CremmaLab project

Transcription guidelines and HTR models for French medieval manuscripts

Jean-Baptiste Camps [1,2] Ariane Pinche [1]

[1]Centre Jean Mabillon, École nationale des chartes | PSL [2]Venice Center for Digital and Public Humanities, Univ. Ca'Foscari

Ancient Documents and handwritten text recognition

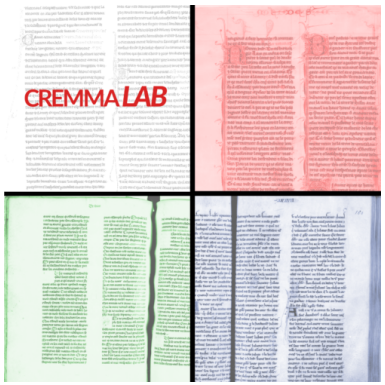
23 juin 2022

Table of Contents

CremmaLab Projects

ENC collective initiatives on HTR, in partnership with **DIM MAP** and **INRIA** :

- CREMMA (Consortium Reconnaissance d'Écriture Manuscrite des Matériaux Anciens)
- CREMMALab.
  - Seminar: **Creation of HTR model(s) for medieval documents in Old French and Middle French between the 10th and 14th centuries** (October 2021-March 2022) : https://cremmalab.hypotheses.org/seminaire-creation-de-modeles-htr
  - Conference : **Ancient documents and automatic recognition of handwriting** (23 and 24 June 2022)



CremmaLab Projects: cremma-medieval

CremmaLab Projects: cremma-medieval

Cremma-medieval is a GitHub repository. It provides HTR ground truth (*Kraken+eScriptorium*) in open access from 13 different manuscripts and different projects :

- 263 different XML files, 21656 lines of transcription
- Data conforming to transcription standards and to *SegmOnto*'s controlled vocabulary for layout description
- Data quality control process :
  - *HTRVX* for Alto XML and respect of the *SegmOnto* ontology
  - *Choco-Mufin* for the uniform use of characters with a table of reference characters
- Data described, shared and made visible through *HTR-United* catalog
- Releases of HTR models (also declared in *Zenodo* and accessible via *Kraken*)

Table of Contents

Birth of the guidelines

The guidelines are the result :

- of the need to create consistent HTR data for french manuscripts to :
  - build shareable and **reusable ground truth data** sets to minimise the collective cost
  - produce **generic models** useful to the scientific community
- of a reflection carried out at the Ecole nationale des chartes during the years 2021 and 2022

Guidelines are available here :
https://hal.archives-ouvertes.fr/hal-03697382

Main principles

Our recommendations are not intended to constitute a definitive transcription or a final edition. The text produced may require the implementation of a multi-step protocol.

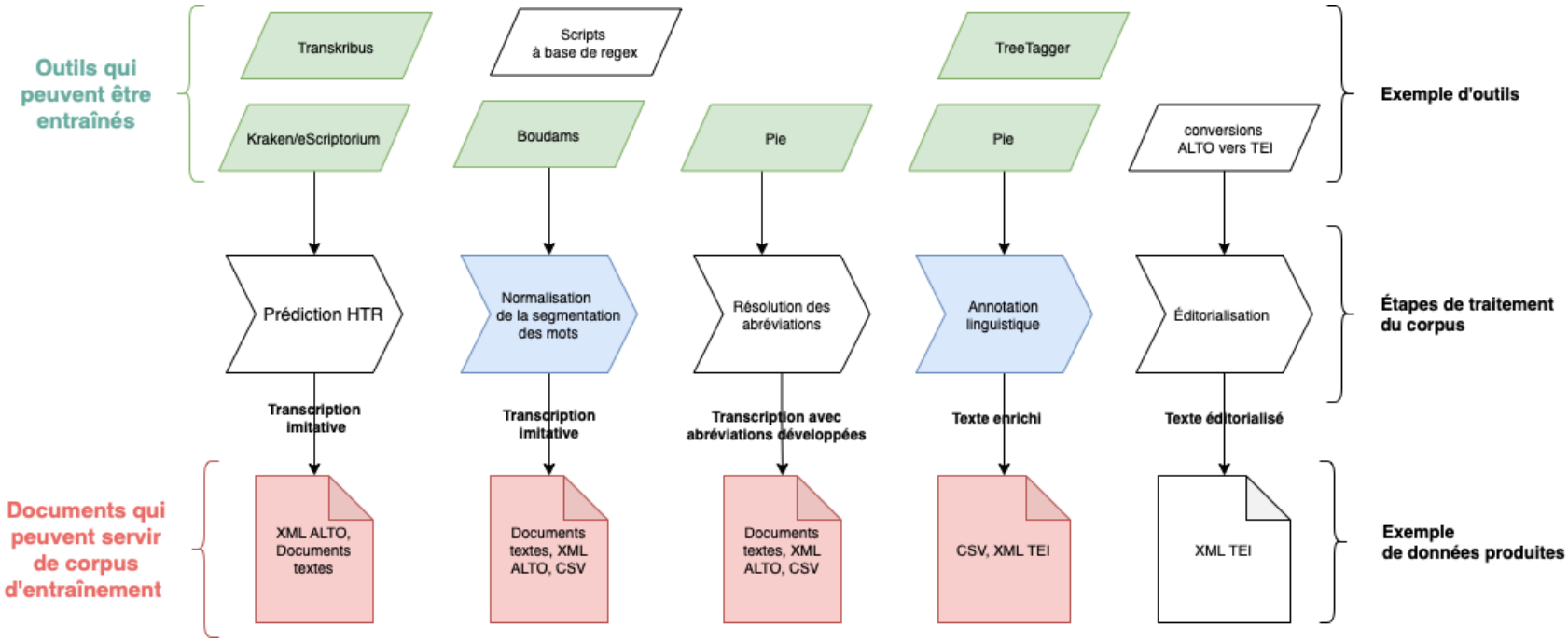


Figure: Example of text production from the HTR output

Main principles

Our system of transcription :

- Preserves abbreviations

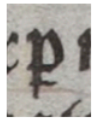
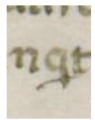
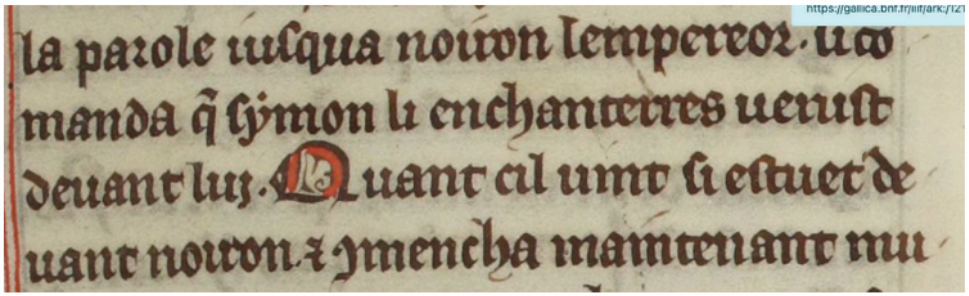
LATIN SMALL LETTER P WITH STROKE (*p barré droit*)		ꝑ	U+A751
LATIN SMALL LETTER P WITH FLOURISH (*p barré courbe*)		ꝓ	U+A753
LATIN SMALL LETTER Q WITH DIAGONAL STROKE (*q barré*)		ꝗ	U+A759
TIRONIAN SIGN ET (*abréviation tironienne de « et »*)		⁊	U+204A
DIVISION SIGN (*abréviation de « est »*)		÷	U+00F7

Figure: Examples of abbreviation representations

Main principles

- Preserves the spelling of text: no normalisation of "u" and "v", or "i" and "j", no normalisation of capital letters
- Reduces each letter form to a standardised representation
- Reduces the complexity of medieval punctuation:
  single sign = ". "and double sign = ":".



BnF, fr. 412

On transcrira :

« la parole iusqua noiron lempereor. li co-
manda q̃ symon li enchanterrres uenist
deuant lui. Quant cil uint si estuet de-
uant noiron ⁊ 9mencha maintenant mu- »

Figure: Example of transcription

Table of Contents

Bicerin model presentation

- Bicerin is an HTR model for medieval French manuscripts.
- It has been trained on 13 manuscripts written between the 13th and 15th centuries mostly in Gothic script (cremma-medieval dataset)
- The last release of the model : https://github.com/HTR-United/cremma-medieval/releases/tag/1.1.0

Figure: Prediction on a manuscript "in-domain" (BnF, ms., fr. 25516, 13^{e} s.)

Bicerin and performance

Releases of HTR models trained on cremma-medieval dataset

- 0.0.1 Arabica, accuracy 89,19% (21/06/21)
- 1.0.0 Bicerin, accuracy 95,49% (21/07/13)
- 1.1.0 Bicerin, accuracy 95,30% (22/06/20)
- (Work in progress) Cortado, a model mixing cremma-medieval dataset with early prints (15^{e} s.) from Gallicorpora project, accuracy 95.54%

Bicerin and performance

Out-of-domain tests, Genève, Comites Latentes 102, 14th c.

Figure: Prediction from Arabica model

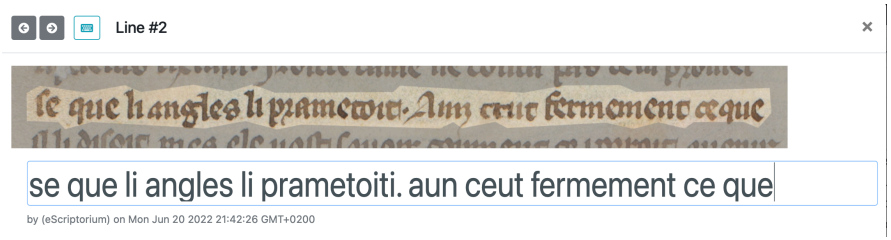


Figure: Prediction from Bicerin 1.0.0 model

Figure: Prediction from Bicerin 1.1.0 model

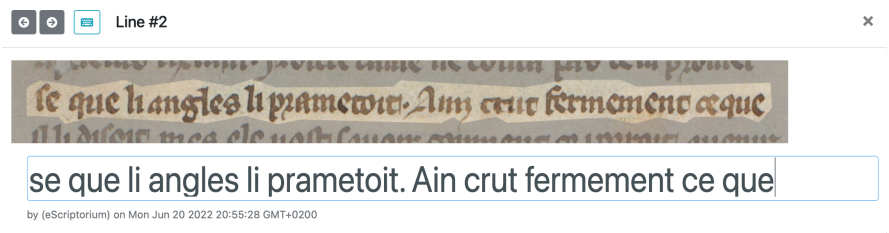


Figure: Prediction from Cortado model

Bicerin and performance

Out-of-domain tests with scores

Three really different documents

Figure: Genève, Comites Latentes 102, 14th c.

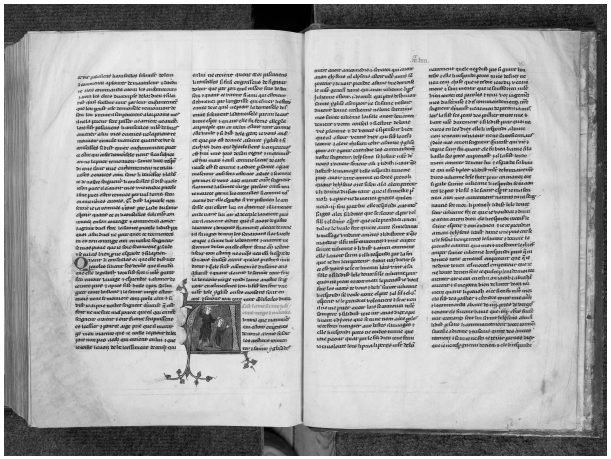


Figure: Chantilly, ms. 734, 14th century

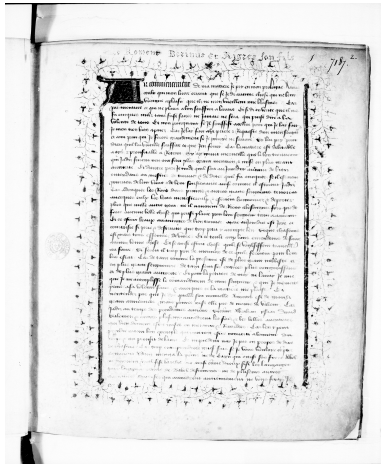


Figure: BnF, ms., fr. 777, 15th century

Bicerin and performance

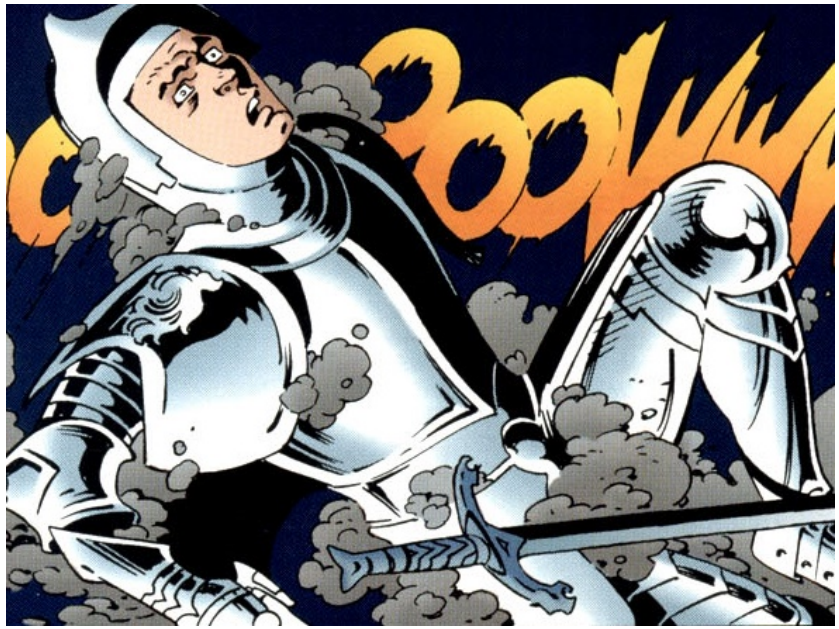
Some "out-of-domain" tests with scores

Source	Bicerin 1.0.0	Bicerin 1.1.0	Cortado
Chantilly, , ms. 734, 14th c.	93.02%	92.58%	93.20%
Genève, Com. Lat. 102, 14th c.	93.84%	95.43%	96.57%,
BnF, ms., fr. 777, 15th c.	43.39%	54.93%	54.36%

Table of Contents

A Thousand Years of French Literary Fictions (1050-...)

Constitute a corpus of chivalric romances
(*chansons de geste*, *romans*, *mises en prose*...)
and their descendants
from the 11th to the 21st century.



Karl der Grosse du Stricker, Saint-Gall, Stadtbibliothek (Vadiana) ms. 302, fol. 52v; Roland s'éveille de la mort pour défendre l'olifant, ill. de Michael Moorcock, Stormbringer (cycle d'Elric de Melniboné), 1965.

JB Camps, Pierre-Carl Langlais, Nicolas Baumard, Olivier Morin, 'From Roland to Conan: First results on the corpus of French literary fictions (1050-1920)', DH Tokyo 2022.

State and Current Goals

- 16th-21st centuries covered in good part;
- Sources: Gallica, Y2 (cotes Nicolas Clément); Google Books.
- Need to re-OCR and process the documents
- Models provided by Jahan & Gabay: excellent results.

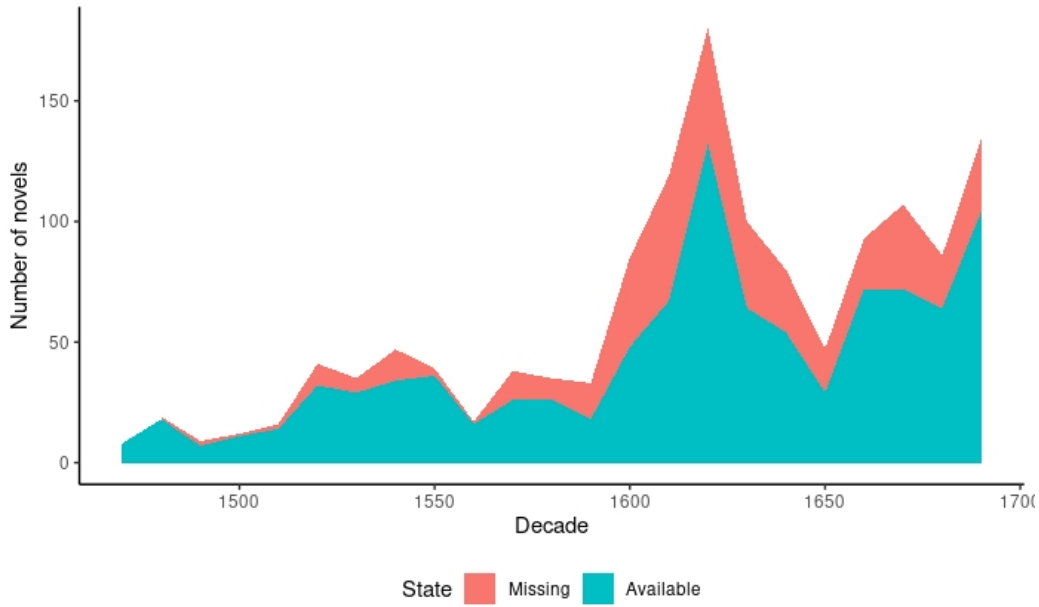


Figure: Number of novels and completeness of the corpus (compared to the Y2 catalogue). Completeness is high, but docs outside of the catalogue have been spotted.

P.C. Langlais, 'Fictions littéraires de Gallica', 2021, doi: 10.5281/zenodo.4751204.
C. Jahan and S. Gabay. 'Ocr17+ - layout analysis and text recognition for 17th c. french prints, 2021' https://github.com/e-ditiones/OCR17plus

17th Century

Test set

- 60 pages,
- randomly selected,
- from Gallica and GB docs,
- (transcr. JB Camps, O. Morin, P.C. Langlais)

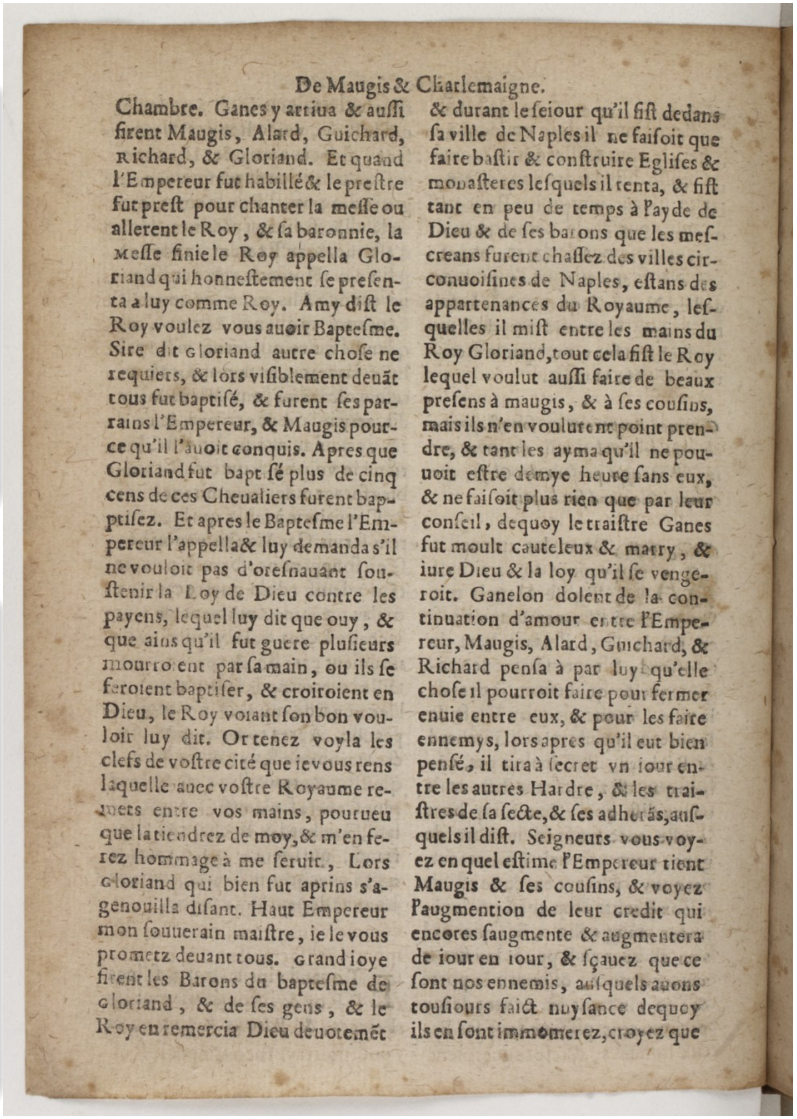
Results

Model `fralatnfc_63.mlmodel`

global acc. 96.77% (96.96 ign. spaces)

gallica acc. 99.44% (99.51 ign. spaces)

Google Books acc. 92.34% (92.71 ign. spaces)

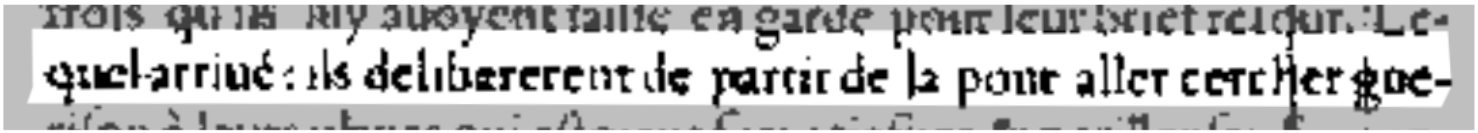


De Maugis & Charlemaigne.

Chambre. Ganes y arriua & auſſi firent Maugis, Alard, Guichard, Richard, & Gloriand. Et quand l'Empereur fut habillé & le preſtre fut preſt pour chanter la meſſe ou allerent le Roy, & ſa baronnie, la meſſe finie le Roy appella Gloriand qui honneſtement ſe preſenta a luy comme Roy. Amy diſt le Roy voulez vous auoir Bapteſme. Sire dit Gloriand autre choſe ne requiers, & lors viſiblement deuãt tous fut baptiſé, & furent ſes parrains l'Empereur, & Maugis pource qu'il l'auoit conquis. Apres que Gloriand fut baptiſé plus de cinq cens de ces Cheualiers furent baptiſez. Et apres le Bapteſme l'Empereur l'appella & luy demanda s'il ne vouloit pas d'oreſnauant ſouſtenir la Loy de Dieu contre les payens, lequel luy dit que ouy, & que ains qu'il fut guere pluſieurs mourroient par ſa main, ou ils ſe feroient baptiſer, & croiroient en Dieu, le Roy voiant ſon bon vouloir luy dit. Or tenez voyla les clefs de voſtre cité que ie vous rens laquelle auec voſtre Royaume remets entre vos mains, pourueu que la tiendrez de moy, & m'en ferez hommage à me ſeruir. Lors Gloriand qui bien fut aprins s'agenouilla diſant. Haut Empereur mon ſouuerain maiſtre, ie le vous prometz deuant tous. grand ioye firent les Barons du bapteſme de Gloriand, & de ſes gens, & le Roy en remercia Dieu deuotemẽt & durant le ſeiour qu'il fiſt dedans ſa ville de Naples il ne faiſoit que faire baſtir & conſtruire Egliſes & monaſteres leſquels il renta, & fiſt tant en peu de temps à l'ayde de Dieu & de ſes barons que les meſcreans furent chaſſez des villes circonuoiſines de Naples, eſtans des appartenances du Royaume, leſquelles il miſt entre les mains du Roy Gloriand, tout cela fiſt le Roy lequel voulut auſſi faire de beaux preſens à maugis, & à ſes couſins, mais ils n'en voulurent point prendre, & tant les ayma qu'il ne pouuoit eſtre demye heure ſans eux, & ne faiſoit plus rien que par leur conſeil, dequoy le traiſtre Ganes fut moult cauteleux & marry, & iure Dieu & la loy qu'il ſe vengeroit. Ganelon dolent de la continuation d'amour entre l'Empereur, Maugis, Alard, Guichard, & Richard penſa à par luy qu'elle choſe il pourroit faire pour fermer enuie entre eux, & pour les faire ennemys, lors apres qu'il eut bien penſé, il tira à ſecret vn iour entre les autres Hardre, & les traiſtres de ſa ſecte, & ſes adherãs, auſquels il diſt. Seigneurs vous voyez en quel eſtime l'Empereur tient Maugis & ſes couſins, & voyez l'augmention de leur credit qui encores ſaugmente & augmentera de iour en iour, & ſçauez que ce ſont nos ennemis, auſquels auons touſiours faict nuyſance dequoy ils en ſont immomerez, croyez que

17th c. errors

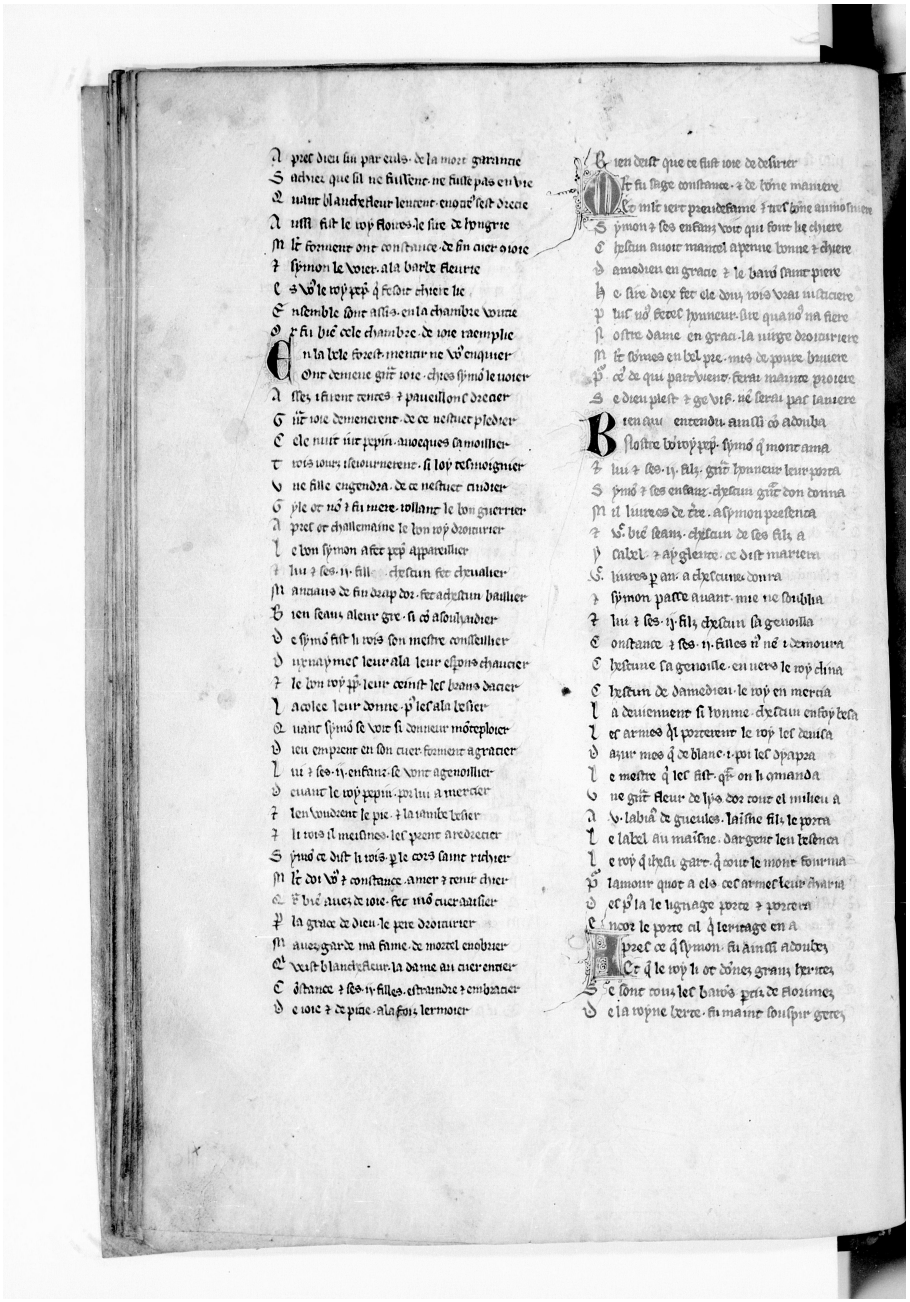
Errors	Correct	Generated
58	SPACE	
54		SPACE
45	e	
36	t	
33		e
30	i	
30	u	
29		t
27	r	

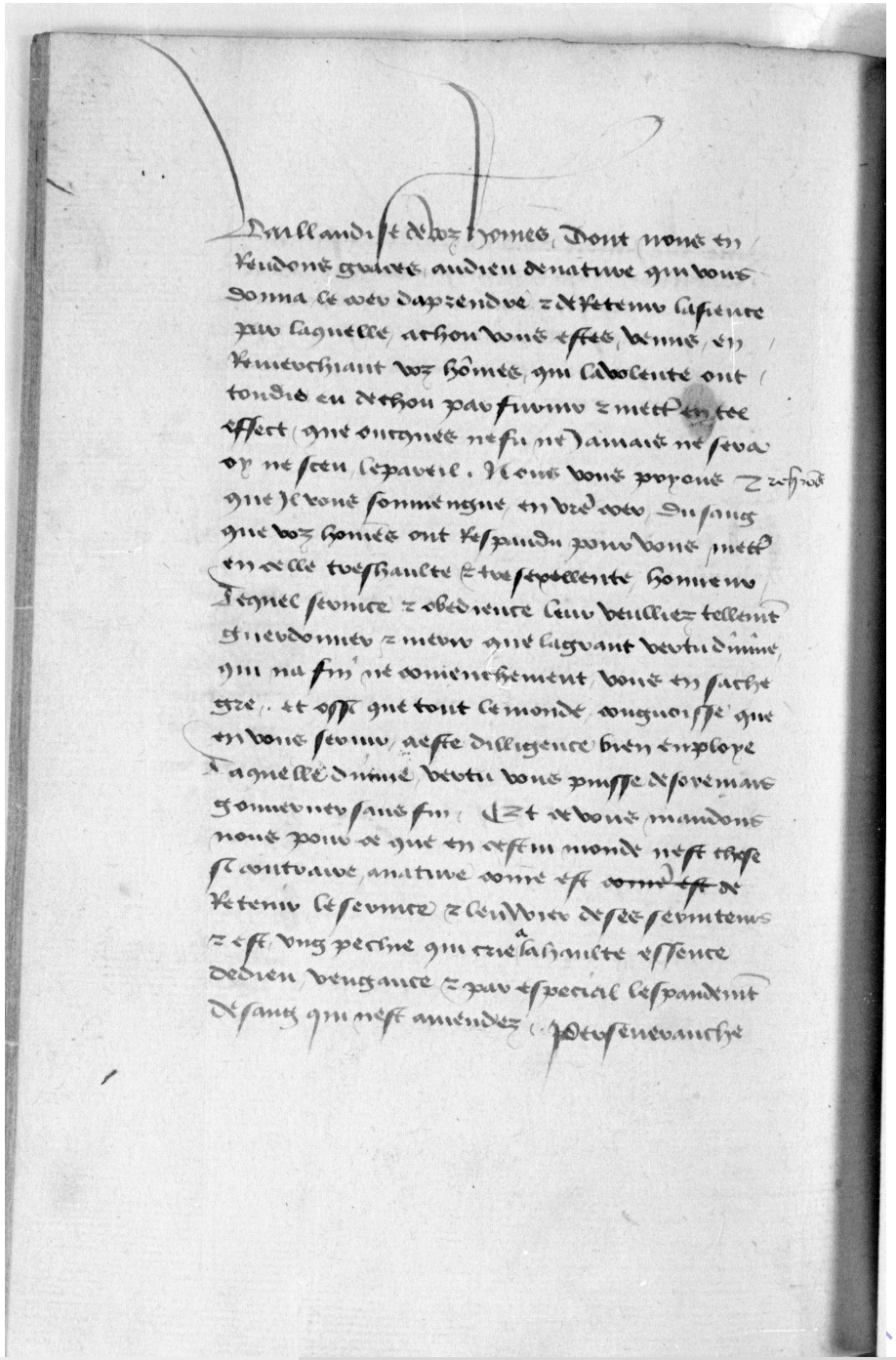


qrel arriué: ils delibererent de partir de la pout aller cercher gœe¬

Medieval documents

- Medieval manuscripts in gothic scripts
- and early prints (incunabula and early 16th c. gothic prints)
- textualis, cursiva, ...
- sample of 30 views (1 to 4 cols),
- from Gallica (uneven qualities),
- transcr. by Svetlana Yatsyk.





Current scores on medieval data (evolving)

- accuracy 93.95%, 95.52% ignoring spaces.

Errors	Correct	Generated
291		SPACE
210	SPACE	
125	.	
59	COMBINING TILDE	
41	i	
41	t	c
35	COMBINING VERTICAL TILDE	
34	r	
28	r	i
26	n	u
25	e	c
23	s	
21		
20		COMBINING TILDE
20	e	
19	t	
18		l
17		;

Table of Contents

References

CHAGUÉ, Alix, CLÉRICE, Thibault et CHIFFOLEAU, Floriane, *HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages*, 2021, [En ligne: https://htr-united.github.io/index.html].
CLÉRICE, Thibault et PINCHE, Ariane, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, 2021, [En ligne: https://github.com/PonteIneptique/choco-mufin].
CLÉRICE, Thibault et PINCHE, Ariane, *HTRVX, HTR Validation with XSD*, 2021, [En ligne: https://github.com/HTR-United/HTRVX].
GABAY, Simon, CAMPS, Jean-Baptiste, PINCHE, Ariane, [et al.], « SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more) », *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland, 2021, [En ligne: https://hal.archives-ouvertes.fr/hal-03336528].
GABAY, Simon, PINCHE, Ariane, LEROY, Noé, [et al.], « Données HTR incunables du 15e siècle », éds. Alix Chagué et Thibault Clérice, *HTR United*, 2022 [En ligne: https://github.com/Gallicorpora/HTR-incunable-15e-siecle].
KIESSLING, B., TISSOT, R., STOKES, P., [et al.], « EScriptorium: An Open Source Platform for Historical Document Analysis », *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 2019, p. 19-19.
PINCHE, Ariane, « CREMMA Medieval, an Old French dataset for HTR and segmentation », éds. Alix Chagué et Thibault Clérice, *HTR United*, 2021, [En ligne: https://github.com/HTR-United/cremma-medieval].
PINCHE, Ariane, *CREMMALAB | Constitution de corpus en ancien français pour l'HTR*, 2022, [En ligne: https://cremmalab.hypotheses.org/].
WILLS, Tarrin, « The Medieval Unicode Font Initiative », *Medieval Unicode Font Initiative*, février 2016, [En ligne: https://skaldic.org//m.php].