To Overfit or Not to Overfit? An Evaluation of HTR Workflow on 17th-18th Century French Corpus

Marine TIGER
Sorbonne-Université, CELLF, CERES, 75006, Paris, France
marine.tiger@sorbonne-universite.fr

Abstract

This paper presents the results of an evaluation of general Handwritten Text Recognition (HTR) models applied to 17th and 18th century corpus written in modern French and the fine-tuning of the models. Our aim was to transcribe a corpus from this period using existing pre-trained models and to assess their performance on such data. While these general models offer a large linguistic coverage, our results demonstrate they are often insufficiently adapted to the specific handwriting nuances and orthographic inconsistencies of early modern French. To improve the results, we fine-tuned a base model to develop a specialized version trained on our dataset. Although the model still encountered difficulties due to highly variable handwriting styles, it significantly improved transcription accuracy and reduced processing time. Following this step, we used a semi-automatic post-correction tool to address remaining errors and integrated Named Entity Recognition (NER) steps for automated TEI-XML encoding. This paper discusses the evaluation results of both the HTR and NER models, and how the overfitting allows to get better transcriptions on a specific corpus.

Keywords: handwritten text recognition, named entity recognition, HTR-postcorrection, modern french, theater history

1. Introduction and related work

This paper presents the methodologies and the results of an evaluation of Handwritten Text Recognition (HTR) models on 17th-18th centuries French corpus, and the processing of the outputs with Named Entities Recognition (NER). This work is part of a project for the online publication of the *Comédie-Française* assembly registers from 1680 to 1921[1]. One of our primary goals is to publish the committee registers. This collection is particularly valuable because it offers insights into the company day-to-day life. The committee are formed by comedians of the troupe called "*sociétaire*"[2], and discuss a wide range of subjects, like rehearsals, budget, programs etc...(Sanjuan and Poirson, 2018) Our aim is not merely to provide access to the text as a digitized document, but also to enable large-scale research (e.g quantitative analysis tools).

The steps of the pipeline can be broken down as follow :

- completing the transcriptions using HTR ;
- post-correcting the transcribed texts ;
- encoding and validating the TEI files according to a TEI schema, and using NER to semi-automate the encoding of names, locations, dates, and plays ;
- and finally, integrating the data into a database for online publication.

To accelerate the transcription process, we use HTR, also known as Automatic Text Recognition (ATR) (Jacsont and Leblanc, 2023). This field of research, which emerged in the early 2000s, has experienced significant growth since the early 2020s. Early approaches of HTR relied on Hidden Markov Models (Marti and Bunke, 2002), while a shift occurred with the introduction of neural network with Connectionist Temporal Classification (CTC) (Graves et al., 2009), Long Short-Term Memory (LSTM) (Garrido-Munoz and Calvo-Zaragoza, 2025) and BLSTM (Granet et al., 2018). More recently, we can find transformer-based models (Ströbel et al., 2022).

Several tools and engines have democratized these technologies, like `eScriptorium` (Kiessling et al., 2019) (using the `Kraken` (Kiessling, 2019) engine), `Transkribus` (Colutto et al., 2019) and `Tesseract` (Smith, 2007). For French historical corpora, initiatives such as *Fondue* (Jacsont and Leblanc, 2023), *CATmUS* (Clérice et al., 2024) and *CREMMA* (Pinche, 2023) have provided essential datasets and generals models. We can mention *HTR-United* (Chagué et al., 2021), a collaborative platforms to collect dataset for ground truth and models. ATR is often considered a resolved task from a computer vision point of view (Hodel et al., 2021; Pinche, 2023), but it remains a challenging issues due to the high variation of hands. Nevertheless, the main challenge remains the ability to handle documents featuring a wide variety of hand-

[1] The *Comédie-Française* is a French theatre company founded in 1680 under the reign of King Louis XIV. It continues today to play a major role in the French theatre landscape(Sanjuan and Poirson, 2018).

[2] Comedians owning shares of the company

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1009–1016
11–16 May 2026.

writing styles. The performance of an HTR model largely depends on the availability of a large and diverse dataset, something difficult to achieve, as it requires the manual transcription of a considerable number of pages for model training. To enrich these transcriptions, we are working on a tool to encode them semi-automatically in XML-TEI format, by using Named Entity Recognition models (NER). Our goal is to recognize relevant person's name, locations, organizations, plays and dates to encode them automatically. In this paper, we present the results of the evaluation of the HTR models and the impact of overfitting on their performances, as well as our first results with the NER models. We are exploring the application of Handwritten Text Recognition (HTR) from a social science perspective rather than a purely technical one. Our approach focuses on evaluating existing tools and workflows and sharing our experience.

2. Corpus

Given that much of the 19th-century assemblies has been previously transcribed, this study focuses on the 18th-century registers : that is approximately 7,000 pages. Our corpus includes several characteristics relevant for HTR training :

- A redundant vocabulary (registers being administrative records);
- French modern language inducing spelling and orthographic variations over the century (Gabay et al., 2024);
- A lot of different hands (e.g. 2-3 hands during the same period).

Figure 1: Three different hands. Two from the same register (R52_3 (1712-1714)), one from the register R52_6 (1719-1721).

Models	1684		1729		1763		1784	
	WER	CER	WER	CER	WER	CER	WER	CER
McFrench	70.09	29.40	77.24	32.77	61.95	22.75	37.21	11.76
McFondue	76.42	30.51	80.26	34.31	68.83	23.39	40.11	10.21
Catmus	72.77	30.68	74.74	31.54	65.60	23.82	38.36	12.03

Table 1: Comparison CER and WER of each model on documents from 1684, 1729, 1763 and 1784.

3. Evaluation of HTR pre-trained models

3.1. Metrics and tool

First, we should state that the work involving HTR models was done using `eScriptorium` interface, as it's a free and open-source web application.

In HTR, as in most machine learning tasks, two main approaches can be distinguished: :

- Training a model from scratch: building and training a model yourself, based on a dataset specifically prepared for that purpose.
- Leveraging a pre-trained model, either by using it directly or by fine-tuning it on a new dataset (Pinche, 2023).

For our periods, we identified three available models at the time, all developed by the CREMMA lab :

- ManuMcfrench: trained on a dataset of 1.148 pairs of XML files and images (Chagué et al., 2023) ;
- ManuMcfondue: an extension of ManuMcfrench, trained on 4604 pages (Gabay et al., 2024) ;
- McCATmUS: trained on 180 manuscripts in 7 differents languages (Clérice et al., 2024).

We first evaluated the performance of this models to see how they performed on our corpus. To evaluate the performance of an HTR model, we generally used two metrics: the Character Error Rate (CER) and the Word Error Rate (WER). The CER calculates the percentage of mistakes at the character level, including letters, punctuation, and spaces. A CER below 10% is usually considered good, while 5% or less is excellent (Hodel et al., 2021). The WER measures errors at the word level. Because a word is counted as incorrect even if a single letter is wrong, it can be three to four times higher than the CER. In our case, the WER could be misleading, because our ground truth mainly consists of modernized transcriptions.

3.2. Overly general models?

These models had been evaluated on a dataset with four hands from four differents assemblies.

As we can see in the table 1, the CER is high, especially on pages from the end of the 17th to the middle of the 18th century. Furthermore, modern French evolved many times during centuries (Gabay et al., 2024), which increases the complexity of developing an efficient HTR model for these periods. The differences mainly concern handwriting styles and spelling, and we can observe that, across the four selected years, the CER varies noticeably. We argue that global models are less efficient while working on a specific corpus. We could even say that the larger the model, the more the CER will deteriorate, as demonstrated by the result of Mcfondue. It also confirms Humphries et al. (2024) statement, that models struggle to perform well on a hand that differs the training data.

The performance of these models is not sufficient to transcribe all the manuscripts. At this stage, there is no significant difference in time or effort between manual transcription and using the existing pre-trained models. Therefore, we've decided to train a custom model that better suited to our documents.

4. Training an HTR model

4.1. Building a ground truth

Training an HTR model requires preparing a dataset that serves as the project's ground truth. The ground truth is the original transcription of the manuscript with its segmentation.

For the assemblies registers, we rely on the transcriptions produced by **?**, which cover the late 17th and early 18th centuries, as well as those of the late 18th century transcribed by Master's students from Victoria University[3]. These transcriptions constitute a valuable basis for our work, as they provide part of the ground truth necessary for both the evaluation and the training of the model.

However, we still need the corresponding image-text alignment and the segmentation of the manuscript pages to be carried out in `eScriptorium`. Establishing a reliable ground truth is a time-consuming process, and it is also important to prepare a separate dataset specifically for evaluation purposes. Although the dataset could be supplemented with material from external sources such as HTR-United, the more heterogeneous the training data becomes, the less the resulting model will be optimized for this specific corpus.

At the time of writing, we prepared a dataset of 431 pages (415 808 tokens). It consists 331 pages from our corpus, supplemented with 100 pages from the CREMMA-lab project[4], to diversify our dataset with more hands (see the details in table 4 in the appendix).

4.2. Testing different approach for fine-tuning

In our case, the easiest way was to fine-tune an existing model, because it allowed us to develop a model with good performance without a large dataset (Jacsont and Leblanc, 2023).

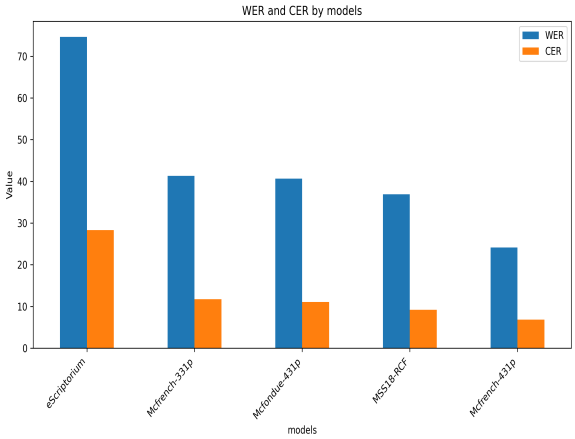

Figure 2: Comparison of the CER and the WER of the different trained models on pages from 1706

The figure 2 shows the WER and CER for the following models :

- **eScriptorium**: Fine-tuned with `eScriptorium`;
- **McFrench-331p**: Fine-tuned from the McFrench base model using 331 pages from the assemblies corpus;
- **Mcfrench-431p**: As above, but with the 431 pages mentioned;
- **Mcfondue-431p**: As above, but with Mcfondue as a base model;
- **MSS18-RCF**: Fine-tuned from the McFrench base model using a hybrid 431 pages set (half from MSS18 and half from our corpus).

According to the documentation, `eScriptorium` uses the default parameters of `kraken`. In fine-tuning case, `Kraken` will inherit the parameters of the base model[5].

[3]https://www.cfregisters.org/#!/equipe

[4]The pages are from 18th manuscripts (MSS18).

[5]For the McFrench model, these parameters include a batch size 16, `NFD` unicode normalization and the following model architecture : `[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3, 13,32 Do0.1,2 Mp2,2`

To have more control over the training process, we fine-tuned the other models in command line using the following `ketos` command:

```
ketos train -f binary -d cuda:0 -
r 0.0001 --resize both \
-i ../ManuMcFrenchV3.mlmodel -
o Mcf_fine_tuned dataset.arrow
```

This command resumes the training of an existing model using the dataset `dataset.arrow`. Training is performed on the GPU (`-d cuda:0`) with a learning rate of `0.0001`, which is appropriate for fine-tuning. The parameter `-f binary` indicates the binary format of the dataset, and `-resize both` allows the resizing of input images and text lines to match the model's expected architecture.

The results can be interpreted as follow :

- Fine-tuning McFrench outperformed McFondue. With the same ground truth, the difference between the two models reached nearly five CER points, suggesting that fine-tuning a base model pre-trained on a smaller dataset may be more effective;
- Diversifying the handwriting styles by incorporating corpora from other projects helps to improve the CER;
- Using `Kraken` directly and not the `eScriptorium` interface gave better performance thanks to the -"-resize both" parameter;

4.3. What about overfitting?

In machine learning, overfitting occurs when a model performs well on the training but not on unseen data. In HTR, this is characterized by a low CER on known hand, while it worsen on "similar but not identical hands" (Hodel et al., 2021).

Based of the results shown in figure 2, the best model is McFrench fine-tuned with 431 pages[6], most of them from our corpus (see the details in table 4 in the appendix).

Using the same evaluation metrics as described above, we used this model with McFrench on two different hands, one from our training dataset, another representing an unseen hand.

As shown in figure 2, the CER increases significantly between the two pages, confirming that the model is overfitted. However, since our goal is to transcribe 7000 pages in a short period of time, the overfitting is acceptable. Even with the deterioration observed for the page from 1729, the fine-tuned model still outperforms general model. We also benefit from working with a highly repetitive corpus in terms of structure: each session begins with a preamble whose content changes very little over the years. This part of the session is very well transcribed by the model.

We could continue to add more training data, but the number of pages already transcribed at our disposal is very limited, and the segmentation and text alignment is very time consuming. To address the remaining errors and potential decreases in performance, we focused on post-correction.

Models	1706		1725	
	WER	CER	WER	CER
McFrench	66.20	20.63	97.33	39.40
MSS18-RCF	21.37	4.83	56.00	18.21

Table 2: Comparison CER and WER of McFrench and MSS18-RCF on one page from 1706 and one page from 1729

5. Processing historical text with NER

5.1. Post-correction

Although our model performs reasonably well, it is important to note that a completely error-free transcription is impossible. Furthermore, as stated by Huynh et al. (2020), the quality of the model's transcriptions has a direct impact on NER performance.

There are three main methods to correct noisy texts: the manual and semi-automatic approaches described by Nguyen et al. (2022), and the machine learning approach. Manually correcting 5,000 pages would be too time-consuming, and fine-tuning a large language model requires substantial resources and a large dataset that we do not possess. Besides, late studies show that using LLM to post-correct historical text doesn't offer good performance (Boros et al., 2024) yet.

We implemented a semi-automatic post-correction tool based on a dictionary built from our HTR ground truth. Corrected errors are saved so that recurring mistakes are automatically fixed, significantly reducing manual intervention, particularly for confusions between similar letters (for example: 'v' and 'u'). However, overcorrections can occur when the same erroneous sequence corresponds to different valid words, depending on the transcription rules applied: should we preserve the writer's original spelling or normalize it?

The post-correction tool allowed us to gain CER points. We got a CER of 1.92 instead of 6.85 with the model. It is particularly efficient to correct the person's names, location or play titles.

`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]` (Chagué et al., 2023)

[6] GitHub repository of the model.

Model	PER	LOC	ORG	MISC	TOT
spaCy					
Ground Truth	230	34	68	113	**445**
model 431p	240	76	81	**129**	**526**
Post correction	222	36	71	**115**	**444**
Mcfrench	338	142	92	**232**	**804**
CamemBERT					
Ground Truth	845	31	41	70	**987**
model 431p	743	21	39	56	**859**
Post correction	844	27	41	60	**968**
Mcfrench	815	28	26	61	**930**

Table 3: SpaCy and CamemBERT results

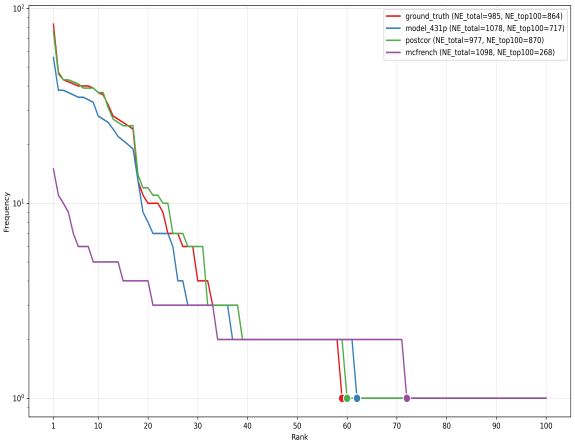

Figure 3: Zipf law for entities recognizes by SpaCy.

Having assessed the intrinsic performance of the tool through these error rate improvements, we then turned to an extrinsic evaluation to measure its impact on named entity recognition (NER) performance.

5.2. NER evaluation

We evaluate the performance on the widely used SpaCy (Honnibal et al., 2020) and CamemBERT (Martin et al., 2020) models. We compare the models outputs on four versions: the HTR ground truth, the Mcfrench transcriptions, the best fine-tuned model and the post-correction of the transcriptions. To compare the results, we appoint the output of SpaCy and CamemBERT on the HTR ground truth as our silver standard because of the lack of gold standard[7].

Our analysis focuses on two informations in the table 3: the total numbers of NEs and the numbers of MISC. As noted by Koudoro-Parfait et al. (2024), a higher WER often deteriorates the quality of the transcriptions, which increases NEs types. If we look back at the table 1, the WER of the Mcfrench model is very high, as is the number of NEs recognized by SpaCy. In contrast, the post-corrected text as a number of NE closed to the silver standard, both for SpaCy and CamemBERT. The latter finds a significantly higher number of NE, likely due to its tokenization, where compound name are often split and counted as multiple entities.

The figure 3 displays the occurrence of entities recognized by SpaCy per texts in a Zipf-like curve. We can observe that the post-correction followed the curve of the ground-truth, while the Mcfrench models decreases quickly. Besides, SpaCy detected 1098 entities, and only 268 are showing in this graph. This indicates that the noisier the output, the higher the number of hapax. Nevertheless, even if the post-correction as a significant suggestd that even reducing the noises in the transcriptions, SpaCy and CamemBERT continue to misidentify tokens as named entities (Parfait, 2025), or failed to label the correct entity. These two models are not trained on historical texts. Consequently, for any historical corpora, NER models require fine-tuning to accurately recognize named entities specific to a given period.

6. Conclusion

To conclude, our study shows that general pretrained HTR models yield poor Character Error Rates (CER) when applied to hands that differs significantly from their training data. Moreover,the transcription rules applied when creating the ground truth should be considered, as they may significantly affect the model's performance.By fine-tuning on a small, specific dataset, we achieved faster and more accurate transcriptions of our corpus through overfitting. However, this specialization also resulted in a decrease in performance when applied on hands different that where not in the training set.

Regarding the NER processing, we found that post-correction can improve the model's ability to recognize named entities. Nevertheless, regardless of the amount of noise present in the data (e.g uppercases that can be misleading as named entities, orthographic variations, etc.), NER models still require fine-tuning to effectively identify historical named entities.

7. Bibliographical References

[7]The NER models identifies four types of name entity (NE): persons (PER), location (ORG), organization (ORG) and others (MISC).

Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. *Journal of Imaging*, 10(1):18. Number: 1.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of Historical Text Transcripts with Large Language Models: An Exploratory Study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159. Association for Computational Linguistics.

Alix Chagué, Thibault Clérice, and Laurent Romary. 2021. HTR-United : Mutualisons la vérité de terrain ! In *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France. MESHS.

Alix Chagué, Thibault Clérice, Jade Norindr, Maxime Humeau, Baudoin Davoury, Elsa Van Kote, Anaïs Mazoue, Margaux Faure, and Soline Doat. 2023. Manu McFrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets. In *Digital Humanities 2023: Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'Connor, Wouter Haverals, Mike Kestemont, Caroline Vandyck, and Benjamin Kiessling. 2024. CATMuS Medieval: A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond. In Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng, editors, *Document Analysis and Recognition - ICDAR 2024*, volume 14806, pages 174–194. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Sebastian Colutto, Philip Kahle, Hackl Guenter, and Guenter Muehlberger. 2019. Transkribus. a platform for automated text recognition and searching of historical documents. In *2019 15th International Conference on eScience (eScience)*, pages 463–466.

Janez Demšar and Blaž Zupan. 2021. Hands-on training about overfitting. *PLOS computational biology/PLoS computational biology*. Place: United States.

Simon Gabay, Ariane Pinche, Peter Nahon, Alix Chagué, Pauline Jacsont, Élodie Paupe, Jean-Claude Rebetez, Maxime Humeau, Christine Payot, Thibault Maillard, Yvan Jauregui, Elina Leblanc, and Loraine Chappuis. 2024. Vers un modèle diachronique pour les mains modernes françaises.

Carlos Garrido-Munoz and Jorge Calvo-Zaragoza. 2025. On the Generalization of Handwritten Text Recognition Models.

Carlos Garrido-Munoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. 2025. Handwritten text recognition: A survey.

Adeline Granet, Emmanuel Morin, Harold Mouchère, Solen Quiniou, and Christian Viard-Gaudin. 2018. Transfer learning for handwriting recognition on historical documents. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, pages 432–439. INSTICC, SciTePress.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

Tobias Hodel, David Schoch, Christa Schneider, and Jake Purcell. 2021. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, 7(0).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy · Industrial-strength Natural Language Processing in Python.

Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2024. Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents. ArXiv:2411.03340 [cs].

Vinh-Nam Huynh, Ahmed Hamdi, and Antoine Doucet. 2020. When to Use OCR Post-correction for Named Entity Recognition? In *Digital Libraries at Times of Massive Societal Transition*, pages 33–42, Cham. Springer International Publishing.

Pauline Jacsont and Elina Leblanc. 2023. L'ATR en pratique : lumière sur les techniques de transcription automatique à Genève.

Benjamin Kiessling. 2019. Kraken - a Universal Text Recognizer for the Humanities. Artwork Size: 953044 Pages: 953044.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

Caroline Koudoro-Parfait, Ljudmila Petkovic, and Glenn Roe. 2024. Analyse multilingue de l'impact de la correction automatique de la ROC sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires. *Traitement Automatique des Langues*, 64(2):43–67.

Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, and Antoine Doucet. 2019. Impact of OCR Quality on Named Entity Linking. In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia.

U.-V. Marti and H. Bunke. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics. [link].

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2022. Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, 54(6):1–37.

Caroline Parfait. 2025. *Des IA au service de l'espace littéraire du XIXe siècle : évaluation et analyse des outils de reconnaissance d'entités nommées spatiales*. Ph.D. thesis, Sorbonne University.

Ljudmila Petkovic, Motasem Alrahabi, and Glenn Roe. 2022. Impact de la correction automatique de l'OCR/HTR sur la reconnaissance d'entités nommées dans un corpus bruité. *Journal of Information Sciences*, 21(2):42–57.

Ariane Pinche. 2023. Generic HTR Models for Medieval Manuscripts. The CREMMALab Project. *Journal of Data Mining & Digital Humanities*, Historical Documents and automatic text recognition.

Ariane Pinche and Peter Stokes. 2024. Historical documents and automatic text recognition: Introduction. *Journal of Data Mining & Digital Humanities*, Historical Documents and automatic text recognition.

Agathe Sanjuan and Martial Poirson. 2018. *Comédie Française. Une histoire du théâtre*, seuil, beaux livres edition. Paris.

Hugo Scheithauer. 2021. Un exemple d'exploitation des données produites grâce à la reconnaissance d'écriture manuscrite : la reconnaissance d'entités nommées.

R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 629–633, Curitiba, Parana, Brazil. IEEE.

Morgan Spinec. 2013. *"Dans l'arrière scène des seuls comédiens du roy"*. Thèse d'établissement, Ecole nationale des chartes.

Phillip Benjamin Ströbel, Simon Clematide, Martin Volk, and Tobias Hodel. 2022. Transformer-based htr for historical documents.

A. Appendix

Id/Name	Years	Total pages
***Comédie-Française* corpus**		
R52_0 (folios)	1682–1686, 1692, 1693, 1695, 1700, 1706	179
R52_12	1729	30
R52_17	1742	20
R52_20	1752	30
R52_21	1756	12
R52_26	1784	40
R124_a	1763–1773	20
MSS18 corpus		
Correspondance de Montfaucon	1730s	29
Archives de la Bastille (police secrète)	1729s	22
Candide de Voltaire	Unknown	5
Les dialogues de Rousseau	Unknown	5
Journal de Mme Genlis	1729s	10
Recueil de pièces diverses	Unknown	21
Vauban	Unknown	8

Table 4: Ground truth dataset content overview.