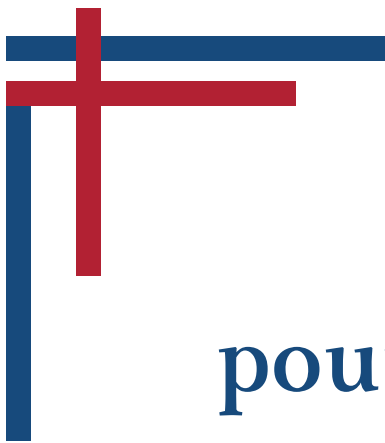


Projet de Fin d'Etudes
Double Diplôme

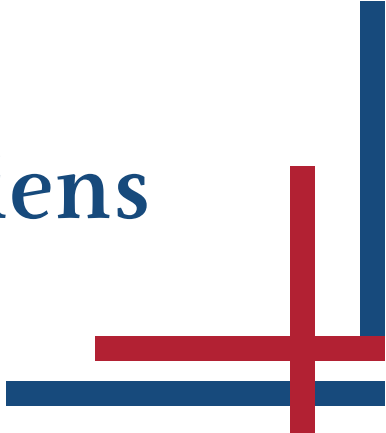
Master : Données et Intelligence Artificielle en Génie Industriel (DIAGI)

Métier : Supplychain Performance des Entreprises



IA et Deep Learning : OCR et LLM

pour la reconnaissance et la structuration de dictionnaires anciens



Devant le jury :

- M. Lyes NECHAK

Présenté par : **Aya RAHOUTI**

2025 - 2026

Tuteurs laboratoire :

- Mme. Véronique EGLIN
- M. Ludovic MONCLA

Tuteur pédagogique :

- M. Mohsen ARDABILIAN

Plan

- 1 Contexte et problématique
- 2 État de l'art
- 3 Expérimentations et résultats
- 4 Conclusion et perspectives

• Contexte - Présentation du LIRIS

LIRIS → une unité mixte de recherche

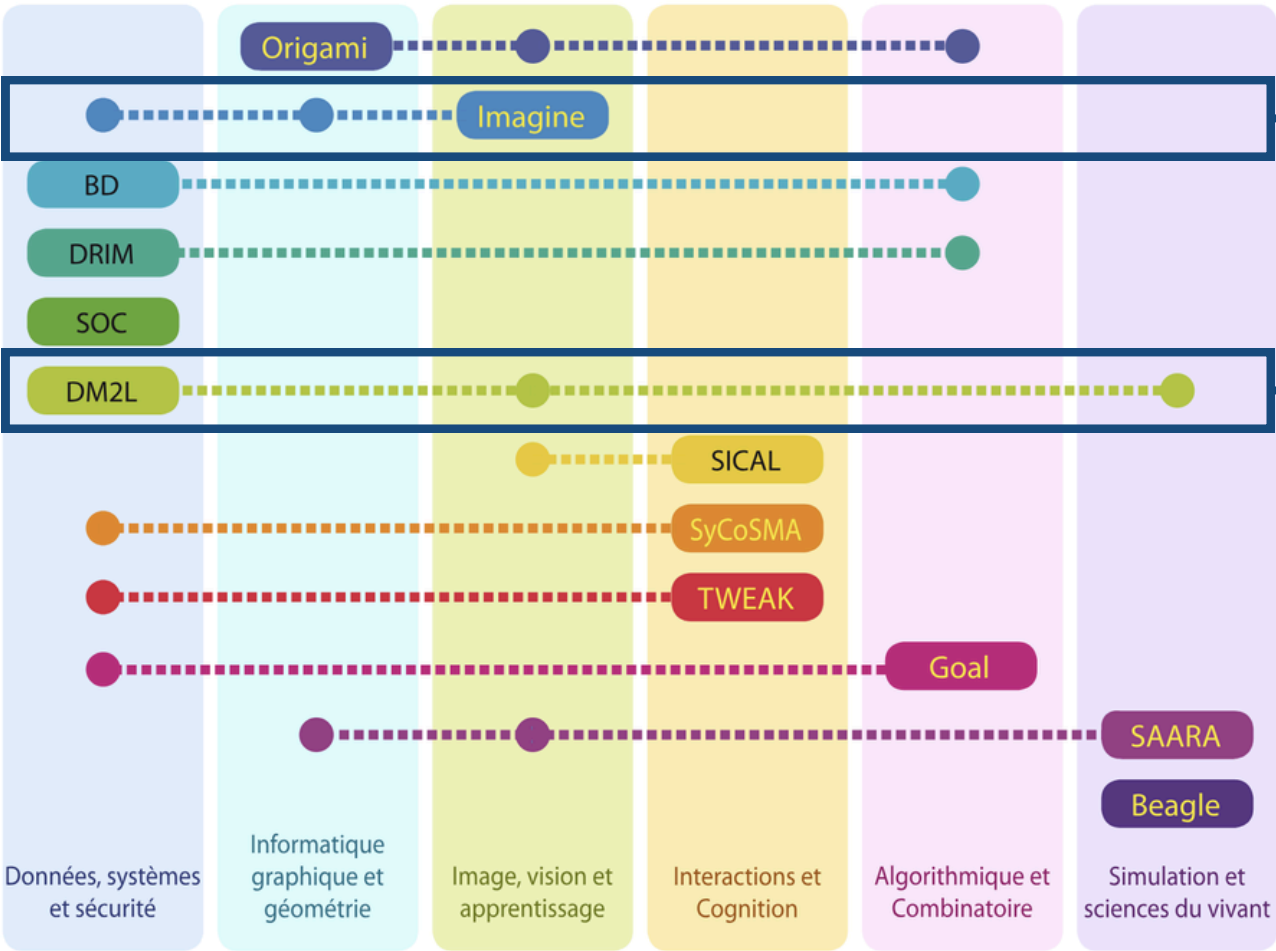


- Membres du LIRIS : 405
 - Chercheurs et enseignants-chercheurs permanents : 147
 - Doctorants : 141
 - Post-doctorants : 21



• Contexte – Le projet OLR-LLM

- 12 équipes de recherche, réparties en 6 pôles thématiques



Projet OLR - LLM

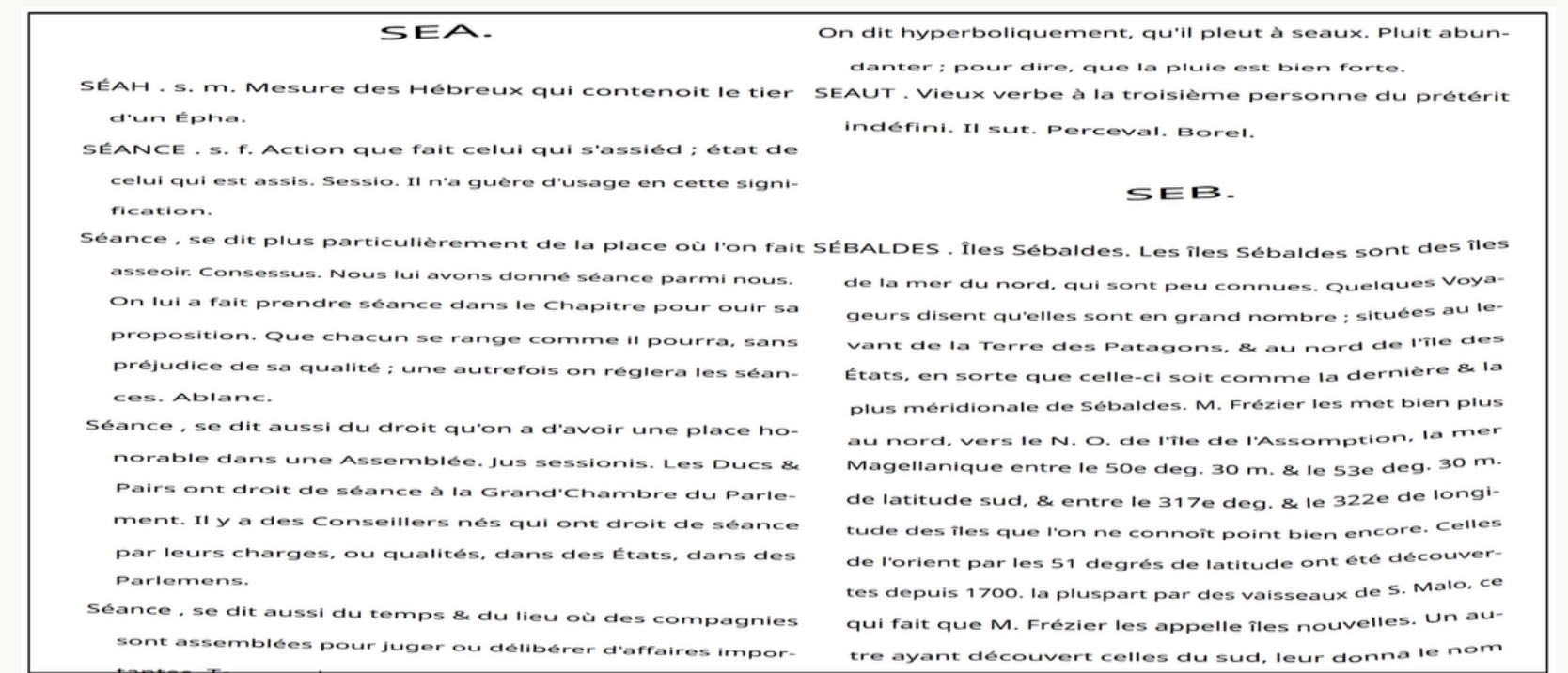
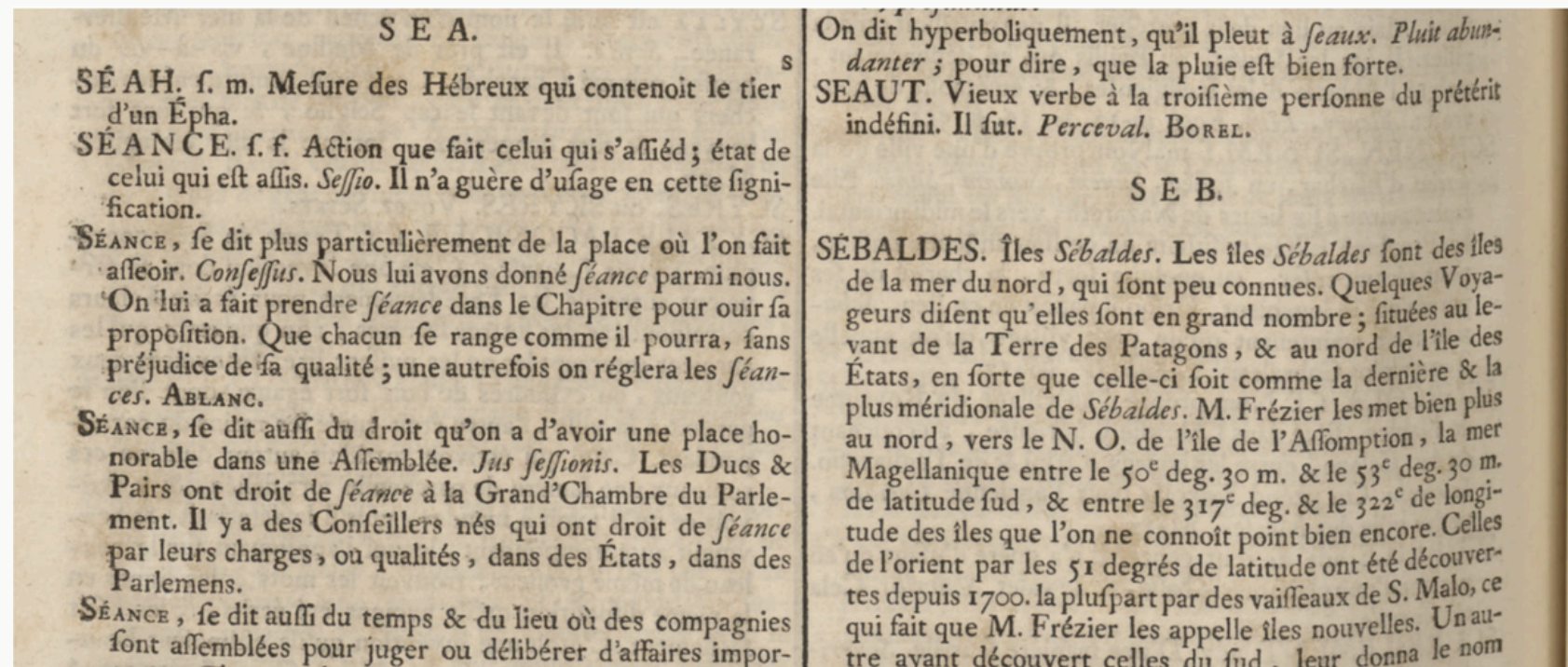
- Financé par l'IXXI et la MSH Lyon-Saint-Étienne, mené en lien avec le laboratoire ICAR
- S’inscrit dans le prolongement du projet **GEODE**, consacré au discours géographique encyclopédique
-

Objectif : mobiliser l’OCR et les grands modèles de langage pour exploiter les dictionnaires anciens

• Contexte – Le corpus : le Dictionnaire de Trévoux

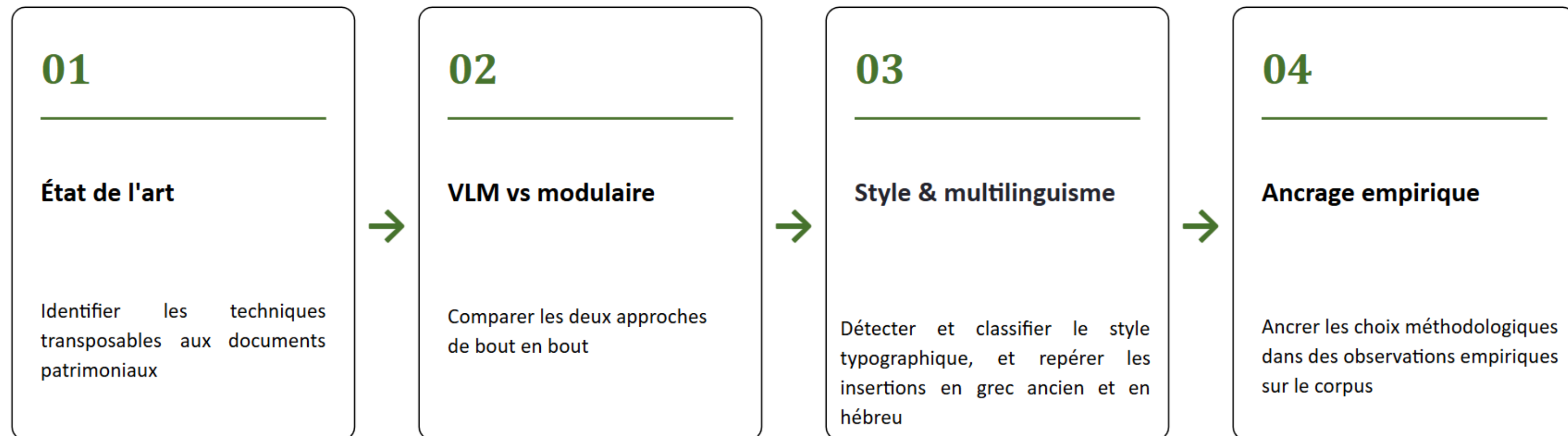
- Environ 956 pages numérisées, mise en page à deux colonnes
- Entrées composites : vedette, sous-entrées, citations ...
- Typographie du XVIIIe siècle : « s long », ligatures, variantes orthographiques
- Insertions multilingues : français, latin, grec ancien, hébreu

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



- **Contexte – Problématique et objectifs**

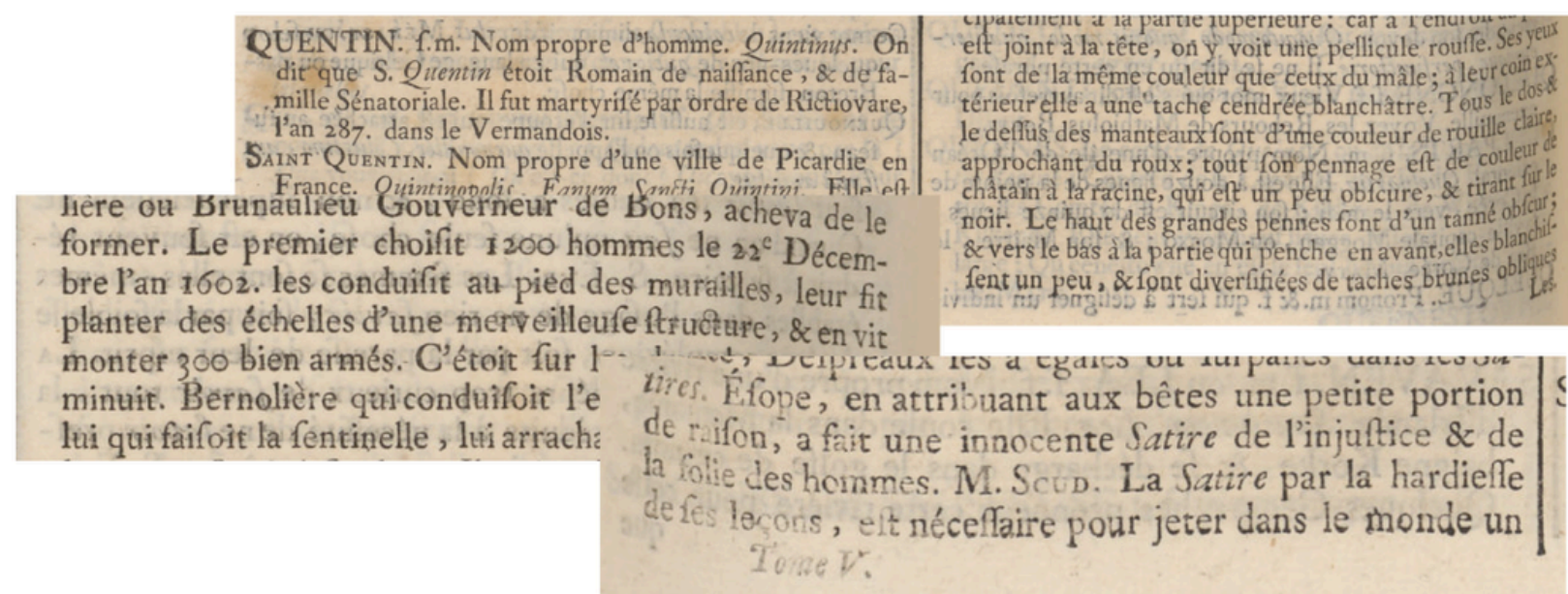
Comment concevoir une chaîne de traitement automatique capable de reconnaître, transcrire et structurer fidèlement les entrées lexicographiques d'un dictionnaire ancien imprimé, à partir d'images de pages numérisées, sans dénaturer les graphies et la structure d'origine ?



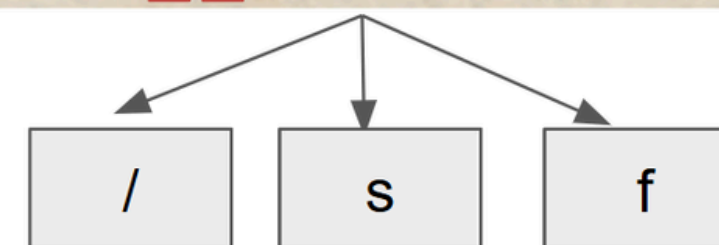
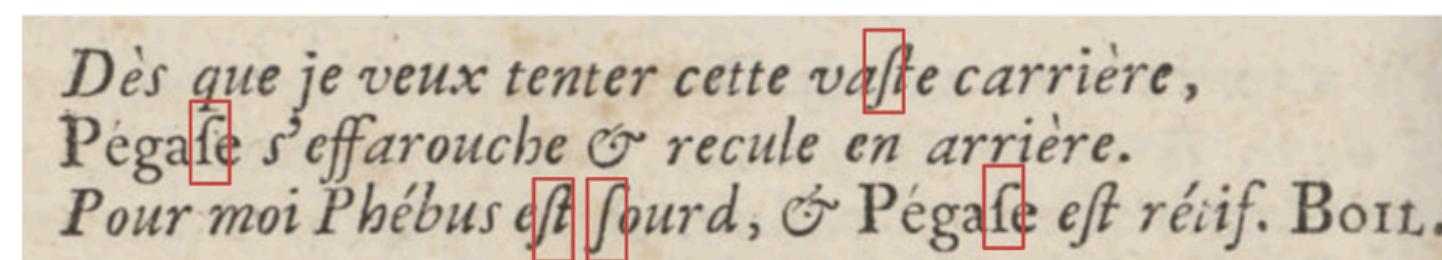
• Défis du corpus :

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

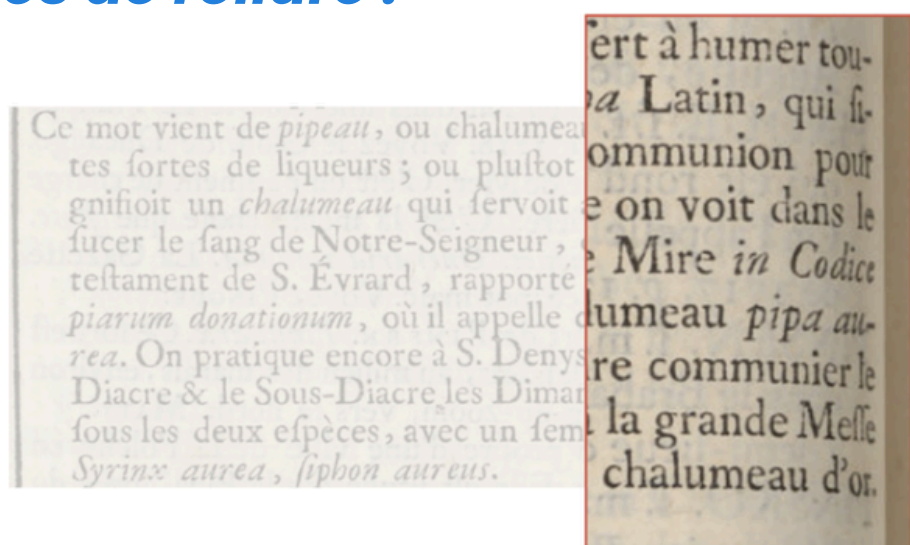
► Qualité du support :



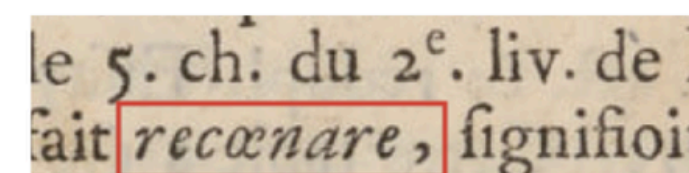
► s long : f f



► Courbures de reliure :



► Ligatures :



recœnare

recœnare

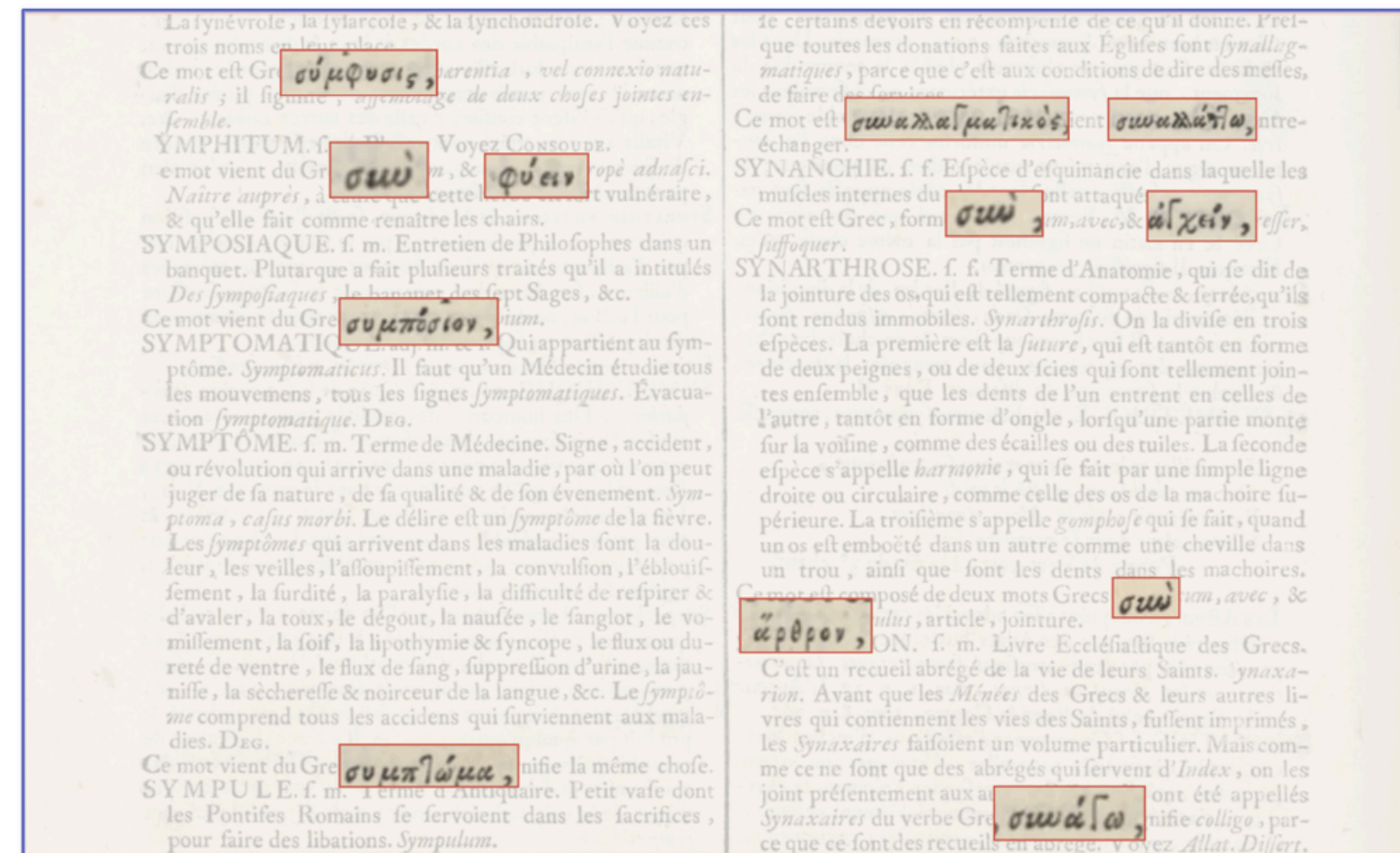
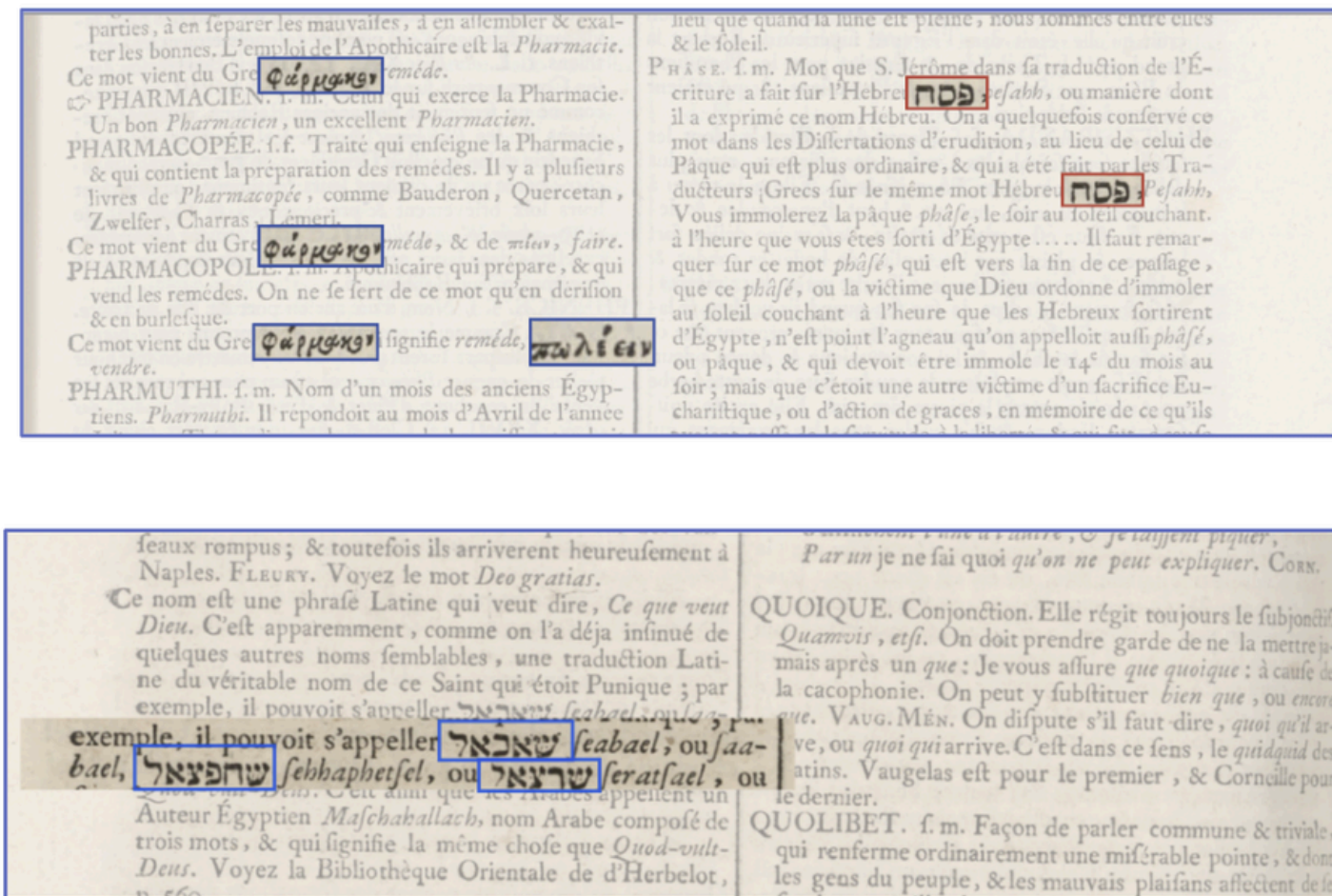
æ = a+e (ex. *præfatio*, *cælum*) æ moderne

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

- **Défis du corpus :**

► *Multilinguisme :*

Français + latin + grec + hébreu parfois



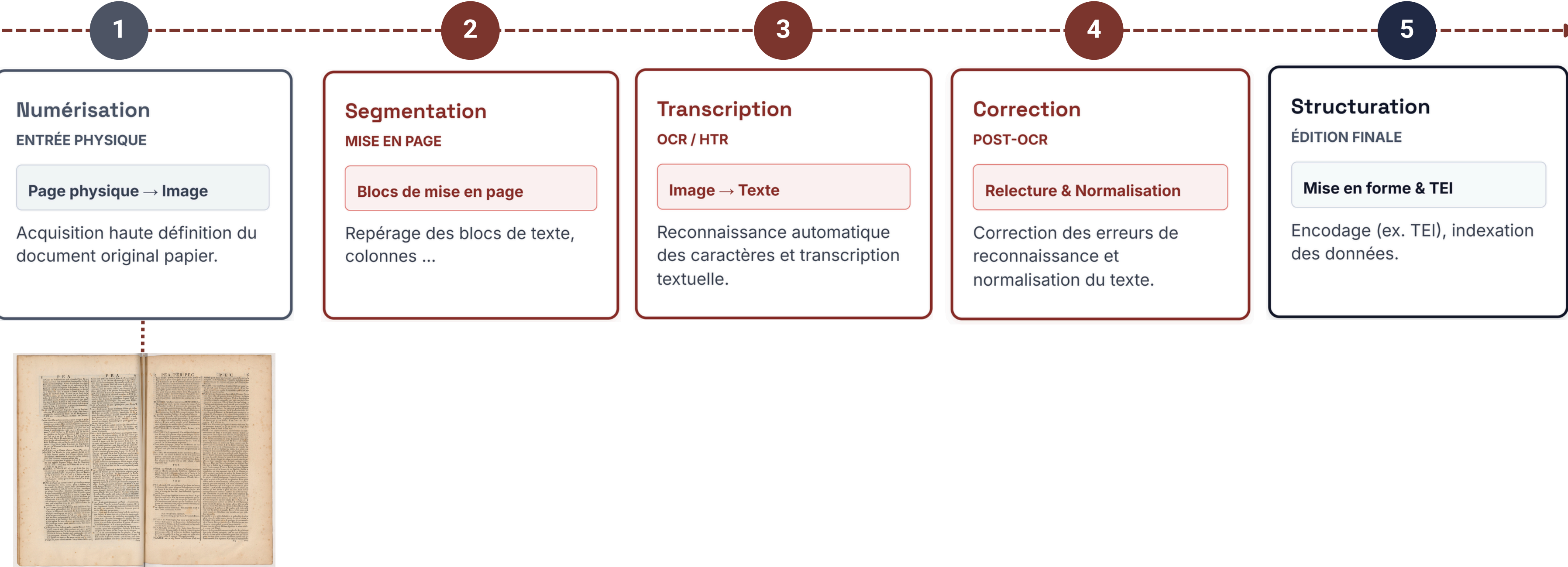
► **Français ancien :**

- franç**ois**
→ franç**ais**
- éto**it** / avo**it**
→ ét**ait** / av**ait**
- conn**oi**tre
→ conn**a**ître

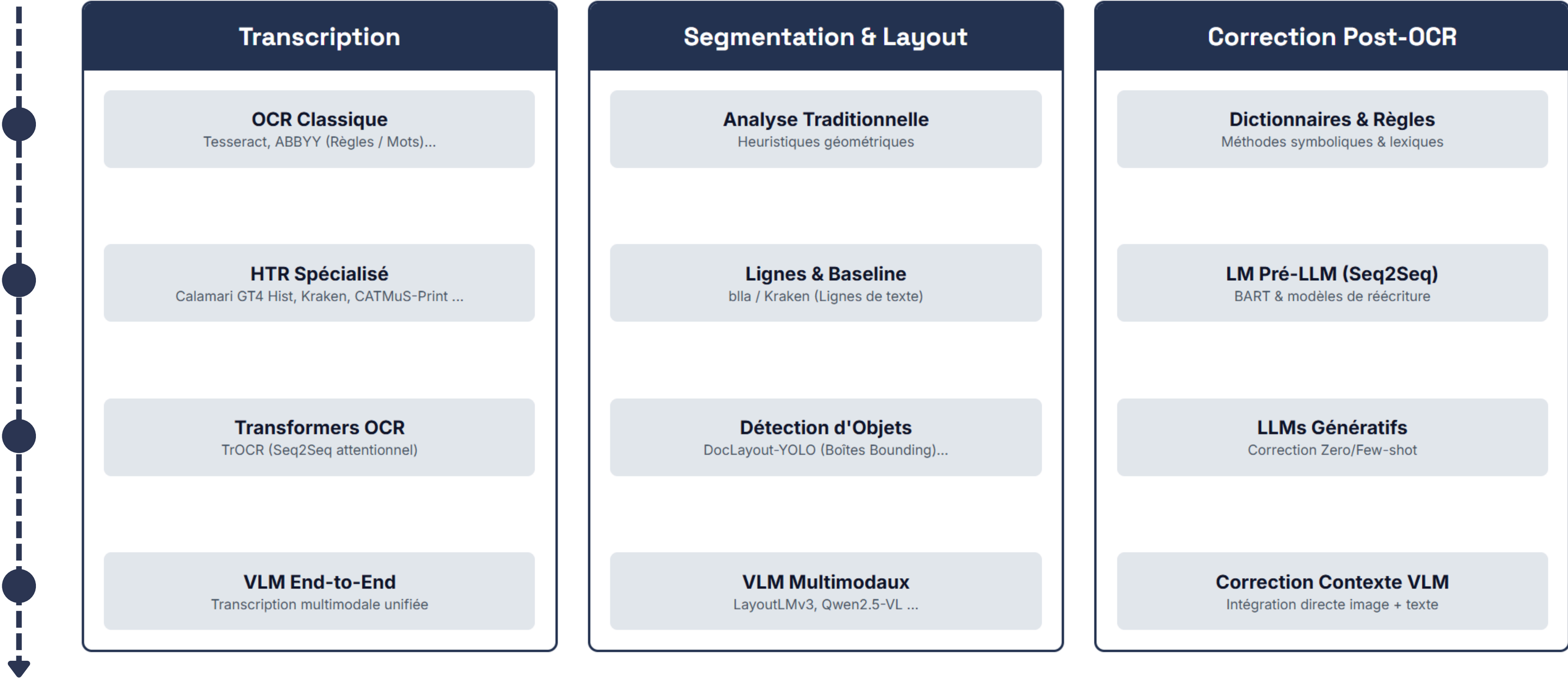
Graphies d'époque à conserver, pas à moderniser

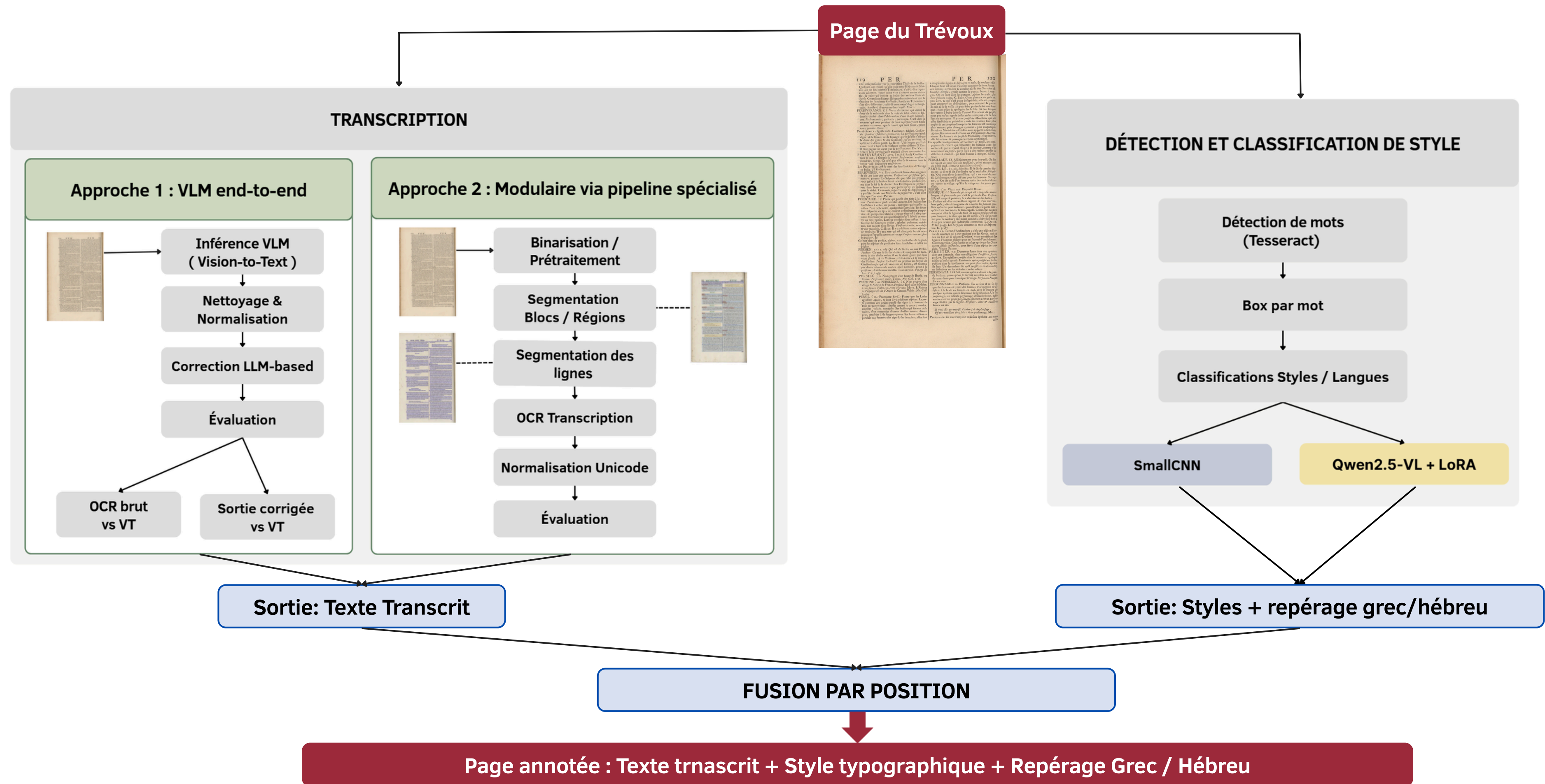
• Contexte – Problématique et objectifs

Comment un document imprimé ancien était-il traité ?



Comment les méthodes ont-elles évolué, tâche par tâche ?





• Protocole expérimental

► 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

Tâche	Expérience	Taille
Transcription	Comparaison des VLM	150 pages
Transcription	Fine-tuning CATMuS-Print	12–45p train, 140p test
Segmentation	DocLayout-YOLO / Qwen2.5-VL	130p train, 40p test
Segmentation	Analyse qualitative (blla)	50p train, 30p test
Classification de style	SmallCNN, Qwen2.5-VL LoRA	100 pages

Transcription

Approches explorées



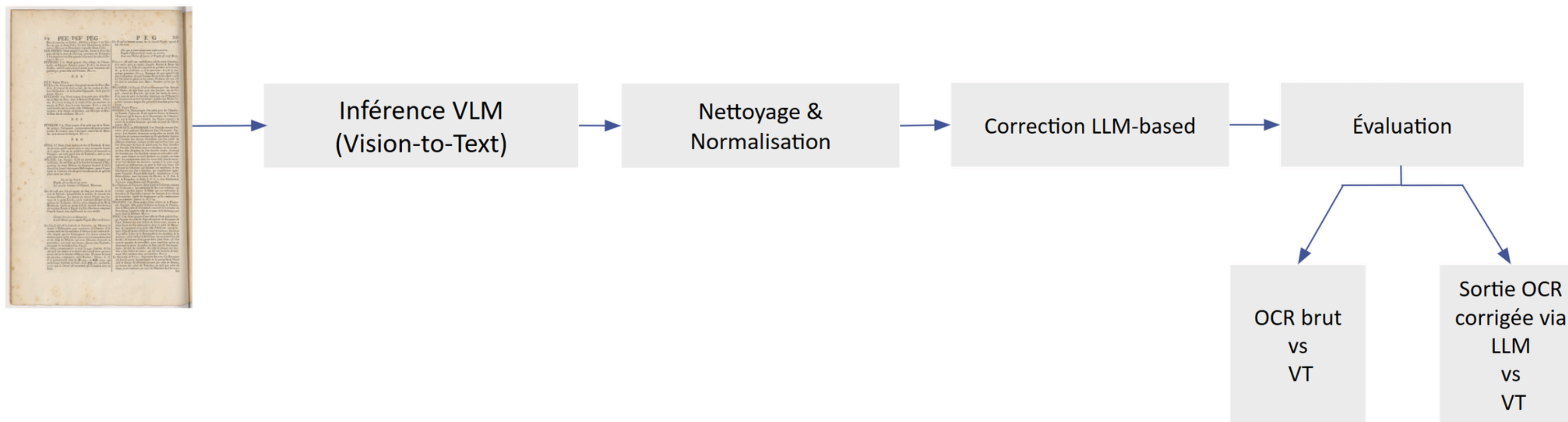
Approche 1 :

Approche globale end-to-end via VLMs

Approche 2 :

Approche modulaire via pipeline spécialisé

• Approche 1 : Approche VLM end-to-end



Expérimentations:

VLMs testés :

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

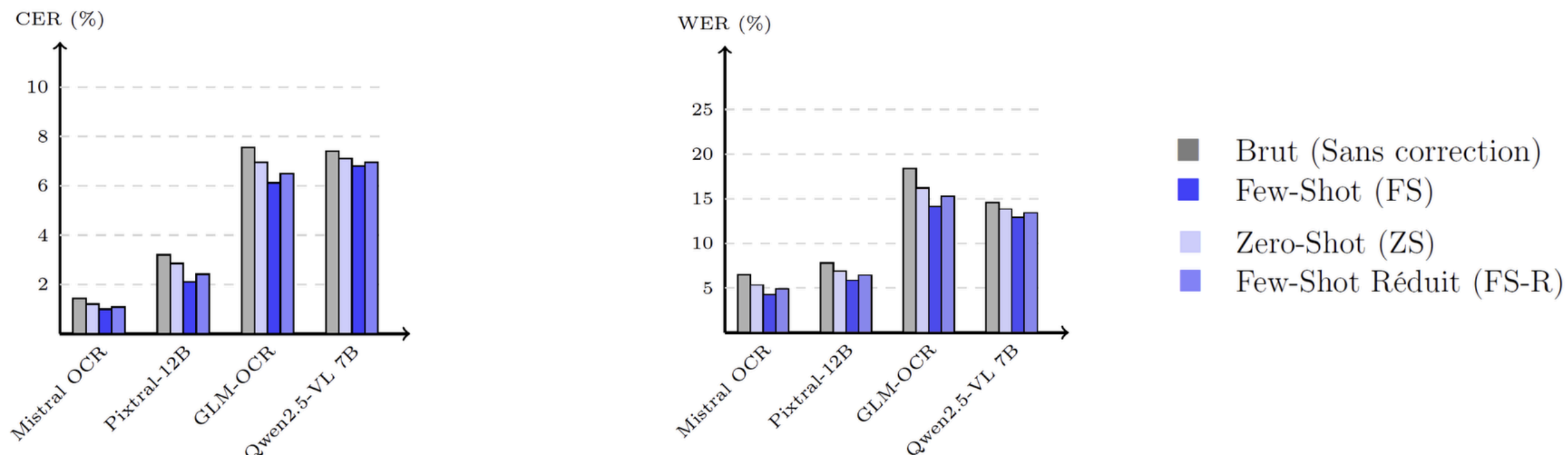
Correction LLM : Llama 3.3 70B (Groq)

3 Stratégies de Prompting Testées:

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

- Approche 1 : Approche VLM end-to-end

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

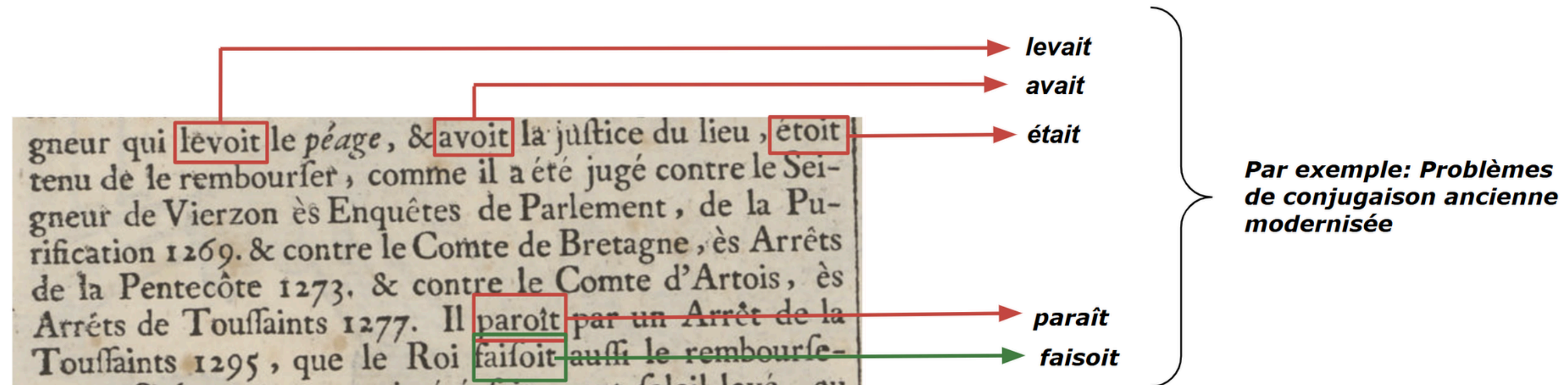
• Approche 1 : Approche VLM end-to-end

Limites du VLM end-to-end

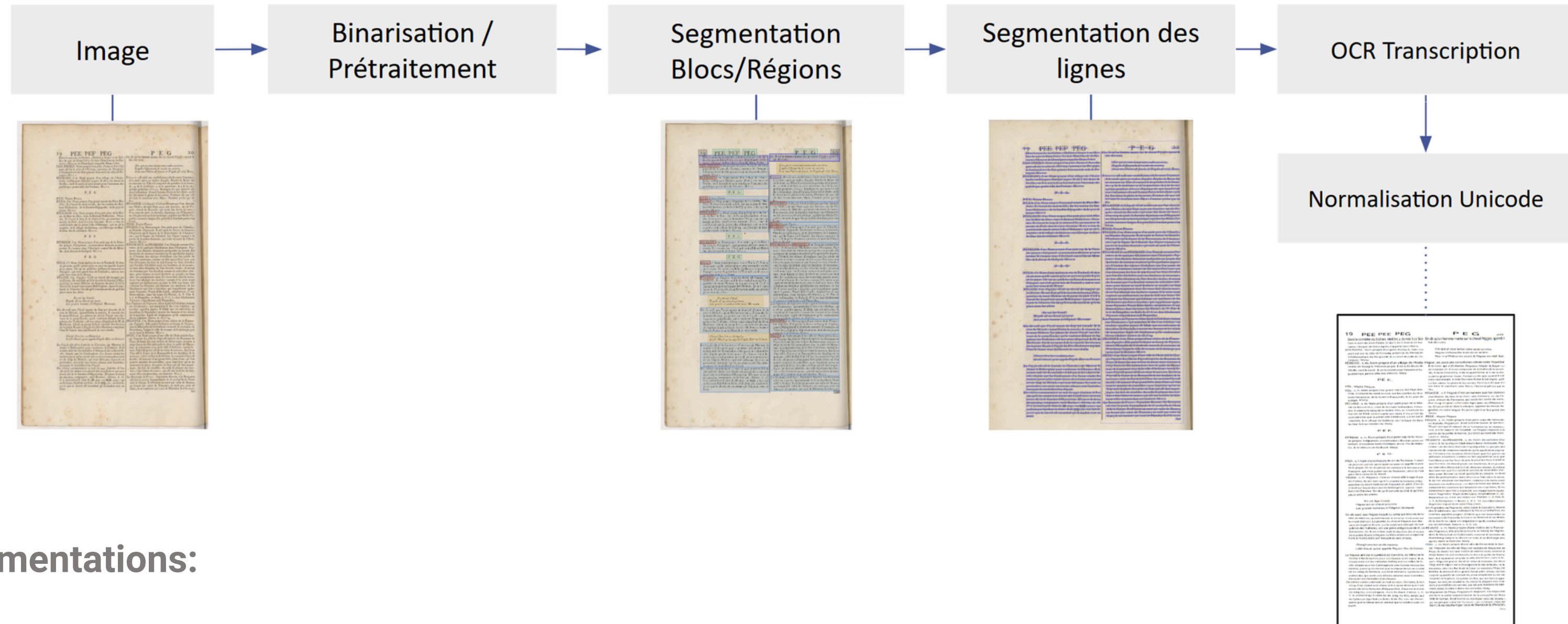
- Hallucination : génère du texte plausible mais faux
- Sources d'erreurs difficiles à diagnostiquer
- Coût des API cloud (956 pages)
- Vocabulaire modernisé à tort

Réponse : approche modulaire

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



• Approche 2 : Approche modulaire via pipeline spécialisé



Expérimentations:

Segmentation blocs :

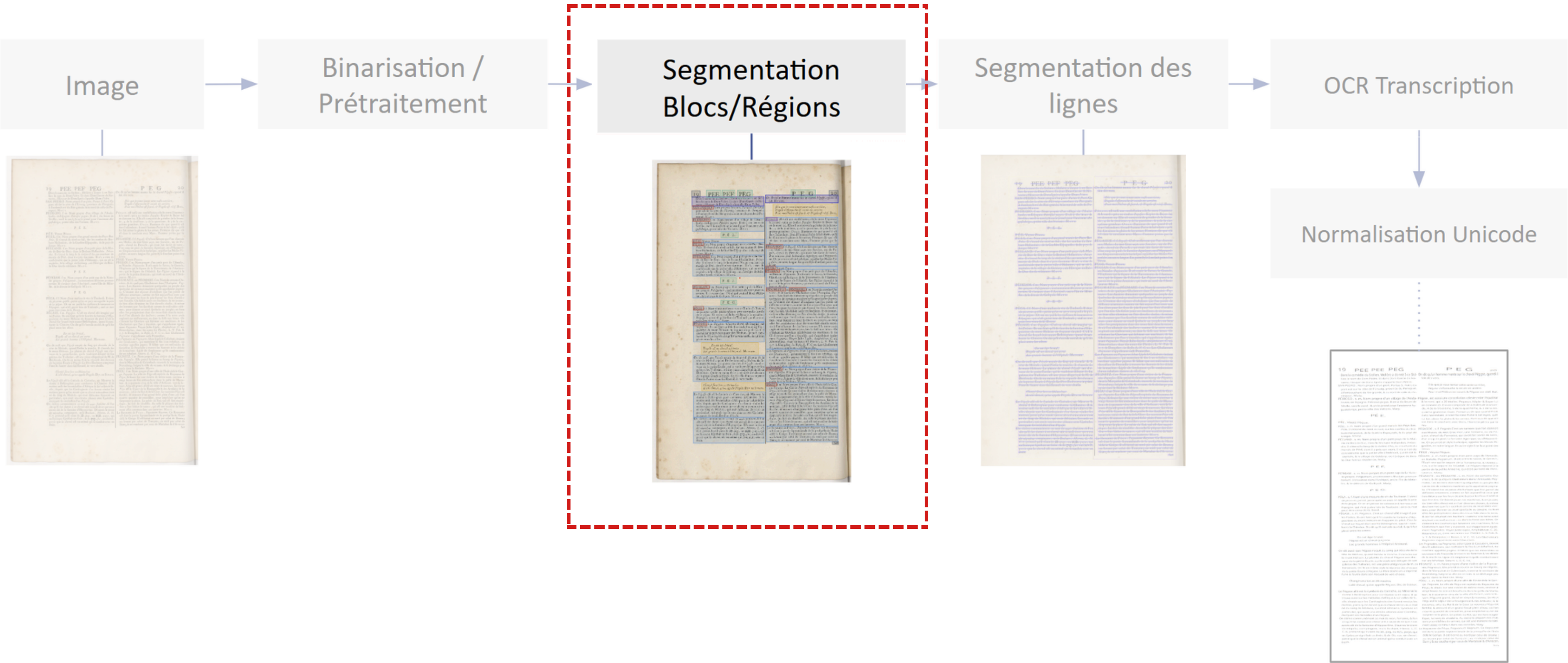
- DocLayout-YOLO
- Qwen 2.5-vl 3B

Segmentation lignes : blla (Kraken, par défaut)

OCR / HTR :

- EasyOCR
- Tesseract 5
- Surya
- Kraken + CATMuS-Print Large

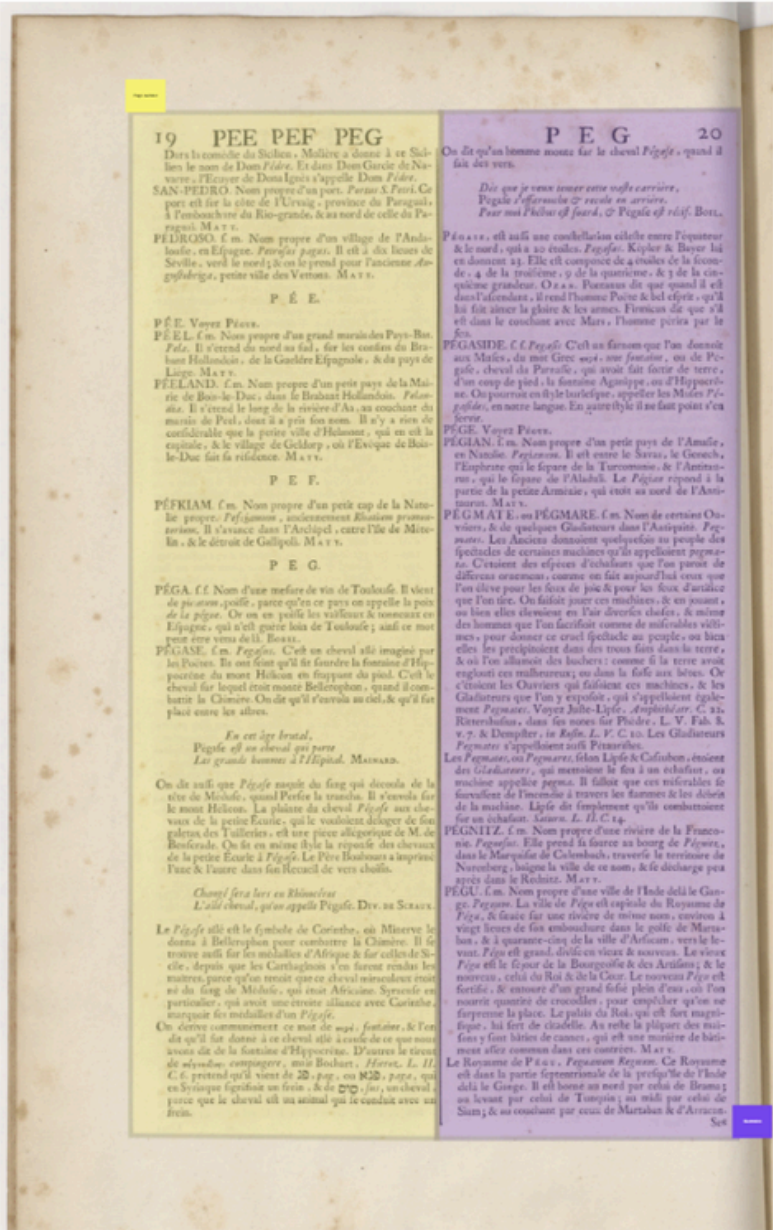
• Approche 2 : Segmentation Blocs / Régions



• Approche 2 : Segmentation Blocs / Régions

Segmentation à 2 niveaux :

- ① Colonnes (gauche / droite)
- ② Régions fines (12 classes)

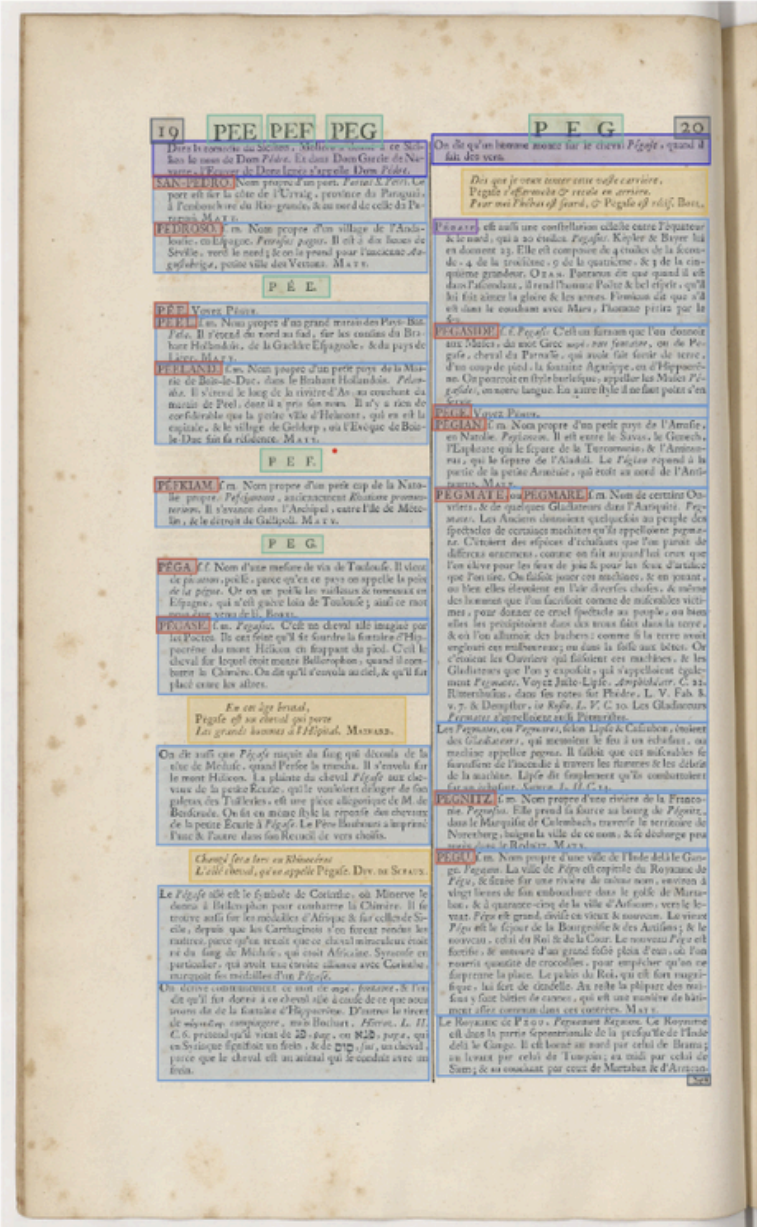


- Séparation gauche / droite
- Ordre de lecture rétabli
- Préalable à la segmentation des lignes



170 pages annotées •

12 classes typées •



→ ordre de lecture

→ structure de l'article

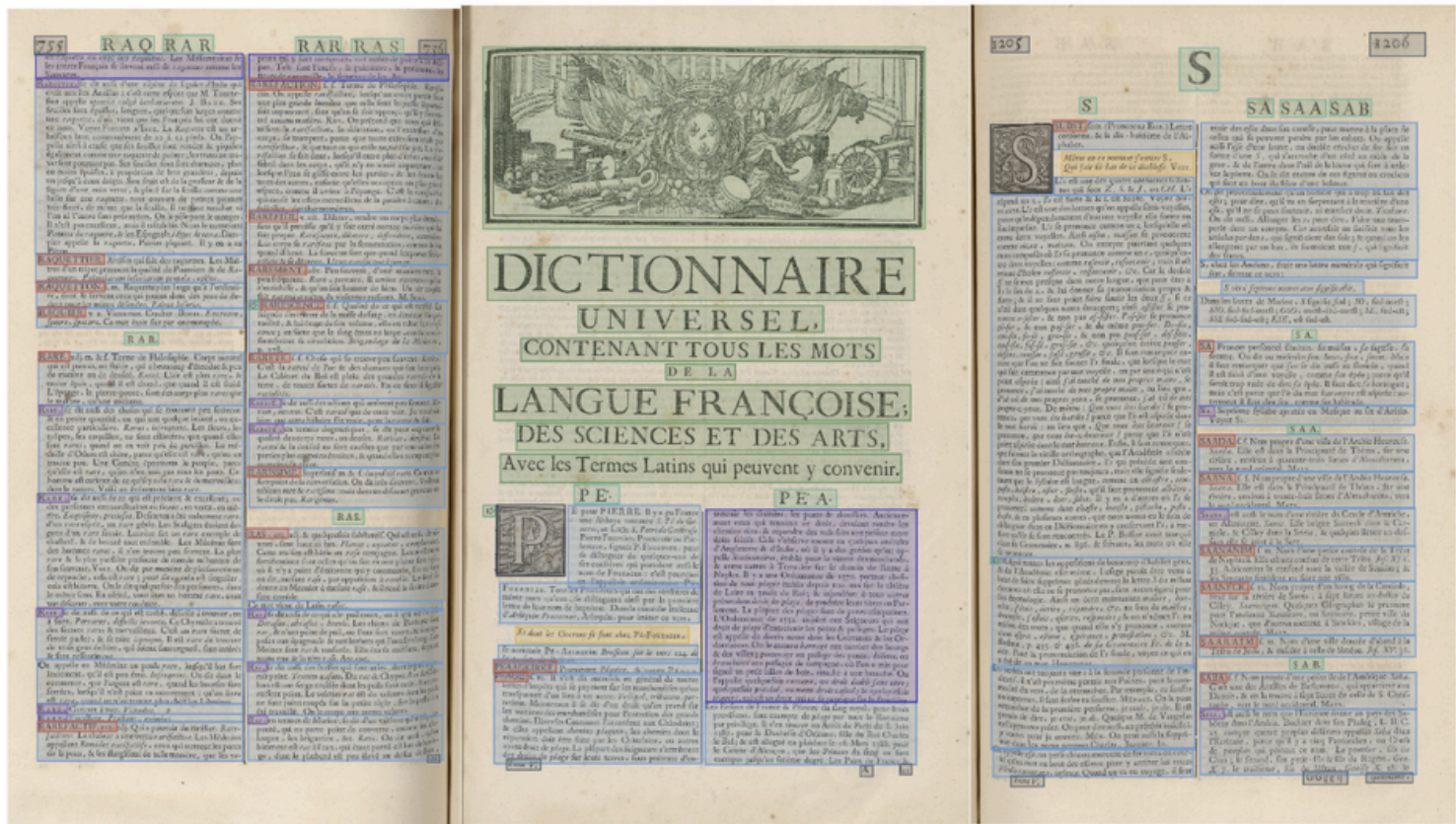
• Approche 2 : Segmentation régions fines

2 approches testées

Approche standard
Fine-tuning modèle DocLayout-YOLO

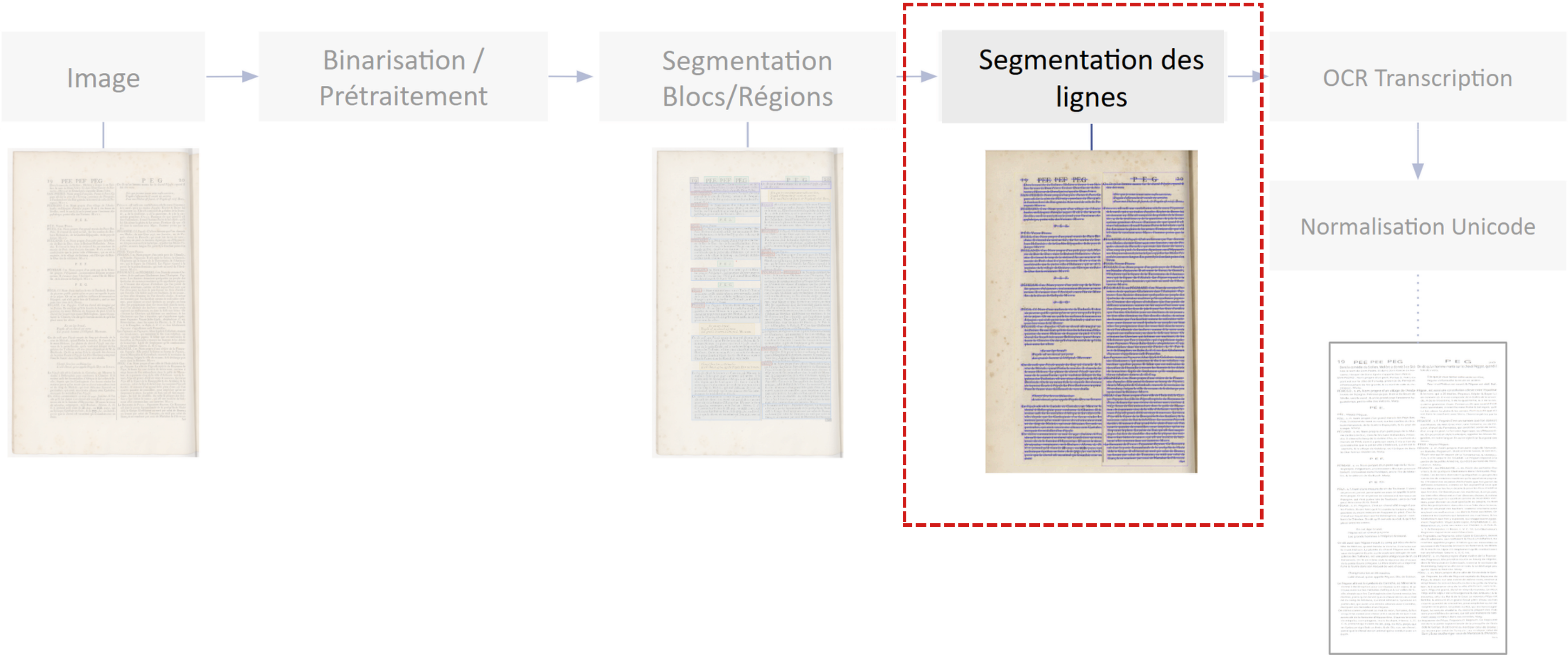
Approche générative
Fine-tuning modèle VLM Qwen2.5-VL

- Mot-vedette
- Vedette-secondaire
- Paragraphe
- Paragraphe-suite
- Citation
- Section-lexicale
- Manicule
- Lettrine
- Illustration
- Titre
- Autre-paratexte
- Tableau



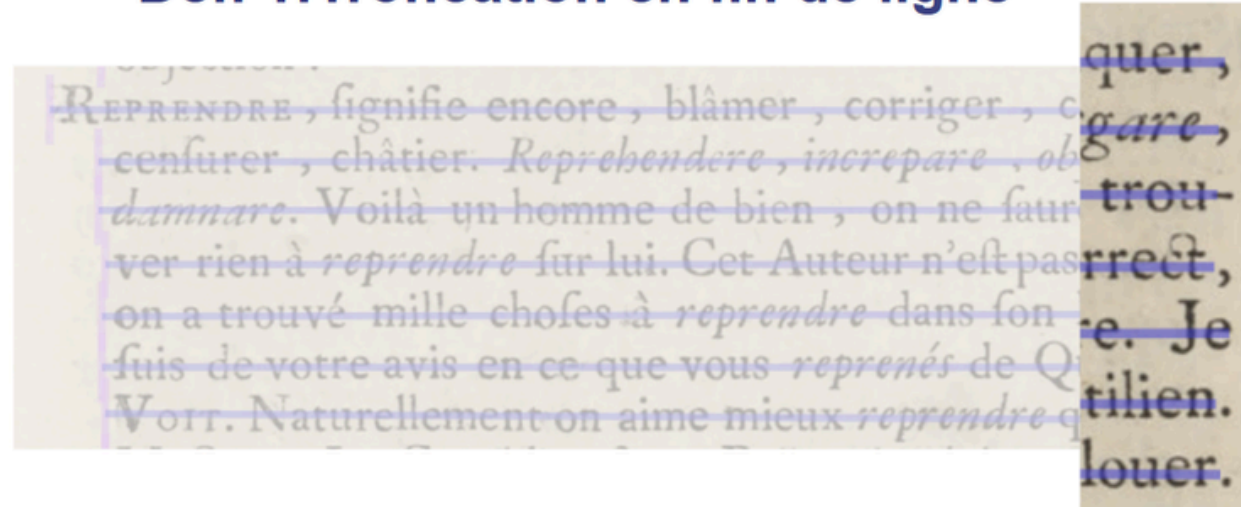
Métrique	DocLayout-YOLO (fine-tuné)	Qwen2.5-VL (fine-tuné)
mAP50 (50 pages)	0,921	0,914
mAP50-95	0,68	0,66
F1-Score	0,90	0,89

• Approche 2 : Segmentation des lignes

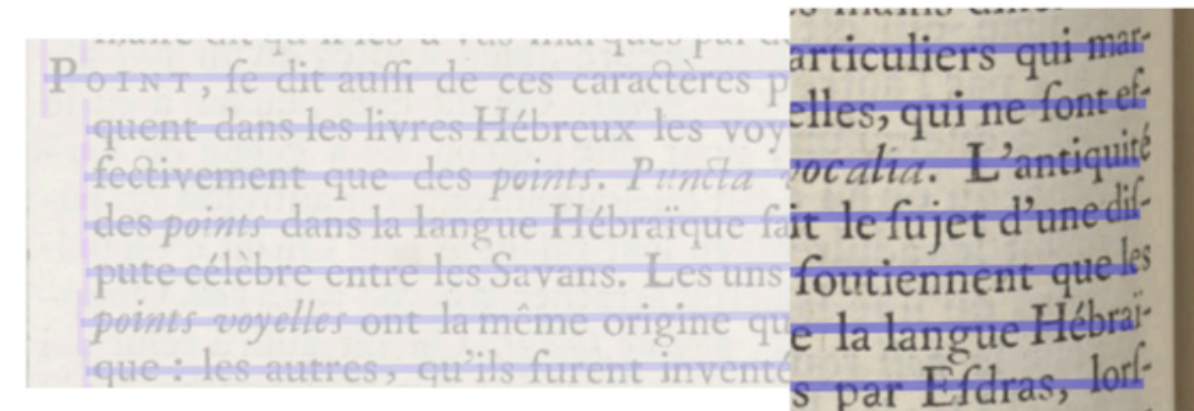


• Approche 2 : Segmentation des lignes

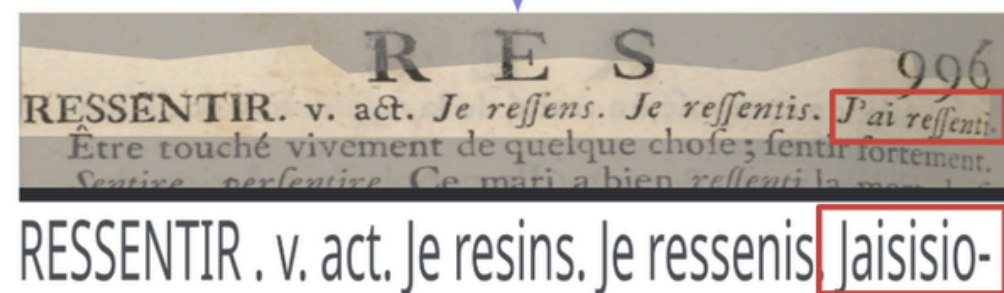
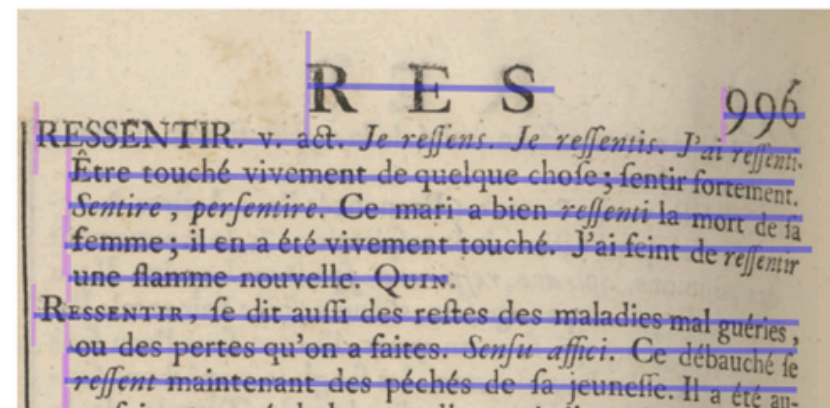
Défi 1: Troncation en fin de ligne



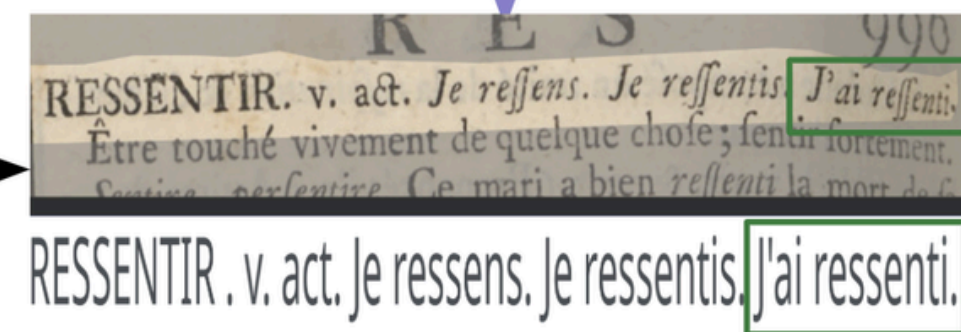
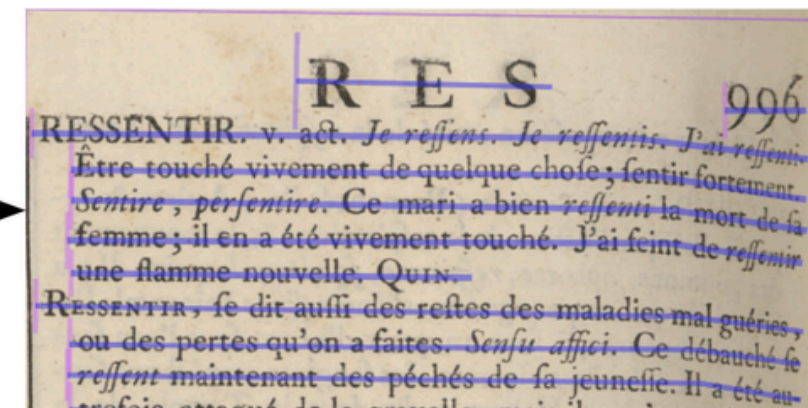
Défi 2: Courbure de reliure



Sans ajustement de segmentation de lignes



Avec ajustement de segmentation de lignes

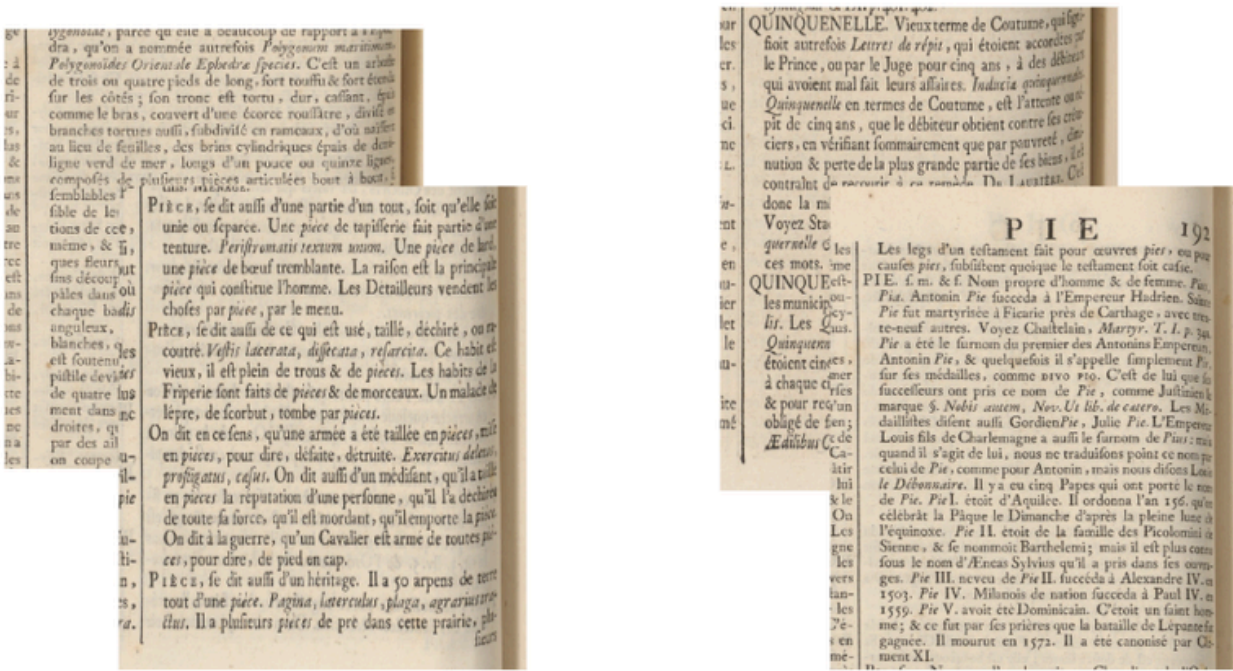
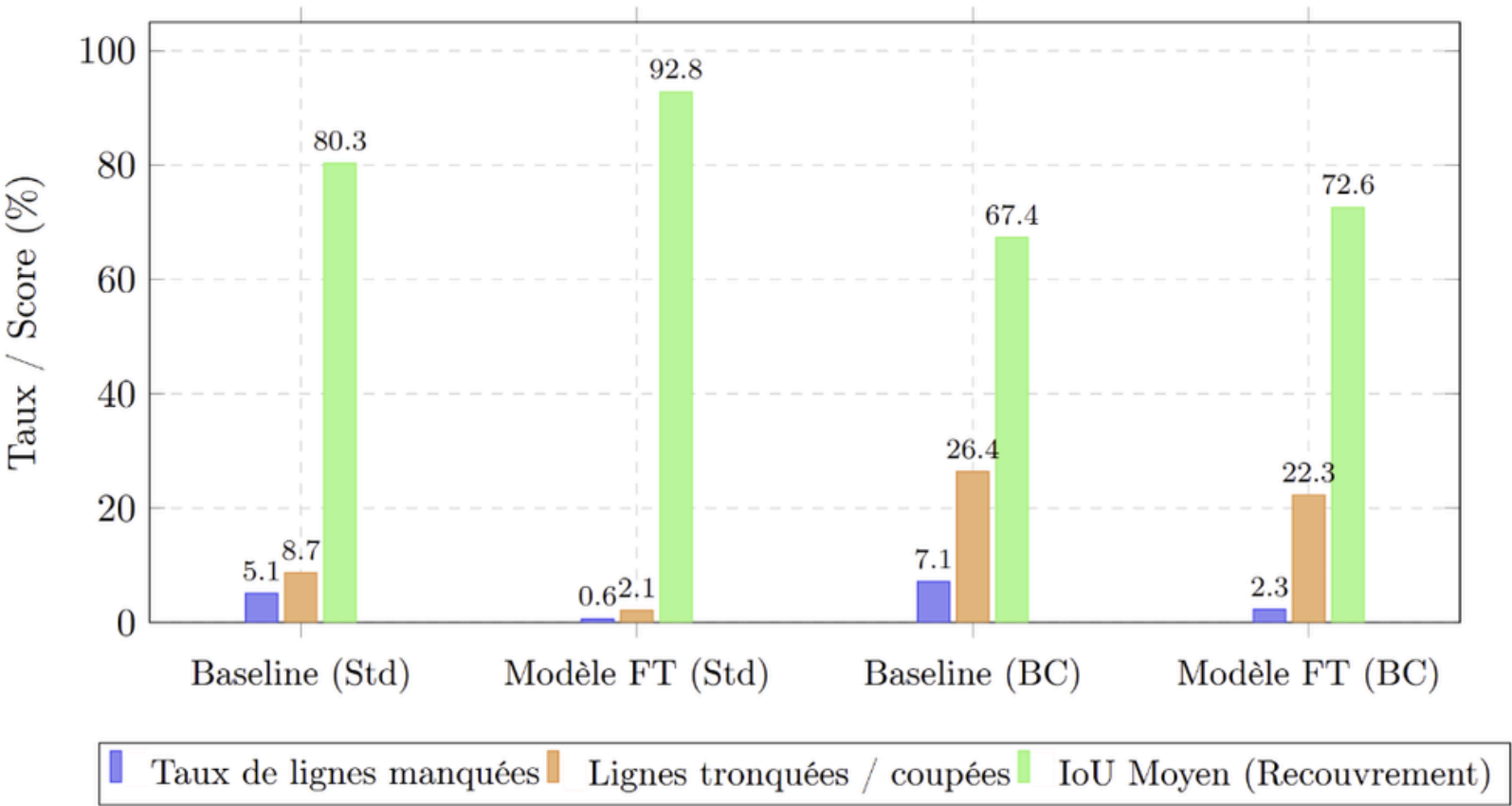


• Approche 2 : Segmentation des lignes

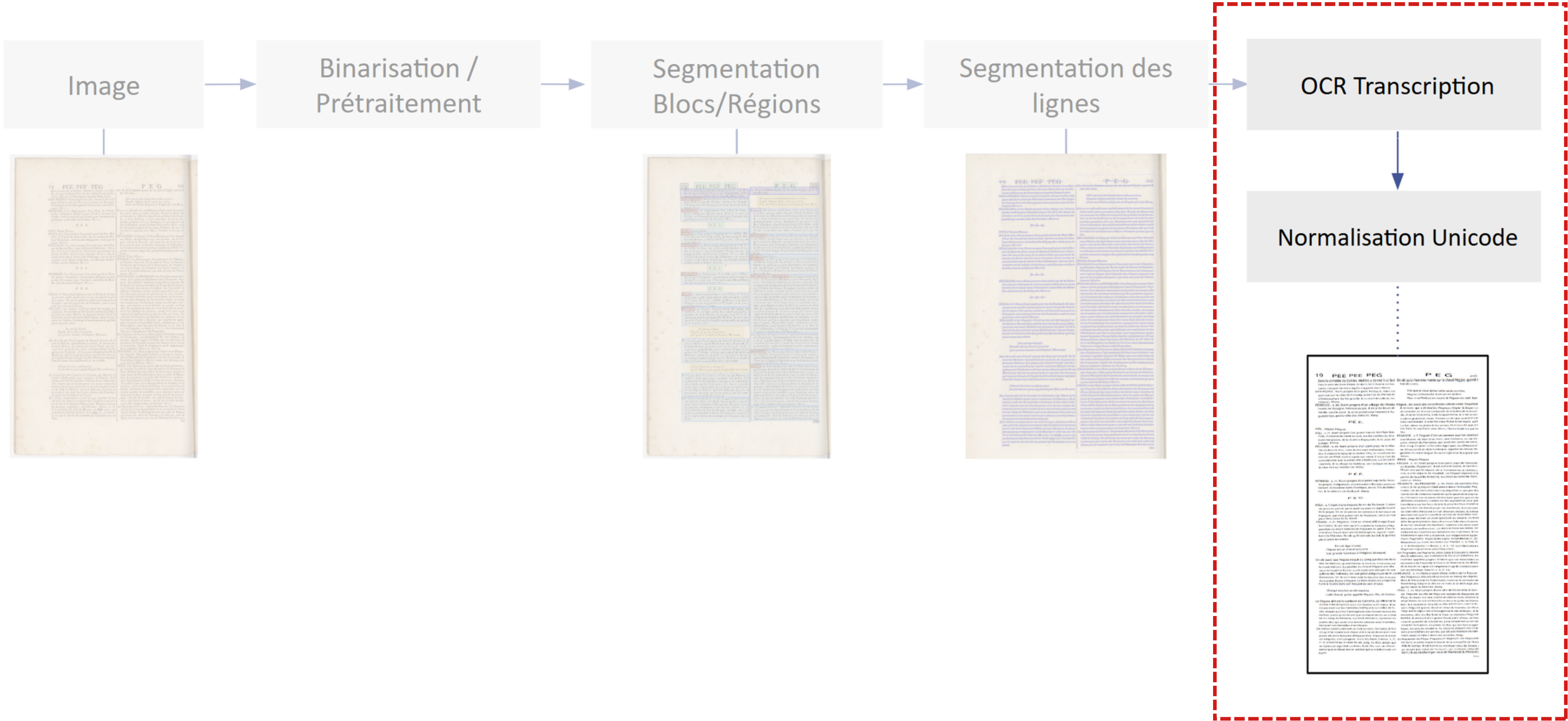
Modèle blla (kraken)

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*)
2. Perte des derniers caractères en fin de ligne



• Approche 2: Évaluation moteurs OCR / HTR traditionnels



• Approche 2: Évaluation moteurs OCR / HTR traditionnels

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 Seigneur de Vierzon ès Enquêtes de Parlement, de la Purification 1269. & contre le Comte de Bretagne, ès Arrêts de la Pentecôte 1273. & contre le Comte d'Artois, ès Arrêts de Toussaints 1277. Il paroît 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 Seigneur de Vierzon ès Enquêtes de Parlement, de la Purification 1269. & contre le Comte de Bretagne, ès Arrêts de la Pentecôte 1273. & contre le Comte d'Artois, ès Arrêts de Toussaints 1277. Il paroît 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.



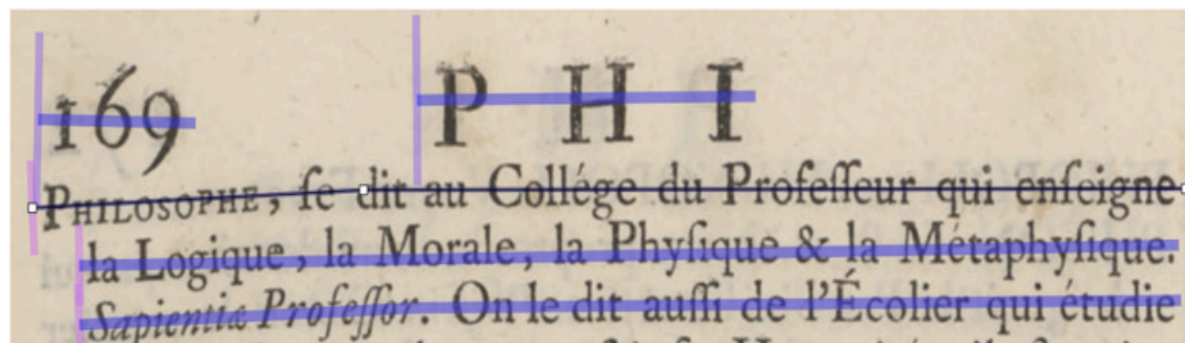
CER= 4,50 % encore trop élevé pour 956 pages → fine-tuning nécessaire

• Approche 2: Évaluation moteurs OCR / HTR traditionnels

eScriptorium : l'environnement de fine-tuning Kraken

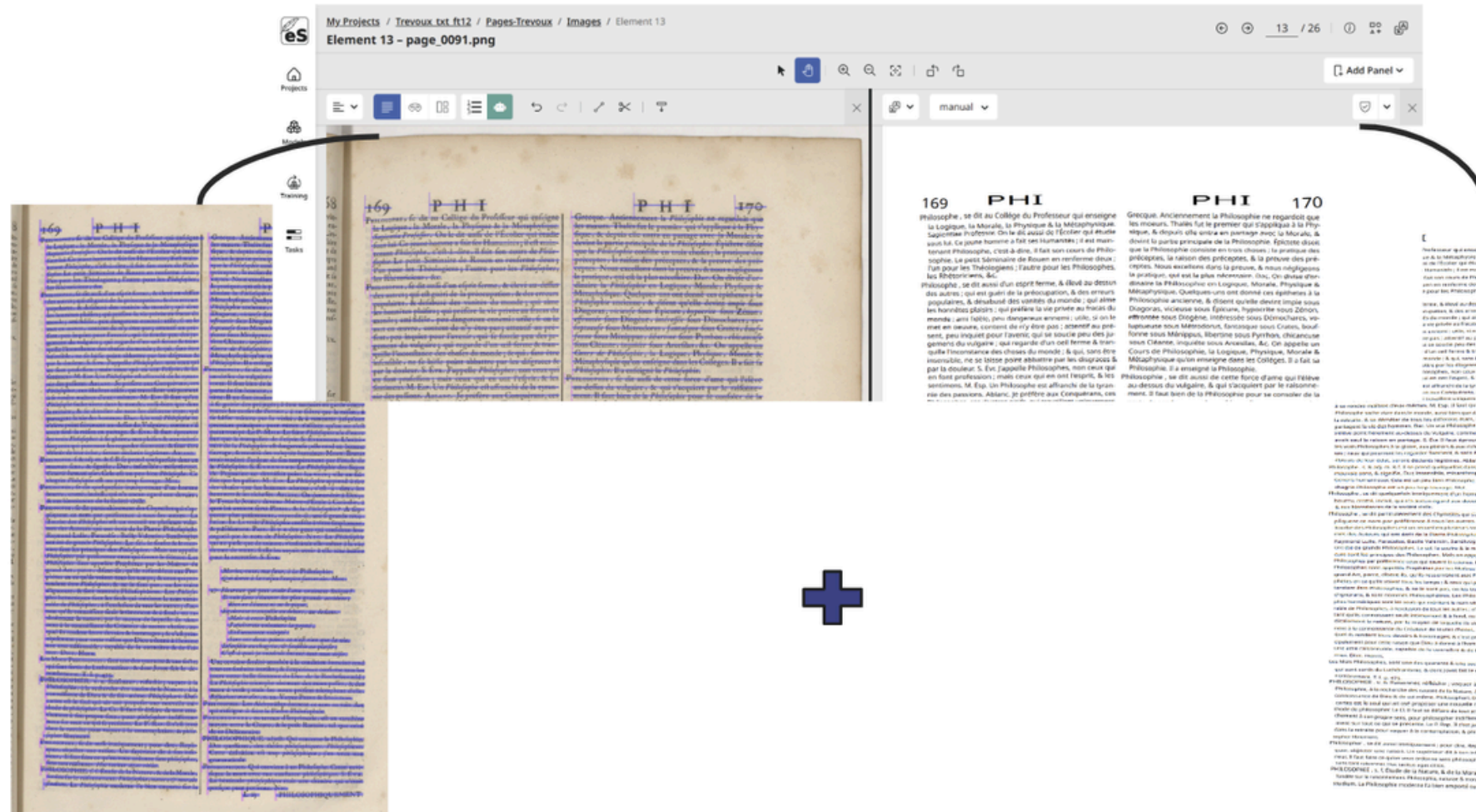


① Segmentation des lignes

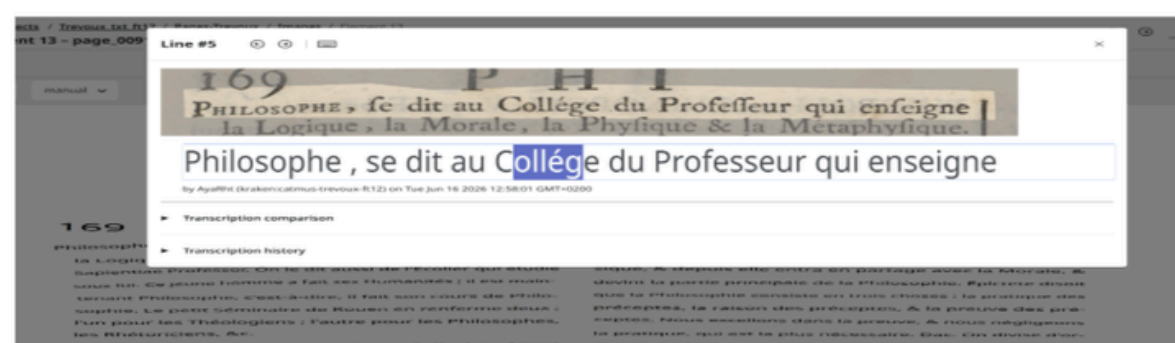


Ajustement des lignes segmentées par blla (modèle de base)

Segmentation, transcription, entraînement réunis:



② Transcription

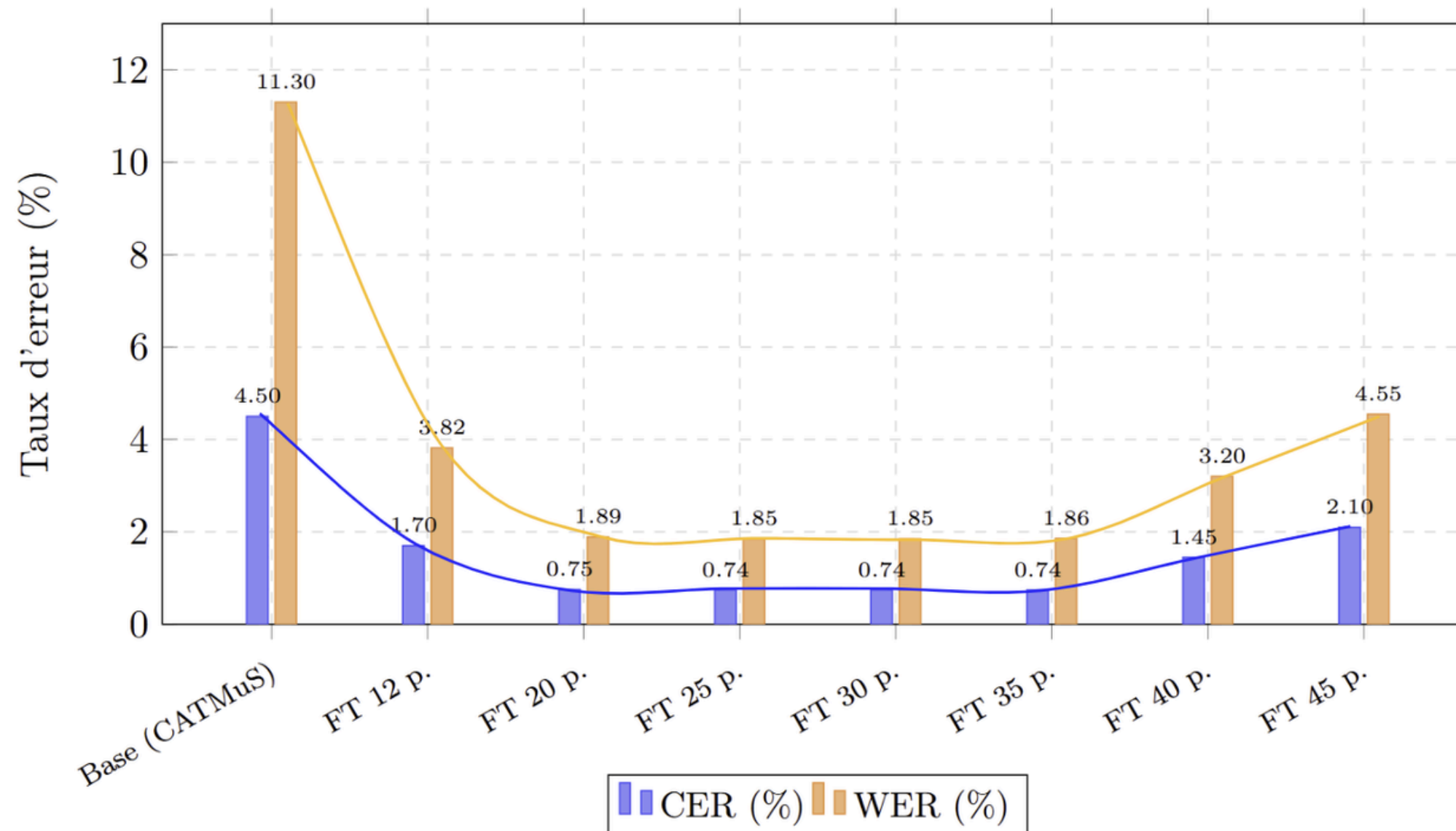


Correction manuelle de la transcription du modèle de base

• Approche 2: Évaluation moteurs OCR / HTR traditionnels

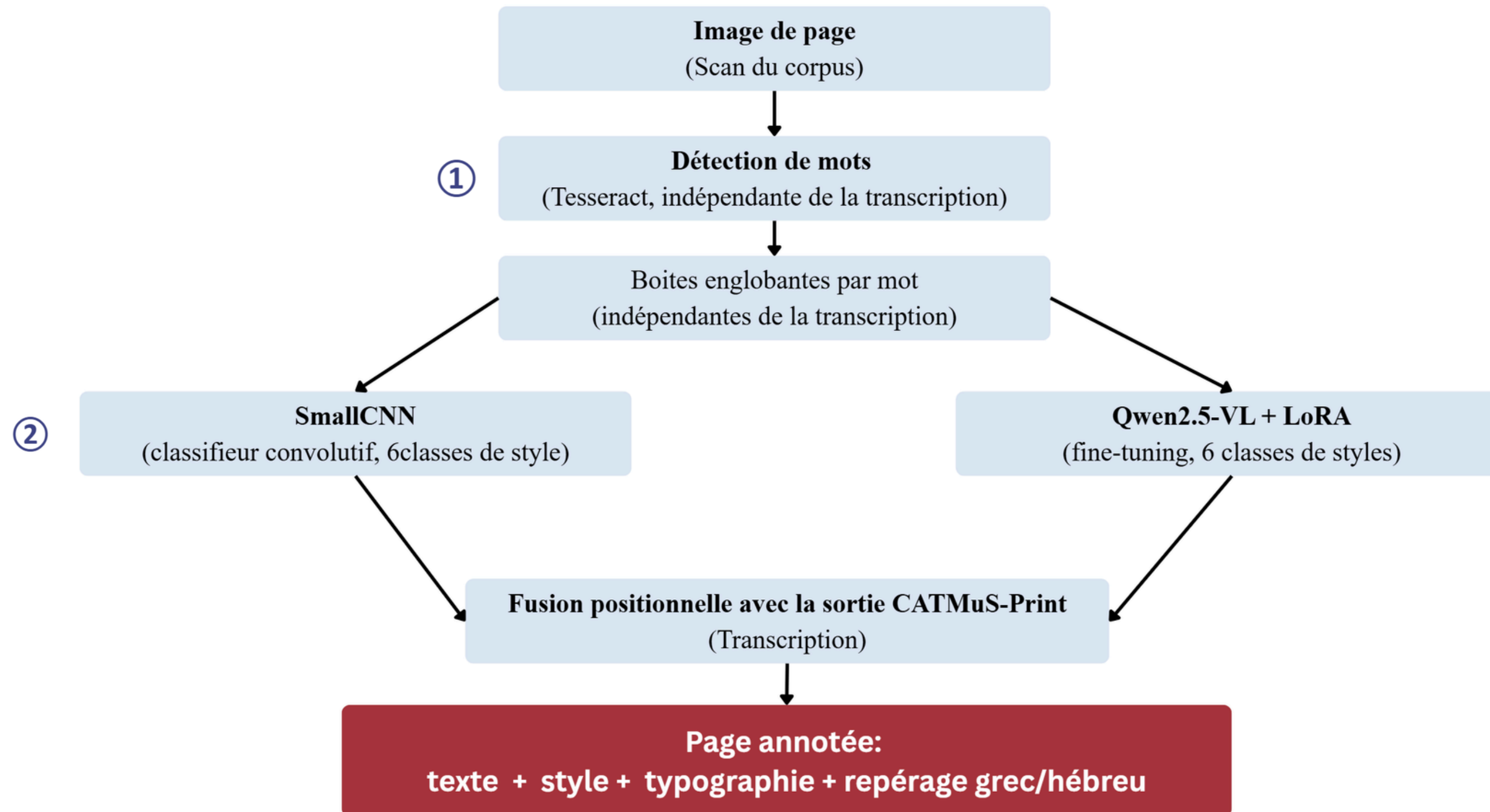
Fine tuning du modèle Kraken+ CATMuS-Print

- 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.



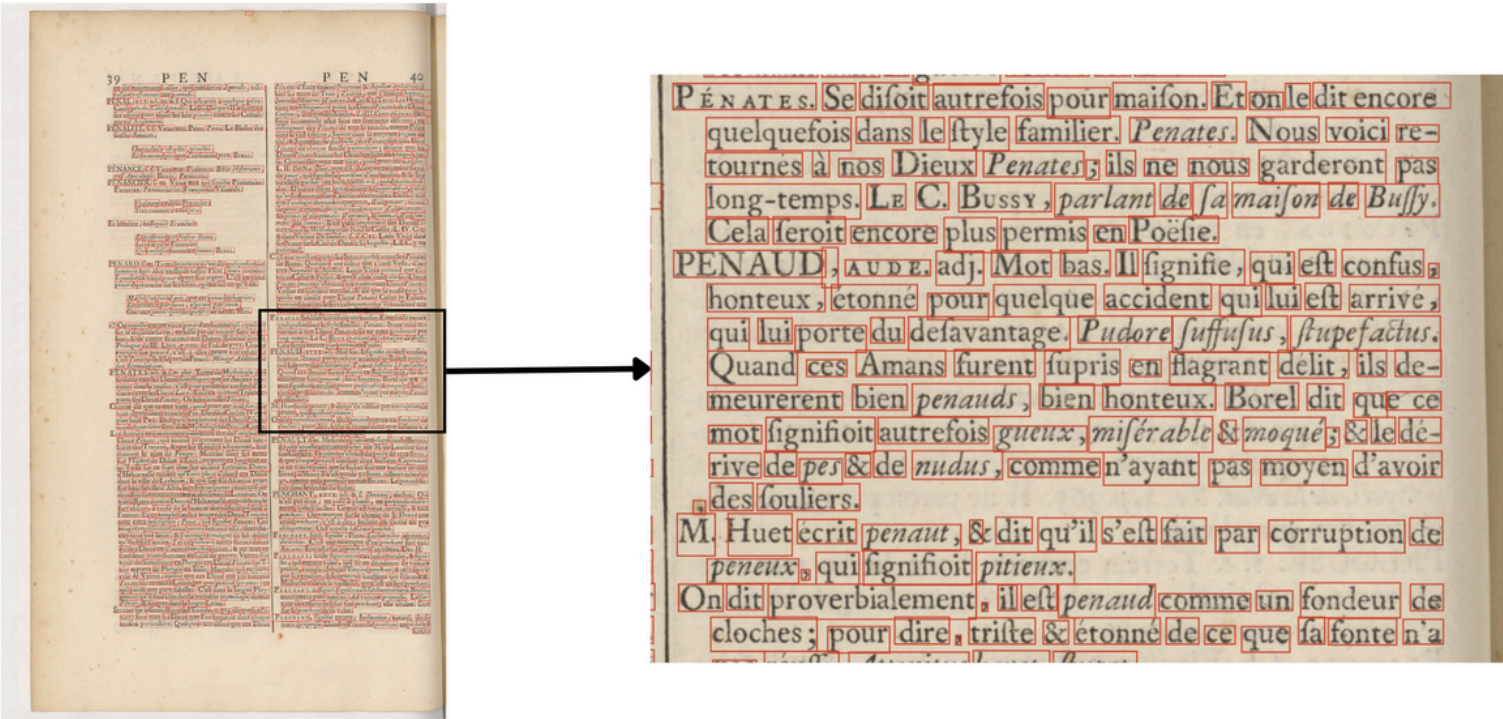
- FT 12 à 25 p. : Optimisation rapide
 - Alignement efficace sur la typographie du *Dictionnaire de Trévoux*.
- FT 25 à 35 p. : Zone de convergence
 - Plafond de performance atteint.
 - Minimum absolu : **CER à 0,738 % / WER à 1,85 %**.
- FT 40 à 45 p. : Overfitting
 - Mémorisation par cœur des bruits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

Détection de mots et classification de style

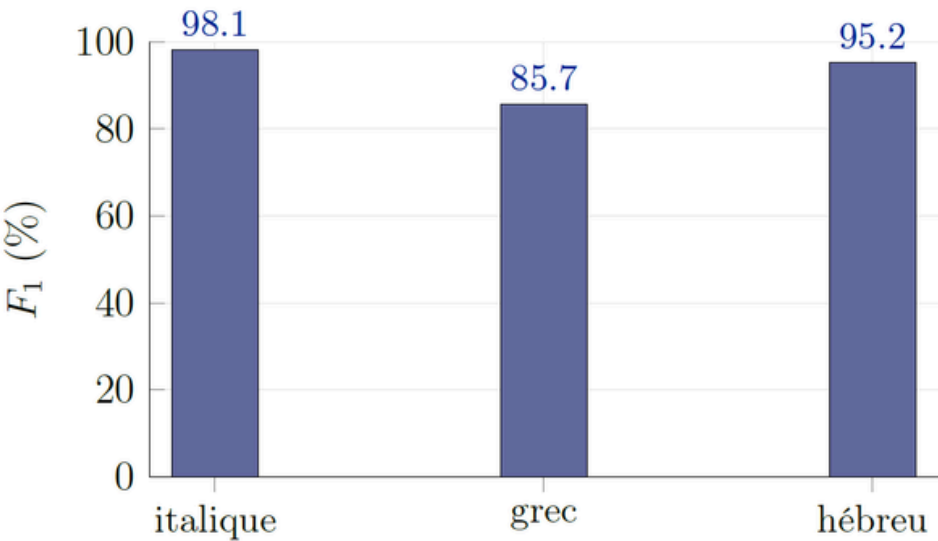


• Détection de mots et classification de style

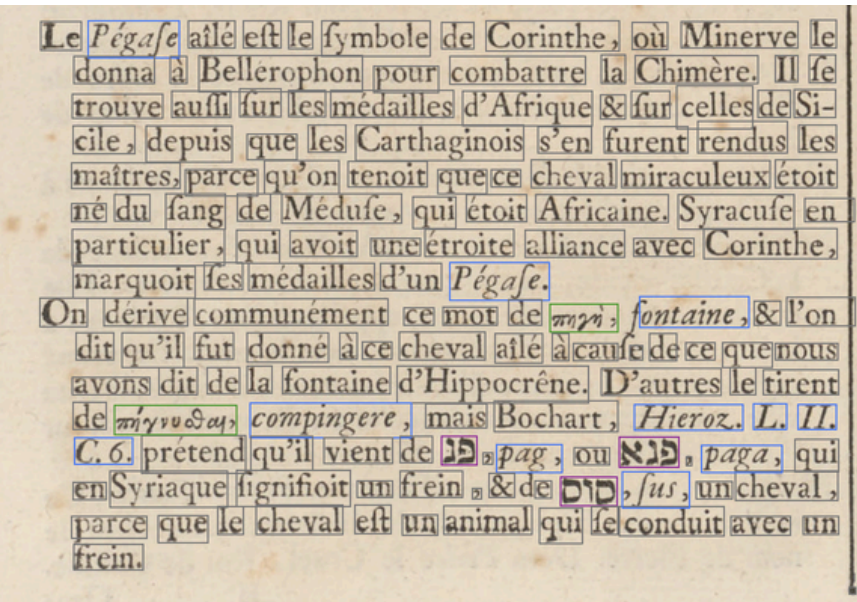
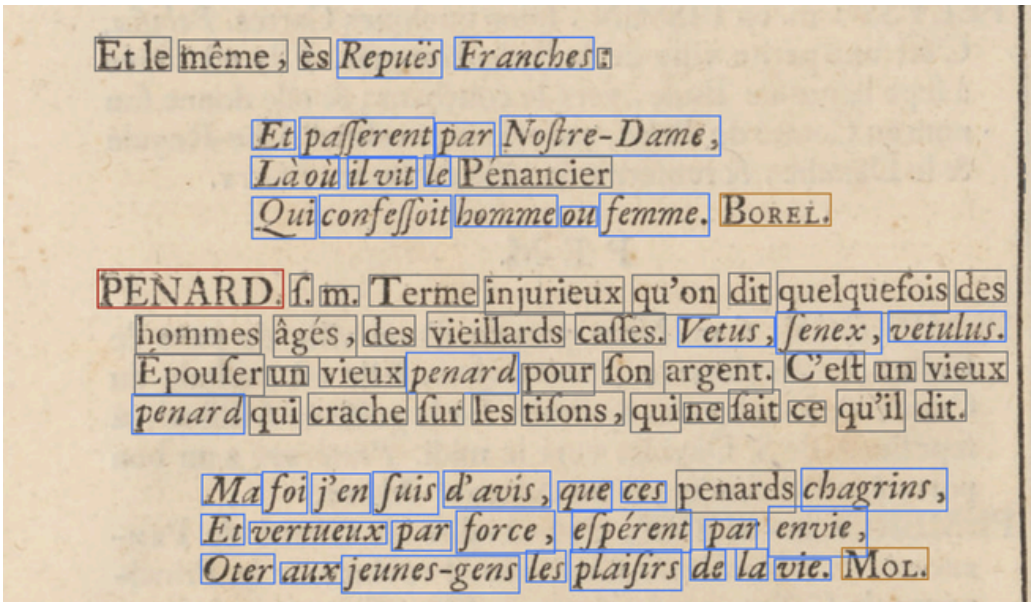
① Détection de mots (Tesseract) :



② Détection de styles :

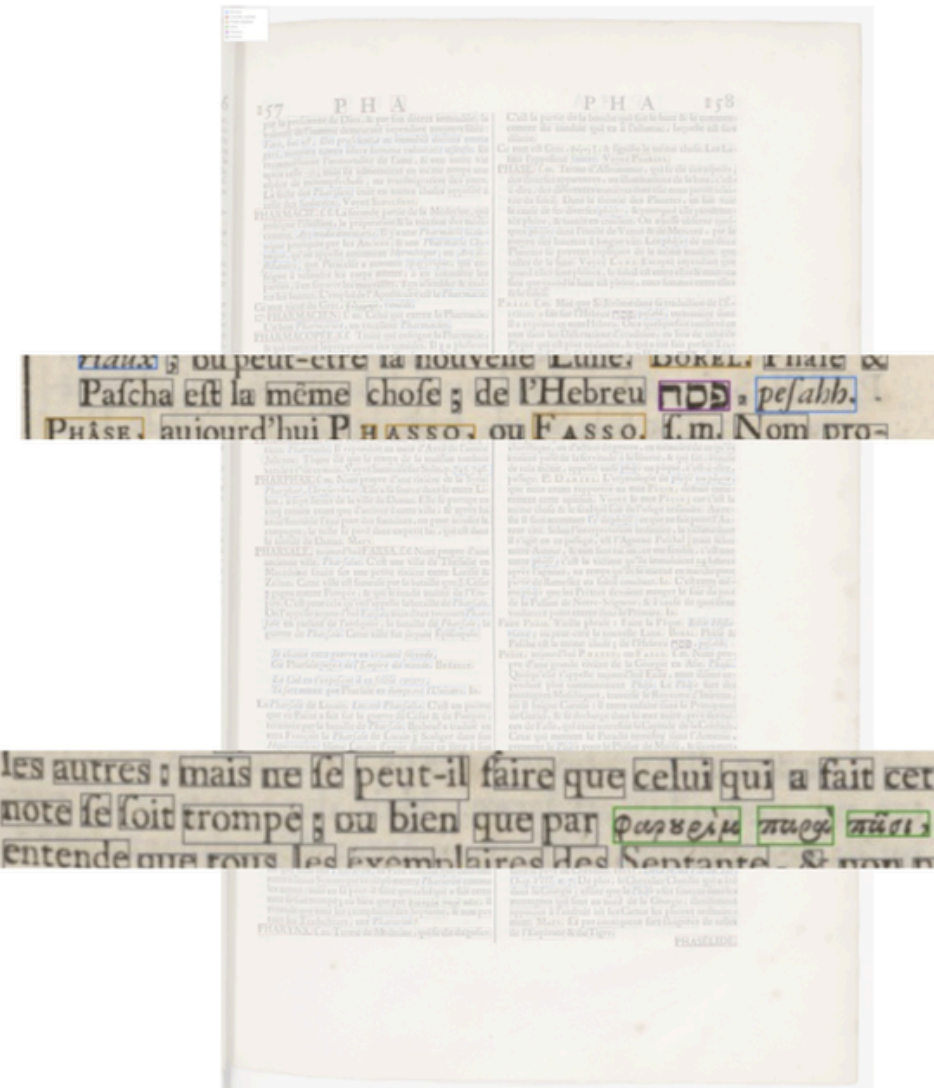
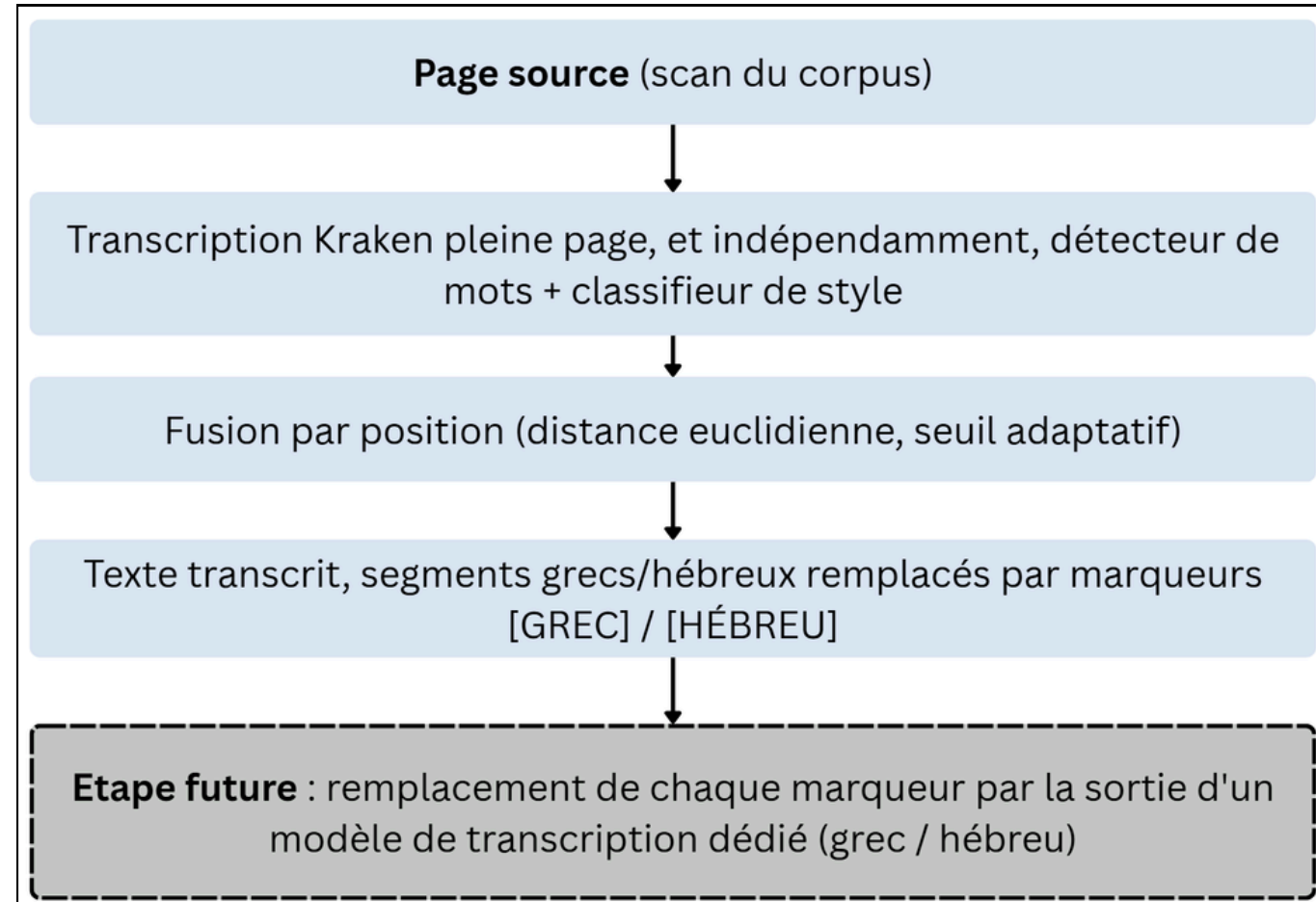


F1-mesure du classifieur de style Qwen2.5-VL sur le jeu de validation



- Détection de mots et classification de style

Fusion positionnelle avec la sortie CATMuS-Print



Pascha est la même chose ; de l'Hebreu **HÉBREU** pesahh.

note se soit trompé, ou bien que par **GREC** **GREC** **GREC** il

