

Journée d'étude

Le Trévoux à l'ère des humanités numériques et de l'IA : expérimentations et résultats préliminaires

IA pour la segmentation et l'OCR de dictionnaires anciens

Aya Rahouti

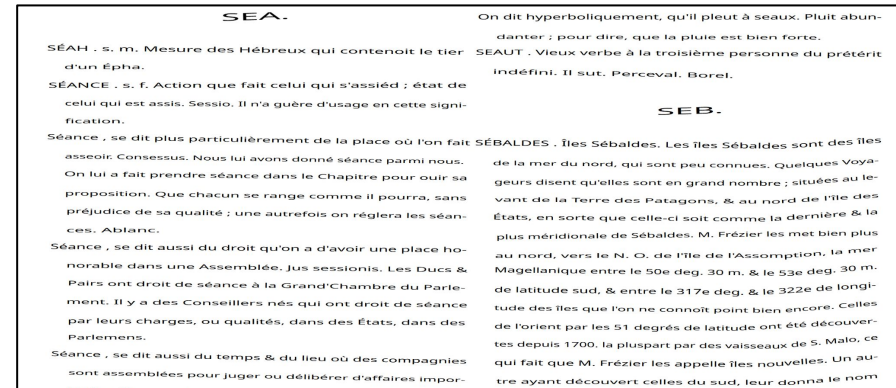
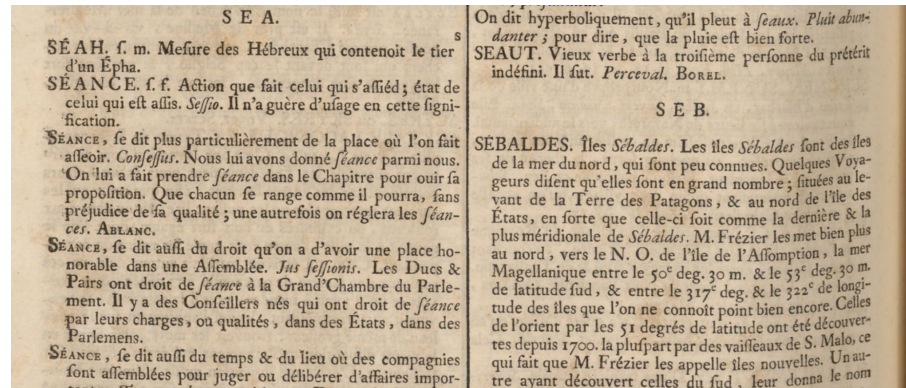
Ludovic Moncla, Véronique Eglin



Introduction

- Édition 1743 du Dictionnaire de Trévoux : environ 956 pages.
- Stage au LIRIS (INSA Lyon / CNRS) : concevoir une chaîne de traitement OCR + segmentation pour le Trévoux.

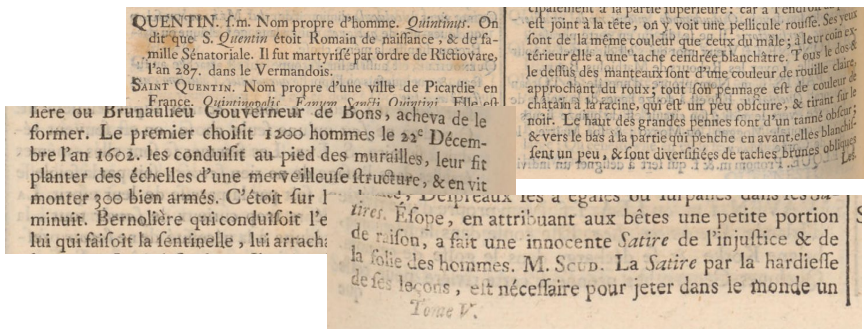
Objectif : passer d'images de pages numérisées à un texte fiable et structuré.



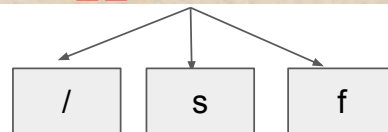
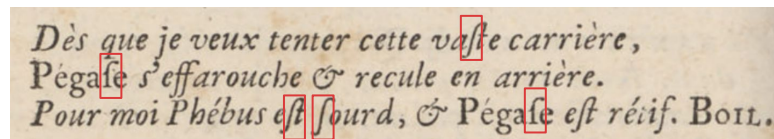
Contexte et problématique

Pourquoi le Trévoux 1743 exige-t-il une chaîne OCR dédiée ?

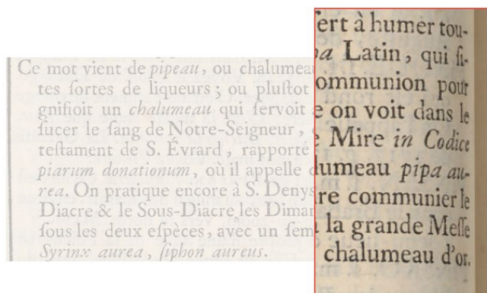
► Qualité du support :



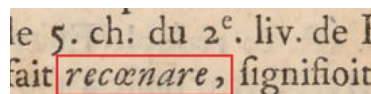
► s long: f f



► Courbures de reliure :



► Ligatures :



→ recœnare

→ recœnare

æ = a+e (ex. *præfatio*, *cælum*) æ moderne

œ = o+e (ex. *cœur*, *œuvre*) œ moderne

► **Français ancien:**

Français + latin + grec + hébreu parfois

parties, à en teparer les mauvaises, à en altérer & exalter les bonnes. L'emploi de l'Apôtre est le Pharmacie. Ce mot vient du Gr. **Φάρμακον** remède.

57. PHARMACIEN. C'est celui qui exerce la Pharmacie. Un bon Pharmacieen, un excellent Pharmacieen.

PHARMACOPÉE. (f. Traité qui enseigne la Pharmacie, & qui contient la préparation des remèdes. Il y a plusieurs livres de Pharmacopée, comme Bauderon, Quercetan, Zwelfer, Charas, &c.)

Ce mot vient du Gr. **Φάρμακον** remède, & de **ποιέω**, faire.

PHARMACOPOLITE. C'est un pharmacien qui prépare, & qui vend les remèdes. On ne se sert de ce mot qu'en dérision & en burlesque.

Ce mot vient du Gr. **Φάρμακον** signifie remède, **πολίτης** remède.

PHARMUTHI. C'm Nom d'un mois des anciens Egyptiens. Pharmuthi. Il répondoit au mois d'Avril de l'année

rien que qu'on la mine est pleine, nous sommes entre elles & le soleil.

P H A R E. f. m. Mot que S. Jérôme dans sa traduction de l'Écriture a fait sur l'Hebreu **פָּרָה** parah, ou manière dont il a exprimé ce nom Hebreu. On a quelquefois confondu ce mot dans les Differtations d'érudition, au lieu de celui de Paque qui est plus ordinaire, & qui est fait sur les Traducteurs Grecs par le même mot Hebreu **פָּרָה** parah. Vous immolez la paque **פָּרָה** le soir au 14^e du mois d'Avril, c'est à l'heure que vous êtes sorti d'Egypte. Il faut remonter plus que ce mot **פָּרָה**, qui est vers la fin de ce passage, & que ce **פָּרָה**, ou la victime que Dieu ordonne d'immoler au soleil couchant à l'heure que les Hebreux sortirent d'Egypte, n'est point l'agneau qu'on appelloit aussi **פָּרָה**, ou paque, & qui devoit être immolé le 14^e du mois au soir; mais que c'étoit une autre victime d'un sacrifice d'Éucharistie, ou d'action de grâces, en mémoire de ce qu'ils

seux rompus; & atefois ils arrivèrent heureusement à Naples. FLURY. Voyez le mot *Doctar*.
Ce nom est une phrase Latine qui veut dire, *Ce que veut Dieu*. C'est apparemment, comme on l'a déjà infinué de quelques autres noms semblables, une traduction Latine du véritable nom de ce Saint qui étoit Punique; par exemple, il pouvoit s'appeler *Quid-ut-vult Deus*, ou *Quid-ut-vult Deus*. Cet autr que les Hébreux appellent un Auteur Egyptien *Maschabalach*, bon Arabe composé de trois mots, et qui signifie la même chose que *Quid-ut-vult Deus*. Voyez la Bibliothèque Orientale de JHerbelot,

Par un je ne sai quoi qu'on ne peut expliquer. CORN.

QUOIQUE. Conjonction. Elle régit tous verbes le subjonctif. *Quantvis, est*. On doit prendre garde de ne la mettre jamais après un que: Je vous assure que quoique; à cause de la cacophonie. On peut y substituer bien que, ou encore car. Vaug. Mén. On dispute s'il faut dire, *quoi qu'il arrive, ou quoi qui arrive*. C'est dans ce sens, le *quidquid* des Latins. Vaugelas est pour le premier, & Cornélius pour le dernier.

QUOLIBET. f. m. Façon de parler commune & triviale, qui renferme ordinairement une misérable plainte, & dont les gens du peuple, & les mauvais plaisants affectent de se servir.

La tynnoire, ou tyndaire, & la tynchoudroite. Voyez ces trois noms en l'article de la tynchoudrite.

Ce mot est Grec **σύνθεσις**, *arēsis*, vel *connexio naturalis*; il signifie l'union de deux choses jointes ensemble.

LYMPHITUM. Voyez Cossoues.

Le mot vient du Grec **σύν**, *syn*, & **φύειν**, *phéi*, *sopé adasci*. Naitre auprès, & par cette raison se dit d'un ulcère aride, & qu'il eût comme renaître les chairs.

SYMPOSIACQUE. f. m. Entretien de Philosophes dans un banquet. Plutarque a fait plusieurs traités qu'il a intitulés *De symposiacis*, ou des convives, &c.

Ce mot vient du Grec **συμπόσιον**.

SYMPTOMATIQUE. Qui appartient au symptôme. *Symptomatiaké*. Il faut qu'un Médecin étudie tous les mouvements, & tous les signes *symptomatiq.ues*. Evacuation *symptomatiq.ue*. Dia.

SYMPTÔME. f. m. Terme de Médecine. Signe, accident, ou révolution qui arrive dans une maladie, par où l'on peut juger de la nature, & de la qualité & de son évènement. *Symptoma, casus morbi*. Le délire est un *symp-tōme* de la fièvre. Les *symptômes* qui arrivent dans les maladies sont la douleur, les veilles, l'apoplexie, la convulsion, l'éblouissement, la furdité, la paralysie, la difficulté de respirer & d'avaler, la toux, le degout, la nausée, le fanglot, & le vomissement, la soif, la lithomye & l'yncope, le flux ou durété de ventre, le flux de sang, suppression d'urine, la jaunisse, la sécheresse & noirceur de la langue, &c. Le *symp-tōme* comprend tous les accidents qui surviennent aux maladies. Dia.

Ce mot vient du Grec **συμπτῶμα**, signifie la même chose.

SYMPOLE. f. m. Vase sacré. Petit vase dont les Pontifes Romains se servaient dans les sacrifices, pour faire des libations. *Sympholum*.

Il se certains devotions en reconnoissance de ce qu'on donne. Prière que toutes les dévotionnaires faites aux Eglises font *synallogmatiques*, parce que c'est aux conditions de dire des messes, de faire des *panaches*.

Ce mot est Grec **συναλλογμάτιστος**, *synallōgmatistos*, entre-échangeant.

SYNACHIE. f. f. Espèce d'effluviaire dans laquelle les muscles internes du bras ont attaque.

Ce mot est Grec, formé du Grec **συνάχαι**, *synachai*, & **ἀλγος**, *algos*, effier, fluyver.

SYNARTHROSE. f. f. Terme d'Anatomie, qui se dit de la jointure des os-qu'il est tellement compacte & serrée, qu'ils forment rendus immobiles. *Synarthrosis*. On la divide en trois espèces. La première est la *sauro*, qui est tantôt en forme de deux peignes, ou de deux faces qui sont tellement jointes ensemble, que les dents de l'un entrent en celles de l'autre, tantôt en forme d'angle, lorsqu'une partie monte par la volute, comme des écailles ou des tuiles. La seconde espèce s'appelle *harmonia*, qui se fait par une simple ligne droite ou circulaire, comme celle des os de la mâchoire supérieure. La troisième s'appelle *gomphose* qui se fait, quand un os est emboîté dans un autre comme une cheville dans un trou, ainsi que sont les dents dans les mâchoires.

Ce mot est composé de deux mots Grecs **συνάχαι**, *synachai*, & **άρθρον**, *arthron*, article, jointure.

SYNON. f. m. Livre Ecclésiastique des Grecs. C'est un recueil abrégé de la vie de leurs Saints. *Synaxarion*. Avant que les *Ménies* des Grecs & leurs autres livres qui contiennent les vies des Saints, fussent imprimés, les *Synaxaires* étoient un volume particulier. Mais comme on ne faisoit que des abréges qu'ils avoient, on les joignoit présentement avec les *Ménies* qu'on appelle *Synaxaire* du verbe Grec **συναξάω**, *synaxáo*, rassembler, parce que ce sont les recueils de tous les saints de l'Eglise. Origène.

- franç**ois** → franç**ais**
- éto**it** / avo**it** → éta**it** / avai**t**
- conno**ître** → conno**aître**

Graphies d'époque à conserver, pas à moderniser

Données et Corpus

Quelles données d'apprentissage et quel corpus de référence constituer ?

► **Corpus source:**

- Pages numérisées du *Dictionnaire de Trévoux* (1743), ~956 pages, deux colonnes.

► **Vérité de terrain:**

- Pages transcrites manuellement (principes diplomatiques)
- Fichiers **.txt** : correspondance image ↔ texte brut
- Fichiers **.xml** (TEI) : texte + structure de l'article

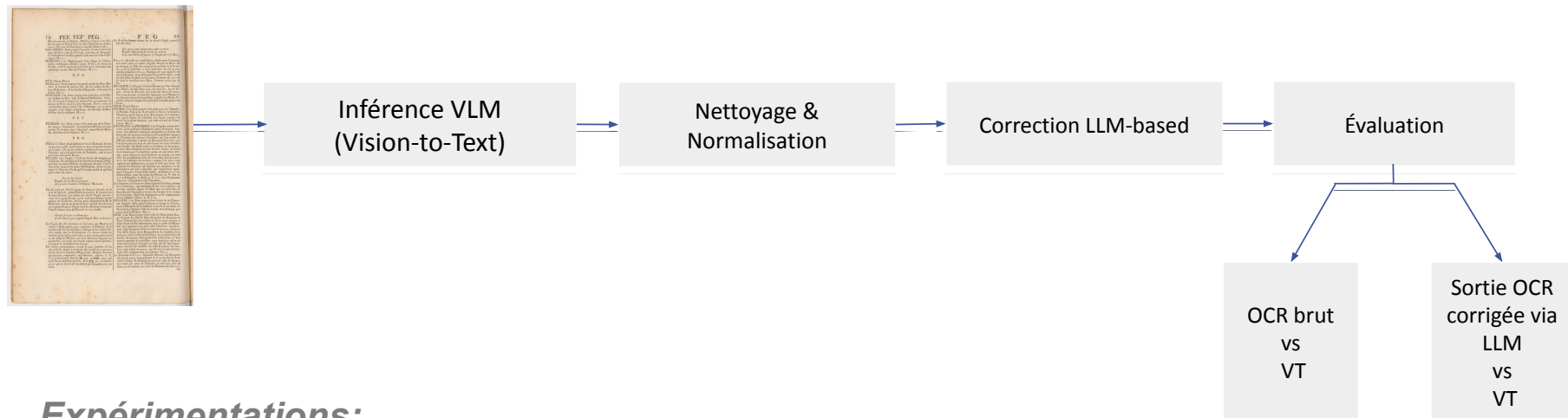
► **Annotation de la mise en page:**

- Schéma à **12 classes** : Mot-vedette, Vedette-secondaire, Paragraphe, Manicule, Lettrine, Tableau, En-tête, Numéro de page...

Approches explorées

- ▶ **Partie 1** : Approche globale end-to-end via VLMs
- ▶ **Partie 2** : Approche modulaire via pipeline spécialisé

Approche VLM end-to-end



Expérimentations:

VLMS testés :

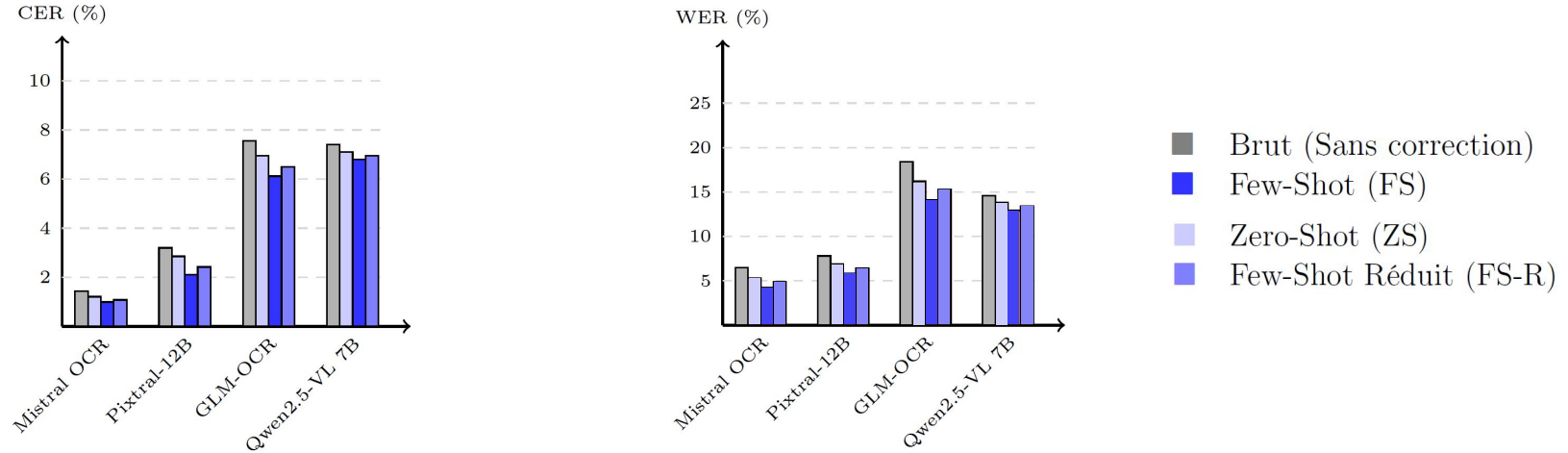
- Mistral OCR
- Pixtral-12B
- Qwen2.5-VL 7B
- GLM-OCR (Zhipu AI)

Correction LLM : Llama 3.3 70B (Groq)

3 Stratégies de Prompting Testées

- Zero-Shot (ZS)
- Few-Shot (FS)
- Few-Shot Réduit (FS-R)

Résultats



Variations du CER & WER selon le vlm et le prompt du LLM de correction

Modèle VLM	Character Error Rate (CER %)				Word Error Rate (WER %)			
	Brut	+ ZS	+ FS	+ FS-R	Brut	+ ZS	+ FS	+ FS-R
Mistral OCR	1,44	1,21	0,99	1,08	6,49	5,30	4,26	4,88
Pixtral-12B	3,20	2,85	2,10	2,42	7,80	6,90	5,85	6,40
GLM-OCR	7,56	6,95	6,12	6,50	18,39	16,20	14,10	15,30
Qwen2.5-VL 7B	7,40	7,10	6,80	6,95	14,60	13,80	12,90	13,40

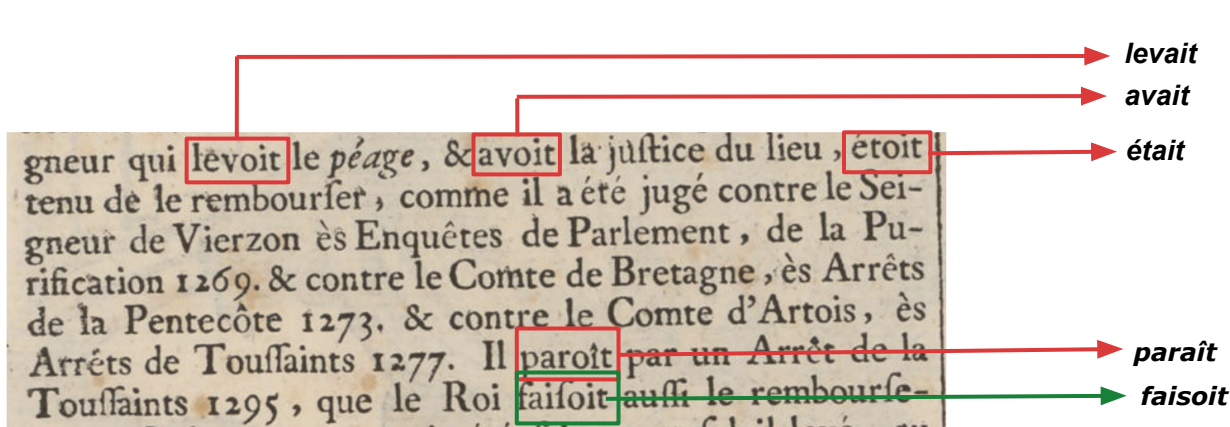
Pourquoi un pipeline dédié ?

Limites du VLM end-to-end

- Hallucination : génère du texte plausible mais faux
- Sources d'erreurs difficiles à diagnostiquer
- Coût des API cloud (956 pages)
- Vocabulaire modernisé à tort

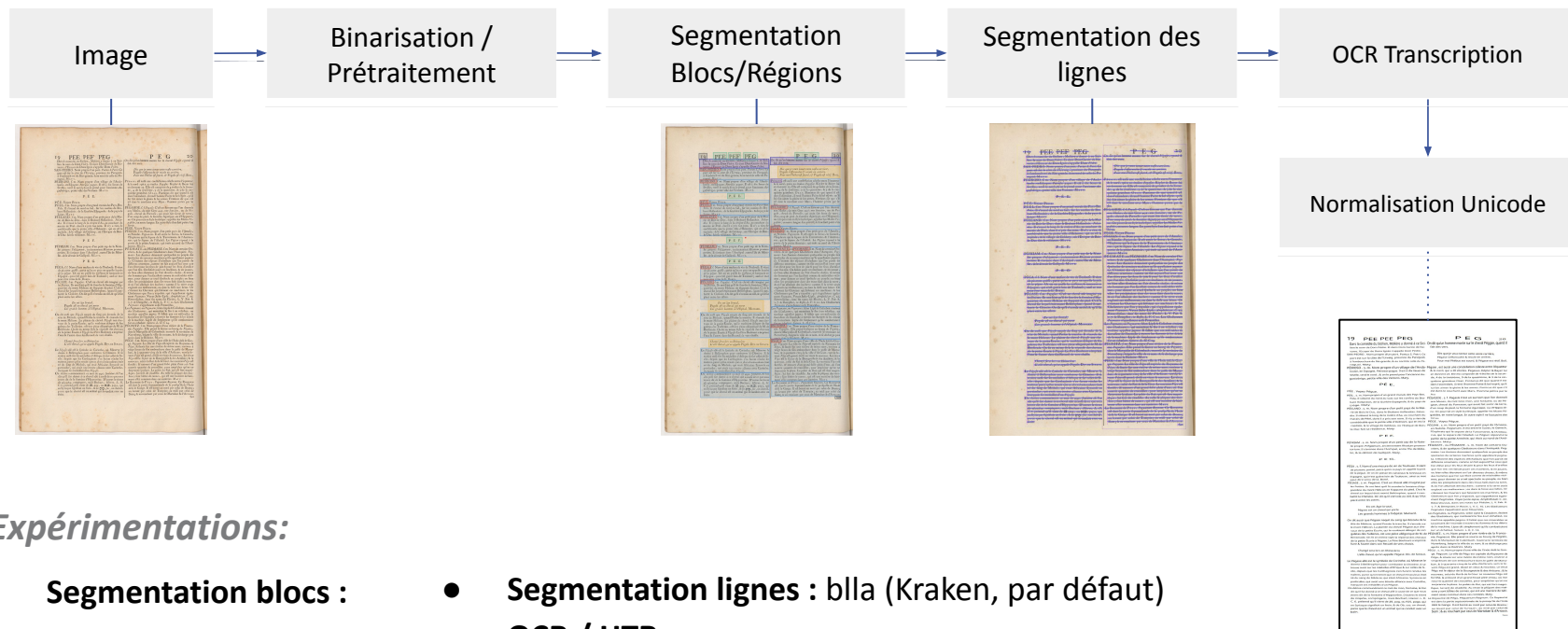
Réponse : approche modulaire

- HTR fidèle : transcrit sans inventer
- Contrôle de chaque étape
- Local, ouvert, gratuit
- Ré-entraînable : segmentation + reconnaissance



Par exemple: Problèmes de conjugaison ancienne modernisée

Approche modulaire via pipeline spécialisé



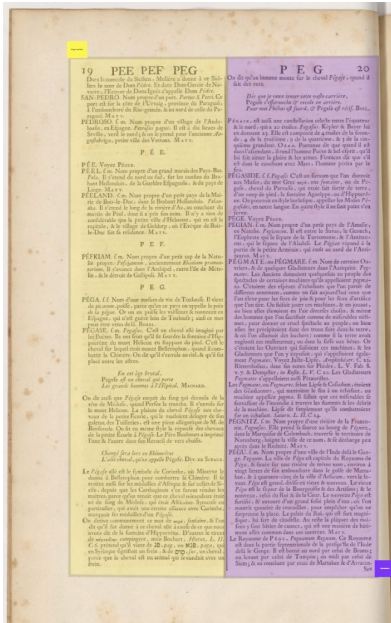
Expérimentations:

- **Segmentation blocs :**
 - DocLayout-YOLO
 - Qwen 2.5-vl 3B
- **Segmentation lignes :** blla (Kraken, par défaut)
- **OCR / HTR :**
 - EasyOCR
 - Tesseract 5
 - Surya
 - Kraken + CATMuS-Print Large

Segmentation Blocs / Régions

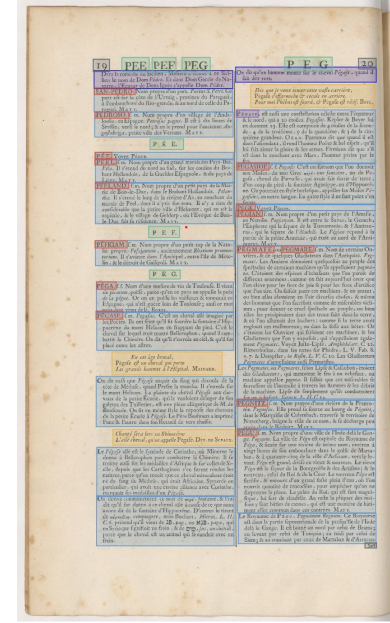
Segmentation à 2 niveaux :

① Colonnes (gauche / droite)



→ ordre de lecture

② Régions fines (12 classes)



→ structure de l'article

- Séparation gauche / droite
- Ordre de lecture rétabli
- Préalable à la segmentation des lignes



170 pages annotées •

12 classes typées •

Segmentation régions fines

2 approches proposées

Approche standard

Fine-tuning modèle DocLayout-YOLO

- 130 pages pour entraînement annotées sur label studio.
- 40 pages pour test annotées.

Expérimentations:

- DocLayout-YOLO
- Qwen2.5-VL

Mot-vedette

Vedette-secondaire

Paragraphe

Paragraphe-suite

Citation

Section-lexicale

Manicule

Lettrine

Illustration

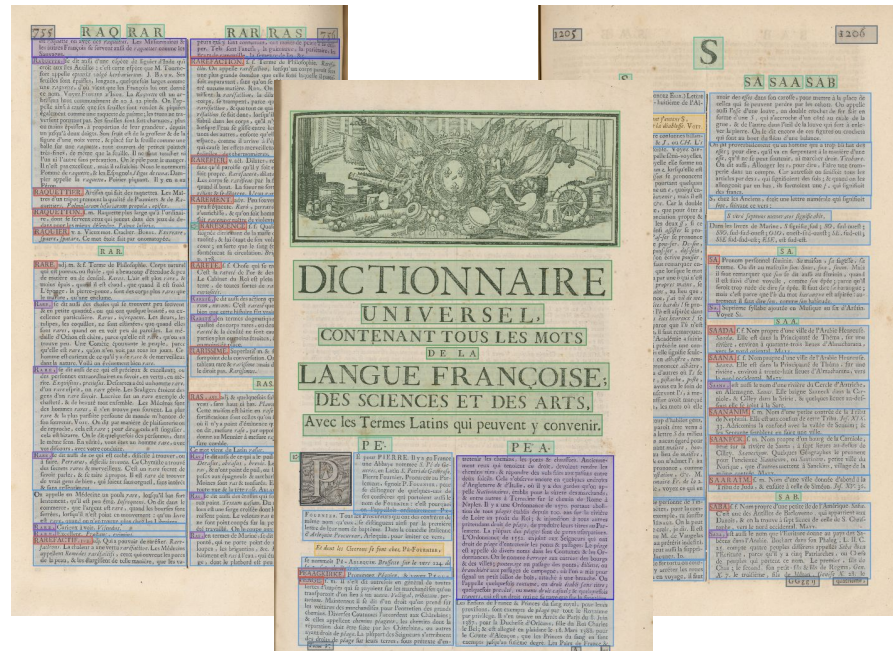
Titre

Autre-paratexte

Tableau

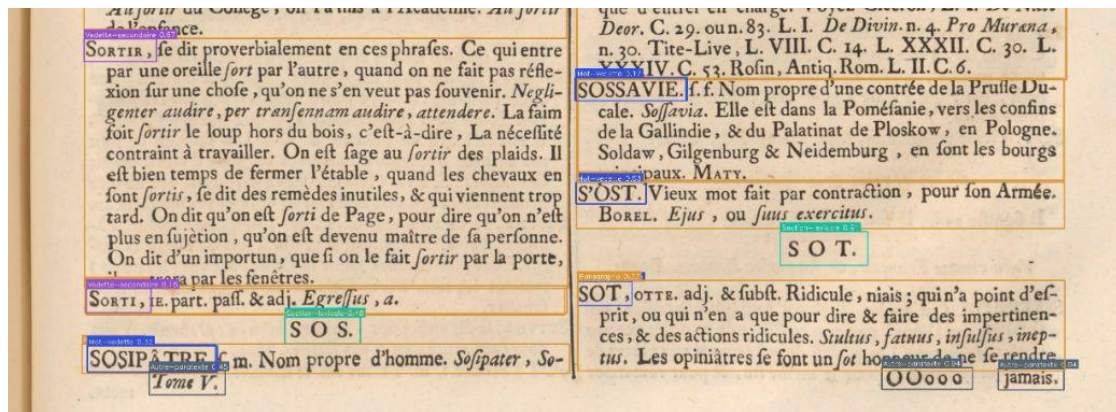
Approche générative

Fine-tuning modèle VLM



Résultats

Métrique	DocLayout-YOLO (fine-tuné)	Qwen2.5-VL (fine-tuné)
mAP50 (50 pages)	0,921	0,914
mAP50-95	0,68	0,66
F1-Score	0,90	0,89
Détection des styles (italique, grandes & petites capitales)	Non – <i>layout seul</i>	Possible

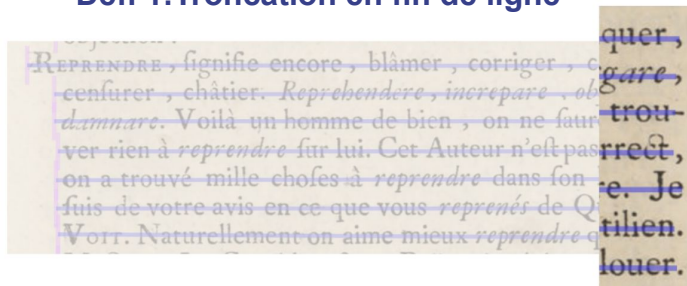


→ **Structures à YOLO** : Mots-vedettes, vedettes secondaires et citations sont délégués à DocLayout-YOLO pour garantir une géométrie maximale.

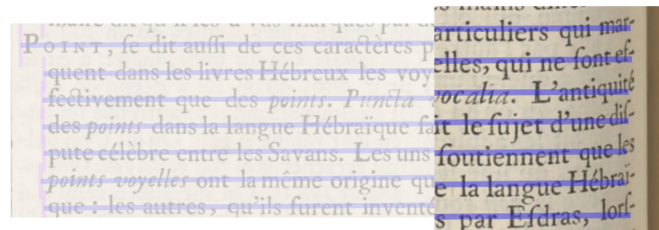
→ **Styles au VLM léger** : Le VLM est conservé *uniquement pour les styles (italique, capitales)*, permettant de basculer vers un modèle plus petit et spécialisé.

Segmentation des lignes

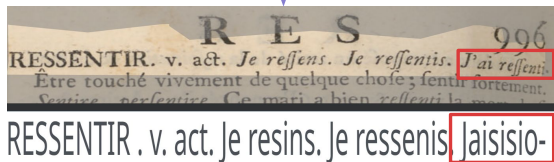
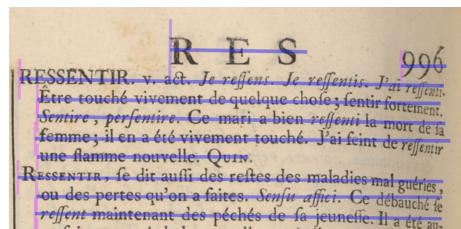
Défi 1: Troncation en fin de ligne



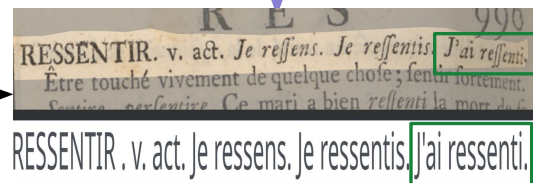
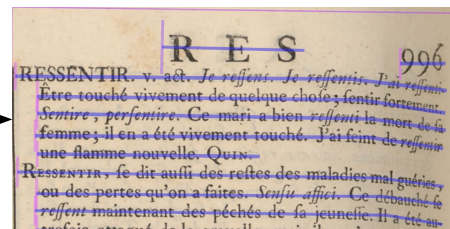
Défi 2: Courbure de reliure



Sans ajustement de segmentation de lignes



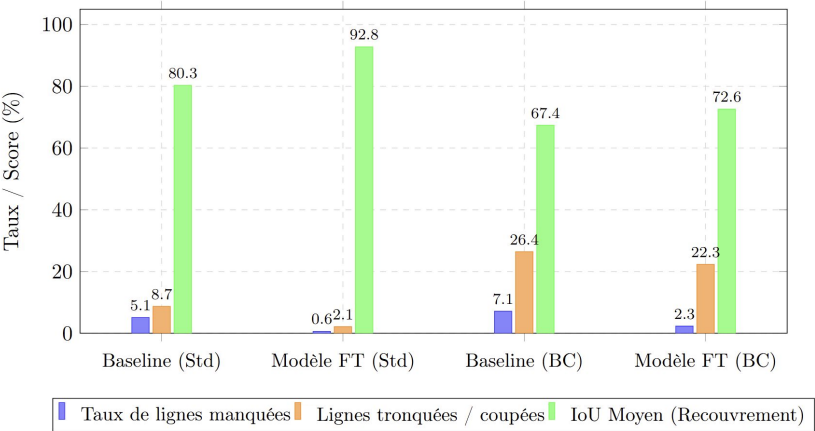
Avec ajustement de segmentation de lignes



Résultats

- Modèle de segmentation : **blla (Kraken)**
- Test indépendant : 30 pages au total (20 pages standards / 10 pages avec *Book Curve*).
- Fine-tuning : 50 pages annotées.

Métrique de segmentation	Pages Standards (20)		Pages avec <i>Book Curve</i> (10)	
	<i>blla Baseline</i>	<i>blla FT</i>	<i>blla Baseline</i>	<i>blla FT</i>
Taux de lignes manquées	5,1 %	0,6 %	7,1 %	2,3 % ¹
Lignes tronquées / coupées	8,7 %	2,1 %	26,4 %	22,3 % ²
IoU Moyen (Recouvrement)	80,3 %	92,8 %	67,4 %	72,6 %



1. **Phénomène de fusion locale de baselines (*under-segmentation*)** lorsque l'angle de la courbure resserre trop l'espacement inter-lignes.
2. **Perte des derniers caractères en fin de ligne**, le modèle *blla* peignant à suivre la trajectoire descendante ou ascendante du texte dans le pli.



Évaluation moteurs OCR / HTR traditionnels

Analyse comparative des performances initiales sur le corps textuel du Dictionnaire de Trévoux:

Moteur OCR / HTR	CER (%)	WER (%)	Observations comportementales
EasyOCR	12,30	28,40	Sensibilité critique au bruit de fond du papier.
Surya	9,15	21,20	Détection des lignes correcte mais hallucinations.
Tesseract 5	8,40	19,80	Échecs multi-colonnes et confusion $\int \rightarrow f$.
Kraken (CATMuS-Print Large)	4,50	11,30	Meilleure robustesse native sur la graphie historique.

Robustesse native sur graphie historique (s long, ligatures, conjugaison ancienne...)

gneur qui levoit le péage, & avoit la justice du lieu, étoit
tenu de le rembourser, comme il a été jugé contre le Sei-
gneur de Vierzon ès Enquêtes de Parlement, de la Pu-
rification 1269. & contre le Comte de Bretagne, ès Arrêts
de la Pentecôte 1273. & contre le Comte d'Artois, ès
Arrêts de Toussaints 1277. Il paroit par un Arrêt de la
Toussaints 1295, que le Roi faisoit aussi le rembourse-

gneur qui levoit le péage, & avoit la justice du lieu, étoit
tenu de le rembourser, comme il a été jugé contre le Sei-
gneur de Vierzon ès Enquêtes de Parlement, de la Pu-
rification 1269. & contre le Comte de Bretagne, ès Arrêts
de la Pentecôte 1273. & contre le Comte d'Artois, ès
Arrêts de Toussaints 1277. Il paroit par un Arrêt de la
Toussaints 1295, que le Roi faisoit aussi le rembourse-

Dès que je veux tenter cette vaste carrière,
Pégase s'effarouche & recule en arrière.
Pour moi Phébus est sourd, & Pégase est rétif. Boil.

Dès que je veux tenter cette vaste carrière,
Pégase s'effarouche & recule en arrière.
Pour moi Phébus est sourd, & Pégase est rétif. Boil.

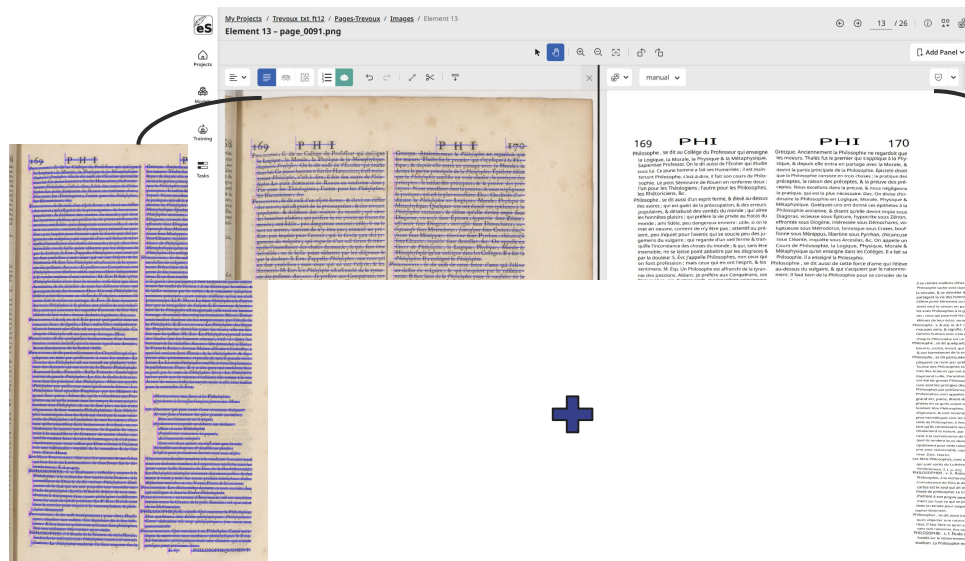
→ CER= 4,50 % encore trop élevé pour 956 pages → fine-tuning nécessaire

eScriptorium : l'environnement de fine-tuning Kraken

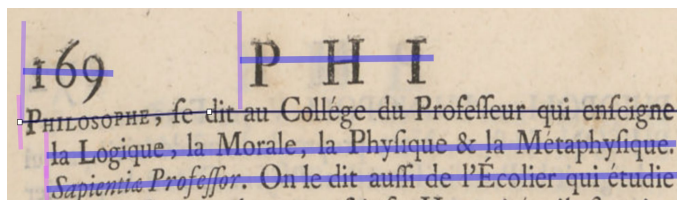


Segmentation, transcription, entraînement réunis:

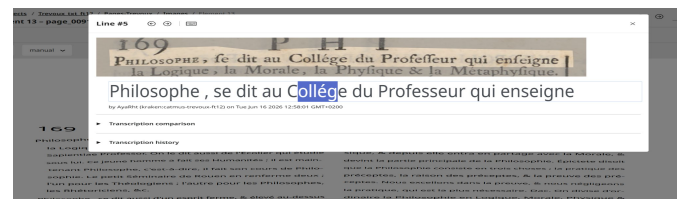
①
Segmentation
des lignes



②
Transcription



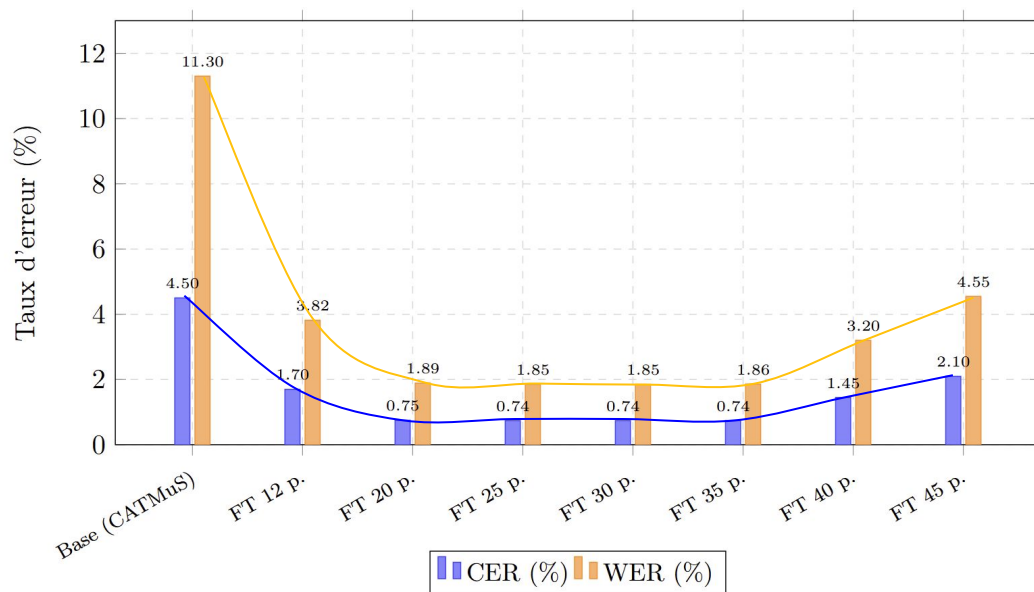
Ajustement des lignes segmentées par blla (modèle de base)



Correction manuelle de la transcription du modèle de base

Fine tuning du modèle Kraken+ CATMuS-Print

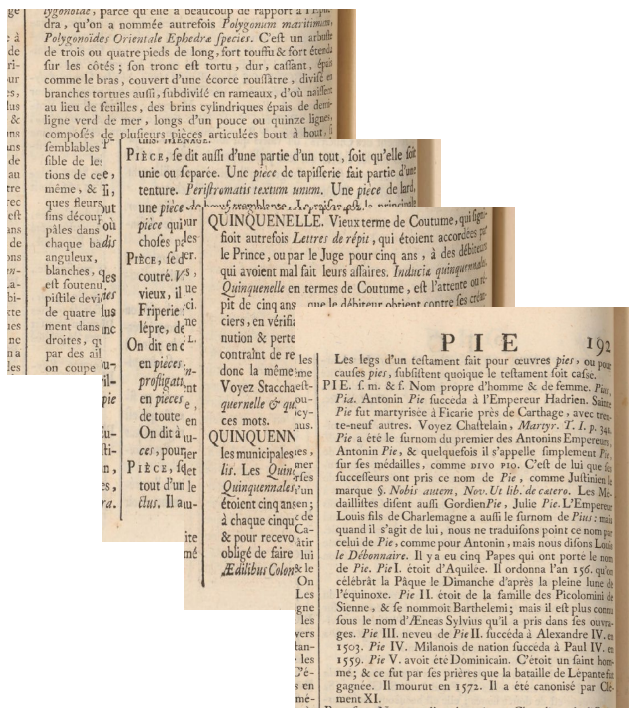
- CATMuS-Print Large (modèle de base)
- Fine-tuning sur les n pages (de 12 à 45 pages) aux CER/WER les plus élevés, Test sur 140 pages.



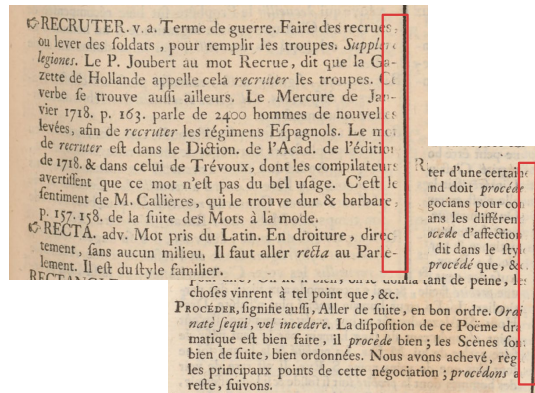
- FT 12 à 25 p. : Optimisation rapide
 - Alignement efficace sur la typographie du *Dictionnaire de Trévoux*.
- FT 25 à 35 p. : Zone de convergence
 - Plafond de performance atteint.
 - Minimum absolu : **CER à 0,738 % / WER à 1,85 %**.
- FT 40 à 45 p. : Overfitting
 - Mémorisation par cœur des bruits physiques (accents, ponctuations,...).

Sources d'erreurs restantes:

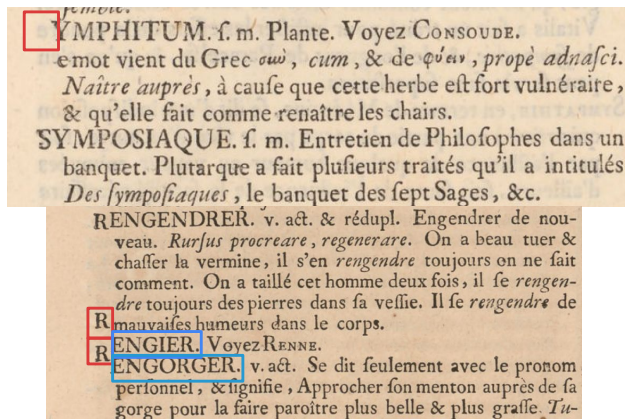
- Courbures de reliure trop sévères dans quelques pages



- Encre peu lisible en fin de ligne (quelques pages)



- Première lettre omise ou décalée (quelques cas)



Conclusion et perspectives

- Détection des styles : grandes capitales, petites capitales, italiques
- Intégration du classifieur typographique dans le flux principal
- Extension aux scripts non latins (grec polytonique, hébreu) : absents de la vérité terrain → annotation dédiée puis reconnaissance

Bibliographie

- Kiessling, B., Tissot, R., Stokes, P., Stökl Ben Ezra, D. (2019). *eScriptorium: An Open Source Platform for Historical Document Analysis*. ICDAR Workshops (ICDARW) 2019, 19–24.
- Kiessling, B. (2019). *Kraken – A Universal Text Recognizer for the Humanities*. Digital Humanities Conference (DH2019), Utrecht. DOI : 10.34894/Z9G2EX.
- Clérice, T., Pinche, A., Vlachou-Efstathiou, M., Chagué, A. et al. (2024). *CATMuS Medieval: A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond*. ICDAR 2024, LNCS 14806, Springer. DOI : 10.1007/978-3-031-70543-4_11.
- Smith, R. (2007). *An Overview of the Tesseract OCR Engine*. ICDAR 2007.
- Zhao, Z., Kang, H., Wang, B., He, C. (2024). *DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception*. arXiv:2410.12628.
- He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). *Mask R-CNN*. ICCV 2017.
- Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. (2020). *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. KDD 2020.
- Shafait, F., Keysers, D., Breuel, T. M. (2008). *Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms*. IEEE TPAMI, 30(6).
- Clausner, C., Antonacopoulos, A., Pletschacher, S. (2019). *Efficient and Effective OCR Engine Training*. IJDAR, 23(1), 73–88. DOI : 10.1007/s10032-019-00347-8.
- Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S., Kaplan, F. (2024). *Post-correction of Historical Text Transcripts with Large Language Models: An Exploratory Study*. LaTeCH-CLfL 2024, 133–159.