Rapport de stage de recherche M2 DIAGI
Laboratoire LIRIS (UMR 5205 CNRS, INSA Lyon)

IA et Deep Learning : OCR et LLM pour la reconnaissance et la structuration de dictionnaires anciens

Élève :
Aya Rahouti

Tuteurs :
Tuteur pédagogique :
Mohsen Ardabilian

Tuteurs laboratoire :
Véronique Eglin
Ludovic Moncla

4 septembre 2026

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens tout d'abord à remercier du fond du cœur mes tuteurs de laboratoire, Véronique Eglin et Ludovic Moncla, pour leur encadrement bienveillant, leur immense disponibilité et leurs précieux conseils qui m'ont accompagnée à chaque étape de ce stage. Leur confiance et leur soutien ont été pour moi une source de motivation constante.

Je remercie également mon tuteur pédagogique, Mohsen Ardabilian, pour son suivi attentif tout au long de ce stage, ainsi que Lyes Nechak, responsable du Master 2 Génie Industriel, parcours DIAGI (Données et Intelligence Artificielle en Génie Industriel) que je suis, pour sa disponibilité et son accompagnement tout au long de cette année de master.

Je souhaite exprimer toute ma gratitude à ma famille, qui m'a soutenue avec tant d'amour tout au long de mes études en France, malgré la distance. Je pense à mon père, à ma mère, à ma sœur Nidal et à mon frère Ammar, dont les encouragements constants ont été une force précieuse à chaque étape de ce parcours. Sans eux, rien de tout cela n'aurait été possible.

Je remercie également mes amies Aya et Hind, toujours présentes pour moi.

Enfin, je remercie chaleureusement l'ensemble du personnel du laboratoire LIRIS, ainsi que mes collègues qui ont toujours été présents pour m'aider et rendre ce stage plus agréable. Je pense en particulier à Maritza Beatriz, que j'ai eu la chance de rencontrer lors de la journée d'étude *« Le Trévoux à l'ère des humanités numériques et de l'IA : expérimentations et résultats préliminaires »*, ainsi qu'à Nicolas Neila et aux autres collègues avec qui j'ai eu le plaisir de travailler sur le Trévoux, dans le cadre de projets différents.

Aya Rahouti

Septembre 2026

Lyon

Résumé

Ce stage de recherche s'inscrit dans le projet **OLR-LLM** (LIRIS, CNRS/INSA Lyon), qui vise la structuration automatique du *Dictionnaire Universel François-Latin* de Trévoux (édition 1743, environ 956 pages, mise en page à deux colonnes). Après un état de l'art des techniques de reconnaissance et de structuration appliquées aux corpus patrimoniaux (OCR spécialisé, segmentation de mise en page, classification de style typographique, correction post-OCR par LLM), deux familles d'approches ont été mises à l'épreuve sur des échantillons du corpus : une approche de bout en bout par modèle de vision-langage (VLM) et une approche modulaire combinant segmentation (DocLayout-YOLO) et OCR spécialisé (Kraken / CATMuS-Print). Une fois le modèle de reconnaissance affiné sur le corpus, la chaîne modulaire obtient un léger avantage (CER de 0,74 % contre 0,99 % pour la meilleure configuration VLM) et un coût de traitement à grande échelle nettement inférieur ; elle a été retenue pour la suite du stage. Un sous-système dédié de détection de mots et de classification de style typographique a par ailleurs été développé pour repérer automatiquement les insertions en grec ancien et en hébreu, que les moteurs OCR spécialisés testés ne savent pas transcrire. Les prochaines étapes consisteront à étendre ces expérimentations à l'ensemble du corpus et à approfondir la transcription spécialisée du grec et de l'hébreu.

Mots-clés : reconnaissance optique de caractères, dictionnaires anciens, Dictionnaire de Trévoux, modèles de vision-langage, segmentation de documents, classification de style typographique, humanités numériques.

Abstract

This research internship is part of the **OLR-LLM** project (LIRIS, CNRS/INSA Lyon), which aims at the automatic structuring of the *Dictionnaire Universel François-Latin* de Trévoux (1743 edition, approximately 956 pages, two-column layout). Following a literature review of recognition and structuring techniques applied to heritage corpora (specialized OCR, layout segmentation, typographic style classification, LLM-based post-OCR correction), two families of approaches were tested on samples of the corpus : an end-to-end vision-language model (VLM) approach and a modular approach combining layout segmentation (DocLayout-YOLO) with specialized OCR (Kraken / CATMuS-Print). Once fine-tuned on the corpus, the modular pipeline achieves a slight edge (0.74 % CER versus 0.99 % for the best VLM configuration) and a substantially lower processing cost at scale, and was therefore selected for the remainder of the internship. A dedicated word-detection and typographic-style classification sub-system was also developed to automatically flag ancient Greek and Hebrew insertions, which the specialized OCR engines tested were unable to transcribe. Next steps include extending these experiments to the full corpus and further developing specialized transcription of Greek and Hebrew.

Keywords : optical character recognition, historical dictionaries, Dictionnaire de Trévoux, vision-language models, document segmentation, typographic style classification, digital humanities.

Liste des abréviations

ALTO *Analyzed Layout and Text Object* : format XML standard décrivant la mise en page et le texte reconnu d'une page numérisée.

BnF Bibliothèque nationale de France.

CER *Character Error Rate* : taux d'erreur caractère.

CNN *Convolutional Neural Network* : réseau de neurones convolutif.

CNRS Centre National de la Recherche Scientifique.

CTC *Connectionist Temporal Classification*.

F1 F-mesure (moyenne harmonique de la précision et du rappel).

ICAR Interactions, Corpus, Apprentissages, Représentations (laboratoire, Université Lyon 2).

INSA Institut National des Sciences Appliquées.

IoU *Intersection over Union* : indice de recouvrement.

IXXI Institut Rhônalpin des Systèmes Complexes.

LIRIS Laboratoire d'InfoRmatique en Image et Systèmes d'information.

LLM *Large Language Model* : grand modèle de langage.

LoRA *Low-Rank Adaptation* : méthode de fine-tuning efficace en paramètres.

mAP *mean Average Precision* : précision moyenne.

METS *Metadata Encoding and Transmission Standard*.

MSH Maison des Sciences de l'Homme.

OCR *Optical Character Recognition* : reconnaissance optique de caractères.

OLR *Optical Layout Recognition* : reconnaissance optique de la mise en page.

TEI *Text Encoding Initiative*.

UMR Unité Mixte de Recherche.

VLM *Vision-Language Model* : modèle de vision-langage.

WER *Word Error Rate* : taux d'erreur mot.

Introduction générale

Les fonds patrimoniaux numérisés par les grandes bibliothèques, au premier rang desquelles la Bibliothèque nationale de France, rendent accessibles en image des dizaines de milliers d'ouvrages anciens. Leur exploitation scientifique suppose cependant que ce contenu soit reconnu, structuré et rendu interrogeable, ce qui reste un enjeu majeur pour les corpus patrimoniaux, en particulier pour les dictionnaires encyclopédiques anciens.

Ces derniers présentent en effet une organisation lexicale composite (vedettes, sous-entrées, exemples, citations) rarement matérialisée de façon univoque par la mise en page, dans une typographie datée (variations de casse, ligatures, caractères multilingues) que les moteurs de reconnaissance actuels peinent à traiter. Structurer automatiquement ce type de document, sans dénaturer sa graphie d'origine ni disposer d'annotations de référence à l'échelle de l'ouvrage complet, en fait un problème particulièrement difficile.

Les approches existantes, qu'il s'agisse de chaînes modulaires combinant segmentation et reconnaissance de caractères ou de modèles de vision-langage traitant en une seule passe l'image et le texte, apportent des réponses partielles à ce problème. Leur performance reste sensible à l'adaptation au domaine patrimonial : les moteurs génériques échouent fréquemment face aux typographies anciennes, et l'apport réel des grands modèles de langage pour la correction et la structuration demeure, selon les protocoles, débattu dans la littérature.

C'est dans ce contexte que s'inscrit le présent stage. Réalisé au laboratoire LIRIS (UMR 5205 CNRS, INSA de Lyon), dans le cadre du projet **OLR-LLM**, financé par l'IXXI et la MSH Lyon-Saint-Étienne et mené en collaboration avec le laboratoire ICAR, ce stage de recherche de Master 2 Données et Intelligence Artificielle en Génie Industriel (DIAGI), intitulé **« IA et Deep Learning : OCR et LLM pour la reconnaissance et la structuration de dictionnaires anciens»**, porte sur le *Dictionnaire Universel François-Latin* de Trévoux (édition de 1743). L'objectif est de dresser un état de l'art des techniques mobilisables pour la reconnaissance et la structuration de ce type de corpus, de mettre à l'épreuve deux familles d'approches (une chaîne modulaire de segmentation et d'OCR spécialisé, et une approche de bout en bout par modèle de vision-langage), et de documenter empiriquement les difficultés typographiques propres au Trévoux.

Ce rapport s'organise comme suit : Le premier chapitre présente l'organisme d'accueil et le cadrage du projet. Le deuxième dresse un état de l'art des techniques de reconnaissance optique de caractères, de segmentation de mise en page, de classification de style typographique et de correction post-OCR, appliquées aux documents patrimoniaux. Le troisième chapitre présente les approches concrètement testées durant le stage, ainsi que leurs résultats.

Chapitre I

Contexte et problématique

I Contexte et problématique

I.1 Introduction

Ce premier chapitre pose les bases nécessaires à la compréhension du travail réalisé durant le stage. Il présente successivement l'organisme d'accueil, le laboratoire LIRIS, et l'équipe au sein de laquelle s'inscrit ce travail, le contexte scientifique et institutionnel du projet OLR-LLM dans lequel s'inscrit ce stage, la problématique de structuration automatique des dictionnaires anciens qui en constitue le fil conducteur, puis les objectifs poursuivis et le plan retenu pour la suite de ce rapport.

I.2 Présentation de l'organisme d'accueil : le LIRIS

Le stage se déroule au **LIRIS** (Laboratoire d'InfoRmatique en Image et Systèmes d'information), une unité mixte de recherche (UMR 5205) rattachée au CNRS, à l'INSA de Lyon, à l'Université Claude Bernard Lyon 1, à l'Université Lumière Lyon 2 et à l'École Centrale de Lyon [1].
Le laboratoire structure son activité autour de **12 équipes de recherche**, regroupées en **6 pôles de compétences** :

- données, systèmes et sécurité.
- informatique graphique et géométrie.
- images, vision et apprentissage.
- interactions et cognition.
- algorithmique et combinatoire.
- simulation et sciences du vivant.

I.2.1 Chiffres clés et gouvernance

La figure 1 résume les principaux indicateurs du laboratoire, d'après sa fiche de présentation [1]. Le tableau 1 en détaille les effectifs, d'après la page dédiée du site du laboratoire [2] : un peu plus d'un tiers des **405 membres** du LIRIS sont des chercheurs et enseignants-chercheurs permanents, auxquels s'ajoutent 141 doctorants, ce qui traduit un laboratoire à forte composante doctorale, cohérent avec son rattachement à cinq tutelles distinctes (CNRS, INSA Lyon, Université Claude Bernard Lyon 1, Université Lumière Lyon 2, École Centrale de Lyon).

FIGURE 1 – Chiffres clés du LIRIS, d'après la fiche de présentation du laboratoire de 2022 [1].

Catégorie	Effectif
Membres du LIRIS (total)	405
Chercheurs et enseignants-chercheurs permanents	147
dont directeurs/chargés de recherche CNRS	12
dont professeurs des universités	49
dont maîtres de conférences et assimilés	86
Doctorants	141
Post-doctorants	21
Personnels administratifs, techniques et ingénieurs	54
Autres (stagiaires, ATER)	42

TABLE 1 – Effectifs du LIRIS par catégorie, d'après la page « Effectifs du LIRIS » [2].

La direction est assurée par Jean-Marc Petit (directeur), Guillaume Damiand et Véronique Eglin (directeurs adjoints) et Djemilia Cavret (directrice administrative et financière) [3]. Véronique Eglin, directrice adjointe du LIRIS, est aussi l'une des deux tutrices laboratoire du présent stage, aux côtés de Ludovic Moncla.

I.2.2 Un laboratoire tourné vers la culture et le patrimoine

Au-delà de la recherche amont, le LIRIS développe des interfaces avec plusieurs enjeux sociétaux (santé, environnement, apprentissage humain...), dont l'axe **culture et patrimoine** : bibliothèque numérique, édition critique, archivage, musée virtuel 3D [1]. Le projet OLR-LLM et le travail sur le Dictionnaire de Trévoux s'inscrivent directement dans cet axe.
Le projet **GEODE** [4] (*discours géographique encyclopédique*), financé par le LabEX ASLAN et mené depuis 2020 en collaboration entre le LIRIS, le laboratoire ICAR (Université Lyon 2) et le laboratoire EVS (Université de Saint-Étienne), avec la participation de l'Alan Turing Institute, illustre concrètement cet ancrage patrimonial : il étudie l'évolution du discours géographique dans les grandes encyclopédies françaises (*Encyclopédie* de Diderot et d'Alembert, *La Grande Encyclopédie*, *Encyclopædia Universalis*, Wikipédia) par classification semi-supervisée de textes et modélisation du langage. Codirigé par Ludovic Moncla (LIRIS), l'un des deux tuteurs laboratoire du présent stage, et Denis Vigier (ICAR), GEODE a étendu son périmètre en 2025 au *Dictionnaire Universel François-Latin* de Trévoux, avec la production d'une première édition numérique de l'édition de 1743 [48] : le présent stage, qui porte sur la reconnaissance et la segmentation de ce même dictionnaire, s'inscrit ainsi dans le prolongement direct des travaux de GEODE et du projet OLR-LLM déjà mentionné en introduction, deux initiatives portées par la même équipe et partageant le même corpus.

I.2.3 Les équipes Imagine et DM2L : un stage à leur intersection

Le stage se situe à l'intersection de deux des douze équipes du LIRIS, chacune rattachée à un pôle distinct : **Imagine** et **DM2L**, présentées ci-dessous avec l'ensemble des équipes du laboratoire (figure 2).

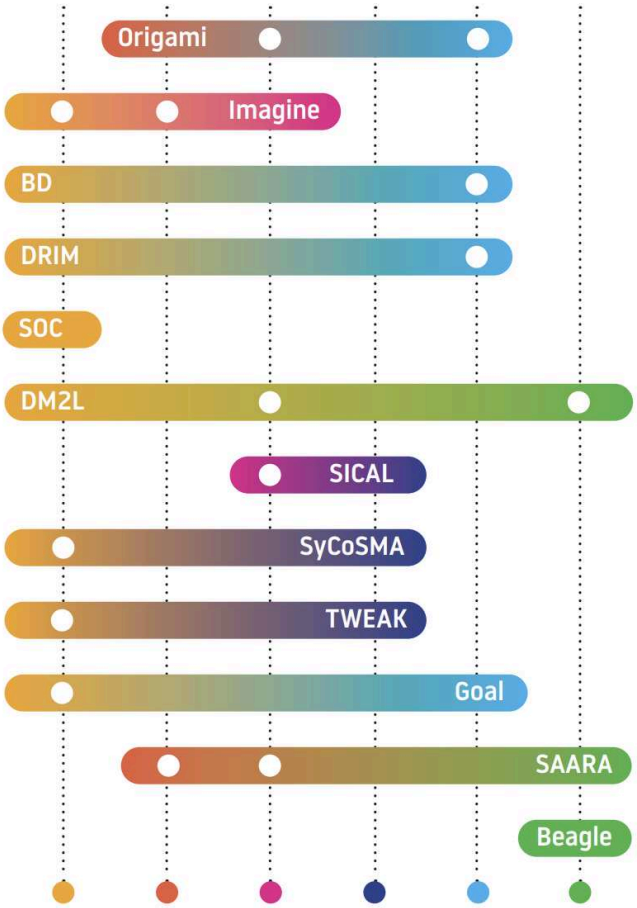


FIGURE 2 – Les douze équipes de recherche du LIRIS [1].

Imagine (*Computer Vision, Machine Learning, Pattern Recognition*), rattachée au pôle « Images, vision et apprentissage », travaille sur la compréhension de contenus multimédias. C'est d'elle que relève la partie « image » du stage : détection de mots, classification de style typographique, reconnaissance optique de caractères.

DM2L (*Data Mining and Machine Learning*), rattachée au pôle « Données, systèmes et sécurité », conçoit des modèles pour l'exploitation de masses de données hétérogènes. C'est d'elle que relève la partie « texte et structuration » du stage : une fois le contenu du dictionnaire reconnu, en structurer les entrées lexicographiques à l'aide de grands modèles de langage.
Le sujet du stage ne relève donc pleinement ni de l'une ni de l'autre équipe prise isolément, mais précisément de leur intersection, ce qui reflète la nature interdisciplinaire du projet OLR-LLM.

I.3 Contexte du stage

Les corpus patrimoniaux numérisés, tels que les dictionnaires et encyclopédies anciennes, constituent des sources précieuses pour la recherche en humanités numériques, en lexicographie historique et en histoire des savoirs. Leur exploitation scientifique est cependant souvent entravée par la complexité de leurs formats d'encodage, issus de

pipelines de numérisation standards, comme celui de la BnF, qui produit des fichiers XML METS/ALTO accompagnant des images haute définition. Ces formats mêlent informations structurelles, spatiales (position des blocs de texte sur l'image) et textuelles issues d'OCR, mais nécessitent un traitement avancé pour permettre une segmentation sémantique fiable des documents.

C'est dans ce contexte que s'inscrit le projet **OLR-LLM**, financé par l'IXXI (Institut Rhônalpin des Systèmes Complexes) et la MSH Lyon-Saint-Étienne, et mené en collaboration entre le laboratoire d'informatique **LIRIS** (UMR 5205 CNRS, INSA Lyon), présenté en section I.2, et le laboratoire de sciences du langage **ICAR**.

Ce projet propose une méthodologie combinant reconnaissance de structure via l'image (OCR/OLR) et interprétation linguistique via le texte (LLM), appliquée au *Dictionnaire Universel François-Latin* de Trévoux (1704 à 1771). Le présent stage porte sur l'édition de 1743 de ce dictionnaire (environ 956 pages numérisées, mise en page à deux colonnes).

I.4 Problématique

Les dictionnaires anciens présentent des structures internes complexes et peu normalisées : une entrée lexicographique peut s'étendre sur plusieurs colonnes et intégrer définitions, exemples, citations, remarques ou variantes orthographiques. Les modèles traditionnels d'analyse de documents peinent à identifier la granularité sémantique pertinente dans ce type de textes, souvent composites, hiérarchisés, voire redondants.
La question centrale de ce stage est donc la suivante : comment concevoir une chaîne de traitement automatique capable d'identifier et de 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 ?

I.5 Objectifs et plan du rapport

Le stage poursuit quatre objectifs complémentaires :

— Établir un état de l'art des techniques mobilisées pour la reconnaissance et la structuration de documents patrimoniaux (dictionnaires, journaux, manuscrits anciens), afin d'identifier celles transposables au Dictionnaire de Trévoux.

— Mettre à l'épreuve, sur un échantillon du corpus, **deux familles de solutions** : des **modèles de vision-langage (VLM)** traitant en une passe OCR et mise en page, et une **chaîne modulaire** (segmentation puis OCR spécialisé), et une chaîne modulaire (segmentation puis OCR spécialisé), avec une piste exploratoire de correction post-OCR par LLM.

— Détecter et classifier le **style typographique** des mots (romain, italique, petites et grandes capitales) ainsi que les **insertions en grec ancien et en hébreu**, deux écritures que les moteurs OCR spécialisés testés ne savent pas transcrire, afin de repérer fidèlement ces segments au sein des entrées lexicographiques.

— Documenter les difficultés propres au corpus du Trévoux (typographie du XVIII^{e} siècle, insertions multilingues, hétérogénéité de numérisation), pour ancrer les choix méthodologiques retenus dans des observations empiriques.

Ces objectifs déterminent le plan retenu : Le chapitre 2 développe l'état de l'art qui sous-tend les choix méthodologiques du stage, en suivant l'ordre des étapes d'une chaîne de traitement (transcription, segmentation, style, correction). Le chapitre 3 présente, de façon exploratoire, les mises en œuvre réalisées sur le corpus, avec leurs résultats quantitatifs et qualitatifs obtenus.

I.6 Conclusion

Ce chapitre a présenté le LIRIS, laboratoire d'accueil du stage, et les équipes Imagine et DM2L à l'intersection desquelles il se situe, avant de poser la problématique centrale : comment structurer automatiquement les entrées lexicographiques d'un dictionnaire ancien comme le Trévoux, sans dénaturer sa graphie d'origine ? Les quatre objectifs qui en découlent structurent la suite du rapport, à commencer par l'étude bibliographique du chapitre suivant.

Chapitre II

Étude bibliographique

II Étude bibliographique

II.1 Introduction

Avant de mettre à l'épreuve les approches retenues pour la reconnaissance et la structuration du Dictionnaire de Trévoux, il convient de dresser un état de l'art des travaux existants sur la reconnaissance et la structuration de documents patrimoniaux. Ce deuxième chapitre constitue le socle théorique sur lequel s'appuie l'ensemble de ce travail, et suit un fil conducteur allant du plus général au plus spécifique.
Après un rappel de la méthodologie de recherche bibliographique, l'étude suit l'ordre même de la chaîne de traitement mise en œuvre au chapitre 3 : transcription, segmentation et structuration de la mise en page, puis style typographique. Le chapitre revient ensuite sur le débat entre pipelines explicites et modèles de fondation, avant d'examiner l'apport encore incertain de la correction post-OCR par grands modèles de langage, puis d'appliquer l'ensemble de ces enseignements à l'objet même du stage, le Dictionnaire de Trévoux.

II.2 Méthodologie de la recherche bibliographique

II.2.1 Espace de recherche

La recherche bibliographique a porté sur plusieurs types de sources :
- Les actes de conférences spécialisées en analyse de documents (ICDAR, workshops DAS et HIP).
- Les travaux du domaine des humanités numériques (Journal of Data Mining and Digital Humanities, ACM Computing Surveys).
- Les ateliers dédiés au traitement automatique des langues appliqué aux corpus patrimoniaux (LaTeCH-CLfL, LT4HALA).
- Des prépublications arXiv récentes évaluant des modèles de vision-langage sur des documents historiques.

Une attention particulière a été portée aux travaux menés au LIRIS et en collaboration avec le laboratoire ICAR, dans le prolongement direct du projet OLR-LLM, notamment ceux du groupe de recherche GEODE consacré à l'étude computationnelle des encyclopédies françaises.

II.2.2 Sélection des mots-clés

Axe thématique	Mots-clés
Lexicographie numérique	*digital lexicography, historical dictionary, TEI encoding*
OCR pour documents anciens	*historical OCR, OCR engine benchmark, layout analysis, object detection*
Style et script	*font classification, typographic emphasis detection, script identification, multilingual document*
Correction post-OCR	*post-OCR correction, LLM correction, historical text normalization*

TABLE 2 – Mots-clés utilisés pour la recherche bibliographique, par axe thématique.

II.2.3 Volume

La combinaison des mots-clés du tableau 2, croisée sur les bases et actes listés en section II.2.1, a produit un premier ensemble d'environ 163 références. Un premier tri sur le titre et le résumé, excluant les travaux ne portant ni sur un corpus historique ou patrimonial, ni sur une évaluation quantitative reproductible, a ramené ce volume à 74 articles. La figure 3 détaille cet entonnoir de sélection jusqu'aux 51 références effectivement citées dans ce rapport.

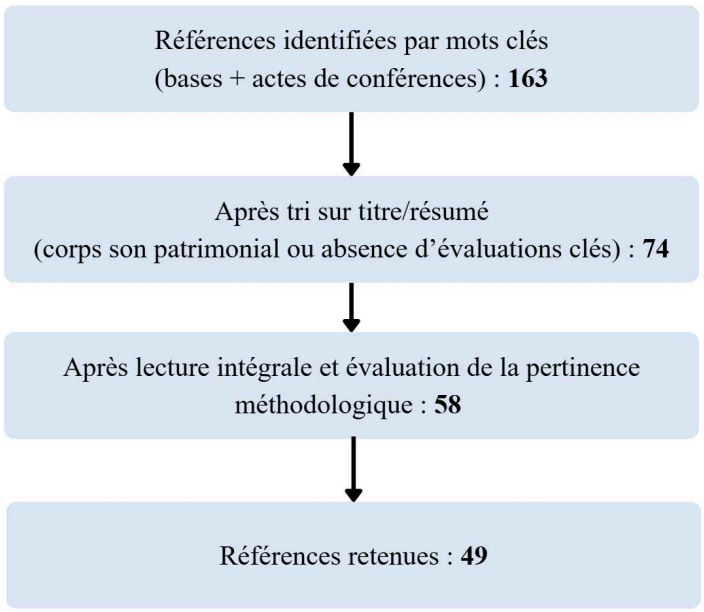


FIGURE 3 – Entonnoir de sélection bibliographique.

II.2.4 Articles étudiés en profondeur

Parmi les 58 articles lus intégralement, une attention particulière a été portée à ceux rapportant des résultats chiffrés directement comparables aux expérimentations menées durant le stage : évaluations de moteurs OCR sur corpus historiques (Romanello et al. [5],

Levchenko [6], Agbeti-Messan et al. [7]), campagnes de comparaison de détecteurs de mise en page (Clérice [8], Zhao et al. [9]), et études de correction post-OCR par LLM (Boros et al. [10], Thomas et al. [11]). Ce sont ces travaux qui fournissent l'essentiel des points de comparaison mobilisés dans le chapitre consacré aux expérimentations et résultats.

II.3 Panorama chronologique des approches

Avant d'entrer dans le détail de chaque étape de la chaîne de traitement, la figure 4 situe les travaux cités dans ce chapitre sur une échelle temporelle, des premiers systèmes de reconnaissance de caractères fondés sur l'apprentissage supervisé classique jusqu'aux modèles de vision-langage les plus récents. Cette mise en perspective montre que les problèmes traités durant le stage, reconnaissance de caractères, analyse de mise en page, identification de style, structuration lexicographique, ne sont pas nouveaux : ce qui change d'une décennie à l'autre est principalement le paradigme d'apprentissage (règles → réseaux de neurones convolutifs → séquentiel récurrent → Transformer → modèles de fondation multimodaux).

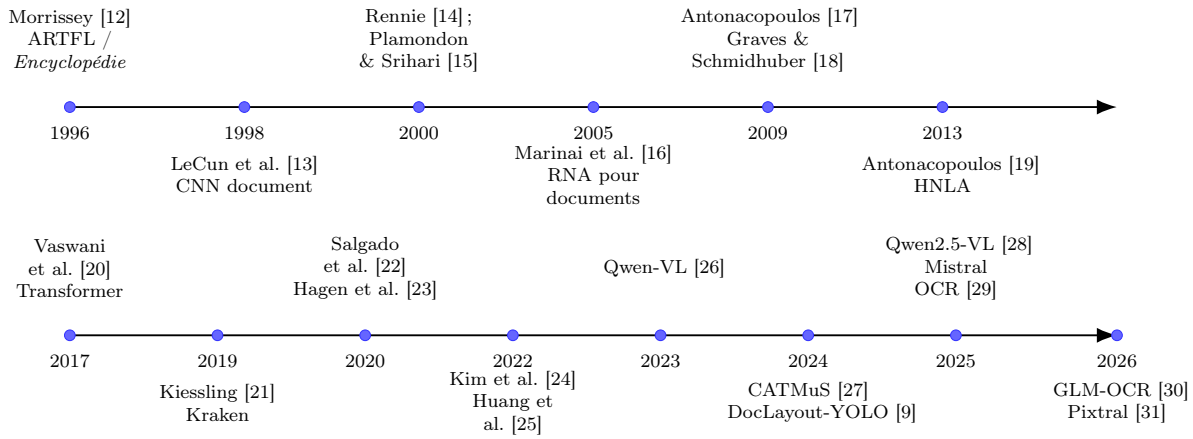


FIGURE 4 – Panorama chronologique des travaux cités.

II.4 Mesures d'évaluation utilisées dans la littérature

Avant de détailler les résultats chiffrés rapportés par les travaux cités dans ce chapitre, il est utile de formaliser les principales mesures d'évaluation qu'ils mobilisent de façon récurrente pour comparer moteurs OCR, systèmes de segmentation de mise en page et méthodes de correction post-OCR. Ces relations mathématiques sont directement tirées de la définition qu'en donnent les articles cités dans la partie bibliographie.

II.4.1 Distance d'édition de Levenshtein

Les taux d'erreur caractère et mot présentés ci-après reposent tous deux sur la distance d'édition de Levenshtein, qui mesure le nombre minimal d'opérations (substitution, suppression, insertion) permettant de transformer une séquence en une autre. Pour deux séquences $a = a_1 \dots a_m$ (transcription produite) et $b = b_1 \dots b_n$ (vérité terrain), cette

distance $d(m, n)$ se calcule par programmation dynamique à l'aide de la récurrence :

$$d(i,j) = \begin{cases} \max(i,j) & \text{si } \min(i,j) = 0 \\ \min\Big\{ d(i-1,j)+1, \quad d(i,j-1)+1, & \\ \qquad d(i-1,j-1) + \mathbb{1}[a_i \neq b_j] \Big\} & \text{sinon} \end{cases} \tag{1}$$

où $\mathbb{1}[a_i \neq b_j]$ vaut 1 si les deux symboles diffèrent (substitution) et 0 sinon. Le nombre de substitutions S, suppressions D et insertions I utilisé dans les équations 2 et 3 correspond au décompte de chacune de ces opérations le long du chemin optimal réalisant $d(m, n)$. C'est cette même récurrence, appliquée au niveau du caractère ou du mot, qu'utilisent Romanello et al. [5], Levchenko [6] et Agbeti-Messan et al. [7] pour calculer respectivement leurs CER et WER.

II.4.2 Taux d'erreur caractère et taux d'erreur mot (CER, WER)

Le CER (*Character Error Rate*) est la mesure de référence employée par la quasi-totalité des travaux cités dans ce chapitre pour évaluer la qualité d'une transcription OCR, notamment Romanello, Najem-Meyer et Robertson [5], Levchenko [6], Agbeti-Messan et al. [7], Greif, Griesshaber et Greif [32], ainsi que les études de correction post-OCR de Boros et al. [10] et Thomas, Gaizauskas et Lu [11]. Il est défini à partir de la distance d'édition de Levenshtein entre la transcription produite $\hat{y}$ et la vérité terrain y, comme le nombre minimal d'opérations d'édition élémentaires (substitutions S, suppressions D, insertions I) nécessaires pour transformer $\hat{y}$ en y, rapporté au nombre total de caractères N de la vérité terrain :

$$\mathrm{CER} = \frac{S + D + I}{N} \tag{2}$$

Le WER (*Word Error Rate*) est défini de façon analogue à l'échelle du mot plutôt que du caractère, S_w, D_w, I_w désignant respectivement le nombre de substitutions, suppressions et insertions de mots, et N_w le nombre total de mots de la référence :

$$\mathrm{WER} = \frac{S_w + D_w + I_w}{N_w} \tag{3}$$

Ces deux mesures, généralement exprimées en pourcentage, sont celles utilisées dans l'ensemble des tableaux comparatifs de ce chapitre (tableaux 6 et 12).

II.4.3 Précision, rappel et F-mesure

Pour les tâches de structuration logique et d'extraction d'information, Gutehrlé et Atanassova [33] ainsi que Huang et al. [25] (LayoutLMv3, évalué sur FUNSD et CORD) rapportent des scores de précision P, de rappel R et de F-mesure F_1, définis à partir du nombre de vrais positifs VP, de faux positifs FP et de faux négatifs FN :

$$P = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{4}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{5}$$

Scius-Bertrand et al. [34] utilisent la même F-mesure pour comparer, sur une tâche d'extraction d'information, une chaîne classique détection-OCR-extraction à une approche end-to-end par modèle de fondation (section II.8). C'est également en précision, rappel et F-mesure que sont évalués, en section II.7, les classifieurs de style typographique de Tensmeyer, Saunders et Martinez [35] et la détection de mots en italique, gras ou tout capitales de Chaudhuri et Garain [36].

II.4.4 Recouvrement (IoU) et précision moyenne (mAP)

L'évaluation des systèmes de segmentation de mise en page et de détection d'objets, en particulier celle de Clérice [8] pour YALTAi et celle de Zhao et al. [9] pour DocLayout-YOLO (section II.6.3), repose sur l'indice de Jaccard, ou IoU (*Intersection over Union*), qui mesure le recouvrement entre une région prédite B_p et la région de vérité terrain correspondante B_g :

$$\mathrm{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{6}$$

Une détection est jugée correcte (vrai positif) si son IoU avec la région de vérité terrain dépasse un seuil fixé (typiquement 0,5). La précision moyenne AP pour une classe donnée est alors définie comme l'aire sous la courbe précision-rappel obtenue en faisant varier le seuil de confiance des détections :

$$\mathrm{AP} = \int_0^1 P(R)\, dR \tag{7}$$

et le mAP (*mean Average Precision*) rapporté par Clérice [8] (tableau 9) correspond à la moyenne des AP sur l'ensemble des K classes de zones considérées :

$$\mathrm{mAP} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{AP}_k \tag{8}$$

II.5 Lexicographie numérique et structuration de dictionnaires anciens

II.5.1 Défis de structuration des dictionnaires anciens

Les dictionnaires encyclopédiques anciens posent un défi spécifique de structuration : leurs entrées suivent une organisation lexicale hiérarchique (vedette, sous-entrées, exemples, citations) qui n'est pas toujours matérialisée de façon univoque par la mise en page. Galleron et Williams [1], à propos du *Dictionnaire universel* de Basnage de Beauval, identifient deux difficultés récurrentes : une typographie non normalisée d'une édition à l'autre, et une frontière entre deux entrées successives qui n'est pas toujours marquée par un signal typographique constant ; ils concluent que la segmentation automatique des entrées reste, à l'échelle d'un dictionnaire complet, une tâche largement non résolue.

Cette même difficulté se retrouve chez Lang [38], qui encode en TEI-XML deux dictionnaires alchimiques anciens (*Lexicon Alchemiae* de Ruland, 1612, et *Lexicon pharmaceutico-chymicum* de Sommerhoff, 1701, soit près de 20 000 entrées au total), avec

un CER de seulement 0,80 % obtenu via le modèle NOSCEMUS GM4 sur Transkribus (équation 2).

Un OCR de bonne qualité ne suffit cependant pas à lui seul à résoudre la structuration : l'auteure a dû recourir à un repérage de marqueurs textuels récurrents par expressions régulières, suivi d'une correction manuelle, pour segmenter les entrées longues du *Lexicon Alchemiae*. Le langage cryptique de l'alchimie, fondé sur des *Decknamen* dépendants du contexte, illustre qu'une segmentation purement lexicale ne suffit pas sans un minimum d'interprétation sémantique, un parallèle avec les renvois et abréviations contextuels du Dictionnaire de Trévoux.

II.5.2 Jalons de la structuration numérique : d'ARTFL à TEI Lex-0

Cette question de la structuration numérique de dictionnaires remonte aux tout premiers projets de corpus textuels numérisés en humanités numériques.
Le projet ARTFL, qui numérise dès le début des années 1990 l'*Encyclopédie* de Diderot et d'Alembert (1751–1772) pour l'université de Chicago, en est l'un des jalons fondateurs : Morrissey [12] y décrit les choix de balisage retenus pour rendre interrogeable un ouvrage de plus de 74 000 articles rédigés par plus de 140 auteurs, une problématique d'échelle comparable à celle du Trévoux et de la Grande Encyclopédie traités dans ce stage. Quelques années plus tard, Rennie [14] documente l'encodage en TEI du *Scottish National Dictionary*, l'un des premiers cas publiés d'application de la TEI à un dictionnaire historique complet, en insistant déjà sur la difficulté à faire correspondre les niveaux hiérarchiques du dictionnaire imprimé (entrée, sous-entrée, citation) à la structure arborescente du XML. Cet effort de normalisation aboutit, deux décennies plus tard, au sous-schéma **TEI Lex-0**, appliqué concrètement par Salgado et al. [22] au *Dictionnaire de l'Académie des sciences de Lisbonne*, puis à grande échelle par Hagen et al. [23] à vingt-deux encyclopédies allemandes et françaises.
Ce jalon est particulièrement pertinent ici puisque TEI Lex-0 constitue justement le format de sortie visé, à terme, pour la structuration du Dictionnaire de Trévoux. Le tableau 3 situe ces différents projets les uns par rapport aux autres.

Projet	Réf.	Année	Corpus	Échelle
ARTFL	[12]	1996	Encyclopédie (Diderot & d'Alembert)	∼74 000 articles
Scottish National Dict.	[14]	2000	Scottish National Dictionary	–
Lexiques alchimiques	[38]	2025	Ruland / Sommerhoff	∼20 000 entrées
TEI Lex-0 (Lisbonne)	[22]	2019	Dict. Académie des sciences	–
22 encyclopédies TEI	[23]	2020	Encyclopédies allemandes/françaises	22 ouvrages
Trévoux / Grande Encyclopédie	[ce stage]	2026	DUFLT (1704–1771) / LGE (1886–1902)	*en cours*

TABLE 3 – Positionnement du corpus d'étude par rapport aux principaux projets antérieurs de structuration numérique de dictionnaires et d'encyclopédies.

Repères historiques :

Avant les moteurs neuronaux comparés ci-après, la reconnaissance de caractères s'est construite en plusieurs vagues successives.

LeCun et al. [13] posent, dès 1998, les bases de l'apprentissage supervisé de bout en bout pour la reconnaissance de documents avec le réseau convolutif LeNet, entraîné directement sur des images de caractères plutôt que sur des primitives dessinées à la main. Plamondon et Srihari [15] dressent en 2000 un état des lieux complet des approches d'écriture manuscrite en ligne et hors ligne, distinguant déjà les approches statistiques (modèles de Markov cachés) des premières approches neuronales.

Marinai, Gori et Soda [16] confirment en 2005 la montée en puissance des réseaux de neurones pour l'ensemble des tâches d'analyse documentaire, de la binarisation à la reconnaissance de mise en page. Graves et Schmidhuber [18] introduisent en 2009 les réseaux récurrents multidimensionnels (MDRNN) pour l'écriture hors ligne, une architecture directement à l'origine des systèmes de reconnaissance séquentielle CTC utilisés aujourd'hui par Kraken [26,31] et par le modèle CATMuS-Print [40].

C'est cette filiation, des réseaux convolutifs de LeCun aux CRNN modernes, qui explique pourquoi les moteurs comparés ci-dessous partagent un même principe d'apprentissage bout en bout, malgré vingt-cinq ans d'écart.

II.5.3 Moteurs spécialisés contre moteurs génériques : le cas du grec polytonique

II.5.3.1 Évaluation sur le grec polytonique

Romanello, Najem-Meyer et Robertson [5] évaluent, dans le cadre du projet *Ajax Multi-Commentary*, trois pipelines d'OCR (Calamari GT4Hist, Tesseract et Kraken + Ciaconna) sur des commentaires classiques du XIX^{e} siècle mêlant grec polytonique et latin. Le tableau 4 résume leurs résultats globaux et par groupe de régions, et le tableau 5 le détail par édition commentée de l'*Ajax* de Sophocle.

Moteur	Global			Grec	Comm.	Faible grec	App. crit.	Struct.	Nombres
	51 186 (29 %)			6 657 (92 %)	23 825 (23 %)	13 322 (2 %)	2 062 (43 %)	3 371 (34 %)	693 (0 %)
	F_1	CER	WER	CER	CER	CER	CER	CER	CER
Calamari GT4Hist	0,61±0,05	0,30±0,04	0,39±0,05	0,84±0,05	0,27±0,04	0,05±0,04	0,46±0,12	0,34±0,01	0,23±0,26
Tesseract	**0,82±0,04**	**0,07±0,02**	0,16±0,03	0,13±0,05	**0,08±0,02**	**0,01±0,00**	0,12±0,01	**0,07±0,01**	**0,13±0,13**
Kraken + Ciaconna	0,82±0,02	0,08±0,02	0,18±0,02	**0,07±0,04**	0,11±0,05	0,04±0,01	**0,07±0,00**	0,07±0,02	0,13±0,17

TABLE 4 – Résultats globaux et régionaux (moyenne pondérée ± écart-type) des pipelines testés, par groupe de régions [5].

Moteur	Lobeck			Schneidewin			Campbell			Jebb			Wecklein		
	F_1	CER	NLD	F_1	CER	NLD	F_1	CER	NLD	F_1	CER	NLD	F_1	CER	NLD
Calamari GT4Hist	0,52	0,37	0,63	0,61	0,28	0,72	0,67	0,27	0,73	0,63	0,31	0,69	0,59	0,32	0,68
Tesseract	0,76	0,11	0,89	**0,82**	**0,08**	**0,92**	**0,87**	**0,05**	**0,95**	**0,80**	**0,08**	**0,92**	0,82	0,05	0,95
Kraken + Ciaconna (base)	0,75	0,11	0,89	0,68	0,17	0,83	0,83	0,07	0,93	0,78	0,12	0,88	**0,83**	**0,05**	**0,95**
Kraken + Ciaconna (réentraîné)	**0,81**	**0,09**	**0,91**	0,82	0,09	0,91	–	–	–	0,82	0,09	0,91	–	–	–

TABLE 5 – Score F_1, CER et distance de Levenshtein normalisée (NLD) par commentaire [5].

Kraken + Ciaconna domine sur les régions les plus riches en grec, texte grec pur (CER de 0,07 contre 0,13 pour Tesseract) et apparat critique (0,07 contre 0,12), mais reste derrière Tesseract sur les régions moins grécisées (commentaire mêlé, faible densité de grec, texte structuré), d'où l'avantage global de Tesseract (0,07 contre 0,08). Ce basculement selon la part de grec confirme que Kraken + Ciaconna n'est pas un moteur uniformément meilleur, mais un moteur spécialisé dont l'avantage se limite aux passages effectivement écrits en grec.
Sa stabilité d'une édition à l'autre (tableau 5) est également moindre avant réentraînement (CER de 0,05 à 0,17, contre 0,05 à 0,11 pour Tesseract), et se resserre autour de 0,09 une fois réentraîné sur le corpus, ce qui suggère qu'une partie de cette instabilité initiale tenait à un manque d'adaptation au style typographique propre à chaque édition.

II.5.3.2 Le moteur Kraken

Le moteur **Kraken**, déjà cité ici pour ses classifieurs grecs, a été conçu à l'origine par Kiessling [21] comme un outil de reconnaissance ouvert et entièrement ré-entraînable, pensé spécifiquement pour les besoins des humanités numériques plutôt que pour les documents administratifs contemporains visés par la plupart des moteurs commerciaux. Sa cinquième version, publiée par Kiessling [39], en modernise l'architecture (segmentation par réseau convolutif, décodage par recherche en faisceau) tout en conservant cette même philosophie d'un outil ré-entraînable sur des corpus réduits.
La figure 5 illustre la densité et l'hétérogénéité typographique de ce type de corpus : plusieurs éditions du même texte (Sophocle, *Ajax*), avec alternance de grec polytonique, de latin et de langues modernes de commentaire (allemand, anglais) sur une même page, une configuration qui rend l'OCR nettement plus difficile qu'un texte monolingue et monopolice.

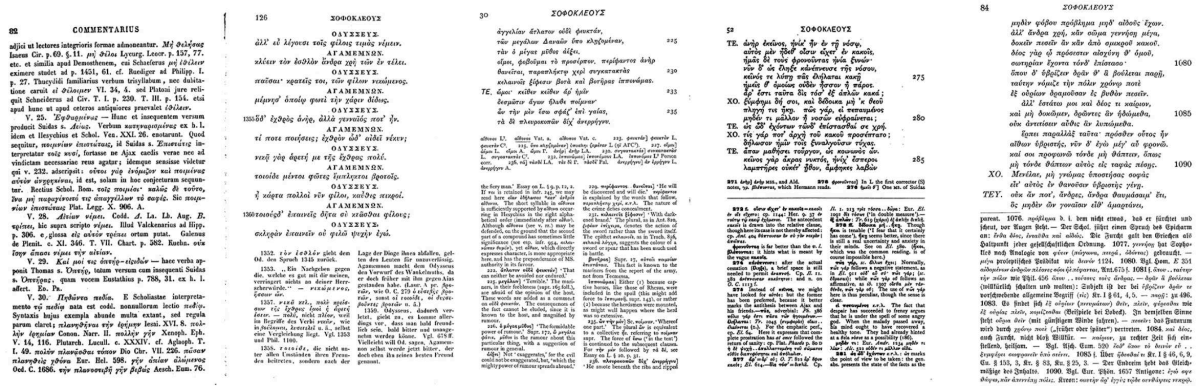


FIGURE 5 – Pages de commentaires du XIX^{e} siècle sur l'*Ajax* de Sophocle [5].

II.5.4 Comparaison de moteurs OCR sur une autre typographie historique

II.5.4.1 Résultats sur le russe imprimé ancien

Levchenko [6] conduit une comparaison similaire sur un corpus de livres russes imprimés en police civile du XVIII^{e} siècle (1 029 pages, 428 ouvrages). Les résultats obtenus par différents moteurs sur un échantillon de 100 pages, résumés au tableau 6, illustrent l'écart de performance entre moteurs génériques et moteurs spécifiquement adaptés à une typographie ancienne.

Moteur OCR	CER (%)	WER (%)
Surya (BT5)	45,96	78,33
Tesseract OCR 4.0	21,55	126,10
Transkribus PyLaia (spécialisé)	26,93	29,07
TrOCR fine-tuné sur le corpus	1,83	7,82

TABLE 6 – Résultats OCR sur un corpus de russe imprimé du XVIII^{e} siècle [6].

Ce tableau confirme, sur une typographie et une langue totalement différentes du Trévoux, un constat déjà établi par Romanello et al. : les moteurs génériques non adaptés (Surya, Tesseract) échouent nettement face à une graphie historique, tandis qu'un moteur fine-tuné sur le corpus cible (TrOCR) réduit le CER d'un ordre de grandeur, un résultat qui plaide plus généralement pour l'adaptation au domaine plutôt que pour l'usage d'un moteur générique.

II.5.4.2 Le modèle CATMuS-Print

CATMuS-Print, diffusé par Gabay et Clérice [40], est un modèle Kraken entraîné de façon diachronique et multilingue sur de l'imprimé occidental allant des premiers imprimés du XVI^{e} siècle aux documents numériques contemporains (français, latin, espagnol, allemand, anglais, italien, corse, catalan), selon des principes de transcription graphématique qui préservent les usages historiques (u/v, i/j) plutôt que de les normaliser silencieusement.

La famille CATMuS ne se limite d'ailleurs pas à l'imprimé : Clérice, Pinche, Gabay et al. [27] étendent la même philosophie de transcription graphématique diachronique et multilingue à l'écriture manuscrite avec **CATMuS Medieval**, un jeu de données et des

modèles Kraken associés couvrant plusieurs siècles et systèmes d'écriture latins pour la reconnaissance de manuscrits médiévaux.
Cette extension, publiée par une large partie de la même équipe que CATMuS-Print (dont Gabay et Clérice, déjà cités pour la version imprimée), illustre la portée plus générale de cette approche de transcription graphématique diachronique.

II.5.5 Le cas spécifique du français et du latin imprimés anciens

Les corpus évoqués jusqu'ici couvrent le grec polytonique [5], le russe imprimé [6], l'allemand en antiqua et en fraktur [7], et, plus loin dans ce chapitre, le mandchou et l'arabe manuscrit, mais aucun ne porte spécifiquement sur le français ou le latin.
Gabay, Clérice et Reul [41], avec le projet **OCR17**, entraînent des modèles de transcription (Kraken et Calamari) sur des imprimés français du XVII^{e} siècle et testent leur généralisation à d'autres siècles, dont le tableau 7 reprend les scores.

Modèle	Test	Imprimés XVI^{e}	Imprimés XVIII^{e}	Imprimés XIX^{e}
KRAKEN				
Modèle de base	97,47 %	97,74 %	**97,78 %**	94,50 %
avec réseau élargi	**97,92 %**	**98,06 %**	**97,78 %**	94,23 %
+ données artificielles	96,65 %	97,26 %	97,74 %	**95,50 %**
avec réseau élargi	97,26 %	97,68 %	97,84 %	94,84 %
CALAMARI				
Modèle de base	98,38 %	98,19 %	97,91 %	92,7 %
Modèle de base (NFC)	98,47 %	98,14 %	98,27 %	93,11 %
Un votant+PT+DA (NFC)	98,76 %	98,49 %	96,47 %	**97,05 %**
5 votants+PT+DA (NFC)	**99,05 %**	**98,68 %**	**98,78 %**	**97,05 %**

TABLE 7 – Scores des modèles Kraken et Calamari selon la période testée [41].

Le score, défini par les auteurs comme 1 moins le CER, dépasse 97 % pour Kraken et 96 % pour Calamari sur les imprimés du XVIII^{e} siècle, non vus à l'entraînement mais contemporains du Trévoux (1704 à 1771) ; Calamari conserve son avantage avec sa meilleure configuration (5 votants, 98,78 % contre 97,84 % pour Kraken).

II.5.6 Approches neuronales de bout en bout : Donut

Donut [24] (*Document understanding transformer*) supprime le module d'OCR intermédiaire : le modèle apprend directement une correspondance entre l'image du document et une séquence de texte structuré en sortie, sans étape de segmentation ni de reconnaissance de caractères distincte (encodeur visuel vers embeddings, décodeur textuel générant directement du JSON structuré). Cette approche *OCR-free* simplifie le pipeline mais reporte sur le décodeur toute la difficulté de la reconnaissance, ce qui la rend sensible à la longueur et à la complexité structurelle du document, une limite critique pour des entrées lexicographiques longues et hiérarchisées comme celles du Trévoux (figure 6).

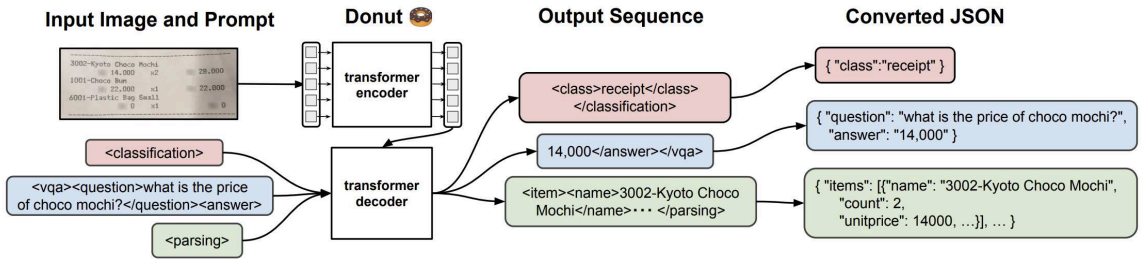


FIGURE 6 – Pipeline de Donut [24].

Agbeti-Messan et al. [7] apportent un indice chiffré de la difficulté de ce transfert vers l'écrit ancien : dans un benchmark de modèles de bout en bout sur des journaux historiques de la Bibliothèque nationale du Luxembourg (avant 1878, typographies antiqua et fraktur), ils confrontent notamment des architectures Transformer à des architectures à espaces d'états (*state-space models*) dérivées de **Mamba** [52], alternative à complexité linéaire aux Transformers pour les séquences longues, un atout potentiel pour des colonnes de dictionnaire de plusieurs centaines de caractères. Parmi les architectures comparées, **TrOCR**, encodeur-décodeur Transformer de la même famille que Donut, obtient les résultats du tableau 8.

Configuration	**CER (%)**
Sans fine-tuning, pages en antiqua	15,46
Sans fine-tuning, pages en fraktur	35,98
Après fine-tuning sur le corpus cible	3,62

TABLE 8 – Résultats de TrOCR sur des journaux historiques de la Bibliothèque nationale du Luxembourg (avant 1878) [7].

La figure 7 (extraite du même benchmark) situe ce gain sur un plan précision/latence, antiqua comme fraktur.

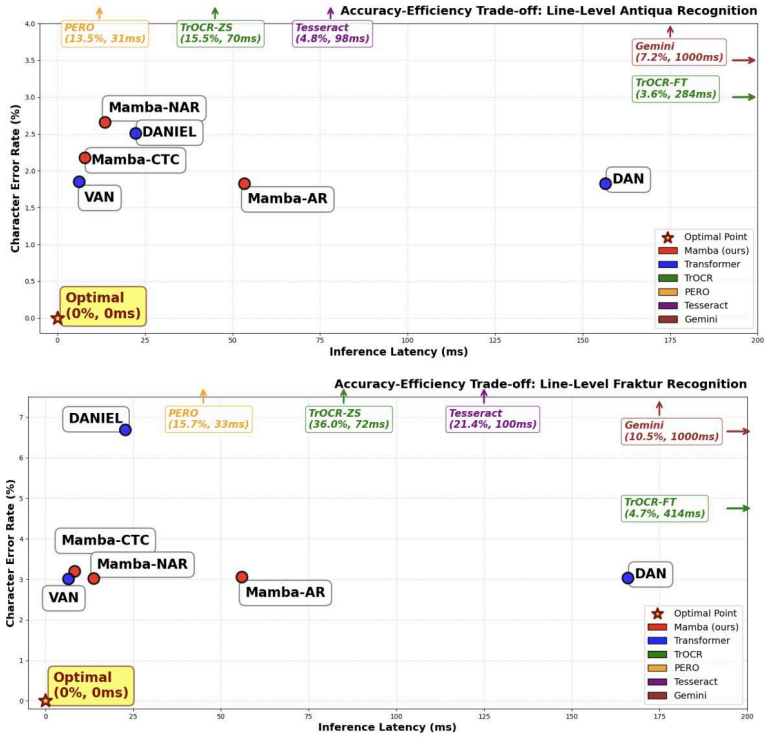


FIGURE 7 – Compromis précision/latence en reconnaissance de ligne, TrOCR zero-shot (ZS) et fine-tuné (FT), écritures antiqua (haut) et fraktur (bas) [7].

Cet écart, comparable à celui observé pour les moteurs OCR classiques (section II.5.3), suggère que les architectures de bout en bout de type Donut/TrOCR ne font pas exception : sans adaptation au domaine, leurs performances se dégradent fortement sur des typographies historiques.

II.6 Segmentation et structuration de la mise en page

Repères historiques :

L'évaluation comparative des méthodes de segmentation de mise en page est elle-même un champ de recherche ancien. Antonacopoulos et al. [17] organisent dès 2009, dans le cadre d'ICDAR, la première compétition internationale de segmentation de page comparant systématiquement plusieurs algorithmes sur un jeu de données commun et un protocole d'évaluation partagé. Les mêmes auteurs reconduisent l'exercice en 2013 sur des journaux historiques [19], un corpus dont les mises en page multi-colonnes irrégulières annoncent directement les difficultés rencontrées sur le Dictionnaire de Trévoux (colonnes désalignées, lettres marginales décalées, voir section III.3). Ces campagnes d'évaluation ont normalisé l'usage de l'IoU (équation 6) comme métrique de référence pour la segmentation, bien avant la généralisation des détecteurs d'objets fondés sur l'apprentissage profond décrits ci-après.

II.6.1 Approches multimodales pour la compréhension de documents : LayoutLMv3

LayoutLMv3 unifie, dans un unique modèle, le traitement du texte et de l'image d'une page de document, sans recourir à un détecteur d'objets pré-entraîné pour l'image. La figure 8 détaille cette architecture : les patches d'image et les tokens de texte sont projetés dans un même espace et encodés conjointement par un Transformer multimodal, entraîné avec trois objectifs de pré-entraînement combinés (masquage de texte, masquage d'image, alignement mot-patch) [25].

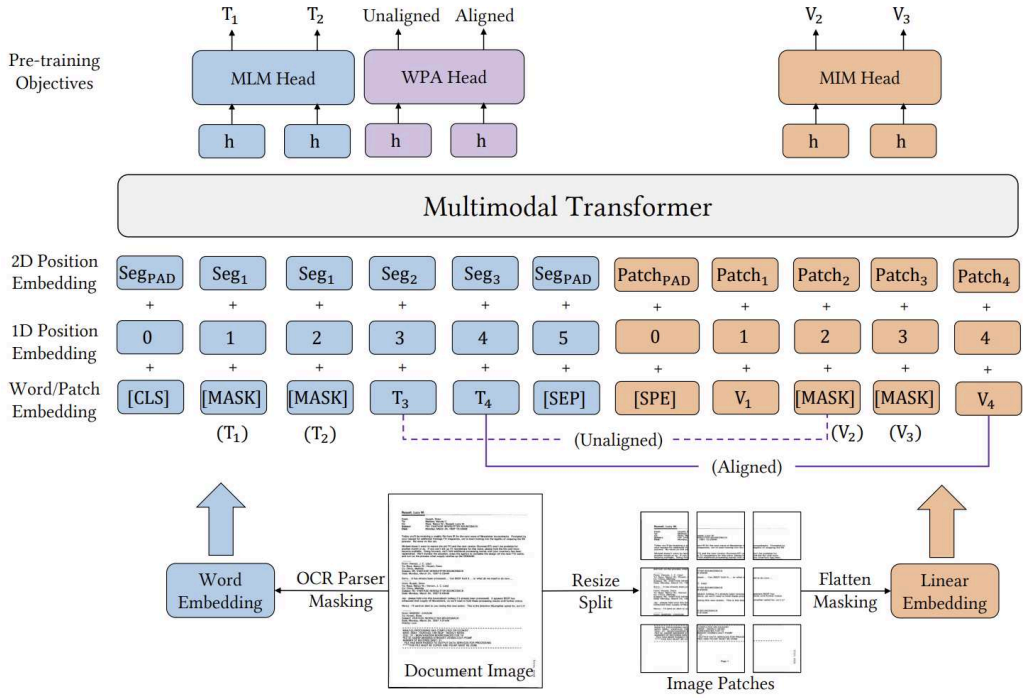


FIGURE 8 – Architecture et objectifs de pré-entraînement de LayoutLMv3 [25].

L'architecture met en oeuvre l'encodage conjoint des patches d'image et des tokens de texte par un Transformer multimodal unique, sans détecteur d'objets. Évalué sur des benchmarks de documents administratifs contemporains (formulaires FUNSD, reçus CORD, questions-réponses documentaires DocVQA, mise en page PubLayNet), il atteint des F-mesures élevées (équation 5 : F_1 d'environ 90 à 92 sur FUNSD, 96 à 97 sur CORD). Ces benchmarks concernent cependant des documents courts et standardisés ; aucune évaluation publiée à ce jour ne porte sur des documents patrimoniaux longs et hiérarchisés comme un dictionnaire ancien, pour lesquels la structure logique ne se déduit pas directement de la position spatiale des blocs.

II.6.2 La segmentation de mise en page reformulée en détection d'objets : YALTAi

Un développement récent concerne la reformulation de la segmentation de mise en page dans le moteur Kraken comme un problème de *détection d'objets*. Kraken segmente nativement la page par classification de pixels suivie d'une polygonisation, une approche

qui se dégrade fortement sur de petits corpus d'entraînement (de l'ordre de 1 000 échantillons annotés ou moins). Clérice [8] propose avec **YALTAi** de remplacer cette étape par un détecteur d'objets à rectangles isothétiques (YOLOv5), injecté dans le pipeline de Kraken. Sur un jeu de données de manuscrits et imprimés anciens (1 110 pages), le tableau 9 résume l'écart de performance observé, mesuré en mAP (équation 8).

Méthode de segmentation	**mAP global (%)**
Kraken (segmentation par pixels, natif)	6,98
YALTAi / YOLOv5x (détection d'objets)	47,75
YALTAi / YOLOv5n (détection d'objets, léger)	34,63

TABLE 9 – Comparaison Kraken natif / YALTAi sur un corpus de 1 110 pages de manuscrits et imprimés anciens [8].

Sur un corpus de documents tabulaires anciens, l'écart est encore plus net (mAP de 0,09 % pour Kraken contre 4,77 % pour YOLOv5x), ce qui suggère que, pour un corpus de taille modeste, remplacer la segmentation native de Kraken par une détection d'objets de type YALTAi peut s'avérer nettement plus robuste, tout en restant compatible avec le reste de son pipeline. La figure 9 montre des exemples de détection obtenus avec YOLOv5x sur des manuscrits enluminés et un document tabulaire, chaque catégorie de zone (illustration, lettrine, texte marginal, corps de texte, colonne) étant repérée par un rectangle coloré.

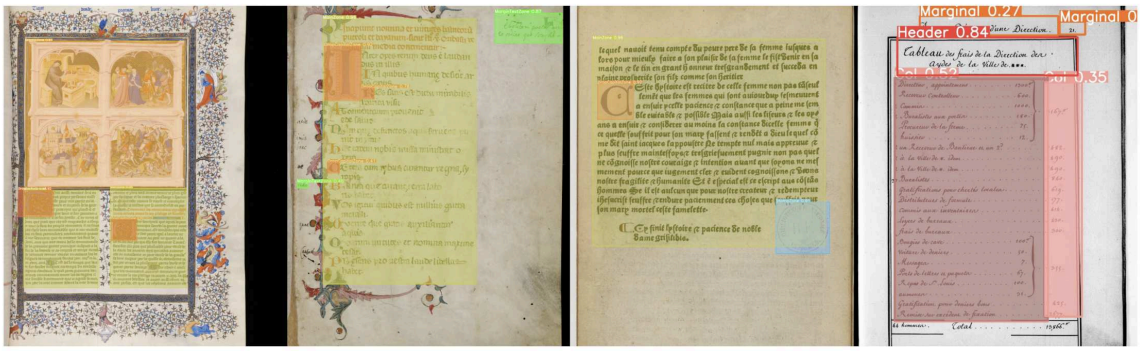


FIGURE 9 – Détections YOLOv5x sur le jeu de données Segmonto (manuscrits enluminés, trois images de gauche) et sur le jeu de données tabulaire (à droite) [8].

Les illustrations en orange, lettrines en orange foncé, texte marginal en vert, corps de texte en jaune.

II.6.3 Au-delà de YOLOv5 : détecteurs de mise en page de nouvelle génération

II.6.3.1 Génération de données synthétiques et architecture.

La piste ouverte par YALTAi, reformuler la segmentation de mise en page comme un problème de détection d'objets, a depuis été prolongée par des détecteurs plus récents. **DocLayout-YOLO** [9], proposé par Zhao, Kang, Wang et He, part du même constat que YALTAi, à savoir la difficulté à réunir suffisamment de données annotées pour entraîner

un détecteur de mise en page robuste, mais y répond par la génération de données synthétiques : les auteurs formulent la composition d'une page de document comme un problème d'empaquetage bidimensionnel (*bin packing*) pour synthétiser automatiquement un grand jeu d'entraînement, DocSynth-300K, puis affinent l'architecture de détection elle-même par un module d'attention « global-to-local » capable de mieux gérer les échelles très variables des éléments d'une page (un titre courant tenant sur une ligne, un tableau occupant une demi-page).

II.6.3.2 Résultats et comparaison

Sur leur benchmark DocStructBench, les auteurs montrent que DocLayout-YOLO conserve la rapidité des détecteurs purement visuels de la famille YOLO tout en approchant la précision des approches multimodales plus coûteuses, un compromis vitesse/précision qui fait écho au débat entre pipelines explicites et modèles de fondation traité plus loin (section II.8) autour d'un modèle de vision-langage généraliste, Qwen2.5-VL [28].

II.6.4 Structuration logique de documents patrimoniaux

II.6.4.1 Comparaison entre système à base de règles et apprentissage automatique.

Au-delà de la reconnaissance de caractères, distinguer titre, en-tête, corps de texte et notes reste un problème largement ouvert. Sur un corpus de journaux historiques français de Franche-Comté (XX^{e} siècle, jusqu'à 48 documents et plus de 51 000 lignes annotées, évalué en précision, rappel et F-mesure, équations 4 et 5), Gutehrlé et Atanassova [33] comparent un système à base de règles conçu par observation du corpus à deux modèles d'apprentissage automatique classiques (RIPPER, Gradient Boosting) entraînés sur des traits géométriques, morphologiques et textuels. La figure 10 détaille les étapes du système à base de règles, depuis les tags TextBlock et TextLine du document ALTO d'entrée jusqu'à la résolution des conflits d'annotation.

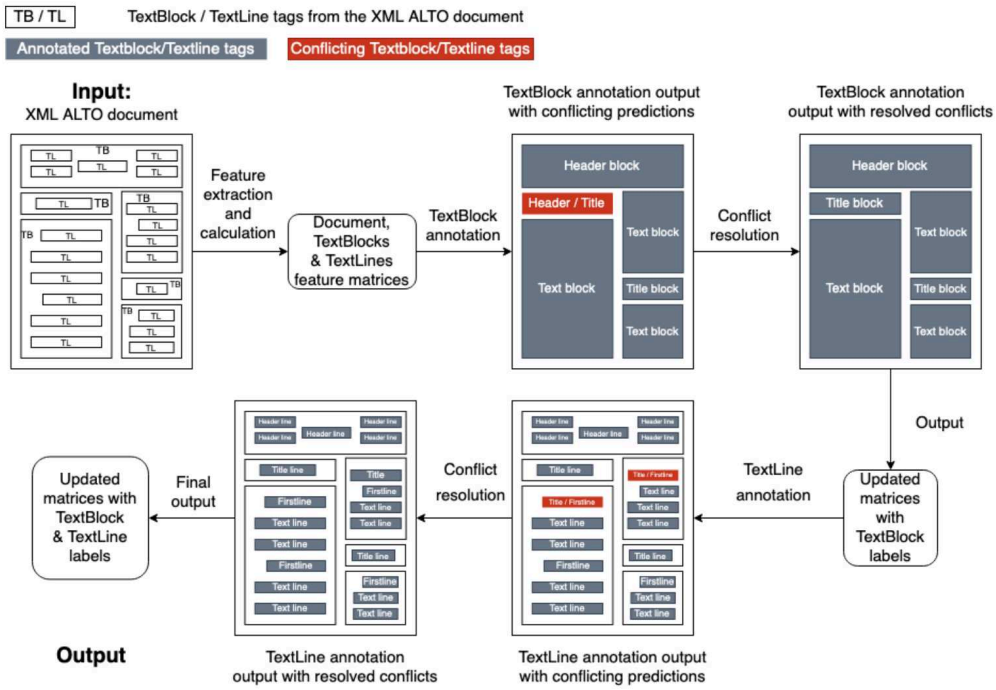


FIGURE 10 – Pipeline du système à base de règles [33].

Leurs résultats montrent que le système à base de règles surpasse les deux modèles d'apprentissage automatique dans la plupart des configurations testées, notamment en rappel, ce qui suggère qu'une connaissance experte du domaine reste précieuse pour amorcer la structuration logique de documents patrimoniaux avant un passage à l'échelle par apprentissage profond — une piste pertinente pour le Trévoux, dont la mise en page suit des conventions typographiques propres au genre lexicographique.

II.7 Style typographique et identification de script

Au-delà de la seule reconnaissance des caractères, le style typographique, romain contre italique, casse normale contre petites ou grandes capitales, porte dans un dictionnaire ancien comme le Trévoux une information structurelle à part entière : c'est souvent l'italique ou les petites capitales, plutôt qu'un marqueur explicite, qui distinguent la vedette d'une entrée, un exemple d'usage ou une citation en langue étrangère. La littérature aborde ce problème selon deux angles complémentaires : la classification du style d'un mot déjà segmenté, et l'identification du script (l'alphabet ou système d'écriture) employé, un problème voisin mais distinct dès lors qu'un document patrimonial mêle plusieurs écritures, comme c'est précisément le cas du Trévoux avec ses insertions grecques et hébraïques.

II.7.1 Classification du style typographique par traits géométriques et par apprentissage profond

II.7.1.1 Approche par traits géométriques explicites

Chaudhuri et Garain [36] proposent une méthode précoce, fondée sur des traits géométriques explicites plutôt que sur de l'apprentissage supervisé, pour détecter

automatiquement les mots en italique, en gras ou tout en capitales dans une image de document : inclinaison des contours de caractères pour l'italique, épaisseur du trait pour le gras, hauteur relative des caractères pour les tout capitales. Cette approche par traits faits main, si elle reste interprétable, est sensible à la qualité de numérisation et à la segmentation préalable des mots, deux conditions rarement réunies sur un corpus comme le Trévoux (voir les défis typographiques documentés en section III.3).

II.7.1.2 Approche par apprentissage supervisé

Tensmeyer, Saunders et Martinez [35] proposent une alternative fondée sur un réseau convolutif entraîné de façon supervisée à classer directement l'image d'un mot ou d'une ligne selon sa police et son style, sans extraction de traits manuels, évalué en précision, rappel et F-mesure (équations 4 et 5) sur plusieurs jeux de polices historiques et contemporaines ; les auteurs montrent que cette approche généralise mieux qu'une méthode à traits géométriques fixes à des polices non vues à l'entraînement, un résultat qui plaide pour un classifieur convolutif entraîné plutôt que pour une détection par règles à traits fixes, en particulier pour un dictionnaire comme le Trévoux où se mêlent styles *normal*, *italique*, petites et grandes capitales, ainsi que des insertions en grec et en hébreu.

II.7.2 Identification de script en contexte multilingue

Le repérage du grec et de l'hébreu au sein d'entrées par ailleurs françaises ou latines relève, plus spécifiquement, de l'identification de script plutôt que de la seule classification de style au sens de romain/italique/capitales.

Yang, Eli, Aysa et Ubul [43] dressent, dans une synthèse récente, un panorama des méthodes d'identification de script en contexte multilingue, depuis les approches par traits géométriques (densité de traits, diacritiques, direction dominante) jusqu'aux réseaux convolutifs et architectures Transformer, et soulignent que les documents mêlant plusieurs systèmes d'écriture au sein d'une même page, comme c'est justement le cas du Trévoux avec ses insertions grecques et hébraïques au sein d'entrées françaises ou latines, restent un cas plus difficile que l'identification de script au niveau document entier, la plupart des jeux de données publics de référence portant sur des documents monoscript.

Cette difficulté, propre aux écritures rares insérées dans un texte par ailleurs latin ou français, lorsque les données annotées manquent pour entraîner un identificateur de script dédié, motive une simplification pragmatique fréquente dans la littérature : traiter ces écritures comme des classes supplémentaires d'un même classifieur de style par mot plutôt que comme un problème d'identification de script à part entière, une piste également suivie par le fine-tuning de Qwen2.5-VL [28](section II.8.4).

Ces travaux (du II.7) montrent que l'identification du style et du script peut compléter une chaîne OCR classique. Une autre possibilité consiste toutefois à confier simultanément la reconnaissance, l'interprétation de la mise en page et, dans certains cas, la structuration du document à un modèle de vision-langage. Nous abordons leur apport dans la section suivante (II.8)

II.8 Modèles de fondation contre pipelines explicites sur documents réels

II.8.1 Pipelines explicites contre modèles de fondation

Scius-Bertrand et al. [34] comparent, sur une tâche d'extraction d'information (photographies d'étiquettes alimentaires), évaluée en F-mesure (équation 5), une chaîne classique (détection de texte, OCR, puis extraction structurée) à une approche end-to-end interrogeant directement un modèle de fondation multimodal sur l'image brute. Leurs résultats nuancent l'idée d'une obsolescence complète des étapes intermédiaires : les modèles de fondation obtiennent de bons résultats sur des documents à structure simple, mais les pipelines explicites conservent un avantage sur des documents plus complexes.

II.8.2 Correction post-OCR multimodale par VLM

Ce constat est nuancé par des travaux plus récents. Greif, Griesshaber et Greif [32] évaluent des modèles de vision-langage multimodaux (GPT-4o, Gemini 2.0 Flash) sur des répertoires urbains allemands imprimés entre 1754 et 1870, en écritures Fraktur et Antiqua mêlées, et les comparent à des moteurs spécialisés (Tesseract, Transkribus Print M1, TrOCR). Surtout, les auteurs testent une correction post-OCR multimodale inédite, où l'image source *et* la transcription bruitée sont fournies conjointement au LLM correcteur (figure 11), avec les résultats résumés au tableau 10.

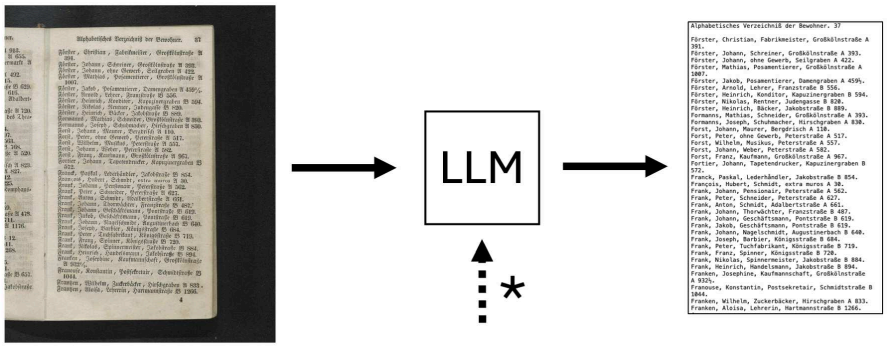


FIGURE 11 – Correction post-OCR multimodale [32].

L'image source et la transcription Transkribus sont fournies conjointement au LLM, qui renvoie un texte corrigé.

Configuration	CER normalisé (%)
Transkribus Print M1 (spécialisé, sans correction)	3,67
Gemini 2.0 Flash (OCR direct, sans fine-tuning)	1,27
Transkribus Print M1 + correction multimodale Gemini 2.0 Flash	0,84

TABLE 10 – CER normalisé sur des répertoires urbains allemands (1754–1870), avant et après correction post-OCR multimodale [32].

Ce résultat, positif contrairement à la plupart des tentatives de correction textuelle seule (voir section II.9), suggère que l'apport d'un LLM correcteur dépend fortement de

l'accès à l'image source et non au seul texte bruité, une question qui illustre plus largement le débat, encore non tranché dans la littérature, entre approches end-to-end par modèle de fondation et pipelines modulaires plus classiques.

II.8.3 Quatre modèles de fondation multimodaux pour l'OCR

Quatre modèles de vision-langage récents, **Mistral OCR** [29], **Qwen2.5-VL** [28], **Pixtral** [31] et **GLM-OCR** [30], prolongent directement cette famille de modèles de fondation multimodaux évaluée ici par Scius-Bertrand et al. et par Greif et al. :

- **Qwen2.5-VL** [28], décrit par l'équipe Qwen d'Alibaba, associe un encodeur visuel à résolution native et un décodeur de langage capable de restituer conjointement le texte reconnu et sa localisation dans la page.
- **Mistral OCR** [29] est un service de reconnaissance de documents dédié, optimisé pour préserver la mise en page d'origine (y compris tableaux et formules) dans sa sortie structurée.
- **Pixtral** [31], également développé par Mistral AI mais diffusé comme modèle ouvert plutôt que comme service dédié, partage avec Mistral OCR l'encodeur visuel mais vise un usage multimodal généraliste dont l'OCR n'est qu'une capacité parmi d'autres.
- **GLM-OCR** [30], enfin, développé par Zhipu AI, illustre la tendance inverse, celle d'un modèle compact (moins d'un milliard de paramètres) entraîné spécifiquement pour l'OCR plutôt que dérivé d'un VLM généraliste.

Ces quatre positionnements différents du même compromis entre reconnaissance de caractères et interprétation de la mise en page recoupent celui déjà observé chez Donut et LayoutLMv3 présentés plus haut.

II.8.4 La famille Qwen-VL et l'usage de VLM généralistes pour la reconnaissance de texte

II.8.4.1 La lecture de texte comme capacité fondatrice

L'intérêt de Qwen2.5-VL [28] pour la reconnaissance de texte s'appuie sur une lignée de travaux spécifiquement construite autour de la lecture de texte dans l'image. Dès la première version de la famille, Qwen-VL [26] inscrit la *lecture de texte* (*text reading*) parmi ses capacités fondatrices de « *grounding* », aux côtés de la localisation d'objets, plutôt que comme une capacité annexe d'un modèle de description d'image généraliste, un positionnement que conservent et approfondissent ses versions suivantes jusqu'à Qwen2.5-VL [28]. Cette orientation explique pourquoi la famille Qwen-VL figure systématiquement parmi les modèles de référence des benchmarks d'OCR par modèles multimodaux, tel OCRBench v2 [44], qui évalue 22 modèles de fondation sur des tâches de localisation et de raisonnement sur du texte visuel.

II.8.4.2 Fine-tuning pour une langue à faibles ressources : le mandchou.

Chung et Choi [45] fine-tunent plusieurs VLM open-source, dont Qwen2.5-VL-3B et Qwen2.5-VL-7B, comme systèmes OCR dédiés pour le mandchou, une langue à faibles ressources s'écrivant dans un script vertical spécifique : le fine-tuning fait passer la reconnaissance de documents manuscrits réels de résultats médiocres en zero-shot à une exactitude de 93,1 % au niveau mot (tableau 11).

II.8.4.3 Qwen2.5-VL comme correcteur post-OCR : le cas de l'arabe manuscrit.

Farazi et al. [46] évaluent quant à eux Qwen2.5-VL-7B *sans fine-tuning* sur un corpus de manuscrits arabes anciens (Muharaf), et obtiennent un résultat plus nuancé, également résumé au tableau 11 : au niveau ligne, Qwen2.5-VL n'est pas statistiquement distinguable de Tesseract seul, mais utiliser Qwen2.5-VL comme **correcteur post-OCR** de la sortie de Tesseract réduit nettement le CER (gain relatif de 12,4 %), avant de se dégrader sur les écritures manuscrites les plus irrégulières.

Étude	Configuration	Résultat
Chung et Choi [45]	Fine-tuning OCR, mandchou (Qwen2.5-VL-3B/7B)	Exactitude mot : 93,1 %
Farazi et al. [46]	Tesseract seul, arabe manuscrit (Muharaf)	CER : 69,2 %
Farazi et al. [46]	Qwen2.5-VL-7B seul, sans fine-tuning	CER : 70,4 %
Farazi et al. [46]	Qwen2.5-VL-7B correcteur post-OCR de Tesseract	CER : 60,6 %

TABLE 11 – Qwen2.5-VL sur deux tâches de reconnaissance de texte historique à faibles ressources, d'après Chung et Choi [45] et Farazi et al. [46].

Ce constat, un VLM généraliste non spécialisé peu compétitif en reconnaissance directe mais utile en correcteur du texte d'un moteur spécialisé, fait directement écho au débat sur la correction post-OCR par LLM traité en section II.9.

Pour le Trévoux, ces résultats suggèrent donc un compromis entre deux stratégies : une chaîne modulaire, permettant d'adapter séparément la segmentation et la reconnaissance au corpus, et une approche multimodale de bout en bout, potentiellement plus simple mais plus dépendante des capacités du modèle et de son adaptation au domaine. Cette opposition constitue le point de départ des expérimentations présentées au chapitre III.

II.9 Correction post-OCR par LLM : un apport encore incertain

François, Eglin et Biou [47] proposent une méthode de détection de texte et de correction post-OCR appliquée à des documents d'ingénierie, associant un module de détection de zones de texte à un module de correction découplé du moteur de reconnaissance utilisé en amont. Ce découplage permet de traiter la correction comme un problème de post-traitement générique, potentiellement transposable à d'autres moteurs OCR et à d'autres types de documents, dont les dictionnaires anciens. L'essor des grands modèles de langage a naturellement conduit à explorer leur usage pour cette même tâche de correction post-OCR. La littérature récente sur ce point est cependant loin d'être unanime, comme le résume le tableau 12, où les résultats sont systématiquement rapportés en CER (équation 2).

Étude	Corpus historique	Constat principal
Boros et al. [10]	8 corpus patrimoniaux (journaux, commentaires classiques grec/latin, manuscrits byzantins, radio)	14 LLM testés (GPT, BLOOM(Z), OPT, LLaMA) ; la correction **dégrade** la transcription dans la grande majorité des cas, quel que soit le prompt
Thomas, Gaizauskas & Lu [11]	BLN600, 600 articles de presse britannique du XIX^{e} siècle	LLaMA 2 (13B) instruction-tuné réduit le CER de 54,51 %, contre 23,30 % pour BART finetuné en seq2seq
Levchenko [6]	Russe imprimé du XVIII^{e} siècle	La correction (image + texte) n'améliore pas le résultat : les modèles re-décodent l'image plutôt que de corriger le texte fourni ; « sur-historicisation » (insertion de caractères archaïques)

TABLE 12 – Synthèse des constats contradictoires sur la correction post-OCR par LLM appliquée à des corpus historiques.

Ces trois études, portant sur des corpus, des langues et des protocoles différents, convergent vers un même constat de prudence : la correction post-OCR par LLM reste, en 2024-2025, une question de recherche non stabilisée, dont l'efficacité dépend fortement du protocole (zero/few-shot contre instruction-tuning, texte seul contre texte et image conjoints) et du corpus.

II.10 Application au Dictionnaire de Trévoux

Moncla, Joliveau et Vigier [48] proposent, dans le cadre du projet OLR-LLM, une étude interdisciplinaire portant sur le traitement du contenu géographique du *Dictionnaire Universel François-Latin* de Trévoux (ainsi que de *La Grande Encyclopédie*). Ce travail, mené au sein même du laboratoire LIRIS présenté au chapitre 1, plus précisément par l'un des deux tuteurs laboratoire du présent stage, souligne que la structuration de ce type de corpus suppose une chaîne complète allant de la reconnaissance optique de caractères jusqu'à l'exploitation du contenu reconnu, et constitue un précédent méthodologique et institutionnel direct pour le présent stage, qui porte sur le même corpus.

II.11 Conclusion

Cette étude bibliographique confirme, à chaque étape de la chaîne de traitement, qu'une adaptation au domaine patrimonial reste nécessaire : les moteurs OCR génériques échouent face aux typographies anciennes, la segmentation par classification de pixels se dégrade sur les petits corpus annotés au profit d'une reformulation en détection d'objets (section II.6), et le style typographique porte une information structurelle à part entière

(section II.7).
Le débat entre pipelines explicites et modèles de fondation reste ouvert, de même que l'apport réel de la correction post-OCR par LLM, dont les résultats publiés sont contradictoires selon le protocole (section II.9). Les travaux menés au LIRIS même sur le Trévoux et l'Encyclopédie constituent enfin un précédent direct pour ce stage, qui se positionne à l'articulation de l'ensemble de ces techniques, comme le montre le chapitre suivant à travers les premières approches testées.

Chapitre III

Reconnaissance et structuration du Trévoux : mise en œuvre et résultats

III Reconnaissance et structuration du Trévoux : mise en œuvre et résultats

III.1 Introduction

Ce chapitre présente, à titre exploratoire, les différentes approches concrètement mises en œuvre et testées durant le stage sur le Dictionnaire de Trévoux (édition 1743). Chaque méthode décrite ci-dessous a été essayée et évaluée séparément, l'ensemble des approches restant à comparer sur le corpus complet, à l'aide des mesures formalisées en section II.4 (CER, WER, IoU, mAP).
Le corpus de travail est constitué de pages numérisées du *Dictionnaire Universel François-Latin* (édition 1743, environ 956 pages, mise en page à deux colonnes), présentant une typographie caractéristique du XVIII^{e} siècle (s long, ligatures, caractères multilingues, français, latin, grec, hébreu).

III.2 Protocole expérimental et cadre de validation

Avant de détailler chacune des approches testées, il est utile de préciser le cadre expérimental commun dans lequel elles s'inscrivent : quelles comparaisons ont été effectivement menées, sur quels échantillons du corpus, par rapport à quelles lignes de référence, et dans quelles conditions de reproductibilité et de ressources de calcul. Ce cadre reste volontairement hétérogène d'une expérience à l'autre, à ce stade exploratoire du stage.

III.2.1 Plan d'expériences et périmètre de comparaison

Trois familles de comparaisons structurent ce chapitre :
— la première oppose, sur un même échantillon de pages, quatre modèles de vision-langage utilisés en bout en bout (Mistral OCR, Pixtral-12B, GLM-OCR, Qwen2.5-VL 7B), avec et sans correction post-OCR par LLM (section III.4.1) ;
— la deuxième oppose l'approche modulaire (Kraken + CATMuS-Print) à des moteurs OCR traditionnels génériques (EasyOCR, Surya, Tesseract), ainsi qu'à un détecteur de mise en page spécialisé (DocLayout-YOLO) comparé à Qwen2.5-VL utilisé comme détecteur (section III.5.2) ;
— la troisième, transversale aux deux précédentes, évalue le pipeline de détection de mots et de classification de style (section III.7) indépendamment de tout moteur de transcription, d'abord sur le jeu de validation issu des pages annotées, puis sur des pages du corpus entièrement inédites.

Ces trois familles de comparaisons ne sont pas encore unifiées en un plan d'expériences unique : elles portent sur des échantillons de tailles et de natures différentes (voir section III.2.2), un choix délibéré à ce stade exploratoire du stage, qui privilégie l'exploration large de plusieurs pistes à la validation approfondie d'une seule.

III.2.2 Jeux de données et échantillons

Le tableau 13 récapitule les différents échantillons du corpus mobilisés selon les expériences, dont la taille varie sensiblement d'un test à l'autre en fonction du coût d'annotation manuelle qu'il suppose.

Expérience	Taille	Nature de l'échantillon
Comparaison des VLM (section III.4.1)	150 pages	Échantillon de test annoté, transcription de référence manuelle
Comparaison DocLayout-YOLO / Qwen2.5-VL (section III.5.2)	170 pages (130 train / 40 test)	Échantillon annoté en zones de mise en page (mot-vedette, corps de texte, citation, tableau ...)
Fine-tuning CATMuS-Print (section III.5.5)	12 à 45 pages (train) 140 pages (test)	Sous-ensembles emboîtés d'un jeu de pages annotées en transcription ligne à ligne
Analyse qualitative de la segmentation (section III.5.6)	50 pages train, 30 pages test (20 pages standards, 10 pages avec courbure)	Pages annotées en polygones de ligne, réparties selon la présence ou non d'une forte courbure de reliure
Classifieur de style, SmallCNN et Qwen2.5-VL LoRA (section III.7)	100 pages annotées	Annotations Label Studio (boîtes et étiquette de style), complétées par minage automatique et synthèse pour les classes rares

TABLE 13 – Échantillons du corpus du Trévoux (956 pages au total) mobilisés selon les expériences du chapitre 3.

III.2.3 Baselines et lignes de référence

Les comparaisons menées dans ce chapitre s'appuient sur le corpus source et sur les vérités de terrain constituées pendant le stage.

- **Corpus source** : pages numérisées du *Dictionnaire de Trévoux* (édition 1743), environ 956 pages, mise en page à deux colonnes.
- **Vérité de terrain** :
  - pages transcrites manuellement, selon des principes diplomatiques.
  - fichiers `.txt` contenant la transcription en texte brut, au fil du texte (sans structuration ni coordonnées).
  - fichiers `.xml` (TEI) encodant le texte et la structure de l'article.

III.2.4 Budget et ressources de calcul

Les solutions testées pendant le stage sont des solutions locales : modèles/moteurs de Transcription, segmentation, ... ont tous été mis en œuvre sur un poste de travail local. Ce choix reflète une priorité donnée, tout au long du stage, à des solutions performantes mais à coût minimal.

III.3 Défis typographiques et corpus de travail

Avant de présenter les approches testées, il est utile d'illustrer les difficultés typographiques concrètes rencontrées sur le corpus du Trévoux, qui motivent une partie

des choix méthodologiques décrits dans la suite du chapitre.

III.3.1 Le « s long » et les confusions de caractères

Le « s long », utilisé systématiquement en position non finale dans la typographie du XVIII[e] siècle, constitue une source majeure d'ambiguïté pour les moteurs de reconnaissance : sa forme graphique proche à la fois du *f* et du *s* rond entraîne des confusions récurrentes. La figure 12 illustre ce phénomène sur un extrait du corpus, où plusieurs occurrences du « s long » sont susceptibles d'être reconnues indifféremment comme /, *s* ou *f* selon le moteur utilisé.

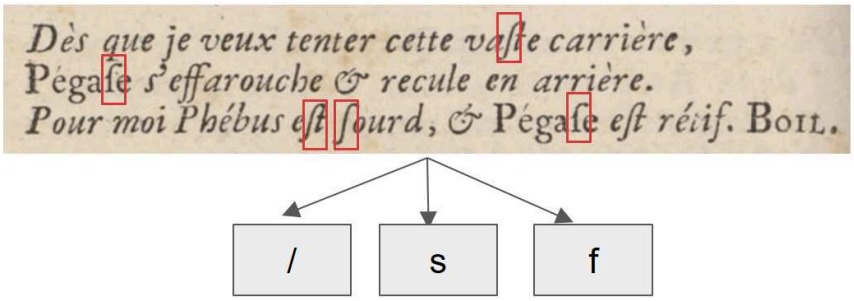


FIGURE 12 – Arbre de confusion du « s long » sur un extrait du corpus [49].

III.3.2 Ligatures et variantes orthographiques

Le corpus présente également des ligatures et variantes orthographiques propres à l'époque, par exemple la ligature « œ » alternant avec la séquence « oe » selon les occurrences d'un même mot, comme illustré à la figure 13 sur l'entrée *recœnare*. Cette variation graphique, si elle n'est pas normalisée en amont, peut conduire à considérer à tort deux graphies d'un même mot comme des erreurs de reconnaissance.

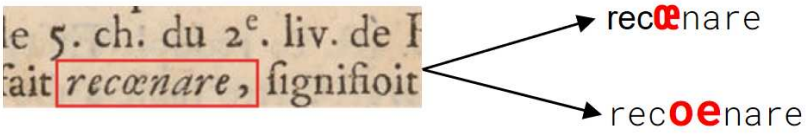


FIGURE 13 – Variante graphique « œ » / « oe » sur le mot *recœnare* [49].

III.3.3 Insertions multilingues

Le Dictionnaire de Trévoux, à vocation encyclopédique, insère fréquemment des mots ou citations en grec ancien et en hébreu au sein des entrées étymologiques, en plus du français et du latin. La reconnaissance de ces segments multilingues pose un défi spécifique, dans la mesure où les moteurs testés jusqu'ici sont essentiellement entraînés ou évalués sur des jeux de données monolingues (français ou latin). Les figures 14 à 16 illustrent plusieurs cas rencontrés dans le corpus : translittération de noms propres hébraïques, insertions grecques et hébraïques au sein d'une même entrée, et exemples grecs isolés.

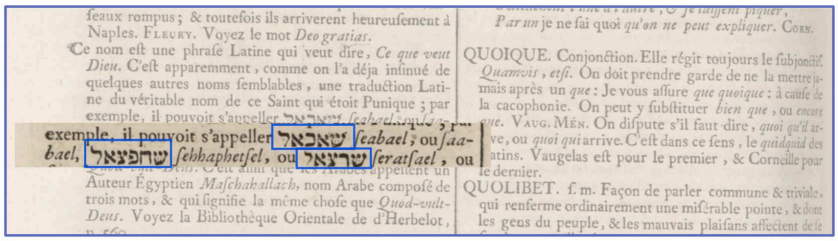


FIGURE 14 – Exemple de translittération de noms propres hébraïques [49].

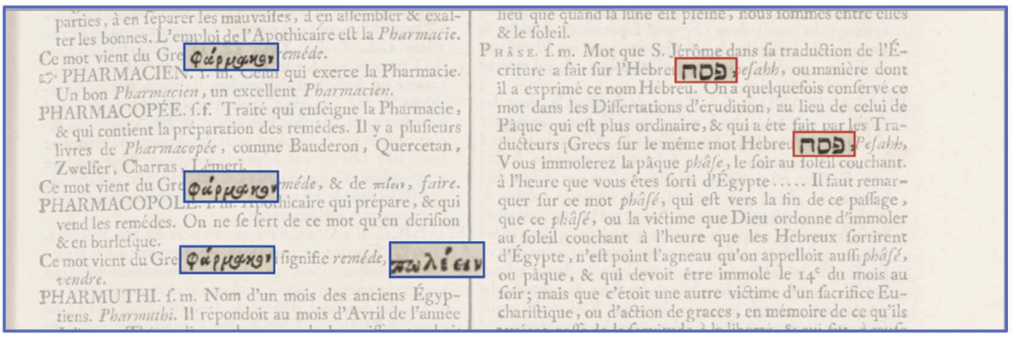


FIGURE 15 – Insertions grecques et hébraïques au sein d'entrées du dictionnaire [49].

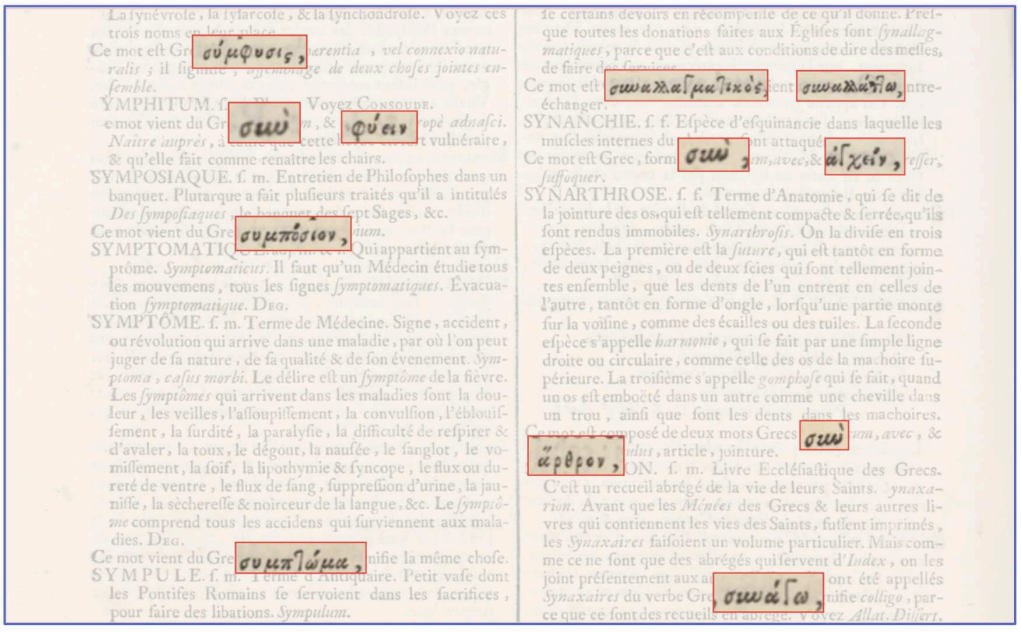


FIGURE 16 – Bloc d'exemples grecs rencontrés dans le corpus [49].

III.3.4 Qualité de numérisation et sources d'erreur récurrentes

Enfin, le corpus présente des conditions de numérisation hétérogènes (pages courbées par la reliure, bords rognés, papier jauni), illustrées aux figures 17 et 18, ainsi que plusieurs sources d'erreur récurrentes rencontrées lors de la transcription : encre effacée en fin de ligne (figure 19), première lettre d'entrée manquante, source d'erreur d'indexation

(figure 20), et lettre marginale de section décalée par rapport à l'entrée qu'elle annonce (figure 21).

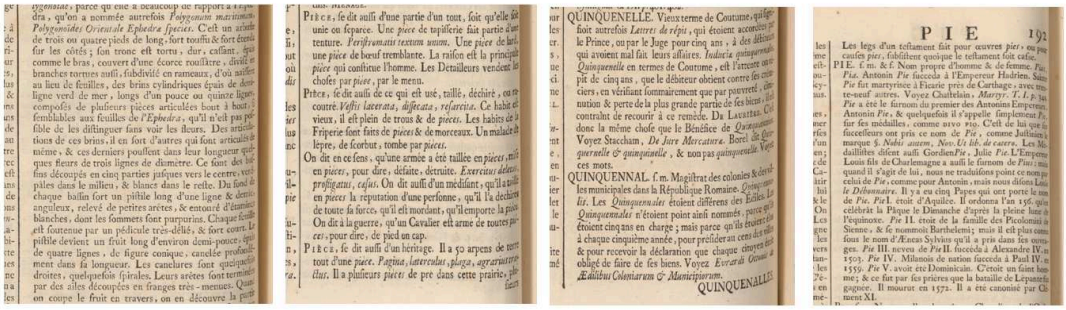


FIGURE 17 – Exemples de pages présentant une forte courbure de reliure, déformant le texte vers la gouttière [49].

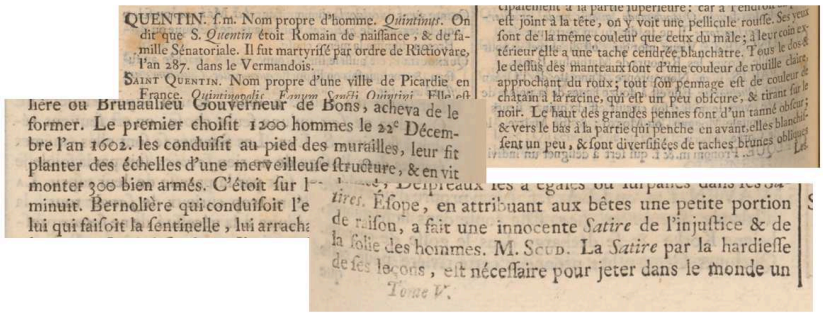


FIGURE 18 – Exemples de zones de mauvaise qualité sur certaines pages du corpus (encre pâle, papier taché ou jauni) [49].

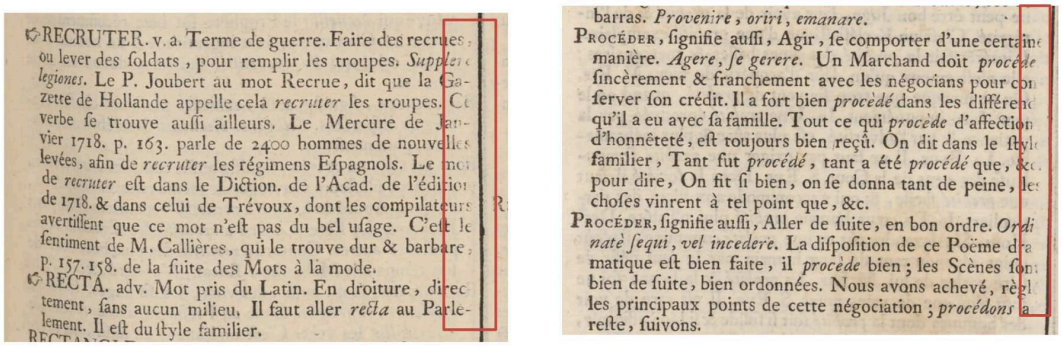


FIGURE 19 – Encre effacée en fin de plusieurs lignes [49].

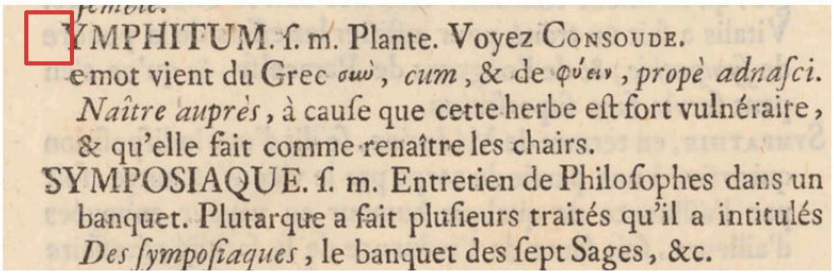

YMPHITUM. ſ. m. Plante. Voyez CONSOUDE.
e mot vient du Grec σὺν, *cum*, & de φύειν, *propè adnaſci*.
Naître auprès, à cauſe que cette herbe eſt fort vulnéraire,
& qu'elle fait comme renaître les chairs.
SYMPOSIAQUE. ſ. m. Entretien de Philoſophes dans un
banquet. Plutarque a fait pluſieurs traités qu'il a intitulés
Des ſympoſiaques, le banquet des ſept Sages, &c.

FIGURE 20 – Première lettre manquante du mot-vedette [49].

La figure si-dessus est un exemple de l'entrée *Symphitum* : le « S » initial du mot-vedette est manquant sur le scan, si bien que l'entrée apparaît comme « YMPHITUM », ce qui complique l'indexation et la recherche de l'entrée dans le texte reconnu.

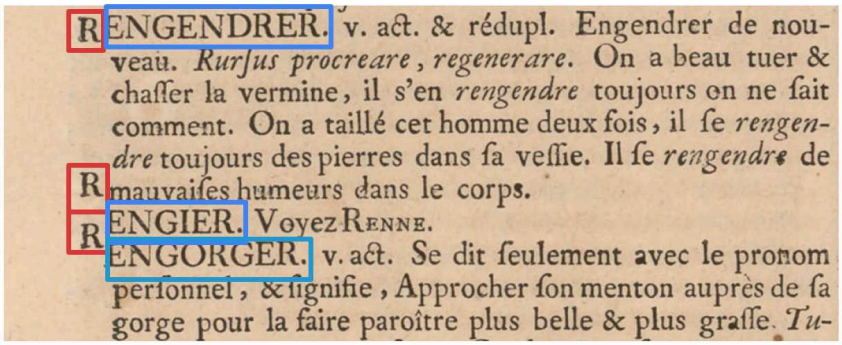

RENGENDRER. v. act. & rédupl. Engendrer de nou-
veau. *Rurſus procreare, regenerare.* On a beau tuer &
chaſſer la vermine, il s'en *rengendre* toujours on ne ſait
comment. On a taillé cet homme deux fois, il ſe *rengen-*
dre toujours des pierres dans ſa veſſie. Il ſe *rengendre* de
R mauvaiſes humeurs dans le corps.
R ENGIER. Voyez RENNE.
R ENGORGER. v. act. Se dit ſeulement avec le pronom
perſonnel, & ſignifie, Approcher ſon menton auprès de ſa
gorge pour la faire paroître plus belle & plus graſſe. *Tu-*

FIGURE 21 – Exemples de ettres marginales décalés [49].

Exemples d'entrées *Rengendrer*, *Rengier*, *Rengorger* : la lettre marginale « R » signalant la section alphabétique est décalée verticalement par rapport à l'entrée qu'elle introduit, ce qui peut perturber la segmentation en colonnes.

III.4 Approche par modèle de vision-langage (VLM)

III.4.1 Principe

Une première approche testée repose sur un modèle de vision-langage (VLM) combinant en une seule passe la reconnaissance optique de caractères, l'analyse de la structure de page et une capacité d'interprétation contextuelle du contenu. Plusieurs VLM ont été comparés : Mistral OCR, Pixtral-12B, GLM-OCR et Qwen2.5-VL 7B. Le pipeline testé suit les étapes présentées à la figure 22.

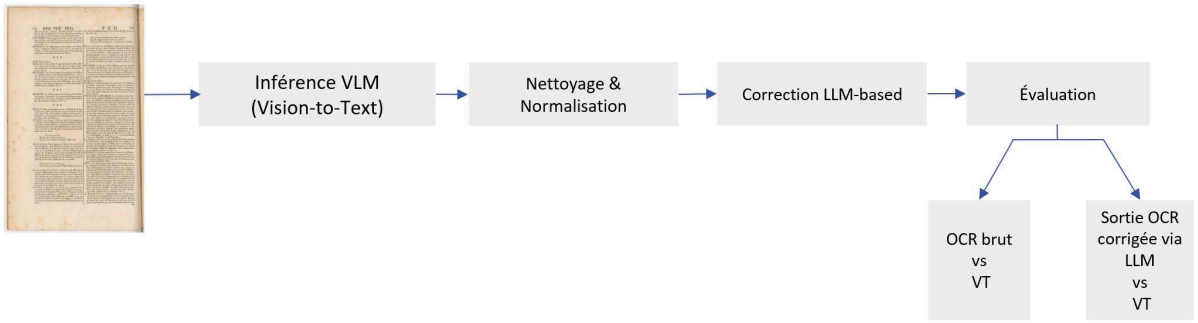


FIGURE 22 – Pipeline de l'approche VLM.

III.4.2 Résultats obtenus

La figure 23 présente les résultats obtenus sur l'échantillon de test, en comparant chaque VLM à quatre configurations de correction : sortie brute, correction zero-shot (ZS), correction few-shot (FS) et correction few-shot réduite (FS-R), évaluées en CER et WER (équations 2 et 3), avec Llama 3.3 70B (Groq) LLM utilisé dans cette partie.

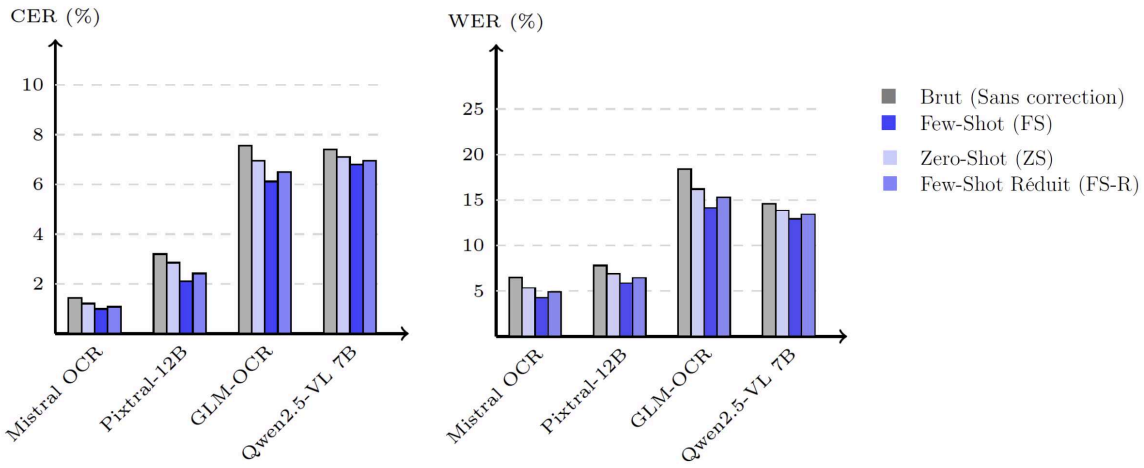


FIGURE 23 – Résultats comparatifs des VLM testés sur le corpus du Trévoux, avec et sans correction post-OCR par LLM.

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

TABLE 14 – Résultats CER et WER (%) des VLM testés sur le corpus du Trévoux, par configuration de correction post-OCR.

Ces résultats confirment plusieurs tendances déjà visibles à la figure 23. D'une part, Mistral OCR obtient systématiquement les meilleurs scores, quelle que soit la configuration testée, avec un CER brut de 1,44 % déjà nettement inférieur à celui

des trois autres modèles ; Pixtral-12B se positionne en second, tandis que GLM-OCR et Qwen2.5-VL 7B affichent des taux d'erreur nettement plus élevés en sortie brute (respectivement 7,56 % et 7,40 % de CER).

D'autre part, pour la correction post-OCR par LLM, la configuration few-shot obtient dans tous les cas les meilleurs résultats, devançant la version few-shot réduite, ce qui suggère que le nombre d'exemples fournis au LLM correcteur joue un rôle non négligeable dans la qualité de la correction.

Enfin, les écarts relatifs entre modèles se resserrent après correction sans toutefois s'annuler : même corrigé en few-shot, GLM-OCR (CER de 6,12 %) reste nettement moins performant que Mistral OCR (CER de 0,99 %), ce qui indique que la correction post-OCR ne compense pas entièrement les différences de qualité de reconnaissance initiale entre modèles.

III.4.3 Correction post-OCR par LLM : une piste exploratoire

Il est important de souligner que cette correction par LLM **ne constitue pas un objectif central du stage** : il s'agit d'une piste explorée en marge de l'approche VLM principale, dans la continuité des travaux de François, Eglin et Biou [47] sur la correction post-OCR.
Les résultats obtenus ici (figure 23) sont encourageants pour la configuration few-shot, mais restent à confirmer sur l'ensemble du corpus : comme le montre la synthèse bibliographique de la section II.9 (tableau 12), l'apport réel de la correction post-OCR par LLM est débattu dans la littérature récente (Boros et al. [10], Thomas et al. [11], Levchenko [6]), et les gains observés sur cet échantillon restreint ne préjugent pas d'un comportement identique à plus grande échelle.

Les figures 25, 26 et 27 illustrent, sur un exemple concret du corpus (entrée *Pêche*), l'effet de cette correction : l'image source, la sortie OCR brute et la sortie après correction LLM.

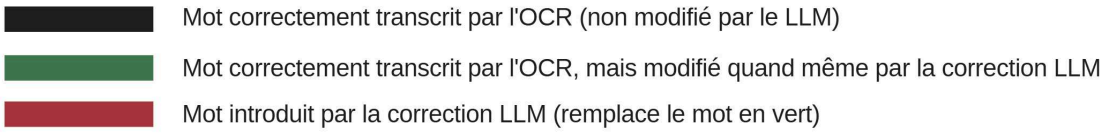
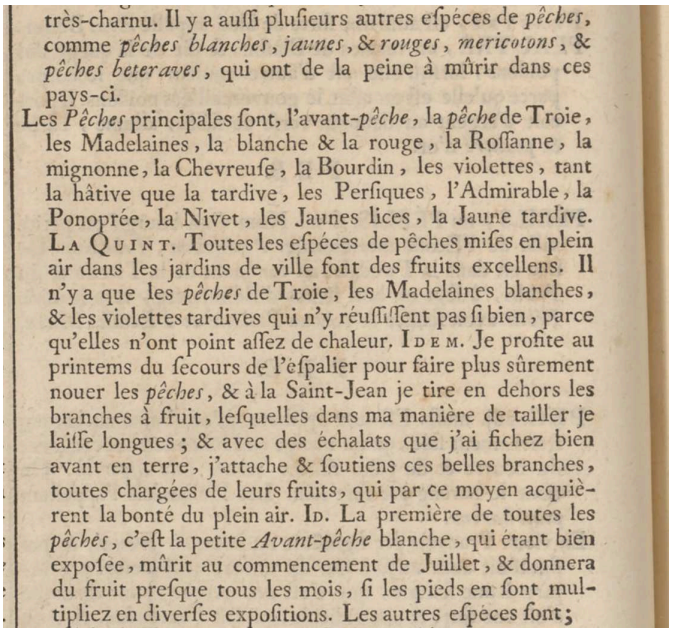


FIGURE 24 – Code couleur utilisé dans les figures 26 et 27.



très-charnu. Il y a auſſi pluſieurs autres eſpéces de *pêches*, comme *pêches blanches*, *jaunes*, & *rouges*, *mericotons*, & *pêches beteraves*, qui ont de la peine à mûrir dans ces pays-ci.

Les *Pêches* principales ſont, l'avant-*pêche*, la *pêche* de Troie, les Madelaines, la blanche & la rouge, la Roſſanne, la mignonne, la Chevreuſe, la Bourdin, les violettes, tant la hâtive que la tardive, les Perſiques, l'Admirable, la Ponoprée, la Nivet, les Jaunes lices, la Jaune tardive. LA QUINT. Toutes les eſpéces de pêches miſes en plein air dans les jardins de ville font des fruits excellens. Il n'y a que les *pêches* de Troie, les Madelaines blanches, & les violettes tardives qui n'y réuſſiſſent pas ſi bien, parce qu'elles n'ont point aſſez de chaleur. IDEM. Je profite au printems du ſecours de l'eſpalier pour faire plus ſûrement nouer les *pêches*, & à la Saint-Jean je tire en dehors les branches à fruit, leſquelles dans ma manière de tailler je laiſſe longues ; & avec des échalats que j'ai fichez bien avant en terre, j'attache & ſoutiens ces belles branches, toutes chargées de leurs fruits, qui par ce moyen acquiè-rent la bonté du plein air. ID. La première de toutes les *pêches*, c'eſt la petite *Avant-pêche* blanche, qui étant bien expoſée, mûrit au commencement de Juillet, & donnera du fruit preſque tous les mois, ſi les pieds en ſont mul-tipliez en diverſes expoſitions. Les autres eſpéces ſont;

FIGURE 25 – Image source, extrait du corpus (entrée *Pêche*), d'après le *Dictionnaire de Trévoux* [49].

Il y a aussi plusieurs autres espéces de pêches, comme pêches blanches, jaunes, & rouges, mericotons, & pêches beteraves, qui ont de la peine à mûrir dans ces pays-ci. Les Pêches principales sont, l'avant- pêche, la pêche de Troie, les Madelaines, la blanche & la rouge, la Rossanne, la mignonne, la Chevreuse, la Bourdin, les violettes, tant la hâtive que la tardive, les Persiques, l'Admirable, la Ponoprée, la Nivet, les Jaunes lices, la Jaune tardive. La Quint. Toutes les espéces de pêches mises en plein air dans les jardins de ville font des fruits excellens. Il n'y a que les pêches de Troie, les Madelaines blanches, & les violettes tardives qui n'y réussissent pas si bien, parce qu'elles n'ont point assez de chaleur. Idem. Je profite au printems du secours de l'espalier pour faire plus sûrement nouer les pêches, & à la Saint-Jean je tire en dehors les branches à fruit, lesquelles dans ma manière de tailler je laisse longues ; & avec des échalats que j'ai fichez bien avant en terre, j'attache & soutiens ces belles branches, toutes chargées de leurs fruits, qui par ce moyen acquièrent la bonté du plein air. Id. La première de toutes les pêches, c'est la petite Avant-pêche blanche, qui étant bien exposée, mûrit au commencement de Juillet, & donnera du fruit presque tous les mois, si les pieds en sont multipliez en diverses expositions. Les autres espéces sont

FIGURE 26 – Sortie OCR brute sur le même extrait.

Il y a aussi plusieurs autres espèces de pêches, comme pêches blanches, jaunes, & rouges, mericotons, & pêches betteraves, qui ont de la peine à mûrir dans ces pays-ci. Les Pêches principales sont, l'avant-pêche, pêche, la pêche de Troie, les Madelaines, la blanche & la rouge, la Rossanne, la mignonne, la Chevreuse, la Bourdin, les violettes, tant la hâtive que la tardive, les Persiques, l'Admirable, la Ponoprée, la Nivette, les Jaunes lisses, la Jaune tardive. La Quint. Toutes les espèces de pêches mises en plein air dans les jardins de ville font des fruits excellents. Il n'y a que les pêches de Troie, les Madelaines blanches, & les violettes tardives qui n'y réussissent pas si bien, parce qu'elles n'ont point assez de chaleur. Idem. Je profite au printemps du secours de l'espalier pour faire plus sûrement nouer les pêches, & à la Saint-Jean je tire en dehors les branches à fruit, lesquelles dans ma manière de tailler je laisse longues; & avec des échalats que j'ai fixés bien avant en terre, j'attache & soutiens ces belles branches, toutes chargées de leurs fruits, qui par ce moyen acquièrent la bonté du plein air. Id. La première de toutes les pêches, c'est la petite Avant-pêche blanche, qui étant bien exposée, mûrit au commencement de Juillet, & donnera du fruit presque tous les mois, si les pieds en sont multipliés en diverses expositions. Les autres espèces sont

FIGURE 27 – Sortie après correction post-OCR par LLM sur le même extrait.

La plupart de ces modifications sont des **modernisations orthographiques non désirées** :

— « beteraves » → « betteraves »
— « printems » → « printemps »
— « lices » → « lisses »

Malgré la consigne du prompt de conserver la graphie d'époque, un travers qui touche aussi, ailleurs dans le corpus, les terminaisons verbales en « -oit» (« levoit », « avoit », «paroit », « faisoit »), parfois normalisées à tort en « -ait ».

Un cas plus préoccupant apparaît également ici : « fichez » (planté, enfoncé en terre) est remplacé par « fixés », une **substitution lexicale** qui altère le sens du passage plutôt qu'une simple modernisation. Ce type d'erreur, absent de la sortie OCR brute et donc introduit par la correction elle-même, illustre concrètement l'apport incertain de la correction post-OCR par LLM discuté en section II.9 : celle-ci ne se limite pas à corriger des erreurs de reconnaissance, elle peut aussi en introduire de nouvelles.

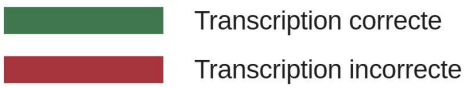


FIGURE 28 – Code couleur utilisé dans la figure 29.

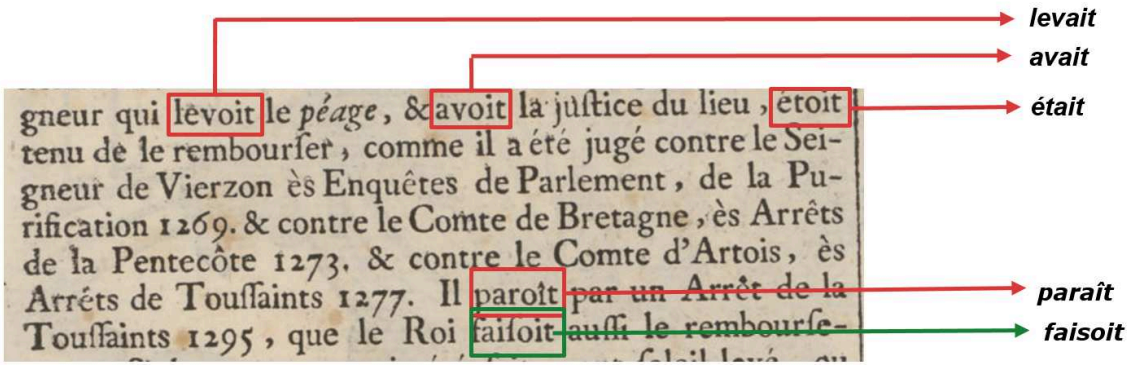


FIGURE 29 – Exemple de correction post-OCR par LLM sur un extrait du corpus du Trévoux, d'après le *Dictionnaire de Trévoux* [49].

III.5 Approche modulaire : segmentation et OCR spécialisé

III.5.1 Principe général

En parallèle de l'approche VLM, une approche plus modulaire a également été testée, consistant à découpler explicitement les étapes de segmentation de la mise en page et de reconnaissance de caractères. Cette approche s'appuie sur la plateforme eScriptorium [50] et sur le moteur de segmentation et de reconnaissance **Kraken** [21], dont le modèle de segmentation de ligne `blla` a été testé sur les pages du corpus, selon le pipeline présenté à la figure 30.

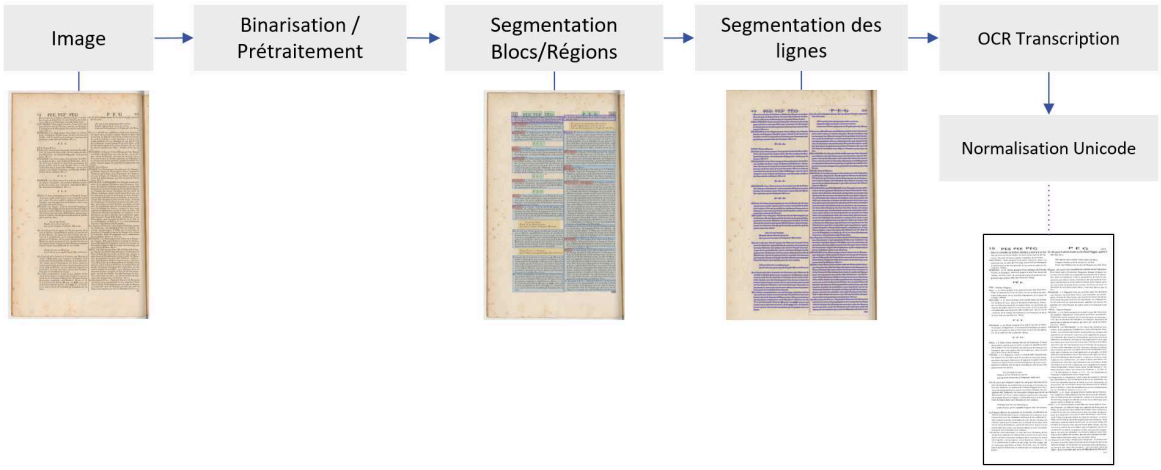


FIGURE 30 – Pipeline de l'approche modulaire.

La segmentation testée distingue plusieurs catégories d'éléments de mise en page propres au genre lexicographique (mot-vedette, vedette secondaire, paragraphe, citation, manicule, lettrine, illustration, titre, tableau, etc.), présentées à la figure 31 et appliquées à un exemple de pages annotées (figure 32).

FIGURE 31 – Catégories utilisées pour la segmentation fine des pages du Trévoux.

FIGURE 32 – Exemple de pages annotées selon les catégories de segmentation.

Cette annotation se fait en deux temps, illustrés à la figure 33 : une première étape sépare les deux colonnes de la page (gauche/droite) afin de rétablir l'ordre de lecture, préalable à la segmentation des lignes ; une seconde étape identifie, au sein de chaque colonne, les régions fines selon les 12 classes typées présentées ci-dessus, ce qui restitue la structure de l'article. Cette annotation en deux étapes a été appliquée à un échantillon de 170 pages.

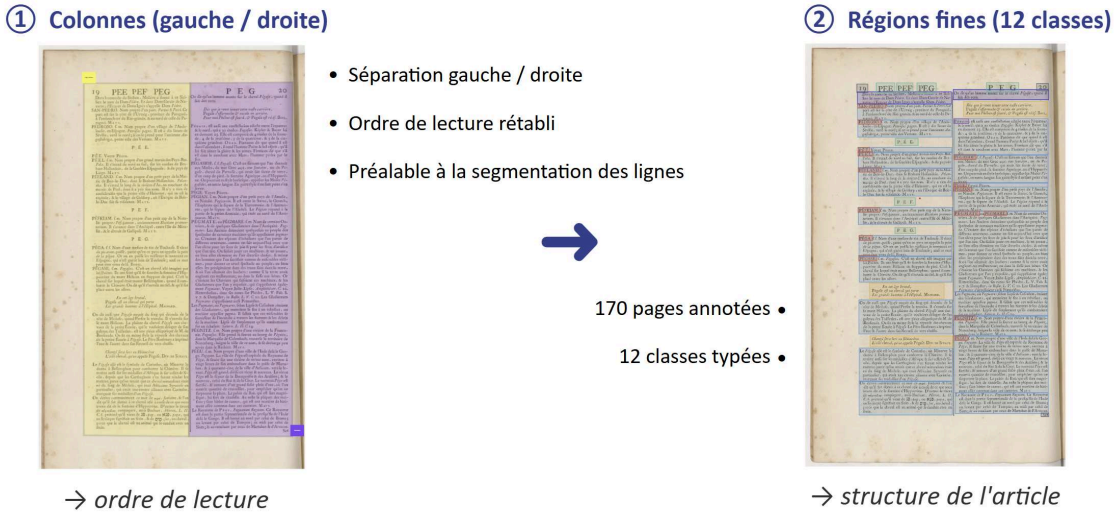


FIGURE 33 – Les deux étapes d'annotation de la mise en page.

III.5.2 Comparaison de modèles de détection de mise en page

Un test comparatif a également été mené entre deux approches de détection de la mise en page sur un échantillon de 50 pages : DocLayout-YOLO [9] (un détecteur d'objets spécialisé) et Qwen2.5-VL [28] utilisé comme détecteur de mise en page. Le tableau ci-dessous résume les 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 – *layout seul*	**Possible**

TABLE 15 – Comparaison DocLayout-YOLO / Qwen2.5-VL pour la détection de mise en page sur un échantillon de 50 pages.

DocLayout-YOLO obtient des scores de détection légèrement supérieurs, conformément au compromis vitesse/précision déjà annoncé en section II.6.3. Sur la base de ce résultat, DocLayout-YOLO est retenu comme modèle de détection de la mise en page (blocs et régions) pour la suite de la chaîne modulaire. Qwen2.5-VL présente néanmoins l'avantage de pouvoir également détecter les styles typographiques (italique, petites et grandes capitales) en plus de la seule mise en page, une information potentiellement utile pour distinguer, au sein d'une entrée, la vedette des exemples ou citations, et qui rejoint directement la problématique de classification de style posée en section II.7.

III.5.3 Comparaison avec les moteurs OCR traditionnels

Une comparaison a également été menée avec des moteurs OCR traditionnels non spécifiquement adaptés au corpus, afin d'établir un choix argumenté d'architecture de reconnaissance pour la suite de la chaîne modulaire. Le tableau ci-dessous résume les résultats obtenus sur l'échantillon de test.

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	Bonne détection des lignes mais hallucinations.
Tesseract 5	8,40	19,80	Échecs multi-colonnes et confusion $\int \to f$.
Kraken (CATMuS-Print Large)	**4,50**	**11,30**	Meilleure robustesse native sur la graphie historique.

TABLE 16 – Comparaison des moteurs OCR traditionnels et de Kraken (CATMuS-Print Large) sur l'échantillon de test.

Ce résultat confirme, sur le corpus du Trévoux, le même constat déjà établi dans la littérature (section II.5.4, Levchenko [6]) : les moteurs OCR génériques (EasyOCR, Surya, Tesseract) sont nettement moins performants qu'un moteur spécialisé et adapté à la typographie ancienne (Kraken avec CATMuS-Print Large), avec un CER pratiquement deux fois inférieur. Sur la base de ce résultat, le couple Kraken + CATMuS-Print est retenu comme architecture de reconnaissance pour la suite du stage ; un fine-tuning spécifique au corpus du Trévoux, détaillé à la section III.5.5 , vient encore affiner ce résultat.

III.5.4 Interface eScriptorium

Le travail d'annotation et de correction manuelle nécessaire à l'entraînement et à l'évaluation des modèles de segmentation et de reconnaissance a été réalisé sur la plateforme eScriptorium [50] (figures 34 à 36), qui permet la visualisation conjointe de l'image source, des polygones de segmentation et de la transcription associée à chaque ligne.

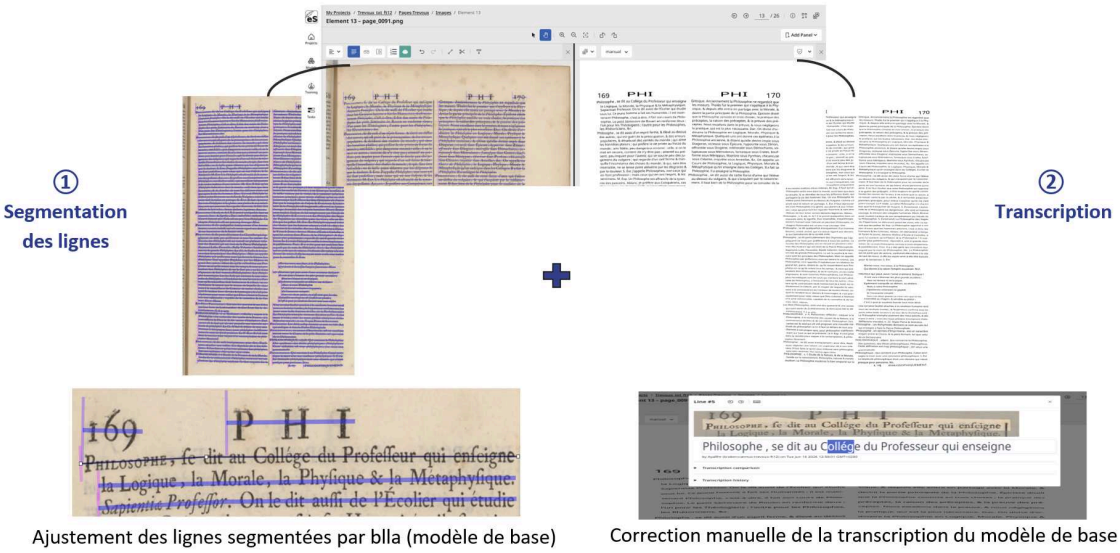


FIGURE 34 – Vue du projet sur une page du corpus (entrée *Philosophe*, p. 169), capture d'écran issue des tests menés sur la plateforme eScriptorium [50].

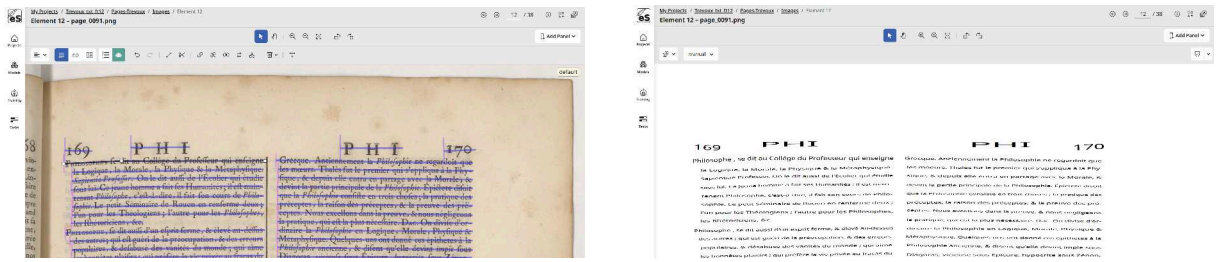


FIGURE 35 – Interface illustrant les deux principales étapes du travail d'annotation [50].

On peut faire la visualisation et la correction de la segmentation des lignes (gauche), puis transcription et historique des modifications associées (droite), captures d'écran issues des tests menés sur la plateforme eScriptorium

FIGURE 36 – Zoom sur l'en-tête de page et le début de l'entrée *Philosophe*, capture d'écran issue des tests menés sur la plateforme eScriptorium [50].

III.5.5 Fine-tuning du modèle de reconnaissance CATMuS

Un fine-tuning du modèle de reconnaissance CATMuS-Print [40] a été testé sur un échantillon annoté de pages du corpus, avec des jeux d'entraînement de taille croissante (12, 20, 25, 30, 35, 40 et 45 pages) aux CER/WER les plus élevés, afin d'évaluer l'apport d'une adaptation au domaine par rapport au modèle générique. La figure 37 présente l'évolution du CER et du WER (équations 2 et 3) en fonction du nombre de pages d'entraînement.

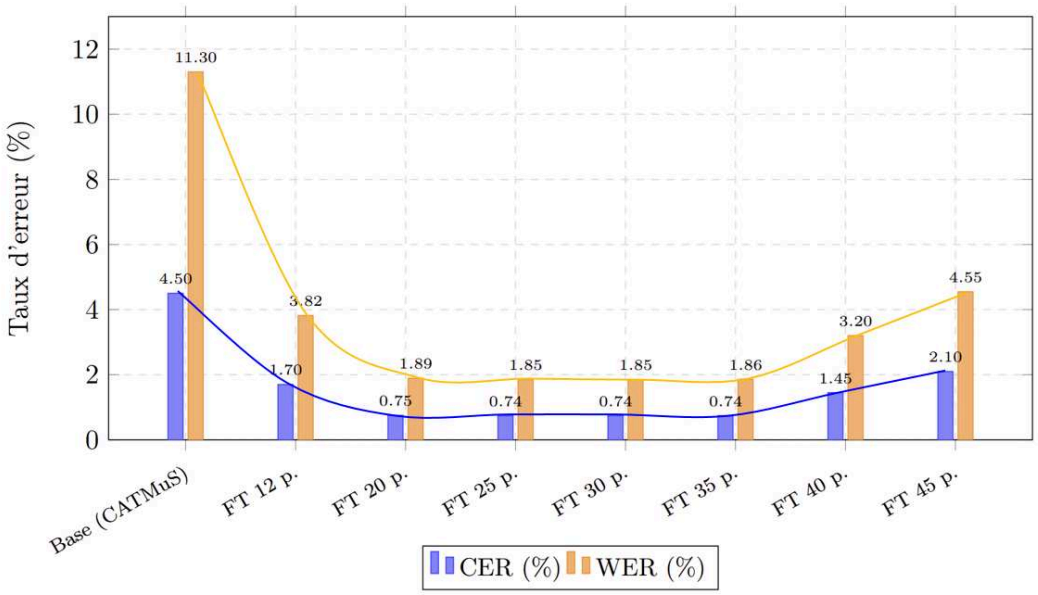


FIGURE 37 – Évolution du CER et du WER de CATMuS-Print selon le nombre de pages annotées utilisées pour le fine-tuning.

Le CER diminue fortement jusqu'à un minimum autour de 25 à 35 pages d'entraînement, puis remonte au-delà de 35 pages (CER de 2,10 % à 45 pages). Trois régimes se distinguent selon la taille du jeu de fine-tuning :

- **FT 12 à 25 pages : optimisation rapide**, où il y a alignement efficace sur la typographie du *Dictionnaire de Trévoux*.
- **FT 25 à 35 pages : zone de convergence**, c'est là où le plafond de performance est atteint ; minimum absolu de CER à 0,738 % (WER à 1,85 %).

— **FT 40 à 45 pages : surapprentissage**, on a donc une possible mémorisation des bruits physiques (accents, terminaisons, etc.).

III.5.6 Analyse qualitative de la segmentation

L'amélioration de la qualité de la segmentation de lignes se traduit directement par une amélioration de la qualité de transcription en aval : une ligne mal détectée ou tronquée par le modèle de segmentation ne peut être correctement reconnue par le modèle de transcription, quelle que soit la qualité de ce dernier. C'est ce lien qui justifie le fine-tuning du modèle `blla` de Kraken, dont l'apport est mesuré dans cette section. La figure 38 illustre ce lien sur l'entrée *Ressentir* : une première segmentation de ligne imprécise produit un recadrage tronqué de l'image source, à l'origine d'une erreur de transcription (« Jaisisio- ») ; une segmentation affinée du modèle `blla` restitue la ligne complète, ce qui permet une transcription correcte (« J'ai ressenti »).

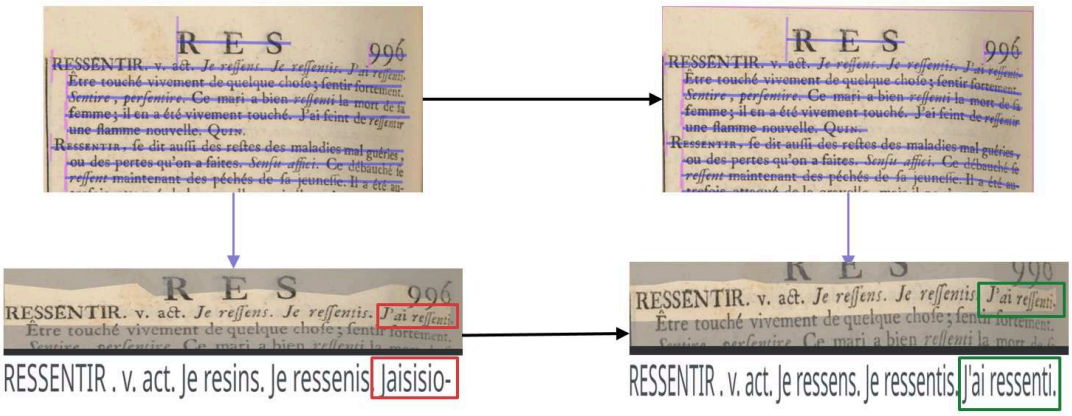


FIGURE 38 – Exemple de segmentation puis transcription itérative sur l'entrée *Ressentir* [50].

L'apport de ce fine-tuning a été mesuré à travers trois indicateurs : le taux de lignes manquées, le taux de lignes tronquées ou coupées, et le recouvrement moyen (IoU, équation 6) entre les lignes détectées et la vérité terrain. Ces indicateurs ont été mesurés séparément sur un sous-ensemble de pages à mise en page standard (20 pages) et sur un sous-ensemble de pages présentant une courbure de reliure prononcée (*book curve*, 10 pages), plus difficile à segmenter. Le tableau ci-dessous et la figure 39 présentent ces résultats.

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

TABLE 17 – Métriques de segmentation de lignes avant/après fine-tuning du modèle *blla* de Kraken.

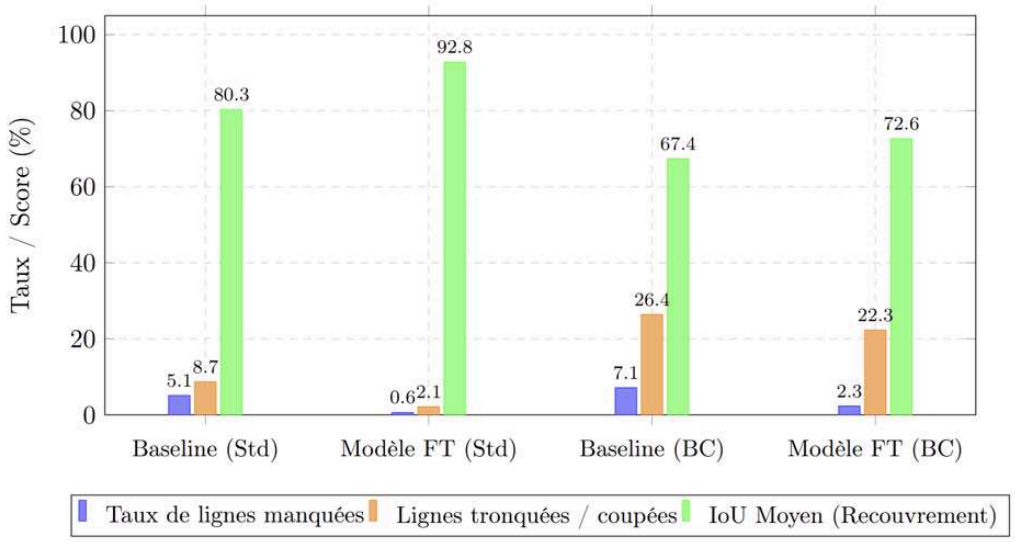


FIGURE 39 – Comparaison visuelle des métriques de segmentation avant/après fine-tuning.

*BC : Book Curve, courbure de reliure. Le fine-tuning améliore nettement les trois indicateurs sur les deux types de pages, mais l'écart entre pages standards et pages *book curve* reste important même après adaptation (IoU moyen de 92,8 % contre 72,6 %), ce qui confirme que la courbure de reliure constitue une source d'erreur structurelle non résolue par le seul fine-tuning du modèle de segmentation de lignes.

III.6 Synthèse comparative

Les deux familles d'approches testées présentent des compromis différents. L'approche VLM (figure 23) offre une mise en œuvre plus directe et de bons résultats dès la sortie brute pour Mistral OCR (CER de 1,44 %), avec un gain supplémentaire en configuration few-shot. L'approche modulaire, plus coûteuse à mettre en œuvre (annotation, fine-tuning), permet en revanche un contrôle plus fin de chaque étape de la chaîne de traitement et une adaptation ciblée au corpus : le fine-tuning de CATMuS-Print ramène le CER à 0,74 % dans sa meilleure configuration (25 à 35 pages d'entraînement), un résultat comparable voire supérieur à celui de l'approche VLM.

Approche	**CER (%)**	**WER (%)**
VLM end-to-end	0,99	4,26
Chaîne modulaire	0,74	1,85

TABLE 18 – Synthèse comparative des deux approches testées.

C'est cette approche modulaire qui est retenue à l'issue de ces tests. D'une part, son fonctionnement par étapes indépendantes (segmentation, puis transcription) permet de localiser précisément l'origine d'une erreur et de cibler la correction ou le réentraînement sur l'étape concernée, alors qu'une approche VLM end-to-end ne permet pas toujours d'identifier directement si une erreur provient d'une mauvaise détection de la structure

de la page ou d'une mauvaise reconnaissance du texte lui-même.

D'autre part, le coût de traitement d'une solution VLM comme Mistral OCR devient un frein important dès lors qu'il faut appliquer le pipeline à l'ensemble du corpus, soit plusieurs centaines de pages : ce coût, cumulé sur un tel volume, est nettement plus élevé que celui de la chaîne modulaire une fois les modèles entraînés. La combinaison de ces deux facteurs, contrôlabilité étape par étape et coût de traitement à grande échelle, motive le choix de l'approche modulaire pour la suite du stage.

III.7 Détection de mots et classification de style typographique

Comme illustré en section III.3.3 (figures 10 à 12), le *Dictionnaire de Trévoux* insère fréquemment des segments en grec ancien et en hébreu au sein d'entrées par ailleurs françaises ou latines, et alterne plus généralement romain, italique et petites ou grandes capitales pour distinguer vedette, exemple et citation. Or le modèle CATMuS-Print [40] utilisé avec Kraken (section III.5.5) ignore totalement cette dimension :

- il produit une simple séquence de caractères, sans information de style.
- il échoue purement et simplement sur le grec et l'hébreu, deux écritures absentes de ses données d'entraînement.

Combler ce manque suppose de disposer, indépendamment de la transcription elle-même, d'un repérage fiable de chaque mot de la page et de son style typographique. Cette section présente, dans l'esprit des classifieurs de police et de style et des travaux d'identification de script décrits en section II.7 (Tensmeyer et al. [35], Yang et al. [43]), les briques développées à cette fin :

- un détecteur de mots générique ;
- deux classifieurs de style comparés entre eux (un réseau convolutif dédié et un fine-tuning léger de Qwen2.5-VL) ;
- leur validation sur des pages du corpus jamais vues à l'entraînement.

L'exploitation de cette classification pour le repérage du grec et de l'hébreu est détaillée dans un second temps. La figure 40 résume l'architecture d'ensemble de ce sous-système, indépendant du pipeline de transcription CATMuS-Print mais fusionné avec lui en aval.

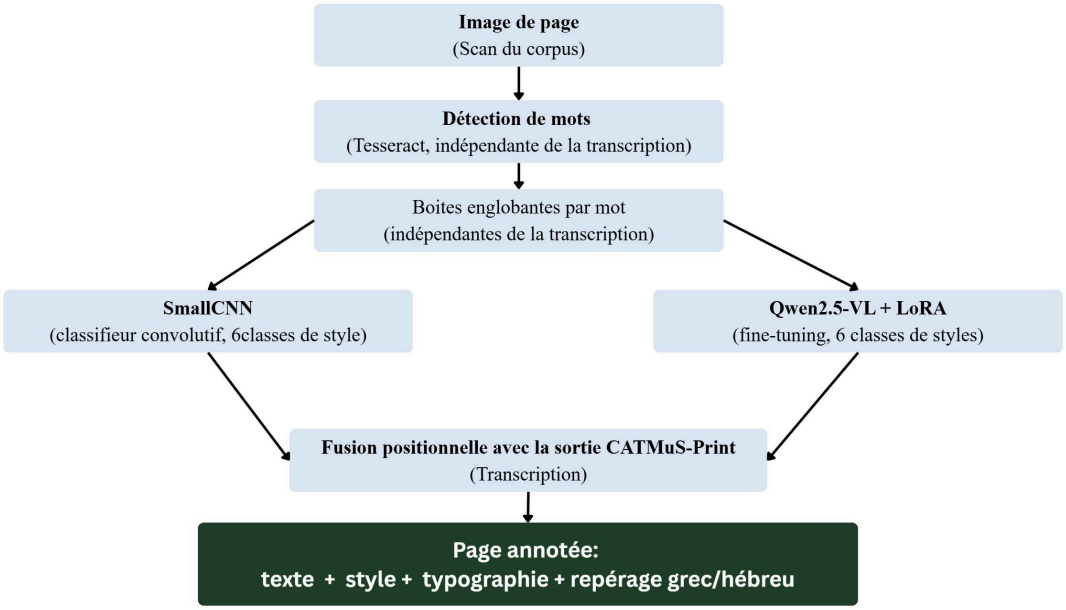


FIGURE 40 – Architecture du sous-système de détection de mots et de classification de style, en parallèle du pipeline de transcription CATMuS-Print.

III.7.1 Détection de mots

La classification de style et le repérage du grec/hébreu supposent tous deux de disposer, pour chaque mot de la page, d'une boîte englobante précise et indépendante de tout moteur de transcription : la sortie de Kraken (section III.5.1) fournit bien des boîtes par mot, mais uniquement pour les mots qu'il reconnaît effectivement, ce qui exclut par construction les mots en écriture inconnue, exactement ceux qu'il s'agit de repérer. Un détecteur de mots dédié, s'appuyant sur Tesseract, a donc été développé indépendamment de tout moteur OCR de production, avec pour seul objectif la localisation spatiale des mots, et non leur transcription. Deux heuristiques de post-traitement corrigent des artefacts récurrents sur une mise en page dense à deux colonnes :

— une fusion des boîtes séparées à tort au sein d'un même mot ;
— un redécoupage des boîtes fusionnées à tort entre deux mots adjacents, en particulier au voisinage de la gouttière entre les deux colonnes du Trévoux.

Un mode de prévisualisation, qui affiche les boîtes détectées directement sur l'image source sans dépendre d'aucun classifieur, a par ailleurs permis, lors de la mise au point du pipeline, de vérifier isolément la qualité de la détection et d'écarter une fausse piste lors du diagnostic d'une anomalie apparue plus en aval. La figure 41 illustre le résultat de cette détection sur une page du corpus, avec les boîtes englobantes obtenues autour de chaque mot.

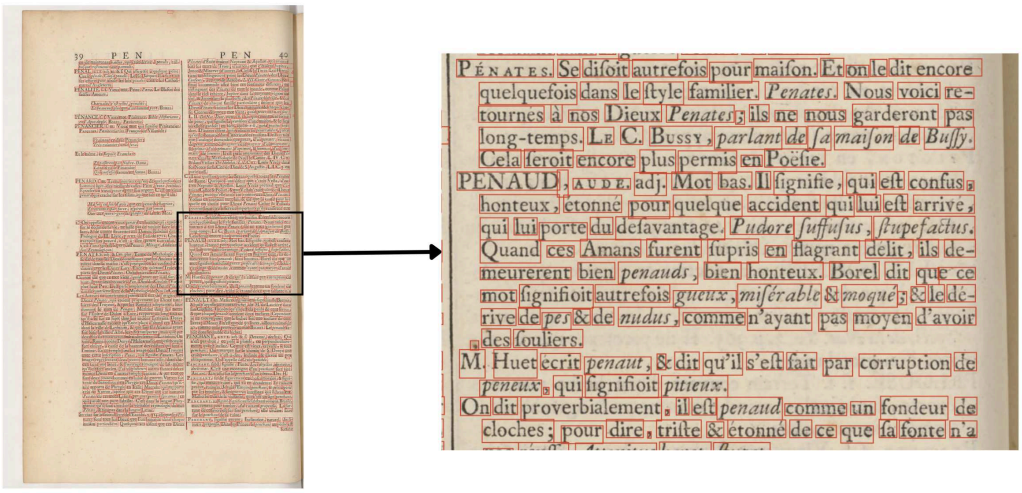


FIGURE 41 – Résultat de la détection de mots (boîtes englobantes) sur une page du corpus (page39-40).

III.7.2 Classification de style par réseau convolutif : SmallCNN

Un premier classifieur, léger et rapide à exécuter sur plusieurs centaines de mots par page, a été entraîné spécifiquement sur ce corpus : *SmallCNN*, un réseau convolutif qui prend en entrée l'image d'un mot détecté et prédit l'un de six styles (normal, italique, petite capitale, grande capitale, grec, hébreu). Le déséquilibre marqué entre classes, la classe *normal* représentant l'essentiel des mots d'une page contre quelques dizaines d'occurrences pour le grec ou l'hébreu, est traité par une fonction de perte et un échantillonnage adaptés qui rééquilibrent l'apprentissage en faveur des classes rares. Le jeu d'entraînement combine trois sources :

— une annotation manuelle de 30 pages du corpus via la plateforme Label Studio ;
— un minage automatique de la classe normale (très majoritaire et coûteuse à annoter exhaustivement) ;
— une génération synthétique de mots grecs et hébreux, nécessaire tant ces styles restent peu représentés même après annotation exhaustive des 30 pages.

Les performances de SmallCNN restent néanmoins limitées par la capacité du réseau convolutif à capturer le contexte global et à généraliser sur les écritures complexes sans fine-tuning profond, comme le montre le tableau 19.

Volume d'entraînement	Précision styles (%)	Rappel styles (%)	F1-score global (%)
20 pages	71,8	68,2	69,9
40 pages	76,4	73,5	74,9
60 pages	81,0	78,1	79,5
80 pages	83,7	81,4	82,5
100 pages	**86,2**	**84,0**	**85,1**

TABLE 19 – Résultats de l'entraînement progressif du réseau SmallCNN selon le volume de pages annotées.

III.7.3 Fine-tuning de Qwen2.5-VL par LoRA pour la classification de style

En complément de SmallCNN, un second classifieur a été exploré : un fine-tuning léger, par LoRA, de Qwen2.5-VL, le même modèle déjà mobilisé comme moteur OCR de bout en bout (section III.4.1) et comme détecteur de mise en page (section III.5.2). L'objectif est double :

— évaluer si un modèle de fondation, adapté sur un volume de données modeste, égale ou dépasse un classifieur convolutif dédié ;
— mutualiser l'outillage avec l'approche VLM déjà testée plutôt que de multiplier les briques logicielles disjointes.

Le jeu de données, d'environ 8 000 exemples issus des mêmes sources que celles de SmallCNN, a été réparti de façon stratifiée par classe afin que chacun des six styles, y compris le grec et l'hébreu malgré leur faible nombre d'exemples, soit représenté à l'entraînement comme à la validation. La figure 42 schématise ce protocole d'entraînement.

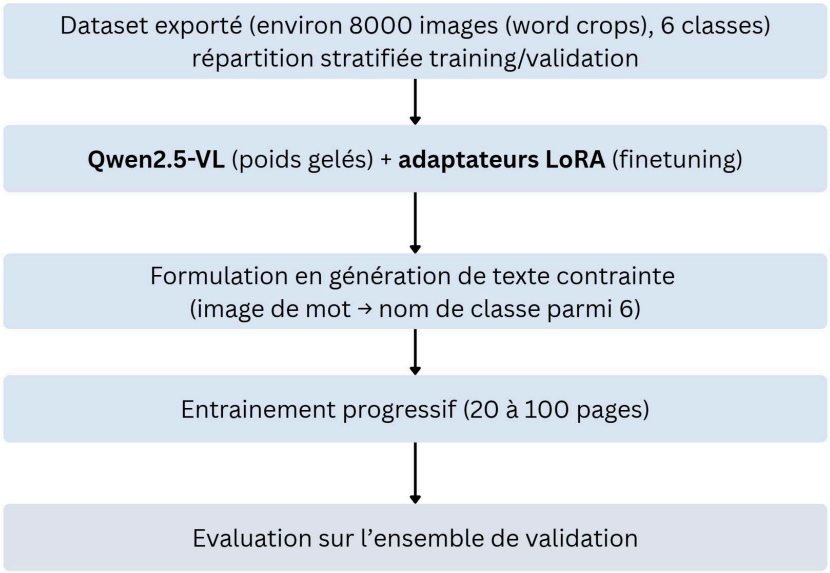


FIGURE 42 – Protocole de fine-tuning LoRA de Qwen2.5-VL pour la classification de style typographique.

Le fine-tuning a été mené de façon progressive, en augmentant par paliers le volume de pages annotées utilisées pour l'entraînement (20, 40, 60, 80 puis 100 pages), afin d'évaluer l'effet du volume de données sur la qualité de détection et de classification. Le tableau 20 résume, pour chaque palier, la précision de classification des styles, le rappel de classification des styles et le F1-score global.

Volume d'entraînement	Précision styles (%)	Rappel styles (%)	F1-score global (%)
20 pages	81,2	78,4	79,8
40 pages	86,5	84,1	85,3
60 pages	90,1	88,7	89,4
80 pages	93,4	92,0	92,7
100 pages	**95,8**	**94,6**	**95,2**

TABLE 20 – Résultats du fine-tuning progressif de Qwen2.5-VL-7B-Instruct par LoRA, selon le volume de pages annotées utilisées à l'entraînement.

Les performances progressent de façon monotone avec le volume d'entraînement, sans signe de saturation nette jusqu'à 100 pages, ce qui suggère qu'un volume d'annotation plus important pourrait encore améliorer les résultats. Ce résultat est néanmoins à interpréter avec prudence dans la mesure où l'ensemble de validation, bien que disjoint du jeu d'entraînement, provient de la même distribution que celui-ci, ce qui motive directement la validation sur des pages totalement inédites décrite ci-après. Les deux classifieurs ont par ailleurs été évalués séparément en précision, rappel et F-mesure par classe (équations 4 et 5). Le classifieur Qwen2.5-VL a par ailleurs été évalué en F-mesure par classe (équation 5) sur trois des styles les plus délicats à distinguer, le grec et l'hébreu du fait de leur faible représentation, l'italique à titre de comparaison : la figure 43 en présente les résultats.

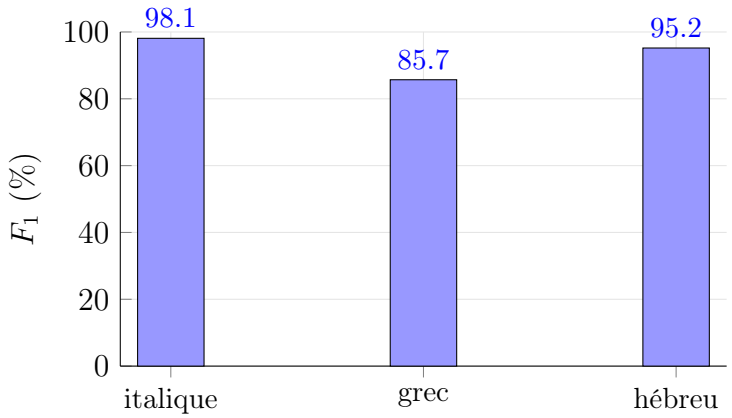
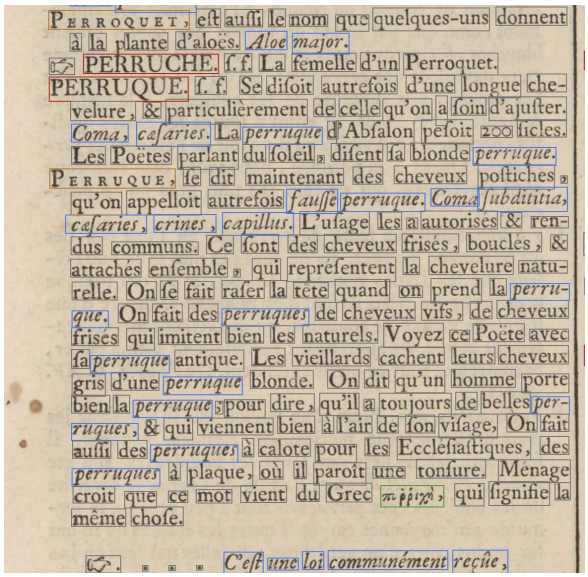
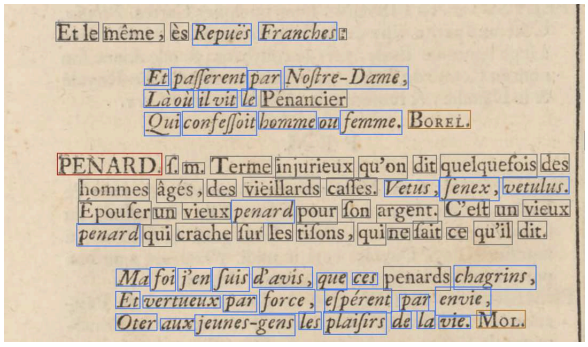


FIGURE 43 – F-mesure du classifieur de style Qwen2.5-VL pour l'italique, le grec et l'hébreu, sur le jeu de validation.

Les figures ci-dessous montre des exemples d'extraits de pages du corpus sur lesquelles le modèle Qwen2.5-VL fine-tuné a été appliqué pour la classification des styles typographiques.

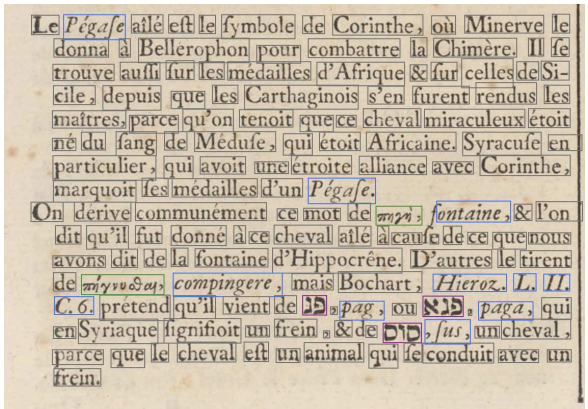


(a) Exemple 1

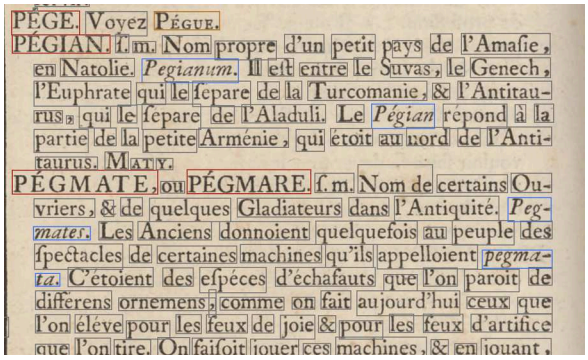


(b) Exemple 2

FIGURE 44 – Exemples de pages du corpus appliqués au modèle Qwen2.5-VL pour la classification de style typographique (exemples 1 et 2) avec la légende des couleurs de style.



(a) Exemple 3



(b) Exemple 4

FIGURE 45 – Exemples de pages du corpus appliqués au modèle Qwen2.5-VL pour la classification de style typographique (exemples 3 et 4) avec la légende des couleurs de style.

Repérage du grec et de l'hébreu :

La figure 46 synthétise l'ensemble du traitement mis en place pour le grec et l'hébreu, depuis la page source jusqu'au texte transcrit par Kraken dans lequel les segments grecs et hébreux sont signalés par les marqueurs `[GREC]` et `[HÉBREU]`. Ce point est essentiel : à

l'état actuel, ces marqueurs ne sont que des repères de position, et non une transcription. L'étape suivante, non encore implémentée et identifiée comme piste de poursuite du stage, consiste à remplacer chacun de ces marqueurs par le résultat d'un modèle de transcription spécifiquement dédié à l'écriture concernée (un moteur entraîné sur du grec pour `[GREC]`, un moteur entraîné sur de l'hébreu pour `[HÉBREU]`), appliqué au *crop* correspondant.

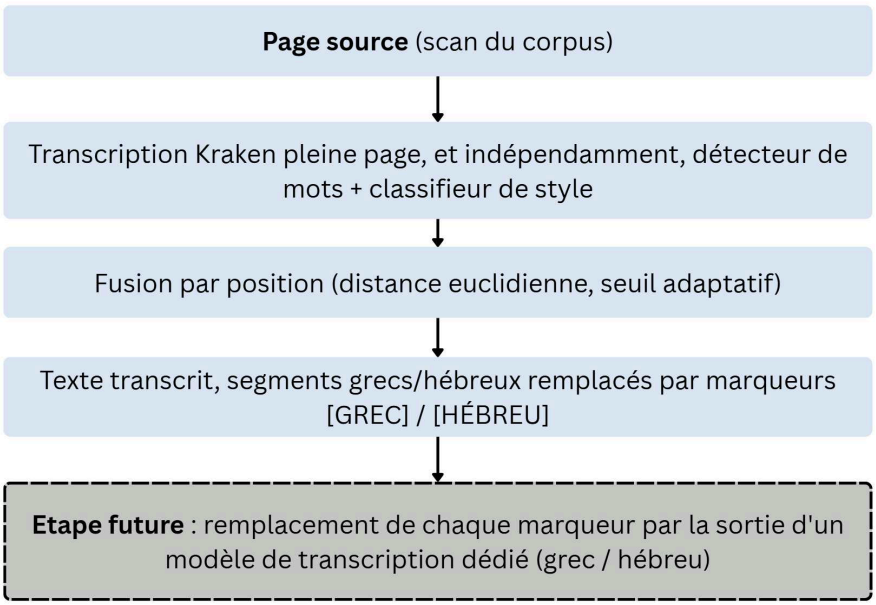


FIGURE 46 – Synthèse du traitement du grec et de l'hébreu.

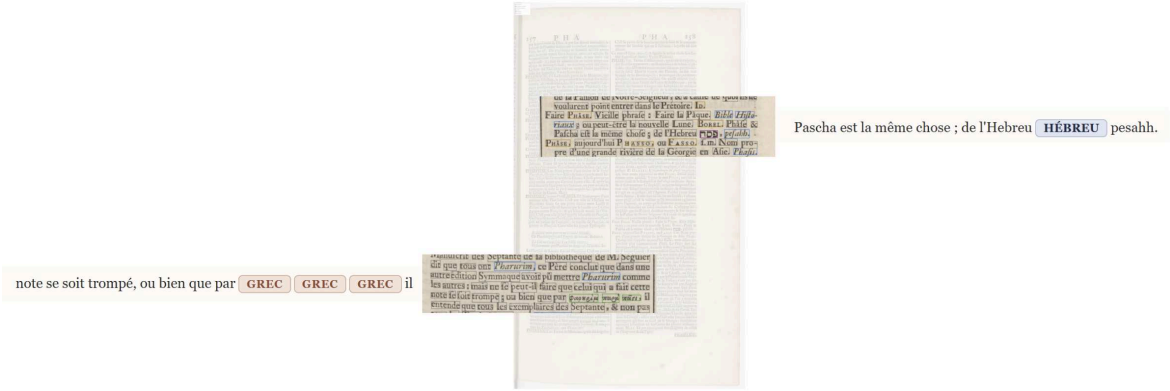


FIGURE 47 – Exemple de détection et de remplacement des segments en écriture grecque et hébraïque par des marqueurs positionnels au sein du texte transcrit.

III.8 Perspectives futures

Les tests exploratoires menés durant le stage ouvrent plusieurs pistes de poursuite, à approfondir dans la suite du stage :

— **Transcription spécialisée du grec et de l'hébreu.** Comme détaillé ci-dessus (figure 46), les marqueurs `[GREC]` et `[HÉBREU]` ne font aujourd'hui que repérer la position de ces segments, sans les transcrire. L'étape suivante consiste à router

les *crops* correspondants vers des modèles de transcription spécifiquement dédiés à chaque écriture (par exemple Kraken avec un modèle entraîné sur des corpus grecs, ou les modèles publics de Transkribus pour l'hébreu). Cette piste suppose au préalable de constituer une *vérité de terrain* propre au corpus du Trévoux, c'est-à-dire un ensemble de pages annotées où le grec et l'hébreu sont effectivement présents et transcrits manuellement : elle est aujourd'hui indisponible, et nécessaire pour entraîner ou, au minimum, évaluer objectivement ces modèles spécialisés avant tout choix définitif.

— **Évaluation quantitative sur l'ensemble du corpus.** Les résultats chiffrés obtenus jusqu'ici (section III.6) portent sur des échantillons de pages ; une évaluation plus complète sur l'ensemble des 956 pages du corpus permettrait d'arbitrer plus solidement entre l'approche VLM et l'approche modulaire.

— **Vers le format cible TEI Lex-0.** Les blocs et mots extraits par DocLayout-YOLO et Qwen2.5-VL fournissent les zones et le texte reconnu, mais pas encore la structure logique de l'entrée lexicographique (vedette, sous-entrée, exemple, citation) requise par TEI Lex-0 (tableau 3). L'étape suivante consiste à faire correspondre les classes de zones détectées à la structure XML-TEI, avant intégration dans la plateforme eScriptorium [50] pour un flux de travail continu accessible aux chercheurs en sciences humaines et sociales.

— **Outils d'assistance à l'analyse de manuscrits patrimoniaux.** Des interfaces récentes comme **InkSight** [51], conçues pour assister par IA l'analyse de manuscrits historiques, pourraient compléter la plateforme eScriptorium déjà mobilisée pour l'annotation et la correction manuelle (section III.5.1), en particulier pour accélérer la constitution de la vérité de terrain nécessaire à la transcription spécialisée du grec et de l'hébreu évoquée ci-dessus.

III.9 Conclusion

Ce troisième chapitre a mis à l'épreuve, sur des échantillons du corpus du Dictionnaire de Trévoux, les deux familles d'approches identifiées au chapitre 2 : une approche de bout en bout par modèle de vision-langage (section III.4.1) et une approche modulaire combinant segmentation et OCR spécialisé (section III.5.1).

Les résultats obtenus (section III.6) montrent un léger avantage à la chaîne modulaire une fois le modèle CATMuS-Print fine-tuné sur le corpus (CER de 0,74 % contre 0,99 % pour la meilleure configuration VLM), un écart qui, combiné à un coût de traitement nettement inférieur à grande échelle, a motivé le choix de cette approche pour la suite du stage.

Ce chapitre a également documenté empiriquement les défis typographiques propres au Trévoux (section III.3) : s long, ligatures, insertions grecques et hébraïques, hétérogénéité de numérisation et développé un sous-système dédié de détection de mots et de classification de style typographique (section III.7), capable de repérer les segments en grec et en hébreu que les moteurs OCR spécialisés testés ne savent pas transcrire. Ces résultats restent toutefois exploratoires, obtenus sur des échantillons de taille limitée plutôt que sur l'ensemble des 956 pages du corpus, et plusieurs pistes de poursuite en découlent (section III.8), notamment l'évaluation à grande échelle et la transcription spécialisée du grec et de l'hébreu.

Conclusion générale

Ce rapport de stage a présenté, en trois temps, l'état d'avancement du stage mené au LIRIS dans le cadre du projet OLR-LLM.

Le premier chapitre a présenté l'organisme d'accueil, le LIRIS, et notamment les équipes Imagine et DM2L à l'intersection desquelles se situe le stage, avant de situer le contexte et la problématique du stage : la structuration automatique du Dictionnaire de Trévoux, à l'articulation entre reconnaissance optique de caractères et interprétation par grands modèles de langage.

Le deuxième chapitre, cœur de ce rapport, a développé une étude bibliographique centrée sur les techniques effectivement testées sur des documents patrimoniaux, moteurs OCR spécialisés, détection de mise en page reformulée en détection d'objets, détection et classification du style typographique, structuration logique par règles expertes, correction post-OCR par LLM, en formalisant au passage les relations mathématiques (CER, WER, précision, rappel, F-mesure, IoU, mAP) qui sous-tendent leurs résultats chiffrés publiés.

Le troisième chapitre a présenté, de façon exploratoire, les approches concrètement testées sur le corpus, VLM et approche modulaire, avec leurs résultats quantitatifs et qualitatifs réels. Au vu de ces résultats, c'est l'**approche modulaire** qui a été retenue pour la suite du stage : elle se décompose en étapes indépendantes, ce qui la rend plus contrôlable qu'une approche VLM de bout en bout, et son coût de traitement à grande échelle est nettement inférieur à celui du VLM.
Les prochaines étapes du stage consisteront à étendre ces expérimentations à l'ensemble du corpus du Trévoux (956 pages), et à approfondir la transcription du grec et de l'hébreu au-delà du seul repérage positionnel.

Bibliographie

[1] LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information), UMR 5205 CNRS / INSA Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / École Centrale de Lyon. *Fiche de présentation du laboratoire*. `https://liris.cnrs.fr` (2022).

[2] LIRIS. *Effectifs du LIRIS*. `https://liris.cnrs.fr/presentation/effectifs-du-liris` (consulté en 2026).

[3] LIRIS. *Organigramme du LIRIS*. `https://liris.cnrs.fr/presentation/organigramme-liris` (consulté en 2026).

[4] Projet GEODE (LIRIS / ICAR / EVS, LabEX ASLAN). *GEODE, discours géographique encyclopédique*. `https://geode-project.github.io/` (consulté en 2026).

[5] Romanello, M., Najem-Meyer, S., Robertson, B. (2021). *Optical Character Recognition of 19th Century Classical Commentaries : the Current State of Affairs*. Proceedings of HIP '21, ACM.

[6] Levchenko, M. (2025). *Evaluating LLMs for Historical Document OCR : A Methodological Framework for Digital Humanities*. arXiv :2510.06743.

[7] Agbeti-Messan, M., Paquet, T., Chatelain, C., Tranouez, P., Nicolas, S. (2026). *A Benchmark of State-Space Models vs. Transformers and BiLSTM-based Models for Historical Newspaper OCR*. arXiv :2604.00725.

[8] Clérice, T. (2023). *You Actually Look Twice At it (YALTAi) : using an object detection approach instead of region segmentation within the Kraken engine*. Journal of Data Mining and Digital Humanities.

[9] 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.

[10] 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*. Proceedings of LaTeCH-CLfL 2024, 133 à 159.

[11] Thomas, A., Gaizauskas, R., Lu, H. (2024). *Leveraging LLMs for Post-OCR Correction of Historical Newspapers*. Proceedings of LT4HALA @ LREC-COLING 2024, 116 à 121.

[12] Morrissey, R. (1996). *The ARTFL Project*. D-Lib Magazine, janvier 1996.

[13] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). *Gradient-Based Learning Applied to Document Recognition*. Proceedings of the IEEE, 86(11), 2278 à 2324.

[14] Rennie, S. (2000). *Encoding a Historical Dictionary with the TEI (With Reference to the Electronic Scottish National Dictionary Project)*. Proceedings of EURALEX 2000, Stuttgart.

[15] Plamondon, R., Srihari, S. N. (2000). *On-Line and Off-Line Handwriting Recognition : A Comprehensive Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 63 à 84.

[16] Marinai, S., Gori, M., Soda, G. (2005). *Artificial Neural Networks for Document Analysis and Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(1), 23 à 35.

[17] Antonacopoulos, A., Pletschacher, S., Bridson, D., Papadopoulos, C. (2009). *ICDAR 2009 Page Segmentation Competition*. Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009).

[18] Graves, A., Schmidhuber, J. (2009). *Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks*. Advances in Neural Information Processing Systems (NeurIPS) 21.

[19] Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S. (2013). *ICDAR2013 Competition on Historical Newspaper Layout Analysis (HNLA 2013)*. Proceedings of ICDAR 2013.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS) 30.

[21] Kiessling, B. (2019). *Kraken, an Universal Text Recognizer for the Humanities*. Digital Humanities Conference 2019 (DH2019 « Complexities »), Utrecht University.

[22] Salgado, A., Costa, R., Tasovac, T., Simões, A. (2019). *TEI Lex-0 In Action : Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa*. Proceedings of eLex 2019, 417 à 433.

[23] Hagen, T., Ketzan, E., Jannidis, F., Witt, A. (2020). *Twenty-two Historical Encyclopedias Encoded in TEI : a New Resource for the Digital Humanities*. Proceedings of LaTeCH-CLfL 2020, Barcelone, 112 à 120.

[24] Kim, G., Hong, T., Yim, M. et al. (2022). *Donut : Document Understanding Transformer without OCR*. ECCV 2022.

[25] Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F. (2022). *LayoutLMv3 : Pre-training for Document AI with Unified Text and Image Masking*. arXiv :2204.08387.

[26] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J. (2023). *Qwen-VL : A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. arXiv :2308.12966.

[27] Clérice, T., Pinche, A., Vlachou-Efstathiou, M., Chagué, A., Camps, J.-B., Gille Levenson, M., Brisville-Fertin, O., Boschetti, F., Fischer, F., Gervers, M., Boutreux, A., Manton, A., Gabay, S., O'Connor, P., Haverals, W., Kestemont, M., Vandyck, C., Kiessling, B. (2024). *CATMuS Medieval : A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond*. Document Analysis and Recognition (ICDAR 2024), LNCS 14806, Springer, 174 à 194.

[28] Qwen Team, Alibaba Group (2025). *Qwen2.5-VL Technical Report*. arXiv :2502.13923.

[29] Mistral AI (2025). *Mistral OCR*. `https://mistral.ai/news/mistral-ocr/` (mars 2025).

[30] Duan, S., Xue, Y., Wang, W. et al. (2026). *GLM-OCR Technical Report*. arXiv :2603.10910.

[31] Agrawal, P., Antoniak, S., Bou Hanna, E. et al. (2024). *Pixtral 12B*. arXiv :2410.07073.

[32] Greif, G., Griesshaber, N., Greif, R. (2025). *Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents*. arXiv :2504.00414.

[33] Gutehrlé, N., Atanassova, I. (2022). *Processing the structure of documents : Logical Layout Analysis of historical newspapers in French*. Journal of Data Mining and Digital Humanities.

[34] Scius-Bertrand, A., Fakhari, A., Vögtlin, L., Cabral, D. R., Fischer, A. (2024). *Are Layout Analysis and OCR Still Useful for Document Information Extraction Using Foundation Models ?*. ICDAR 2024.

[35] Tensmeyer, C., Saunders, D., Martinez, T. (2017). *Convolutional Neural Networks for Font Classification*. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017). arXiv :1708.03669.

[36] Chaudhuri, B. B., Garain, U. (1998). *Automatic detection of italic, bold and all-capital words in document images*. Proceedings of the 14th International Conference on Pattern Recognition (ICPR 1998).

[37] Galleron, I., Williams, G. C. (2022). *Tenir la promesse du Dictionnaire universel : l'esprit encyclopédique d'Henri Basnage de Beauval*. Langue française, 214(2), 27 à 42.

[38] Lang, S. (2025). *Towards a Data-Driven History of Lexicography : Two Alchemical Dictionaries in TEI-XML*. Journal of Open Humanities Data, 11(1).

[39] Kiessling, B. (2025). *Version 5 of the Kraken ATR Engine for the Humanities*. Document Analysis and Recognition (ICDAR 2025), LNCS 16025, Springer, 443 à 458.

[40] Gabay, S., Clérice, T. (2024). *CATMuS-Print [Large]*, modèle Kraken de reconnaissance de texte imprimé occidental diachronique et multilingue. Zenodo. `https://zenodo.org/records/10592716`.

[41] Gabay, S., Clérice, T., Reul, C. (2023). *OCR17 : vérité de terrain et modèles pour les imprimés français du XVII^{e} siècle (voire un peu plus)*. Journal of Data Mining and Digital Humanities, 2023.

[42] Pinche, A., Stokes, P. A. (2024). *Historical Documents and Automatic Text Recognition : Introduction*. Journal of Data Mining and Digital Humanities.

[43] Yang, Y., Eli, E., Aysa, A., Ubul, K. (2025). *Script identification in multilingual environment : a survey in recent years*. Artificial Intelligence Review, 58(10), article 294.

[44] Fu, L., Yang, B., Kuang, Z., Song, J., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Huang, M. et al. (2025). *OCRBench v2 : An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning*. arXiv :2501.00321.

[45] Chung, Y. H. M., Choi, D. (2025). *Finetuning Vision-Language Models as OCR Systems for Low-Resource Languages : A Case Study of Manchu*. arXiv :2507.06761.

[46] Farazi, M., Alam, F., Maaradji, A., Maamar, Z., Mubarak, H., Zaghouani, W. (2026). *When Do VLMs Help Arabic Manuscript OCR ? A Cross-Dataset Study*. arXiv :2608.22366.

[47] François, M., Eglin, V., Biou, M. (2022). *Text Detection and Post-OCR Correction in Engineering Documents*. DAS 2022, LNCS 13237, Springer.

[48] Moncla, L., Joliveau, T., Vigier, D. (2024). *Propositions pour une étude interdisciplinaire de la géographie dans un dictionnaire universel et une encyclopédie du XVIIIe siècle*. 1er colloque du réseau METALEX.

[49] *Dictionnaire universel françois et latin, vulgairement appelé Dictionnaire de Trévoux, contenant la signification et la définition des mots de l'une et de l'autre langue....* 6e édition, Paris, 1743, 6 vol. in-folio. Corpus source de ce stage ; des exemplaires numérisés sont consultables à la Bibliothèque nationale de France (Gallica).

[50] École Pratique des Hautes Études (EPHE). *eScriptorium : plateforme de production de textes numériques (OCR/HTR)*. `https://www.sofer.info/` (consulté en 2026).

[51] Sakhovskiy, A. et al. (2026). *InkSight : Towards AI-Aided Historical Manuscript Analysis*. Proceedings of EACL 2026 (Demonstrations), 271 à 281.

[52] Gu, A., Dao, T. (2023). *Mamba : Linear-Time Sequence Modeling with Selective State Spaces*. arXiv :2312.00752.

Annexes

Annexe 1 : Glossaire

CATMuS-Print : Modèle Kraken de reconnaissance de texte imprimé occidental, entraîné de façon diachronique et multilingue selon des principes de *transcription graphématique* (voir cette entrée) [40].

CER / WER : *Character Error Rate* et *Word Error Rate* : taux d'erreur caractère et taux d'erreur mot, mesurant l'écart entre une transcription produite et la vérité terrain (équations 2 et 3).

CTC (*Connectionist Temporal Classification*) : Fonction de perte qui aligne une séquence d'images (une ligne de texte) avec une séquence de caractères sans exiger de segmentation préalable au caractère près ; utilisée notamment par Kraken et CATMuS-Print.

Entrée lexicographique : Ensemble des informations rattachées à un mot-vedette dans un dictionnaire (définitions, exemples, citations, variantes), qui en constitue l'unité de base.

Fine-tuning : Ré-entraînement, sur un jeu de données restreint et spécifique à une tâche ou un corpus, d'un modèle déjà pré-entraîné sur un corpus plus large et généraliste.

Gouttière : Bord intérieur d'une page reliée, du côté de la reliure ; source fréquente de déformation de l'image lors de la numérisation (figure 17).

Grec polytonique : Grec comportant plusieurs signes diacritiques (accents aigu, grave, circonflexe, esprits doux/rude) hérités de la tradition éditoriale grecque, par opposition au grec moderne monotonique qui les a abandonnés.

IoU (*Intersection over Union*) : Indice de Jaccard mesurant le recouvrement entre une région détectée et la région de vérité terrain correspondante (équation 6) ; sert de base au calcul du mAP.

Ligature : Caractère unique représentant graphiquement deux lettres fusionnées (par exemple « œ » pour « oe »), fréquent dans la typographie ancienne.

LoRA (*Low-Rank Adaptation*) : Méthode de fine-tuning efficace en paramètres (PEFT) : au lieu de réentraîner tous les poids du modèle, elle insère et n'entraîne que de petites matrices de rang réduit, ce qui réduit fortement la mémoire GPU nécessaire.

mAP (*mean Average Precision*) : Moyenne, sur l'ensemble des classes de zones considérées, de la précision moyenne d'un détecteur de mise en page (équation 8).

Manicule : Symbole typographique en forme de main pointeuse (index tendu), utilisé dans la marge des imprimés anciens pour signaler un passage notable.

Mot-vedette (ou vedette) : Forme sous laquelle une entrée est répertoriée et mise en évidence typographiquement (ici en petites capitales) en tête d'article.

OCR (*Optical Character Recognition*) : Reconnaissance optique de caractères : conversion de l'image d'un texte en caractères numériques éditables.

OLR (*Optical Layout Recognition*) : Reconnaissance optique de la mise en page : détection et classification des zones structurelles d'une page (titre, colonne, illustration, etc.), indépendamment du texte qu'elles contiennent.

s long : Forme historique de la lettre « s » utilisée systématiquement en position non finale dans la typographie du XVIII^{e} siècle, dont la forme graphique proche du *f* et du *s* rond est une source fréquente de confusion pour l'OCR.

TEI (*Text Encoding Initiative*) / TEI Lex-0 : Norme XML de référence pour l'encodage de textes numériques en humanités numériques ; TEI Lex-0 en est le sous-schéma dédié aux ressources lexicographiques (dictionnaires, glossaires), format de sortie visé à terme pour la structuration du Dictionnaire de Trévoux.

Transcription graphématique : Transcription qui restitue fidèlement la graphie d'origine du document (formes historiques des lettres, ponctuation, abréviations non résolues), par opposition à une transcription normalisée qui régulariserait ces usages vers l'orthographe moderne.

VLM (*Vision-Language Model*) : Modèle de vision-langage traitant conjointement l'image et le texte, ici utilisé en bout en bout pour reconnaître et structurer une page sans étape de segmentation distincte (par exemple Qwen2.5-VL, Mistral OCR).

Annexe 2 : Outils d'annotation utilisés : Label Studio et eScriptorium

Deux plateformes d'annotation web ont été mobilisées durant le stage pour constituer les vérités de terrain nécessaires à l'entraînement et à l'évaluation des différents modèles.

Label Studio

Sert à l'annotation des tâches de classification, pour lesquelles eScriptorium n'est pas adapté. Deux projets distincts y ont été menés : l'un pour la segmentation de mise en page destinée à l'entraînement de DocLayout-YOLO (section III.5.2), l'autre pour la classification de style typographique destinée à SmallCNN et au fine-tuning LoRA de Qwen2.5-VL (sections III.7.2 et III.7.3).

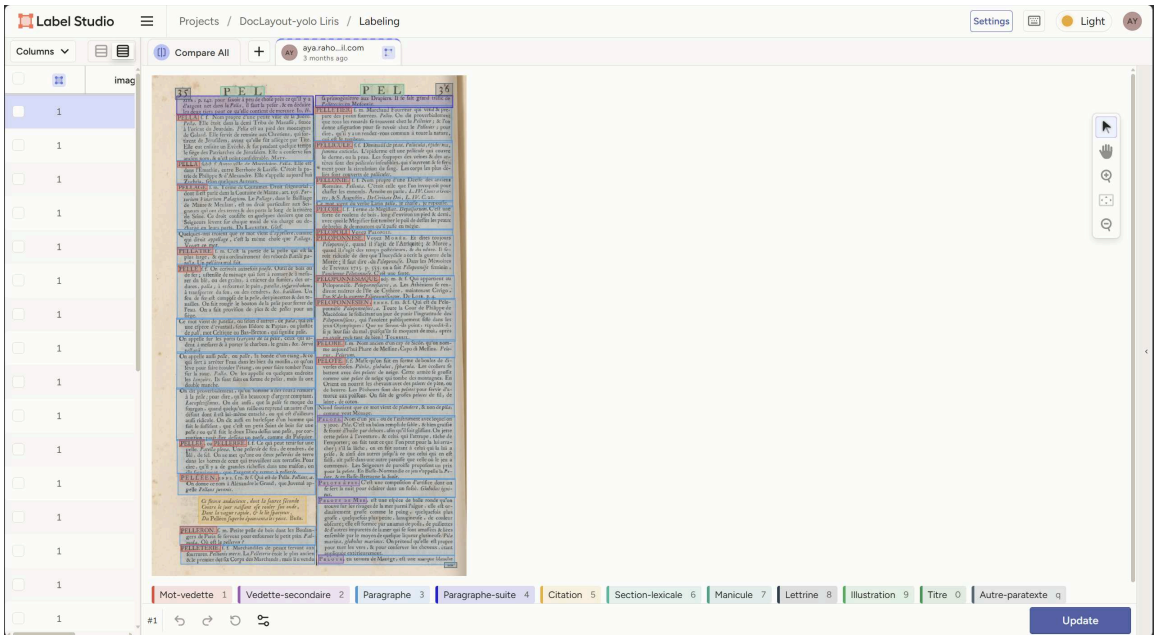


FIGURE 48 – Interface Label Studio : annotation de la mise en page (mot-vedette, vedette secondaire, paragraphe, citation, section lexicale, manicule, lettrine, illustration, titre, autre paratexte) pour l'entraînement de DocLayout-YOLO.

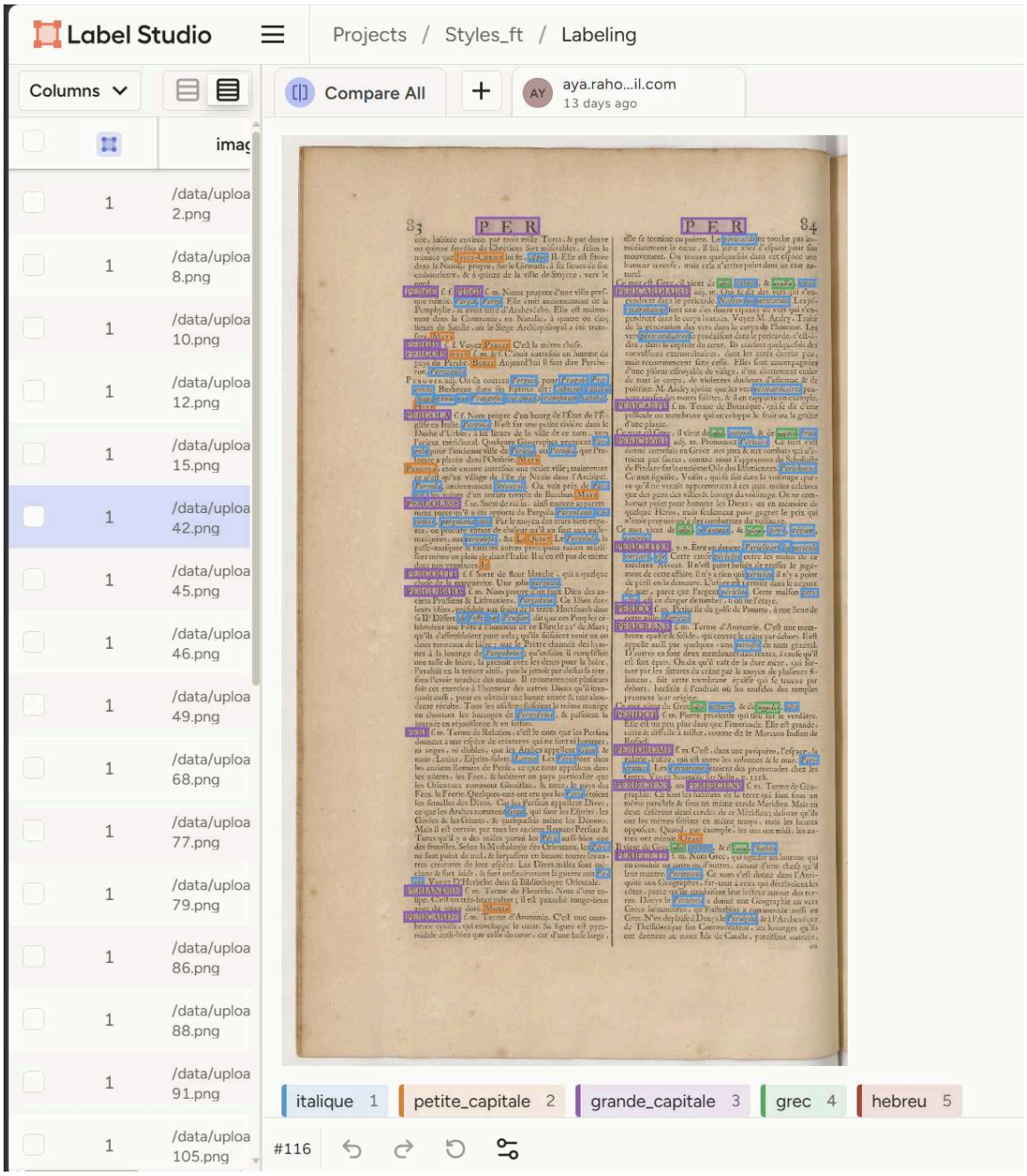


FIGURE 49 – Interface Label Studio : annotation du style typographique par mot (italique, petite capitale, grande capitale, grec, hébreu).

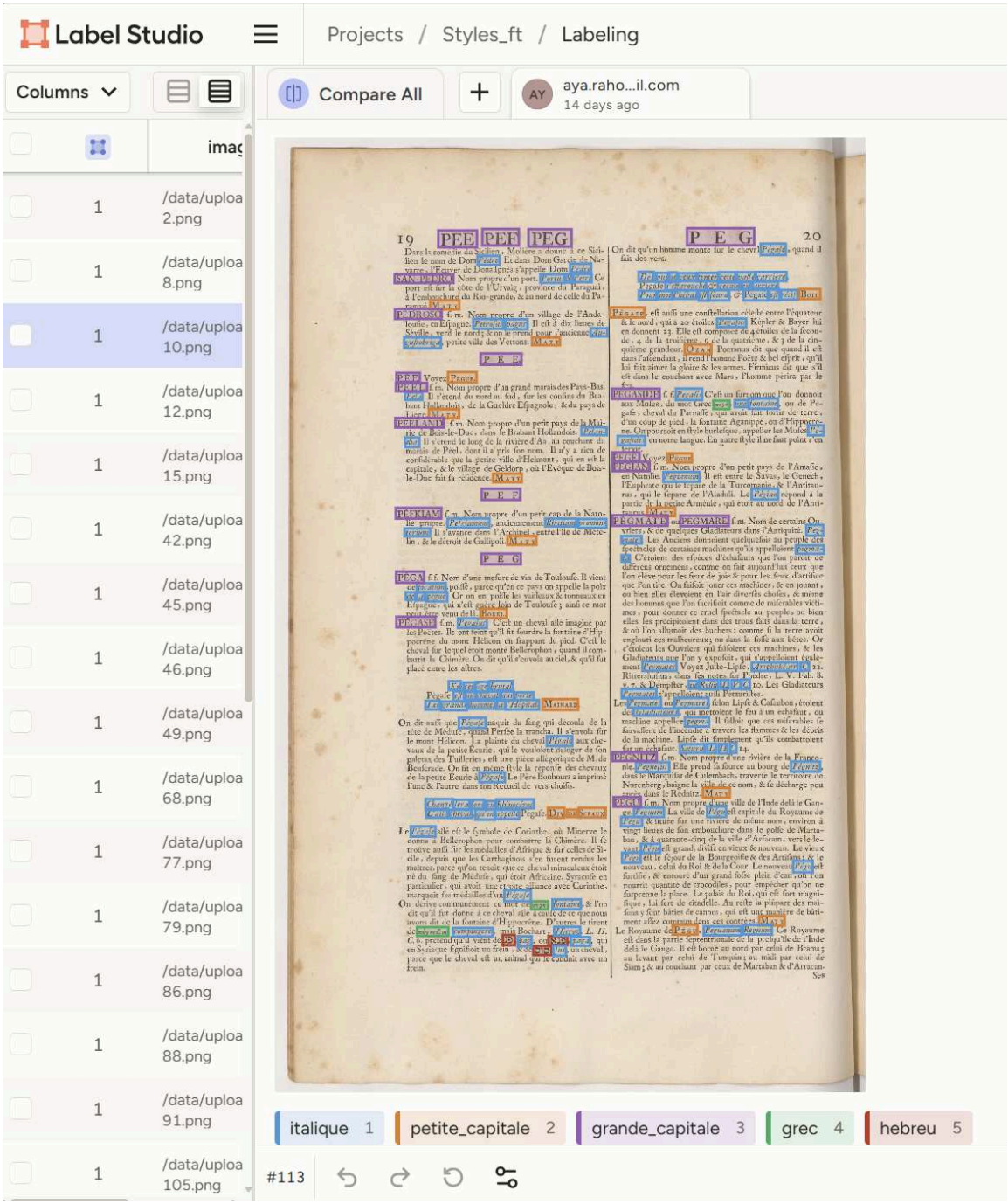


FIGURE 50 – Interface Label Studio : annotation du style typographique sur une autre page du corpus.

eScriptorium

Déjà présenté en section III.5.1 [50], a été utilisé pour la segmentation de ligne et la transcription du corpus avec le moteur Kraken. La figure 51 montre la vue de segmentation, avec les lignes de base détectées ; la figure 52 montre la vue de transcription correspondante, une fois le texte reconnu associé à chaque ligne.

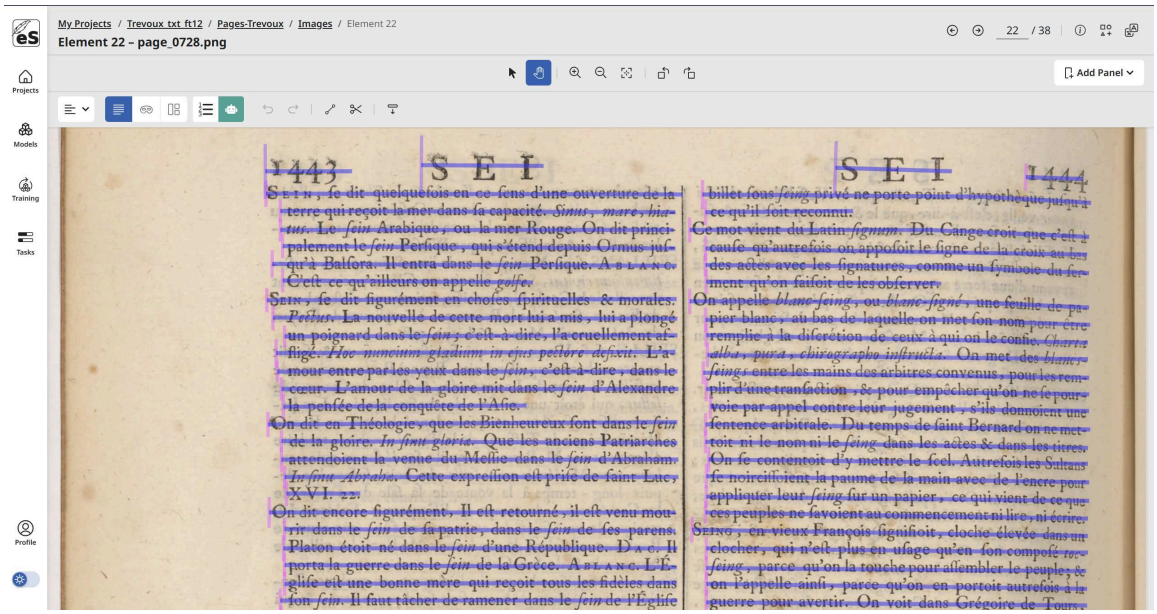


FIGURE 51 – Interface eScriptorium : vue de segmentation de ligne sur une page du corpus.

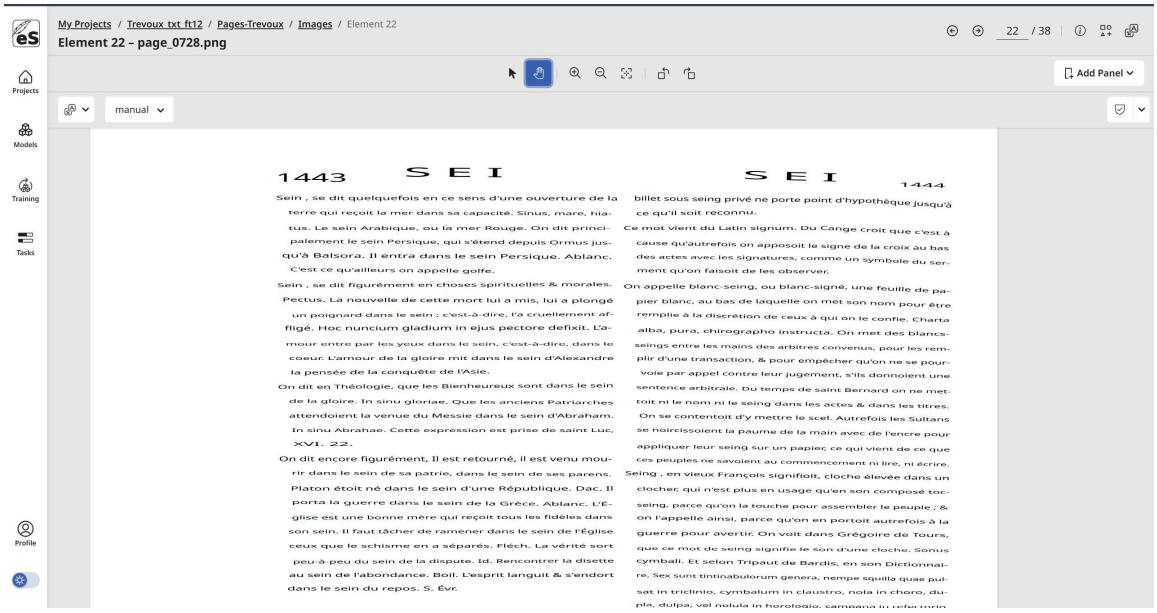


FIGURE 52 – Interface eScriptorium : vue de transcription correspondante, texte reconnu ligne à ligne.