

Text editing in the digital age

Automatic text recognition

Ariane Pinche¹², Matthias Gille Levenson³²

¹CNRS, ²CIHAM, ³ÉNS Lyon

9 avril 2024

Table of Contents

1 Text editing in the digital age :

2 Automatic Handwriting Recognition

- Definition
- A Bit of History...
- ATR and Historical Documents Today
- How Does It Work ?
- Evaluating an ATR Model
- Fine-Tuning Models

3 ATR and Specific Challenges in Historical Documents

- Variety of Scripts and Graphic Systems
- Describing Layout
- How to Transcribe ?
- Presentation of Kraken and eScriptorium

4 References

In the beginning was the text...

Every edition is the product of an era and a philological school. Traditional editions, heirs to a centuries-old tradition, are limited by their materiality and present an normalised and fixed text.



VIE DE SAINT MARTIN

De saint Martin

1. Mout¹ doit on doucement et volentiers le bien oïr et entendre, car par le bien savoir et retenir [fol. 103b] puet l'en sovent a bien venir. Qui bien ne seit ne bien n'entent de bien faire n'a nul talent. Mes del bien nest sovent li biens, del mal li maussu com dist l'Ecriture. Por ce se doit l'en au bien avoier et le bien feire, si com li saint home furent çā en arriere de cui nos trovons les oevres et les vies [es] Escriptures. Et bien sacent tuit cil qui vivent qe ja n'auront tant de bien fet en totes lor vies qe, qant la mort dont nule rien n'escape les poindera au cuer, q'il ne cuident petit avoir fait. Dex ! Qe feront dont cil qui riche sont et aise de l'avoir de cest siecle, ne en eus n'om doucor ne humilité ne misericorde, ainz sont plein d'angoisse et de traïsson et de felonie et de si grant

5

10

15

¹La Vie de saint Martin s'ouvre sur une majuscule historiée (m) qui représente Saint Martin sur un cheval, l'épée à la main, prêt à couper son manteau pour le partager avec un nécessiteux également présent dans l'illustration.

1 De saint Martin | Ci commence la vie de monseigneur saint Martin C°. Ci commence la vie saint Martin C° || ¶ es Escriptures C° C° | scriptures C°

In the beginning was the text...

Digital editing allows us to overcome these limitations

- Data can be enriched
 - ▶ Linguistic annotations
 - ▶ Graphical variants
 - ▶ Text versions
- Preliminary work on text establishment is preserved
- Textual data can be utilized beyond the scope of editing
 - ▶ Reuses (correction of the edition, insertion into another corpus, etc.)
 - ▶ Multiplication of visualizations (imitative, standardized, etc.)
 - ▶ Statistical analyses (stemmatology, stylometry, data visualization, etc.)

In the beginning was the text...

- How to acquire your text ?
 - ▶ By manual transcription
 - ▶ Starting from a natively digital text
 - ▶ Using automatic text acquisition
- How to enrich your text ?
 - ▶ Using a markup system (XML TEI)
 - ▶ Manually annotating
 - ▶ Using automatic annotation tools (NLP, NER)

Table of Contents

1 Text editing in the digital age :

2 Automatic Handwriting Recognition

- Definition
- A Bit of History...
- ATR and Historical Documents Today
- How Does It Work ?
- Evaluating an ATR Model
- Fine-Tuning Models

3 ATR and Specific Challenges in Historical Documents

- Variety of Scripts and Graphic Systems
- Describing Layout
- How to Transcribe ?
- Presentation of Kraken and eScriptorium

4 References

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.

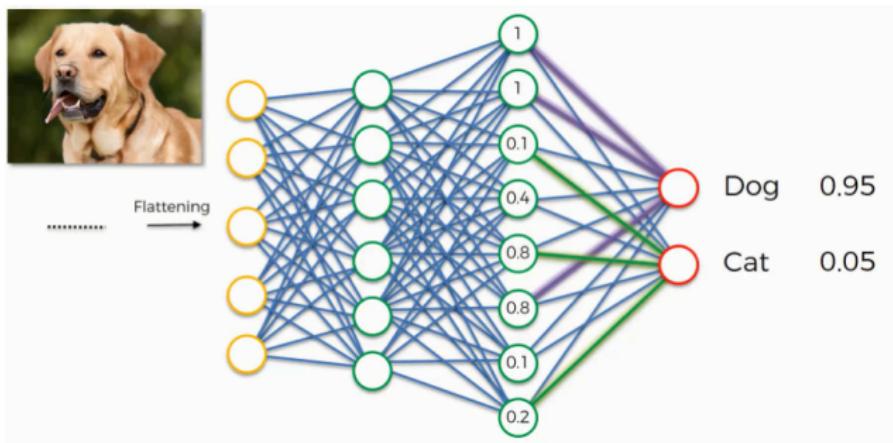


Figure – CNN, source :

<https://www.superdatascience.com/blogs/the-ultimate-guide-to-convolutional-neural-networks-cnn>

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 ATR 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 (1^{er} déc. 2013), p. 603-614. DOI : 10.1093/linc/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 ATR

- 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	4	60	157	236	255	255	177	95	61	32	0	0
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255
13	217	243	255	155	33	226	52	2	0	10	15	232	255	255
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137
0	87	252	250	248	215	10	0	112	252	255	248	244	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61
0	107	251	241	255	230	98	55	19	116	217	248	253	255	52
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0

A set of small, semi-transparent navigation icons located at the bottom of the slide, including arrows for navigation, a magnifying glass for search, and other symbols.

Classification

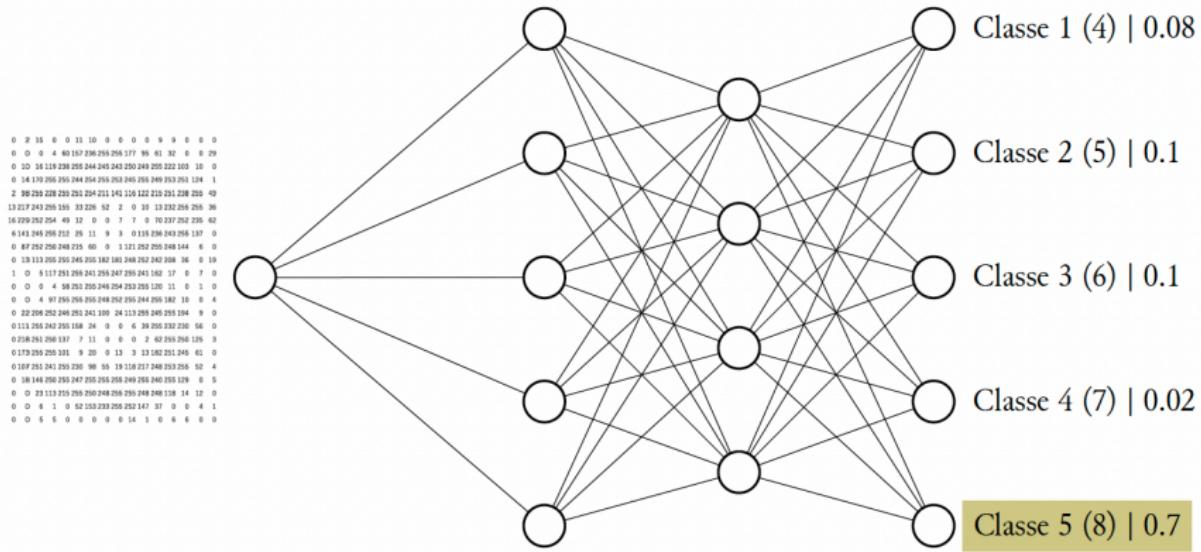


Figure – Simplified representation of a neural network

Training an ATR Model

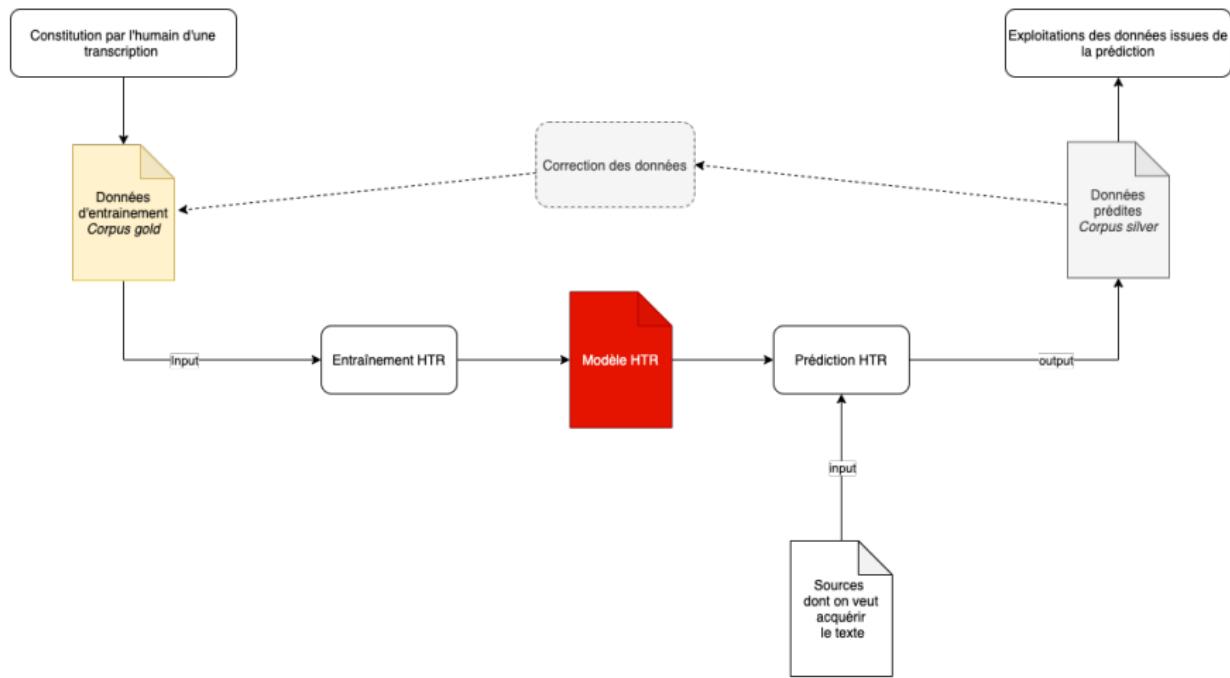


Figure – Representation of a training cycle

Training an ATR Model

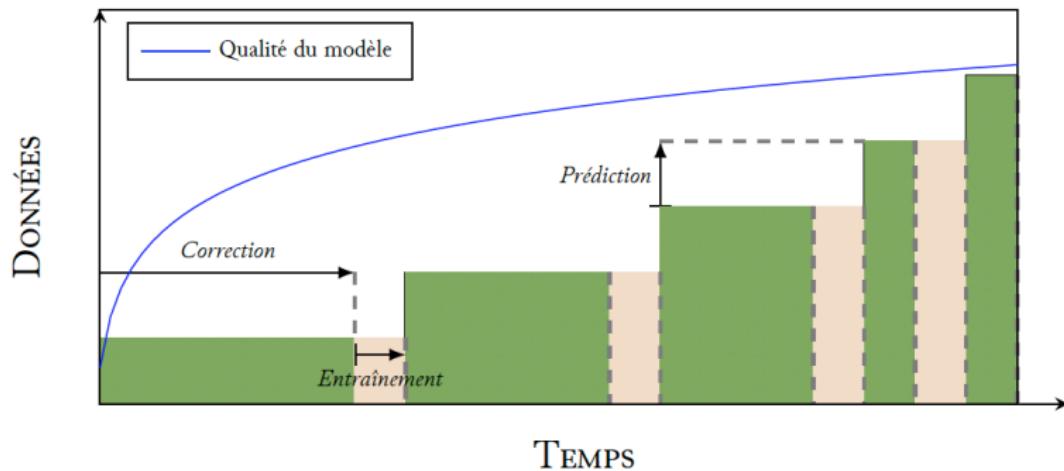


Figure – Evolution of data correction time according to the model's quality

Evaluating an ATR Model

To evaluate an ATR 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

Line #2

se que li angles li prametoit. Ain crut fermement ce que
se que li angles li prametoit. Lui ctut fermement ce que.

by (escriptorium) on Mon Jun 20 2022 21:44:17 GMT+0200

Figure – Prediction from Arabica model

Line #2

se que li angles li prametoit. Ain crut fermement ce que
se que li angles li prametoiti. aun ceut fermement ce que.

by (escriptorium) on Mon Jun 20 2022 21:42:26 GMT+0200

Figure – Prediction from Bicerin 1.0.0 model

Line #2

se que li angles li prametoit. Ain crut fermement ce que
se que li angles li prametoit. Ai ctut fermement ce que.

by (escriptorium) on Mon Jun 20 2022 20:48:13 GMT+0200

Figure – Prediction from Bicerin 1.1.0 model

Line #2

se que li angles li prametoit. Ain crut fermement ce que
se que li angles li prametoito. Ain crut fermement ce que.

by (escriptorium) on Mon Jun 20 2022 20:59:28 GMT+0200

Figure – Prediction from Cortado model

Performance of Bicerin and Cortado “out-of-domain”

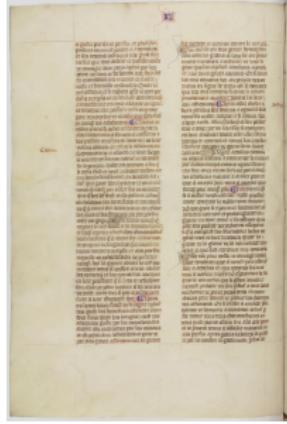


Figure – BnF,
NAF 27401,
14th century



Figure – Arras, BM,
861, 14th century

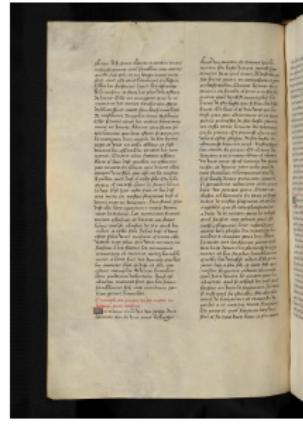


Figure – Brussels,
KBR, 9232,
15th century



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 CREMMLA 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

1 Text editing in the digital age :

2 Automatic Handwriting Recognition

- Definition
- A Bit of History...
- ATR and Historical Documents Today
- How Does It Work ?
- Evaluating an ATR Model
- Fine-Tuning Models

3 ATR and Specific Challenges in Historical Documents

- Variety of Scripts and Graphic Systems
- Describing Layout
- How to Transcribe ?
- Presentation of Kraken and eScriptorium

4 References

Diverse Sources



Figure – BnF, Latin,
8001, 13th century



Figure – Strasbourg,
ms. 1.916, 13th century



Figure – BnF, French, 777,
15th century

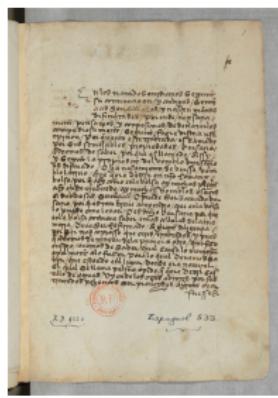


Figure – BnF,
Spanish, 533, 15th century

Various Documents and Layouts



Figure – Turin Manuscript,
"Sécurant 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://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



jmfraudejas.bsky.social José Manuel Fradejas
@JMFradeRue

...

Examination of the output of the previous ms shows one of the problems with this model. Being a snowball model, it mixes transcription criteria. Lines 24 and 25 show that there're models that don't develop abbreviations (ñ); line 27 tells that some use the HSMS system (q<ue>) ->

[Traduire le post](#)

1-23 fechura no deue paran mjero

1-24 ala color \$da ñ qere\$ a los

1-25 fralcons ñ soy cntrados o

1-26 f Fuara a manallos.

1-27 oq<ue> torna contra umneio prriua

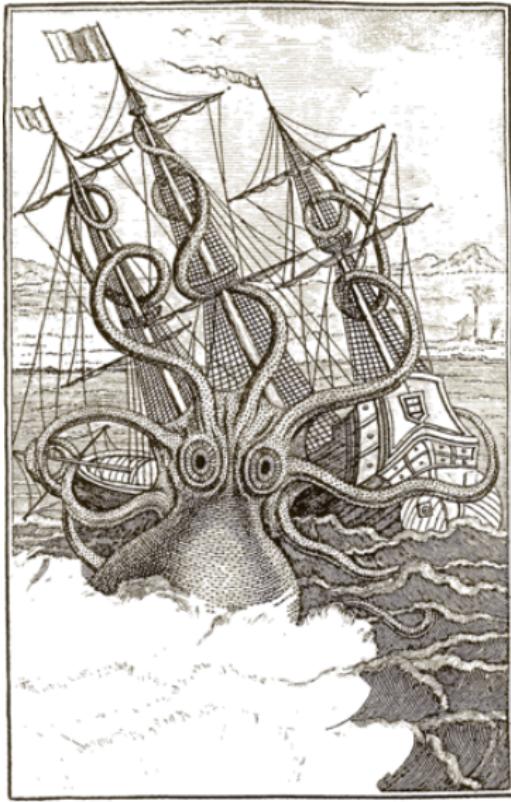
12:46 PM · 30 nov. 2023 · 134 vues

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

Kraken



Kraken

- Layout analysis and ATR tool
- Based on deep learning
- Developed by Ben Kiessling in the Scripta project (PSL)
- Python module, <https://github.com/mittagessen/kraken>
- Documentation : kraken.re
- It can be used directly from the command line or via the eScriptorium interface

eScriptorium

- Free software for segmenting a document, detecting lines, transcribing, training an ATR model, and applying it to sources
- Developed as part of the Scripta project (PSL)
- Relies on Kraken for layout analysis and ATR
- Requires deployment on a server by an institution
- Code : <https://gitlab.inria.fr/scripta/escriptorium>
- Video demos : <https://scripta.hypotheses.org/escriptorium-video-gallery>

Utiliser eScriptorium nécessite :

- D'ouvrir un compte sur une instance d'eScriptorium ou d'installer une instance locale d'eScriptorium
- D'avoir accès aux fichiers images de ses sources : des fichiers locaux ou téléchargés directement depuis un site institutionnel en utilisant un manifeste IIIF :
<https://gallica.bnf.fr/iiif/ark:/12148/btv1b84259980/manifest.json>

eScriptorium

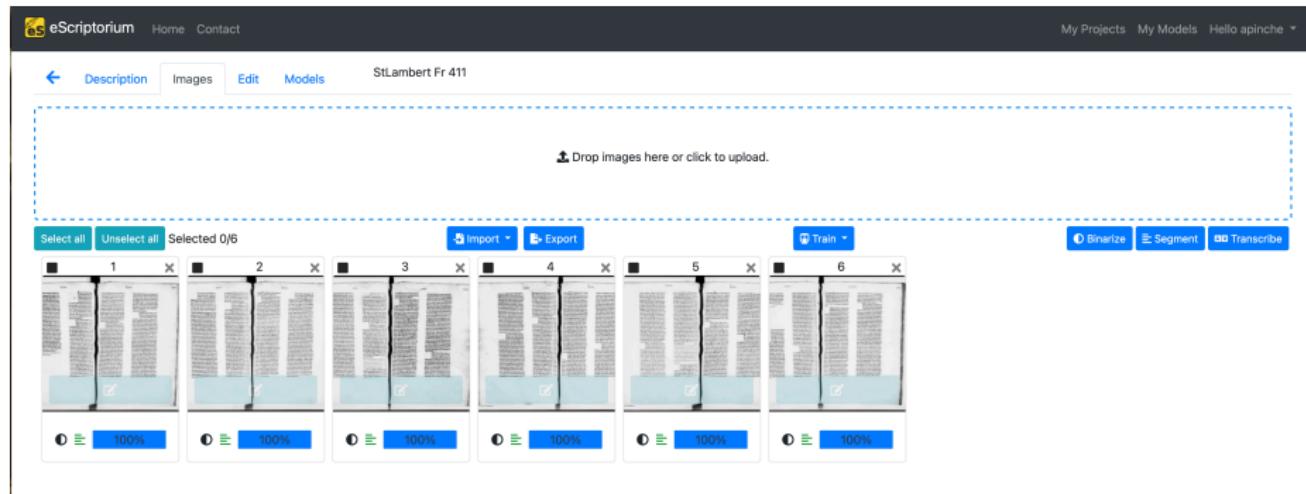


Figure – eScriptorium Interface

eScriptorium is a web interface that allows :

- Segmenting the page of a document (zones and lines)
- Transcribing documents to create training data
- Training an ATR model
- Applying an ATR or segmentation model to a document



MARTIN

103

Auoit non leomouicna Eti

de la fontaine cor me les castors [■] les
etats de la fontaine cor me le castor
par lequel le castor cor le chene de la fontaine par
corneau voulz en avoir un arbre.

mais il n'avoit pas de la force de la fontaine
dans la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

Puet len souent abien ueni

renent auvent neudre. Deben hore ha ne
neudre neudre. Deben hore ha ne
neudre neudre. Deben hore ha neudre neudre.

mais il n'avoit pas de la force de la fontaine de la fontaine
dans la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

et il n'avoit pas de la force de la fontaine de la fontaine
de la fontaine de la fontaine de la fontaine.

Figure – Segmentation and transcription of a document using eScriptorium

Table of Contents

1 Text editing in the digital age :

2 Automatic Handwriting Recognition

- Definition
- A Bit of History...
- ATR and Historical Documents Today
- How Does It Work ?
- Evaluating an ATR Model
- Fine-Tuning Models

3 ATR and Specific Challenges in Historical Documents

- Variety of Scripts and Graphic Systems
- Describing Layout
- How to Transcribe ?
- Presentation of Kraken and eScriptorium

4 References

Références I

- [1] Maciej EDER. « Mind Your Corpus : Systematic Errors in Authorship Attribution ». In : *Literary and Linguistic Computing* 28.4 (1^{er} 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 (HTK/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. KISSLING, 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.2019.90032.
- [10] Benjamin KISSLING. « 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 CREMMLab 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, Noémie 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.