Handwritten Text Recognition for Documentary Medieval Manuscripts

Sergio Torres Aguilar, Vincent Jolivet

Handwritten Text Recognition for Documentary Medieval Manuscripts

Sergio Torres Aguilar[1] **and Vincent Jolivet**[2]

[1]Université du Luxembourg, Institut für History, Luxembourg
[2]École nationale des chartes, Centre Jean-Mabillon, France

Corresponding author: Sergio Torres , sergio.torres@uni.lu

Abstract

Handwritten Text Recognition (HTR) techniques aim to accurately recognize sequences of characters in input manuscript images by training artificial intelligence models to capture historical writing features. Efficient HTR models can transform digitized manuscript collections into indexed and quotable corpora, providing valuable research insight for various historical inquiries. However, several challenges must be addressed, including the scarcity of relevant training corpora, the consequential variability introduced by different scribal hands and writing scripts, and the complexity of page layouts. This paper presents two models and one cross-model approach for automatic transcription of Latin and French medieval documentary manuscripts, particularly charters and registers, written between the 12th and 15th centuries and classified into two major writing scripts: Textualis (from the late-11th to 13th century) and Cursiva (from the 13th to the 15th century). The architecture of the models is based on a Convolutional Recurrent Neural Network (CRNN) coupled with a Connectionist Temporal Classification (CTC) loss. The training and evaluation of the models, involving 120k lines of text and almost 1M tokens, were conducted using three available ground-truth corpora : The e-NDP corpus, the Alcar-HOME database and the Himanis project. This paper describes the training architecture and corpora used, while discussing the main training challenges, results, and potential applications of HTR techniques on medieval documentary manuscripts.

Keywords

medieval charters, HTR for historical documents, HTR for medieval Latin manuscripts, digital diplomatics, medieval digital studies

I INTRODUCTION

Recent advancements in Handwritten Text Recognition (HTR) techniques, along with increased availability of ground-truth corpora, have facilitated the development of powerful models for handwriting recognition for various sources of interest in historical research. The transition from experimental to production phase has been achieved in the past five years, thanks to the widespread use of general-purpose HTR toolkits and turn-key annotation interfaces. Many institutions now consider automatic acquisition of text as a crucial step for studying and disseminating their digitized collections in a structured format. This allows for unrestricted querying, indexing, and sharing operations. As a result, there is an increasing demand for ground-truth data aligned with specific needs, as well as robust pre-trained models that can be adapted to unique requirements.

In an effort to address this pressing need, this paper presents three HTR models that can efficiently transcribe a wide range of manuscripts including charters collections, registers, and serial documents. These types of manuscripts are often the most abundant yet understudied sources. The proposed general-purpose models are specifically made to be reused, and ready

to be deployed and fine-tuned for specific uses. They focus on two script families : *Textualis*, used in manuscripts between the late-11th and 13th centuries, and *Cursiva*, which emerged in the mid-13th century and persisted into the Modern Age. The ground-truth data for training the models is sourced from collections that encompass two of the most significant written sources for studying the Middle Ages: cartularies, which are copies of deeds related to the transfer of rights (sales, donations, exchanges, etc.); and registers, i.e. serial documents pertaining to the internal management and public activities of an institution.

This paper is organized into three distinct sections. The first section presents the training corpus and outlines the architecture used for modeling historical writing. The second section conducts a comprehensive analysis of the obtained results, including an in-depth examination of the most prevalent errors. Lastly, the paper discusses the implications and challenges raised by the application of HTR in the field of medieval paleography and diplomatics.

II RELATED WORKS

The field of HTR techniques applied to manuscripts has undergone significant transformation over the past decade, transitioning from Markov engine-based models (Bunke et al. [1995]) to deep-learning approaches using neural networks at character level (Graves and Schmidhuber [2008]). Recent advancements such as the introduction of CTC mechanisms and bidirectional networks (Graves et al. [2008]) have considerably enhanced accuracy and reduced the dependency on ground-truth data for the development of generalist models (de Sousa Neto et al. [2020]). In the realm of medieval manuscripts, pioneering efforts have been made by projects such as Himanis (Stutzmann et al. [2017]) and Home (Stutzmann et al. [2021]), supported by Transkribus, in creating extensive aligned and documented corpora for effective HTR and NLP model training on medieval sources. Similarly, the Scripta-PSL and Cremma projects (Chagué [2021]), supported by Kraken and the eScriptorium platform, have introduced widely adopted scientific practices and annotation guidelines (Pinche [2022]) for model production applied to medieval Latin and Hebrew manuscripts.

Moreover, in recent years, the application of HTR techniques to historical manuscripts has increasingly intersected with paleography and diplomatics. This symbiotic relationship is natural, as generating ground-truth data for modeling ancient scripts is not only resource-intensive and time-consuming, but also requires specialized expertise in reading and deciphering ancient languages and scripts. It also requires proposing intellectual methodologies to clarify ambiguities and define annotation guidelines. Paleographers and historians have actively contributed to the creation and selection of reliable ground-truth data, as well as the analysis of common errors of prediction results, providing valuable feedback for improving models.

In contrast to OCR, which typically deals with a limited set of typographic variations, HTR requires the development of specialized models, facing a higher degree of variability: chronological aspect of the texts, typology of the documents, script families, regional practices, even customs and personal writing practices of the people who produced it: scribes, court scribes and notaries. Those issues have sparked discussions that go beyond simple accuracy-based results, delving into the nature of the ground truth data itself. Questions arise on the necessity to account for abbreviations (Camps et al. [2021]), on the transcription conventions –graphemic or diplomatic, i.e. imitative, as proposed by Driscoll (Driscoll [2006]), and on the representativeness of a training corpus while mitigating potential biases (Schoen and Saretto [2022]).

Writing, as a normative system, can be effectively modeled using modern HTR techniques. Recent studies have demonstrated that HTR applied to medieval manuscripts can achieve im-

pressive prediction rates, with less than 5% Character Error Rate (CER). However, it is worth noting that this rate may significantly decrease when models are confronted with documents outside their domain or belonging to a different script family. This is particularly challenging in the case of ancient scripts, where readers must restore missing information encoded in abbreviations, formulae, and implicit knowledge that cannot be fully modeled using current state-of-the-art techniques.

In previous years, modeling proposals for handwriting recognition were often focused on individual hand-level and author-centered approaches. However, due to the significant variability in writing styles throughout the medieval centuries and the limited availability of domain-specific ground truth, it has become evident that a more global approach of handwriting classification is more appropriate (Hodel et al. [2021]). In our case, we adopt the classification based on Latin script families, as proposed by the CLAMM corpus, which encompasses 12 book-script families spanning the period from the 9th to the 15th centuries (Kestemont et al. [2017]). Our proposed models specifically focus on the two most numerous script families, namely Cursiva (mid-13th to late 16th centuries) and Textualis (late 11th to mid-13th centuries), for which sufficient ground truth data is available. It is important to note that this division is purely consensual, as the number of script families and sub-families increases significantly towards the end of the 14th century.

III CORPORA DESCRIPTION

3.1 The e-NDP corpus

The corpus of the registers of Notre-Dame-de-Paris is one of the few available corpora dealing with the HTR of documentary manuscripts from the late medieval period (Claustre and Smith [2022]). These registers, consisting of minutes of decisions made during the weekly meetings held by the canons, pertain to the management of the institution and its assets. A key objective of the e-NDP project was to obtain the text from over 14,000 pages constituting the entirety of the records from the medieval period, housed in the Archives nationales (LL105 to LL128). Subsequently, the inferred text has been structured and transformed into a research engine capable of indexing and facilitating quick access to valuable information for studying one of the most significant urban institutions of medieval France. For this purpose, a total of 500 pages from 26 registers, dating from 1326 to 1504, were transcribed by historians and paleographers using eScriptorium. These transcriptions were then used to train HTR models specifically adapted to produce high-quality transcriptions of all the digitized pages.

The registers in question are characterized by their usage of a family of cursive scripts, predominantly in Latin, with occasional pages and formulations in medieval French. Their page layout, typical of documentary manuscripts, evolves with time, with patterns such as lists of names, margin notes, titles, and other peritextual additions. These manuscripts were often produced for the purpose of daily consultation and administration, rather than the intention of long-term preservation, resulting in a less meticulous design compared to literary manuscripts.

3.2 The HOME-Alcar corpus

The HOME-Alcar corpus (Stutzmann et al. [2021]) was created as a component of the European research initiative ”HOME History of Medieval Europe.” This project was coordinated by the Institut de Recherche et d’Histoire des Textes (IRHT-CNRS). The corpus offers a collection of images of medieval manuscripts that are aligned with their scholarly editions at the line level, along with comprehensive annotations of named entities such as persons and places. This

Table 1 List of training and testing manuscripts.

Set	Manuscript	code name	script type	nº lines	nº tokens
train	Cartulary of the Notre-Dame de la Roche abbey	Roche	Cursiva antiquior	2103	19164
train	Cartulary of Saint-Denis abbey	S_Denis	Textualis	18363	132854
train	Cartulary of Charles II of Navarre	Navarre	Cursiva	6777	94262
train	Cartulary of Notre-Dame de Vauluisant abbey	Vauluisant	Textualis	12642	69364
train	Cartulary of Notre-Dame de Fervaques abbey	Fervaques	Textualis	4661	45251
train	Cartulary of Saint Nicaise of Reims	S_Nicaise	Textualis	7404	99526
train	Cartulary of Notre-Dame de Clairmarais	Clairmarais	Semi-hybrida & Cursiva	8554	77478
train	Formulary of Odart of Morchesne	Morchesne	Cursiva	10515	110033
train	Registers of the chapter of Notre-Dame de Paris	e-NDP	Cursiva	33735	202348
test	Cartulary of Nesle seigneury	Nesle	Cursiva & Textualis	3562	37756
test	Cartulary of the Pontigny abbey	Pontigny	Textualis & Cursiva antiquior	10717	78045
test	Cartulary of the Cathedral of Notre-Dame de Chartes	Chartres	Textualis	1636	14564
test	Register of the French Royal Chancery	Himanis	Cursiva	485	8441
	Total Train			104754	850280
	Total Test			16400	138806
	Train + Test			**121154**	**989086**

corpus serves as a valuable resource for training HTR and Named Entity Recognition (NER) models synchronously.

The corpus contains 17 French cartularies, which are volumes containing medieval copies of original documents, dating from the 12th to the 14th centuries. These cartularies belong to at least 4 distinct script families: Textualis, Cursiva, Cursiva Antiquior, and Semi-Hybrida. Cartularies were commonly produced in ecclesiastical institutions since the 11th century, and in civil institutions from the 13th century onwards. These volumes are highly valuable in medieval studies as they contain documents that were often not preserved in their original form, such as property transfers, wills, land and debt disputes, as well as rarer documents such as treaties, indemnities, or successions. All of these documents are included in the HOME-Alcar corpus. In summary, this corpus comprises 3090 acts, with 2760 written in Latin and 330 in Old and Middle French, totaling almost 1 million tokens.

3.3 The Himanis project

The Himanis project, in collaboration with the READ consortium (Recognition and Enrichment of Archival Documents), has developed the most advanced model to date for late medieval script recognition, known as Himanis Chancery M1+. This model is based on the partial edition of the registers produced by the French Royal Chancery between 1302 and 1483, specifically the Archives nationales, JJ35 to JJ211. These registers, also referred to as cartulary-registers, contain copies of various types of charters, including letters of remission, mandates, amortizations, ennoblements, and property confirmations.

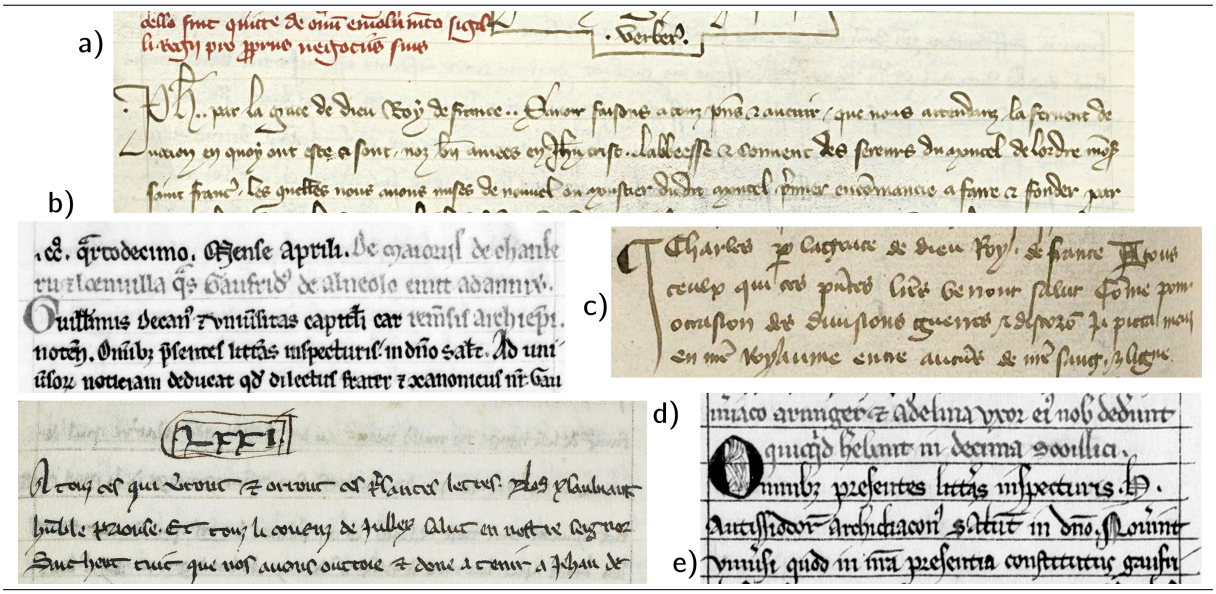
The Himanis model was developed by aligning the digitized images line by line with the partial semi-diplomatic edition of Paul Guerin (Guérin [1881]), and then encoding the data using the HTR+ engine in Transkribus (Leifert et al. [2016]). This model achieved a validation Character Error Rate (CER) of 0.08. Using this model, over 70,000 pages were transcribed, and a querying tool with indexing and word-spotting functionality was built to facilitate access to the entire French Royal Chancery corpus. In 2021, the training dataset for the Himanis model, known as HIMANIS-Guérin, was made freely available on Zenodo. This dataset includes 1,500 images and 30,000 text lines of ground-truth data, which were primarily written in Latin and Old French using a Cursiva script.

3.4 Datasets configuration

During the training phase, the manuscripts were distributed into two groups (see Table 1) according to the predominant type of writing:

- Cursiva (**G1**) contains 5 elements: The e-NDP registers; and the cartularies of Navarre, Clairmarais, Notre dame de la Roche and the formulary of Odart de Morchesne for a total of 61684 lines.
- Textualis (**G2**) contains 4 elements: The cartularies of Saint Denis, Fervaques, Saint Nicaise and Valuisant for a total of 43070 lines.

Figure 1 Five examples of act protocols from Himanis (a), Chartres (b), e-NDP (c), Nesle (d) and Pontigny (e). In both Latin and French, all five follow a similar act opening containing an intitulatio, a general address ("to all who will read this letter") and a salutation.



Four manuscripts that were not used during the training phase were set aside for the purpose of testing. These manuscripts include: (1) the Nesle cartulary, which predominantly features Cursive script but also contains some pages in Textualis script; (2) the Chartres cartulary, which is entirely written in Textualis script; (3) the Pontigny cartulary, which exhibits alternating sections in Cursiva Antiquior and Textualis script; and finally, (4) a random set of 30 pages from Himanis, selected from various volumes of registers and written in Cursiva script.

Both the training and test corpora are bilingual, with an overall ratio of 4:1 (Latin/French) for the training set and 5:2 for the testing set. Furthermore, as corpora come from diplomatic editions, the transcriptions can be defined as semi-diplomatic, that means all abbreviations have been expanded, punctuation standardized, named entities capitalized and all variants of a given letter (allographs) reduced to the canonical letter without distinctions between them.

Additionally, the commas in the ground truth were adjusted in all documents to conform with their usage in the original manuscript. In instances where modern editors introduced a comma (or semicolon) to indicate a pause in the sentence, even though such punctuation marks were not present in the manuscript, the comma was removed. In other cases, commas were replaced with periods (".") as it is often the period that serves the three functions (soft, middle, and final) of separating sentences or clauses in most manuscripts.

The complete set of 110 characters used in training transcription is listed below:

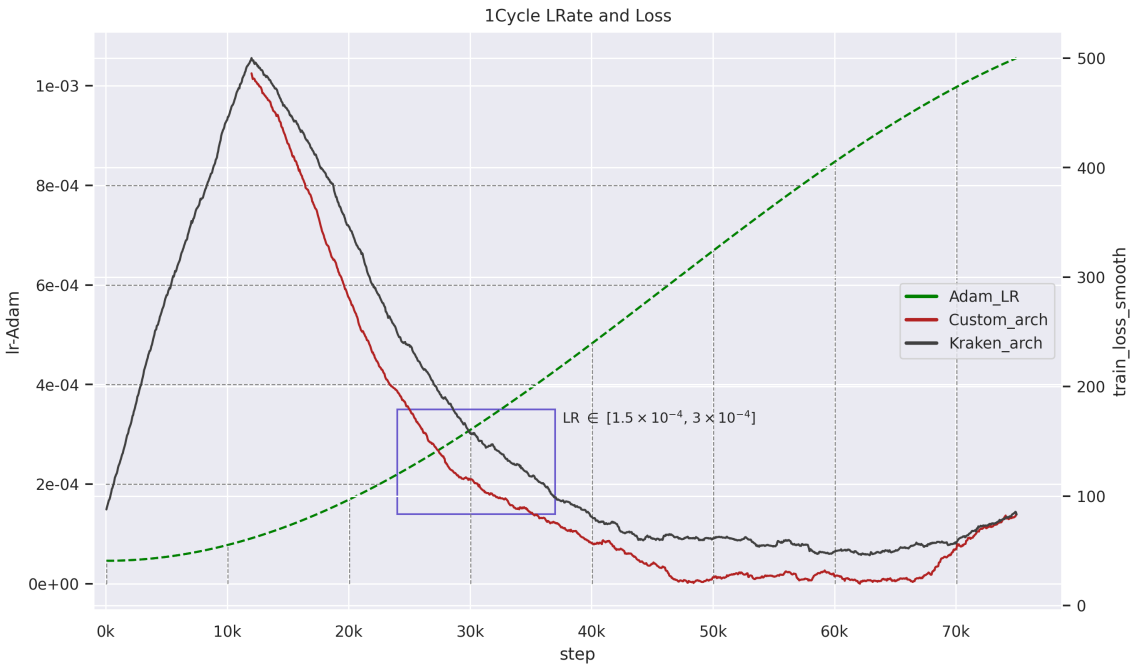
- lowercase letters : abcdefghijklmnopqrstuvwxyz
- capital letters : ABCDEFGHIJKLMNOPQRSTUVWXYZ
- numbers : 0123456789
- punctuation marks : .,;:!?’-| blank
- diacritical marks : ÇÉÏÜàâæçèéëęïüÿōœñ
- other glyph : #%$*+-§“”°[](){}/ ><

IV HTR ARCHITECTURE

The HTR architecture employed in this study can be characterized as a classical Convolutional Recurrent Neural Network (CRNN) pattern recognition approach, consisting of 5.9 million trainable parameters. The architecture, which seems to achieve a slightly smaller loss on our documents compared to Kraken’s setup default (See Figure 2), operates in three steps: (i) gray-scale images are inputted into four Convolutional Neural Network (CNN) layers for feature extraction and encoding; (ii) the features are then propagated through three Recurrent Neural Network (RNN) layers in a bidirectional manner to capture contextual information; and finally (iii) a Connectionist Temporal Classification (CTC) algorithm is employed to compute the model’s loss value and render the inferred text as a UTF-8 string.

(i) The convolutional block consists of layers with varying kernel sizes (4x16, 3x8) and 16n filters per layer (32, 32, 64, 64). Following each convolutional layer, the neurons pass through a MaxPooling layer with a 2x2 kernel. Additionally, a 2D dropout with a probability of 0.1 is applied to each layer, and Rectified Linear Units (ReLU) are used as the activation function. Prior to applying the recurrent block, a reshape layer is used to collapse the non-1 height dimensions into a single value, with the height of the images fixed at 128 pixels, allowing for variable width to accommodate varying sizes of manuscript lines more effectively.

Figure 2 The OneCycle scheduler was applied to both our architecture and Kraken’s v4 default setup. The Adam learning rate was cycled between $\{5 \times 10^{-5} : 1 \times 10^{-3}\}$ values for 5 epochs. Results indicate that a starting learning rate between $\{1.5 \times 10^{-4} : 3 \times 10^{-4}\}$ is optimal for a batch size of $\{4\}$ which confirms a LR value as the $sqrt$(batch-size) . Data was smoothed using the Savitzky-Golay filter.



After the reshape layer, the recurrent block is activated (ii). It consists of three BiLSTM networks, each using a 1D dropout with a probability of 0.3. The number of hidden units in all the LSTMs is set to 256. Subsequently, (iii) the output passes through a dense layer with softmax activation, with a size equal to the charset size + 1 (CTC blank symbol). A 1D dropout with a probability of 0.3 is also applied before this last layer.

In Kraken, which uses the VGSL network specification, this training architecture can be fully replicated by using : `-s '[1,128,0,1 Cr4,16,32 Do0.1,2 Mp2,2 Cr4,16,32 Do0.1,2 Mp2,2 Cr3,8,64 Do0.1,2 Mp2,2 Cr3,8,64 Do0.1,2 S1(1x0)1,3 Lbx256 Do0.3,2 Lbx256 Do0.3,2 Lbx256 Do0.3]'`

4.1 Hyperparameters

During the training process, we use a batch size of $\{4\}$; a learning rate of $\{2.5 \times 10^{-4}\}$ as indicated by the LR finder (See figure 2); and a pad size of $\{24\}$ pixels. The pad size is useful in providing more space to the kernel for better coverage of the image. Additionally, we ran a $\{$ReduceOnPlateau$\}$ optimizer with patience $\{3\}$. Data augmentation techniques, such as distortion, blur, and rotation, were also applied to the images to enhance training performance.

All the training cycles were realized using a RTX 3090 (24GB) coupled to a 64GB of RAM for about 6-40 hours depending of the size of the corpus.

V EXPERIMENTS

We formulated our experiments with three primary objectives in mind: (i) Developing generic models for each script family; (ii) Exploring the correlation between performance accuracy and the quantity and diversity of ground truth data; and (iii) Assessing the advantages and limitations of creating a composite model by combining resources from both script families. To achieve these goals, we conducted four distinct experiments :

1. A step-wise training using batches of documents for each script family. (Tables 2 and 3)
2. A cross-script and cross-lingual training on quartiles (Q1 is composed of 400 pages : 200 random pages from each script family; Q2 contains 400 of each one, etc.) using the entire ground-truth (G1 + G2)
3. A fine-tuning exercise (retraining on initialized weights) for the models developed in points 1 and 2.
4. A testing exercise was conducted on four multilingual and mixed-families manuscripts that were not included in the training dataset. The purpose of this exercise was to assess the models' generalization ability and sensitivity to retraining.

The performance of each model will be evaluated using widely recognized metrics, including accuracy (calculated as errors per characters), CER (Character Error Ratio), and WER (Word Error Ratio), without applying normalization to the results. Additionally, for the fine-tuning exercises, we used 10 reserved pages from each manuscript.

VI RESULTS

The results presented in Tables 2 and 3 reveal that the proto-Gothic and early Gothic scripts (Textualis) achieve a convergence to an accuracy over 0.9, much faster compared to the Late-Gothic scripts (Cursiva), which generally require a significantly higher number of pages to reach the same level of accuracy. However, fine-tuning on the Textualis manuscripts (Chartres) only yields a modest improvement of +1.2% compared to the generalist model. In contrast, the Cursiva manuscripts (Nesle and Himanis) are more responsive to fine-tuning, resulting in

substantial accuracy improvements of +6% (Table 2) and +8% in three of the test experiments (as shown in Table 4). The fitting on Pontigny, which combines Textualis and Cursiva antiquior, naturally falls in the middle with a +4.5% improvement (Table 4).

Table 2 Evaluation results for G1 models. *val_acc*: validation accuracy during training. *test_acc*: testing accuracy on the cartulary of Nesle

model_name	Content	pages	val_acc	test_acc	CER	WER
G1_test_1	Clairmarais, Roche	121 + 66	0.935	0.696	0.296	0.629
G1_test_2	Clairmarais, Roche, +Navarre	187 + 208	0.939	0.798	0.192	0.518
G1_test_3	Clairmarais, Roche, Navarre, +Morchesne	395 + 176	0.941	0.797	0.194	0.514
G1_test_4	Clairmarais, Roche, Navarre, Morchesne, +e-NDP/5	571 + 100	0.939	0.807	0.183	0.501
G1_test_5	Clairmarais, Roche, Navarre, Morchesne, +e-NDP full	671 + 400	0.936	0.840	0.156	0.448
G1_FineTuning	+10 pages Nesle	1071 + 10	**0.942**	**0.901**	**0.095**	**0.277**

Table 3 Evaluation results for G2 models. *test_acc*: testing accuracy on the cartulary of Chartres.

model_name	Content	pages	val_acc	test_acc	CER	WER
G2_test_1	S_Denis, Fervaques	199 + 90	0.929	0.885	0.097	0.278
G2_test_2	S_Denis, Fervaques, +S_Nicaise	289 + 109	0.932	0.906	0.078	0.225
G2_test_3	S_Denis, Fervaques, S_Nicaise, +Vauluisant	398 + 106	0.935	0.914	0.071	0.213
G2_FineTuning	+10 pages Chartres	504 + 10	**0.938**	**0.926**	**0.060**	**0.163**

Table 4 Evaluation results for cross-scripts models by quartile. Each quartile (Q*n*) contains 25% of the entire training corpus (G1 + G2 : 1575 pages). The Fine-tuning experiments (Q*n* FT) were performed using 10 reserved pages from each manuscript. For Nesle and Chartres, these pages and the test-set remain the same as the previous experiments.

Manuscript /	Nesle (G1)			Chartres (G2)			Pontigny (G1 + G2)			Himanis (G1)		
metric	test_acc	CER	WER	test_acc	CER	WER	test_acc	CER	WER	test_acc	CER	WER
Q1 (25%)	0.841	0.153	0.430	0.911	0.073	0.209	0.821	0.166	0.454	0.789	0.222	0.580
Q1 FT	0.890	0.106	0.292	0.917	0.068	0.183	0.879	0.110	0.301	0.863	0.136	0.416
Q2 (50%)	0.862	0.133	0.386	0.919	0.065	0.179	0.843	0.146	0.397	0.803	0.202	0.544
Q2 FT	0.904	0.090	0.256	0.926	0.060	0.163	0.888	0.101	0.275	0.884	0.115	0.361
Q3 (100%)	0.868	0.127	0.363	0.923	0.062	0.170	0.848	0.142	0.373	0.832	0.171	0.468
Q3 FT	0.910	0.084	0.234	0.928	0.057	0.158	0.892	0.097	0.264	0.893	0.104	0.324

HTR modelizations are often conducted on entire corpora of manuscripts. However, as evident from the cross-family models in Table 4, using smaller batches with a diverse range of origins can give satisfactory results. The cross-family models, trained with 200 pages from each script (Q1, 25%) from 9 manuscripts, are able to identify dominant trends more quickly and achieve comparable performance to the Textualis (0.914 vs 0.917) and Cursiva (0.840 vs 0.841) models (Tables 2 and 3) that have more than twice the amount of ground truth data. However, the improvements become more costly beyond the inflection point of around 85% accuracy in Cursiva and 90% in Textualis, where the increase in quantity of ground truth introduces only discrete gains and leads to a plateau in the accuracy model. This is evident from the Q2 and Q3 models (Table 4), where the improvement of 1 to 2 points in accuracy on the four test manuscripts required incorporating over a thousand new pages of ground truth data.

The tests conducted to assess the impact of a multilingual and multiscript corpus revealed that cross-training does not result in any loss of accuracy. On the contrary, it leads to a slight improvement in Chartres (0.914 vs 0.923) and a significant improvement in Nesle (0.840 vs 0.868), which is predominantly written in Cursiva with some pages in Textualis. Despite the existence of a bilingual record, as seen in Nesle, Himanis, and Pontigny, the quality of inference

does not seem to be dramatically reduced, although it could be a potential source of errors due to the differences in the Latin abbreviation systems used in French and Latin. Using such multilingual and multifamily models is not only an effective way to reduce bias and increase variance, but also aligns with the characteristics of medieval documentary textuality. On one hand, medieval documentary compilations often contain documents written in diverse hands and script families, sometimes within the same page. On the other hand, handwriting during that period was evolutionary and transitional, with scripts like Cursiva antiquior being widely practiced. The ability of HTR models to handle multilingual sources is crucial for late-medieval documents, as acts in vernacular and Latin coexist within the same manuscript or collection, and Latin formulations in legal value documents are commonly used in the late-medieval period (Glessgen [2004]).

Table 5 The ten characters accumulating the most errors (substitutions + deletions + insertions) according to the hypotheses of the Q3 and Q3 FT models on the test datasets. The updating of the models on new data allows for significant improvements in the reading of the most problematic characters. *+diff(%)*: Precision improvement after FT for each character. *impr_acc*: Total percentage of precision improvement after FT on the ten characters.

model /	Nesle (205064 chars)			Chartres (97129 chars)			Himanis (52110 chars)			Pontigny (464660 chars)		
char	Q3	Q3 FT	+diff (%)	Q3	Q3 FT	+diff (%)	Q3	Q3 FT	+diff (%)	Q3	Q3 FT	+diff (%)
space	2706	2057	24	1066	1022	4	536	347	35	5721	4822	16
s	2383	863	64	435	392	10	707	335	53	5987	3876	35
i	2211	1734	35	675	613	9	958	598	38	7870	5303	33
e	2307	1492	22	643	598	7	1098	546	50	6370	4891	23
n	2045	1143	44	379	346	9	731	434	41	5365	3457	36
r	1775	1190	33	301	308	-2	823	401	51	4461	2925	34
u	1426	930	35	392	369	6	620	455	27	4473	2911	35
t	1390	795	43	359	327	9	605	326	46	2852	2036	29
a	1087	784	28	328	300	9	587	248	58	3981	2232	44
m	979	738	25	294	265	10	449	252	44	3949	2945	25
c	866	684	21	424	321	24	437	279	36	2614	1448	45
Total	19175	12410	35	5296	4861	8	7551	4221	44	53643	36846	31
impr_acc	205064 / 6765		**+3.30**	97129 / 435		**+0.45**	52110 / 4221		**+6.39**	464660 / 16797		**+3.61**

According to existing literature on the topic of *dirty* OCR, it has been demonstrated that a minimum threshold of 80% Character Error Rate (CER) is required for the effective utilization of natural language processing tools (Eder [2013]). Conversely, human legibility, and an optimal level of recall during text indexing (Evershed and Fitch [2014]) are not ensured unless a range between 85%-90% transcription precision is reached (Mühlberger et al. [2014]). From here, the time investment required for text collation and text post-processing stages remains significant (Springmann et al. [2016]). In modern OCR engines, a 90% CER result is deemed subpar. However, in HTR, characterized by its intricate nature and insufficient data availability, this same outcome provides a preliminary basis for accelerating the development of tailored training resources. Generic models, such as ours, are not intended for direct application. Instead, their purpose lies in addressing data scarcity challenges by serving as a foundational model for analogous HTR tasks. Results indicate that a first fine-tuning exercise on generic transcriptions using only a few pages can quickly improve CER by a range of 2% to 7%, which in terms of WER (Word Error Ratio) can equate up to 20% fewer erroneous words.

Regarding the analysis of errors, a consistent pattern is observed in all cases, as shown in Table 5. The most common errors are related to insertions of a limited set of characters, including *i*, *e*, *s*, *r*, *n*, *m*, *t*, *u*, *a*, and *white-space*. Upon closer examination, these errors often coincide with misrecognition of typical phenomena in medieval handwriting. These include indistinctness in characters composed of successive *minims* (single strokes), such as *n*, *m*, *i*, *u* (e.g., *indiuidue*,

mandauimus); misrecognized ligatures, such as *st* and *ct*; undeveloped or incompletely developed abbreviations by suspension, which are easy to fit (e.g., *no[-bis]*, *franc[-orum]*); by contraction, which are harder to fit (e.g., *m[a]g[is]t[er]*, *d[o]m[i]n[us]*); abbreviations whose expansion depends on the declension, (v.g. 'Par.' which can be resolved as Par[is], Par[isiensis] or Par[isiense] depending on the context); final form variants, such as *s* ronds and longs (ſ) or allographs: v.g. single and two-compartment *a*; diacritical marks, such as *é* and *à*; and unresolved diphthongs, such as *ae* and *oe*.

Figure 3 Transcribed line in Himanis (AN, JJ073, 44v, l.22). GT: ground-truth; Q3: prediction of the Q3 model; FT: Prediction of the Q3 fine-tuned model.

GT : royal de grace especial et certaine science au dit Guillaume à ses hoirs et à ceulx qui auront cause de
Q3 : royal de quice especial et cavine science ausdit Guille a ses hoirs et aceube qui auront cai se de
FT : royal de grace especial et certauine science au dit Guillame a ses hoirs et a ceulx qui auront cause de

Figure 4 Transcribed line in cartulary of Nesle (f. 106r, l.16).

GT : estoient en sa justise et en sa seignerie et des quex il estoit saisiz quant nos
Q3 : estoient en la justile et en la leignerie et desquet il estoit lailiz qoant nes
FT : estoient en sa justise et en sa seignerie et des quer il estoit saisiz quant nos

Figure 5 Transcribed line in cartulary of Navarre using normalized (GT) and abbreviated (PL) transcription modes (f. 86v, l.18).

GT : Jehan de Contelery chevalier faiz assavoir a tous que comme tres excelllent prince le roy de Navarre
PL : Jeh[an] de Contelery , ch[eva]li[er] faiz assauoir atous que com[m]e tres [e]xcelle[n]t prince le R[o]y de na[ua]rre

The case of *white-space* errors is special because it encompasses several phenomena. One is editorial, as modern editions, from which most transcriptions are derived, often introduce punctuation marks and spaces that may not be reflected in the original manuscript. Others are intrinsic to late-medieval writing itself, such as ligatures (letters sharing the same stroke) or fake ligatures, and long end strokes, especially after prepositions or articles (as seen in Figure 3 and Q3 transcription), which hinder the automatic isolation of words. Additionally, the use of *scripta continua* (more common in manuscripts prior to the 12th century) or the dissimilar practice of blank space and period by scribes to separate words and blocks of sentences in documentary manuscripts adds to the complexity of *white-space* errors.

It is essential to investigate the specific characteristics acquired by the model through the fine-tuning process. It is evident that certain optimizations stem from the simultaneous incorporation of new data points. The refined model facilitates a more accurate comprehension of the fluent handwriting style and the abbreviation practices employed by scribes. Consequently, it enhances word segmentation, thereby reducing errors related to whitespace, and improves the treatment of abbreviations, thereby mitigating errors in word endings and within words themselves. This effectively prevents literal transcription errors, such as the instance of *Guille*

instead of *Guillaume*. This is exemplified in the cases of Himanis and Pontigny, where the model progresses synchronously in learning the nuances of *i*, *e*, *m*, *u*, *r*, *b* and the *blank* (see Figure 3). In some rare instances, the model may also learn specific style features of individual scribes. As an example, consider Nesle (refer to Figure 4), where a particular scribe employs an unconventional letter "s" that bears a striking resemblance to the "l" utilized in the authentic manuscripts. Initially, the Q3 model transcribes it as an "l," but through the process of fine-tuning, this inconsistency is rectified. This straightforward modification yields a noteworthy enhancement in the overall accuracy, with an approximate increase of nearly 1%. Notably, this adjustment addresses two of the most frequently employed characters in the manuscript.

It's evident that a significant number of errors are inherent in a model trained on semi-diplomatic transcriptions. The model will probabilistically replicate normalization practices and diplomatic conventions, producing clearer and legible texts but introducing a deliberate disparity between the graphical representation and textual content of each manuscript line (refer to Figure 5). The standardization process carried out by editors involves using a restricted set of characters, expanding abbreviations (which often rely on inflection), replacing glyphs and normalizing spacing, punctuation, and ligatures whose use varies widely from one tradition to another. This series of choices made by the editors to modernize the texts actually constitutes a set of manuscript reading levels that lies between interpretation and adherence to common but not always rigid text restoration principles. Nevertheless, even though multiple reading levels are possible, due to its probabilistic nature, the model cannot accommodate multiple interpretations of the same graphical series and will try to infer unique replacement and replication rules. As a result, its automatic transcriptions, based on interpreted human text editions, will inevitably carry a noticeable ratio of false negatives, especially when dealing with heavily ligatured or abbreviated late-medieval and pre-modern scripts where the number of potential readings and expansions increases.

VII DISCUSSION

There are several explanations for these results from the perspectives of paleography and medieval diplomatics. Firstly, the Textualis model has been exclusively trained on cartularies that closely adhere to the bookhand script. Monks who wrote books and copied cartularies were meticulous in their letters layout and avoided embellishments to ensure readability. While there may be occasional ornamentations or stylistic preferences introduced by professional pen-writers, the scribal scripts and their abbreviative systems during the 12th century were largely consistent across different locations (Hasenohr [1998]).

In this context where individual variability is constrained by the script type, the unseen test data align closely with the training data, resulting in models with high generalization capacity but still bearing a notable level of bias. This could explain why updating weights on a robust generalist Textualis model only leads to modest improvements in handwriting recognition. It can be presumed that this model would perform well on book manuscripts, but may not be as effective on original charter collections that exhibit a more diverse graphical apparatus.

The process of gotization in writing, which emerged during the 13th century and continued thereafter, brought about significant changes in terms of extreme angularity, pen tools, and writing speed. Subsequently, the introduction of the Cursive variant, characterized by continuous writing without lifting the pen between characters, resulted in more profound alterations, including a shift in the axis of writing from vertical to horizontal, simplification of the ductus,

and considerable development of ligatures (Poulle [1966], Guyotjeannin et al. [1993]). The cursive style rendered the shape and size of letters more reliant on the individual hand of the scribe, thereby introducing a greater degree of variability. Additionally, with the proliferation of non-monastic orders and the emergence of new written instruments, such as multilingual accounts, fiefs, and registers, open documents with varying handwriting styles and page setups became more prevalent among the laity. Consequently, handwriting recognition (HTR) work on Cursiva necessitates the consideration of a much wider diversity of written records, hands, languages, and documentary typologies, resulting in an increased number and complexity of features and trends that need to be learned by the model. Models trained on such data may encounter underfitting issues, underscoring the importance of collecting and providing new specific information through fine-tuning to mitigate these challenges.

VIII CONCLUSION

In this article, we have presented specialized models for handwriting recognition that are tailored for documentary and serial manuscripts dating from the 12th to the 15th centuries. Our robustness experiments have demonstrated that a single model trained on semi-diplomatic transcriptions is capable of achieving high accuracy ($\geq$85 points without fine-tuning, $\geq$90 with fine-tuning) in recognizing multiple handwriting script families and effectively handling diverse linguistic and documentary registers. Moreover, our training experiments have revealed that even small batches of ground-truth data from varied sources yield results comparable to those obtained from training on large uniform corpora. Furthermore, our findings indicate that fine-tuning of the models gives positive outcomes when dealing with manuscripts that exhibit a high number of variations deviating from the regularity of the script, which is more commonly observed in late-Gothic scripts or late-medieval documentary manuscripts.

IX MODEL REPOSITORIES

The Zenodo repositories containing training logs and models for this work available at:

https://doi.org/10.5281/zenodo.7547438

https://doi.org/10.5281/zenodo.7401833

References

Horst Bunke, Markus Roth, and Ernst Günter Schukat-Talamazzini. Off-line cursive handwriting recognition using hidden markov models. *Pattern recognition*, 28(9):1399–1413, 1995.

Jean-Baptiste Camps, Chahan Vidal-Gorène, and Marguerite Vernet. Handling heavily abbreviated manuscripts: Htr engines vs text normalisation approaches. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 306–316. Springer, 2021.

Alix Chagué. Cremma: Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique. In *Sciences du patrimoine-sciences du texte. Confrontation des méthodes*, 2021.

Julie Claustre and Darwin Smith. e-ndp notre-dame de paris et son cloitre (1326-1504). *Revue Mabillon*, 2022.

Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. Htr-flor: A deep learning system for offline handwritten text recognition. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 54–61. IEEE, 2020.

Matthew Driscoll. Levels of transcription. In *Electronic textual editing*, pages 254–261. Modern Language Association of America, 2006.

Maciej Eder. Mind your corpus: systematic errors in authorship attribution. *Literary and linguistic computing*, 28 (4):603–614, 2013.

John Evershed and Kent Fitch. Correcting noisy ocr: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 45–51, 2014.

Martin-D Glessgen. L'écrit documentaire dans l'histoire linguistique de la france. *La langue des actes. Actes du XIe Congrès international de diplomatique*, 2004.

Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems*, 21, 2008.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2008.

Paul Guérin. *Recueil de documents concernant le Poitou contenus dans les registres de la Chancellerie de France*, volume 11. Société des archives historiques de Poitou, 1881.

Olivier Guyotjeannin, Jacques Pycke, and Benoît-Michel Tock. *Diplomatique médiévale*. Brepols, 1993.

Geneviève Hasenohr. Abréviations et frontières de mots. *Langue française*, pages 24–29, 1998.

Tobias Mathias Hodel, David Selim Schoch, Christa Schneider, and Jake Purcell. General models for handwritten text recognition: Feasibility and state-of-the art. german kurrent as an example. *Journal of open humanities data*, 7(13):1–10, 2021.

Mike Kestemont, Vincent Christlein, and Dominique Stutzmann. Artificial paleography: computational approaches to identifying script types in medieval manuscripts. *Speculum*, 92(S1):S86–S109, 2017.

Gundram Leifert, Tobias Strau, Tobias Gr, Welf Wustlich, Roger Labahn, et al. Cells in multidimensional recurrent neural networks. *Journal of Machine Learning Research*, 17(97):1–37, 2016.

Günter Mühlberger, Johannes Zelger, and David Sagmeister. User-driven correction of ocr errors: combining crowdsourcing and information retrieval technology. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 53–56, 2014.

Ariane Pinche. Guide de transcription pour les manuscrits du xe au xve siècle. 2022. URL https://hal.archives-ouvertes.fr/hal-03697382.

Emmanuel Poulle. *Paléographie des Écritures Cursives en France du XVe au XVIIe*. Librairie Droz, 1966.

Jenna Schoen and Gianmarco E Saretto. Optical character recognition (ocr) and medieval manuscripts: Reconsidering transcriptions in the digital age. *Digital Philology: A Journal of Medieval Cultures*, 11(1):174–206, 2022.

Uwe Springmann, Florian Fink, and Klaus U Schulz. Automatic quality evaluation and (semi-) automatic improvement of ocr models for historical printings. *arXiv preprint arXiv:1606.05157*, 2016.

Dominique Stutzmann, Jean-François Moufflet, and Sébastien Hamel. La recherche en plein texte dans les sources manuscrites médiévales: enjeux et perspectives du projet himanis pour l'édition électronique. *Médiévales. Langues, Textes, Histoire*, 73(73):67–96, 2017.

Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. HOME-Alcar: Aligned and Annotated Cartularies, 2021. URL https://zenodo.org/record/5600884. Zenodo: https://doi.org/10.5281/zenodo.5600884.