

Teaching the Machine How to Read Medieval Manuscripts

Ariane Pinche

▶ **To cite this version:**

Ariane Pinche. Teaching the Machine How to Read Medieval Manuscripts. Master. France. 2021. hal-03259075

HAL Id: hal-03259075
https://hal.science/hal-03259075

Submitted on 13 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Teaching the Machine How to Read Medieval Manuscripts

Ariane Pinche, Univ. Jean Moulin Lyon 3 et École nationale des chartes, UMR 5648 (CIHAM)

Seminar *Teaching the Machine How to Read Medieval Manuscripts*: May 5, 2021.

From digital born editing to automatic text recognition

- XML TEI edition : scientific edition, linguistic annotation, paleographic edition
- Building searchable corpus for distant reading
- To enrich large corpora, artificial intelligence's use has increased
- To acquire a lot of text faster, we use automatic text recognition

How to acquire a large amount of textual data?

- Traditional method: Transcription of witnesses according to scientific criteria: allographic, graphematic, normalised transcription.
- Downloading texts from online corpora
- Computer-assisted transcription using an optical character recognition (OCR) or handwritten text recognition (HTR) tools.

OCR and HTR

OCR	HTR
Old technology	New technology
Performance : - Caracter Error Rate below 2 %, work only on print documents	Caracter Error Rate between 5 and 10 %, work on handwritten documents
- Tools : Abby (adobe), but commercial, no share code / tesseract 4 (free, share code)	- Tools : Transkribus (commercial, no share code) or Kraken (free, share code)
Working system : Generic models per language	Working system : AI needs training.

Using HTR in a research project: new questions arising

- Creating and sharing of training datasets for manuscripts
  - Define transcription methods
  - Creating homogeneous data sets
- Implementing new pipelines for correcting transcriptions of editions
- Evaluating HTR models

Kraken performance : use case

Model train on manuscript BnF, fr. 412 :

https://gallica.bnf.fr/ark:/12148/btv1b84259980/f237.item.zoom

Train : 10 folios

Evaluation : 27 folios, 95,961 characters

Accuracy : 95.38%

error type	Number	error/line
Insertions	1,883	0.76
Deletions	725	0.29
Substitutions	1,823	0.73

eScriptorium presentation

- Transcription interface
- Train text segmentation model
- Train HTR model
- Segmentation and HTR service

Documentation :
https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en

Some references

GABAY, Simon, CLÉRICE, Thibault et REUL, Christian, « OCR17: Ground Truth and Models for 17th c. French Prints (and hopefully more) », 2020, https://hal.archives-ouvertes.fr/hal-02577236.

KIESSLING, Benjamin, TISSOT, R., STOKES, P. « 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.

KIESSLING, Benjamin, « Kraken - an Universal Text Recognizer for the Humanities », Ultrecht, CLARIAH, 2019, https://dev.clariah.nl/files/dh2019/boa/0673.html.

STUTZMANN, Dominique, KERMORVANT, Christopher, VIDAL, Enrique,« Handwritten Text Recognition, Keyword Indexing, and Plain Text Search in Medieval Manuscripts », 2018, https://dh-abstracts.library.cmu.edu/works/6324.

Thank you for your attention !