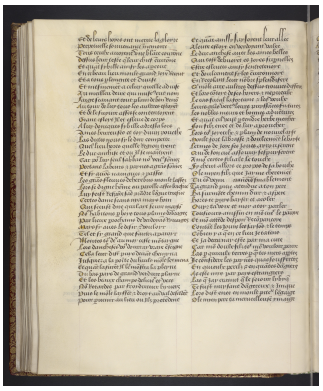

Figure 6: Uni. of Pennsylvania, codex 909, 15th c.

The *Cremma Medieval* dataset (see table 1) was produced between July 2021 and September 2022. It was created with *eScriptorium* and *Kraken*. It consists of fifteen French manuscripts written between the 13th and 15th centuries, mainly digitized in high definition and in colour, with the exception of one manuscript (Vatican) which is a black and white document, and BnF fr. 17229, 13496 and 411 which are from microfilms. The different digitization qualities have introduced some noise to help manage variations in image qualities, as this factor can impact the performance of the model (see subsection 4.3). The initial datasets are mainly made up of pre-existing transcribed texts and the sample sizes can be very different from one source to another. For the data recently added, we try to limit the sample to about ten image files.[6]

Manuscripts	**Date**	**Number of transcribed lines**
BnF, ms fr. 412	13th	**6324**
BnF, Arsenal, 3516	13th	1991
Cologny, Bodmer, 168	13th	1976
BnF, ms fr. 24428	13th	1328
BnF, ms fr. 25516	13th	717
BnF, ms fr. 844	13th	224
BnF, ms fr. 17229	13th	164
BnF, ms fr. 13496	13th	161
BnF, Arsenal 3516	13th	**105**
BnF, ms fr. 22549	14h	2682
Vaticana, Reg. Lat., 1616	14th	1772
University of Pennsylvania, codex 660	14th	368
BnF, ms fr. 411	14th	179
BnF, ms fr. 1728	14th	622
University of Pennsylvania, codex 909	15th	2513
ALL		21126

Table 1: Composition of the *Cremma Medieval* dataset

[6]Ten may seem arbitrary, but, according to my experiments, this is the number of files needed for a two-column manuscript from the 13th-14th c. to finetune a model. It was also easier to give a number of image files than a number of written lines.

As the data come from different projects, transcriptions have been standardized to strengthen the HTR models (see subsection 3.3). We also standardized the layout description using the SegmOnto ontology[7], separating columns, margin notes, numbering, drop capitals, etc. The dataset is shared and made visible through HTR-united thanks to Alix Chagué and Thibault Clérice.[8] Models and data are under a Creative Commons Licence CC-BY 4.0.

3.2 Gathering data and data quality

The first step was to collect data from previous projects. We did not use previous data from *Transkribus*, as we identified a compatibility problem with *Kraken*. Indeed, using the same transcription of the same manuscript extract[9] to build two datasets, one of which was aligned with the image using *Transkribus* and the other made with *eScriptorium* (see figure 7), we saw a difference in the performance of *Kraken* models depending on the data set.

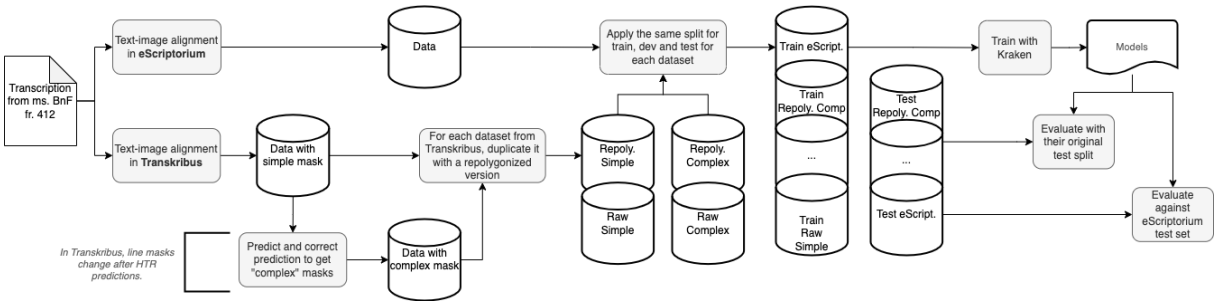

Figure 7: Experiment setup to compare performance between models trained on *Transkribus* and *eScriptorium* data.

Dataset	Mask	Repolygon.	Dev. Avg	Test Avg		Delta	Dev. Med.	Test Med		Delta
				eScript.	Transk.			eScript.	Transk.	
eScriptorium	N/A	N/A	90.2	**89.8**	N/A	N/A	90.2	**89.8**	N/A	N/A
Transkribus	Simple	Yes	85.1	84.3	83.6	-0.7	85.1	84.7	83.7	-1.0
Transkribus	Complex	Yes	88.4	84.4	88.9	-4.5	88.5	84.3	88.9	-4.6
Transkribus	Simple	No	**93.3**	16.6	**92.9**	**-76.3**	**93.4**	16.9	**93.8**	**-76.9**
Transkribus	Complex	No	91.2	79.5	90.7	-11.2	91.2	79.5	**93.8**	-11.3

Table 2: Results of models trained with *Kraken* on test sets made with *eScriptorium* and *Transkribus*, **Scores are average and median accuracies calculated on the top 10 Kraken models, dev scores are given to show that there is no divergence between dev and test scores on the same dataset, and no warning of compatibility discrepancy between the eScriptorium and Transkribus datasets depending on the parameters.*

When the *Transkribus* train data had a simple mask[10] and the repolygonisation option[11] was turned on, the *Kraken* model scored almost the same on both the *eScriptorium* and *Transkribus* test set, with the top 10 models averaging about 0.69% better accuracy on the *eScriptorium* test set. But, using as training set the *Trankribus* data with an automatically generated segmentation and complex masks, the performance of the model is about 4.5% lower accuracy with repolygonisation and 11.2% without repolygonisation on the *eScriptorium* dataset (see table 2 and figure 8).

[7]The complete controlled vocabulary is available at https://segmonto.github.io.

[8]HTR-United "aims at gathering HTR/OCR models for and transcriptions of all periods and style of writing, mostly but not exclusively in French".

[9]Manuscript fr. 412, fol. 103r to 127r.

[10]The mask is the area around the written lines.

[11]With the repolygonisation option, Kraken recalculates masks around letters using the baseline.

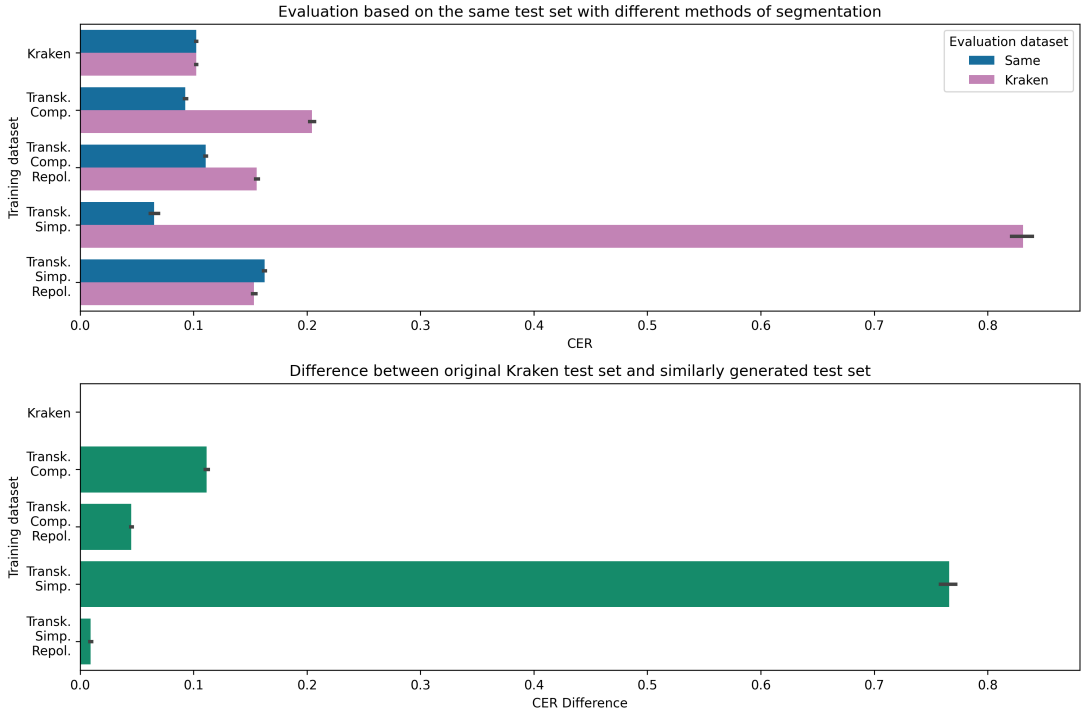

Figure 8: Comparison of the results of models trained with *Kraken* on test sets made with *eScriptorium/Kraken* and *Transkribus*. **In the first diagram, "same" means a test set made with the same interface and parameters as the training set. In both diagrams, the black lines represent the dispersion of the score between the last ten training models, the shorter they are, the less dispersion there is.*

These results do not mean that *Transkribus* data cannot be used when working with *Kraken*, but that, at the time of writing, they need to be adjusted to be fully compatible with *kraken/eScriptorium* (baseline, bounding box adjustments, etc.). This example also shows that before collecting and integrating data into a corpus, it is necessary to test its compatibility and check the constitution of the data: format, number of transcribed lines, segmentation engine, language, date, document type, transcription guidelines.

We did not use automatic text alignment, because we had no expertise in the field and because we wanted to have full control over the data and a diplomatic transcription (see sections 2.3 and 3.3). Therefore, the dataset has been built from scratch, because we felt that a manual alignment from previous transcriptions into the *eScritptorium* interface would be faster, in consideration of the needed adjustments to our own transcription standards. We started with the *Vie de Saint Martin* de Wauchier de Denain from the manuscript BnF fr. 412 for which we had a diplomatic transcription [Pinche, 2021]. We, then, aligned transcriptions from "transcribathons" organized by Laura Morreale (Stanford Libraries projects) or from projects hosted by the Ecole nationale des chartes[12], adapting if necessary the pre-existing transcription. After this first step, the trained HTR model was already relatively accurate (95.49% on the test set [Pinche and Clérice, 2021]). It was then used to help external projects to automatically acquire the text from their handwritten sources. In exchange for our help, they returned ten ground truth pages[13], which had otherwise been used to fine-tune a model on their source. The addition of more

[12]Thanks to the work of Jean-Baptiste Camps (Otinel Edition), Viola Mariotti (Maritem project), and all the transcribers from the Stanford projects.

[13]Many thanks to our transcribers : C. Carnaille, P. Deleville, L. Dugaz, S. Lecomte, A. Meylan,

and more diverse data then produced a model that was increasingly resistant to changes of hands, scripts and documents. The process worked well and in one year we collected 21,126 transcribed lines, all following the same transcription rules.

In order to share our data and ensure its reuse, we need to guarantee its quality. This is why a data control pipeline has been set up. Thanks to T. Clérice, continuous integration tools ensure the homogeneity of XML data.

HTRVX [Clérice and Pinche, 2021b] is a tool based on an XML schema for checking ALTO files (imbrication, empty lines, etc.). In our case, it also allows us to check compliance with the segmOnto ontology for the naming of zones and lines during segmentation. This first step guarantees the uniformity of the XML files of the dataset.

The second step is the verification of the characters used in the transcription. The transcriptions are standardised with the *Choco-Mufin* tool [Clérice and Pinche, 2021a] that uses a table to map the characters of transcriptions, check whether they match those allowed in the table and, if not, replace the non-conforming characters with a corresponding equivalent declared in the table. This practice avoids the use in the dataset of characters that are identical for the (human) transcriber but very different for the machine such as "ꝑ" (Armenian lower-case letter Ké, U+0584) instead of p with stroke (ꝑ, U+A751). The Cremma-Medieval table is based on a restricted selection of MUFI[14] characters.

3.3 Transcription guidelines

To further unify the dataset, transcription guidelines have been written to harmonize the production of new data [Pinche, 2022b].[15] Our goal was to find a solution to translate the way the text is delivered in its original medium into a system that can be interpreted by a machine and that supports its learning. The proposed solutions are necessarily reductive and interpretative, since it is impossible to render the full variety of handwriting by means of a computer with a limited number of characters.[16]

In order to propose an accessible transcription system, the idea of producing allographic transcriptions[17] has been discarded. It seemed impossible to make general recommendations for all medieval documents from the twelfth to the fifteenth century, taking into account each character variation. To push the imitation too far would risk making the transcription impossible to complete and unusable. It would have been too time-consuming, but it would also have generated too many conflicts in transcriptions[Robinson and Solopova, 1993]. Producing normalized transcriptions did not seem to be appropriate either, because of the loss of information with regard to the source, and because the resolution of abbreviations is an interpretative act linked to each specific documents. Finally, since our aim was to produce generic models, the resolution of abbreviations could prevent

[14]MUFI: The Medieval Unicode Font Initiative, <https://mufi.info/m.php?p=mufi>.

[15]The transcription guidelines are the results of reflections lead during the seminar (2021-2022) : "Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le X^{e} et le XIV^{e} siècle". Many thanks to my colleagues J.B. Camps and F. Duval and to all the participants without whom those guidelines couldn't have been made. All the compte-rendus are available here : <https://cremmalab.hypotheses.org/seminaire-creation-de-modeles-htr>.

[16]"Transcription for the computer is a fundamentally interpretative activity, composed of a series of acts of translation from one system of signs (that of the manuscript) to another (that of the computer)" [Robinson and Solopova, 1993].

[17]Transcription that aims to give access to several different forms of each letter or sign.

the model from being applied to a wide variety of documents that would require different resolutions.

We have therefore chosen diplomatic transcriptions. Each letter is reduced to a standard representation. To avoid ambiguity in the representation of the medieval punctuation system and to ensure consistency in transcription, its complexity has been synthesized into three signs:

- Single full stops are transcribed as ".";
- Double signs are transcribed by ";";
- Commas are rendered by ",".

All punctuation marks are transcribed directly after the preceding sign without spaces.

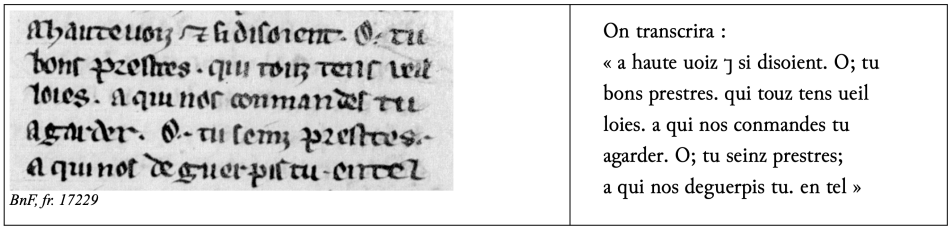

Figure 9: Example of transcription, extract from the transcription guidelines

For instance, the spelling of the text and abbreviations are also preserved : no distinction of "u" and "v", or "i" and "j", and no normalisation of capital letters.

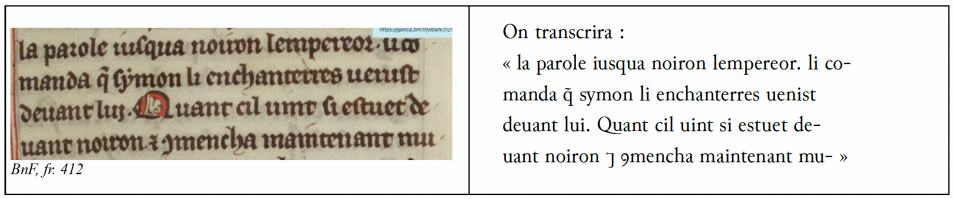

Figure 10: Example of transcription, extract from the transcription guidelines

To ensure uniformity of transcriptions, a set of recommended characters has been designed for special characters such as "tironian et".

LATIN SMALL LETTER P WITH STROKE (*ꝑ barré droit*)		ꝑ	U+A751
LATIN SMALL LETTER P WITH FLOURISH (*ꝓ barré courbe*)		ꝓ	U+A753
LATIN SMALL LETTER Q WITH DIAGONAL STROKE (*ꝗ barré*)		ꝗ	U+A759
TIRONIAN SIGN ET (*abréviation tironienne de « et »*)		⁊	U+204A
DIVISION SIGN (*abréviation de « est »*)		÷	U+00F7

Figure 11: Extract of the table with recommended characters set

All these protocols and recommendations have been made with the idea that the more consistent the data, the less is needed and the better the results. Having principles in the

constitution of the corpus also allows others to better understand our HTR predictions. The full transcription guidelines have been released, as mentioned above, and are available online.[18]

IV EXPERIMENTAL SET UP

4.1 *Bicerin* trainings with the *Cremma Medieval* dataset

Using *Cremma Medieval* dataset, we trained a model, called *Bicerin*. We worked with *Kraken* (version 4.2.0) and tried different configurations. The neural network was first trained with the default *Kraken* learning rate (0.001) and then with a lower learning rate of 0.0001.[19] Both configurations have been set up with parameters : `--lag 20` and `--augment`[20]. For each configuration, we lunch an experiment with the "raw" corpus and another with a harmonized corpus using *Choco-Mufin*.

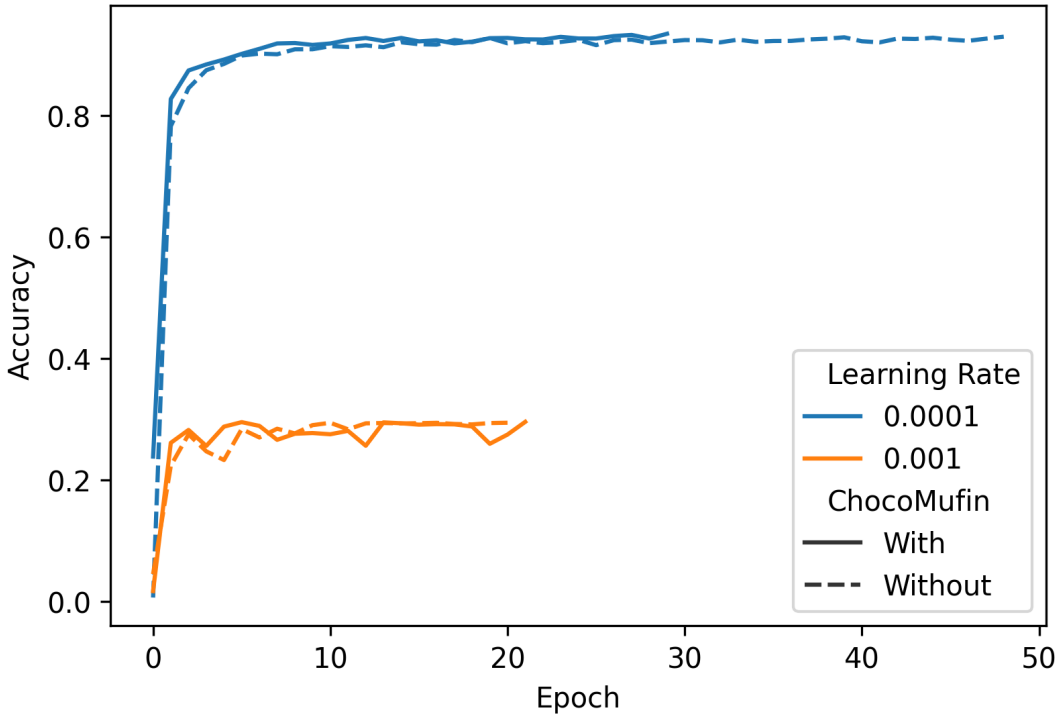

Figure 12: Comparison of scores between models with harmonized and unharmonised data using default or lower learning rate during training process

All the scores in table 3 are calculated with the *Kraken* best model on the same *Cremma Medieval* test set with harmonized transcription.

[18] Ariane Pinche. *Guide de transcription pour les manuscrits du X^e au XV^e siècle*. 2022. hal-03697382 .

[19] Recommended learning rate for manuscripts in Kraken documentation, see https://kraken.re/master/ketos.html.

[20] The parameter *lag* provides the number of epochs without score improvement required before ending the learning cycle. The parameter *augment* activates the option *(automatic) data augmentation* to improve the training process.

Best Models	Test score
Bicerin low LR with CM	94.41 %
Bicerin low LR without CM	93.36 %
Bicerin default LR with CM	28.83 %
Bicerin default LR without CM	28.64 %

Table 3: Scores of the best *Bicerin* models according to *Kraken* configuration, **CM = Choco-Mufin, LR = Learning rate.*

4.2 Cortado : training a mixed model

Next, we diversified our dataset to improve the efficiency of the model. To do so, we trained a mixed model by augmenting the *Cremma Medieval* dataset with 15th-century manuscripts [Pinche et al., 2022b][21] and incunabula [Pinche et al., 2022a][22] from the *Gallic(orpor)a* project (see table 4). All documents are identified by their BnF ark identifier and all *Gallic(orpor)a* data were unified with *Choco–Mufin* before being uploaded to the repository. In the manuscript dataset, we have different types of writing, mainly *textualis* and *hybrida*.[23]

Documents	Type	Nb of lines
btv1b90076543	manuscript	878
btv1b10549431k	manuscript	719
btv1b100737746	manuscript	280
btv1b90069505	manuscript	638
btv1b55008562q	manuscript	576
bpt6k15223596	incunabula	1292
bpt6k15260973	incunabula	424
btv1b8600143n	incunabula	374
btv1b8600164t	incunabula	1383
btv1b8626779r	incunabula	324

Table 4: List of the Documents From *Gallic(orpor)a* Project

The Cortado model has been trained with *Kraken* (version 4.2.0), a learning rate of 0.0001 and parameters : `--lag 20` and `-- augment`. All the scores in the table 5 are calculated with the *Kraken* best model on the *Cremma Medieval* test set with harmonized transcription and on the *Cortado* test set (test set from the documents in the table 4 has been added to the previous test set *Cremma Medieval*).

Test set	Bicerin	Cortado	Improvement
Cremma Medieval test set	94,41%	93.46%	-0.95
Cortado test set	90,88%	94.17%	+3.29

Table 5: Comparison between *Bicerin* and *Cortado* scores, **Bicerin model with low learning rate and harmonized train data.*

[21] https://github.com/Gallicorpora/HTR-MSS-15e-Siecle

[22] https://github.com/Gallicorpora/HTR-incunable-15e-siecle

[23] The main distinctions between those writing styles is respectively the form of the “a” with two or one compartments and the “s” standing on the line or with descender, see [Derolez, 2003].

4.3 Comparison of *Bicerin* and *Cortado* models on out-of-domain documents.

To test the ability of our models to work on a wide range of documents, we apply it on four out-of-domain manuscripts.[24]

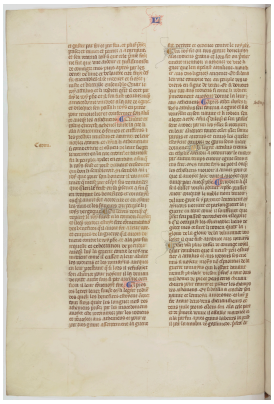

Figure 13: BnF, NAF 27401, 14th c.

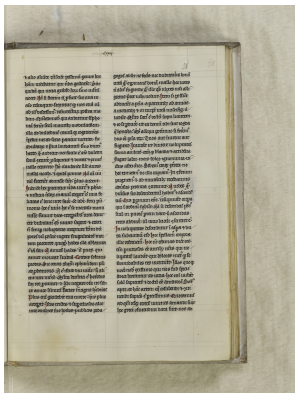

Figure 14: Arras, BM, 861, 14th c.

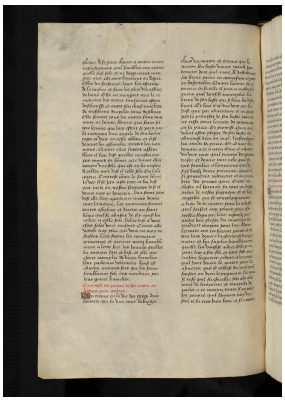

Figure 15: Bruxelles, KBR, 9232, 15th c.

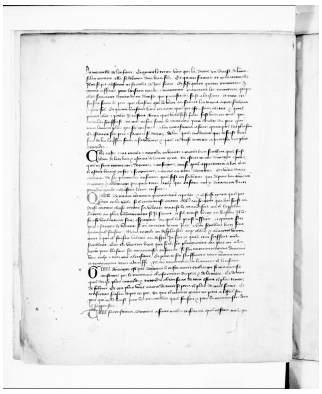

Figure 16: BnF, fr. 777, 15th c.

Each of the selected documents has its own properties. Document n°1 (see figure 13) is a French manuscript similar to the documents in the *Cremma Medieval* dataset. Document n°2 (figure 14) is a document from the same period as document n°1 with a similar script, but in Latin, which induces a greater number of abbreviations and diacritical signs. Document n°3 (figure 15) is a 15th-century manuscript with a hybrid script that is close to that of the codex 909 of the University of Pennsylvania from the *Cremma Medieval* dataset. Last, document n°4 (figure 16) is a manuscript with a completely different script and a black and white digitalisation. This document serves to assess the breaking point of the *Bicerin* model and to see if more diversity in the training corpus can help to overcome these difficulties. All the scores in table 6 have been calculated on one XML file for each document.

N°	Manuscripts	Date	Script	Lang.	Bicerin acc.	Cortado acc.	Improvement
1	BnF, NAF 27401	14th	textualis	Old Fr.	91.25%	91.40%	**+0.15**
2	Arras, Bibliothèque municipale, ms. 861	14th	textualis	Latin	82.99%	83.95%	+0.96
3	Bruxelles, Bibliothèque royale, ms. 9232	15th	hybrid	Old Fr.	91.34%	95.93%	+4.59
4	BnF, fr. 777	15th	cursiva	Old Fr.	63.96%	82.80%	**+18.84**

Table 6: *Bicerin* and *Cortado* accuracy scores on out-of-domain documents

4.4 Generic model and fine-tuning

The purpose of a generic model can be multiple, including to quickly produce data that does not need to be perfect for distant reading on a large-scale dataset or to train with a small dataset an accurate model perfectly adapted to a particular corpus, for example to start an edition. In this second case, it will be necessary to train a fine-tuned model from the generic model with a sample of 5 to 10 pages depending on the complexity of the document, the accuracy required or the degree of difference between the new dataset and the generic dataset. In this last experiment, we tested the effectiveness of fine-tuning the *Bicerin* and *Cortado* models. To do so, we used 4 pages of each document out-of-domain to train the new fine-tuned model and one to test it. To evaluate the performance of the

[24]Both models have been trained with low learning rate and harmonized data.

model, we used the same test set as in the previous experiment in order to facilitate the comparison between the results.[25]

N°	Manuscripts	date	script	Lang.	Bicerin FT acc.	Cortado FT acc.
1	BnF, NAF 27401	14th	textualis	Old Fr.	98.83% (+7.58)	98.08% (+6.68)
2	Arras, Bibliothèque municipale, ms. 861	14th	textualis	Latin	92.16% (+9.17)	92.81% (+8.86)
3	Bruxelles, Bibliothèque royale, ms. 9232	15th	hybrid	Old Fr.	98.70% (+7.36)	99.04% (+3.11)
4	BnF, fr. 777	15th	cursiva	Old Fr.	98.73% (**+34.77**)	98.88 (**+16.08**)

Table 7: *Bicerin* and *Cortado* fine-tuned scores, **numbers in parentheses represent the improvement between the score of the "default" model and the fine-tuned one.*

V RESULTS ANALYSIS

5.1 Benefits of parameters in training

Parameters used to train a model can lead to considerable changes in HTR results. In our case, the default *Kraken* learning rate provides models that are not usable (see table 3). However, it is quite possible that the huge difference between the two training is related to the composition or the size of our corpus and that it is not a systematic phenomenon.

With the default configuration, the training fails to provide a model above 30%, and can even go completely down and reach 0% accuracy. In view of the erratic learning progression (see figure 12), the best hypothesis is that the default learning rate of 0.001 is too high for our corpus. This is not surprising considering that the diversity of characters (diacritics, abbreviations) and spelling (lack of spelling rules) is higher in manuscripts than in printed documents, so reducing it to 0.0001 allow reaching better results. The number of epochs is also very important, because of the high number of classes in the handwritten corpora: overwritten letters, abbreviations, and so on. We recommend a minimum number of epochs of about 20. Other parameters could help optimise recognition, such as customising the height of the bounding box to create masks that better encompass letters in order to increase the results.

5.2 Benefits of normalization

With the low learning rate, the harmonization of the transcription allows us to gain in accuracy: 94.41% against 93.36% (see table 3). Our model also converges faster, in 50 epochs compared to 69 epochs for the version without *Choco-Mufin* (CM, as shown in figure 12). During training, the number of parameters decreased (4.1 M versus 4.0 M). Finally, using CM also limits the number of characters only present in the training set (39 without CM and 17 with CM). However, the reduced difference between the two trainings may be biased by the fact that most of the *Cremma Medieval* dataset was aligned by the same transcriber and that all recent additions follow the transcription guidelines. We can assume that without these two factors, the impact of the CM could have been higher.

Among the models trained with and without CM, different types of errors appear (see table 8). With the CM model, the most frequent type of error (about 15%) is misplaced spaces, which is quite a normal error because even for a human the perception can vary and the notion of space is really only meaningful for printed documents [Saenger, 1997]. The error may therefore be due to the difficulty of identifying whether or not there is a space, or to

[25]Tables with the most common errors with *Cortado* and *Cortado FT* for each test set are available in the annexes.

heterogeneous transcriptions in the corpus. However, we try to minimize the variations by advising transcribers in case of doubt, to follow a semantic segmentation of the text. The other most important errors are related to traditional palaeographic difficulties such as counting minims (confusing "n" and "m", for instance) or the distinction between "u" and "n" ([Rochebouet, 2009]). Without CM, other types of errors occur due to the lack of normalization, such as variation of signs between "et" tironian with the stroke or not, or variations between tildes and macrons.

Thus, harmonization of the transcription reduces the number of character classes, and this in turn optimises HTR results and reduces training time. It also increases the consistency of the prediction, as a sign on the manuscript is always represented by the same character, which will ultimately improve the digital reuse of the data.

Bicerin Complex Model with CM (74453 characters – 4165 errors)			**Bicerin Complex Model without CM (74453 characters – 4946 errors**		
Nb	*Correct*	*Generated*	*Nb*	*Correct*	*Generated*
656	{ SPACE }	{ }	588	{ SPACE }	{ }
414	{ }	{ SPACE }	438	{ }	{ SPACE }
68	{ . }	{ }	318	{ }	{ 0xf158 }
63	{ i }	{ }	138	{ }	{COMBINING ACUTE ACCENT }
62	{ }	{ i }	92	{ COMBINING TILDE }	{ COMBINING MACRON }
57	{ n }	{ }	83	{ i }	{ }
55	{ e }	{ }	82	{ . }	{ · }
53	{ }	{ e }	82	{ COMBINING TILDE }	{ }
50	{ t }	{ }	64	{ . }	{ }
48	{ u }	{ n }	62	{ n }	{ }
46	{ n }	{ u }	60	{ }	{ i }

Table 8: Selection of the most common errors

5.3 Benefits of the variety of the dataset

To estimate the gain of a training corpus including a certain documentary diversity, the results of the *Bicerin* and *Cortado* models on documents out-of-domain were compared (see table 6). On each document, the *Cortado* model obtains better results, on average +6% accuracy. However the difference varies significantly between manuscripts, with only +0.15% accuracy for document 1 which is very similar to the *Cremma Medieval* training set and up to +18.84% for document 4 which is a break point for the *Bicerin* model (63.96% accuracy). The more different the handwriting of the document is from the Gothic handwriting of the 13th and 14th century manuscripts, the higher is the difference between *Cortado* and *Bicerin*.

In the case of document 2, which is in Latin, it seems that the change of language leads to lower scores, with less than 85% accuracy for both models. The results could have been even lower, but the text comes from a manuscript of the 14th century in medieval Latin (*Sermones super Cantica canticorum* by Bernard de Clairvaux) which is not that far from Old French. There is no significant difference in performance between the two models on this document, probably because the *Cortado* training set is not multilingual, nor does it include specific Latin characters, such as abbreviations or diacritics (see table 9), that are not included in the Bicerin training set.[26]

[26]We already have worked on a model mixing Latin and French form the medieval period (8th-15th c.) with promising results (92% accuracy), see [Clérice et al., 2022].

5.4 Benefits of Generic Models

Depending on the use of HTR predictions, the accuracy of a generic model may have more or less impact. For text mining, accuracy above 85% might be sufficient [Eder, 2013], but to produce an edition and speed up the transcription phase, it is more helpful for the user to reach scores equal to or above 95%.

However, generic models are not always intended to be used directly, but also to accelerate the creation of a suitable model and to avoid training a model from scratch. Fine-tuning can be very effective, as shown in table 7, achieving mostly more than 98% accuracy with only four pages. The models are slightly better when fine-tuned from *Cortado*, on average +0.1% accuracy over the four documents compared to *Bicerin*. Both models saw a large improvement between their results before and after fine-tuning, on average +15% for *Bicerin* and +8% for *Cortado*. The process particularly benefited manuscripts that were not well recognized at the beginning, such as document 4 (+34% for *Bicerin* and +16.08% for *Cortado*, see table 12) and document 2 (+9.17% for *Bicerin* and +8.86% for *Cortado*). With this type of model, research teams can focus exclusively on adapting the generic model to their own material, whether it is adapting it to a corpus in Latin, written in another script, or even to a collection composed of low quality images.[27]

Thus, fine-tuning can be used to customize a model in a different script or language and doing so from a generic model is an effective way to quickly produce a fitting model on a particular document.

VI CONCLUSION

The aim of this paper was not to propose a computational or mathematical approach to HTR performance, but to offer a data-driven exploration of the results of this technology. We have shown that the quality of the training corpus can improve the results of HTR (e.g. *Choco-Mufin* harmonization). We also highlighted the scientific aspect of data preparation: (i) through the implementation of transcription guidelines adapted to the research objectives and (ii) through the establishment of a long-term sustainability process, notably by providing accurate documentation, which is a prerequisite for data sharing and, consequently, for the acceleration of textual acquisition using HTR for the whole scientific community. The objective was also to prove that the production of generic models and datasets is now crucial to allow research teams in Humanities to use them and to easily and quickly create or fine-tuned new models for their own corpora. The results we can achieve today with generic or fine-tuned models open up new research approaches. The use of generic HTR models would make it possible, for example, to carry out stylometric prospections on a collection composed of several manuscripts from different hands and periods in order to search for coherent blocks not only in a single source but in a set of witnesses to reinforce hypotheses, cross-check results and reduce variations in results due to differences in scribal hands[28].

[27] In our experience, most of the time, working with an adapted generic model and on a one-handed document, 10 pages of a two-column and 42-line manuscript are enough to achieve at least 95% accuracy. After that, we reach a plateau, and the last few percent requires a number of GTs that makes the time investment unprofitable in terms of the little improvement in the model.

[28] Our objective in the next few years is to study a collection of about thirty French hagiographic manuscripts using generic HTR models in order to analyse the composition of successive compilations, and also perhaps to attribute some anonymous Lives of saints to a known author, in the continuity of the work conducted with my colleagues J.B. Camps and T. Clérice on hagiographic collections in BnF

Our project naturally has limitation, including that the dataset has a relatively limited variety of scripts and so it does not yet provide a truly generic model for all medieval documents. In further research, we will open the dataset to Latin manuscripts to build a less language-dependent model or to documents such as charters, registers, account books with more complex scripts and layout. To improve the use of HTR in medieval manuscripts, we should also enhance our results for layout analysis. Future development should go towards more efficient models for the recognition of zones and lines on documents, as default models are today not completely effective for complex medieval materials.[29]

VII ACKNOWLEDGEMENTS

Thanks to our funders: DIM MAP through the CREMMALab project, and to the CREMMA and INRIA infrastructure who gave us access to servers for trainings and to an *eScriptorium* interface for the preparation of the GTs. Thanks to all the transcribers who participated in the constitution of the Cremma Medival database. Our thanks also go to our colleagues who have accompanied this year of reflection and work: J. B. Camps, Th. Clérice, F. Duval and L. Romary.

Datasets and Models

Pinche, A. (2022). Cremma Medieval (Version Cortado 2.0.0) [Data set]. https://github.com/HTR-United/cremma-medieval

Pinche, Ariane. (2022). HTR model Cremma Medieval (Bicerin 1.1.0). Zenodo. https://doi.org/10.5281/zenodo.6669508

Pinche, A., Gabay, S., Leroy, N., & Christensen, K. Données HTR incunables du 15e siècle [Data set]. https://github.com/Gallicorpora/HTR-incunable-15e-siecle

Pinche, A., Gabay, S., Leroy, N., & Christensen, K. Données HTR manuscrits du 15e siècle [dataset]. https://github.com/Gallicorpora/HTR-MSS-15e-Siecle

References

Théodore Bluche, Sebastien Hamel, Christopher Kermorvant, Joan Puigcerver, Dominique Stutzmann, Alejandro H. Toselli, and Enrique Vidal. Preparatory KWS Experiments for large-scale Indexing of a vast medieval Manuscript Collection in the HIMANIS Project. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 311–316, November 2017.

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. 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, October 2021a. URL https://doi.org/10.1093/llc/fqab033.

Jean-Baptiste Camps, Chahan Vidal-Gorène, and Marguerite Vernet. Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches. In Elisa H. Barney Smith and Umapada Pal, editors, *Document Analysis and Recognition – ICDAR 2021 Workshops*, pages 306–316. Springer International Publishing, 2021b.

Alix Chagué, Thibault Clérice, and Laurent Romary. HTR-United : Mutualisons la vérité de terrain ! October 2021. URL https://hal.archives-ouvertes.fr/hal-03398740.

Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. July 2022. URL https://hal-enc.archives-ouvertes.fr/hal-03723208.

Thibault Clérice and Ariane Pinche. Choco-Mufin, a tool for controlling characters used in OCR and HTR projects, September 2021a. URL https://github.com/PonteIneptique/choco-mufin.

manuscript fr. 412: see [Camps et al., 2021a].

[29]Further research is already being done to provide a better model using object detection, see Clérice [2022]

Thibault Clérice and Ariane Pinche. HTRVX, HTR Validation with XSD, September 2021b. URL https://github.com/HTR-United/HTRVX.

Thibault Clérice, Ariane Pinche, and Malamatenia Vlachou-Efstathiou. Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th century, October 2022. URL https://doi.org/10.5281/zenodo.7234166.

DH2022 Local Organizing Committee, editor. *Digital Humanities 2022, Conference Abstracts*. Tokyo, Japan, 2022. URL https://dh2022.dhii.asia/dh2022bookofabsts.pdf.

James Cummings. *TEI2022 Conference Book*. Zenodo, Newcastle, UK, September 2022. URL https://zenodo.org/record/7071026.

Albert Derolez. *The palaeography of Gothic manuscript books: from the twelfth to the early sixteenth century*. Royaume-Uni, 2003. ISBN 978-0-521-80315-1. ISSN: 1746-2282.

Maciej Eder. Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4):603–614, December 2013. URL https://doi.org/10.1093/llc/fqt039.

Andreas Fischer, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. Automatic Transcription of Handwritten Medieval Documents. In *2009 15th International Conference on Virtual Systems and Multimedia*, pages 137–142, September 2009. doi: 10.1109/VSMM.2009.26.

Andreas Fischer, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. Ground truth creation for handwriting recognition in historical documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 3–10, New York, USA, June 2010. Association for Computing Machinery. URL https://doi.org/10.1145/1815330.1815331.

Andreas Fischer, Emanuel Indermuhle, Volkmar Frinken, and Horst Bunke. HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents. In *International Conference on Document Analysis and Recognition*, pages 53–57, September 2011.

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, and Claire Jahan. SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more). In *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland, September 2021. URL https://hal.archives-ouvertes.fr/hal-03336528.

Alex Graves and Jürgen Schmidhuber. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://papers.nips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html.

Tobias Hodel, David Schoch, Christa Schneider, and Jake Purcell. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, 7:13, July 2021. ISSN 2059-481X. URL http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24, November 2017.

B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, September 2019.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. Utrecht, July 2019. CLARIAH. URL https://dev.clariah.nl/files/dh2019/boa/0673.html.

Benjamin Kiessling. The Kraken OCR system, April 2022. URL https://kraken.re.

Ariane Pinche. *Edition nativement numérique du recueil hagiographique "Li Seint Confessor" de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France*. Thèse de doctorat, Université Lyon3, Lyon, France, May 2021. URL http://www.theses.fr/s150996.

Ariane Pinche. Cremma Medieval, June 2022a. URL https://github.com/HTR-United/cremma-medieval.

Ariane Pinche. Guide de transcription pour les manuscrits du Xe au XVe siècle, June 2022b. URL https://hal.archives-ouvertes.fr/hal-03697382.

Ariane Pinche and Thibault Clérice. HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI), August 2021. URL https://zenodo.org/record/5235186.

Ariane Pinche, Simon Gabay, Noé Leroy, and Kelly Christensen. Données HTR incunables du 15e siècle, May 2022a. URL https://github.com/Gallicorpora/HTR-incunable-15e-siecle.

Ariane Pinche, Simon Gabay, Noé Leroy, and Kelly Christensen. Données HTR manuscrits du 15e siècle,

July 2022b. URL https://github.com/Gallicorpora/HTR-MSS-15e-Siecle.
Stephen Rice, Junichi Kanai, and Thomas Nartker. An evaluation of OCR accuracy. Information Science Research Institute. *Information Science Research Institute*, 9:20, 1993.
Peter Robinson and Elizabeth Solopova. Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. July 1993. doi: 10.5281/zenodo.4050360. URL https://zenodo.org/record/4050360.
Anne Rochebouet. Une « confusion » graphique fonctionnelle? sur la transcription du u et du n dans les textes en ancien et moyen français. *Scriptorium*, 63(2):206–219, 2009. URL https://www.persee.fr/doc/scrip_0036-9772_2009_num_63_2_4057.
Verónica Romero, Alejandro Héctor Toselli, and Enrique Vidal. *Multimodal Interactive Handwritten Text Transcription*. World Scientific, 2012.
Verónica Romero, Alejandro H. Toselli, Enrique Vidal, Joan Andreu Sánchez, Carlos Alonso, and Lourdes Marqués. Modern vs Diplomatic Transcripts for Historical Handwritten Text Recognition. In Marco Cristani, Andrea Prati, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *New Trends in Image Analysis and Processing – ICIAP 2019*, pages 103–114. Springer International Publishing, 2019.
Paul Saenger. *Space Between Words: The Origins of Silent Reading*. Stanford University Press, 1997.
Benoît Sagot, Laurent Romary, Rachel Badwen, Pedro Ortiz Suárez, Jean-Baptiste Camps, Simon Gabay, and Ariane Pinche. Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue, September 2022. URL https://github.com/Gallicorpora/Gallicorpora.github.io.
Dominique Stutzmann, Jean-François Moufflet, and Sébastien Hamel. La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique. *Médiévales*, 73(73):67–96, December 2017. URL https://journals.openedition.org/medievales/8198.
Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. HOME-Alcar: Aligned and Annotated Cartularies, 2021. URL https://hal.archives-ouvertes.fr/hal-03503062.

VIII ANNEXES

8.1 Generic model and fine-tuning

Cortado model tested on BnF, NAF 27401			Cortado fine-tuned model tested on BnF, NAF 27401		
Nb	correct	Generated	Nb	correct	Generated
29	{n}	{u}	11	{ COMBINING TILDE }	{}
27	{}	{ COMBINING DOT ABOVE }	3	{ u }	{ n }
15	{ }	{ i }	2	{ u }	{}
11	{i}	{}	2	{ o }	{}
9	{ SPACE }	{}	2	{ n }	{u}
8	{}	{n}	2	{}	{ i}
8	{}	{ SPACE }	2	{ COMBINING LATIN SMALL LETTER I }	{}
8	{ m }	{i}	2	{ u }	{ r }
7	{m}	{ u }	2	{ s }	{}
6	{u}	{n}	2	{ i }	{ o}

Table 9: Selection of the 10 most common errors

Cortado model tested on Arras, ms. 861			Cortado fine-tuned model tested on Arras, ms. 861		
Nb	correct	Generated	Nb	correct	Generated
44	{ SPACE }	{}	98	{ COMBINING TILDE }	{}
34	{}	{ COMBINING TILDE }	24	{ a }	{ ã }
20	{ ĩ }	{ i }	19	{ o }	{ õ }
18	{ i }	{}	18	{ u }	{ ũ }
16	{ COMBINING VERTICAL TILDE }	{}	17	{ e }	{ ẽ }
15	{ 0xf1ac }	{}	13	{ i }	{ ĩ }
12	{ d }	{ o }	7	{ n }	{ ñ }
12	{ COMBINING TILDE }	{}	2	{}	{ SPACE }
10	{}	{ SPACE }	2	{ d }	{ o }
10	{ . }	{}	1	{ õ }	{ o }

Table 10: Selection of the 10 most common errors

Cortado model tested on KBR, ms. 9232			Cortado fine-tuned model tested on KBR, ms. 9232		
Nb	correct	Generated	Nb	correct	Generated
12	{i}	{}	12	{ COMBINING TILDE }	{}
11	{}	{ SPACE }	1	{}	{ m }
7	{r}	{}	1	{ j }	{ i }
5	{r}	{n}	1	{ D }	{ d }
5	{v}	{u}	1	{ s }	{ l }
4	{i}	{m}	1	{ I }	{ i }
4	{n}	{}	1	{ u }	{ v }
4	{i}	{n}	1	{ U }	{ u }
4	{i}	{ u }	1	{ b }	{ l }
4	{n}	{m}	1	{ s }	{ s}

Table 11: Selection of the 10 most common errors

Cortado model tested on BnF, fr. 777			Cortado fine-tuned model tested on BnF, fr. 777		
Nb	correct	Generated	Nb	correct	Generated
26	{r}	{m}	7	{ o }	{ õ }
23	{u}	{ m }	7	{ COMBINING TILDE }	{}
20	{t}	{}	2	{ , }	{ }
19	{ SPACE }	{}	2	{ o }	{ e }
17	{u}	{n}	1	{ ẽ }	{ l }
13	{i}	{}	1	{ }	{ , }
12	{i}	{n}	1	{ SPACE }	{}
9	{u}	{i}	1	{ r }	{ }
8	{}	{ SPACE }	1	{ }	{ o }
8	{e}	{}	1	{u}	{n}

Table 12: Selection of the 10 most common errors