Manu McFrench, from zero to hero

Impact of using a generic handwriting recognition model for smaller datasets

Alix Chagué (ALMAnaCH, UdeM, ÉPHE)

Thibault Clérice (ALMAnaCH, CJM) et al.

July 12 2023

DH2023, Graz, Austria

Manu McFrench

Full list of authors

- Jade Norindr
- Maxime Humeau
- Baudoin Davoury
- Elsa Van Kote
- Anaïs Mazoue
- Margaux Faure
- Soline Doat

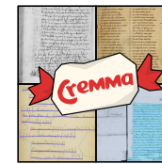
Overall, the problem

- OCR and HTR are great opportunities to access collections of documents and create textual corpora
- but transcription models are costly to produce because they require training examples
- luckily, we can rely on pre-existing models and pre-existing data!

The CREMMA Project

1 Introduction

- initiated in 2021 with a 50 000 € funding from DIM MAP
- objectives:
  1. build a server to offer eScriptorium to researchers and students in the Paris area (42 000€)
  2. produce training data (9th to 21st centuries) + generic model(s) for French and Latin mss (8 000 €)
- complemented by CREMMALab
  - more data for medieval French + a reference dataset (CREMMA Medieval)
  - a seminar gathering experts to agree on good practices and guidelines for medieval HTR



The issue

1 Introduction

- One of the objectives of CREMMA was the “creation of a generic model that would allow users to not start their transcription from zero”;
- But the budget alloted to this part was only 8 000 € (which is both plenty and not a lot);
- AND it had to be split between middle ages and modern / contemporary times;

With a budget of 8 000 €, and non-expert transcribers, what can you do ?

Table of Contents

2 Datasets behind Manu McFrench

Making the best out of a low resource situation

2 Datasets behind Manu McFrench

3 ways to optimize collecting training data

- Reuse existing datasets (few of them, but luckily we had HTR-United)[a];
- Reuse existing public digital or printed edition to create new training data;
- Create our own data but optimizing the transcription time (we explain more later)

Figure: Looking for solutions and datasets

[a] Some researchers identified an issue with the portability of Transkribus data to Kraken. *Cf.* 3.2 of A. Pinche, *Generic HTR Models for Medieval Manuscripts*, 2023

Reusing datasets: our transcription rules

The CREMMA guidelines are diplomatic: they respect the specificities of the document, as much as possible:

- no correction of the text;
- no abbreviation revolution;
- reproduction of some typographic phenomenons;
- no text normalization.

So: **we can't massively align datasets with images** because most printed and digitized edition do at least one of the above.

Reusing dataset: three level of reusable datasets

Outside of rights (of course), we can grade editions or transcriptions according to their (im)portability:

1. Abbreviation kept, no spelling correction, images available online (Extremely rare);
2. Abbreviation resolved but no spelling correction, images avail. (scarce);
3. Abbreviation resolved, spelling correction, images avail. (common);
4. Same as the above but no images :sadface:
5. Images, but no transcription

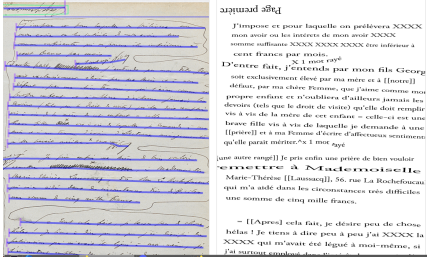
Level 1 dataset: the *Testaments de poilus*

2 Datasets behind Manu McFrench

Emmanuelle de Champs, Florence Clavaud, Pauline Charbonnier, Christine Nougaret, *et al.*, "Testaments de Poilus", 2022. `https://edition-testaments-de-poilus.huma-num.fr/`

1. Extract *batches* using metadata (dpt. of birth) to ensure diversity
2. Export plain text and images
3. Import into eScriptorium for segmentation and alignment
4. Adapt to CREMMA's *guidelines*

Easy to harvest because the metadata is stable, the transcriptions are clean and it's open (thus licensed)!



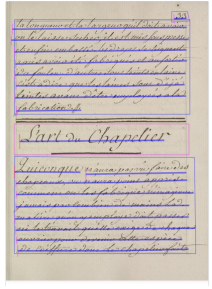
Github: HTR-United/CREMMA-AN-TestamentDePoilus

Level 2-3: Abrégé des Desc. des Artes

2 Datasets behind Manu McFrench

CREMMA-MSS-18 includes data reused from an American crowdsourcing campaign (Smithonian, using what looks like FromThePage software)

1. 179X Manuscript, single hand
2. Random selection of 20 pages
3. Manually downloaded images and copies transcription
4. Import into eScriptorium for segmentation and alignment
5. Adapt to CREMMA's *guidelines*



Github:
HTR-United/CREMMA-MSS-18

In this case the transcription practices can change from one to contributor to another.

Others ?

2 Datasets behind Manu McFrench

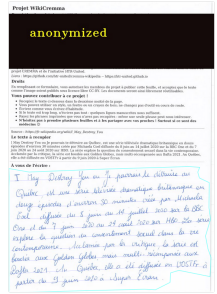
- We focused on French manuscripts $->$ not many already existing data to transform at the time
- We found a way to create more data at a lower cost

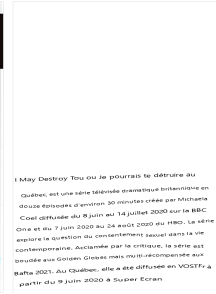
Low-cost and diverse: the CREMMA Wiki dataset

2 Datasets behind Manu McFrench

CREMMA-Wikipedia is a dataset made from scratch, by us and some of you!

1. Random pages selected from Wikipedia
2. Turned into a form that we ask volunteers to copy
3. Anonymized then imported in eScriptorium for segmentation and transcription
4. Published along with metadata




I May Destroy Tou ou je pourrais te détruire au Québec, est une série télévisé dramatique britannique en douze épisodes d'environ 30 minutes créée par Michaela Coel diffusée du 8 juin au 14 juillet 2020 sur la BBC One et du 7 juin 2020 au 24 août 2020 du HBO. La série explore la question du consentement sexuel dans la vie contemporaine. Acclamée par la critique, la série est récipiendaire aux Golden Globes mais multi-récompensée aux Bafta 2021. Au Québec, elle a été diffusée en VOSTFr à partir du 9 juin 2020 à Super Ecran

Github:
HTR-United/CREMMA-Wikipedia

No reuse limitation, no transcription problem or ambiguity, a varied vocabulary and varied handwritings.

Recap on datasets

2 Datasets behind Manu McFrench

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

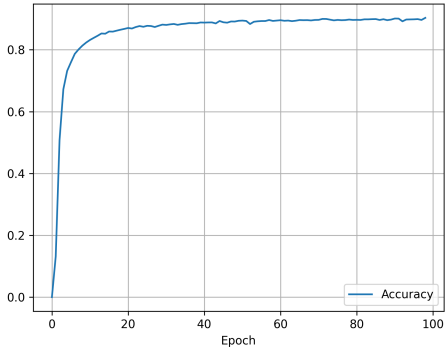
Table of Contents

3 Training Manu McFrench

Training Manu McFrench

3 Training Manu McFrench

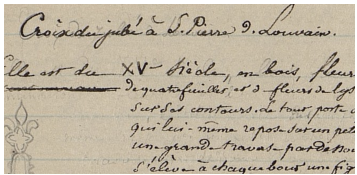
- 73,964 lines for train; 8,881 lines for dev (5%)
- Batch: 16, Padding: 16
- Augmentation (image distortion, etc.)
- Unicode norm.: NFKD
- Learning rate $1e^{-4}$
- Lag (Patience): 10
- Accuracy: 90.32%



[1,120,0,1 Cr4,2,32,4,2 Gn32 Cr4,2,64,1,1 Gn32 Mp4,2,4,2 Cr3,3,128,1,1 Gn32 Mp1,2,1,2 S1(1x0)1,3 Lbx256 Do0.5 Lbx256 Do0.5 Lbx256 Do0.5]

Raw results: the example of Quicherat manuscript

3 Training Manu McFrench



Croix du pjiulbé a CSe. Pierre d. Louvain.
Elle est du XV^re Siècle, en bois, fleuronnee ton>t<s
>< de quaturefecuilles ,, et de fleieurs de lays et engreâlée
sur daes contours. de tour porte dsurt un picolt
qui lui-même repos-e sutr un petist rocher et
uenne- grandeu travearse par dessous, d'ouǹ
Ls'élève à chaque bous une figure.; d'eun côté
la vierge et de l'autre S. Jean. JSours la
otiméme traverses, un tableau à
trois nvoolets, avec deux pieceds qui
descendent poceusr faire pa>r< l'aoffice de
pociontes d'appui Ssutr lae jube.
dLes feuillures du tableau deu
côté de la nef sont occupéues
Mpar Ttrois figures sculptéues.
Dans les bfleurons de la croitx
sont les animaux symbeoleiques.

	CER
Default	13.629
Ignoring digits	13.629
Ignoring case	13.312
Ignoring punctuation	13.311
Ignoring diacritics	12.479
Combining all options	11.764

Table of Contents

4 Fine-tuning Manu McFrench

All paths lead to a model

4 Fine-tuning Manu McFrench

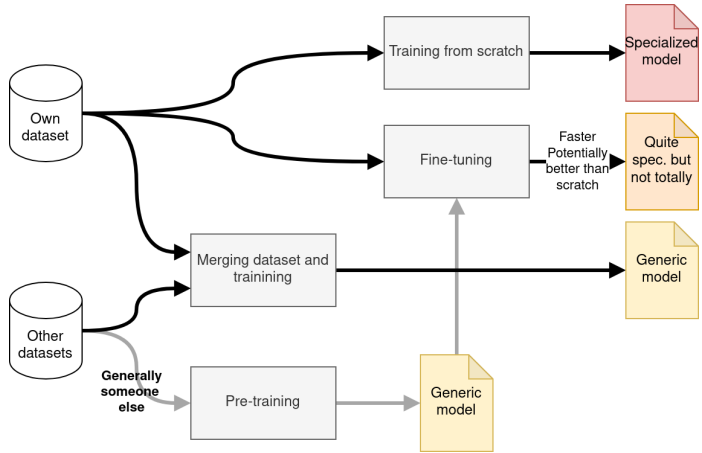


Figure: Possibilities regarding training a model for a specific project

Experimental set-up

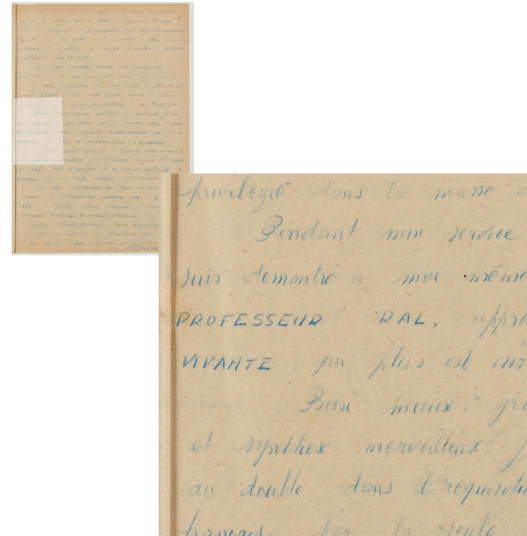
4 Fine-tuning Manu McFrench

We tested finetuning on 3 different datasets

- We split dataset in evenly(ish) subsets. We keep one of those subset for testing.
- We train a model on the first subset 1, with and without Manu McFrench as base (we use the same configuration for patience, learning rate, etc.)
- We train a model accumulating subset 1 until subset n.
- We keep track of the accuracy on the test set
- We evaluate a best scenario by training test (90/10) on itself and testing (unbeatable score).

Experiments: Peraire

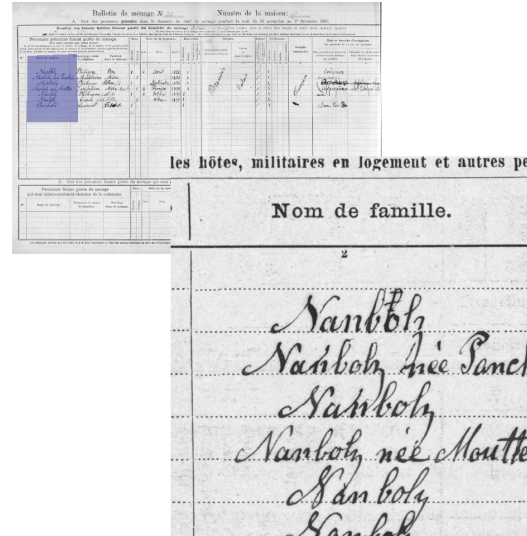
4 Fine-tuning Manu McFrench



- Dataset:
  - Peraire's travel notes
  - 1 hand, French, blue on school paper
  - Same guideline, "same" language as Manu McFrench

Experiments: Archives du Valais

4 Fine-tuning Manu McFrench

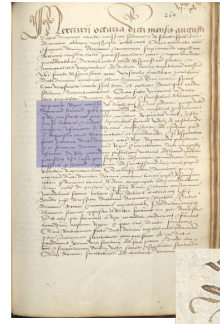


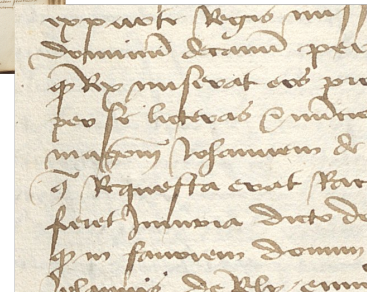
- Dataset:
  - Census form of 1880
  - 1 hand / form
  - 90% French, 10% German
  - Mostly proper names, some job names, and numbers
  - Same guideline, "same" language as Manu McFrench

Experiments: e-Notre Dame de Paris

4 Fine-tuning Manu McFrench

- Dataset:
  - Administrative documents (Middle Age)
  - Many hands
  - Mostly Latin, some Old/Middle French
  - Very repetitive
  - Different language, but also different guideline: normalization and abbreviation resolution embedded in the transcription





Conclusion of the experiments: Speed

4 Fine-tuning Manu McFrench

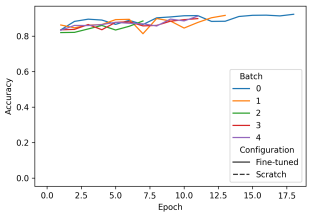


Figure: Peraire

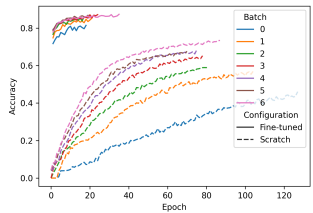


Figure: Valais

Figure: e-NdP data was lost for this metric

Conclusion of the experiments: Accuracy

4 Fine-tuning Manu McFrench

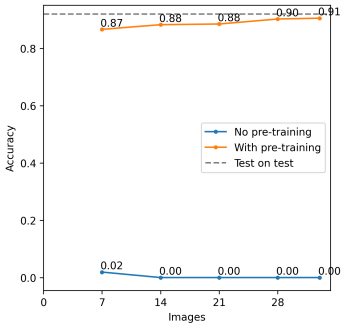


Figure: Peraire

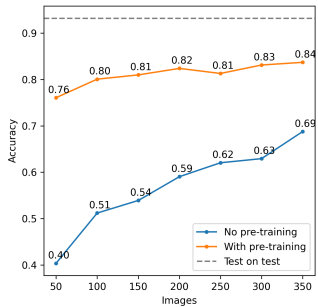


Figure: Valais

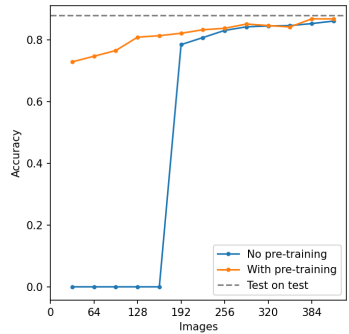


Figure: e-NdP

Table of Contents

5 Conclusion

Conclusion

5 Conclusion

- Open-source data means open-source and “massive” models are possible;
- Generic models provide healthy bases for training things, and speed-up the training for better results (in general);
- It can even speed-up training on what a human reader would consider a very different dataset using the same character ranges (but a different writing style).
- Other initiatives currently on-going: HTRomance & HTRogène

Q & A

Thank you for listening!
Your feedback will be highly appreciated!