Transcribing Your Corpus in the Digital Age : Automatic Handwriting Recognition

Julie Bordier[3,4], Olivier Brisville[2,3], Matthias Gille Levenson[1,3,4], Ariane Pinche[2,3]

[1]ENS-PSL, [2]CNRS, [3]CIHAM, [4]ENSL

Transcribing Your Corpus in the Digital Age, April 5th, 2024, Lyon

Table of Contents

Key Terms

- **OCR** : Optical Character Recognition
- **HTR** : Handwritten Text Recognition
- **ATR** : Automatic Text Recognition

Examples of platforms/tools :

- eScriptorium [B. Kiessling, Tissot, Stokes et Ezra 2019 and Benjamin Kiessling 2019]
- Transkribus [Kahle, Colutto, Hackl et Mühlberger 2017]
- Calfa [Vidal-Gorène et al. 2021]

What is ATR ?

Figure – ATR Prediction

- Prediction of text content
- from an image of the source by an
- artificial intelligence trained by a human
- in an alternating process
  - involving human interventions
  - and computational phases

Rapid Technological Progress

- 1951, pioneering work by Gustav Tauschek : "Pattern Recognition by Machine"
- 1970-1980, improvement of OCR algorithms also based on layout, used by postal services.
- 2000-present, advent of machine learning and deep learning models with CNNs.

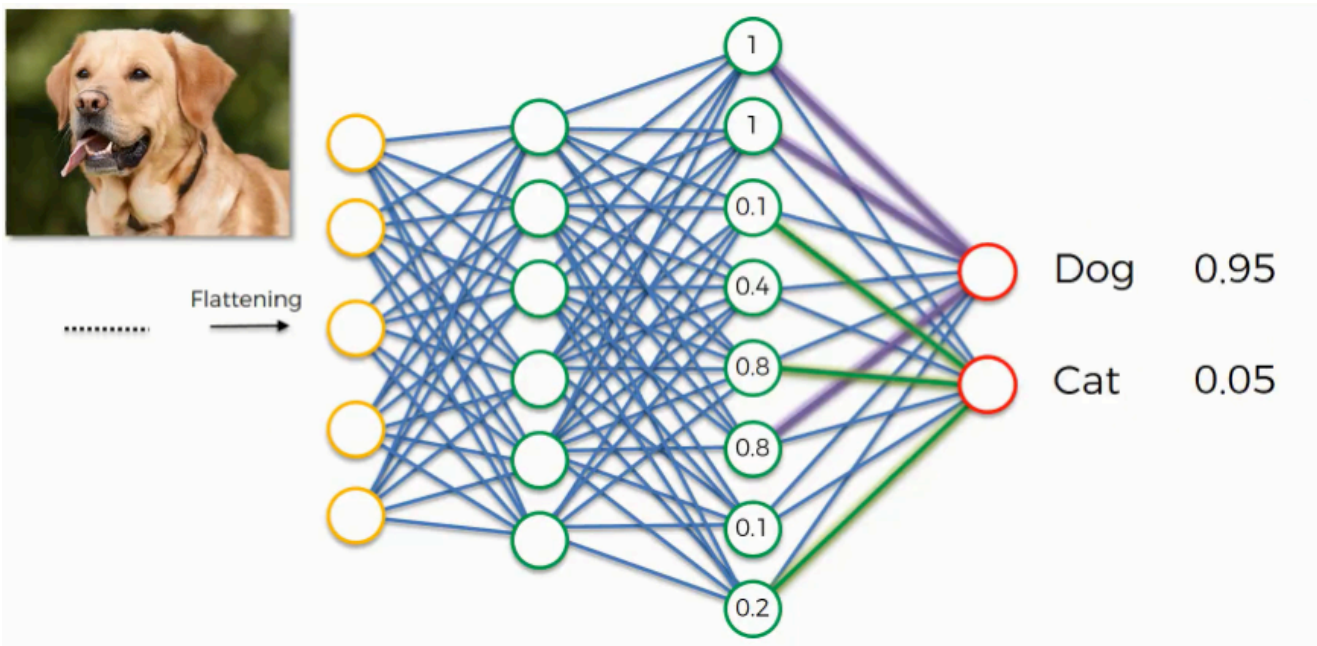


Figure – CNN, source :

ATR and Humanities (2000-2020)

- Handwritten texts on historical documents presents unprecedented challenges :
  - Non-standardized layouts
  - Degraded support
  - Irregular writing
  - Graphical and/or dialectal variations
- Pioneering research :
  - Alex Graves et Jürgen Schmidhuber. « Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks ». In : *Advances in neural information processing systems* 21 (2008) and Andreas Fischer et al. « Ground Truth Creation for Handwriting Recognition in Historical Documents ». In : *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. DAS '10 : The Eighth IAPR International Workshop on Document Analysis Systems. Boston Massachusetts USA : ACM, 9 juin 2010, p. 3-10. DOI : 10.1145/1815330.1815331.
- Pioneering projects : Himanis project (2015), the ANR Horae project (2017) led by Dominique Stutzmann .

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 et Purcell 2021.
- Emergence of intuitive platforms : eScriptorium and Transkribus.
- Organization of international conferences
  - ICDAR : International Conference on Document Analysis and Recognition
  - HIP : Historical Document Imaging and Processing workshop
- ATR is becoming a common step in more and more research projects, as seen in the DH and TEI conference programs.

Why Use HTR Today ?

- 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 (1er déc. 2013), p. 603-614. DOI : 10.1093/llc/fqt039)

Terminology

- **Corpus** : set of hand-labeled data
- **Supervised Learning** : machine learning technique based on pairs data/labels
- **Model** : adaptable computer file that, based on input data, provides an output, the *prediction*. One can also think of the model as a large mathematical function that, given numerical input data, proposes numerical output data.
- **Prediction** : production of data based on a model and input data
- **Training** : set of cycles of adapting a model to a data corpus

The Steps of HTR

- Loading images
- Segmenting image areas
- Segmenting lines containing text
- Predicting text on images
- Exporting data (txt, alto, page)

Digitizing the Data

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Classification

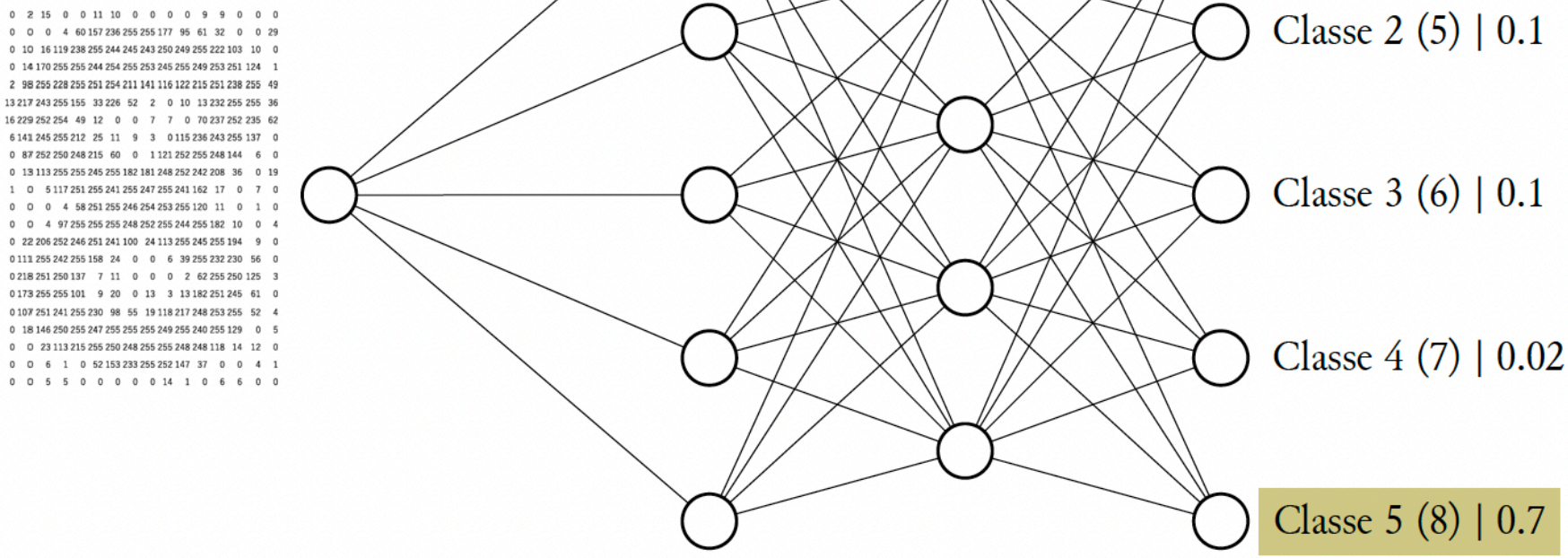


Figure – Simplified representation of a neural network

Training an HTR Model

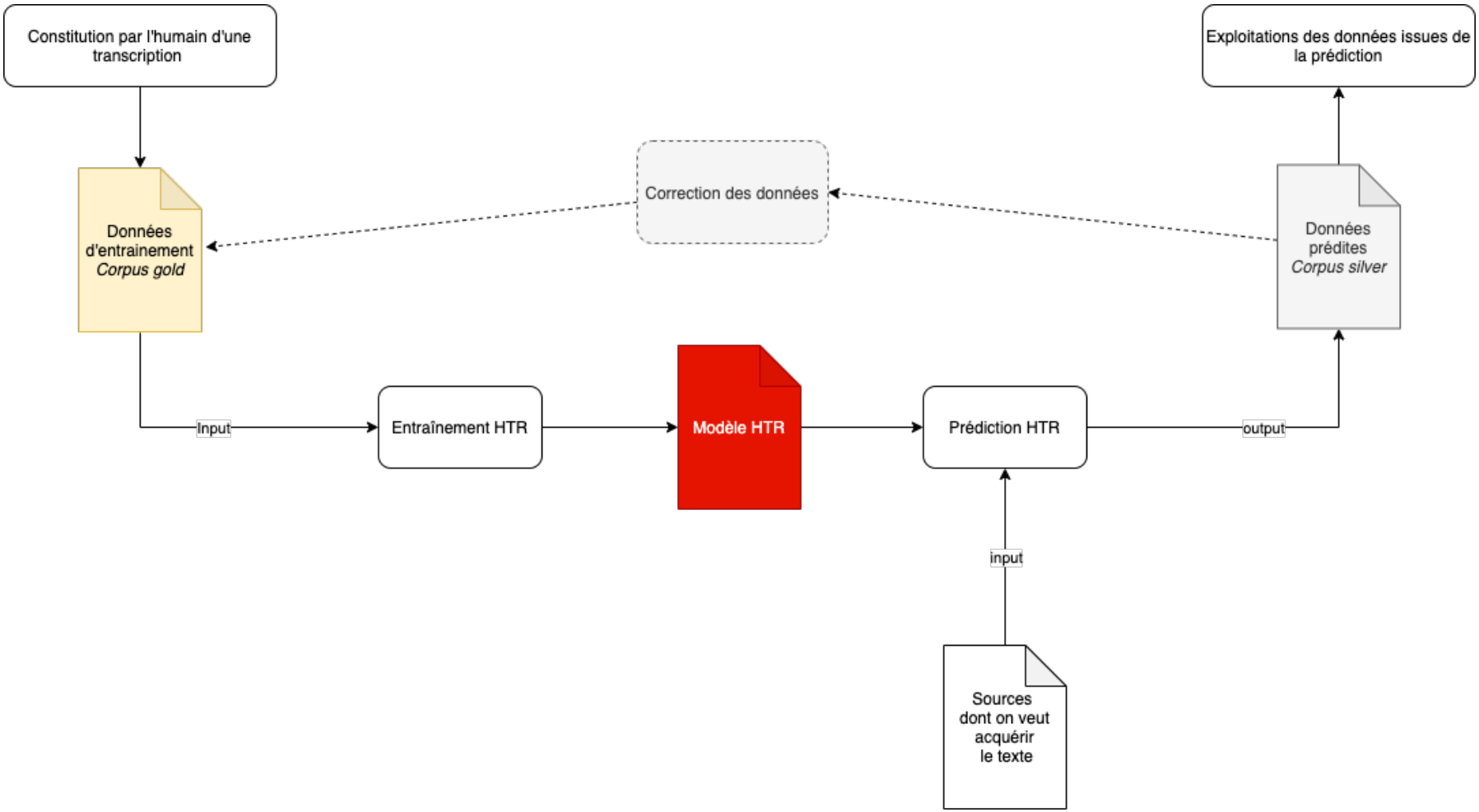


Figure – Representation of a training cycle

Training an HTR Model

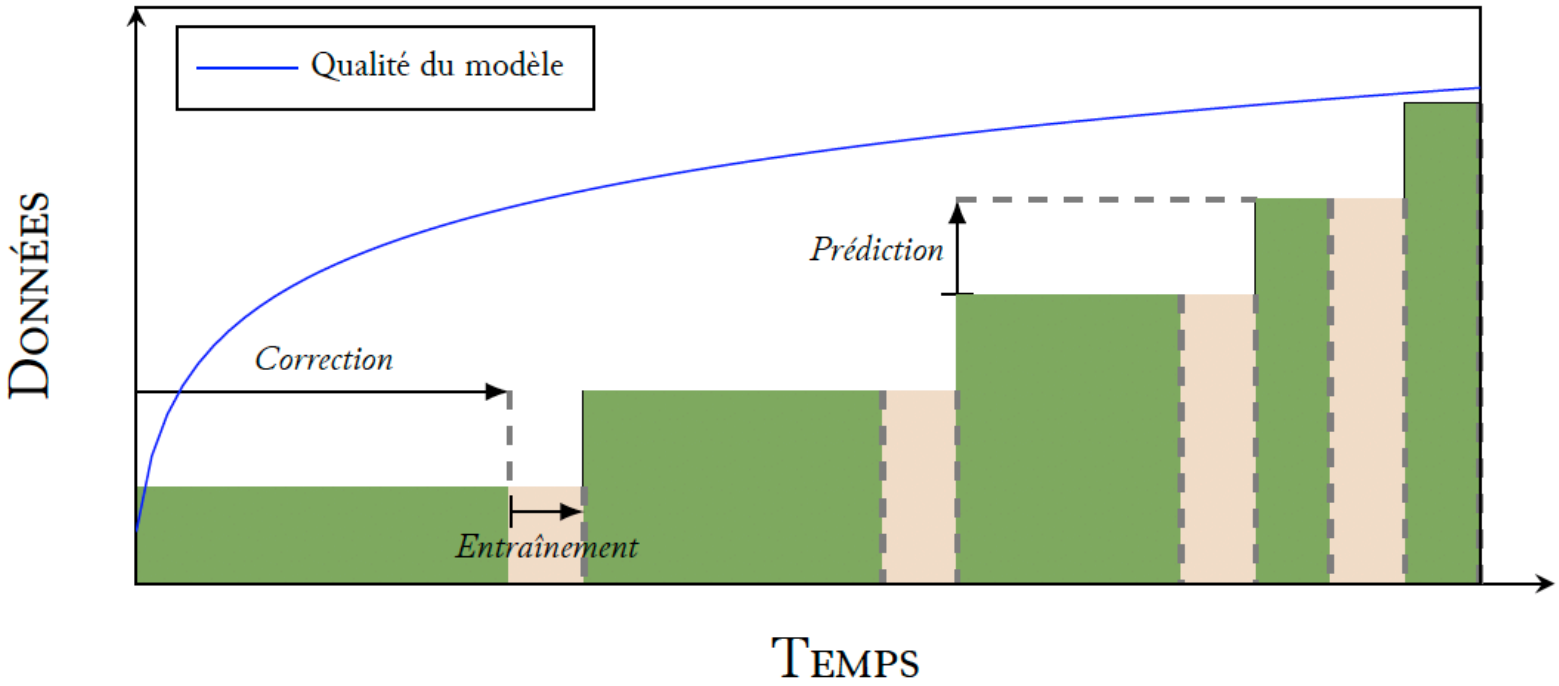


Figure – Evolution of data correction time according to the model's quality

Evaluating an HTR Model

To evaluate an HTR model :

- Compare :
  - Ground truth (GT) produced by a human (test set)
  - To the model's prediction of the same lines
  - To calculate a score that takes the form of either :
    - CER (Character Error Rate)
    - Accuracy (the percentage of the model's success)

Types of Errors

Calculating CER

$$CER = \frac{S + D + I}{N}$$

Performance of Bicerin and Cortado "out-of-domain"

Focus on Predictions : Geneva, Comites Latentes 102, 14th century

Figure – Prediction from Arabica model

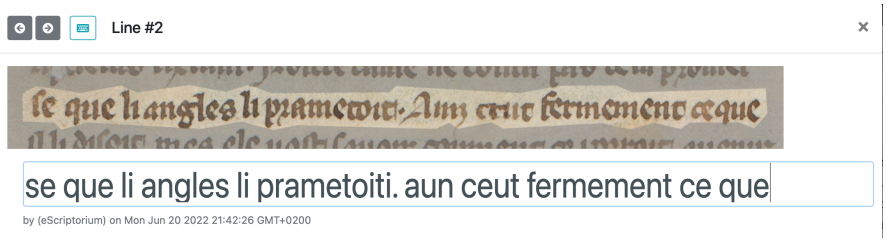


Figure – Prediction from Bicerin 1.0.0 model

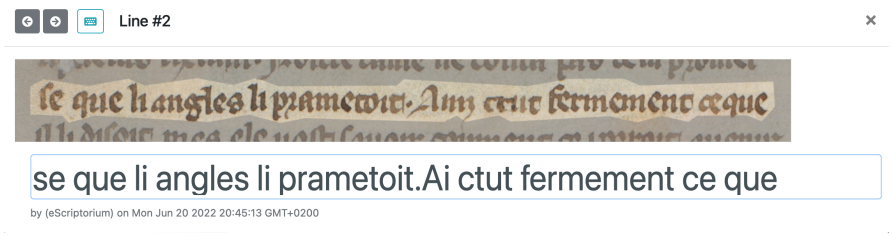


Figure – Prediction from Bicerin 1.1.0 model

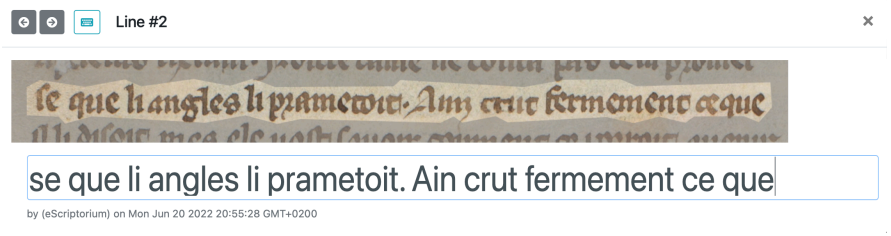


Figure – Prediction from Cortado model

Performance of Bicerin and Cortado "out-of-domain"

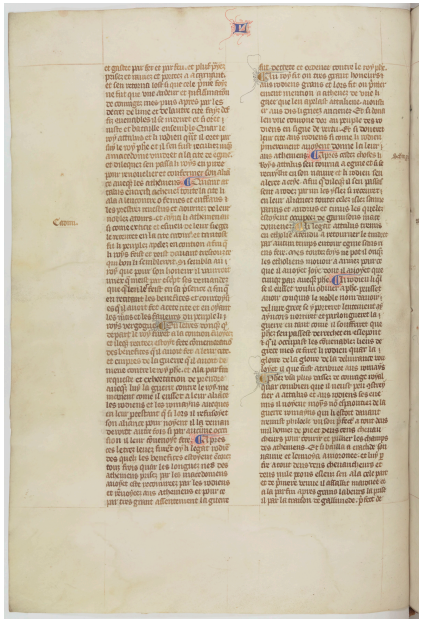


Figure – BnF, NAF 27401, 14th century

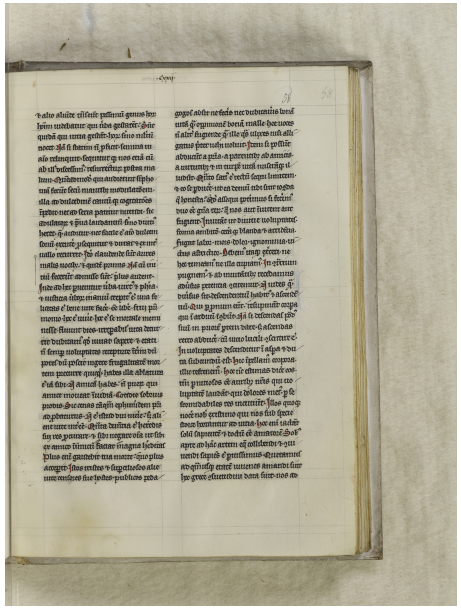


Figure – Arras, BM, 861, 14th century

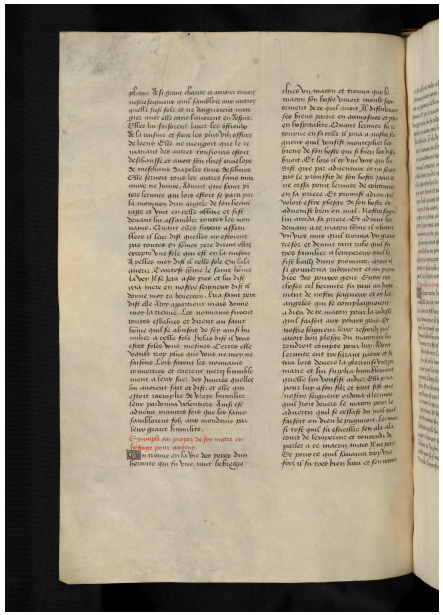


Figure – Brussels, KBR, 9232, 15th century

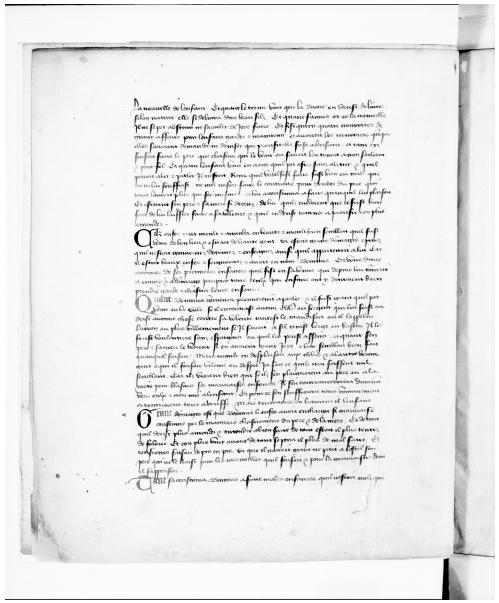


Figure – BnF, fr. 777, 15th century

Performance of Bicerin and Cortado "out-of-domain"

Scores of different models on four "out-of-domain" manuscripts

N°	Manuscripts	Date	Script	Lang.	Bicerin acc.	Cortado acc.	Improvement
1	BnF, NAF 27401	14th	textualis	Old Fr.	91.25%	91.40%	**+0.15**
2	Arras, Bibliothèque municipale, ms. 861	14th	textualis	Latin	82.99%	83.95%	+0.96
3	Bruxelles, Bibliothèque royale, ms. 9232	15th	hybrid	Old Fr.	91.34%	95.93%	+4.59
4	BnF, fr. 777	15th	cursiva	Old Fr.	63.96%	82.80%	**+18.84**

Tableau – Performance de Bicerin et Cortado "out-of-domain"

Performance of Bicerin and Cortado with Fine-Tuning

Scores of different fine-tuned models (4 pages) on four "out-of-domain" manuscripts

N°	Manuscripts	date	script	Lang.	Bicerin FT acc.	Cortado FT acc.
1	BnF, NAF 27401	14th	textualis	Old Fr.	98.83% (+7.58)	98.08% (+6.68)
2	Arras, Bibliothèque municipale, ms. 861	14th	textualis	Latin	92.16% (+9.17)	92.81% (+8.86)
3	Bruxelles, Bibliothèque royale, ms. 9232	15th	hybrid	Old Fr.	98.70% (+7.36)	99.04% (+3.11)
4	BnF, fr. 777	15th	cursiva	Old Fr.	98.73% (**+34.77**)	98.88 (**+16.08**)

Tableau – Performance des modèles affinés à partir de Bicerin et Cortado

Towards the Creation of "Large-Scale" ATR Models

- 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/
- 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
- Generic model from Transkribus : Medieval_Scripts_M2.4

Training Generic Models

- A general model for medieval manuscripts : Ariane Pinche et al. « CATMuS Medieval ». lat. In : (nov. 2023). Publisher : Zenodo. URL : https://zenodo.org/records/10066219 (visité le 08/01/2024)
- A general model for gothic prints : Sonia Solfrini et Simon Gabay. « CATMuS Gothic Print ». frm. In : (jan. 2024). Publisher : Zenodo. URL : https://zenodo.org/records/10599911 (visité le 27/03/2024)
- A general model for prints : Simon Gabay et Thibault Clérice. « CATMuS-Print [Large] ». fra. In : (jan. 2024). Publisher : Zenodo. URL : https://zenodo.org/records/10592716 (visité le 27/03/2024)

Table of Contents

Diverse Sources

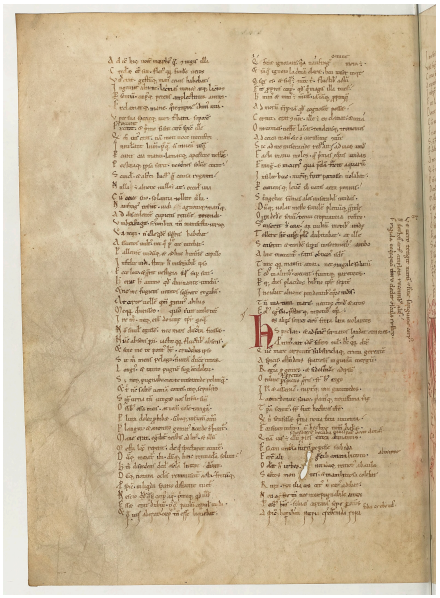


Figure – BnF, Latin, 8001, 13th century

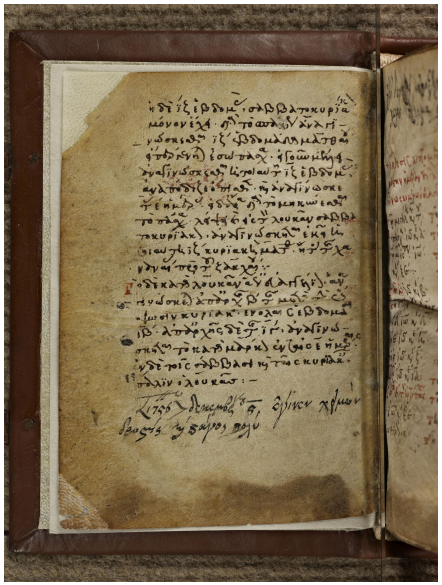


Figure – Strasbourg, ms. 1.916, 13th century

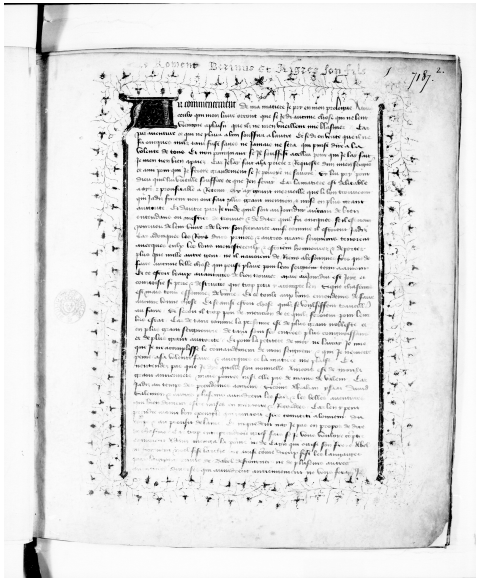


Figure – BnF, French, 777, 15th century

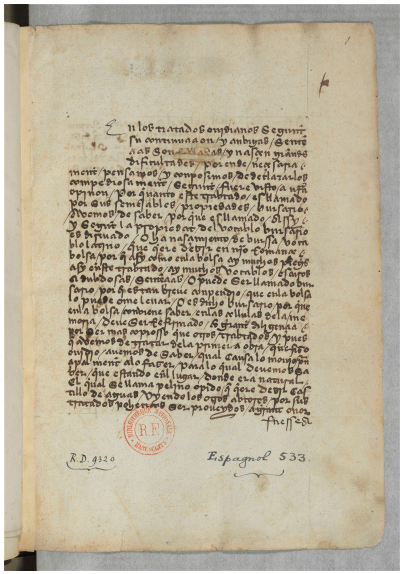


Figure – BnF, Spanish, 533, 15th century

Various Documents and Layouts

Figure – Turin Manuscript, "Ségurant le chevalier au dragon", 15th century

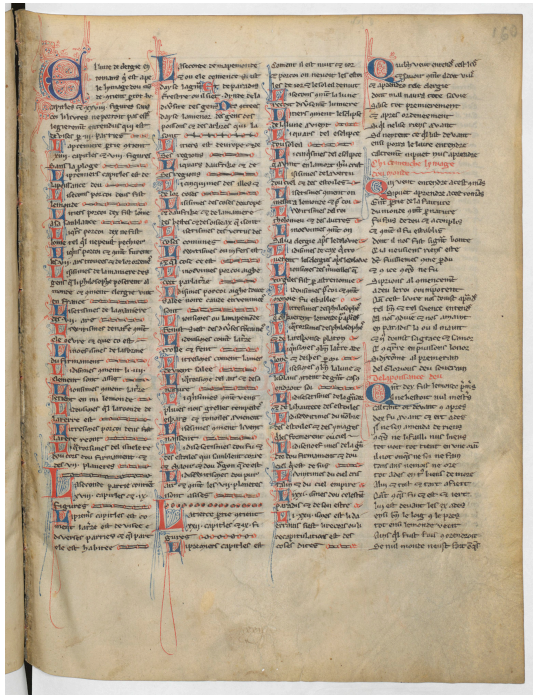


Figure – BnF, Arsenal, 3516, 12th century

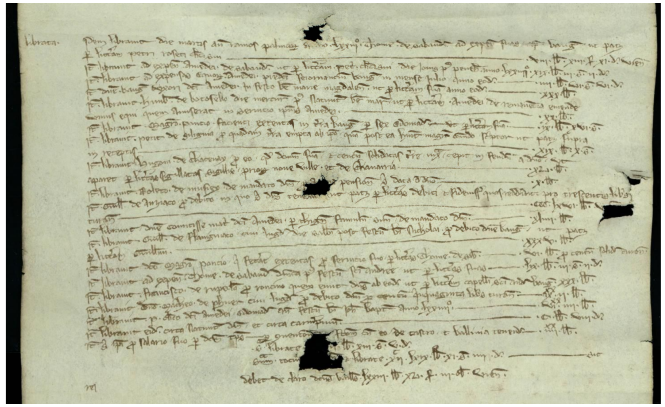


Figure – Departmental Archives of Côte d'Or, B6739, 13th century

Layout Analysis : Segmentation

- Identification of different areas of the document : using controlled vocabulary, such as SegmOnto.

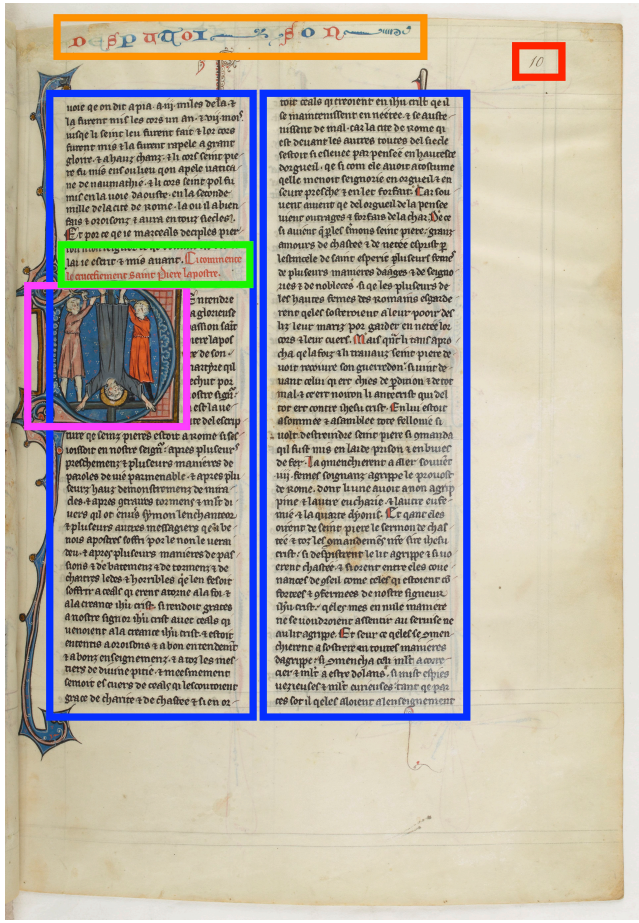


Figure – BnF, fr. 412, fol.10r

SegmOnto

Simon Gabay, Ariane Pinche, Kelly Christensen et Jean-Baptiste Camps. « SegmOnto : A Controlled Vocabulary to Describe and Process Digital Facsimiles ». working paper or preprint. Déc. 2023. URL : https://hal.science/hal-04343404

https://segmonto.github.io

Page

- DamageZone
- DropCapitalZone
- FigureZone
- MainZone
- MarginZone
- MusicNotationZone
- NumberingZone
- QuireMarksZone
- RunningTitleZone

Line

- DefaultLine
- DropCapitalLine
- Interlinear
- MusicLine
- HeadingLine

Definitions

https://github.com/SegmOnto/examples

How to Transcribe Manuscripts ?

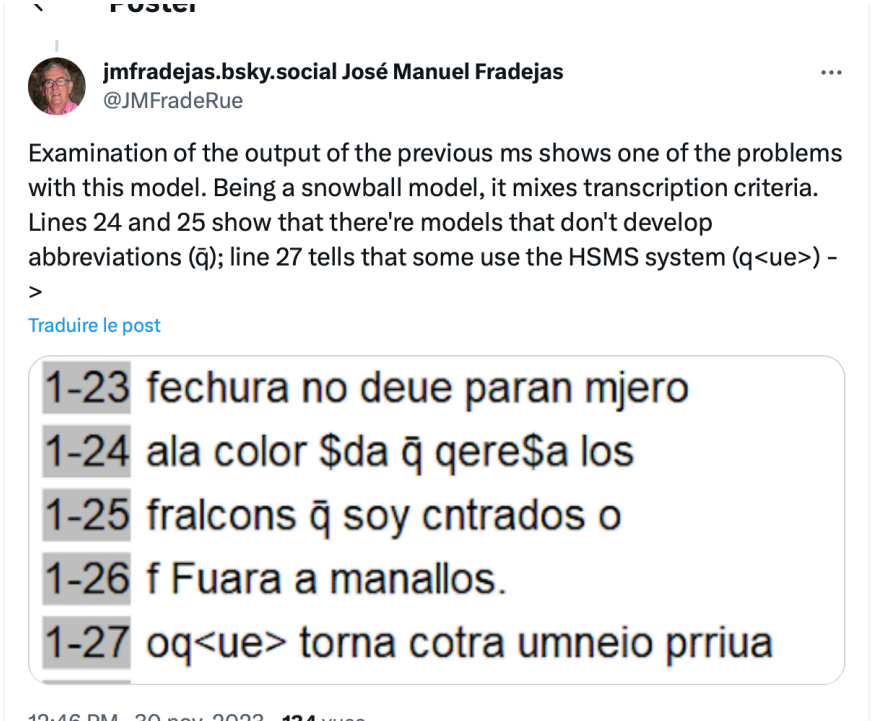
- How to transcribe manuscripts for the machine ?
- How to transcribe consistently within a project ?
- How to transcribe for reusable data ?

  - "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 et 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), p. 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.
  - See the MUFI (Medieval Unicode Font Initiative) initiative
  - See the transcription recommendations proposed as part of CREMMA for medieval texts
- Ensure compatibility of transcription data



Conclusion

ATR is a technology :

- becoming common as a first step in textual acquisition
- evolving rapidly in recent years
- becoming increasingly easy to use
- enabling the study of unprecedented corpus sizes

Table of Contents

Références I

[1] Maciej Eder. « Mind Your Corpus : Systematic Errors in Authorship Attribution ». In : *Literary and Linguistic Computing* 28.4 (1er déc. 2013), p. 603-614. DOI : 10.1093/llc/fqt039.

[2] Andreas Fischer, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser et Michael Stolz. « Ground Truth Creation for Handwriting Recognition in Historical Documents ». In : *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. DAS '10 : The Eighth IAPR International Workshop on Document Analysis Systems. Boston Massachusetts USA : ACM, 9 juin 2010, p. 3-10. DOI : 10.1145/1815330.1815331.

[3] Simon Gabay et Thibault Clérice. « CATMuS-Print [Large] ». fra. In : (jan. 2024). Publisher : Zenodo. URL : https://zenodo.org/records/10592716 (visité le 27/03/2024).

[4] Simon Gabay, Ariane Pinche, Kelly Christensen et Jean-Baptiste Camps. « SegmOnto : A Controlled Vocabulary to Describe and Process Digital Facsimiles ». working paper or preprint. Déc. 2023. URL : https://hal.science/hal-04343404.

[5] 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.

[6] Alex Graves et Jürgen Schmidhuber. « Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks ». In : *Advances in neural information processing systems* 21 (2008).

[7] Tobias Mathias Hodel, David Selim Schoch, Christa Schneider et 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), p. 1-10.

[8] Philip Kahle, Sebastian Colutto, Günter Hackl et 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)*. T. 4. IEEE, 2017, p. 19-24.

Références II

[9] B. Kiessling, R. Tissot, P. Stokes et D. Stökl Ben Ezra. « eScriptorium : An Open Source Platform for Historical Document Analysis ». In : *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). T. 2. Sept. 2019, p. 19. DOI : 10.1109/ICDARW.2019.10032.

[10] Benjamin Kiessling. « Kraken - an Universal Text Recognizer for the Humanities ». In : DH2019 : Complexity. Utrecht, 2019. URL : https://dev.clariah.nl/files/dh2019/boa/0673.html.

[11] 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/.

[12] Ariane Pinche et al. « CATMuS Medieval ». lat. In : (nov. 2023). Publisher : Zenodo. URL : https://zenodo.org/records/10066219 (visité le 08/01/2024).

[13] Sonia Solfrini et Simon Gabay. « CATMuS Gothic Print ». frm. In : (jan. 2024). Publisher : Zenodo. URL : https://zenodo.org/records/10599911 (visité le 27/03/2024).

[14] Chahan Vidal-Gorène, Noëmie Lucas, Clément Salah, Aliénor Decours-Perez et Boris Dupin. « RASAM. A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi ». In : *International Conference on Document Analysis and Recognition*. Springer, 2021, p. 265-281.