CREMMALab Project: Handwritten text recognition (HTR) for medieval manuscripts

Ariane Pinche

CREMMALab Project

Ariane Pinche
École nationale des chartes, Centre Jean Mabillon
DH-Tokyo 2022

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 French medieval literary manuscripts for HTR (see the *cremma-medieval* repository on Github : https://github.com/HTR-United/cremma-medieval).

The challenge of HTR is still to be taken up because of the great variety of writings, the poor readability, even for a human reader, of handwritten documents, whether due to the degradation of the source or the lack of homogeneity of the writing. Finally, the production of training data is extremely costly. However, the technical progress of the last few years in AI and neural networks allows us to produce data that significantly reduces manual work. In the Humanities, if some tools have emerged such as Transkribus (Kahle and al.), 2017) or Kraken (Kiessling, 2019) and its interface eScriptorium (Kiessling et al., 2019), developed at the EPHE, PSL university, we lack data on medieval documents to train performant models. So, producing training data is today a major issue.

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 on HTR-united (Chagué et al., 2021). The CREMMALab project implements transcription protocols to optimize the training of HTR models and to eventually produce homogeneous and shareable data. As a first step towards the FAIR principles, through the *cremma-medieva*l repository, we have set up some tools to ensure the citability, the durability and the quality of the data. Thus, data are described (language, date, type of document, transcription method, number of transcribed lines etc.). Then thanks to continuous integration tools, the compatibility of the XML data is checked (HTRUX), as well as the uniformity of the character sets used in the corpus (chocoMufin). Through the gathering of a corpus of medieval manuscripts, the learning process of the HTR algorithms is examined to evaluate the problems related to the constitution of training data : how to transcribe, handle abbreviations, segment words and so on. We also seek to determine thresholds of ground truth lines needed to meet quality goals, and also to evaluate the impact of the training corpus on the quality and genericity of the models.

Bicerin, an HTR model for medieval French manuscripts, is already available. It has been trained on eleven manuscripts written between the 13th and 14th centuries in Gothic script (about 18400 transcribed lines, see table 1). Its accuracy (CER) on this corpus is 95.38 % (dev score). This generic model, which still needs to be improved produces predictions for similar out-of-domain sources (Chantilly, Bibliothèque du Château, ms. 734, 14^{e} siècle) with an accuracy of 83 % (test score, see figure 1). A quick customization, requiring the addition of only two folios (336 transcribed lines) achieves an accuracy of 91 % (test score with the same sample as the first experiment, see figure 2). Most of the recognition problems come from specific difficulties due to

1 Presentation project : <https://cremmalab.hypotheses.org>

the manuscript : word segmentation, distinction between *u* and *n*, abbreviations and reproduction quality.

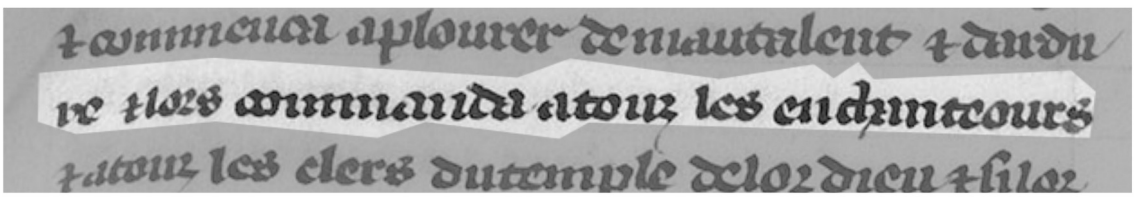
The model can also be customized on a document which, at first sight, might seem incompatible with the training corpus. The manuscript codex 909 from the University of Pennsylvania was written in France at the end of the 15th century, or at the very beginning of the 16th century. It is written in *Bâtarde* (this script is a hybrid of the formal style with a cursive script). Applied directly, the model shows an accuracy of 80 % (test score, figure image 3). However, *Bicerin* is very flexible, and with a customization from the addition of two folios (320 manually transcribed lines), we get an accuracy of 97 % (test score with the same sample as the precedent experience, see figure 4). The high recognition of this writing is certainly related to the regularity of the writing and the quality of the support and the reproduction, two criteria that seem to be more important for HTR than the type of writing.

In the future, we hope to increase the number and diversity of our training data to improve the genericity of the model and its robustness. We also hope to determine its breakpoints and to delimit contexts in which a generic model is enough and those in which it is relevant to create a personalization or another model from scratch.

Annexes

Table 1 : Cremma medieval corpus

Manuscript	**Date**	**Transcribed Lines**
BnF, ms fr. 17229	13th	161
BnF, ms fr. 13496	13th	159
BnF, ms fr. 411	14th	153
BnF, Arsenal 3516	13th	1991
BnF, ms fr. 22549	14th	411
BnF, ms fr. 24428	13th	1295
BnF, ms fr. 412	13th	4551
BnF, ms fr. 844	13th	1026
Cologny, bodmer, 168	13th	1927
Vaticane, Reg. Lat., 1616	14th	1726



re nlors counmnanda a touz les eu chiceours

Figure 1: Bicerin model prediction on Chantilly, ms. 734

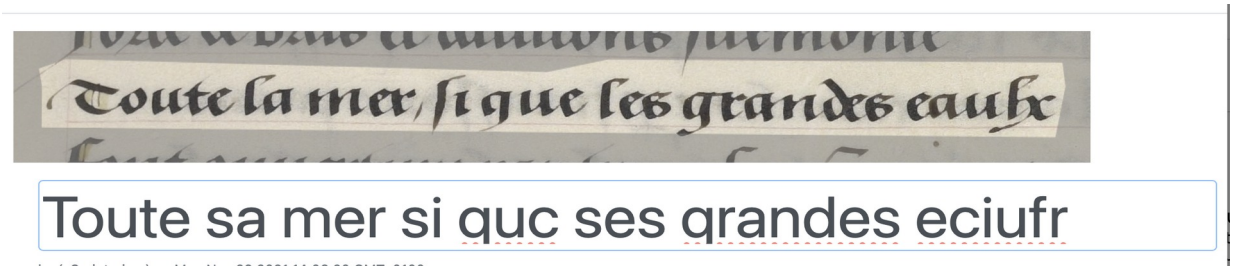


Figure 2: Bicerin model prediction on Philadelphy, university of pennsylvania, ms codex 909

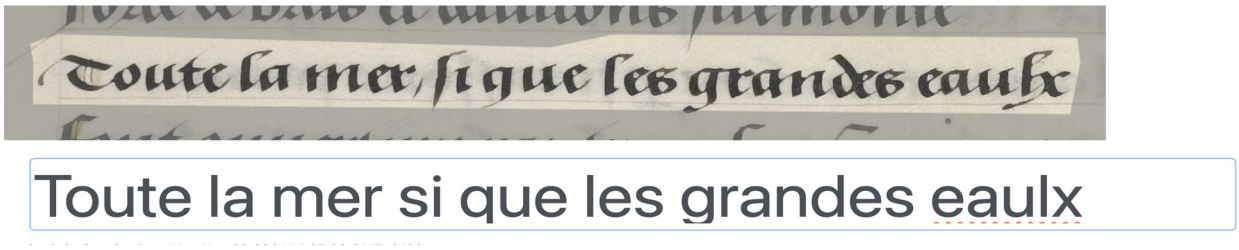


Figure 3: Finetuned model prediction on Philadelphy, university of pennsylvania, ms codex 909

Figure 4: Finetuned model prediction on Chantilly, ms. 734

References

Bulacu, M., & Schomaker, L. (2007), "Automatic Handwriting Identification on Medieval Documents", *14th International Conference on Image Analysis and Processing (ICIAP)*, 2007, 279-284, <https://doi.org/10.1109/ICIAP.2007.4362792>.

Chagué, A., Clérice, T., & Chiffoleau, F. (2021), *HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages*, <https://github.com/HTR-United/htr-united> (Original work published 2020)

Dome, S., & Sathe, A. P. (2021), "Optical Charater Recognition using Tesseract and Classification", *International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, 153–158. https://doi.org/10.1109/ESCI50559.2021.9397008

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., & Stolz, M. (2009). "Automatic Transcription of Handwritten Medieval Documents", *15th International Conference on Virtual Systems and Multimedia*, 2009, 137–142, <https://doi.org/10.1109/VSMM.2009.26>.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). “Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”, *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, 4, 19–24, <https://doi.org/10.1109/ICDAR.2017.307>.

Kestemont, M., Christlein, V., & Stutzmann, D. (2017), « Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts. *Speculum* », 92(S1), S86–S109. https://doi.org/10.1086/694112

Kiessling, B. (2019), “Kraken—An Universal Text Recognizer for the Humanities”, *DH2019*, Utrecht, <https://dev.clariah.nl/files/dh2019/boa/0673.html>.

Kiessling, B., Tissot, R., Stokes, P., & Stoekl, D. (2019), “eScriptorium: An Open Source Platform for Historical Document Analysis”, *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, 19–19, <https://doi.org/10.1109/ICDARW.2019.10032>.

Pinche, A., & Clérice, T. (2021), “HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI)”, *Zenodo*, <https://doi.org/10.5281/zenodo.5235186>.

Ströbel, P. B., Clematide, S., & Volk, M. (2020), « How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR », *Proceedings of the 12th Language Resources and Evaluation Conference*, 3551–3559. https://aclanthology.org/2020.lrec-1.436