Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR)

Matthias Gille Levenson[1]

[1]École Normale Supérieure de Lyon
[1]CIHAM, UMR 5648

Corresponding author: Matthias Gille Levenson , matthias.gille-levenson@ens-lyon.fr

Abstract

This paper introduces a first HTR/OCR open access gold corpus for spanish late medieval sources, based on the allographetic transcription of more than 300 pages of several manuscripts of the *Regimiento de los Prínçipes*, as well as a first set of general transcription and regions/lines segmentation models trained with Kraken. These models are evaluated with in-domain and out-of-domain data.

Keywords

HTR models; HTR dataset; medieval castilian; out-of-domain evaluation; zones and lines segmentation; manuscript transcription; incunabula

I BACKGROUND AND STATE OF THE ART

1.1 Datasets of Spanish medieval and early modern sources

If Handwitten Text Recognition and Optical Text Recognition (from now on HTR and OCR) technology is now mature[1] and can be fully integrated into publishing and philology projects, little work has been done on automatic transcription for medieval and pre-modern Castilian context. Back in 2010, a corpus has been produced for the 16th century, from a manuscript of 1545, the RODRIGO corpus (Serrano et al. [2010]), but there is no open access and documented datasets for earlier periods[2]. Moreover, this initial work is more on the machine learning side than on the philological side and does not yet aim to produce models that can be used in the context of publishing projects or studies of written sources. Thus, this RODRIGO corpus is used in several works presenting architectures for processing handwritten text[3]. It is the machine learning process and architecture that is at the centre of the researchers' interest here, and not the production and publication of a corpus for philological purposes[4].

In 2022, the journal *Historias Fingidas* published a special issue[5] including four articles dealing with

[1]Ocr4all, Transkribus, eScriptorium or Calfa are some examples of automatic platforms that are commonly used in philology: see Reul et al. [2019], Kahle et al. [2017], Kiessling et al. [2019] and Kindt and Vidal-Gorène [2022].

[2]According to the latest datasets survey: see Nikolaidou et al. [2022].

[3]See for instance Romero et al. [2011], Granell et al. [2018], Martínez-Hinarejos et al. [2018], Xamena et al. [2021].

[4]We can also mention the GERMANA corpus, which contains 21.000 lines of a handwritten text produced at the end of the 19th century. This dataset has been created with the same idea of allowing the comparison of architecture and tools of analysis and extraction of the handwritten text. See Pérez et al. [2009].

[5]Bazzaco [2022].

text recognition in a Spanish context (Bazzaco et al. [2022], Blasut [2022], Aranda García [2022], Ayuso García [2022]). The material described in these articles consists of printed sources from the 16th century. Unfortunately, at the time of writing, the data described in these articles is either not accessible, or not freely accessible and adaptable[6].

In conclusion, concerning handwritten sources, there is a lack of open data for the castilian Middle Ages and Early Modern period. In this article I describe the HTR/OCR dataset I have produced using the eScriptorium application (Kiessling et al. [2019]), which consists of ten manuscripts and a 15th century incunabulum, as well as models created with kraken (Kiessling [2019]) along with their performance. The models are tested with in-domain and out-of-domain data which is described and documented. This work is intended to be part of the creation of meta-corpora such as HTR-United (Chagué et al. [2021]), and the dataset will be aggregated into it.

1.2 Description of the text

The training corpus is taken from the *Regimiento de los príncipes*, also known – erroneously, because it does contain parts of the original translation – as the *Glosa castellana al "Regimiento de prínçipes"*, scholarly edited in 1947 by Juan Beneyto Pérez(Beneyto Pérez [2005]). It is a text of political literature, a translation of the *De Regimine Principum* of Giles of Rome (1247-1316), which played an important role in the diffusion of medieval Aristotelianism in Castile and in the tripartition of the practical philosophy into moral, economic and political branches. (Bizzarri [2000]). The transcribed corpus is essentially taken from the last book (political and military matters), and from the first book (moral matters).

II THE DATASET

This section describes the in-domain dataset only; for the out-of-domain dataset, see below (section 4.1.1, page 9 and table 7 page 16).

2.1 Quality of the reproductions

The digital reproductions are generally of good quality[7] and in high definition and the documents are clearly legible. However, one manuscript (manuscript L, figure 1) is a bit more problematic. Some folios are damaged, the image is sometimes a bit blurred; during the digitization process, the page was not correctly lit, so a portion of the image is darkened because of a shadow. This happens on large parts of each page (more than half of them, about 30 images). The manuscript J also shows imperfections caused by ink-bleeding.

2.2 Description of the corpus

All manuscripts are known to have been produced in the 15th century. The incunabula was printed in 1494. The corpus is a good basis for the production of a general corpus for the late medieval period (14th-15th or even 13th-15th centuries), given its extension and the diversity of hands represented. Even though the train corpus focuses on a restricted area, the resulting models produce good predictions, as I will show below. It is a good start for general models that can handle a large diversity of hands and *scripta*[8]. However, it has some limitations, since it is thematically and chronologically biased, because it is produced from a single work.

[6]Because of the no-derivative clause of the licence (CC BY-NC-ND 4.0) in the case of Bazzaco et al. [2022].

[7]The images are in jpeg or png. Their resolution is variable: the DPI goes from 72 to 650. Mean DPI: 284; median DPI: 184.

[8]See Hodel et al. [2021], which shows the efficency of general HTR models.

Figure 1: Manuscripts L, fol. 387v, and manuscript J, fol. 388r. Universidad de Salamanca (Spain), Biblioteca General Histórica.

2.2.1 Hands, types and location

Table 1 shows the distribution of the corpus. This corpus is composed of an incunabula, three main manuscripts (A, L and S), and seven secondary manuscripts. I have identified fourteen different hands in the handwritten corpus, some more important in volume than others (A_1, L_1 S_3, S_4). A large number of elements in the corpus are more poorly represented but allow for a first generalisation of the model to a large number of distinct hands. According to Derolez classification (Derolez [2003]), the hands of manuscripts A, L, S and U correspond to professional gothic textualis scripts. B, G, J and M can be described as gothic hybrida scripts: they have loopless ascenders, their straight ſ and f go under the line. Manuscript Q, represented by two folios (244 lines of very tight handwriting), is a cursive gothic[9]. R shows some cursive features too. In this manuscript, some d have loops, and the f and ſ go under the line. Figure 11, in the appendix, shows some transcribed lines. The types used for the production of the incunabula by Meinardo Ungut and Estanislao Polono are classified as 97G (Martín Abad and Moyano Andrés [2002]). This set of types is used between 1494 and 1500 in a large number of editions[10].

[9]This hand is added to reinforce the robustness of the model, and the dataset will be extended with more cursive *scripta* in a future version.

[10]For other incunabula produced at this time, see [Martín Abad and Moyano Andrés, 2002, 112-121].

Hand[11]	Location	Folios	Number of pages	Number of lines
A_1	II, 2, 36 – III, 3, 23	fols. 236v-273v	75	6771
A_2	*Definición de nobleza*[12]	fols. 274r-275r	3	227
B_1	III, 2, 36* – III, 3, 2*	fols. 302r-304r	5	415
G_1	III, 2, 36* – III, 3, 1*	fols. 407r-409v	6	408
J_1	III, 2, 36* – III, 3, 1	fols. 386v-389r	6	404
L_1	III, 2, 27* – III, 2, 35*	fols. 365r-376r	24	1778
	III, 3, 1* – III, 3, 8*	fols. 377r-391r	30	2179
M_1	II, 2, 3*	fol. 228r	1	32
Q_1	III, 3, 1 – III, 3, 2*	fols. 141v-142r	2	244
R_1	III, 2, 36* – III, 3, 2*	fols. 245v-247r	4	316
S_1	I, 1, 1 – I, 1, 2*	fols. 1r-2v	4	378
S_2	I, 1, 2* – I, 1, 5*	fols. 3r-6r	7	654
S_3	I, 1, 5* – I, 2, 11*	fols. 6v-27r	42	3923
S_4	III, 3, 1 – III, 3, 23	fols. 176r-197r	44	4031
U_1	III, 2, 36* – III, 3, 2*	fols. 171v-172v	3	277
Total (manucripts)			**256**	**22,037**
Z	III, 3, 1* – III, 3, 23	fols. 220r-249v	62	5556
Total (manucripts + incunabula)			**318**	**27,593**

Table 1: Corpus distribution by hand or identified *scripta*. The title appears if it is not the *Regimiento de los prínçipes*. Otherwise the book, part and chapter are indicated. An asterisk means the chapter is incomplete. All pages are fully transcribed.

2.2.2 Words and forms count

It is not easy to count the vocabulary of a non-standardised corpus: hyphenation is never indicated, except in the incunabulum, not in a systematic way. If we ignore the hyphenations and count the forms starting and ending a line as full words, the corpus contains 181,240 words, and 28,309 unique forms[13].

2.2.3 Zones and lines annotation principles

The model is adapted to one column or two column texts without glosses (or with occasional marginal glosses). There is no heavily glossed manuscript in the corpus. The zoning is consistent with the Segmonto shared vocabulary (Gabay et al. [2021]). Each document was validated with HTRVX (Clérice and Pinche [2021b]). Five classes are used for zoning and two for lines labelling (see table 2). When dealing with drop capitals, I did not distinguish between realized capitals and spaces left blank (figure 2). The distinction will be made in a future version of the dataset.

Regarding the lines' typology, the class `HeadingLine:rubric` aims to represent any heading line

[11] I reuse the acronyms used in Díez Garretas et al. [2003]: A, mss 289, Fundación Lázaro Galdiano, Madrid; B: ms. 26.I.5, Instituto Valencia de don Juan, Madrid; G: ms. II/215, Biblioteca Real, Madrid; J: mss. 2097, Biblioteca Universitaria, Salamanca; L: ms. 2709, Biblioteca Universitaria, Salamanca; M: ms. h.I.8, Biblioteca del Escurial, San Lorenzo de el Escorial; Q: ms. K.I.5, Biblioteca del Escurial, San Lorenzo de el Escorial; R: ms. 332/131, Biblioteca Universitaria, Sevilla; S: ms. 251, Biblioteca de Santa Cruz, Valladolid; U: ms. 482/2, Rosenbach Foundation, Philadelphia. The incunabulum Z corresponds to the exemplar INC/901 of the Biblioteca Nacional de España.

[12] Díez Garretas and Dietrick [2007].

[13] This metric is accurate for this particular transcription task, because the recognizer works line-per-line.

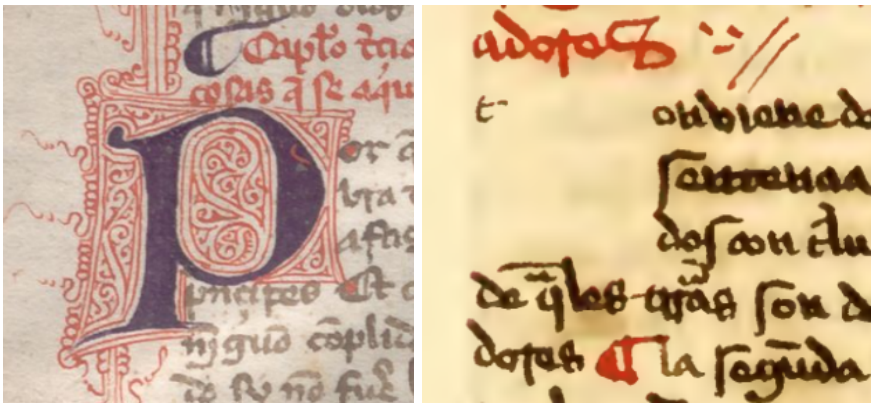

Figure 2: Left: a drop capital, S, fol. 3v; right: a space left blank for a drop capital, R, fol. 247r. Both are annotated as `DropCapitalZone`.

with or without change of ink. In my annotation, I consider the line as a semantic element and not a physical one, when and only when a physical element marks a difference (changes in ink colour, font or character size). For example, when a rubric starts in the middle of the line, I create two different lines (figure 3, p. p. 5). The line-fillers are ignored, and reading order is not corrected.

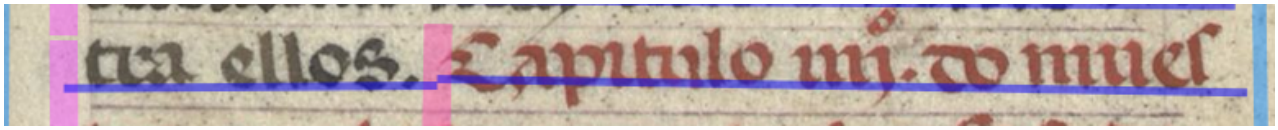

Figure 3: Two "semantic" lines segmented on the same physical line. The first line is typed as `DefaultLine`, and the second one as `HeadingLine:rubric`. Ms A, fol. 243r.

Zone type	Quantity
MainZone	671
RunningTitleZone	203
NumberingZone	148
DropCapitalZone	131
MarginTextZone	119
QuireMarksZone	29
Total	**1301**

Line type	Quantity
DefaultLine	26,775
HeadingLine:rubric	818
Total	**27,593**

Table 2: Zones and line quantification.

2.3 What is annotated and what is not

Marginal glosses are included in `MarginTextZone`'s and fully transcribed, except for one large marginal gloss on manuscript A (folio 240r), which deviates too much from the layout of the rest of the corpus and would not be useful for training the baselines and transcription models[14]. Interlinear glosses and additions outside the baseline are ignored. Strikethrough text is transcribed as if it was normal text, if it is legible. Otherwise, it is ignored[15].

[14] This marginal gloss has been included in a `MarginZoneText` only.

[15] It will be added between double square brackets "[[...]]" in a future version of the dataset as recommended in Pinche [2022a].

III TRANSCRIPTION NORMS

Peter Robinson and Elisabeth Solopova's manual shows that the choice of the level of granularity in the transcription is very problematic and cannot be answered definitively[16]. The choice of keeping some allographs only (see below) may be debatable, but it is a first step in the production of allographetic recognition models, and the corpus can be modified and refined later. The detailed description of the encoding method should make it possible to identify the problematic elements and to correct them if necessary. I consider that the important thing is to properly document the classification choices.

An allographic transcription could help to detect collective scribal profiles, to identify groups of copists who share the same *usus scribendi*, and could even help dating the manuscripts (Camps [2016]). HTR methods are helpful for this kind of transcriptions that would require a lot of time if they were to be realized by hand. In particular, it makes possible to highlight the positional value of some allographs (the "r" and the "s" in particular[17]), information that is lost in case of an overly standardized transcription. Moreover, recent HTR engines allow to extract a large volume of palaeographic data, such as allographs (or grapheme, depending on the chosen transcription standards). This is of great support for *scripta* description[18]. For instance, all the graphemes' plates (figures 4 to 8) have been produced automatically with a script of my own that uses kraken predictions and the eScriptorium API[19].

Figure 4: Realizations of "s" (left) and "z" (right) in manuscript A.

3.1 Allographs

Listed below are the rules I followed to transcribe the corpus. Ligatures are not taken into account.

Capital letters. I make a distinction between capitals and small letters.

d allographs:. Straight "d" and insular "ꝺ" were distinguished.

[16] Robinson and Solopova [1993].

[17] See Pinche et al. [2022], though the authors normalize the allographs.

[18] We can cite the DigiPal project for example: Stokes [2011].

[19] https://github.com/matgille/escriptorium_graphemes_retrieval. As the script uses kraken 4.20.0 to identify the coordinates of the graphemes, it is dependant on the accuracy of the model, and can identify wrong graphemes: it needs good models to produce accurate results.

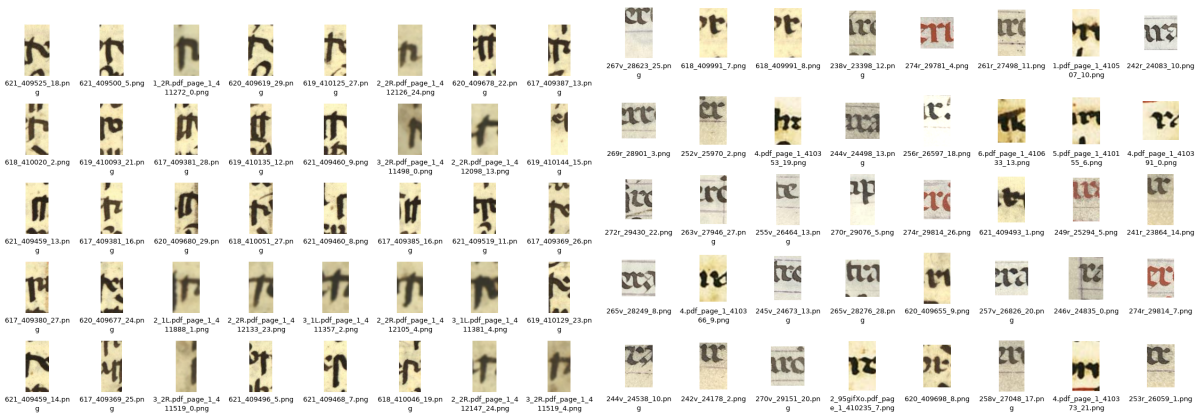

Figure 5: Realizations of long and short legged “r” in various manuscripts.

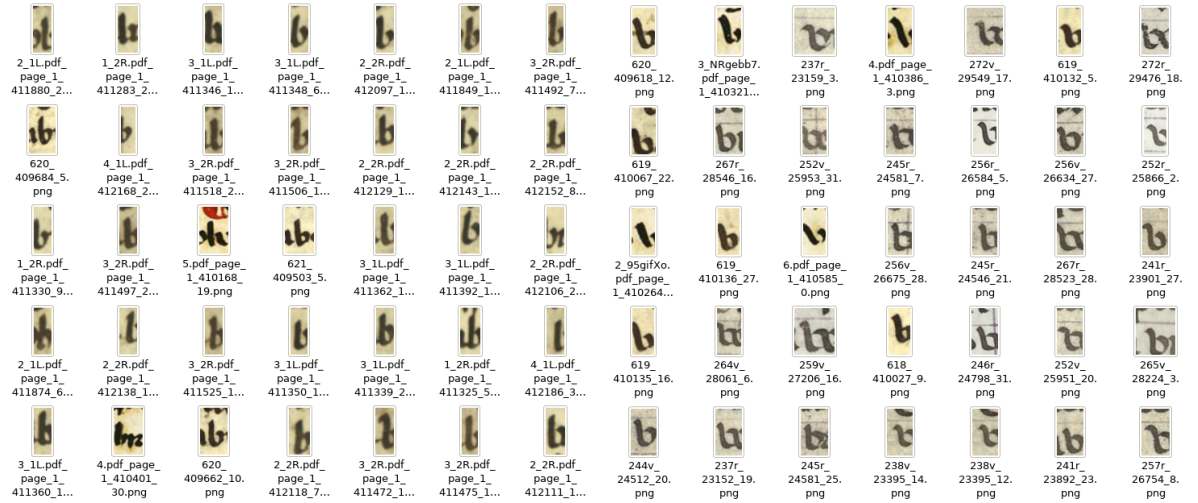

Figure 6: Realizations of “b” (left) and “v” (right) in different manuscripts.

i allographs:. I make a distinction between i and j used for the short i “i” and the descending i “j”. The ascending and descending J’s have been noted as “J” and were distinguished from the descending “j”s.

n allographs:. N with descending legs “ŋ” (rare) and common “n” were distinguished.

r allographs:. Descending legged “ɾ”, straight “r”, and round “ɾ” were distinguished.

s allographs:. “ſ” and the round “s” were distinguished.

t allographs:. Straight “t” and insular “ꞇ” were distinguished.

y allographs:. Dotted “ẏ” and non dotted y were distinguished. The vast majority of y are dotted.

z allographs:. A round and an elongated form of the “z” were distinguished: “z” shown in figure 4[20], and “ʒ”.

3.2 Abbreviations

When it is posible, the subscribed dot – used essentially in manuscript A – (figure 7), and the tilde, which appears only in the printed edition, were distinguished. As for other abbreviations, the main ones are: the trema, the grave accent (which will be expanded as “ri” or “i”) or the hook (figure

[20]This realization of the letter in this corpus seems specific to the Spanish context is described in [Millares Carlo, 1983, 212]; it is the most common in the corpus.

8), and some superscript letters (i, e, s). Regarding the place of the abbreviation sign, as much as possible, it is transcribed where it appears in the image.

Figure 7: Some superscript macrons (left) and dots (right) in manuscripts A and B.

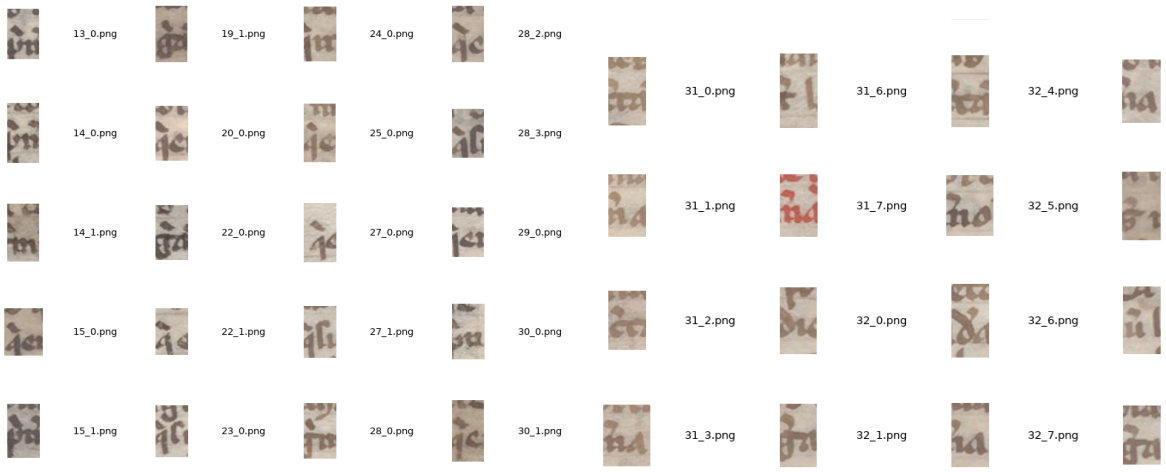

Figure 8: Realizations of grave accent (left) and hook (right) in manuscript S.

3.3 Punctuation

I distinguished between the comma ",", the solidus "/", the full stop ".", the middot "·", the colon (used extensively by the incunabulum) ":", the vertical line "|" which is used in manuscript G as a semantic and syntactic pause and as an indication of direct speech.

3.4 Public/private domain character uses

Almost the entire corpus has been transcribed using the public domain of unicode, except for a few characters: the ser "ꝛ", that only appears in the manuscript Q, about a dozen times. The same applies to a few specific punctuation marks, in the out-of-domain corpus in particular (*punctus elevatus* "⁖", for example).

3.5 Spaces

There is only one class of spaces in the dataset. Word segmentation is not normalized, meaning that when a modern space is missing in the source, it is not added in the transcription. This task has clear hermeneutical limitations: the description of segmentation is sometimes difficult (see figure 9). The normalization of the segmentation could be performed by the transcription engine, but provided that (as for now) the current HTR tools – at least, Kraken – work line by line, it is not possible to perform accurate hyphenation using one of them[21].

Figure 9: A case of ambiguous word segmentation. “laſegunda” or “la ſegunda”?

3.6 Data uniformity

The file `table.csv` was produced by Choco-Mufin[22], and helped uniformizing the transcription and the encoding by making sure no similar characters were used to transcribe a grapheme. It lists all the characters in the dataset.

IV MODELS AND RESULTS

The results and models described are produced with NFD (which splits the diacritic and the main form) and NFC (which combines the diacritic with the main form) normalizations. A seed value of 1234 has been chosen for splitting, training and testing phases. The other hyperparameters are taken from Ariane Pinche's work[23]; data augmentation; batch size: 16; lag: 20; learning rate: 0.0001; architecture (using the VGSL norm): ‘[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do]’.

4.1 In-domain and out-of-domain tests sets

The training dataset is about 80% of the global dataset described above. Three datasets are used to test the models during and after the training:

- the dev set, which corresponds to 10% of the global dataset and is used by kraken for best model selection,
- the in-domain test set, which corresponds to 10% of the global dataset, and is used for testing the best model;
- an out-of-domain test set, with more thant 30 pages (about the same size than the preceding dataset) taken from other castilian manuscript sources, used to test the accuracy of the best model with unseen data.

4.1.1 Out-of-domain

The out-of-domain contains 34 pages, that is, around 1500 lines. The texts proposed are of varying genres: political literature, historiography, legal literature, but also didactic and narrative poetry.

[21] The detection of hyphenated words at the end of the line, which is in fact a tokenisation task (if we merge the lines two by two), must be performed by an ad-hoc tool. See for example Gille Levenson [2021], a program largely based on the work of (Clérice [2020]) from the point of view of the problem to be solved and of the network architecture. It also offers the possibility of detecting the hyphenations.

[22] Clérice and Pinche [2021a].

[23] See Pinche [2022b].

The corpus is precisely described in table 7. The pages are taken from the 13th, 14th and 15th centuries manuscripts, with a predominance of the 15th century. Some scripts are quite different from the ones of the in-domain dataset : the *Liber Regum*, for example, which can be described as a *cursiva libraria/formata*[24] (see figure 10).

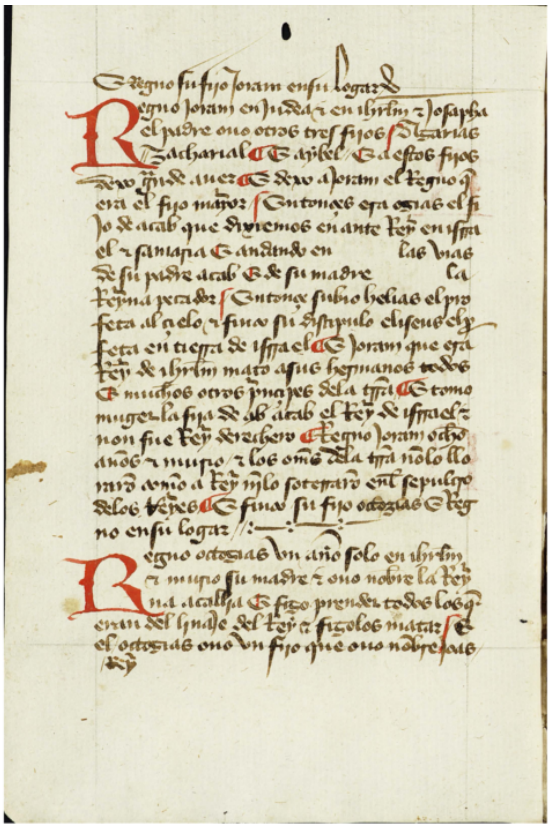

Figure 10: *Liber regum*, Mss BG 2011, fol. 6v.

4.2 Zones and lines recognition

Results are shown in table 3 and table 4. The regions model is acceptable, but the baseline model seems to be less performant than the default model (`blla.mlmodel`). It is possible that the test and out-of-domain dataset are too small to produce reliable results, especially for less frequent zones.

4.3 Transcription

The global in-domain accuracy is 96.30% (see table 5). The models are usable to pre-annotate documents, in order to produce new data faster. As for scholarly editing, they probably need some prior finetuning to be perfectly adapted to a particular hand. Moreover, but they should be usable, for distant reading: the current accuracy is good enough to produce good results in stylometry,

[24] [Derolez, 2003, pl. 81].

		In-domain			Out-of-domain		
Category	Class	Pixel Acc.	IOU	Objects	Pixel Acc.	IOU	Objects
regions	MainZone	0.976	0.952	71	0.961	0.903	53
regions	NumberingZone	0.998	0.102	17	0.999	0.000	13
regions	RunningTitleZone	0.997	0.681	21	0.999	0.266	6
regions	QuireMarksZone	1.000	0.000	4	1.000	0.000	1
regions	DropCapitalZone	0.999	0.620	18	0.993	0.257	24
regions	MarginTextZone	0.998	0.158	9	0.991	0.100	11
Mean Accuracy		0.996			0.993		
Mean IOU		0.564			0.441		
Frequency-weighted IOU		4.512			3.527		
Class-independent Region IOU		0.943			0.895		

Table 3: Zones recognition results.

		In-domain			Out-of-domain		
Category	Class	Pixel Acc.	IOU	Objects	Pixel Acc.	IOU	Objects
baselines	DefaultLine	0.940	0.632	2768	0.904	0.296	1099
baselines	HeadingLine:rubric	0.998	0.500	81	0.999	0.303	32
Mean Accuracy		0.980			0.972		
Mean IOU		0.486			0.248		
Frequency-weighted IOU		1.942			0.991		
Class-independent Region IOU		0.631			0.298		

Table 4: Lines recognition results.

	Allographetic data		
Unicode Norm.	**Model**	**In-domain**	**Out-of-domain**
NFC	`allogr-nfc.mlmodel`	96,30%	93,15%
NFD	`allogr-nfd.mlmodel`	96,14%	92,94%

Table 5: Global results with in and out-of-domain test sets.

for instance (Eder [2013]). It is a first basis for refining a model with little data using transfer learning[25], and more particularly by finetuning the model (see below, p. 12).

4.3.1 Unicode normalization

Table 5 (p. 11) and 6 (p. 12) show that unicode normalization has an influence on the final results[26], and that it should be taken into account when producing datasets, to speed up data production. I recommend testing each model on a few pages (4-5) first, and then deciding which normalization works best with the target corpus. Regarding the difference in results between NFD and NFC normalizations, my hypothesis was that there could be a link between the normalization, global results and the number of abbreviations in the corpus, as unicode normalization has an impact in terms of number of output classes[27]. However I did not find a statistical link between the difference in results and the number of abbreviations (retrieved with the number of combining characters for

[25] Yosinski et al. [2014].

[26] It is important to note that the same unicode normalization has been used for training and evaluating the models.

[27] See Vidal-Gorène [2023].

Text	Nb. of lines	Columns	NFC	NFD
Partidas	117	2	91.65	90.52
Fuero	138	2	89.88	88.29
Zifar	75	2	90.87	89.98
Ordenamiento	68	2	91.54	91.77
Regum	65	1	85.52	85.1
Mugeres	58	1	93.94	94.21
Privilegio	28	1	94.56	94.18
$Calila_1$	32	1	96.14	95.88
Sem	68	1	92.86	92.99
Tesoretto	25	1	89.96	92.22
Confesional	70	1	93.96	94.01
$Calila_2$	28	1	95.4	95.33
Cronica	165	2	96.02	96.35
Flores	48	1	94.12	94.19
Geografia	57	1	91.07	91.45
Mocedades	177	2	93.98	94.22
Monteria	34	1	93.94	93.52
Morales	73	2	93.59	93.12
Soliloquio	70	1	95.17	94.44
Vita	48	1	96.47	97.26
Average			93.03	92.96

Table 6: Detailed results with out-of-domain test set, ordered by production date (increasing order).

each text). A more systematic study should be conducted on the matter.

4.3.2 Finetuning

A table with finetuned results is attached in appendix (table 8, p. 17). As can be seen, the general model is a good base that can be easily finetuned to produce models fitted to a particular script, within the limits of the original model (which is unsurprisingly less accurate on cursive scripts). The individual results are not as solid as the global statistics and should be taken really carefully, as some of the finetuned models seem to have overfitted the data, however they can give an idea of the ability of the general model to be finetuned and to adapt to new scripts with quite a small amount of data (see in particular the 14th century examples: *Partidas*, *Fuero real*, *Zifar*, *Ordenamiento*, and the *Liber Regum*).

V CONCLUSIONS AND FURTHER WORKS

This dataset constitutes a first step towards a generalist model of transcription of castilian manuscripts. It lacks source diversity but the model already performs well on out-of-domain, and finetuning allows it to be strengthened at little cost. The production of a graphematic corpus from the corpus described in this paper is underway and will be available on zenodo. It is planned to extend the generic and temporal diversity of the dataset.

Kraken is really helpfull for the production of results, which are easily publishable and comparable[28].

[28]The version used to perform the trainings and the tests is 4.2.0: link.

To my view, this is missing in eScriptorium[29], which allows the user to perform trainings, but is limited in terms of hyperparameters tuning and of results production. However, it should be noted that this lack of parameterisation is the result of a conscious choice on the part of the development team. eScriptorium is a very efficient data production platform, but I would recommend to use kraken directly to train the models.

VI DATASETS, MODELS, SCRIPTS

The annotations are freely available in XML (ALTO) format under the CC-BY-SA-NC license. The dataset and the scripts used to produce and test the models[30] are available on a zenodo repository[31], with the permission of the libraries holding the original documents. The final NFC and NFD models are published alongside the data. Each pair of finetuned models is available too, but their performance on unseen data is not guaranteed. The results of the tests have been saved in the `logs` directory.

VII ACKNOWLEDGEMENTS

I would like to thank the reviewers for their reading and comments which helped improving this paper. Many thanks to Ariane Pinche for her help, her attentive reading of the paper and her plentiful advice. I'm responsible for all inaccuracies, errors and imprecisions that may be found in it. I thank Daniel Meharg for his proofreading and corrections; thanks to the Centre Blaise Pascal (CBP) of the ENS of Lyon for granting me the access to a powerful machine to perform the trainings. I thank the following libraries for granting me the publication rights of the images: the Biblioteca General Histórica of Salamanca and the Biblioteca de la Fundación Lázaro Galdiano.

References

Nuria Aranda García. Humanidades digitales y literatura medieval española: la integración de transkribus en la base de datos comedic. *Historias Fingidas*, (Special Issue 1):127–149, 2022. ISSN 2284-2667. doi: 10.13136/2284-2667/1107. URL https://historiasfingidas.dlls.univr.it/article/view/1107.

Manuel Ayuso García. Las ediciones de arnao guillén de brocar de beclar transcritas con ayuda de transkribus y ocr4all: creación de un modelo para la red neuronal y posible explotación de los resultados. *Historias Fingidas*, (Special Issue 1):151–173, 2022. ISSN 2284-2667. doi: 10.13136/2284-2667/1102. URL https://historiasfingidas.dlls.univr.it/article/view/1102.

Stefano Bazzaco. (ed.) *Historias fingidas*, "Digital Humanities e studi letterari ispanici". (Special Issue 1), 2022. ISSN 2284-2667.

Stefano Bazzaco, Ana Milagros Jiménez Ruiz, Ángela Torralba Ruberte, and Mónica Martín Molares. Sistemas de reconocimiento de textos e impresos hispánicos de la edad moderna. la creación de unos modelos de htr para la transcripción automatizada de documentos en gótica y redonda (s. xv-xvii). *Historias Fingidas*, (Special Issue 1): 67–125, 2022. ISSN 2284-2667. doi: 10.13136/2284-2667/1190. URL https://historiasfingidas.dlls.univr.it/article/view/1190.

Juan Beneyto Pérez, editor. *Glosa Castellana al "Regimiento de Príncipes" de Egidio Romano*. Centro de Estudios Políticos y Constitucionales, 2005.

Hugo Óscar Bizzarri. Fray Juan García de Castrojeriz receptor de Aristóteles. *Archives d'Histoire Doctrinale et Littéraire du Moyen Âge*, 67:225–236, 2000.

Giada Blasut. Los modelos de htr silves1549_bne y spanish gothic como herramientas de la labor ecdótica. *Historias*

[29] I work on a local server on the `develop` branch. Last pulled commit: `82d3e8d46f72afbe`.

[30] A main python script (`produce_results.py`) has been used to perform all training and test phases. The script `count_vocab.py` helped to produce stats on the number word forms in the corpus. `count_lines.py` was used to count the lines in the corpus. `upload_api.py` and `upload_on_request.py` have been used to easily upload files to my eScriptorium instance, making use of its API.

[31] https://doi.org/10.5281/zenodo.7386489.

Fingidas, (Special Issue 1):175–193, 2022. ISSN 2284-2667. doi: 10.13136/2284-2667/1178. URL https://historiasfingidas.dlls.univr.it/article/view/1178.

Jean-Baptiste Camps. *La Chanson d'Otinel: édition complète du corpus manuscrit et prolégomènes à l'édition critique*. PhD thesis, 2016.

Alix Chagué, Thibault Clérice, and Laurent Romary. HTR-United: Mutualisons la vérité de terrain!, 2021. URL https://hal.science/hal-03398740.

Thibault 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/pdf.

Thibault Clérice and Ariane Pinche. Choco-Mufin, a tool for controlling characters used in OCR and HTR projects, 9 2021a. URL https://github.com/PonteIneptique/choco-mufin.

Thibault Clérice and Ariane Pinche. HTRVX, HTR Validation with XSD, 9 2021b. URL https://github.com/HTR-United/HTRVX.

Albert Derolez. *The Palaeography of Gothic Manuscript Books: From the Twelfth of the Early Sixteenth Century*. Cambridge University Press, 2003. ISBN 978-0-521-80315-1 978-0-521-68690-7.

María Jesús Díez Garretas and Deborah Anne Dietrick. Otra definición de nobleza de Perafán de Ribera. In *Actas Del XI Congreso Internacional de La Asociación Hispánica de Literatura Medieval:(Universidad de León, 20 al 24 de Septiembre de 2005)*, pages 469–479. Servicio de Publicaciones, 2007.

María Jesús Díez Garretas, Isabel Acero-Durántez, José Manuel Fradejas-Rueda, and Deborah Dietrick Shmithbauer. *Los Manuscritos de La Versión Castellana Del "De Regimine Principum" de Gil de Roma*. Universidad de Valladolid, 2003.

Maciej Eder. Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4): 603–614, 2013. ISSN 0268-1145. doi: 10.1093/llc/fqt039. URL https://doi.org/10.1093/llc/fqt039.

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, and Claire Jahan. SegmOnto: Common vocabulary and practices for analysing the layout of manuscripts (and more). In *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, 2021.

Matthias Gille Levenson. Boudams-like segmenter, 2021. URL https://github.com/matgille/boudams_like_tokenizer.

Emilio Granell, Edgard Chammas, Laurence Likforman-Sulem, Carlos-D. Martínez-Hinarejos, Chafic Mokbel, and Bogdan-Ionuţ Cîrstea. Transcription of Spanish Historical Handwritten Documents with Deep Neural Networks. *Journal of Imaging*, 4(1):15, 2018. ISSN 2313-433X. doi: 10.3390/jimaging4010015. URL https://www.mdpi.com/2313-433X/4/1/15.

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.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter 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, pages 19–24. IEEE, 2017.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. 2019. URL https://dev.clariah.nl/files/dh2019/boa/0673.html.

Benjamin Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, 2019. doi: 10.1109/ICDARW.2019.10032.

Bastien Kindt and Chahan Vidal-Gorène. An Automated Process for Ancient Armenian or Other Under-Resourced Languages of the Christian East. *Armeniaca*, 2022.

Julián Martín Abad and Isabel Moyano Andrés. *Estanislao Polono*. Universidad de alcalá edition, 2002.

Carlos-D. Martínez-Hinarejos, Emilio Granell-Romero, and Verónica Romero-Gómez. Comparing Different Feedback Modalities in Assisted Transcription of Manuscripts. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 115–120, 2018. doi: 10.1109/DAS.2018.13.

Agustin Millares Carlo. *Tratado de paleografía española*. Madrid : Espasa-Calpe, 1983. ISBN 978-84-239-4986-1.

Konstantina Nikolaidou, Mathias Seuret, Hamam Mokayed, and Marcus Liwicki. A Survey of Historical Document Image Datasets, 2022.

Ariane Pinche. Guide de transcription pour les manuscrits du xe au xve siècle, 2022a.

Ariane Pinche. HTR Models and genericity for Medieval Manuscripts, 2022b.

Ariane Pinche, Frédéric Duval, and Jean-Baptiste Camps. Création de modèle (s) HTR pour les documents médiévaux en ancien français et moyen français, Xe-XIVe siècles, 2022.

Daniel Pérez, Lionel Tarazón, Nicolás Serrano, Francisco Castro, Oriol Ramos Terrades, and Alfons Juan. The GERMANA Database. In *2009 10th International Conference on Document Analysis and Recognition*, pages 301–305,

2009. doi: 10.1109/ICDAR.2009.10.
Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences*, 9(22):4853, 2019. ISSN 2076-3417. doi: 10.3390/app9224853. URL https://www.mdpi.com/2076-3417/9/22/4853.
Peter Robinson and Elizabeth Solopova. Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. 1993. doi: 10.5281/zenodo.4050360. URL https://zenodo.org/record/4050360.
Veronica Romero, Nicolas Serrano, Alejandro H Toselli, Joan Andreu Sanchez, and Enrique Vidal. Handwritten Text Recognition for Historical Documents. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pages 90–96, 2011.
Nicolas Serrano, Francisco Castro, and Alfons Juan. The RODRIGO Database. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/477_Paper.pdf.
Peter A. Stokes. Digital Resource and Database for Palaeography, Manuscripts and Diplomatie. *Gazette du livre médiéval*, 56(1):141–142, 2011. doi: 10.3406/galim.2011.1991. URL https://www.persee.fr/doc/galim_0753-5015_2011_num_56_1_1991.
Chahan Vidal-Gorène. La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées. *Programming Historian en français*, 2023. doi: 10.46430/phfr0023.
Eduardo Xamena, Héctor Emanuel Barboza, and Carlos Ismael Orozco. End-to-end platform evaluation for Spanish Handwritten Text Recognition. *Ciencia y Tecnología*, 21:81–95, 2021.
Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

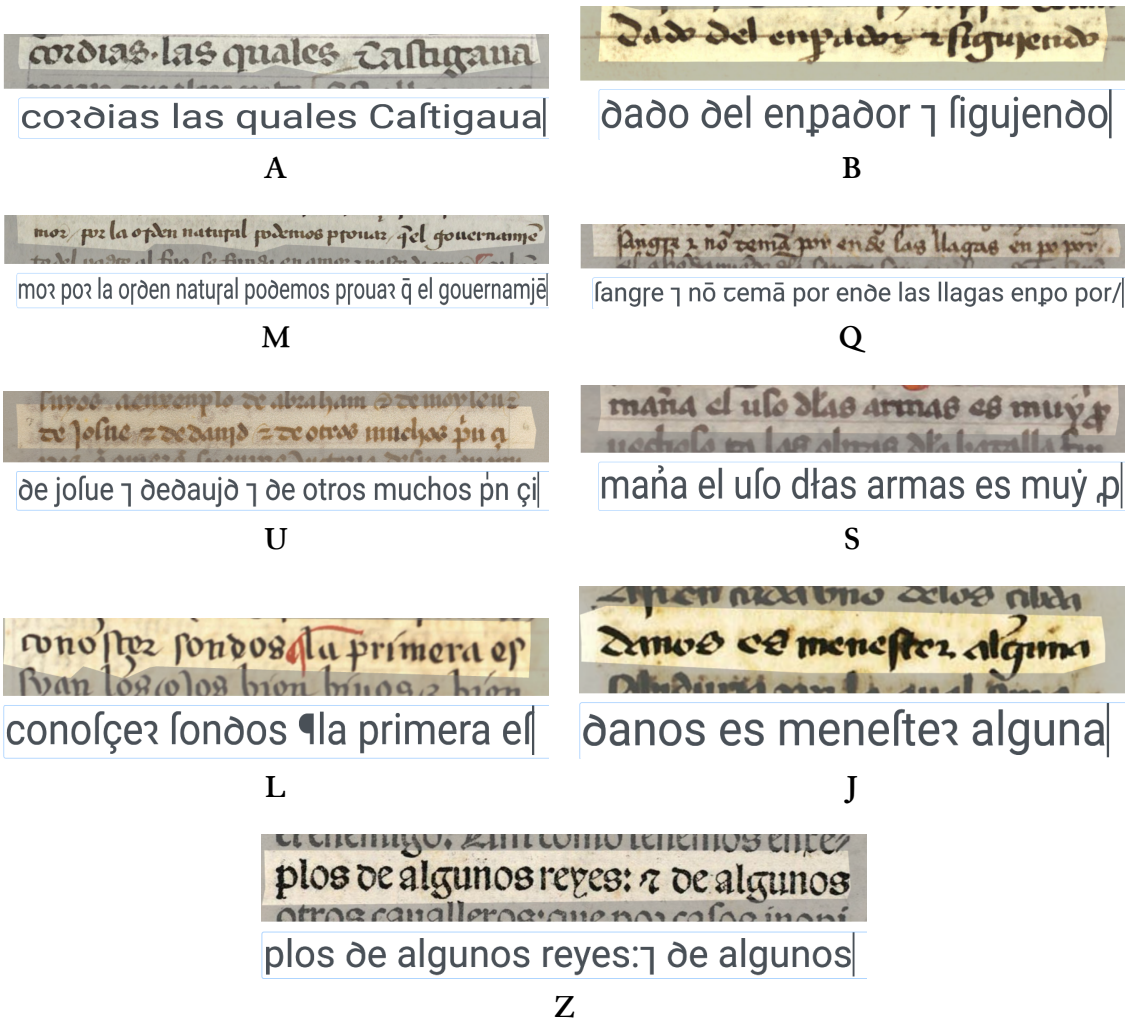

Figure 11: Some ground truths taken from the in-domain dataset (eScriptorium application).

Short title	Title	Author	Date	Library & Signature	Loc.	Pages	Lines	link
Partidas	Siete Partidas	Alfonso X de Castilla	1300-1310	Escurial RBME Z-I-12	fol. 15v	1	117	link
Fuero	Fuero Real	ø	1325	Salamanca BG 2673	fols. 34r-35v	2	138	link
Zifar	Libro del Caballero Zifar	ø	14th cent.	BnF, Espagnol 36	fol. 154v	1	75	link
Ordenamiento	Ordenamiento de Alcalá	ø	14th cent.	Escurial RBME Z-III-9	fol. 2r and fol 12r	2	68	link
Regum	Liber Regum	ø	1401-1450	Salamanca BG 2011	fols. 6v-7r	2	56	link
Mugeres	Libro de las clara y virtuosas mugeres	Álvaro de Luna	1446	Salamanca BG 2654	fols. 88r-89v	2	58	link
Privilegio	Carta de privilegios otorgados a la ciudad de Sevilla	ø	1447	Bodleian Library MS. Span. d. 2/1	fol. 2r	1	28	link
Calila$_1$	Calila e Dimna	ø	1467	Escurial RBME X-III-4	fol. 7r	1	32	link
Sem	Versos al rey D. Pedro del rabí don Sem Tob	Sem Tob de Carrión	1460-1480	Escurial RMBE b-IV-21	fols. 77v-79r	4	68	link
Tesoretto	Tesoreto	Brunetto Latini	1467	Bodleian Library MS. Span. d. 1	fol. 49r	1	25	link
Confesional	Confesional	Alfonso de Madrigal	1472	BNE MSS/4183	fol. 3v-4r	2	70	link
Calila$_2$	Calila e Dimna	ø	15th cent.	Escurial RBME h-III-9	fol. 5r	1	28	link
Crónica	Cronica de España	Alfonso X de Castilla	15th cent.	BnF. Espagnol 12	fol. 111v and 142v	2	165	link
Flores	Flores de las leyes	ø	15th cent.	RMBE b-IV-15	fols. 6v-7r	2	48	link
Geografía	Tratado de geografía	ø	15th cent.	Salamanca BG 2086	fols. 18v-19r	2	57	link
Mocedades	Mocedades de Rodrigo	ø	15th cent.	BnF, Espagnol 12	fol. 196r and 198r	2	177	link
Montería	Libro de la montería de Alfonso XI	ø	15th cent.	BnF, Espagnol 218	fol. 7v	1	34	link
Morales	Morales de sant Gregorio	ø	15th cent.	Escurial RBME b-II-11	fol. 11r	1	73	link
Soliloquio	Soliloquio de San Agustín	ø	15th cent.	Escurial RMBE a-II-17	fols. 90v-91r	2	60	link
Vita	De vita beata	Séneca (Alonso de Cartagena)	15th cent.	Escurial RBME T-III-5	fol. 2v-3r	2	48	link
Total						**34**	**1425**	

Table 7: Out-of-domain test set, ordered by production date.

			Global model		Finetuned models		Δ Finetuning	
Text	**Nb. of lines**	**Columns**	**NFC**	**NFD**	**NFC**	**NFD**	Δ NFC	Δ NFD
Partidas	117	2	91.65	90.52	95.06	99.2	3.41	8.68
Fuero	138	2	89.88	88.29	94.46	96.9	4.58	8.61
Zifar	75	2	90.87	89.98	96.82	96.85	5.95	6.87
Ordenamiento	68	2	91.54	91.77	93.79	96.81	2.25	5.04
Regum	65	1	85.52	85.1	93.32	90.99	7.8	5.89
Mugeres	58	1	93.94	94.21	96.84	96.62	2.9	2.41
Privilegio	28	1	94.56	94.18	98.48	99.09	3.92	4.91
$Calila_1$	32	1	96.14	95.88	96.72	96.82	0.58	0.94
Sem	68	1	92.86	92.99	94.03	96.57	1.17	3.58
Tesoretto	25	1	89.96	92.22	95.23	95.52	5.27	3.3
Confesional	70	1	93.96	94.01	98.78	99.12	4.82	5.11
$Calila_2$	28	1	95.4	95.33	96.3	97.16	0.9	1.83
Cronica	165	2	96.02	96.35	99.2	97.36	3.18	1.01
Flores	48	1	94.12	94.19	97.32	95.61	3.2	1.42
Geografia	57	1	91.07	91.45	96.41	96.62	5.34	5.17
Mocedades	177	2	93.98	94.22	98.65	98.82	4.67	4.6
Monteria	34	1	93.94	93.52	94.53	94.62	0.59	1.1
Morales	73	2	93.59	93.12	94.84	95.14	1.25	2.02
Soliloquio	70	1	95.17	94.44	98.02	99.11	2.85	4.67
Vita	48	1	96.47	97.26	99.01	97.75	2.54	0.49
Average			93.03	92.95	96.39	96.83	3.36	3.88

Table 8: Finetuned models with a small amount of data, tested on the train set only, ordered by production date (increasing).