CATMuS-Medieval: Créer un jeu de données multilingue pour un modèle générique de reconnaissance automatique de texte

Ariane Pinche [1,2]

[1]CNRS, [2]CIHAM

3 mai 2024

Table of Contents

ATR and Humanities (2020-...)

- ATR is a well-mastered task from a computer science perspective.
  - Nowadays, with models that can achieve a Character Error Rate (CER) between 8% and 2% for manuscripts, "from a computer science point of view, the recognition of handwriting seems to be a resolved task." Hodel, Schoch, Schneider, and Purcell 2021.
- Emergence of intuitive platforms: eScriptorium and Transkribus.
- ATR is becoming a common step in more and more research projects, see DH and TEI conference programs.

Why Use ATR in Humanities ?

- To accelerate the text acquisition phase. Prediction can be useful for:
  - serving as a basis for editing: high precision level, exceeding 95% accuracy
  - providing raw text: medium precision level, between 90% and 95%
  - serving as a basis for quantitative analysis: low precision level, exceeding 80% (see Maciej Eder. "Mind Your Corpus: Systematic Errors in Authorship Attribution". In: *Literary and Linguistic Computing* 28.4 (Dec. 1, 2013), pp. 603–614. DOI: 10.1093/llc/fqt039)

ATR and Humanities (2000-2020)

- Handwritten texts on historical documents presents unprecedented challenges:
  - Irregular writing
  - special characters
  - Graphical and/or dialectal variations

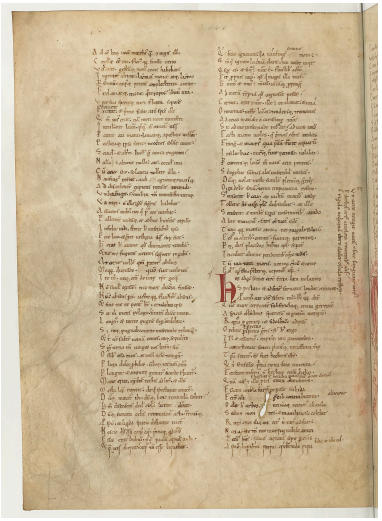


Figure: BnF, Latin, 8001, 13th century

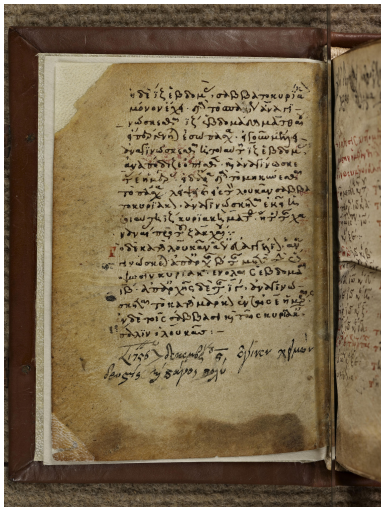


Figure: Strasbourg, ms. 1.916, 13th century

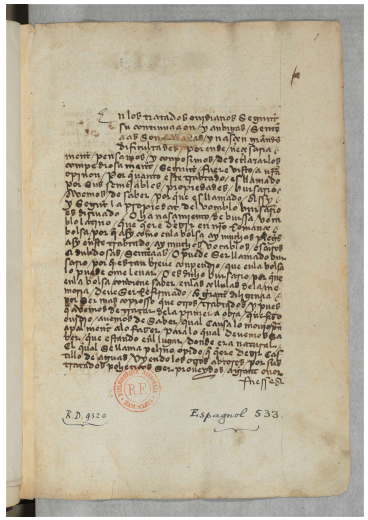


Figure: BnF, Espagnol, 533, 15th century

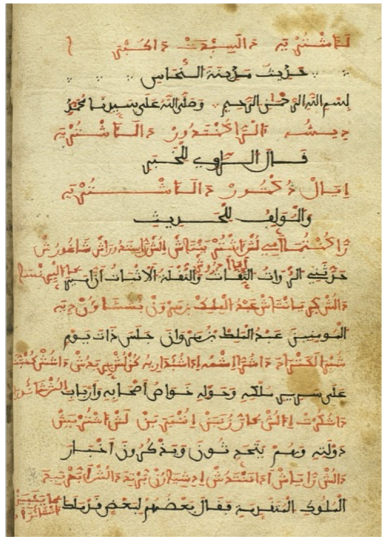


Figure: FDHCA, L536, 15th century

Towards the Creation of "Large-Scale" ATR Models

- Matthias Gille Levenson. "Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR)". In: *Journal of Data Mining and Digital Humanities* (2023). DOI: 10.46298/jdmdh.10416. URL: https://zenodo.org/records/8340483
- Ariane Pinche. "Generic HTR Models for Medieval Manuscripts The CREMMALab Project". In: *Journal of Data Mining & Digital Humanities* (2023). URL: https://univ-lyon3.hal.science/hal-03837519/
- Generic model from Transkribus: Medieval_Scripts_M2.4

Table of Contents

CATMuS Project Overview

- CATMuS stands for Consistent Approaches to Transcribing ManuScripts
- An international initiative involving collaborators from Europe and North America: France, Italy, Switzerland, Canada.
- Provides guidelines, datasets, and models for transcription of medieval manuscripts.
- To know more about the project, see Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, et al. "CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts". In: *DH2024*. ADHO. Washington DC, United States, Aug. 2024. URL: https://inria.hal.science/hal-04346939

How to Transcribe Manuscripts?

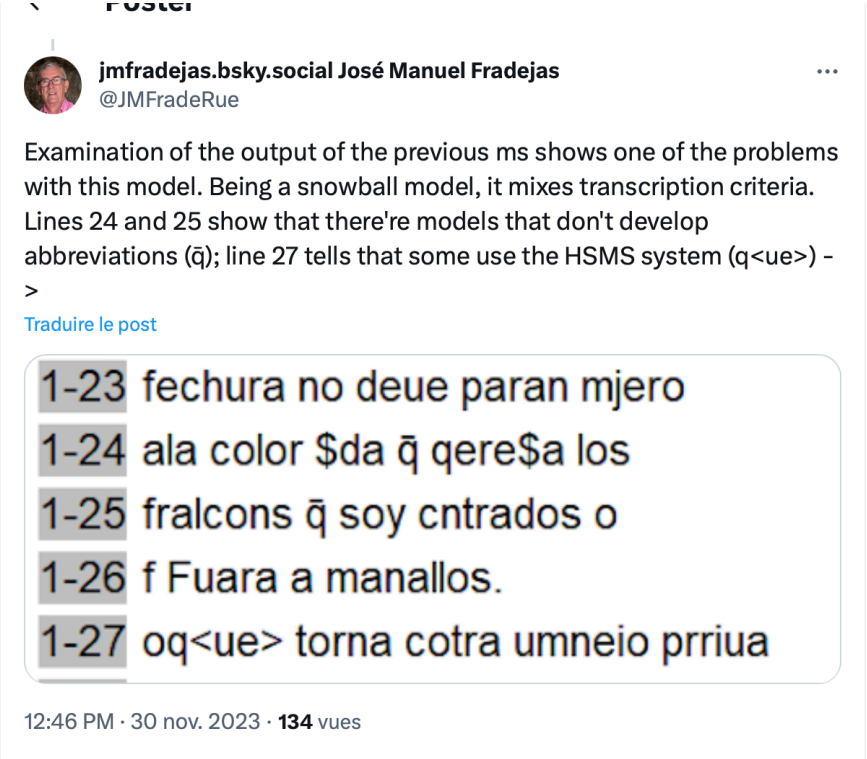
- How to transcribe manuscripts for the machine?
- How to transcribe consistently within a project?
- How to transcribe for reusable data?

  - "Well prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently."

    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". In: *Journal of open humanities data* 7.13 (2021), pp. 1–10

How to Transcribe Manuscripts?

- Define transcription methods suitable for machine learning.
- Determine the desired level of detail in transcription.
- Use a predefined character set and document your choices.
- Ensure compatibility of transcription data

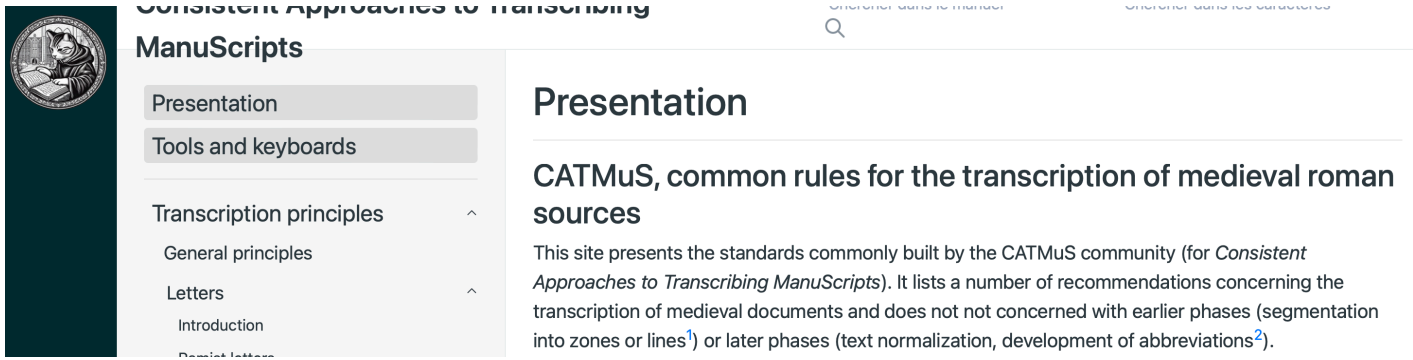


CATMuS Transcription Guidelines

- CATMuS transcription guidelines evolve from pre-existing guidelines for medieval French manuscripts (Pinche 2022)
- thore guidelines are based on basic principles :
  - restricting the characters used for specific purposes according to the MUFI;
  - applying a graphematic transcription;
  - keeping abbreviations (and thus, reducing the part of the language-specific traits the model has to learn).
- Decisions on abbreviations, diacritics, and phonetic representations aim to minimise characters in the HTR model.

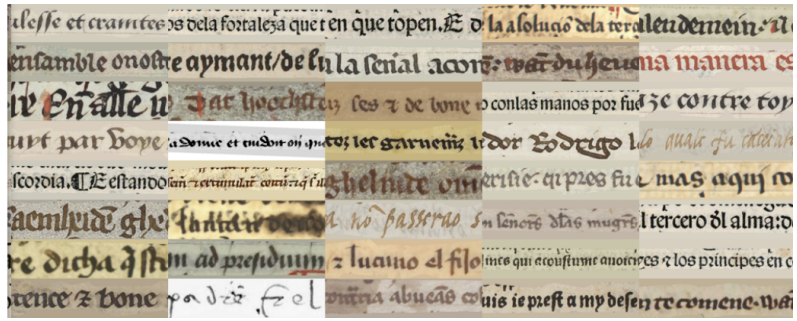
CATMuS transcription guidelines

- New languages and new challenges :
  - Latin involves new abbreviation to represent, such as <ꝛ> for <rum> ;
  - Middel English involves new signs for new phonetic realisation, like <ð> (eth) representing the voiced and voiceless dental fricative.
- The establishment of transcription rules remains an ongoing and evolving effort for ensuring data quality and homogeneity.
- The guidelines are available at the following link : `https://catmus-guidelines.github.io`



CATMuS dataset

CATMuS Medieval



- CATMuS dataset is published and documented on hugging face : `https://huggingface.co/datasets/CATMuS/medieval`
- built upon 17 different repositories
- c. 115.000 lines and 3.4M characters
- contains 180 different documents-hands
- Its most represented centuries are the 15th century (74 document-hands), the 14th (45), and the 13th (34).
- Over-representation of some genre and language : French-Narratives (29.2% of the dataset), Treatises-Latin (24.2%) and Treatises-Castillan (23%), linked to the history of the data gathering

CATMuS medieval model

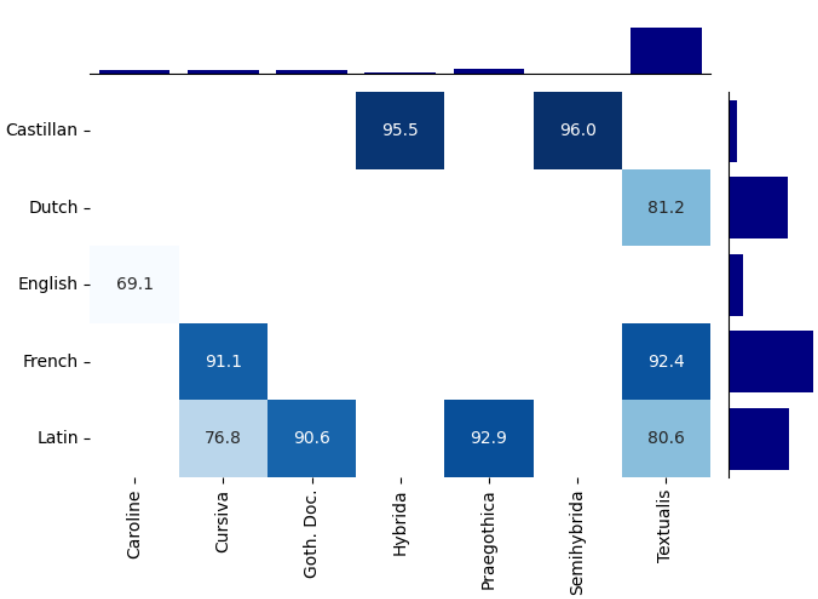


Figure: Absolute accuracy

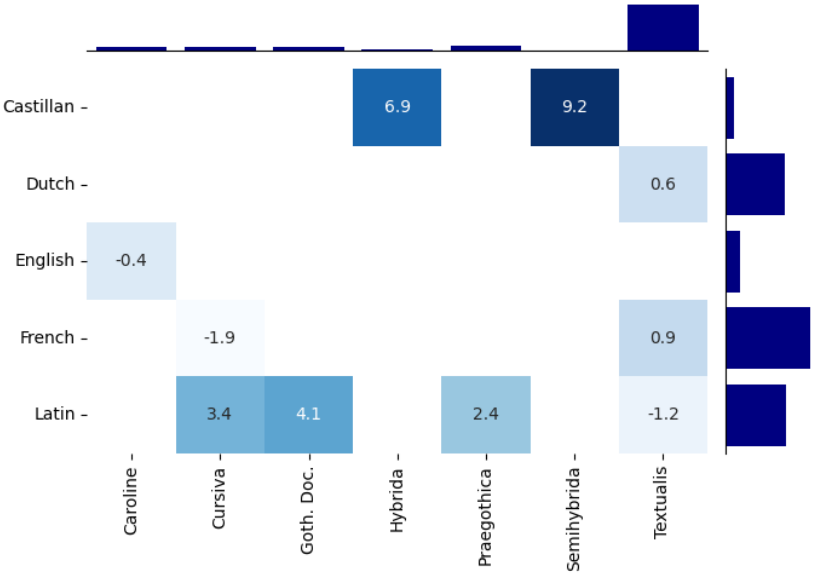


Figure: Improvement over CREMMA Generic

Figure: Test micro-accuracy per language and scripta. Bars represents the total number of characters per categorical feature.

CATMuS Generic Models

- A general model for medieval manuscripts : Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, and Simon Gabay. "CATMuS Medieval". lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024)
- A general model for gothic prints : Sonia Solfrini and Simon Gabay. "CATMuS Gothic Print". frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024)
- A general model for prints : Simon Gabay and Thibault Clérice. "CATMuS-Print [Large]". fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024)

Table of Contents

Conclusion

- ATR is becoming common as a first step in textual acquisition in humanities projects
- CATMuS project aims : providing consistent transcription standards :
  - Helps with data sharing
  - Improves the quality of predictions
  - Enables training of generic models
- For the future ? We want to expand the linguistic area with Old English and Middle Dutch, the type of documents treated to documents of the practice, but also the time period from Middle Age to the contemporary era.

Table of Contents

Références I

[1] Maciej Eder. "Mind Your Corpus: Systematic Errors in Authorship Attribution". In: *Literary and Linguistic Computing* 28.4 (Dec. 1, 2013), pp. 603–614. DOI: 10.1093/llc/fqt039.

[2] Simon Gabay and Thibault Clérice. "CATMuS-Print [Large]". fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024).

[3] Matthias Gille Levenson. "Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR)". In: *Journal of Data Mining and Digital Humanities* (2023). DOI: 10.46298/jdmdh.10416. URL: https://zenodo.org/records/8340483.

[4] 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". In: *Journal of open humanities data* 7.13 (2021), pp. 1–10.

[5] Ariane Pinche. "Generic HTR Models for Medieval Manuscripts The CREMMALab Project". In: *Journal of Data Mining & Digital Humanities* (2023). URL: https://univ-lyon3.hal.science/hal-03837519/.

[6] Ariane Pinche. "Guide de transcription pour les manuscrits du Xe au XVe siècle". June 2022. URL: https://hal.archives-ouvertes.fr/hal-03697382.

[7] Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, and Simon Gabay. "CATMuS Medieval". lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024).

[8] Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, et al. "CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts". In: *DH2024*. ADHO. Washington DC, United States, Aug. 2024. URL: https://inria.hal.science/hal-04346939.

[9] Sonia Solfrini and Simon Gabay. "CATMuS Gothic Print". frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024).