Generic HTR Models for Medieval Manuscripts The CREMMALab Project

Ariane Pinche

Generic HTR Models for Medieval Manuscripts
The CREMMALab Project

Ariane Pinche[1]

[1]CNRS, CIHAM (UMR 5846)

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

Abstract

In the Humanities, the emergence of digital methods has opened up research questions to quantitative analysis. This is why HTR technology is increasingly involved in humanities 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 handwritten text recognition 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 HTR 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 easily develop an HTR model for medieval manuscripts, create and gather training data by knowing the issues underlying their choices. The objective is also to show the importance of the constitution of consistent data as a prerequisite to allow their gathering and to train efficient HTR 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 (25). 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 (13; 10). At the time, these technologies were mostly experimental and required the support of an expert in image processing and deep learning. Only large-scale projects could afford to use this type of technology, such as *Oriflams* (31). Since the late 2010s and early 2020s, huge improvements have been made with HTR engines, such as *Pylaia*, *HTR+*, *Kraken* and their interfaces: *Transkribus* and *eScriptorium* (15; 18; 16). Nowadays, models can reach a character error rate (CER) between 8% and 2% 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" (14).

These advances made it accessible to any humanities research project and 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 HTR in the DH 2022 (5) and TEI 2022 (7) conferences (7) both in the GLAM sector and in academia (*stricto sensu*).

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 (Consortium for Handwriting Recognition of Ancient Materials) supported by the DIM (research fund of the Île-de-France Region) MAP (Matériaux anciens et patrimoniaux), 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 (20) and listed in the *HTR-united* catalogue (3). 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 have been examined to evaluate the impact of the training corpus on the robustness and genericity of the models.

Regarding the need for HTR data and models in the community, as producing training data is extremely time consuming (10), this paper proposes, through the work and experience of the *CREMMALab* project, to present some recommendations for building consistent data for medieval manuscripts that could be shared with other projects. 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 genericity, scoring and fine-tuning of HTR models, so that they can reuse it 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 they react when fine-tuned. The results will be analysed in Section V. Finally, Section VI summarizes the work presented and draws conclusions.

II BACKGROUND AND RELATIVE WORKS

2.1 General principles of the HTR

Handwritten text recognition is the ability of a machine to accept pixels from images of manuscripts containing text as input and render each of the characters in the digitization as code points 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 (15) or Kraken (17) and its interface eScriptorium (16) offer graphical interfaces to facilitate the use of HTR by non-specialists,

whether for applying or train 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 HTR, there are in fact two distinct steps : segmentation and text recognition. 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 word or character level (10). In contrast to the project of the 2000s, it is not necessary to prepare the images upstream for line segmentation before the actual phase of text recognition (11). Indeed, binerisation or colour processing to reduce noise in the images is no longer required (18) if the digitization is of high quality, as it is the case with *Gallica* or *e-codices* digital portals. Thanks to advances in the 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, as proposed by the *SegmOnto* project (12), 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 (30).

Text recognition implies to train and/or apply a model that fits a given collection of historical documents. To train a model, GT have to be produced. It 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 model with new data representative of the corpus to be acquired.

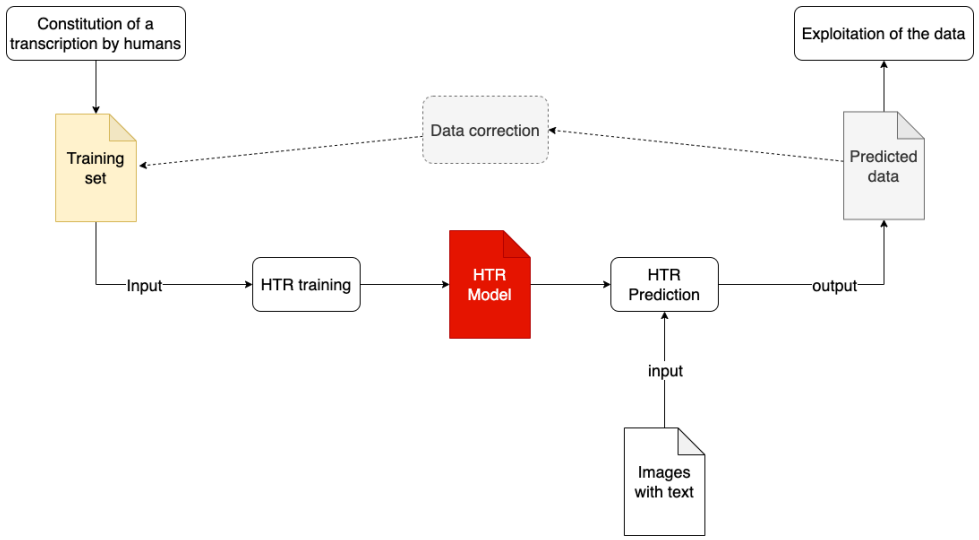


Figure 2: Example of an HTR training process

2.2 HTR and ground truth production

Making GT must be a major focus in the next few years if the user community is to grow. There are several ways to produce those data: (i) aligning pre-existing transcriptions with images and (ii) production of manual transcriptions to build one's own corpus.

(i) Alignment: this method makes it possible to quickly produce enough data to train a performing model, while the availability of ground truth data is still very limited. This is the path chosen by projects such as *Himanis* or *Home* (32). 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. These transcriptions were converted in XML-TEI format" (1). 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 (11; 10; 9). However, advanced computer skills are required to align edited text with an image, whether at lines, words or characters level. 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, information about original spellings are lost in HTR predictions.

(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 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 to share more data and models well documented about their conception to help teams reuse 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" (2), 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 (28). In terms of HTR performance, 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). 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 word key spotting (WKS), 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 scriptural variation and 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 give 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. Some systems allow producing both character-level diplomatic transcripts and the corresponding modernized versions using tools like CATTI (29)

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 and not to multiply the 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" (14, 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 *Cremma Medieval* dataset

How 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 it is 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.

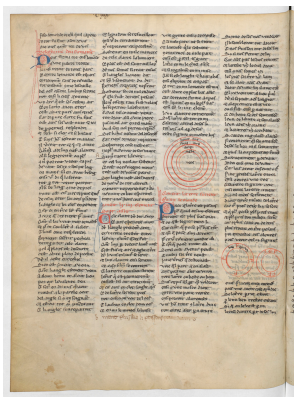


Figure 4: BnF, Arsenal, 3516, 13th c.

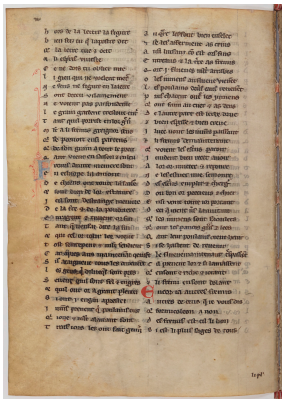


Figure 5: BnF, ms fr. 24428, 13th c.

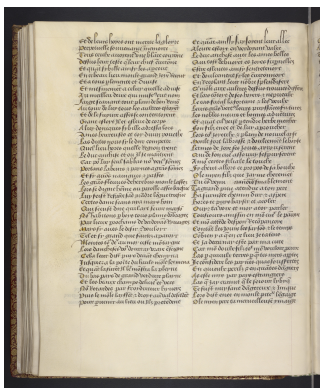


Figure 6: Uni. of Pennsylvania, codex 909, 15th c.

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 on the performance of the model. 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[1].

[1]The number may seem arbitrary, but it is the amount of files needed for a two-column manuscript

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
Vaticane, 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		22278

Table 1: Composition of the *Cremma Medieval* dataset

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[2], 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[3].

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[4] 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.

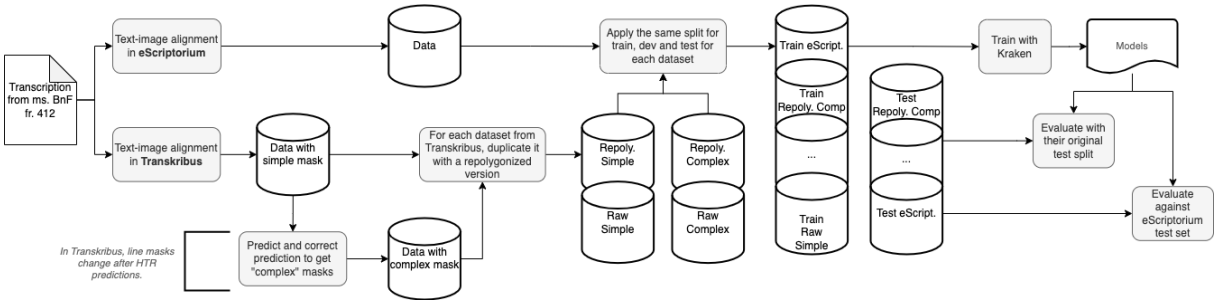


Figure 7: Experiment setup to compare performance between models trained on Transkribus and eScriptorium data.

from the 13th-14th c. to finetuned a model. It was also easier to give a number of image files than a number of written lines.

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

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

[4]Manuscript fr. 412, fol. 103r à 127r.

When the *Transkribus* train data had a simple mask and the repolygonisation option was turned on, the *Kraken* model scored almost the same on both the *Kraken* and *Transkribus* test set , with the top 10 models averaging about 0.69% better accuracy on the *Kraken* 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 4.52% lower accuracy with repolygonisation and 18.38% without repolygonisation on the *Kraken* dataset (see figure 8).

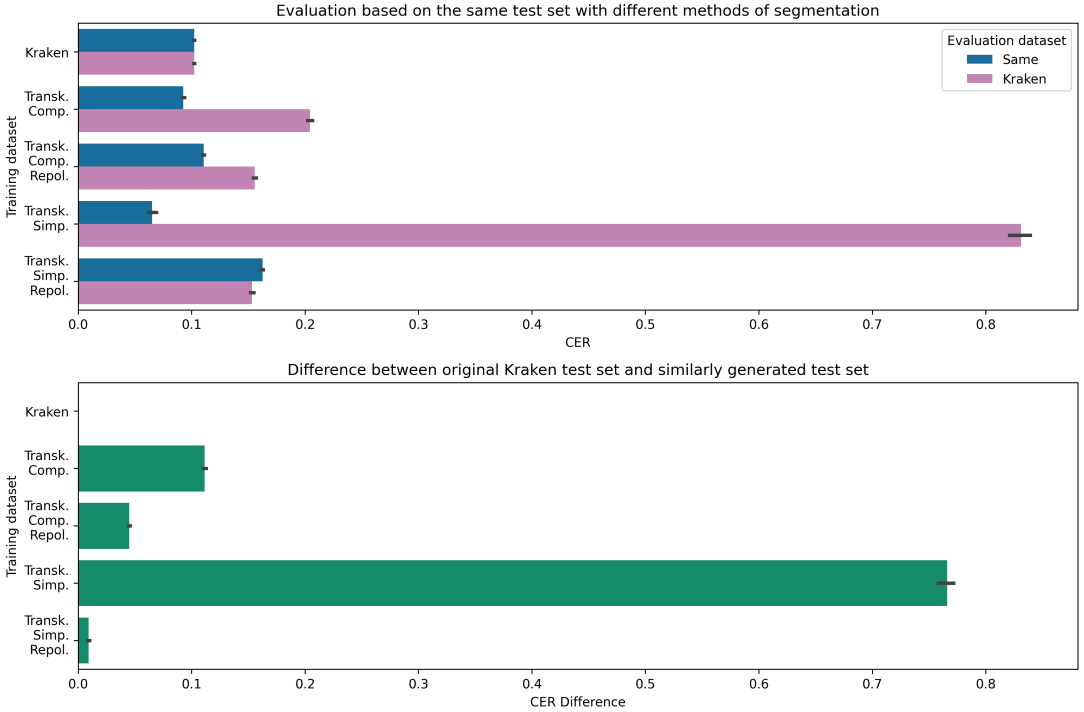


Figure 8: Comparison of the results of models trained with *Transkribus* and *Kraken* on test sets made with *eScriptorium* and *Transkribus*. **In the first diagram, "same" means a test set made with the same interface as the training set. In both diagrams, the black lines represent the dispersion of 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 today it needs to be adjusted to be fully compatible (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 (19). Then, we aligned transcriptions from "transcribathons" organized by Laura Morreale (Stanford Li-

braries projects) or from projects hosted by the Ecole nationale des chartes[5], adapting if necessary the pre-existing transcription. After this first step, the trained HTR model was already relatively efficient (95.49% accuracy on the test set (22)). 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[6], which had otherwise been used to fine-tune a model on their source. The addition of more 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 22,278 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. (i) *HTRUX* (6) 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. (ii) Thanks to the *ChocoMufin* tool (5) associated with a reference characters table, the transcriptions are standardized. This practice avoids the use of alternative characters such as "ꝑ" (Armenian lower-case letter Ké, U+0584) instead of p with stroke (ꝑ, U+A751). The list is based on a restricted selection of MUFI[7] characters.

3.3 Transcription guidelines

To go further in the uniformization of the dataset, transcription guidelines (21) have been written[8] to harmonize the production of new data. Our goal was to find a way of translating 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[9].

In order to propose an accessible transcription system, the idea of producing allographic transcriptions[10] 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 characters variations. 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 (26). Producing normalized transcriptions did not seem to be appropriate either, because of:

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

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

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

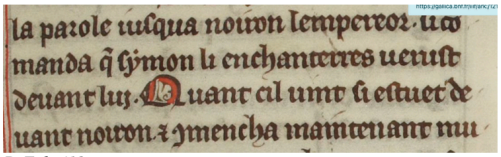
[8]The transcription guidelines are the results of reflections lead during the seminar (2021-2022) : "Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le X^{e} et le XIV^{e} siècle"(). Many thanks to my colleagues J.B. Camps and F. Duval and to all the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

[9]"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)" (26)

[10]Transcription that aims to give access to all forms of each letter or sign.

(i) the loss of information with regard to the source and because (ii) the resolution of abbreviations is an interpretative act linked to the specificity of each of the documents. Finally, since our aim was to produce generic models, the resolution of abbreviations could have been detrimental to its extension.

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, its complexity has been synthesized into two signs: single sign = "."and double sign = ":". The spelling of the text and abbreviations are also preserved : no distinction of "u" and "v", or "i" and "j", no normalization of capital letters.



BnF, fr. 412

On transcrira :
« la parole iusqua noiron lempereor. li co-
manda q̃ symon li enchanterrres uenist
deuant lui. Quant cil uint si estuet de-
uant noiron ⁊ ymencha maintenant mu- »

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

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

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

Figure 10: Extract of the table with recommended characters set

All these protocols and recommendations have been made with the idea that creating consistent data is a way to produce fewer data for better results. Having principles in the constitution of the corpus also allows others to better understand our HTR predictions.

IV EXPERIMENTAL SET UP

4.1 *Bicerin* trainings with *Cremma Medieval* dataset

Using *Cremma Medieval* dataset, we train 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. Both configurations have been set up with parameters : `--lag 20` and `--augment`. For each configuration, we lunch an experiment with the "raw" corpus and another with a harmonized corpus using *ChocoMufin*.

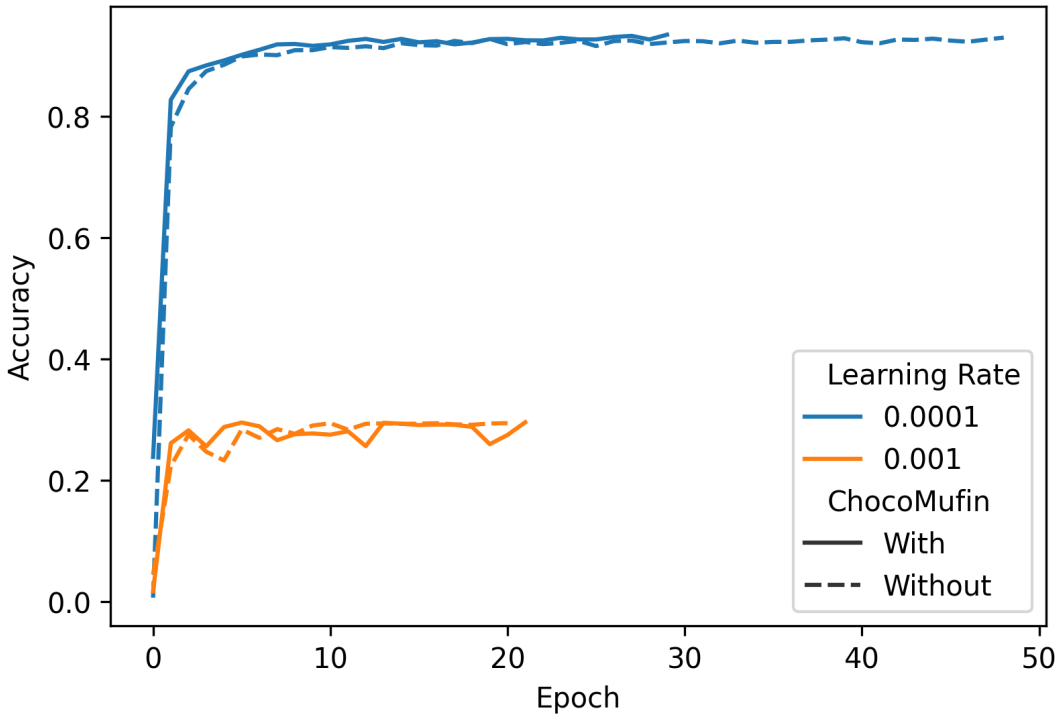


Figure 11: Comparison of scores between models with harmonized and unharmonised data using default or lower learning rate during training process

All the scores in the table 2 are calculated with the *kraken* best model on the same *Cremma Medieval* test set with harmonized transcription.

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 2: Scores of the best *Bicerin* models according to *Kraken* configuration, **CM = ChocoMufin, LR = Learning rate.*

4.2 Cortado : training a mixed model

Next, we diversify our dataset to improve the efficiency of the model. To do so, we trained a mixed model by adding to the *Cremma Medieval* dataset 15th-century manuscripts (24)[11] and incunabula (23)[12] from *Gallic(orpor)a* project (see table 3). All documents are identified by their BnF ark identifier and all Gallic(orpor)a data were unified with *ChocoMufin* before being uploaded to the repository. In the manuscript dataset, we have different types of writing, mainly textualis and hybrida.

[11] https://github.com/Gallicorpora/HTR-MSS-15e-Siecle
[12] https://github.com/Gallicorpora/HTR-incunable-15e-siecle

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 3: List of the Documents From *Gallicorpora* 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 4 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 3 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 4: Comparison between *Bicerin* and *Cortado* scores, **Bicerin model with low learning rate and harmonized train data.*

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

To test the ability of our models[13] to work on a wide range of documents, we apply it on four out-of-domain manuscripts.

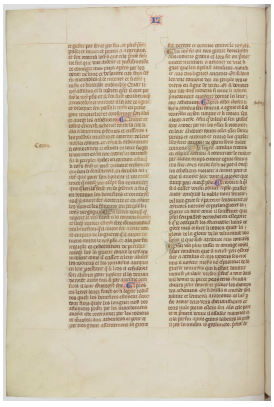


Figure 12: BnF, NAF 27401, 14th c.

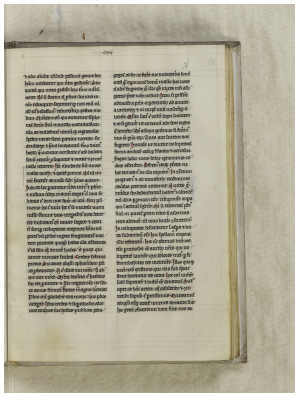


Figure 13: Arras, BM, 861, 14th c.

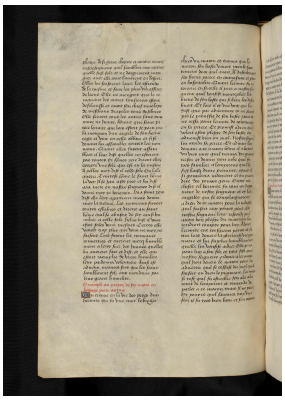


Figure 14: Bruxelles, KBR, 9232, 15th c.

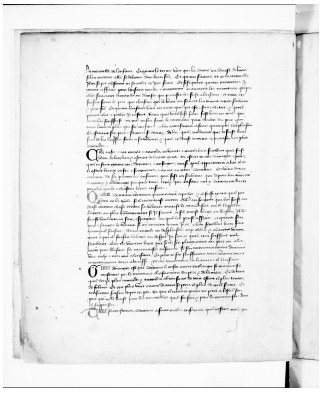


Figure 15: BnF, fr. 777, 15th c.

Each of the selected documents has its own properties. Document n°1 (see figure 12) is a French manuscript similar to the documents in the *Cremma Medieval* dataset. Document

[13]both models have been trained with low learning rate and harmonized data

n°2 (figure 13) 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 14) is a 15th-century manuscript with a hybrid script close to the script of the codex 909 of the University of Pennsylvania from *Cremma Medieval* dataset. Last, document n°4 (figure 15) is a manuscript with a completely different script and a black and white digitalization. This document serves to assess the breaking point of the *Bicerin* model and to see if more diversity of the training corpus can help to overcome these difficulties. All the scores in the table 5 have been calculated on one XML files of 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 5: *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: (i) to quickly produce data that does not need to be perfect for distant reading on a large-scale dataset, (ii) to train with a small dataset an efficient 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 differences between the new dataset and the generic dataset. In this last experiment, we tested the fine-tuning capacity of 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 model, we used the same test set as in the previous experiment in order to facilitate the comparison between the results.

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 6: *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 does not provide good scores (see table 2). 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 progression of learning (11, 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 2). Our model also converges faster, in 50 epochs compared to 69 epochs for the no *chocoMufin* (CM) version (figure 11). During training, the number of parameters decreased (4.1 M versus 4.0M). 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 7). With the CM model (i), 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 has a full meaning only for printed documents. The error may therefore be due to the difficulty of identifying them or to 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 leg counting or the distinction between "u" and "n" ((27)). With the no CM model, other types of errors occur (ii), such as variation of signs between "et" tironian with the stroke or not, or variations between tildes and macrons, due to the lack of normalization.

Thus, harmonization of the transcription, because of its inherent reduction in the number of character classes, optimizes 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 7: 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 5). On each document, the *Cortado* model obtains better results, on average +6% accuracy.

But 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* model.

In the case of document 2, which is in Latin, it seems that the change of language leads to lower scores: less than 85% accuracy for both models. 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.

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 ((8)), but to produce an edition and speed up the transcription phase, it is more comfortable 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 6, 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*) and 2 (+9.17% for *Bicerin* and +8.86% for *Cortado*).

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. *ChocoMufin* 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/or datasets is now crucial to allow research team in Humanities to use them

and to easily and quickly create or fine-tuned new models for their own corpora. The results, based on our experiments, are promising both in terms of general performance and when fine-tuned.

Our project has some restrictions due to a relatively limited variety in our dataset to provide a truly generic model for the whole medieval document corpus. In further research, we will open the dataset to Latin manuscripts to build a less language-dependent model or to documents of the practice with more complex scripts and layout. To improve the use of HTR in medieval manuscripts, we should also completely solve the problems of 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[14].

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 the 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. My thoughts are with my colleagues who have accompanied this year of reflection and work: J. B. Camps, Th. Clérice, F. Duval and L. Romary.

References

[1] Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Dominique Stutzmann, Toselli, A.H., Vidal, E.: 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). vol. 01, pp. 311–316 (Nov 2017)

[2] Camps, J.B., Vidal-Gorène, C., Vernet, M.: Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches. In: Barney Smith, E.H., Pal, U. (eds.) Document Analysis and Recognition – ICDAR 2021 Workshops. pp. 306–316. Springer International Publishing (2021)

[3] Chagué, A., Clérice, T., Romary, L.: HTR-United : Mutualisons la vérité de terrain ! (Oct 2021), https://hal.archives-ouvertes.fr/hal-03398740

[4] Clérice, T.: You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine (Jul 2022), https://hal-enc.archives-ouvertes.fr/hal-03723208

[5] Clérice, T., Pinche, A.: Choco-Mufin, a tool for controlling characters used in OCR and HTR projects (Sep 2021). https://doi.org/10.5281/zenodo.5356154, https://github.com/PonteIneptique/choco-mufin

[6] Clérice, T., Pinche, A.: HTRVX, HTR Validation with XSD (Sep 2021). https://doi.org/10.5281/zenodo.5359963, https://github.com/HTR-United/HTRVX

[7] Cummings, J.: TEI2022 Conference Book. Zenodo, Newcastle, UK (Sep 2022), https://zenodo.org/record/7071026

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

[9] Fischer, A., Indermuhle, E., Frinken, V., Bunke, H.: HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents. In: International Conference on Document Analysis and Recognition. pp. 53–57 (Sep 2011)

[10] Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., Stolz, M.: Ground truth creation for handwriting recognition in historical documents. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. pp. 3–10. Association for Computing Machinery, New York, USA (Jun 2010), https://doi.org/10.1145/1815330.1815331

[14] New research are already made to provide better model using object detection, see (4)

[11] Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M.: Automatic Transcription of Handwritten Medieval Documents. In: 2009 15th International Conference on Virtual Systems and Multimedia. pp. 137–142 (Sep 2009). https://doi.org/10.1109/VSMM.2009.26
[12] Gabay, S., Camps, J.B., Pinche, A., Jahan, C.: 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 (Sep 2021), https://hal.archives-ouvertes.fr/hal-03336528
[13] Graves, A., Schmidhuber, J.: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In: Advances in Neural Information Processing Systems. vol. 21. Curran Associates, Inc. (2008), https://papers.nips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html
[14] Hodel, T., Schoch, D., Schneider, C., Purcell, J.: General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. Journal of Open Humanities Data **7**, 13 (Jul 2021), http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/
[15] Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 04, pp. 19–24 (Nov 2017)
[16] Kiessling, B., Tissot, R., Stokes, P., Ezra, D.S.B.: eScriptorium: An Open Source Platform for Historical Document Analysis. In: International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19 (Sep 2019)
[17] Kiessling, B.: Kraken - an Universal Text Recognizer for the Humanities. CLARIAH, Utrecht (Jul 2019), https://dev.clariah.nl/files/dh2019/boa/0673.html
[18] Kiessling, B.: The Kraken OCR system (Apr 2022), https://kraken.re
[19] Pinche, A.: 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), http://www.theses.fr/s150996
[20] Pinche, A.: Cremma Medieval (Jun 2022), https://github.com/HTR-United/cremma-medieval
[21] Pinche, A.: Guide de transcription pour les manuscrits du Xe au XVe siècle (Jun 2022), https://hal.archives-ouvertes.fr/hal-03697382
[22] Pinche, A., Clérice, T.: HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI) (Aug 2021). https://doi.org/10.5281/zenodo.5235186, https://zenodo.org/record/5235186
[23] Pinche, A., Gabay, S., Leroy, N., Christensen, K.: Données HTR incunables du 15e siècle (May 2022), https://github.com/Gallicorpora/HTR-incunable-15e-siecle
[24] Pinche, A., Gabay, S., Leroy, N., Christensen, K.: Données HTR manuscrits du 15e siècle (Jul 2022), https://github.com/Gallicorpora/HTR-MSS-15e-Siecle
[25] Rice, S., Kanai, J., Nartker, T.: An evaluation of OCR accuracy. Information Science Research Institute. Information Science Research Institute **9**, 20 (1993)
[26] Robinson, P., Solopova, E.: Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue (Jul 1993). https://doi.org/10.5281/zenodo.4050360, https://zenodo.org/record/4050360
[27] Rochebouet, A.: 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), https://www.persee.fr/doc/scrip_0036-9772_2009_num_63_2_4057
[28] Romero, V., Toselli, A.H., Vidal, E., Sánchez, J.A., Alonso, C., Marqués, L.: Modern vs Diplomatic Transcripts for Historical Handwritten Text Recognition. In: Cristani, M., Prati, A., Lanz, O., Messelodi, S., Sebe, N. (eds.) New Trends in Image Analysis and Processing – ICIAP 2019. pp. 103–114. Springer International Publishing (2019)
[29] Romero, V., Toselli, A.H., Vidal, E.: Multimodal Interactive Handwritten Text Transcription. World Scientific (2012)
[30] Sagot, B., Romary, L., Badwen, R., Ortiz Suárez, P., Camps, J.B., Gabay, S., Pinche, A.: Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue (Sep 2022), https://github.com/Gallicorpora/Gallicorpora.github.io
[31] Stutzmann, D., Moufflet, J.F., Hamel, S.: 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 (Dec 2017), https://journals.openedition.org/medievales/8198
[32] Stutzmann, D., Torres Aguilar, S., Chaffenet, P.: HOME-Alcar: Aligned and Annotated Cartularies (2021), https://hal.archives-ouvertes.fr/hal-03503062