Ground-truth Free Evaluation of HTR on Old French and Latin Medieval Literary Manuscripts

Thibault Clérice[1,*,†]

[1]Centre Jean Mabillon, École nationale des Chartes, & INRIA

Abstract

As more and more projects openly release ground truth for handwritten text recognition (HTR), we expect the quality of automatic transcription to improve on unseen data. Getting models robust to scribal and material changes is a necessary step for specific data mining tasks. However, evaluation of HTR results requires ground truth to compare prediction statistically. In the context of modern languages, successful attempts to evaluate quality have been done using lexical features or n-grams.This, however, proves difficult in the context of spelling variation that both Old French and Latin have, even more so in the context of sometime heavily abbreviated manuscripts. We propose a new method based on deep learning where we attempt to categorize each line error rate into four error rate ranges ($0 < 10\% < 25\% < 50\% < 100\%$) using three different encoder (GRU with Attention, BiLSTM, TextCNN). To train these models, we propose a new dataset engineering approach using early stopped model, as an alternative to rule-based fake predictions. Our model largely outperforms the n-gram approach. We also provide an example application to qualitatively analyse our classifier, using classification on new prediction on a sample of 1,800 manuscripts ranging from the 9th century to the 15th.

Keywords

HTR, OCR Quality Evaluation, Historical languages, Spelling Variation

1. Introduction

Handwritten Text Recognition (HTR) technologies have come a long way over the last five years, to the point where data mining of medieval manuscripts and HTR-supported critical editions is getting less rare nowadays, thanks in part to the user-friendliness of interfaces such as Transkribus[1] and eScriptorium[2]. HTR, however, often shows limits in its ability to adapt to other scribes or periods, as it seems to fit specific scripts and languages. For example, Schoen and Saretto [3] has shown that a model trained over 1,330 lines of the 15th-century manuscript CCC 198[1] produces around 8.73% CER over test lines of the same manuscripts, drops to 14% on the same text in another manuscript from the same decade, and can go as low as 73.23% CER for a manuscript of a different text[2] even though it is at most 20 years "younger" and in the same language.

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

*Corresponding author.

thibault.clerice@chartes.psl.eu (T. Clérice)

"https://github.com/ponteineptique" (T. Clérice)

0000-0003-1852-9204 (T. Clérice)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Oxford, Corpus Christi College 198.

[2]Oxford, Corpus Christi College 201.

In order to evaluate the consistency of a model on an out-of-domain document such as another manuscript or a new hand, researchers usually have to create new ground-truth transcriptions to which the model predictions are compared. In this context, it seems out of reach to leverage with confidence the amount of data that remains dormant in the open data vaults of libraries such as the Bibliothèque Nationale de France (BnF) for statistical studies, making the number of 50,149 IIIF manifests catalogued by Biblissima's portal[4] promising while leaving a bitter taste of unavailability: it would require the manual transcription of at least a few hundred lines for each manuscript[3].

To address this, we can approach this issue not as an HTR one but rather as a Natural Language Processing (NLP) task, evaluating the apparent "correctness" of the acquired text rather than its direct relationship with the digital picture of the manuscript. Evaluating new transcriptions without ground truth has been done, but mainly for OCR and non-historical documents. For modern languages, where spelling is fixed and grammar stable, a dictionary approach in combination with some n-gram statistics have provided a solid framework for establishing the probability that a document is well transcribed. However, for languages such as old French or medieval Latin, both evolving over the span of few centuries, the issue is different. For example, Camps et al. [6] has catalogued 36 forms of the word *cheval* (horse) in the largest available Old French corpus. A Dictionary approach would already prove to be complex, but to make things worse, the abbreviated nature of medieval texts would require taking into account several abbreviation systems, making it unsustainable.

HTR is most often, in the humanities, not a task in itself but rather a preliminary step for corpus building (such as digital editions) or corpus analysis. In this context, HTR quality can be of primordial importance, depending on the task at hand. While Eder [7] has suggested that good classification in stylometry is still possible for corpora with noise levels as high as 20%, even for the smallest feature sets, Camps et al. [8] demonstrated that, for HTR, noise leads to accumulating errors throughout its post-processing (word segmentation, abbreviation resolution, lemmatization and POS-tagging), making the post-processed textual features less reliable than original character n-grams. For some other tasks, such as in corpus linguistics (e.g. semantic drift studies), the study of abbreviation systems such as the one performed by Honkapohja and Suomela [9] or even the training of large language models such as MacBerth[10] might require a higher level of precision.

As such, evaluating the textual quality of an automatic transcription "from afar" is extremely useful, as it provides solid grounds to either exclude documents from analysis or help guide ground-truth creation campaigns in well-funded projects. For cultural heritage institutions, it can also provide a welcome indicator for the document that could be ingested by a research engine. We can even imagine situations where these institutions transcribe only a sample of each element of their collection, and only fully and automatically transcribe the ones that reach a certain level of quality, thus saving energy and ultimately budget on the computation front.

From a human reader's perspective, Springmann et al. [11] and Holley [12] have set a limit of a CER below 10% for a good OCR quality. Recently, Cuper [13] has proposed the evaluation of

[3]Five million lines would be required for the mentioned set of manifests of the BnF with only 100 lines per manuscript. As a comparison point, the accumulated number of lines of manuscript dataset, regardless of the script or language, publicly available on the HTR-United catalog[5] is 164,418 at the end of August 2022.

OCR quality for heritage text collections, specifically Dutch newspapers from the 17th century, to distinguish good OCR from bad, using the aforementioned threshold. They provide a tool, *QuPipe*, which offers binary classification capacities, putting text in either the range $[0; 10]$% of CER or in the remaining range of "bad" OCR. In 2022 as well, Ströbel et al. [14] addressed this issue regarding HTR of cultural heritage documents, specifically from the 16th century. They provide a strong argument for using lexical features and (pseudo-)perplexity scores for HTR quality estimation, with the specific limitation that the texts they studied, 16th-century Latin correspondence, does not provide as much variation as older languages such as historical German. We also note that correspondence may be less abbreviated, and that this dataset spans a very short period.

In this paper, we address this issue as a supervised classification task, based on a dataset of around 50,000 lines of ground truth spanning from the 9th through to the 15th century. Following the conclusion of Cuper [13], we augment the number of categories we want to find: we distinguish *Good* ($[0, 10)$%), *Acceptable* ($[10, 25)$%), *Bad* ($[25, 50)$%), and *Very Bad* ($\geq$ 50%) rates of OCR. This provides a more fine-grained evaluation of the transcription and allows for guided transcription campaigns, by addressing either the low-hanging fruits (*Acceptable*) or the rotten ones. We evaluate three kinds of basic architectures (GRU with attention, BiLSTM and TextCNN) on line classification using real-life "bad" transcriptions and precomputed CER scores.

The resulting models have shown promising results, with quality levels such as *Very Bad* and *Good* being well recognized. In order to evaluate the models and showcase their usefulness, we also provide an example of a real-life classification application, where 1800 manuscripts were randomly selected from the BnF and classified by our best model.

In summary, the contributions of this paper are:

1. a new approach for HTR evaluation of historical languages with variable spellings;
2. a new method to produce ground truth for OCR evaluation that does not rely on artificially and manually tuned generation;
3. an initial evaluation of the output and a quick glance at the state of HTR for Old French and Medieval Latin over six centuries.

The remainder of this paper is organised as follows. We start by addressing the background in Section 2, specifically regarding the specifics of Old French and Medieval Latin and the idea of readability. In Section 3, we describe the HTR datasets we used and their particularities. In Section 4, we describe the architecture of the models, their feature engineering and the process behind the generation of bad predictions. In Section 5, we describe the set-up of our model selection and evaluation. Finally, in Section 6, we analyse the result both on the dataset produced *ad hoc* (described in Sections 3 and 4), but also on completely unseen documents from the BnF, to showcase the capacities of such models.

2. Background and Related Work

Handwritten Text Recognition, a sibling or sub-task of Optical Character Recognition, aims at recognising text from digitised manuscripts. In the last five years, the digital humanities

landscape has seen a surge in HTR engines, as well as transcription interfaces that connect and work well with these engines, from the dominant Transkribus[1] to the open-source pair of eScriptorium[2] and Kraken[15]. To be able to recognize text, users have to provide models, which are themselves the result of supervised training on ground truth data (human provided transcriptions).

Printed books have been, over the last few decades, the focus in terms of remediation, from their analogue form to a digitized picture and finally to a machine-readable (and human searchable) text. With the advances in HTR over the last five years, the focus can now shift or be shared with materials that have, for the most part, remained inaccessible from a digital point of view, except for pictures. Latin manuscripts are present during the whole period of manuscript production in western Europe. Literary Old French manuscripts exist from the 12^{th} century onward, with only a hundred known surviving manuscripts in the 12^{th} century[16]. Over the span of these seven centuries, multiple forms of handwritten scripts have existed, for both French and Latin. As an example, the 2016 ICHFR *Competition on the Classification of Medieval Handwritings in Latin Script*[17] provided ground truth for the classification of 12 main families, of which at least six are represented in our datasets. This diversity makes training models for HTR quite complex but also a reachable goal, as they tend, specifically for literary manuscripts, to be more readable and stable between different hands.

Medieval French and Latin present both dialectal and scriptural variation in synchrony on top of diachronic evolution. Old French's syntax varies chronologically and geographically. The spelling is simply variable. While Latin shows some level of variation, it differs from Old French mostly in its higher rate of abbreviation. These observations are limited to the context of the datasets at hand, which are literary works (including scholastic, theological and medical works). The Old French CREMMA Medieval dataset[18] has 0.97% of horizontal tildes and 0.16% of vertical ones, which are markers used in the dataset guidelines to indicate various similar abbreviation diacritics[19]. Using the same guidelines, the CREMMA Medieval Latin dataset shows a rate of 5.63% and 1.52% for the same characters. This difference could be due to the nature of the transcribed texts.

The question of abbreviation and the specificity of medieval literary manuscripts has provoked many discussions in terms of how to transcribe documents, from a completely "diplomatic" approach with variants of letters to "semi-diplomatic" approaches. In the last year, three authors have provided guidance or thoughts around guidelines for transcriptions: Pinche [19] focusing on Old French, Schoen and Saretto [3] on Middle English, and Guéville and Wrisley [20] on Latin. The CREMMA guidelines have been used by 5 other datasets for a total of 1.15 millions of characters over fifty manuscripts, which make them the most diverse and comprehensive ones for HTR of medieval manuscripts in Latin and Old French.

The most traditional metrics for HTR and OCR are both Word Error Rate (WER) and Character Error Rate (CER). The first one proves to be complicated to apply in Old French and Medieval Latin, as spaces in medieval manuscripts tend to vary in size or simply be nonexistent from a modern perspective, relying on the knowledge of the reader to separate words—or the ability of NLP models to separate them[21]. The second one works well, with the limitation that spaces are often the first source of mistakes. CER corresponds to the sum of character insertion, removal and replacement over the total number of characters, thus providing a fine-grained metric.

As mentioned earlier, in the introduction, both CER and WER require ground truth, and other metrics currently discussed as alternatives, such as the (pseudo-)perplexity or lexical measures proposed by Ströbel et al. [14]. The other approach to evaluating quality without ground truth is to predict a class of CER, such as the work done by Bazzo et al. [22]. These approaches rely on features such as n-grams, word statistics and language classifier outputs which are difficult to leverage in the present context. In order to train their classifier, Bazzo et al. [22] and Nguyen et al. [23] engineered bad predictions by creating rules to reproduce the most common errors in OCR, such as "rn" becoming "m". These bad predictions are then fed to their model along with the metrics both papers want to predict.

Nguyen et al. [23] provide an innovative approach to the issue of noise in OCR by shifting from a CER/WER problem to a readability one: if the reader "can reod a txt with miffpelling" without having to refer back to the picture, at least one of the goals of OCR has been achieved. As simply put by Martinc et al. [24], "Readability is concerned with the relation between a given text and the cognitive load of a reader to comprehend it". It is even more important in the context of handwritten documents where a somewhat badly but readable HTR output can be easier for non-specialists to read than the original. In the field of readability assessment, Martinc et al. [24] has shown that supervised models perform adequately, while Nguyen et al. [23] has shown that this translates to the OCR issues as well. This has not been applied to any medieval dataset that we know of.

3. Dataset

To train different models, we reused the data from various projects, aligned with the same guidelines used by Pinche [19]. Our experiment was made possible by the open release of many projects' datasets, including one MA thesis and one student project[25, 26]. We used the ground truth of the CREMMA[27] and CREMMALab[18] projects, the Rescribe[28] project, and the GalliCorpora[29, 30] projects, for a total of 42,292 lines (see Table 1). We include one dataset of incunabula, which use graphical shapes similarly to literary manuscripts (but with more regularity), while also using an abbreviation system.

Dataset name	Project or company	Coverage	Language	Lines	Characters	Manuscripts
Eutyches	*MA Thesis*	850-900	Latin	2,828	86,832	2
Caroline Minuscule	Rescribe	800-1199	Latin	457	17,155	17
CREMMA Medieval	CREMMALab	1100-1499	French	21,656	579,368	14
CREMMA-Medieval-LAT	CREMMA	1100-1599	Latin	6,648	240,291	18
DecameronFR	*Homework*	1430-1455	French	751	19,821	1
Données HTR manuscrits du 15e siècle	GalliCorpora	1400-1499	French	5,937	169,221	11
Incunables du 15e siècle	GalliCorpora	1400-1499	French	7,608	244,958	13

Table 1
Training material for our models and our future bad transcription dataset.

The datasets present not only two main languages but also many different levels of digitization quality (including old binarization), different kinds of handwriting families, different abbreviation levels and different genres. For example, while the CREMMA Medieval dataset focuses more on literary texts, specifically hagiographical and *chanson de geste* texts, the CREMMA Medieval LAT corpus offers theological commentaries and medicinal recipes, each

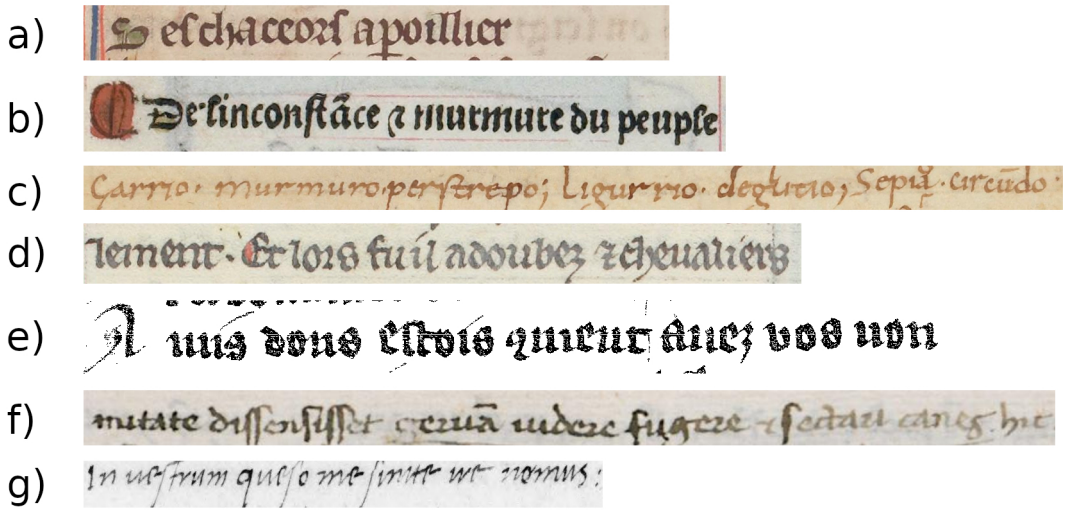


Figure 1: Example of lines. (a) comes from the GalliCorpora manuscript dataset, (b) from the *incunabula* one. (c) is drawn from the Eutyches MA Thesis, (d) and (e) from the CREMMA Medieval (French) dataset. (f) and (g) are both taken from the CREMMA Medieval Latin repository.

genre having its own specific vocabulary. The dataset in general is skewed towards French and the *gothica* handwritten family.

The transcription guidelines of Pinche [19] provide simplification rules: allographic approaches are forbidden (different shapes of s such as long s and "modern" s are not differentiated), macrons and general horizontal-line diacritics over the letters such as tildes are represented by horizontal tildes, any "zigzag"[4] or similarly shaped forms are simplified into superscript vertical tildes, etc. This allows for simpler transcriptions and also limited diversity of characters for the machine to learn, satisfying both the human transcriber in terms of the learning curve of the guidelines, and the HTR engine in terms of complexity. Each corpus was passed through the ChocoMufin software [31] using project-specific character translation tables. This software, along with these tables, allows each dataset to be controlled at the character level and adapted to guideline modifications. It also allows project-specific transcription standards to be translated to a more common one, such as Pinche's.

4. Proposed Method

Our goal is to be able to predict a quality class for any HTR output on medieval French and Latin. First, we design a way to generate ground truth for the quality assessment of HTR output. Then, we propose three supervised text-based models, with specific adaptations to handle both languages with a single classifier.

[4]Official name from the Unicode specifications for the character U+299A.

4.1. "Bad Prediction" Ground Truth

In order to train our classification model, we require ground truth material along a CER class: *Good* ($[0; 10)\%$), *Acceptable* ($[10; 25)\%$), *Bad* ($[25; 50)\%$) and *Very Bad* ($\geq 50\%$). In order to have real life errors, and to reproduce the rather difficult to predict capacity of a model to confuse certain characters with others in specific settings, we propose a three-step method:

1. We train Kraken[15] models based on the complete dataset, or on a subset. We voluntarily stop some of the training in very early stages, when the CER on the validation dataset remains high. We also keep one "best" model[32] trained on the full dataset.
2. We run each model on our two biggest and most diverse repositories, *CREMMA Medieval* and *CREMMA Medieval LAT*. We also run a model trained on modern and contemporary scripts, Manu McFrench[33] to create garbage-level transcriptions.
3. We evaluate each line's CER and store it alongside the line. We also keep the ground truth, whose CER is estimated as 0. We remove short lines (fewer than 15 characters) and duplicated predictions across models for the same line.

Regarding the final models for prediction production, we have 16 models, allowing for a maximum of 16 versions of each line, if none of the models predict the same text (see Table 2 for examples):

1. 4 models trained on the same train and validation dataset as *best* with a validation CER of 55.9, 28.3, 23 and 20.8% according to Kraken.
2. 5 models trained on the *CREMMA Medieval LAT* dataset only, from the 1^{st} to the 6^{th} epochs, ranging from 86% to 46% of CER.
3. 1 model trained on the *Eutyches* (Latin, Carolingian of the 9^{th} century) and the *Decameron* (French, 16^{th} century) datasets with a 98.5% CER on its validation set.
4. 3 models trained on the *CREMMA Medieval* (Old French) dataset only, fine-tuned from the *Manu Mc French* Model, from 11% of CER down to 8.2%.
5. *Manu McFrench*, the *best* model and the ground-truth data.

These provide variable CER on unseen data from the test set of both CREMMA dataset but also on training and validation sets as they did not reach their full capacities during the training phase. After filtering small and repeated predictions, we have access to 322,903 lines of "HTR Predictions, CER" couples (see in appendix Figure 6). We then translate that into each bin of CER to produce the four established classes.

4.2. Model Architecture

We applied three model architectures, common to many NLP task, with an embedding-sentence encoder-linear classifier structure where only the sentence encoder changes from one model to another (see Figure 2). The embedding layer takes into account special tokens (Padding, Unknown char, Start of Line, End of Line) and each character according to the Unicode NFD[5] normalization of the line, for which characters and their diacritics are separated, e.g. [é]

[5]Normalization Form Canonical Decomposition.

transcription	CER
ůra on de q̃l vertu ses petis pies sont que vous	0.0
Bra on de q̃l vertuses petis pies sont que vous	6.1
Fra on de q̃l vertuses petis pies sont que vou	8.2
Bra on de q̃l vertuses petis pies sont que uous	8.2
Pra on de ql vertuses petis pies sont que dons	12.2
ura on de q̃l vertu ses petis pies font grre op	16.3
ura on de q̃l uertu ses petis pies font re dory	16.3
ura on de ql vertu ses petis pies font itce ir	20.4
Ard ondegl ratules nus mes sont que ls	42.9
a on de at etn le peos pes os e	49.0
a om de ał vrtir sot oliš pa sosisinos	57.1
ꝛs cm dec uł vrtr fe pdp̃ pns ots pte	61.2

Table 2
Example of pairs of predictions for the same line for a file of CREMMA Medieval (University of Pennsylvania 660, *Le Pélerinage de Mademoiselle Sapience*). The first line is the ground truth, the second our best model trained on the full dataset for production, the 4th from the bottom is from Manu McFrench. Note that the diacritics are not consistently transcribed.

becomes `[e]+[´]`. The linear layer is a simple (Encoding Output Dimension, Class Count) decision layer. Each model uses a cross-entropy loss function[6] and reduces its learning rate at plateau using the validation set's macro averaged recall metric. Optimization of the model is done through the Ranger optimizer[34].

The encoding layer varies between three different forms:

- The first version uses a single BiLSTM network where the sentence encoding is the result of the concatenation of the start-of-line token (BOS) and end-of-line token (EOS) hidden state.
- The second version follows the architecture of sentence-level attention proposed by Yang et al. [35], using a bidirectional GRU. The encoded sentence vector is the sum of products of the hidden state of each token with its attention. Attention is also provided as an output for human interpretation of the results.
- The last one, TextCNN[36], uses the concatenation of the Max Pooling of each n-gram size (2, 3, 4, 5, 6) taken into account by a convolutional neural network.

As we deal with two different languages, we added another special token, following the work of Martin et al. [37] and Gong et al. [38]: for each encoding variation we add one variation of the codec where the first token after the beginning-of-string is a metadata token indicating the language. Thus, a line such as `Fra on de q̃l vertuses petis pies sont que vo` will be encoded as `<BOS><FRO>Fra on de q̃l vertuses petis pies sont que vo<EOS>`.

[6]Code available at https://github.com/PonteIneptique/neNequitia.

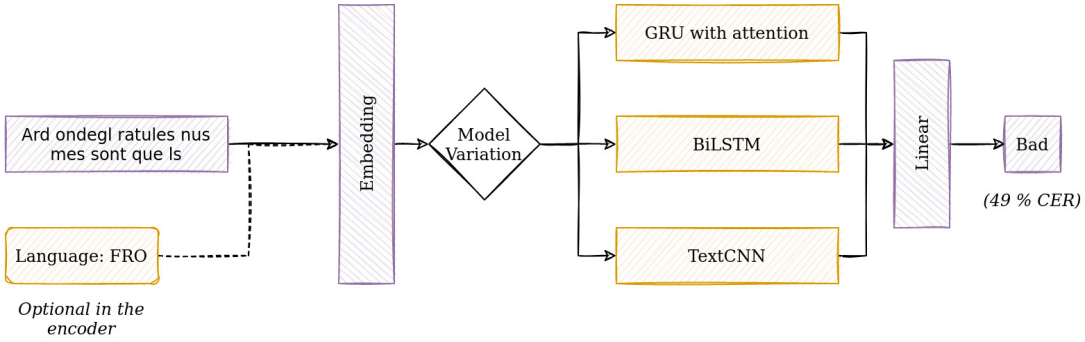


Figure 2: Available model architectures. Elements in orange are optional or varying elements, elements in blue are common to all models.

K		1	2	3	4	5
Validation	French	BnF fr. 17229, BnF fr. 25516	BnF fr. 3516, BnF fr. 25516	BnF fr. 24428, BnF. Arsenal 3516	BnF fr. 24428, BnF fr.844	Pennsylvania Codex 909, BnF fr.844
	Latin	Arras 861, CCCC Ms 165	CLM 13027, CCCC 165	CLM 13027, Montpellier, H318	BnF lat. 6395, Montpellier, H318	BnF lat. 6395, Laur. Plut.33.31
Test	French	BnF Arsenal 3516, BnF fr. 13496	BnF fr. 24428, BnF fr. 411	BnF fr. 844, BnF fr. 22549	BnF fr.412, Phil., Col. of Phys. 10a 13	Bodmer 168, Vat. reg. lat. 1616
	Latin	Sorbonne Fr. 193, CLM 13027	CCCC Ms. 236, H318	BnF lat. 6395, Egerton 821	BnF fr. 16195, Laur. Plut. 33.31	Laur. Plut. 53.08, BnF lat. 8236
Train	Good	80,056	76,564	65,764	39,165	39,165
	Acceptable	44,346	41,769	34,429	35,803	35,803
	Bad	60,381	59,265	51,637	41,793	41,793
	Very Bad	71,008	71,053	60,898	52,212	52,212
Validation	Good	4,246	98,57	12,770	11,625	11,625
	Acceptable	3,933	10,377	12,496	8,492	8,492
	Bad	4,338	10,884	13,430	10,250	10,250
	Very Bad	4,867	15,428	18,386	11,461	11,461
Test	Good	9,165	7,046	14,933	42,677	42,677
	Acceptable	9,744	5,877	11,098	13,728	13,278
	Bad	12,763	7,333	12,415	25,439	25,439
	Very Bad	18,056	7,350	14,647	30,258	30,258

Table 3
Composition of K-Folds set, based on manuscript selection.

5. Experimental Setup

In order to avoid lexical bias and to ensure the strength of our analysis, we propose a 5-Fold-like experiment, where each subset for train, validation and test are the results of split across manuscripts. For each K, two French manuscripts and two Latin ones are used for the validation set and the test set, and they differ by at least one manuscript from one K to another, leaving three K completely different (K1, K3, K5; see Table 3). Each test set also contains a Latin manuscript that was not used in any of the HTR model training or validation: *Berlin, Hdschr. 25.* This manuscript was then used for model evaluation, to have a stable pillar for evaluation. Models are then evaluated using class-specific precision and recall, as well as macro averaged precision and recall.

For our baseline, we use the relative frequency of the 2000 most common n-grams of size 3, 4 and 5 as features and feed them to a linear classifier, with cross entropy loss and the Adam optimizer. We run each model architecture once for each K, resulting in 7 different results with the baseline (presence/absence of language token for the three encoding modules + baseline).

Our whole pipeline uses pandas for data preparation[39], PyTorch[40] for model development, and Pytorch Lightning[41] for the training, evaluation and prediction wrapping.

Lang	Encoder	Good				Acceptable				Bad				Very bad			
		Precision		Recall		Precision		Recall		Precision		Recall		Precision		Recall	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
No	*Baseline*	33.87	35.24	33.84	33.76	36.81	37.63	7.56	8.67	37.34	37.11	19.15	18.27	60.27	59.73	97.24	97.32
Yes	Attention	65.31	65.61	**41.62**	**41.88**	**45.29**	44.01	26.72	26.32	49.36	49.66	49.23	47.70	75.74	75.33	95.53	95.48
Yes	BiLSTM	**67.00**	**66.82**	38.02	37.31	41.75	41.29	21.89	21.05	47.13	47.77	51.79	51.09	76.17	74.18	94.20	94.51
Yes	TextCNN	57.78	59.06	31.52	26.14	41.97	43.24	20.87	22.29	43.66	44.88	35.93	33.26	68.09	65.95	**96.73**	**97.61**
No	Attention	58.08	57.00	39.85	41.62	44.10	**44.40**	**35.60**	**34.21**	**51.98**	51.34	49.98	47.16	**80.01**	78.53	94.41	94.51
No	BiLSTM	60.30	57.60	39.70	36.55	42.87	42.90	31.15	28.17	50.95	**51.37**	**52.39**	**52.63**	79.96	**80.40**	94.19	93.80
No	TextCNN	50.85	49.35	38.43	38.32	40.24	40.10	24.77	26.63	48.59	49.14	47.94	47.16	76.92	77.48	94.46	94.44

Table 4
Test results statistics for each K and each model configuration.

6. Experiments

6.1. Model Classification Results

The first conclusion we can draw from the experience is that our models always beat the baseline (see Table 4 and, in the Appendix, Table 4 for more details). No RNN-based architecture clearly beats the other, but TextCNN clearly underperforms. The introduction of the language metadata token helps when detecting *Good* transcriptions (delta $\approx$ +7% for attention's median precision, $\leq$ +1% for the recall) for both RNN based models. Models without a language marker tend to outperform models with language markers, except for the *Very Bad* class where the delta is up to +6% in favour of models without language tokens (using median precision scores).

Regarding the variability of results, we found that the length of the string had an impact on the prediction, no matter the model architecture. Surprisingly, none of the models withstand long noisy lines: the accuracy of the *Very Bad* class is inversely correlated with line size. On the contrary, depending on the encoder, some classes benefit from longer strings: *Good* lines benefit from it with all models except the baseline. TextCNN is the only model to really correlate accuracy on the *Bad* and *Acceptable* classes with line length.

Finally, for all models except the baseline, the most common confusion is always in the "adjacent" class(es) (see Figure 4). For the classes *Acceptable* and *Bad*, which have two neighbours, the error rate is evenly split between them: the class *Acceptable* tends to be confused with either *Good* or *Bad*. This shows the model's ability to understand cleanness or noise, but also shows the limit of these classes: for a line with 50 characters, such as "quãt tel e**u**fa**u**t gist en te**s** lieu. **D**erite respo**i**nt", 6 mistakes are enough to swing into the *Acceptable* category (Ground truth: "quãt tel enfant gist en tel lieu . Uerite respon", one space has been removed before the dot).

Overall, with an accuracy for the *Good* and *Very Bad* classes around 50% on these languages, and considering that most of the confusions are from adjacent classes (*e.g.* *Good* is confused with *Acceptable*, *Acceptable* with *Good* and *Bad*, etc.), the solution performs well either at filtering badly read manuscripts, or keeping only the very good ones. The *Acceptable* class and the *Bad* class have stable performance facing variable line length, although the *Acceptable* class shows the worst classification performance.

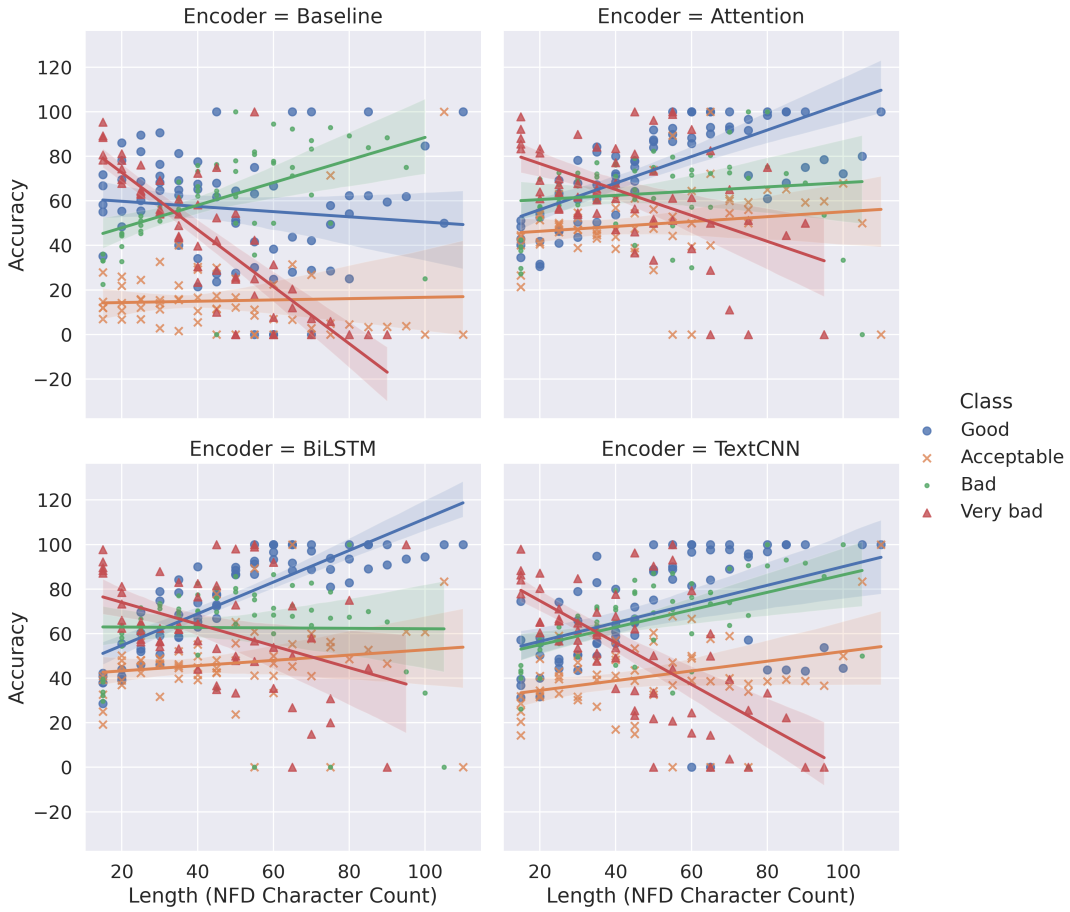


Figure 3: Regression of accuracy based on lines' length overall 5-Fold test sets. Common manuscript not included (*Berlin, Hdschr. 25*).

6.2. Application on a Real-World Library Dataset

As a real-world application, we wanted to apply one of our best models to an unseen dataset, in the same way that we envision cultural institutions might use the tool. We describe the set-up for this particular experiment below, and then evaluate the results of the classification model with regard to the capacity of the HTR model; we also study some randomly sampled elements.

6.2.1. Set-up

To evaluate on as much unseen data as possible, we crawled the Biblissima IIIF collection portal[4]. We searched individually for each combination of language (French, Latin) and century (9^{th} to 15^{th}), limiting the number of samples retrieved to 500 manuscripts. We then sam-

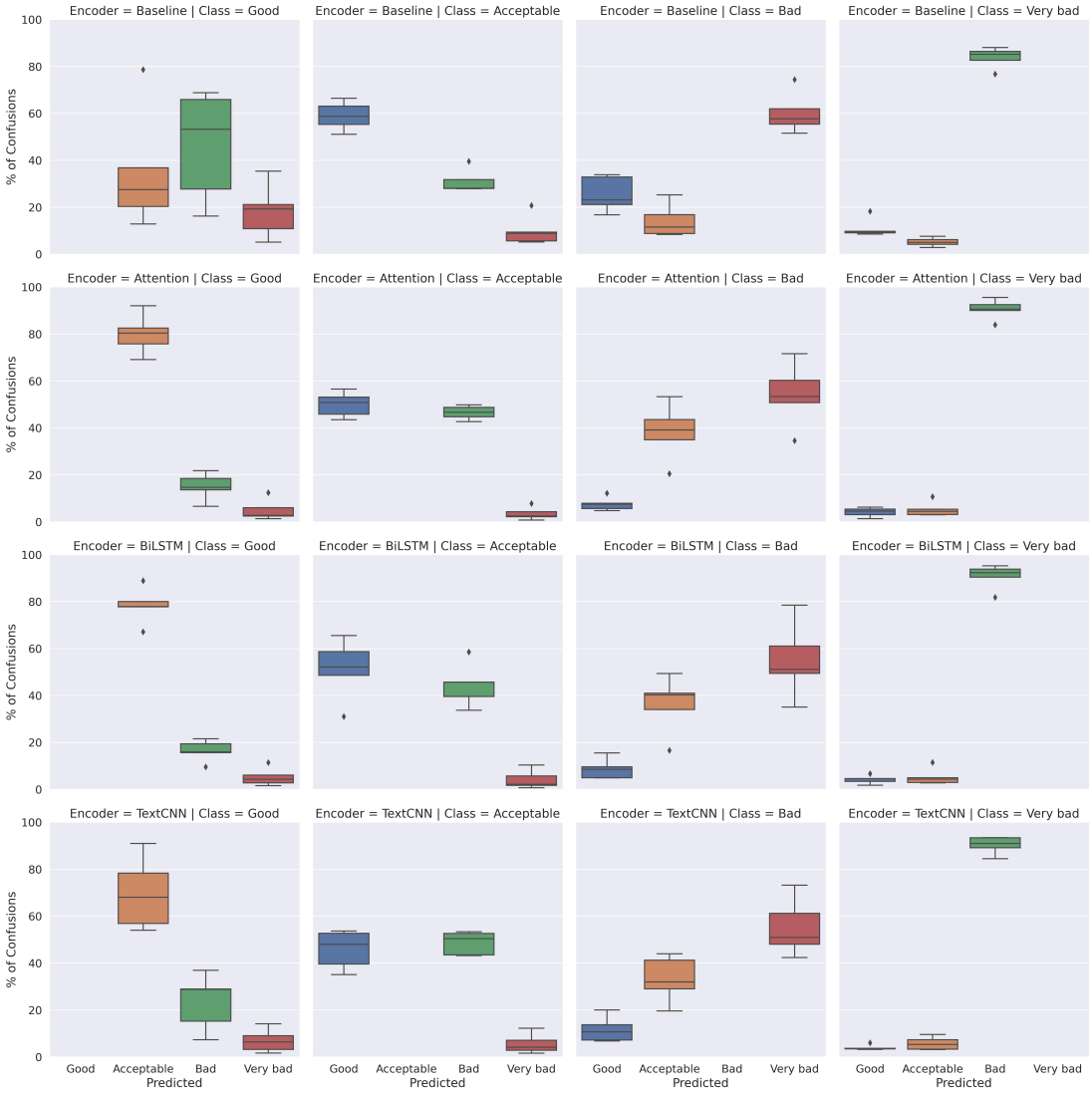


Figure 4: Confusion rate dispersion in the errors made by each model. Only confusion that happens more than 50 times is taken into account, as well as the total number of errors greater or equal to 300. The graph can be read as follows: for the baseline, 40% of the errors for the ground truth class *Good* are *Acceptable* predictions.

pled 10 sequential pictures from each manuscript.[7] To avoid empty pages (which tend to be at the start and the back of each book's digitization or IIIF manifest at the BnF), we take either the ten first pictures from the second decile of the manifest, or from the 20^{th} up to the 30^{th} if there are fewer than 100 pictures, or the 10 last if there are fewer than 20 pictures.

[7]Note that we are not talking about pages but about pictures: in some cases, most commonly in the case of digitised microfilms, one picture can contain two pages.

Each downloaded sample is then segmented using YALTAi[42] with the included model designed for cultural heritage manuscripts and the base Kraken BLLA segmenter[43]. As YALTAi provides different zones—from the margin to main body of text—through numbering, we only consider lines that are part of the main bodies of text of each model, thus excluding any marginal or paratext. We then use Kraken to predict a transcription for each line, using the best trained model as described in our first experiment. Next, we feed each line to our best BiLSTM model (K-Fold 1 has the best recall/precision on *Good*) while keeping the line metadata: language, century, manuscript identifier, and page identifier.

Finally, we provide three different evaluations of the transcriptions. The first is based strictly on the number of lines predicted in each class (*Good*, *Acceptable*, etc.). The second is page-based: we take the most common prediction for all lines. The last one is manuscript based: we take the most common page prediction, using the previous page-based metric.

6.2.2. Evaluation

Overall, the HTR prediction results produced by our BiLSTM module are in line with the HTR strength on the dataset (see Figure 5). The model performs extremely well on early manuscripts thanks to the presence of two datasets of early manuscripts (*Eutyches* and *Caroline Minuscule*) It performs well on Old French except for the 13TH century, where *Bad* predictions are more common. The relative frequency of *Very Bad* predictions tends to grow as we get closer to the 16TH century: from the data we have seen, this could be due to the presence of non-literary manuscripts written in cursive, for which our model has no ground truth.

If we look at the sampled predictions (Appendix, Table 2), most *Good* predictions seem correct or nearly correct. However, we can see that the metadata from Biblissima and the BnF has some limitations when used automatically, as it can produce problematic results: most 12th century *Acceptable* predictions are probably in Latin, which would indicate a multilingual manuscript or a badly catalogued one. This issue also arises in the crawler for the century, as some manuscripts were catalogued as French but with a production date that is before the first known French document: these are most likely multilingual documents, with either a collection of various leaves from previous manuscripts, or the inclusion of the language used for marginal notes. 3 out of the 6 *Acceptable* predictions between the 13th and the 14th century are definitely readable and understandable, and we cannot but wonder if the lack of spaces in "q̃ merueilles fu lacitebiengarne mlt" is responsible for its classification as *Acceptable* rather than *Good*. We note that at least one *Very Bad* prediction in French, "OU EtE L. Cheualier de Monifort, son Oncle, Gles", seems rather readable, albeit with more corrections than for a *Good* transcription. Latin shows the same trend, in being accurate over *Good* and *Acceptable*.

7. Conclusion

The ability to filter, without pre-transcribing samples, automated transcriptions of manuscripts in Latin, Old French or any other Western historical language, might lead to the production of datasets designed for analysis that relies on better transcriptions, or to guiding cultural heritage institutions and their partners in the production of new ground truth. Producing HTR ground truth does indeed require time, skilled transcribers and, last but not least, budget. However,



Figure 5: Predictions distribution per line (first two rows), per page (row 3 & 4), per manuscript (last rows) over languages and centuries filtering.

most current error rate prediction or HTR output analysis models rely on n-gram frequencies and lexical features—two approaches that are often less viable for languages such as Old French which "suffers" from a highly variable spelling system or for languages like Latin which are potentially highly abbreviated, with abbreviations changing even within a single manuscript, depending on the context, the topic and the scribe.

In this context, we chose to treat CER range predictions as a sentence-like classification problem, for which we implemented three basic models, using either a single BiLSTM encoder, an attention-supported GRU, or a TextCNN encoder. These three tools show stronger results than an n-gram based baseline. On top of this, we include a language metadata token which can improve the reliability of the lowest range of CER (between 0 and 10%, the *Good* class) while worsening the classification's reliability for the highest range (over 50%, the *Bad* class). For the purpose of training these models, we propose a new way to generate real life "bad

transcriptions", using early-stopped HTR models, or models trained on small samples of data: this provides an alternative to previous rule-based generation of "bad transcription" ground truths.

We show that on a completely unknown dataset of around 1,800 manuscripts, analysed with a new HTR model specifically trained on medieval Latin and French, the number of well-transcribed manuscripts predicted is on par with the ground truth for that dataset. The quality assessment predictions provide quick insights for larger collections, and could be run relatively often by cultural heritage institutions.

In the future, hyper-parameter fine-tuning and other encoders could be used in the architecture. Specifically, with more correctly transcribed manuscripts, including the abbreviations in their transcriptions, fine-tuning larger language models could allow the application of (pseudo-)perplexity ranking such as the one proposed by Ströbel et al. [14], while allowing for partial noise in the training data. We hope to see such classification of manuscripts used by ground truth producers in order to enhance the robustness of openly available HTR models.

Acknowledgments

I want to thanks Jean-Baptiste Camps, Ariane Pinche and Malamatenia Vlachou-Efstathiou for their constant feedback and replies on some particular questions regarding manuscripts or HTR data. Thank you to Ben Nagy for his proof-reading of the pre-print version.

This work was funded by the Centre Jean Mabillon and the DIM MAP (https://www.dim-map.fr/projets-soutenus/cremmalab/).

References

[1] P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, Transkribus-a service platform for transcription, recognition and retrieval of historical documents, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 4, IEEE, 2017, pp. 19–24.

[2] B. Kiessling, R. Tissot, P. Stokes, D. S. B. Ezra, escriptorium: an open source platform for historical document analysis, in: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, IEEE, 2019, pp. 19–19.

[3] J. Schoen, G. E. Saretto, Optical character recognition (OCR) and medieval manuscripts: Reconsidering transcriptions in the digital age, Digital Philology: A Journal of Medieval Cultures 11 (2022) 174–206. URL: https://muse.jhu.edu/article/853521. doi:10.1353/dph.2022.0010.

[4] E. Frunzeanu, E. MacDonald, R. Robineau, Biblissima's choices of tools and methodology for interoperability purposes, CIAN. Revista de historia de las universidades 19 (2016) 115–132.

[5] A. Chagué, T. Clérice, HTR-United: Ground Truth Resources for the HTR and OCR of patrimonial documents, 2022. URL: https://htr-united.github.io.

[6] J.-B. Camps, T. Clérice, F. Duval, L. Ing, N. Kanaoka, A. Pinche, Corpus and models for

lemmatisation and POS-tagging of old french, 2022. URL: https://halshs.archives-ouvertes.fr/halshs-03353125.

[7] M. Eder, Mind your corpus: systematic errors in authorship attribution, Literary and Linguistic Computing 28 (2013) 603–614. URL: https://doi.org/10.1093/llc/fqt039. doi:10.1093/llc/fqt039.

[8] J.-B. Camps, C. Vidal-Gorène, M. Vernet, Handling heavily abbreviated manuscripts: HTR engines vs text normalisation approaches, 2021. URL: https://hal-enc.archives-ouvertes.fr/hal-03279602.

[9] A. Honkapohja, J. Suomela, Lexical and function words or language and text type? Abbreviation consistency in an aligned corpus of Latin and Middle English plague tracts, Digital Scholarship in the Humanities 37 (2021) 765–787. URL: https://doi.org/10.1093/llc/fqab007. doi:10.1093/llc/fqab007.

[10] E. Manjavacas, L. Fonteyn, Adapting vs. Pre-training Language Models for Historical Languages, Journal of Data Mining and Digital Humanities NLP4DH (2022). URL: https://hal.inria.fr/hal-03592137. doi:10.46298/jdmdh.9152.

[11] U. Springmann, F. Fink, K. U. Schulz, Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings, 2016. URL: http://arxiv.org/abs/1606.05157. doi:10.48550/arXiv.1606.05157.

[12] R. Holley, How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs, D-Lib Magazine 15 (2009).

[13] M. Cuper, Examining a multi layered approach for classification of ocr quality without ground truth, DH Benelux Journal (2022) 17.

[14] P. B. Ströbel, S. Clematide, M. Volk, R. Schwitter, T. Hodel, D. Schoch, Evaluation of htr models without ground truth material, arXiv preprint arXiv:2201.06170 (2022). URL: http://arxiv.org/abs/2201.06170.

[15] B. Kiessling, The Kraken OCR system, 2022. URL: https://kraken.re.

[16] M. Careri, C. Ruby, I. Short, Livres et écritures en français et en occitan au XIIe siècle: catalogue illustré, Viella, 2011.

[17] F. Cloppet, V. Eglin, V. C. Kieu, D. Stutzmann, N. Vincent, ICFHR2016 competition on the classification of medieval handwritings in latin script, in: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2016, pp. 590–595. URL: http://ieeexplore.ieee.org/document/7814129/. doi:10.1109/ICFHR.2016.0113.

[18] A. Pinche, Cremma medieval, 2022. URL: https://github.com/HTR-United/cremma-medieval. doi:10.5281/zenodo.5235185.

[19] A. Pinche, Guide de transcription pour les manuscrits du Xe au XVe siècle, 2022. URL: https://hal.archives-ouvertes.fr/hal-03697382.

[20] E. Guéville, D. J. Wrisley, Transcribing medieval manuscripts for machine learning, arXiv preprint arXiv:2207.07726 (2022). URL: https://arxiv.org/abs/2207.07726.

[21] T. Clérice, Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin, Journal of Data Mining & Digital Humanities 2020 (2020). URL: https://jdmdh.episciences.org/6264. doi:10.46298/jdmdh.5581.

[22] G. T. Bazzo, G. A. Lorentz, D. Suarez Vargas, V. P. Moreira, Assessing the impact of OCR errors in information retrieval, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Lecture Notes

in Computer Science, Springer International Publishing, 2020, pp. 102–109. doi:10.1007/978-3-030-45442-5_13.

[23] H. T. T. Nguyen, A. Jatowt, M. Coustaty, A. Doucet, Readocr: A novel dataset and readability assessment of ocred texts, in: International Workshop on Document Analysis Systems, Springer, 2022, pp. 479–491.

[24] M. Martinc, S. Pollak, M. Robnik-Šikonja, Supervised and unsupervised neural approaches to text readability, Computational Linguistics 47 (2021) 141–179. URL: https://doi.org/10.1162/coli_a_00398. doi:10.1162/coli_a_00398.

[25] M. Vlachou-Efstathiou, Voss.lat.o.41 - eutyches "de uerbo" glossed, 2022. URL: https://github.com/malamatenia/Eutyches.

[26] S. Biay, V. Boby, K. Konstantinova, Z. Cappe, Tnah-2021-decameronfr, 2022. URL: https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-DecameronFR. doi:10.5281/zenodo.6126376.

[27] T. Clérice, M. Vlachou Efstathiou, A. Chagué, Cremma manuscrits médiévaux latins, 2022. URL: https://github.com/HTR-United/CREMMA-Medieval-LAT.

[28] N. White, A. Karaisl, T. Clérice, Caroline minuscule by rescribe, 2022. URL: https://github.com/rescribe/carolineminuscule-groundtruth.

[29] A. Pinche, S. Gabay, N. Leroy, K. Christensen, Données HTR incunables du 15e siècle, 2022. URL: https://github.com/Gallicorpora/HTR-incunable-15e-siecle.

[30] A. Pinche, S. Gabay, N. Leroy, K. Christensen, Données HTR manuscrits du 15e siècle, 2022. URL: https://github.com/Gallicorpora/HTR-MSS-15e-Siecle.

[31] T. Clérice, A. Pinche, Choco-Mufin, a tool for controlling characters used in OCR and HTR projects, 2021. URL: https://github.com/PonteIneptique/choco-mufin. doi:10.5281/zenodo.5356154.

[32] T. Clérice, A. Pinche, M. Vlachou-Efstathiou, Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th century, 2022. URL: https://doi.org/10.5281/zenodo.7234166. doi:10.5281/zenodo.7234166.

[33] A. Chagué, T. Clérice, HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French), 2022. URL: https://doi.org/10.5281/zenodo.6657809. doi:10.5281/zenodo.6657809.

[34] L. Wright, New deep learning optimizer, ranger: Synergistic combination of RAdam + LookAhead for the best of..., 2019. URL: https://medium.com/@lessw/new-deep-learning-optimizer-ranger-synergistic-combination-of-radam-lookahead-for-the-best-of-2dc83f7

[35] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 1480–1489. URL: http://aclweb.org/anthology/N16-1174. doi:10.18653/v1/N16-1174.

[36] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181. doi:10.3115/v1/D14-1181.

[37] L. Martin, É. Villemonte de La Clergerie, B. Sagot, A. Bordes, Controllable Sentence Simplification, in: LREC 2020 - 12th Language Resources and Evaluation Conference, Mar-

seille, France, 2020. URL: https://hal.inria.fr/hal-02678214, due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http://www.lrec-conf.org/proceedings/lrec2020/index.html.

[38] H. Gong, S. Bhat, P. Viswanath, Enriching word embeddings with temporal and spatial information, in: Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2020, pp. 1–11. URL: https://aclanthology.org/2020.conll-1.1. doi:10.18653/v1/2020.conll-1.1.

[39] W. McKinney, et al., pandas: a foundational python library for data analysis and statistics, Python for high performance and scientific computing 14 (2011) 1–9.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[41] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019. URL: https://github.com/Lightning-AI/lightning. doi:10.5281/zenodo.3828935.

[42] T. Clérice, You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine, 2022. URL: https://hal-enc.archives-ouvertes.fr/hal-03723208, working paper or preprint.

[43] B. Kiessling, A modular region and text line layout analysis system, in: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2020, pp. 313–318.

A. Appendix

The software has been archived at the following address: https://doi.org/10.5281/zenodo.7233984. A good chunk of the data is available here: https://github.com/PonteIneptique/neNequitia/releases/tag/chr2022-release.

Manuscripts metadata and the predictions in XML ALTO formats for Section 6 are available at https://doi.org/10.5281/zenodo.7234399. The same repository contains also the XML data for training the classifier.

Lang	Century	Prediction	Transcription
fro	12	Good	ra monstre de couf de uoudenay
fro	12	Good	uucuns pair er que Iehuz de le chaulre le ieuue uor de teuteur dicelui office oy a este priuoz et de loucez pour msen et acaise de cereus cu ⅂ decpa
fro	12	Good	seriant estoit exilliez en laueniance de sacolpe. li poures
fro	13	Good	les cõdurroit car il sauoit crop bien. coz lespas. ⅂
fro	13	Good	Se il lẽ set dire nouele
fro	13	Good	tiseras. ⅂ ieres sempres amendes. ⅂ en un au
fro	14	Good	Procureur du Roi du mẽme iour qui ne l empeche. lOrdon.
fro	14	Good	sonpere tous les rodais et les tartcites

Lang	Century	Prediction	Transcription
fro	14	Good	sacies ꝺl nestoit deriens i cant desirans �ume de
fro	15	Good	quil ne lui celast mit ains lui deist qui
fro	15	Good	miere pour ce que par la renue de cest
fro	15	Good	sauoit puis fait. Et il lui cõte cõme�attached
lat	9	Good	cumppriae accipiť tab naculum belli res est. Adtem pus enim cumd- abolo dimicamꝰ. & tunc opusẽ
lat	9	Good	auť comminati : miscrunteuminex
lat	9	Good	ce Detempore ordinat ionum.
lat	10	Good	epm̃ ñaccipiant xccraxui a gererĩ ꝓcur auerint
lat	10	Good	babeaťˊ. sicastigat: psatis faccionẽ uenia ab epo noluerit ꝓmereri.
lat	10	Good	prima creatrix : posterior
lat	11	Good	cer cũ fratrib; in labore manuu
lat	11	Good	la tricem illã uiris armisq; nobilẽ hispadua: illam semi
lat	11	Good	ꝓ motionẽ dare debebit Postumianus eps̃ dixit:
lat	12	Good	minus.¶ Vmmasculo ñ cõmiscebis contu femineo: ꝺa
lat	12	Good	non tenuerit eccłiastice ficlei caritatisq; cõ
lat	12	Good	que fuerant futura damnantur. Deinde si eisad ꝑcipiendũ bapti
lat	13	Good	diei. q̃sitꝰ a ꝑentib; infans inuentꝰ est ⁊ sublatꝰ defouea obnolutꝰ cenoⁿ
lat	13	Good	fit. ꝑ adumantibo utust lactis q st iťˊ costa
lat	13	Good	sub sarcinis adoriri. Qua pulsa inpedimentisq; direptis. futurũ
lat	14	Good	seqꝝ. dicã i vit. fiałs et
lat	14	Good	do rerũ. Que disciplina: Que grã
lat	14	Good	utusque ⁊ siauł deiusto ⁊ułto tubłiliau quostã ꝑtrła
lat	15	Good	a tlium �usilus extraneus audeat discre pare
lat	15	Good	ꝑparata pena.S; qd cica; : Duodici; fatemur xpm̃ apostolos habuis
lat	15	Good	absoloe oñm et c̓ ꝑ ꝓcessũ aut et t̓ lu ꝺom et c̓ Ncessus de cai et pre uacãtib;
fro	12	Acceptable	hoc michi uircus caritacis ex
fro	12	Acceptable	poue; sg̃uis not arcã de roy nosta siro l gñt dixur de gendy
fro	12	Acceptable	uideliet qꝝ. Vluifxix ꝑ iii obo ł ddẽ debił monsõ daentũ erignita t ån- decim f derẽ ꝺ cuo forep pn hune medum ui
fro	13	Acceptable	q̃ merueilles fu lacitebiengarne mlt
fro	13	Acceptable	eceual �umanda a .i. desgrũs baillies
fro	13	Acceptable	⁊ aumanġ loea lonseigne inporcee
fro	14	Acceptable	en excepter aucuns : quĩ dit les aroits, sans en excepter aucuns, dir tous
fro	14	Acceptable	beancoipe ⁊ de nofimeeeEt chastellaus du chosirur diu hur d ursarce Confe sfout anen en ilirur
fro	14	Acceptable	¶ Oedee est alber de chyam de
fro	15	Acceptable	grans coupz sur leitargt du foy des orgueilleux
fro	15	Acceptable	cau en ueritayp cest grant ⁊ Iouff
fro	15	Acceptable	nophanes eracleopolites ꝺ ceste
lat	9	Acceptable	septies. sedusque septuagies septies.
lat	9	Acceptable	aestuat. Dehac rcriptũ. ẽ :
lat	9	Acceptable	to hostem patriae redire iubet ad propria. Iune
lat	10	Acceptable	bilis sit deuotio. Consttt qu uram dilec tione magna remune
lat	10	Acceptable	sustinebť salus auť mea insẽpiť nũ crit.
lat	10	Acceptable	sorac cae plũr tm̃ ut ħierusolima. quasi ut hic narrabo plũr tm̃ uthutreueri utroque
lat	11	Acceptable	bitatem ipsiis omino ugor
lat	11	Acceptable	diccũ ẽ. ego dns exaudiã eos. dř istł ñderelinquã eos: ñidõ diccũ ẽ. cãquã gen
lat	11	Acceptable	ait.A ẽsis hicuobis micummm̃siũ.primus est uobis irm̃si

Lang	Century	Prediction	Transcription
lat	12	Acceptable	ꝑ secutio leuist adcauedũ.s beticor seductioꝑni
lat	12	Acceptable	ierłm & uide. immo iudicent inť
lat	12	Acceptable	surci :reccutores repu :. lic: etť migriin
lat	13	Acceptable	Meseach ⅂ tafari ⅂ Rrasis sič dicč
lat	13	Acceptable	orit lui. ⅂ termo optimus est
lat	13	Acceptable	quilibet sp̃s. omĩno
lat	14	Acceptable	potior conditio pp̃e.facit de rxp. duobus .li. bl.
lat	14	Acceptable	se dm̃ habere. et pmic⁹ sibimet satiffaciens.
lat	14	Acceptable	ualent vuã breuẽ. ⅂ ultia ualet tũi
lat	15	Acceptable	sup s compaꝛõ ioñ prudẽtes 9quas
lat	15	Acceptable	metermuim.et rẽgm euis non erit finis
lat	15	Acceptable	L e carnalis ht, qm pater ip̃s parentis.
fro	12	Bad	orailleo .poulleer xv lib
fro	12	Bad	Rbir les bartres
fro	12	Bad	deaute lqu ques creppt Eentiferoi rece ny⁹ seelle ces liea aa mn pie do d ce ee lu moasum
fro	13	Bad	atourne. giest sibo lans quil qsui
fro	13	Bad	Carde peo eequie auoit dyonde eane de adtus edtoit pao coudequaiẽ
fro	13	Bad	Mol edito se vtan di or icutts
fro	14	Bad	Q Anne Autiron. Que ledit saques de Lancrau, epousa en premieres noces, le
fro	14	Bad	stallis eudinor 9pu es uiai fugdutu padeur i uo; ferras pu uea puis ai
fro	14	Bad	uolent ipitur ⅂atia rertace, quan ineestaeq aleeeclere, et decõuy ny s ã ã
fro	15	Bad	deianarr de bbrdide
fro	15	Bad	⅂uribz allegate, Sed epclusissent ab uitestate Ipsi
fro	15	Bad	msuoĨ ipousaultis anonuen natucõ auol
lat	9	Bad	lus . necnonalu acquealii fundatores ecdlesiae atque erudito
lat	9	Bad	us prae erat ut ꝑhoc. P̃sedemtis
lat	9	Bad	crea turar quues upra
lat	10	Bad	eruc tucins quat tuor an
lat	10	Bad	utunde positum eleganter concin
lat	10	Bad	aecenim consid rtio suasit qnm manifestum. ẽ. omnemutabile
lat	11	Bad	UERBuai : FIuERBũ.
lat	11	Bad	mus qm ipse anns nř animã posuit suã
lat	11	Bad	fra si tua foret roma.to
lat	12	Bad	et arbusta eius cedros dei.
lat	12	Bad	rit i audalunt dñt surrebeř scm̃q; marie
lat	12	Bad	relecti mansueti.
lat	13	Bad	qurdr uicba .i. quit est
lat	13	Bad	de inim̃. m. ñ quãu
lat	13	Bad	usqi io intintoẽm amcti delendi sumuis
lat	14	Bad	cui subsunc becm̃braa laceiicelligas lřãm. siue sint plati or
lat	14	Bad	Sĩ mõlał ãt ibs
lat	14	Bad	fult ad nol in eigilia lanooe orucil oroit ꝓuril crauns
lat	15	Bad	⅂ Artaiita mons cum flumi-
lat	15	Bad	ꝑ te maiorz pñt corrumpe
lat	15	Bad	lo sunt ni locu unu. ⅂ appare
fro	12	Very bad	I guille choeneau
fro	12	Very bad	mnl cct quarantẽ dope
fro	12	Very bad	nullo cappic d bii uigr
fro	13	Very bad	noulonoe rolissicanuꝑ, Rudauu/, 9garobanu
fro	13	Very bad	L a nuis ÷ eueuue sihat ipponses

Lang	Century	Prediction	Transcription
fro	13	Very bad	diqe uť le diuiɔ inr
fro	14	Very bad	oe consepeedeetante cemere
fro	14	Very bad	OU EtE L. Cheualier de Monifort, son Oncle, Gles
fro	14	Very bad	Bussoy Iaguio dar Rnauex a eedamet dunin
fro	15	Very bad	aximiun oiuinca s apenalriuuuirõuo uutonli
fro	15	Very bad	libas �uadiorandapio sidimił
fro	15	Very bad	Euuon lan uiii fut
lat	9	Very bad	IV Mtru&E Rlnꝯrdo¬
lat	9	Very bad	eo locus sp atiosus admanen
lat	9	Very bad	ie godñpsormanilues
lat	10	Very bad	dule hanu curde uut lato
lat	10	Very bad	arnals de seruull
lat	10	Very bad	sbib liotheca tsede
lat	11	Very bad	s ec tanie ca uis uirtusq;
lat	11	Very bad	Don de N. le duc de la Tremoille . MV
lat	11	Very bad	uitr fuit rtimuli
lat	12	Very bad	minu Benedlicat uos clns exsy
lat	12	Very bad	sad mumnoui hominꝭ
lat	12	Very bad	uanni addanť ad aunos aᷤms cabłam aiomsn ita uidt
lat	13	Very bad	ngeũ drãs mudtistã
lat	13	Very bad	fit cum eo emplin ypocondrus
lat	13	Very bad	mauuse ɔ de ala mri nream
lat	14	Very bad	G terie eni ītuiueẽ sã
lat	14	Very bad	ni quo coła nig ꝯsurg
lat	14	Very bad	uino dť. uł nat
lat	15	Very bad	duabus ncibus ¶ uel syr ma-
lat	15	Very bad	Orano mlelll pam cess aa qra mfenus
lat	15	Very bad	Poleuae me oblous ons

Table 6: Examples of HTR Prediction on unseen documents and their classification by the model.

Lang	Encoder	Good		Acceptable		Bad		Very bad	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Yes	Attention	63.35	40.36	43.04	26.32	49.66	47.70	75.33	96.25
Yes	Attention	65.61	31.47	41.80	32.35	51.81	51.75	77.74	95.41
Yes	Attention	63.32	51.27	**49.17**	18.27	47.84	45.95	71.96	95.99
Yes	Attention	66.00	41.88	48.45	33.90	52.85	56.78	80.47	94.51
Yes	Attention	68.27	43.15	44.01	22.76	44.67	43.98	73.20	95.48
Yes	TextCNN	59.06	38.07	44.31	11.46	40.16	32.17	64.50	**97.87**
Yes	TextCNN	54.70	25.13	38.16	22.45	44.88	44.09	72.30	95.41
Yes	TextCNN	51.54	42.39	44.37	20.74	45.34	27.13	64.88	97.61
Yes	TextCNN	60.23	26.14	39.78	27.40	45.54	43.00	72.81	95.16
Yes	TextCNN	63.35	25.89	43.24	22.29	42.40	33.26	65.95	97.61
Yes	BiLSTM	62.33	46.19	44.07	12.07	44.10	51.09	74.11	94.51
Yes	BiLSTM	71.75	32.23	41.29	30.80	50.41	53.94	78.07	94.06
Yes	BiLSTM	60.56	**55.33**	47.55	21.05	49.94	48.14	74.18	94.64
Yes	BiLSTM	**73.53**	19.04	36.58	30.80	47.77	**58.53**	81.51	91.41
Yes	BiLSTM	66.82	37.31	39.26	14.71	43.42	47.26	72.96	96.38
No	Attention	56.05	44.67	45.02	30.80	49.16	44.75	76.96	95.16
No	Attention	61.79	33.25	43.23	**43.50**	54.51	58.21	**86.30**	92.76
No	Attention	57.00	43.40	45.66	34.21	50.41	47.16	78.53	94.51
No	Attention	58.57	41.62	44.40	38.08	54.50	53.72	82.40	94.06
No	Attention	56.97	36.29	42.20	31.42	51.34	46.06	75.85	95.54
No	TextCNN	54.51	36.80	41.15	26.63	47.59	43.22	74.30	95.41
No	TextCNN	47.37	38.83	40.10	26.01	49.14	47.16	77.48	94.25
No	TextCNN	54.71	38.32	41.80	28.02	49.66	55.58	81.19	92.83
No	TextCNN	48.31	39.85	39.09	28.02	50.12	47.26	78.94	94.44
No	TextCNN	49.35	38.32	39.04	15.17	46.45	46.50	72.71	95.35
No	BiLSTM	70.79	31.98	42.90	24.30	47.44	55.80	78.07	94.96
No	BiLSTM	57.60	36.55	42.23	35.76	52.28	52.63	80.84	93.22
No	BiLSTM	60.25	36.55	41.84	28.17	51.37	57.44	80.40	93.80
No	BiLSTM	56.69	49.49	42.90	40.71	**54.52**	47.48	82.85	93.60
No	BiLSTM	56.17	43.91	44.47	26.78	49.12	48.58	77.64	95.35

Table 5
Results of each model on the Berlin, Hdschr. 25 manuscript.

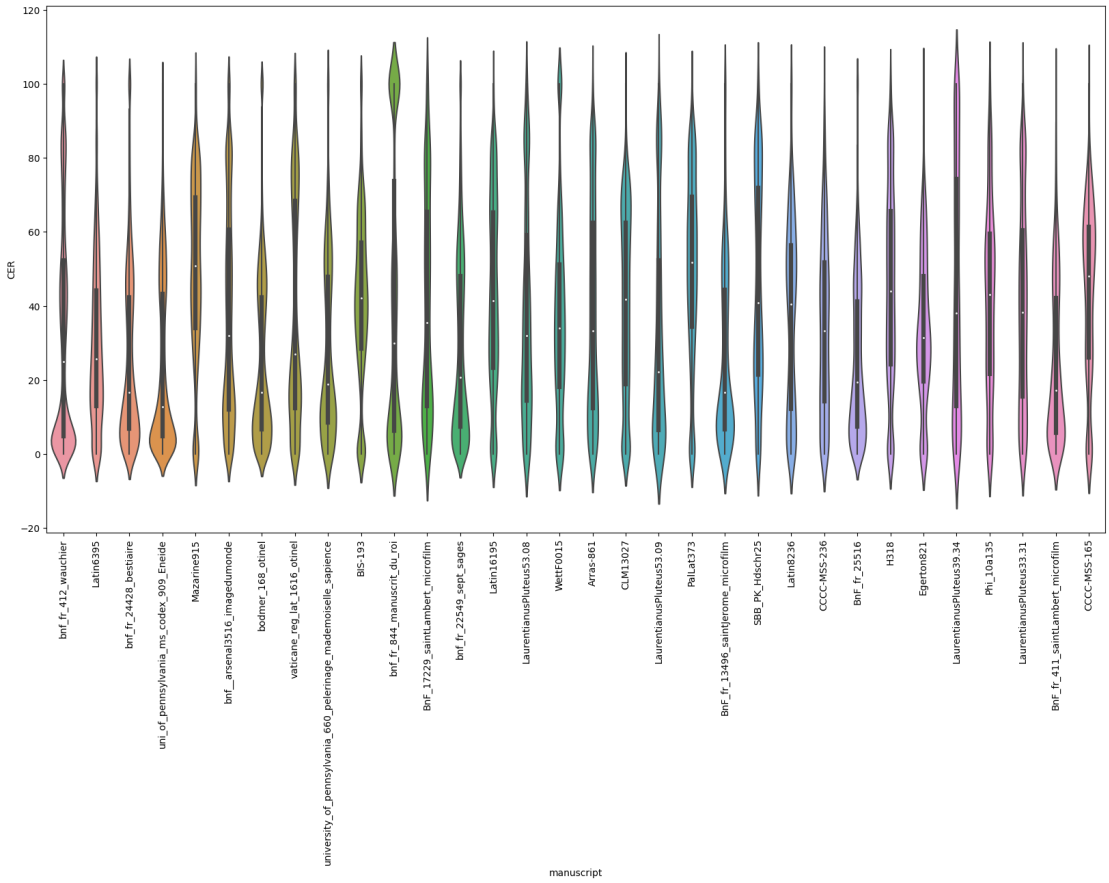


Figure 6: "Bad transcriptions" CER Violin plot, per manuscript. Most manuscript have a strong enough diversity of CER to train upon.