CIHAM (UMR 5648), CNRS

Automatic Text Recognition and Historical Documents

Ariane Pinche
ariane.pinche@cnrs.fr

EnExDi, Lyon, 3 juin 2025

Table of Contents

What is ATR?

Figure: ATR Prediction

- Prediction of textual content
- from an image of the source by
- an artificial intelligence trained by a human
- through a process alternating
  - phases of human intervention
  - and computation phases

Some Vocabulary

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

Some Tools

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

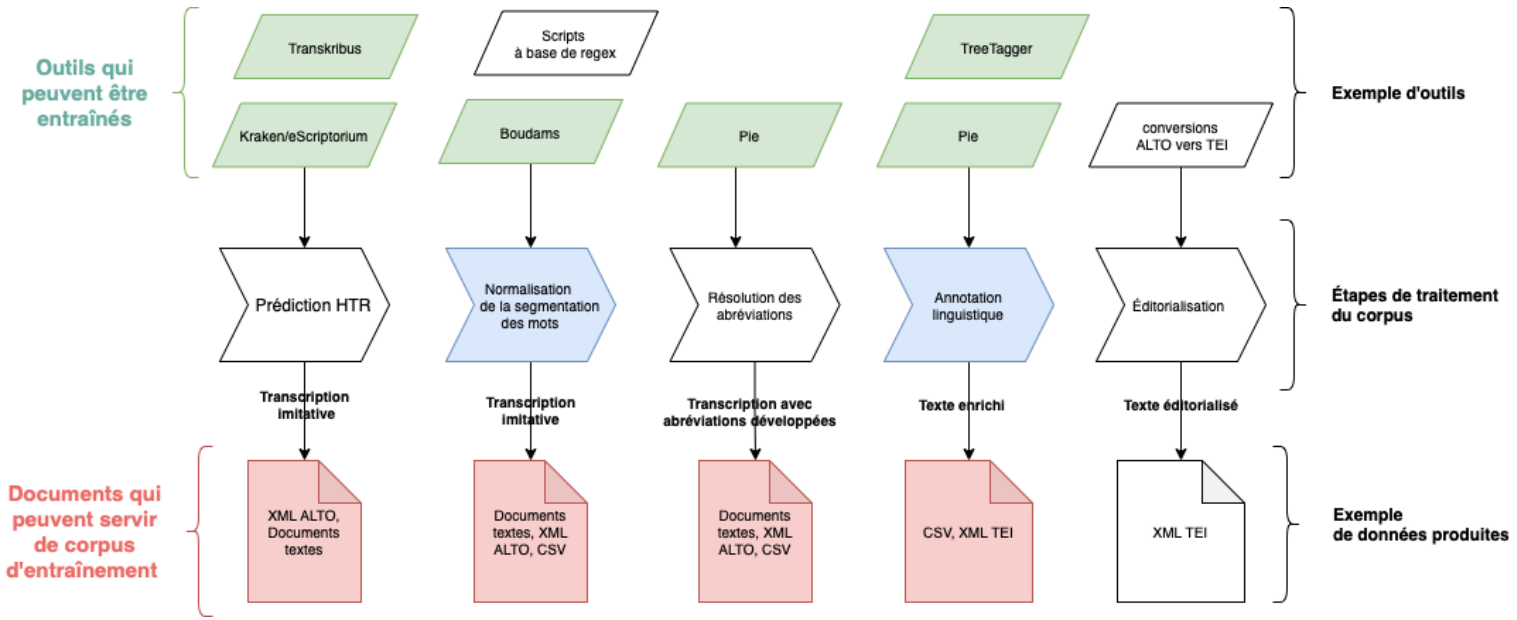
A bit of History...

- 2000–2010: Development of HTR usage (experimental phase)
- Early research and pioneering projects:
  - Alex Graves and Jürgen Schmidhuber, “Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks,” *Advances in Neural Information Processing Systems*, 2008.
  - Andreas Fischer, Emanuel Indermühle, Horst Bunke, [et al.], “Ground truth creation for handwriting recognition in historical documents,” 2010.
  - Pioneering projects: Himanis project (2015), ANR Horae project (2017), both led by Dominique Stutzmann.
- Late 2010s: Growing use of HTR in research projects, driven by the development of two main tools: Transkribus (using Pylaia and HTR+ engines) and eScriptorium (using the Kraken engine)
- Organization of international conferences:
  - ICDAR: International Conference on Document Analysis and Recognition
  - HIP: Historical Document Imaging and Processing Workshop

Today...

- ATR is a well-mastered task from a computer science point of view
  - Nowdays, with model that can reach a CER (character error rate) between 8% and 2% for manuscripts, “from a computer science point of view, the recognition of handwriting seems to be a resolved task. The latest recognition engines allow for the successful recognition of specifically trained hands producing a text as reusable data” , *Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021. “General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example.” Journal of Open Humanities Data 7(0):13. doi: 10.5334/johd.46.*
- L’ATR est devenu une étape courante dans les pipelines d’acquisition textuelle qui intéresse aussi bien les institutions patrimoniales que les projets de recherche.

Example of textual acquisition workflow

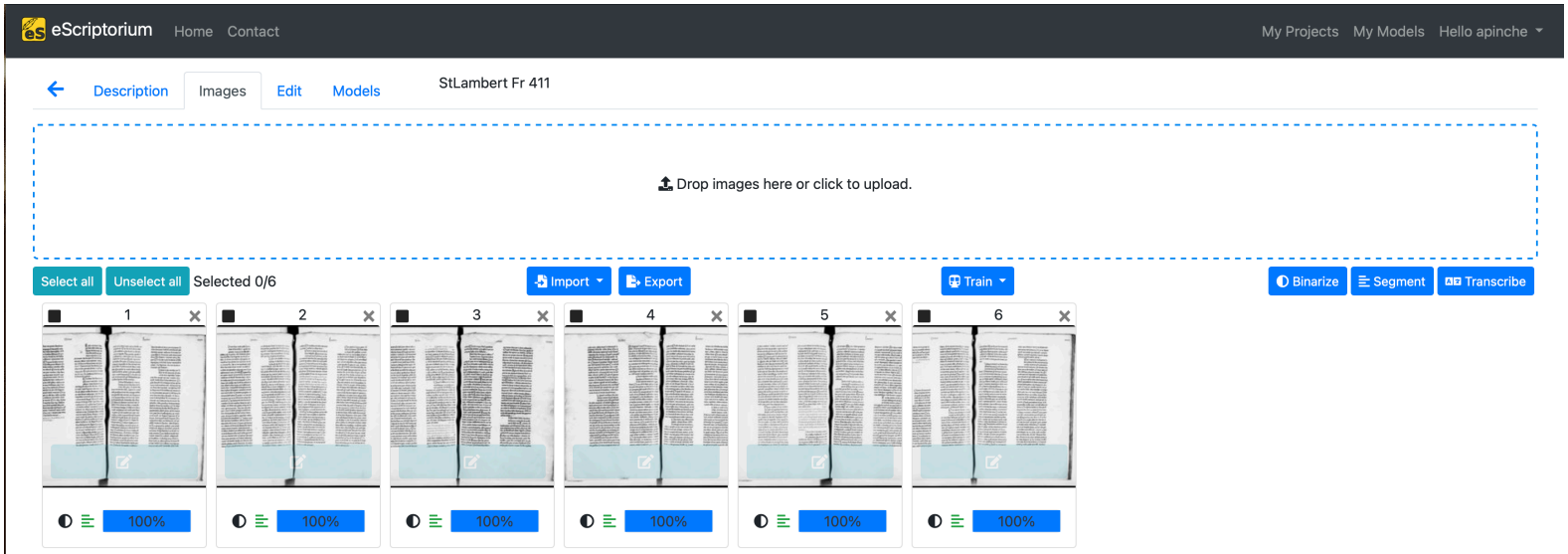


Why Use ATR Today?

- To accelerate the text acquisition phase. The prediction can serve:
  - as a basis for an edition: high level of accuracy, above 95%
  - for providing raw text: medium accuracy, between 90% and 95%
  - as a basis for quantitative analyses: lower accuracy, above 80% (see EDER, Maciej, "Mind your corpus: systematic errors in authorship attribution," *Literary and Linguistic Computing*, vol. 28/4, December 2013, pp. 603–614.)

The Steps of ATR

- Image loading
  - Load a collection of JPG or TIF images locally
  - Load from an IIIF manifest (e.g., collections from Gallica or e-Codices)
- Image preprocessing (optional)
  - 300 dpi resolution
  - Colour or grayscale
  - Optional linearisation to reduce noise
  - Multispectral imaging (in the case of heavily damaged documents)



The Steps of ATR

- Segmentation of the zones

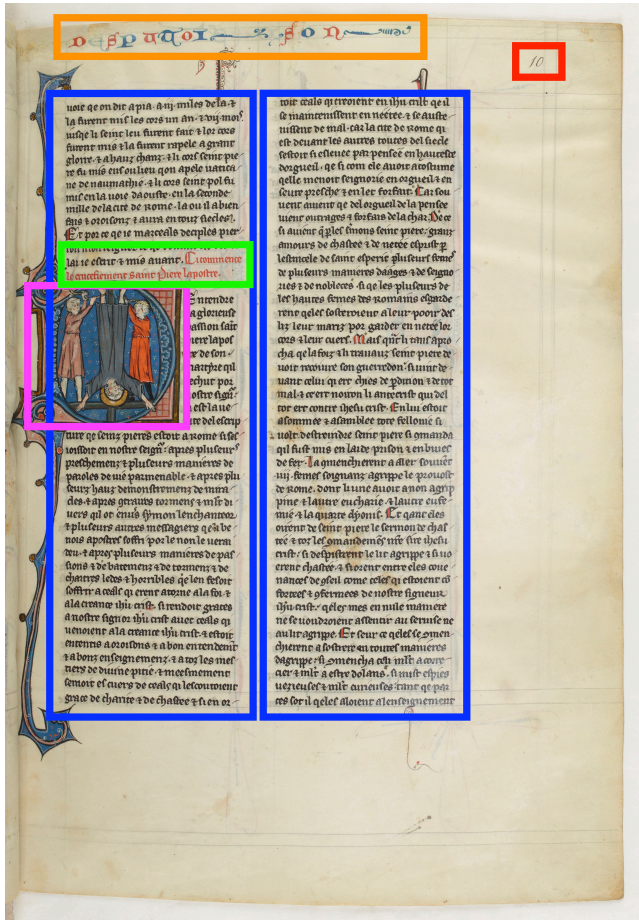


Figure: Bnf, fr. 412, fol.10r

The Steps of ATR

- Segmentation of the Text Lines

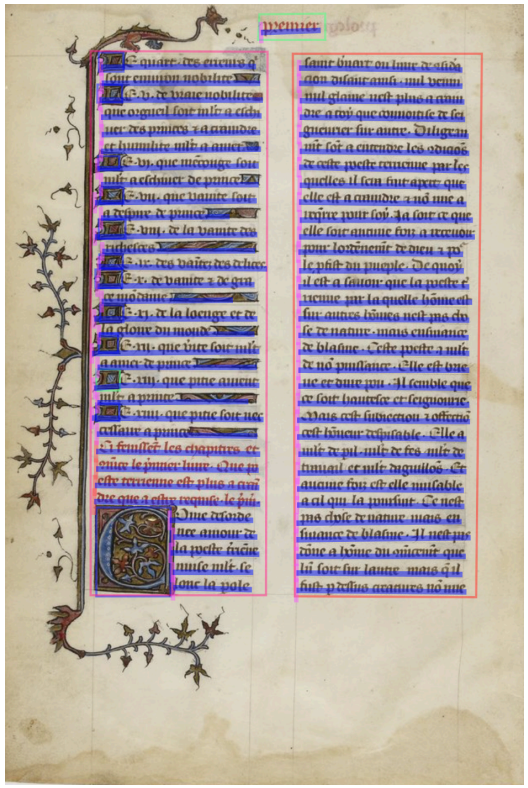
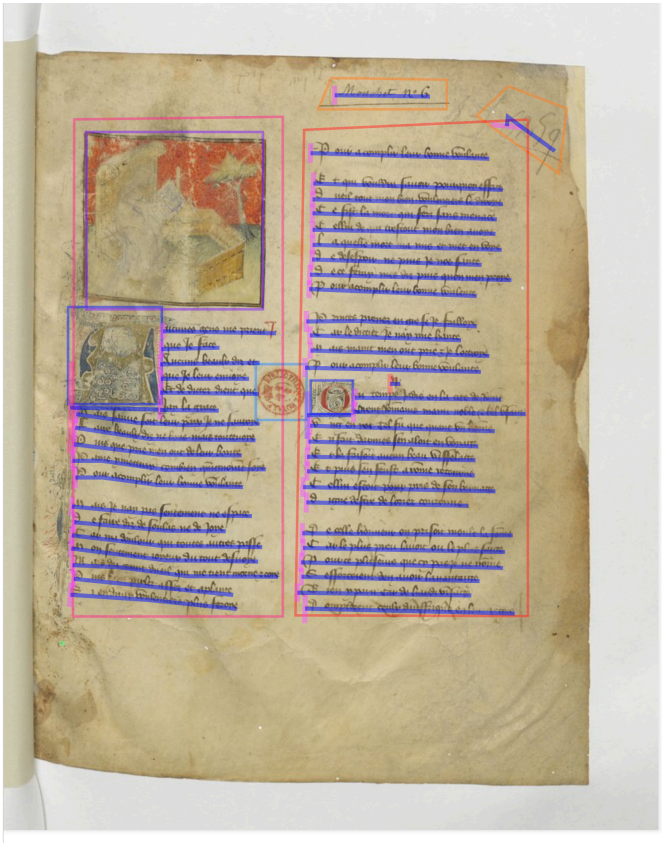


Figure: Bnf, fr. 1728, fol.8v

The Steps of ATR

- Text prediction



1	ucunes gens me prient ⁊.
2	que le face
3	Aucuns beaulx diz et
4	que le leur enuoye
5	Et de ditter dient que
6	A
7	Mais sauue soit leur paix le ne sauroye
8	lay la grace
9	Faire beaulx diz ne bons. mais touteuoye
10	Puis que prie men ont de leur bonte
11	Paine y mettray combien quĩgnouĩt soye
12	Pour acomplir leur bonne uoulante
13	Mais le n ay pas sentement ne espace
14	De faire diz. de soulas ne de loye
15	Car ma doulour qui toutes autres passe
16	Mon sentement ioyeux du tout desuoye
17	Mais du grant dueil qui me tient morne ⁊coye
18	Puis bien parler assez et aplante
19	Si en diray uoulentiers plus feroye
20	6259
21	Pour acomplir leur bonne uoulante
22	Et qui uouldra sauoir pourquoy efface
23	Dueil tout mon bien. uoulentiers le diroye
24	Ce fist la mort qui feri sans menace
25	Cellui de qui trestout mon bien auoye
26	Laquelle mort ma mis et met en uoye
27	De desespoir ne puis le nos sante
28	De ce feray mes diz puis quon men proye
29	Pour acomplir leur bonne uoulente
30	Princes prenez en gre se le failloye
31	Car le dittier le nay mie hante
32	Mais maint men ont prie ⁊ le lottroye
33	Pour acomplir leur bonne uoulante
34	u temps ladis en le cite de Rõme
35	.ii.
36	O
37	ung en yot. Tel fu que quant un hõme
38	Orent Rõmains maint noble ⁊ bel usaige

Figure: Bnf, fr. 12779, fol.9r

The Steps of ATR

- Data export (txt, alto, page)

```
<Layout>
        <Page WIDTH="4648" HEIGHT="3407" PHYSICAL_IMG_NR="8" ID="eSc_dummypage_">
                <PrintSpace HPOS="0" VPOS="0" WIDTH="4648" HEIGHT="3407">

                        <TextBlock HPOS="693" VPOS="321" WIDTH="1701" HEIGHT="2451"
                                ID="eSc_textblock_08b9f915" TAGREFS="BT3852">
                                <Shape>
                                        <Polygon
                                          POINTS="693 413 693 2772 2394 2772 2254 321"/>
                                </Shape>


                                <TextLine ID="eSc_line_d939596f" TAGREFS="LT1299"
                                        BASELINE="746 476 2143 428" HPOS="743" VPOS="352"
                                        WIDTH="1400" HEIGHT="156">
                                        <Shape>
                                          <Polygon
                                          POINTS="2078 388 2050 388 2021 386 1993 383 1964 383 1936 380 1908 377 1876 374 1848 374 1820 371 1811
                                          />
                                        </Shape>
                                        <String
                                          CONTENT="fors de la uille. Tant fut lassault merueilleux et"
                                          HPOS="743" VPOS="352" WIDTH="1400" HEIGHT="156"/>
                                </TextLine>
```

Figure: Example of ALTO xml

A Bit of Vocabulary

- **Training corpus**: A dataset composed of images and their corresponding line-by-line transcriptions, used to teach the model how to recognise text.
- **Model**: A file generated at the end of the training process, containing all optimised parameters. It allows the Automatic text recognition tool to predict text from previously unseen images.
- **Prediction**: The process by which the model uses images to predict the text, based on the parameters learned during training.
- **Training**: A sequence of cycles during which the model adjusts its parameters to improve its ability to correctly recognise and predict text from images.

Training an ATR Model

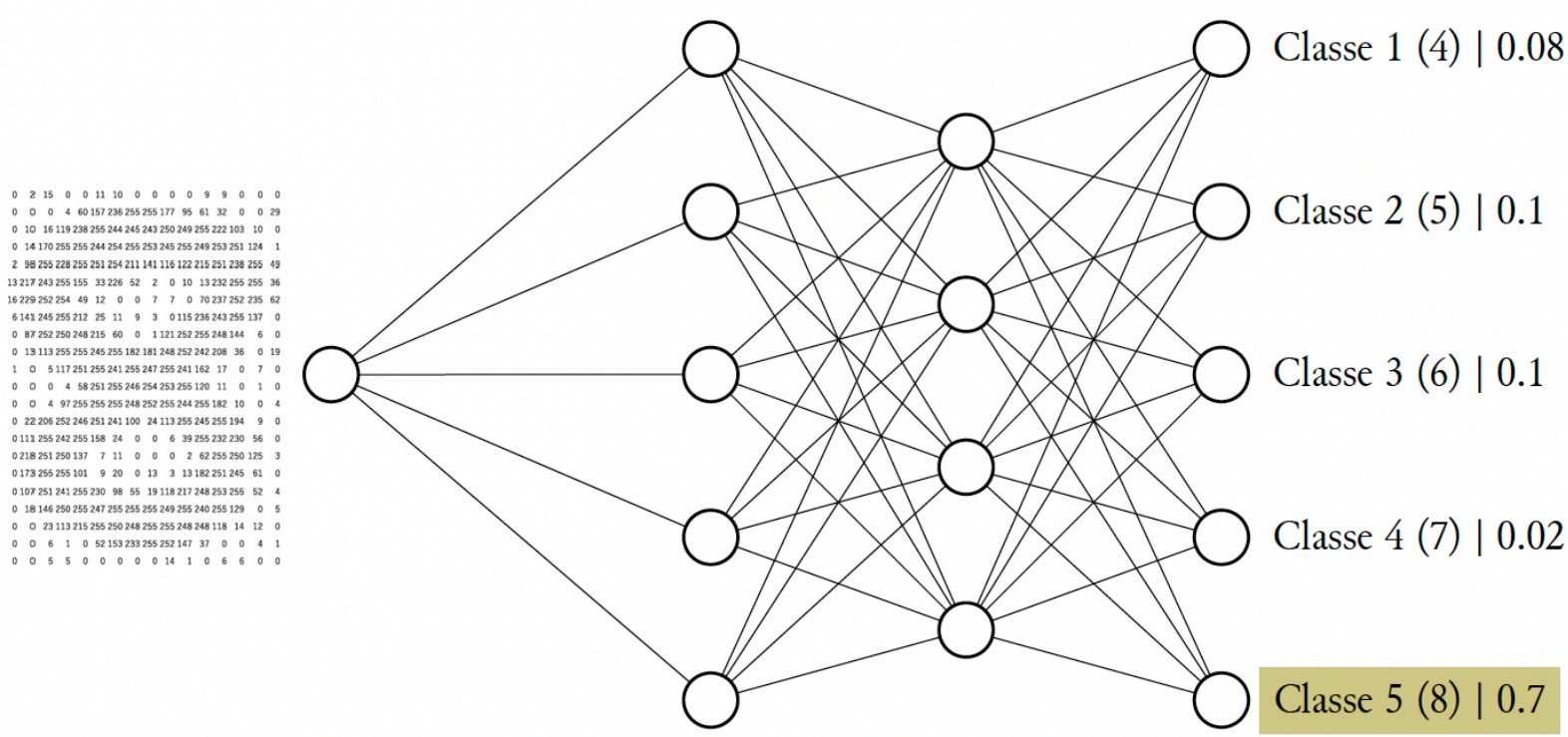


Figure: Simplified representation of a neural network

Training an ATR Model

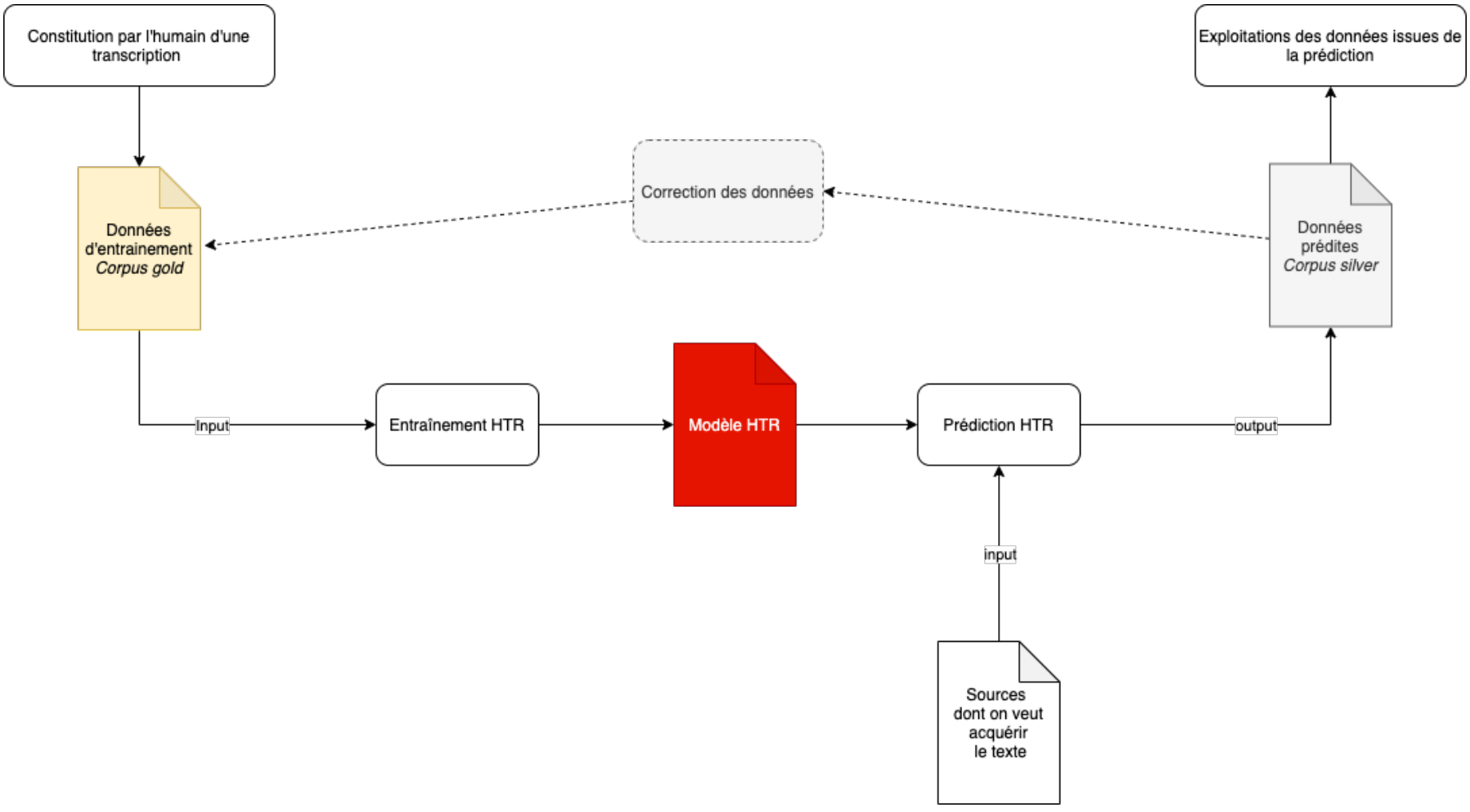


Figure: Representation of a training cycle

Training an ATR Model

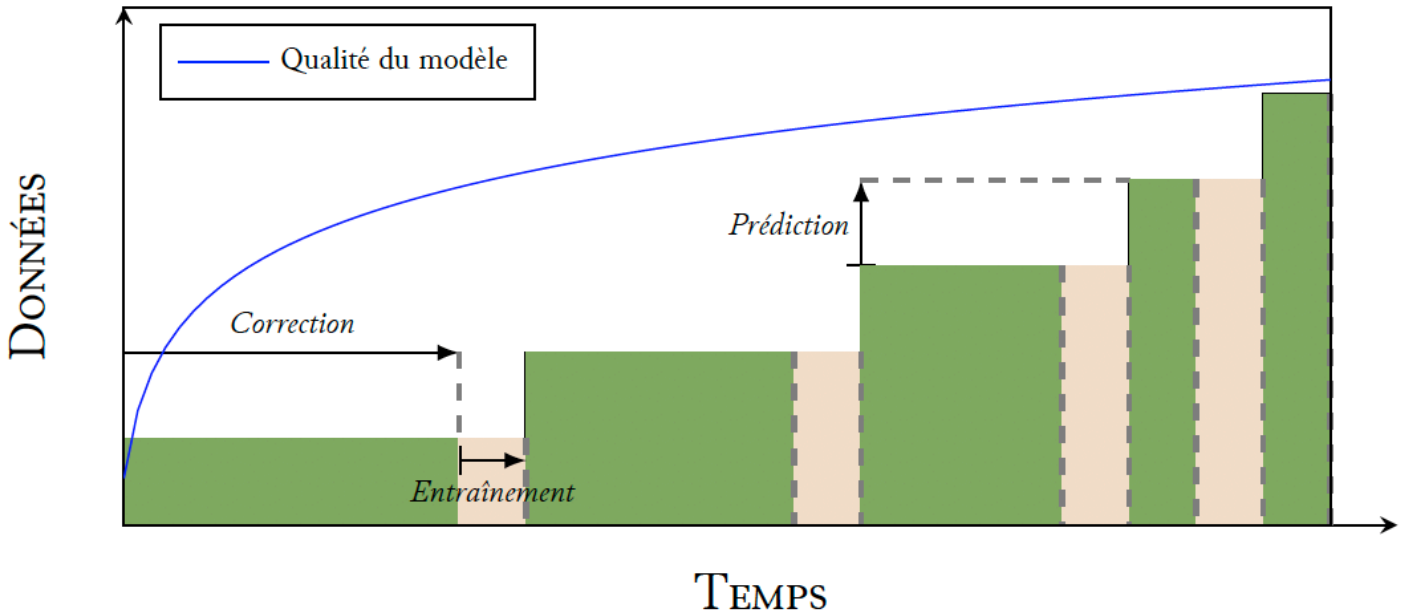
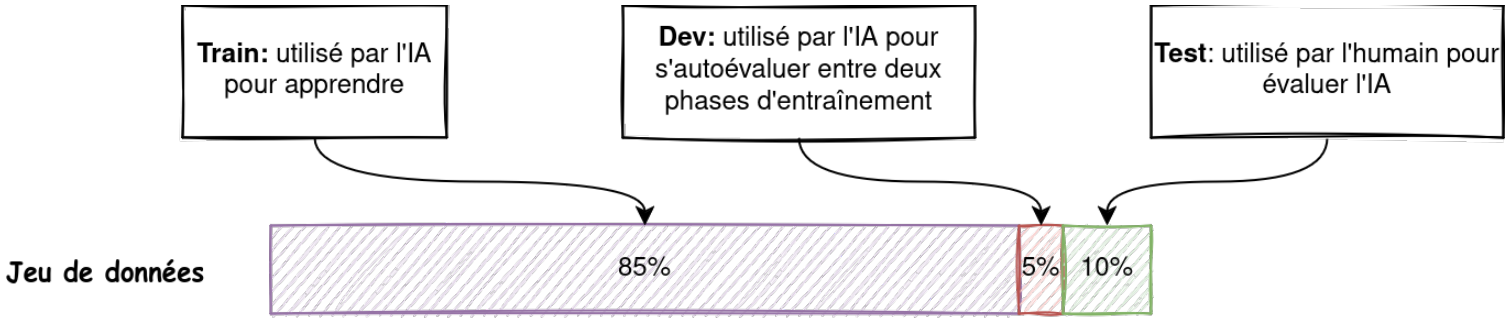
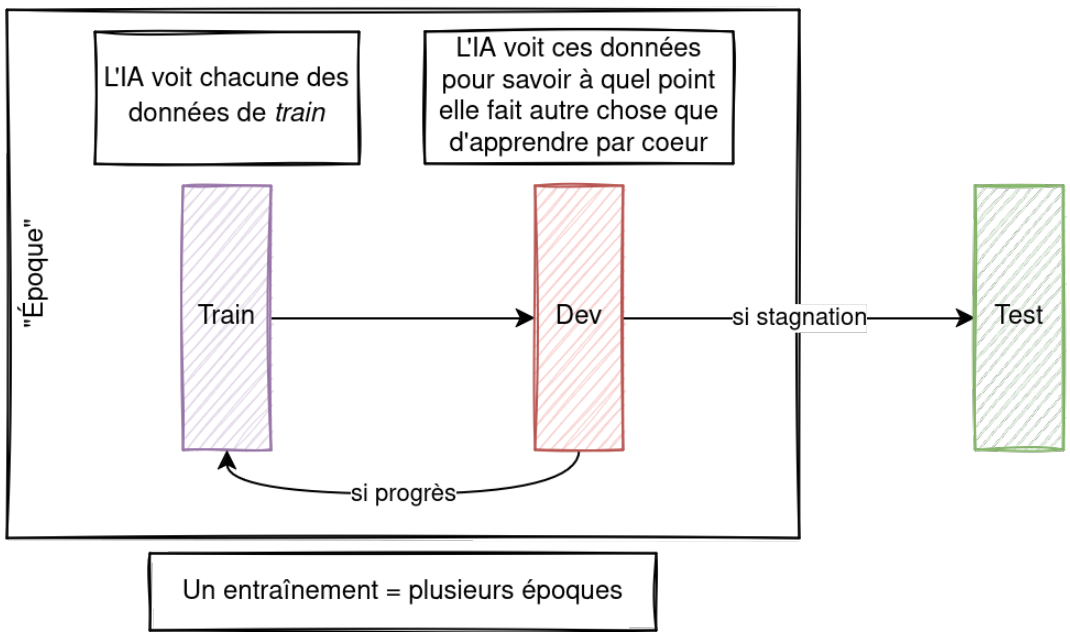


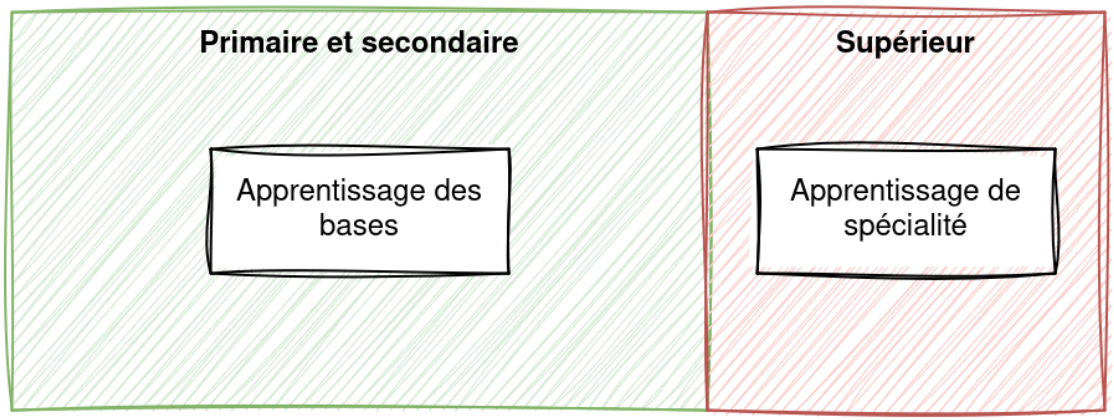
Figure: Evolution of data correction time depending on the model quality, figure from the work of M. Gille Levenson

Train? Dev? Test? The splits.





Fine-tuning?



Vous rappelez vous de tout ce que vous avez appris au lycée ?

En particulier ce qui n'a aucun rapport avec votre spécialité ?

Oubli d'une partie de (1) mais amélioration sur (2)

Evaluating an ATR Model

To evaluate an ATR model:

- prepare the dataset
  - train set (80%) – training
  - dev set (10%) – evaluation during training cycles
  - test set (10%) – data never seen during training
- compare:
  - a ground truth (GT) produced by a human (test set)
  - with the model's prediction of the same lines
  - to calculate a score which can be either:
    - a CER (Character Error Rate)
    - or an Accuracy (the percentage of correct predictions by the model)

Types of Errors

Calculating the CER

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

Where to Find Models?

- A performant model already exists. Where to find it? Where to find data to train your own?

  Data HTR-united

  Models Zenodo Community “OCR/HTR Models”

Table of Contents

ATR and Historical Documents

- Deciphering handwritten scripts in historical documents presents unique challenges:
  - Non-standardised layouts
  - Degraded supports
  - Irregular handwriting(s)
  - Graphic and/or dialectal variations

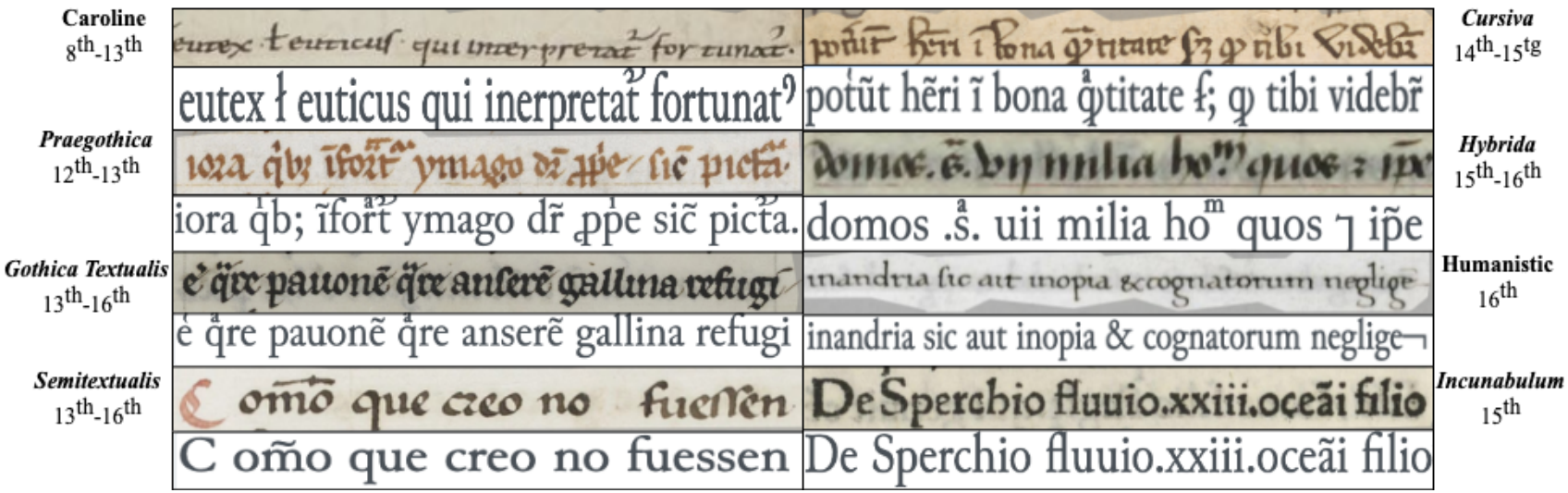


Figure: Sample of medieval scripts

Non-unified Sources

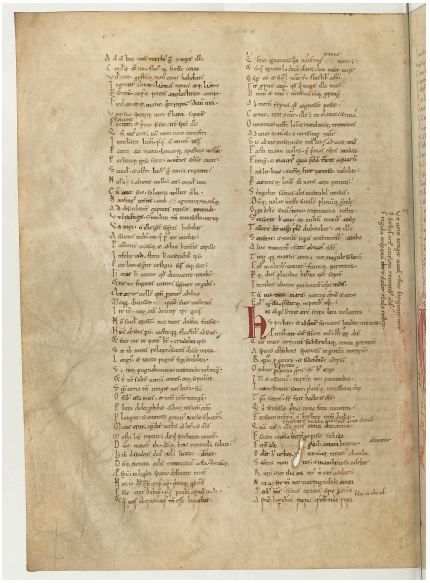


Figure: BnF, Latin, 8001, 13th century

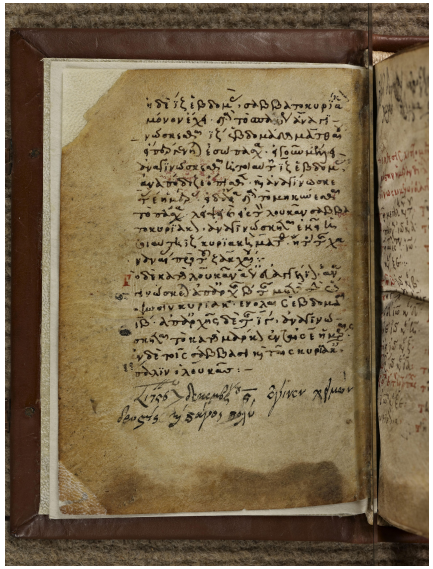


Figure: Strasbourg, ms. 1.916, 13th century

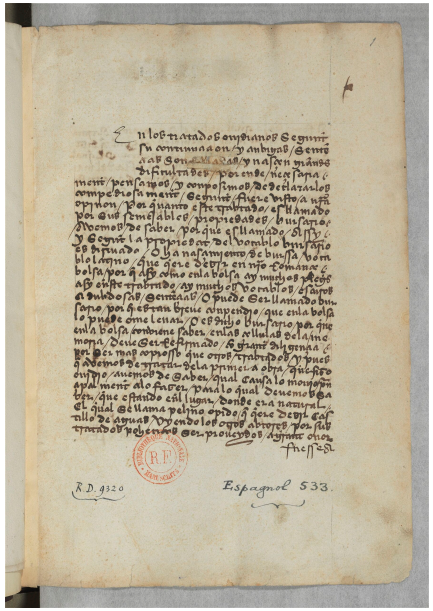


Figure: BnF, Spanish, 533, 15th century

How to Transcribe Documents?

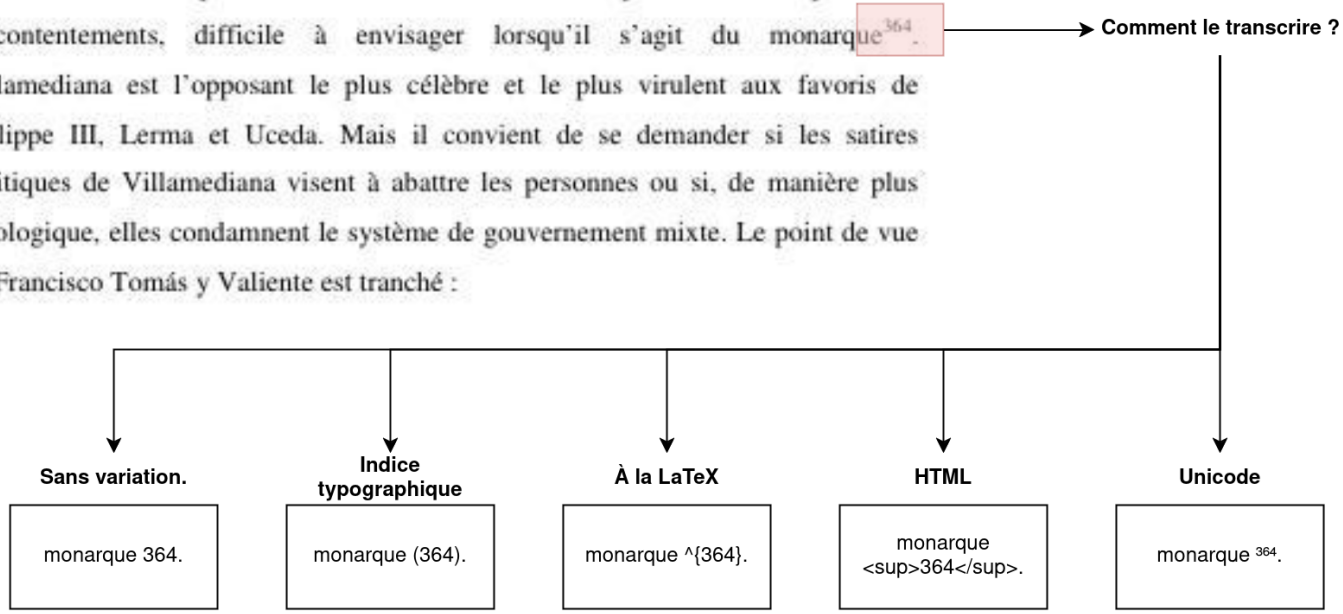
- Transcribing means describing a source
- Transcribing means translating the source for non-palaeographers
- Transcribing means interpreting the text
- Transcribing means making choices

  - "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 Hodel, David Schoch, Christa Schneider, [et al.], "General Models for Handwritten Text Recognition: Feasibility and State-of-the-Art. German Kurrent as an Example," 2021.

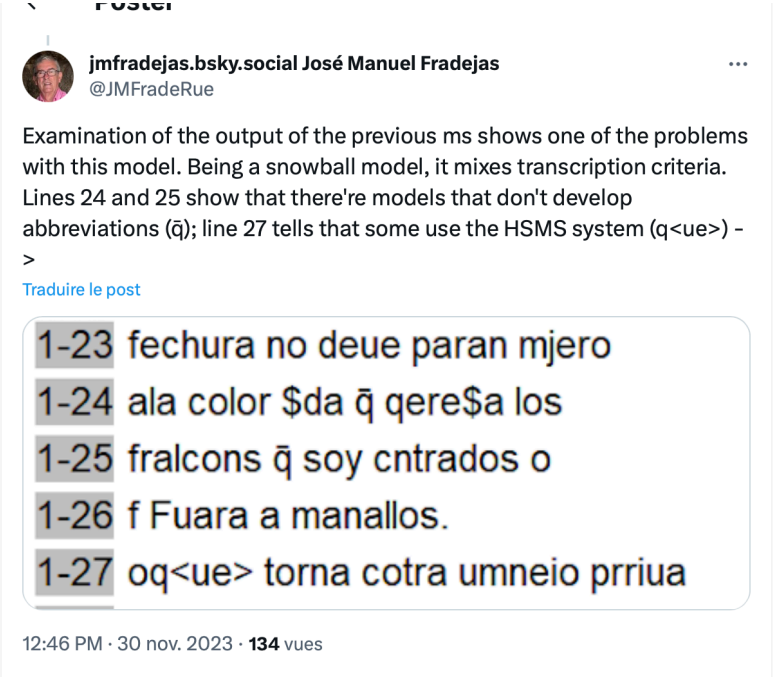
Making Choices and Documenting Them

Par ailleurs, comme le fait remarquer Francisco Tomás y Valiente, cette instance humaine de pouvoir, non sacralisée, favorise l'expression de critiques et de mécontentements, difficile à envisager lorsqu'il s'agit du monarque[364]. Villamediana est l'opposant le plus célèbre et le plus virulent aux favoris de Philippe III, Lerma et Uceda. Mais il convient de se demander si les satires politiques de Villamediana visent à abattre les personnes ou si, de manière plus idéologique, elles condamnent le système de gouvernement mixte. Le point de vue de Francisco Tomás y Valiente est tranché :



How to Transcribe Documents?

- How to transcribe documents for the machine?
- How to transcribe consistently within a project?
- How to transcribe so that my data can be reused?



How to Transcribe Documents?

- Define transcription methods adapted to the research question and machine learning.
- Define the desired level of precision in transcription
- Use a predefined character set and document your choices.
- Ensure compatibility of transcription data.

Documents and Various Layouts

Figure: Turin Manuscript, Ségurant the Knight and the Dragon, 15th century

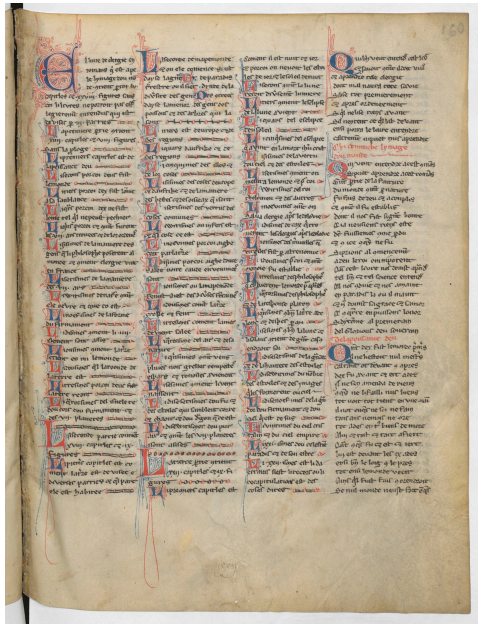


Figure: BnF, Arsenal, 3516, 12th century

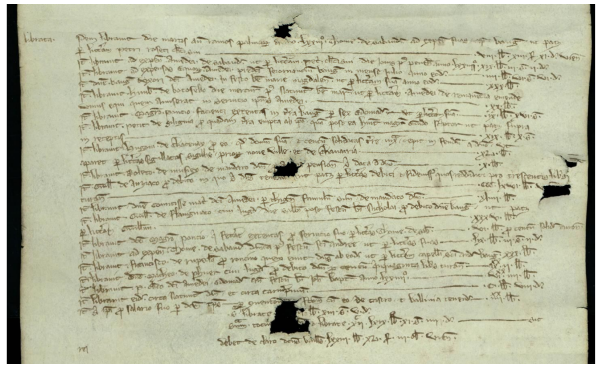


Figure: Departmental Archives of Côte d'Or, B6739, 13th century

Layout Analysis: Segmentation

- Identification of the different zones in the document: use controlled vocabulary such as SegmOnto.

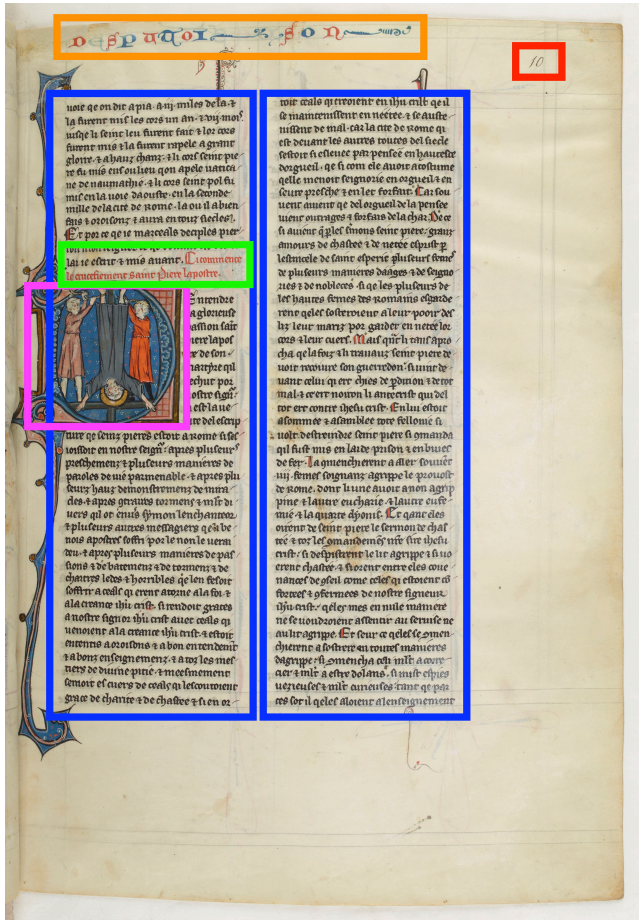
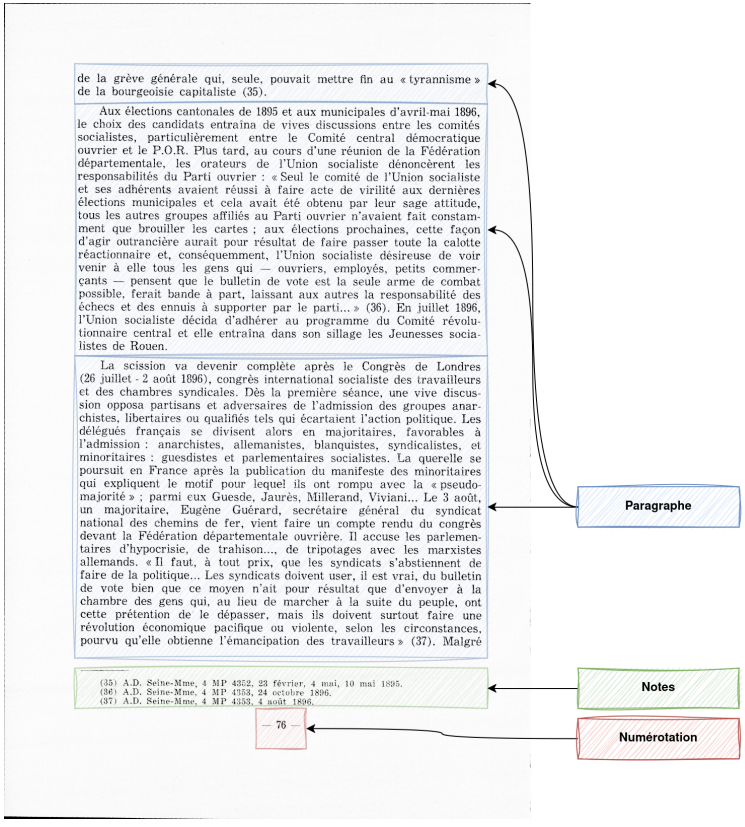
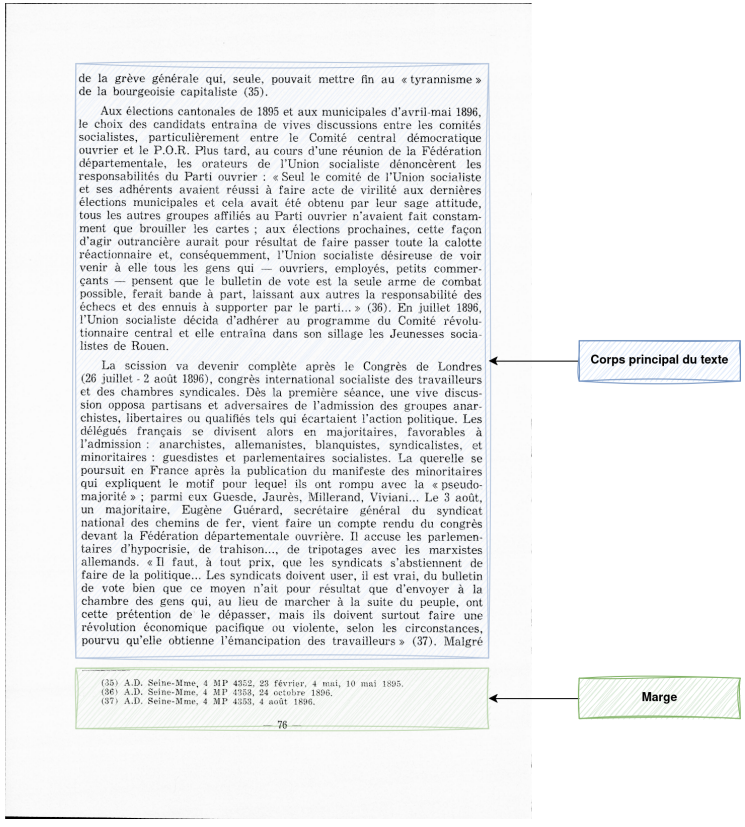


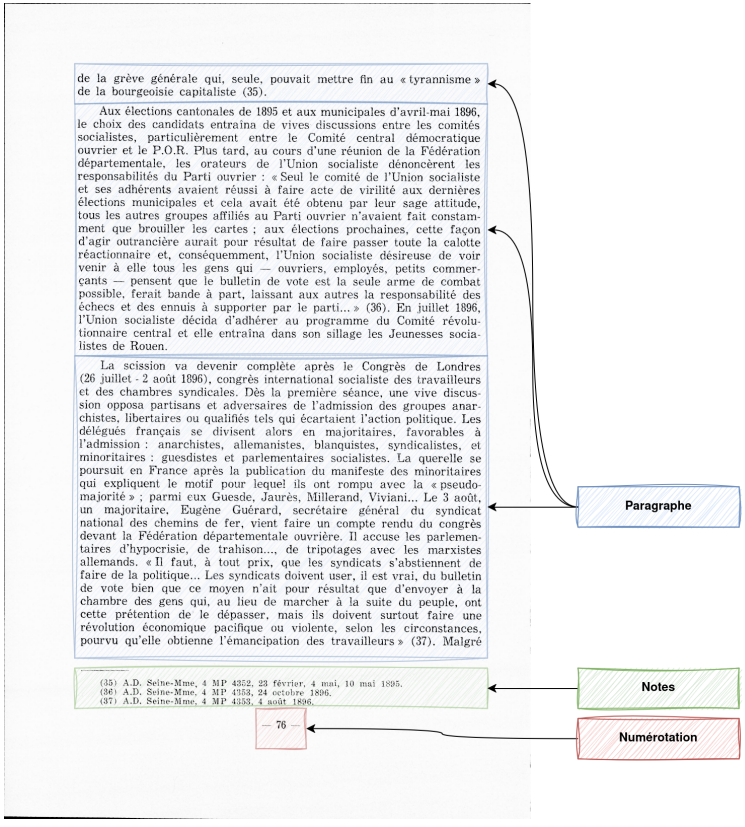
Figure: BnF, fr. 412, fol. 10r

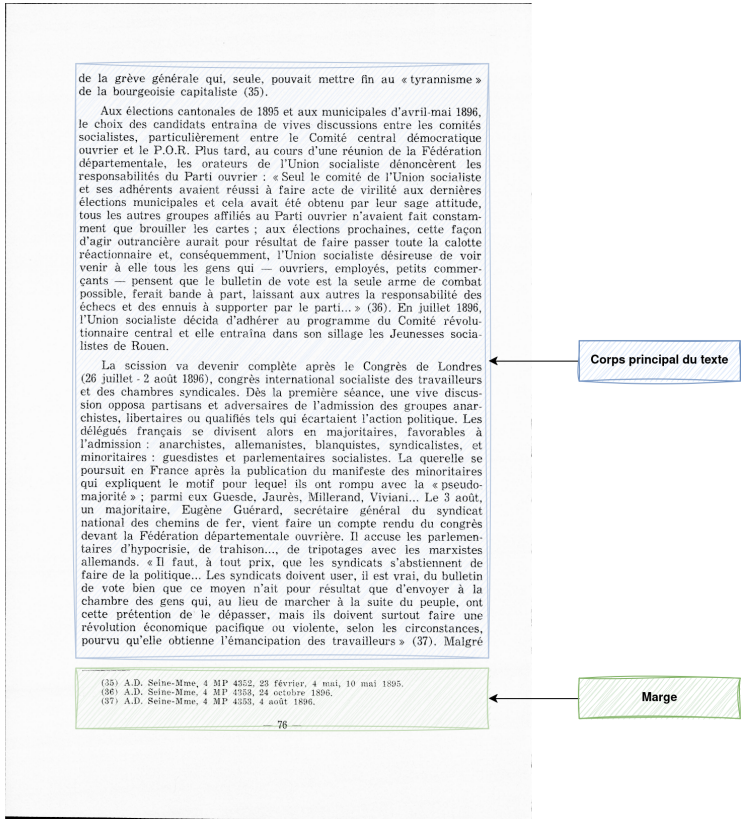
Options?


de la grève générale qui, seule, pouvait mettre fin au « tyrannisme » de la bourgeoisie capitaliste (35).
Aux élections cantonales de 1895 et aux municipales d'avril-mai 1896, le choix des candidats entraîna de vives discussions entre les comités socialistes, particulièrement entre le Comité central démocratique ouvrier et le P.O.R. Plus tard, au cours d'une réunion de la Fédération départementale, les orateurs de l'Union socialiste dénoncèrent les responsabilités du Parti ouvrier : « Seul le comité de l'Union socialiste et ses adhérents avaient réussi à faire acte de virilité aux dernières élections municipales et cela avait été obtenu par leur sage attitude, tous les autres groupes affiliés au Parti ouvrier n'avaient fait constamment que brouiller les cartes ; aux élections prochaines, cette façon d'agir outrancière aurait pour résultat de faire passer toute la calotte réactionnaire et, conséquemment, l'Union socialiste désireuse de voir venir à elle tous les gens qui — ouvriers, employés, petits commerçants — pensent que le bulletin de vote est la seule arme de combat possible, ferait bande à part, laissant aux autres la responsabilité des échecs et des ennuis à supporter par le parti... » (36). En juillet 1896, l'Union socialiste décida d'adhérer au programme du Comité révolutionnaire central et elle entraîna dans son sillage les Jeunesses socialistes de Rouen.
La scission va devenir complète après le Congrès de Londres (26 juillet - 2 août 1896), congrès international socialiste des travailleurs et des chambres syndicales. Dès la première séance, une vive discussion opposa partisans et adversaires de l'admission des groupes anarchistes, libertaires ou qualifiés tels qui écartaient l'action politique. Les délégués français se divisent alors en majoritaires, favorables à l'admission : anarchistes, allemanistes, blanquistes, syndicalistes, et minoritaires : guesdistes et parlementaires socialistes. La querelle se poursuit en France après la publication du manifeste des minoritaires qui expliquent le motif pour lequel ils ont rompu avec la « pseudo-majorité » ; parmi eux Guesde, Jaurès, Millerand, Viviani... Le 3 août, un majoritaire, Eugène Guérard, secrétaire général du syndicat national des chemins de fer, vient faire un compte rendu du congrès devant la Fédération départementale ouvrière. Il accuse les parlementaires d'hypocrisie, de trahison..., de tripotages avec les marxistes allemands. « Il faut, à tout prix, que les syndicats s'abstiennent de faire de la politique... Les syndicats doivent user, il est vrai, du bulletin de vote bien que ce moyen n'ait pour résultat que d'envoyer à la chambre des gens qui, au lieu de marcher à la suite du peuple, ont cette prétention de le dépasser, mais ils doivent surtout faire une révolution économique pacifique ou violente, selon les circonstances, pourvu qu'elle obtienne l'émancipation des travailleurs » (37). Malgré
(35) A.D. Seine-Mme, 4 MP 4352, 23 février, 4 mai, 10 mai 1895.
(36) A.D. Seine-Mme, 4 MP 4353, 24 octobre 1896.
(37) A.D. Seine-Mme, 4 MP 4353, 4 août 1896.
— 76 —
Paragraphe
Notes
Numérotation


de la grève générale qui, seule, pouvait mettre fin au « tyrannisme » de la bourgeoisie capitaliste (35).
Aux élections cantonales de 1895 et aux municipales d'avril-mai 1896, le choix des candidats entraîna de vives discussions entre les comités socialistes, particulièrement entre le Comité central démocratique ouvrier et le P.O.R. Plus tard, au cours d'une réunion de la Fédération départementale, les orateurs de l'Union socialiste dénoncèrent les responsabilités du Parti ouvrier : « Seul le comité de l'Union socialiste et ses adhérents avaient réussi à faire acte de virilité aux dernières élections municipales et cela avait été obtenu par leur sage attitude, tous les autres groupes affiliés au Parti ouvrier n'avaient fait constamment que brouiller les cartes ; aux élections prochaines, cette façon d'agir outrancière aurait pour résultat de faire passer toute la calotte réactionnaire et, conséquemment, l'Union socialiste désireuse de voir venir à elle tous les gens qui — ouvriers, employés, petits commerçants — pensent que le bulletin de vote est la seule arme de combat possible, ferait bande à part, laissant aux autres la responsabilité des échecs et des ennuis à supporter par le parti... » (36). En juillet 1896, l'Union socialiste décida d'adhérer au programme du Comité révolutionnaire central et elle entraîna dans son sillage les Jeunesses socialistes de Rouen.
La scission va devenir complète après le Congrès de Londres (26 juillet - 2 août 1896), congrès international socialiste des travailleurs et des chambres syndicales. Dès la première séance, une vive discussion opposa partisans et adversaires de l'admission des groupes anarchistes, libertaires ou qualifiés tels qui écartaient l'action politique. Les délégués français se divisent alors en majoritaires, favorables à l'admission : anarchistes, allemanistes, blanquistes, syndicalistes, et minoritaires : guesdistes et parlementaires socialistes. La querelle se poursuit en France après la publication du manifeste des minoritaires qui expliquent le motif pour lequel ils ont rompu avec la « pseudo-majorité » ; parmi eux Guesde, Jaurès, Millerand, Viviani... Le 3 août, un majoritaire, Eugène Guérard, secrétaire général du syndicat national des chemins de fer, vient faire un compte rendu du congrès devant la Fédération départementale ouvrière. Il accuse les parlementaires d'hypocrisie, de trahison..., de tripotages avec les marxistes allemands. « Il faut, à tout prix, que les syndicats s'abstiennent de faire de la politique... Les syndicats doivent user, il est vrai, du bulletin de vote bien que ce moyen n'ait pour résultat que d'envoyer à la chambre des gens qui, au lieu de marcher à la suite du peuple, ont cette prétention de le dépasser, mais ils doivent surtout faire une révolution économique pacifique ou violente, selon les circonstances, pourvu qu'elle obtienne l'émancipation des travailleurs » (37). Malgré
(35) A.D. Seine-Mme, 4 MP 4352, 23 février, 4 mai, 10 mai 1895.
(36) A.D. Seine-Mme, 4 MP 4353, 24 octobre 1896.
(37) A.D. Seine-Mme, 4 MP 4353, 4 août 1896.
— 76 —
Corps principal du texte
Marge

Options ?


de la grève générale qui, seule, pouvait mettre fin au « tyrannisme » de la bourgeoisie capitaliste (35).
Aux élections cantonales de 1895 et aux municipales d'avril-mai 1896, le choix des candidats entraîna de vives discussions entre les comités socialistes, particulièrement entre le Comité central démocratique ouvrier et le P.O.R. Plus tard, au cours d'une réunion de la Fédération départementale, les orateurs de l'Union socialiste dénoncèrent les responsabilités du Parti ouvrier : « Seul le comité de l'Union socialiste et ses adhérents avaient réussi à faire acte de virilité aux dernières élections municipales et cela avait été obtenu par leur sage attitude, tous les autres groupes affiliés au Parti ouvrier n'avaient fait constamment que brouiller les cartes ; aux élections prochaines, cette façon d'agir outrancière aurait pour résultat de faire passer toute la calotte réactionnaire et, conséquemment, l'Union socialiste désireuse de voir venir à elle tous les gens qui — ouvriers, employés, petits commerçants — pensent que le bulletin de vote est la seule arme de combat possible, ferait bande à part, laissant aux autres la responsabilité des échecs et des ennuis à supporter par le parti... » (36). En juillet 1896, l'Union socialiste décida d'adhérer au programme du Comité révolutionnaire central et elle entraîna dans son sillage les Jeunesses socialistes de Rouen.
La scission va devenir complète après le Congrès de Londres (26 juillet - 2 août 1896), congrès international socialiste des travailleurs et des chambres syndicales. Dès la première séance, une vive discussion opposa partisans et adversaires de l'admission des groupes anarchistes, libertaires ou qualifiés tels qui écartaient l'action politique. Les délégués français se divisent alors en majoritaires, favorables à l'admission : anarchistes, allemanistes, blanquistes, syndicalistes, et minoritaires : guesdistes et parlementaires socialistes. La querelle se poursuit en France après la publication du manifeste des minoritaires qui expliquent le motif pour lequel ils ont rompu avec la « pseudo-majorité » ; parmi eux Guesde, Jaurès, Millerand, Viviani... Le 3 août, un majoritaire, Eugène Guérard, secrétaire général du syndicat national des chemins de fer, vient faire un compte rendu du congrès devant la Fédération départementale ouvrière. Il accuse les parlementaires d'hypocrisie, de trahison..., de tripotages avec les marxistes allemands. « Il faut, à tout prix, que les syndicats s'abstiennent de faire de la politique... Les syndicats doivent user, il est vrai, du bulletin de vote bien que ce moyen n'ait pour résultat que d'envoyer à la chambre des gens qui, au lieu de marcher à la suite du peuple, ont cette prétention de le dépasser, mais ils doivent surtout faire une révolution économique pacifique ou violente, selon les circonstances, pourvu qu'elle obtienne l'émancipation des travailleurs » (37). Malgré
(35) A.D. Seine-Mme, 4 MP 4352, 23 février, 4 mai, 10 mai 1895.
(36) A.D. Seine-Mme, 4 MP 4353, 24 octobre 1896.
(37) A.D. Seine-Mme, 4 MP 4353, 4 août 1896.
— 76 —
Paragraphe
Notes
Numérotation


de la grève générale qui, seule, pouvait mettre fin au « tyrannisme » de la bourgeoisie capitaliste (35).
Aux élections cantonales de 1895 et aux municipales d'avril-mai 1896, le choix des candidats entraîna de vives discussions entre les comités socialistes, particulièrement entre le Comité central démocratique ouvrier et le P.O.R. Plus tard, au cours d'une réunion de la Fédération départementale, les orateurs de l'Union socialiste dénoncèrent les responsabilités du Parti ouvrier : « Seul le comité de l'Union socialiste et ses adhérents avaient réussi à faire acte de virilité aux dernières élections municipales et cela avait été obtenu par leur sage attitude, tous les autres groupes affiliés au Parti ouvrier n'avaient fait constamment que brouiller les cartes ; aux élections prochaines, cette façon d'agir outrancière aurait pour résultat de faire passer toute la calotte réactionnaire et, conséquemment, l'Union socialiste désireuse de voir venir à elle tous les gens qui — ouvriers, employés, petits commerçants — pensent que le bulletin de vote est la seule arme de combat possible, ferait bande à part, laissant aux autres la responsabilité des échecs et des ennuis à supporter par le parti... » (36). En juillet 1896, l'Union socialiste décida d'adhérer au programme du Comité révolutionnaire central et elle entraîna dans son sillage les Jeunesses socialistes de Rouen.
La scission va devenir complète après le Congrès de Londres (26 juillet - 2 août 1896), congrès international socialiste des travailleurs et des chambres syndicales. Dès la première séance, une vive discussion opposa partisans et adversaires de l'admission des groupes anarchistes, libertaires ou qualifiés tels qui écartaient l'action politique. Les délégués français se divisent alors en majoritaires, favorables à l'admission : anarchistes, allemanistes, blanquistes, syndicalistes, et minoritaires : guesdistes et parlementaires socialistes. La querelle se poursuit en France après la publication du manifeste des minoritaires qui expliquent le motif pour lequel ils ont rompu avec la « pseudo-majorité » ; parmi eux Guesde, Jaurès, Millerand, Viviani... Le 3 août, un majoritaire, Eugène Guérard, secrétaire général du syndicat national des chemins de fer, vient faire un compte rendu du congrès devant la Fédération départementale ouvrière. Il accuse les parlementaires d'hypocrisie, de trahison..., de tripotages avec les marxistes allemands. « Il faut, à tout prix, que les syndicats s'abstiennent de faire de la politique... Les syndicats doivent user, il est vrai, du bulletin de vote bien que ce moyen n'ait pour résultat que d'envoyer à la chambre des gens qui, au lieu de marcher à la suite du peuple, ont cette prétention de le dépasser, mais ils doivent surtout faire une révolution économique pacifique ou violente, selon les circonstances, pourvu qu'elle obtienne l'émancipation des travailleurs » (37). Malgré
(35) A.D. Seine-Mme, 4 MP 4352, 23 février, 4 mai, 10 mai 1895.
(36) A.D. Seine-Mme, 4 MP 4353, 24 octobre 1896.
(37) A.D. Seine-Mme, 4 MP 4353, 4 août 1896.
— 76 —
Corps principal du texte
Marge

SegmOnto

https://segmonto.github.io

Page

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

Line

- DefaultLine
- Interlinear
- HeadingLine

To go further...

See extended dataset LADaS :

Thibault Clérice et al. *Layout Analysis Dataset with SegmOnto (LADaS)*. URL: https://github.com/DEFI-COLaF/LADaS

Segmentation and Historical Documents

Today, segmentation is the main source of errors in Handwritten Text Recognition (HTR):

- Incorrect identification of zones
- Incorrect labelling of zone
- Merging of nearby zones
- Lines not linked to the correct zone
- Since 2011. This task has been the focus of competitions held at ICDAR and HIP conferences.

Sharing Data

Harmonising data allows the exchange of data and HTR models. How to achieve this?

- Define transcription methods adapted to your research questions and machine learning.
- Use a predefined character set and document your choices.
  - See the Medieval Unicode Font Initiative (MUFI)
  - See the transcription guidelines proposed by CREMMALab for medieval texts, and the CREMMA guidelines for modern transcriptions.
- Use controlled vocabulary to describe the layout and document it.

Making Your Data Available

- Deposit your data on an accessible online repository:
  - Github
  - GitLab
- Document your data:
  - Data format
  - Number of transcribed lines
  - Segmentation tools
  - HTR engine
  - Corpus language
  - Date
  - Document and handwriting type
  - Transcription method
- Increase data visibility: integrate into a catalogue, see HTR-united.