HTR Models and genericity for Medieval Manuscripts

Ariane Pinche

HTR Models and genericity for Medieval Manuscripts

Written by Ariane Pinche for Digital humanities conference, Tokyo 2022.
Poster presentation: Ariane Pinche. CREMMALab Project: Handwritten Text Recognition for medieval manuscripts. *Digital Humanities*, Jul 2022, Tokyo, Japan, ⟨hal-03724041⟩,

Within the infrastructure of the CREMMA project (Consortium for Handwriting Recognition of Ancient Materials) supported by the DIM (research funded by the Île-de-France Region) MAP (Ancient and Heritage Materials), the CREMMALab[1] project combines research questions, creation and release of data from medieval French literary manuscripts for HTR. The objective of the CREMMALab project is to propose open training data and HTR models for medieval documents. All data and models produced by the project are already available in the *CREMMA Medieval* repository (Pinche 2022) on HTR-united catalogue (Chagué, Clérice, and Chiffoleau, 2021). In accordance with this objective, the project implements transcription protocols to optimise the training of HTR models and to produce homogeneous and shareable data and models.

I- CREMMA Medieval dataset

The CREMMA Medieval dataset was created and has been enlarged in the years 2021 and 2022 with the eScriptorium interface (Kiessling et al. 2019). Initially, the dataset was focussed on 13th and 14th century manuscripts in Old French and Gothic Textualis, and then extended to the 15th century with the addition of the University of Pennsylvania codex 909, written in Hybrida script ("Burgundian Bastard", hybrid script between the formal Gothic style and a cursive script) (Derolez 2003).

Manuscript	**Date**	**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. 17,229	13th	164
BnF, ms fr. 13,496	13th	161
BnF, Arsenal 3516	13th	105
BnF, ms fr. 22549	14th	2682
Vaticane, Reg. Lat., 1616	14th	1772
University of Pennsylvania, codex 660	14th	368
BnF, ms fr. 411	14th	179
University of Pennsylvania, codex 909	15th	2513
All		21656

Table 1 : CREMMA Medieval dataset (last state)

In our opinion, the way in which corpora are produced is one of the keys to building a robust and consistent model with a reasonable amount of data. Indeed, their quality guarantees the

1 Project presentation : <https://cremmalab.hypotheses.org>

intelligibility and the coherence of the HTR model predictions. This is why we have implemented transcription rules to help transcribers produce transcriptions that are as consistent as possible. We chose a graphematic method, following the definitions of D. Stutzmann (Stutzmann 2011), so that a sign in the image corresponds to a sign in our text. All data produced conform as far as possible to the following rules:

- Each letter form is reduced to a standardized representation.
- The spelling of the text is preserved
- All abbreviations are kept
- u/v or i/j are not distinguished
- No standardization of capital letters is done

These years of reflection on the production of data for HTR have led to the writing of transcription guidelines for medieval manuscripts (Pinche 2022).

For the description of the document's layout, all data follow SegmOnto controlled vocabulary[2] to describe the different zones of a folio, such as main zone, running title zone, margin or numbering zone (Gabay et al. 2021).

At last, thanks to T. Clérice, continuous integration tools have been built to ensure the homogeneity of the XML data:

- ChocoMufin (Clérice and Pinche 2021a) for the uniform use of characters in the dataset[3]
- HTRUX (Clérice and Pinche 2021b) for verification of Alto XML and respect of the segmOnto ontology

II- HTR Models for Medieval manuscripts

We will present here three different models all trained with CREMMA Medieval dataset and Kraken (Kiessling 2019). Each model was trained with the following more efficient architecture for manuscripts : [1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do] and an NFD character uniformity.

1. Models

All scores given here are scores calculated by Kraken on 10% of the global training corpus (test set) that were never seen during training. The accuracy is based on the characters error rate. Caution should be exercised with these results, as none of the models presented here were trained on the same training set, and thus tested on the same test set.

- **Bicerin 1.0.0** (DOI : 10.5281/zenodo.5235186, 21/07/13), accuracy 95.49%, The model is based on the first CREMMA Medieval dataset and specialized on 13th and 14th century manuscripts.
- **Bicerin 1.1.0** (DOI : 10.5281/zenodo.6669553, 22/06/22), accuracy 95.30%. The model is based on CREMMA Medieval extended to 15th c. manuscripts (see table 1).

2 <https://segmonto.github.io>

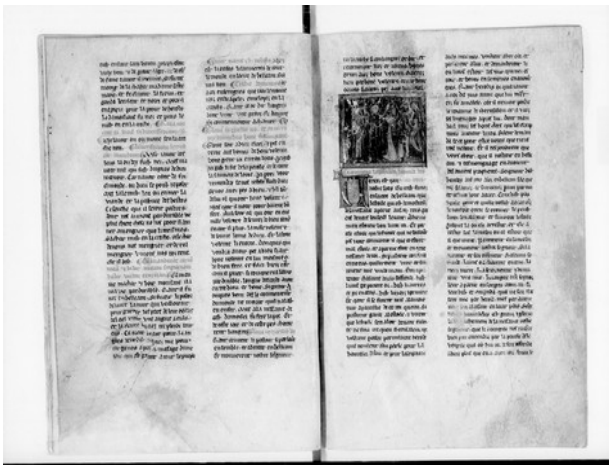
3 A specific table has been made during *CremmaLab* project in conjunction with the transcription guide specifically for Cremma Medieval, <https://github.com/HTR-United/cremma-medieval/blob/main/table.csv>.

- **Cortado 2. 0.0** (22/06/22), accuracy 95.54 %. It is a model that mixes CREMMA Medieval dataset with early prints (15th c.) from *Gallic(orpor)a project* (Pinche et al. 2022). This is still a work in progress.

2. Test set out-of-domain

In order to estimate the robustness of the different models, an "out-of-domain" test set was constructed. An extract of French manuscripts was chosen for each of the following centuries: 13th, 14th and 15th centuries. Each extract was segmented and transcribed according to the rules of the CREMMA Medieval dataset.

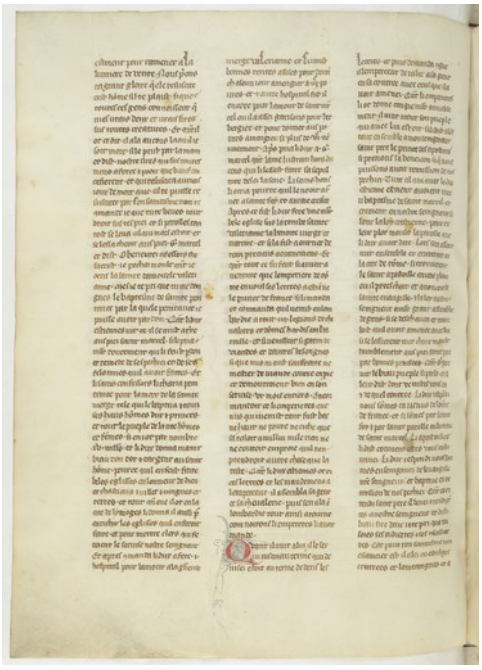
2.1 BnF, manuscripts, fr. 17229, 13th c.



The digitization of the manuscript is in black and white. This manuscript is a two-column manuscript written in Textualis Gothic script. The dark ink tends to lighten in the body of the text. This document is very similar to the documents that make up the medieval CREMMA dataset. Two folios of this manuscript have been transcribed, containing 10178 characters.

Source gallica.bnf.fr / Bibliothèque nationale de France. Département des Manuscrits. Français 17229

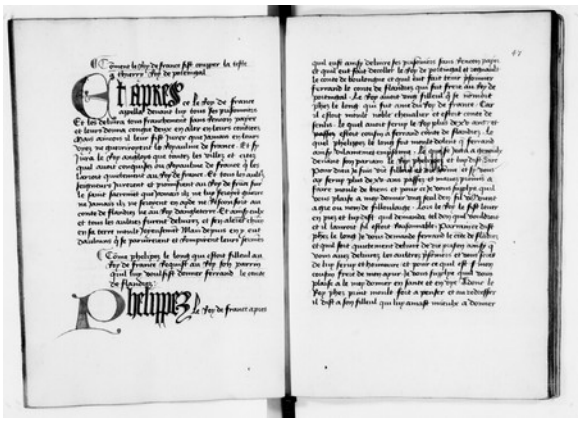
2.2 BnF, manuscripts, fr. 185, 14th c.



The digitization of the manuscript is in colour and in high definition. The manuscript is written in Textualis Gothic script. It has many headings in red ink. The black ink tends to fade in some places. This document is very similar to the documents from the medieval CREMMA dataset. For this manuscript, three folios of one column, i.e. 26353 characters, have been transcribed.

Source gallica.bnf.fr / Bibliothèque nationale de France. Département des Manuscrits. Français 185

2.3 BnF, manuscripts, nouv. Acq. fr. 6213, 15th c.



The digitization of the manuscript is in black and white. Unlike the other documents, the text was not written on parchment, but on paper. The manuscript is written in Hybrida script. The first words of some paragraphs are in larger modules and mostly in capital letters. This document is the one that differs most from the training dataset. For this manuscript, we have transcribed three folios of three columns, i.e. 13701 characters.

Source gallica.bnf.fr / Bibliothèque nationale de France. Département des Manuscrits. Nouvelles acquisitions françaises 6213.

3. Prediction and scores

A test on each document was performed with the three different models to see how they behave and how robust they are on documents that the HTR engine has never seen during its training. The aim of this experiment is to evaluate the best way to build a generic model directly reusable by the community.

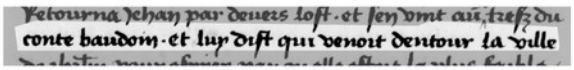
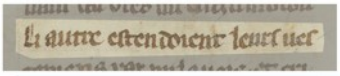
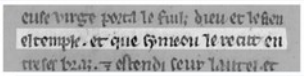
3.1 score table (accuracy based on CER)

All scores given here were calculated by the Kraken testing tool.

	BnF, Ms, fr. 17229	Bnf, Ms, fr. 185	BnF, Ms, NAF 6213	ALL
Cortado 2.0.0	**92.71%**	**92.07%**	**87.48%**	**90.95%**
Bicerin 1.1.0	91.64%	91.34%	83.40%	89.23%
Bicerin 1.0.1	90.66%	88.45 %	79.67%	86.50%

3.2 Examples of Cortado model predictions

All predictions were made through the eScriptorium platform.

BnF, ms., fr. 17229	 conte baudoi. et luy dist qui uenoit dentour la ville
BnF, ms., fr. 185	 li autre estendoient leurs ues
BnF, NAF, 6213	 eltemple. et que symeou le recut en

3.3 Most common errors in prediction

All the information given here was provided by the Kraken test tool.

3.3.1. Table 2 : BnF, manuscripts, fr. 17229, 13th c.

Cortado 2.0.0		**Bicerin 1.1.0**		**Bicerin 1.0.1**	
10178 Characters		10178 Characters		10178 Characters	
742 Errors		851 Errors		951 Errors	
92.71% Accuracy		91.64% Accuracy		90.66% Accuracy	
NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}
110	{ SPACE } - { }	146	{ } - { SPACE }	239	{ } - { SPACE }
89	{ } - { SPACE }	124	{ SPACE } - { }	84	{ SPACE } - { }
30	{ } - { COMB. TILDE }	33	{ } - { COMB. TILDE }	30	{ } - { i }
29	{ n } - { u }	27	{ } - { i }	23	{ u } - { v }
20	{ } - { i }	23	{ u } - { v }	22	{ u } - { n }

3.3.2. Table 3 : BnF, manuscripts, fr. 185, 14th c.

Cortado 2.0.0		**Bicerin 1.1.0**		**Bicerin 1.0.1**	
26353 Characters		26353 Characters		26353 Characters	
2091 Errors		2283 Errors		3044 Errors	
92.07% Accuracy		91.34% Accuracy		88.45% Accuracy	
NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}
352	{ SPACE } - { }	439	{ SPACE } - { }	368	{ SPACE } - { }
145	{ i } - { }	123	{ . } - { }	191	{ . } - { }
111	{ . } - { }	116	{ i } - { }	176	{ i } - { }
94	{ } - { SPACE }	97	{ n } - { }	124	{ } - { SPACE }
78	{ } - { COMB. TILDE }	90	{ } - { SPACE }	116	{ n } - { }

3.3.3. Table 4 : BnF, manuscripts, nouv. Acq. fr. 6213, 15th c.

Cortado 2.0.0		**Bicerin 1.1.0**		**Bicerin 1.0.1**	
13701 Characters		13701 Characters		13701 Characters	
1715 Errors		2275 Errors		2785 Errors	
87.48% Accuracy		83.40% Accuracy		79.67% Accuracy	
NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}
84	{ a } - { o }	211	{ a } - { o }	136	{ a } - { o }
79	{ } - { COMB. TILDE }	111	{ } - { i }	125	{ } - { i }
59	{ } - { i }	85	{ } - { COMB. TILDE }	107	{ } - { SPACE }
59	{ . } - { }	71	{ } - { SPACE}	89	{ r } - { i }
56	{ SPACE } - { }	66	{ . } - { }	85	{ . } - { }

3.3.4. Table 5 : All manuscripts

Cortado 2.0.0		**Bicerin 1.1.0**		**Bicerin 1.0.1**	
50232 Characters		50232 Characters		50232 Characters	
4548 Errors		5409 Errors		6780 Errors	
90.95% Accuracy		89.23% Accuracy		86.50% Accuracy	
NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}	NB	Type {expected-value}-{error}
518	{ SPACE } - { }	625	{ SPACE } - { }	523	{ SPACE } - { }
232	{ } - { SPACE }	307	{ } - { SPACE }	470	{ } - { SPACE }
204	{ i } - { }	212	{ a } - { o }	297	{ . } - { }
187	{ } - { COMB.TILDE }	202	{ . } - { }	252	{ i } - { }
176	{ . } - { }	189	{ } - { COMB. TILDE }	188	{ } - { i }

4. Interpretation of the results

Looking at tables 2, 3, 4, 5 (see *3.3 Most common errors in prediction*) which contain the five most common errors made for each model, it can be seen that the biggest problem is the prediction of spaces, which is to be expected since even for a human being it is difficult to determine where they are. For the 13th and 14th century manuscripts, the errors are classic palaeographic errors related to

the counting of the legs of the letters. Distinguishing between “u” and “n” can be difficult, problem even for palaeographers. In table 4 for the BnF, NAF, 6213, we see clearly that this manuscript has a distinct way of writing “a”, as it uses single-looped round “a” whereas in the other two documents “a” are double-looped. Thus, the round “a” of the Hybrida script is the cause of the confusion between “a” and “o”, which is the most recurrent error in the three models for this manuscript.

For the two releases of Bicerin, results can be confusing, seeing that Bicerin 1.1.0 has a lower accuracy “in-domain”. The reason is that the model is less specific due to its openness to a manuscript from the 15th. Thus, the more specific the model, the higher the score “in the domain”, but this is not indicative of its robustness “out of the domain”, as the result of Bicerin 1.1.0 in Table 1 proves. In fact, even if its accuracy “in-domain” is lower than its first release; “out-of-domain”, the second version is always between one and three points higher, showing a better robustness. It thus seems that the variety of training data improves the robustness of a model.

This theory is reinforced by the results of Cortado model. Its in-domain accuracy shows that it is the best model, which is also confirmed by the “out-of-domain” tests. The variety of its dataset, extended to early 15th century prints, makes it more robust on “out-of-domain” documents, even on the 13th century manuscript. The difference in performance is really evident on the 15th century manuscript, with 4 points more accuracy than Bicerin 1.1.0 and almost 8 points more than Bicerin 1.0.1. Certainly, if we want to extend our model to 15th century manuscripts, we will have to add Hybrida manuscripts in the set to increase its adaptability to the variation of letters in the different variants of the Gothic script and avoid problems like those encountered with the confusion between “a” and “o”. This will certainly quickly allow us to have reusable predictions for the manuscripts of this period.

In conclusion, we can deduce that the “in-domain” accuracy score does not tell us everything about the performance of the model. Thus, the more specific the model is, the higher the score will be “in-domain”, but this is not indicative of its robustness “out-of-domain”. Indeed, the latest version of Bicerin is more robust on unknown documents, even if its score is lower than the one from the previous version. We can therefore conclude that for a generic model, the variety of the training set is important, even in our case with an early prints for Cortado model. Finally, a generic model can always be quickly fine-tuned on a given corpus to improve the results on a particular document if necessary.

References

Chagué, Alix, Thibault Clérice, and Floriane Chiffoleau. 2021. *HTR-United, a Centralization Effort of HTR and OCR Ground-Truth Repositories Mainly for French Languages*.

Clérice, Thibault, and Ariane Pinche. 2021a. *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects* [computer program]. Version 0.0.4. ttps://github.com/PonteIneptique/choco-mufin.

Clérice, Thibault, and Ariane Pinche. 2021b. *HTRVX, HTR Validation with XSD* [computer program]. Version 0.0.1. https://github.com/HTR-United/HTRVX.

Derolez, Albert. 2003. *The palaeography of Gothic manuscript books: from the twelfth to the early sixteenth century*. Cambridge, Royaume-Uni .

Fischer, Andreas, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2010. “Ground Truth Creation for Handwriting Recognition in Historical Documents.” pp. 3–10 in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, *DAS ”10*. New York, NY, USA: Association for Computing Machinery.

Fischer, Andreas, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2009. “Automatic Transcription of Handwritten Medieval Documents.” pp. 137–42 in *2009 15th International Conference on Virtual Systems and Multimedia*.

Kiessling, B., R. Tissot, P. Stokes, and D. Stökl Ben Ezra. 2019. “EScriptorium: An Open Source Platform for Historical Document Analysis.” pp. 19–19 in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2.

Kiessling, Benjamin. 2019. “Kraken - an Universal Text Recognizer for the Humanities.” UItrecht: CLARIAH.

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

Pinche, Ariane. 2022. « Guide de transcription pour les manuscrits du X^{e} au XV^{e} siècle. », (https://hal.archives-ouvertes.fr/hal-03697382).

Pinche, Ariane. 2021. « Projet CREMMALAB. » (https://cremmalab.hypotheses.org/23).

Pinche, Ariane, Simon Gabay, Noé Leroy, and Kelly Christensen. 2022. *Données HTR Incunables u 15^{e} siècle*.