Manu McFrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets

Alix Chagué, Thibault Clérice, Jade Norindr, Maxime Humeau, Baudoin Davoury, Elsa Van Kote, Anaïs Mazoue, Margaux Faure, Soline Doat

Proposal to DH2023: Collaboration as Opportunity

Manu McFrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets

Alix Chagué[1,2,3], Thibault Clérice[1,4], Jade Norindr[5], Maxime Humeau[5], Baudoin Davoury[5], Elsa Van Kote[5], Anaïs Mazoue[5], Margaux Faure[5], and Soline Doat[5]

[1]ALMAnaCH - Automatic Language Modelling and Analysis & Computational Humanities, Inria, Paris, France
[2]UdeM - Université de Montréal, Montréal, Canada
[3]EPHE - École Pratique des Hautes Études, Paris, France
[4]CJM - Centre Jean Mabillon, Paris, France
[5]Énc - École nationale des chartes, Paris, France

November 2022

Since the mid-2010s, Handwritten Text Recognition (HTR) has become an important opportunity for digital humanists and cultural institutions to explore and retrieve textual information from handwritten documents. The creation of software equipped with graphical user interfaces (GUI) like Transkribus[15] and Kraken-eScriptorium [14] facilitates the annotation of ground truths (perfect transcriptions which can be used for training models) which can later be exported in the form of pairs of images and XML files (ALTO XML or PAGE XML) containing the text equivalent as well as the location of the text on the image. The *Consortium pour la Reconnaissances d'Écritures Manuscrites des Matériaux Anciens*[1] (CREMMA) project was initiated in 2021 with the aim of funding a regional server capable of supporting fast training of HTR models, for students and researchers of the Paris region. It consisted in a

[1]Consortium for handwritten text recognition on ancient materials.

starting grant of 42,000€ covering the cost of the hardware (graphic cards, servers, etc) as well as an evaluation grant dedicated to providing base models for the users of CREMMA (8,000€), in particular for two languages: French and Latin, from the 9th to the 21st centuries. A postdoctoral position, CREMMAlab, provided the infrastructure with complementary time for building a dataset (CREMMA Medieval[16]) and expertise around transcribing medieval manuscripts.

Simultaneous to the creation of CREMMA, the HTR-United[4] initiative offers a solution to facilitate conformity to the FAIR principles[2] when HTR users create and share datasets of ground truth. It consists in both a catalog of machine-actionable metadata on open datasets of HTR ground truth and a toolkit to strengthen the control of the documentation as well as the validity of the data. As of early November 2022, it comprehends 58 datasets, composed of 18,155 pairs of images and XML files, which represent over 41.5 millions of characters, covering 13 languages and 6 scripts[3]. While designing HTR-United, we became aware of the importance of spending part of the CREMMA budget in the creation of new corpora and models.

1 Manu McFrench and its datasets

As of November 2022, 9 CREMMA datasets are described in the HTR-United catalog: *CREMMA Medii Aevi*[8], *CREMMA Medieval*[16], *CREMMA Manuscrits du 17e*[9], *CREMMA Manuscrits du 18e*[5], *CREMMA Manuscrits du 19e*[10], *CREMMA Manuscrits du 20e*[7], *CREMMA-Wikipedia*[11], *CREMMA-AN Testament De Poilus*[12] and *CREMMA Early Modern Books*. They gather ground truth for, in order, Latin and Old French manuscripts from the medieval period, French manuscripts from the 17th, 18th, 19th, 20th and 21st centuries, French manuscripts from the *Testament de Poilus* corpus[6] as well as early modern books (printed) in Latin and modern French. Put together, these datasets amount to 1,148 pairs of XML files and images, spanning over 1.3 million characters. These datasets were contributed by Thibault Clérice and Alix Chagué, as well as students from the Master programs of the École nationale des chartes (Paris) hired within the frame of CREMMA to execute transcription or alignment tasks.[4]

The first version of a transcription model for French modern and contemporaneous texts (called "Manu McFrench") was trained with Kraken[13] in June 2022. We used the data generated through CREMMA for the corresponding periods as well as datasets signaled in HTR-United[3] which shared the same transcription guidelines

[2]Findable, Accessible, Interoperable, Reusable.

[3]Not all projects provide fine-grained descriptive statistics about their datasets.

[4]In some cases, we provided original transcriptions from in-house projects, which had to be proof-read and realigned with the original material.

Dataset name	Project or company	Century	Language	Lines	Characters	Hands
CREMMA Manuscrits du 17e	CREMMA	17	French	2,245	81,909	Few
CREMMA Manuscrits du 18e	CREMMA	18	French	4,017	141,747	Few
Notaires de Paris - Bronod	LECTAUREP	18	French	3,708	359,676	Few
CREMMA Manuscrits du 19e	CREMMA	19	French	1,807	55,581	Few
Projet Correspondance Berlioz	ENC - BPDC	19	French	367	13,474	Few
Projet Notre-Dame	ENC - BPDC	19	French	735	29,286	Few
TIMEUS Corpus	ANR TIME US	19	French	7,701	746,997	Many
Notaires de Paris - Mariages et Divorces	LECTAUREP	19-20	French	20,305	1,969,585	Many
Notaires de Paris - Répertoires	LECTAUREP	19-20	French	29,410	525,619	Many
CREMMA Manuscrits du 20e	CREMMA	20	French	224	5,764	Few
CREMMA-AN Testaments de Poilus	CREMMA	20	French	1,353	33,652	Many
CREMMA Wikipedia	CREMMA	21	French	1,353	33,652	Many
Araucania	Araucania	19	Spanish	3,932	117,155	Few
Memorials for Jane Lathrop Stanford	ENC - BPDC	20	English	770	18,063	Few
Total manuscript				77,927	4,132,160	
Données imprimés du 16e siècle	Gallicorpora	16	French	4,918	186,202	N/A

Table 1: Datasets used for Manu McFrench v3. Hands categories: Few means that at most there is less than 10 hands, Many means that there is nearly one hand per image. All datasets are described and available on HTR-United.

and were developed in eScriptorium.[5] The latest model, v3, has been trained using the materials shown in Table 1. The final model reaches a character recognition accuracy (CER) of 90.56% on our development set (*cf.* Figure 1).

2 Testing

We introduce two case studies to demonstrate how useful such models can be for the community of HTR users, specifically for project with low data yield or small budgets:

1. The *Recensement du Valais*[1], produced through the Valais Time Machine and the Sion Archives, proposes a set of census forms from 1880. Generally, each form is filled by a single person, which means that the dataset has nearly as many hands as it has files. The dataset is composed, at the time of writing, of 396 images, of which around 103 are in German. Only the manuscript portion of each form was transcribed, adding up to a total of 23,394 lines (lines are very short: they are similar to a table cell).

2. The Peraire Ground Truth dataset[2] was produced using images of Lucien Peraire (1906-1997)'s handwritten diaries, held at the Bibliothèque Sébert, Espéranto-France (Paris), during an exploratory experimentation for the Digital Peraire project. The documents are all written in French and date from the

[5]About this limitation, see the experiment by Pinche[16], section 3.2.

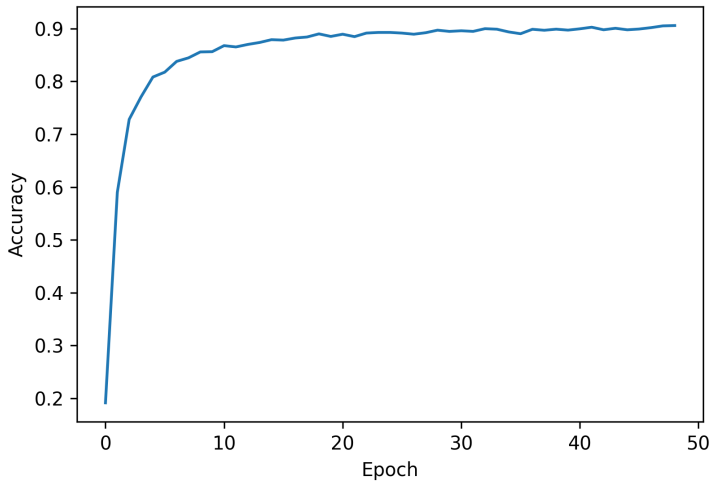


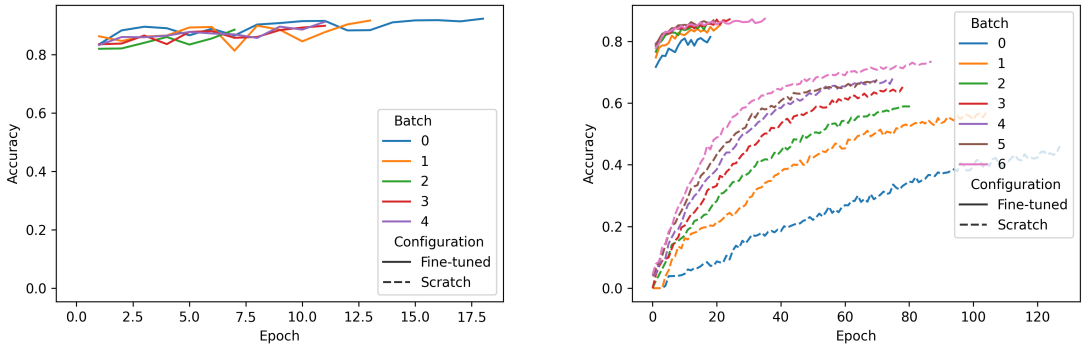
Figure 1: Training logs for Manu McFrench v3.

second half of the 20th century. The dataset is made of 33 images containing a total of 1,059 lines associated to 4 images for test purposes.

To evaluate the impact of Manu McFrench, each dataset is cut in smaller subsets. The *Recensement du Valais* is split in 8 subsets of a maximum of 50 images (around 3000 lines): each subset is composed of an equivalent amount of German and French (9 images in German, 41 in French), except for the single one we keep aside for test purposes (40% of German, 60% of French). The *Peraire* dataset kept its testing dataset (4 images) and the rest of its 33 images were split in subset of size 7 (the last one being 5 images). We then train model such as each model is trained with the same parameters[6], one using Manu McFrench for fine-tuning, the other without ("from scratch"). The training set are accumulated, so that subset 1 is used alone, subset 2 is used in addition of subset 1, etc.: in the end, the last trained model is composed of all training images.

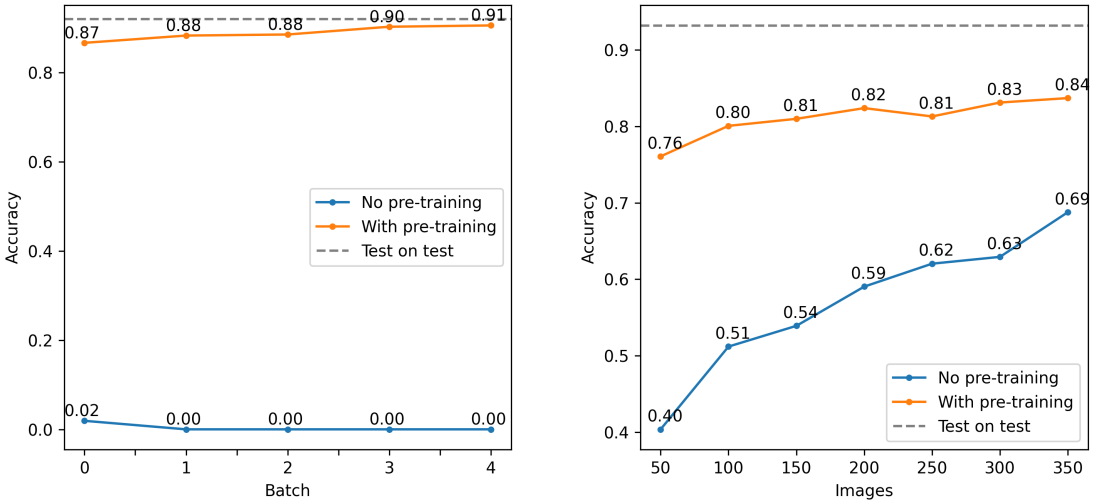
Overall, the training yielded much better results with Manu McFrench, both from a scoring point of view (Figure 2) and a training time one (Figure 3). This shows both the importance of generic, big models, that can then be used by smaller project to accelerate and lower the costs of transcription.

[6]Parameters: unicode normalization:`NFD`; Data Augmentation; Batch size: `16`; Learning Rate: `0.0001`, Model architecture: `[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do0.1,2 Lbx200 Do]`. Other parameters are the defaults from Kraken 4.1.2: we expect the low `lag` value (5) to be responsible for the absence of good accuracy for *Peraire*.



(a) Peraire Dataset (b) Valais Dataset

Figure 2: Training time (in epochs) based on the amount of data and the use of a pre-trained model (Manu McFrench v3). For the *Peraire* dataset, accuracy scores without Manu McFrench stay at 0 with this configuration.



(a) Peraire Dataset (b) Valais Dataset

Figure 3: Accuracy based on the amount of data and the use of a pre-trained model (Manu McFrench v3).

During the DH2023 conference, we would like to further introduce the CREMMA datasets and the strategies put in place to train the Manu McFrench model. We believe it is essential for the community to have access to similar robust and generic models: they can be costly to produce since they require a lot of ground truth and computation capacities, yet they are extremely effective in reducing the amount of ground truth later necessary to reach good performances when they can be fine-tuned on a specific handwriting.

References

[1] Dubois Alain et al. *Tables du recensement du Valais*. URL: https://github.com/PonteIneptique/valais-recensement.

[2] Alix Chagué. *Peraire Ground Truth*. Oct. 2022. DOI: 10.5281/zenodo.7185907. URL: https://github.com/alix-tz/peraire-ground-truth.

[3] Alix Chagué and Thibault Clérice. *HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French)*. Version 1.0.0. June 2022. DOI: 10.5281/zenodo.6657809. URL: https://doi.org/10.5281/zenodo.6657809.

[4] Alix Chagué and Thibault Clérice. "Sharing HTR datasets with standardized metadata: the HTR-United initiative". In: *Documents anciens et reconnaissance automatique des écritures manuscrites*. The recording of the conference is available at: https://www.canal-u.tv/chaines/enc/25-sharing-htr-datasets-with-standardized-metadata-the-htr-united-initiative. CREMMALab. Paris, France, June 2022. URL: https://hal.inria.fr/hal-03703989.

[5] Alix Chagué et al. *CREMMA Manuscrits du 18e*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/CREMMA-MSS-18.

[6] Florence Clavaud. "Testament de Poilus, une plateforme de transcription participative pour le grand public". In: *Archives participatives : d'une logique de guichet à une logique de co-construction*. Master 2 Gestion des archives et de l'archivage (Université de Versailles-Saint-Quentin-en-Yvelines). Pierrefitte sur Seine, France, Mar. 2019. URL: https://hal.archives-ouvertes.fr/hal-02076555.

[7] Thibault Clérice and Alix Chagué. *CREMMA Manuscrits du 20e*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/CREMMA-MSS-20.

[8] Thibault Clérice, Alix Chagué, and Malamatenia Vlachou-Efstathiou. *CREMMA Medii Aevi*. URL: https://github.com/HTR-United/CREMMA-Medieval-LAT.

[9] Thibault Clérice et al. *CREMMA Manuscrits du 17e*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/CREMMA-MSS-17.

[10] Thibault Clérice et al. *CREMMA Manuscrits du 19e*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/CREMMA-MSS-19.

[11] Thibault Clérice et al. *WikiCremma*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/cremma-wikipedia.

[12] de Champs Emmanuelle et al. *CREMMA-AN Testament De Poilus*. Ed. by Alix Chagué and Thibault Clérice. URL: https://github.com/HTR-United/CREMMA-AN-TestamentDePoilus.

[13] Benjamin Kiessling. *The Kraken OCR system*. Version 4.1.2. Apr. 2022. URL: https://kraken.re.

[14] Benjamin Kiessling et al. "eScriptorium: an open source platform for historical document analysis". In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. IEEE. 2019, pp. 19–19.

[15] Guenter Muehlberger et al. "Transforming scholarship in the archives through handwritten text recognition". In: *Journal of Documentation* 75.5 (Jan. 2019), pp. 954–976. ISSN: 0022-0418. DOI: 10.1108/JD-07-2018-0114. URL: https://doi.org/10.1108/JD-07-2018-0114.

[16] Ariane Pinche. "Generic HTR Models for Medieval Manuscripts The CREMMALab Project". working paper or preprint. Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03837519.