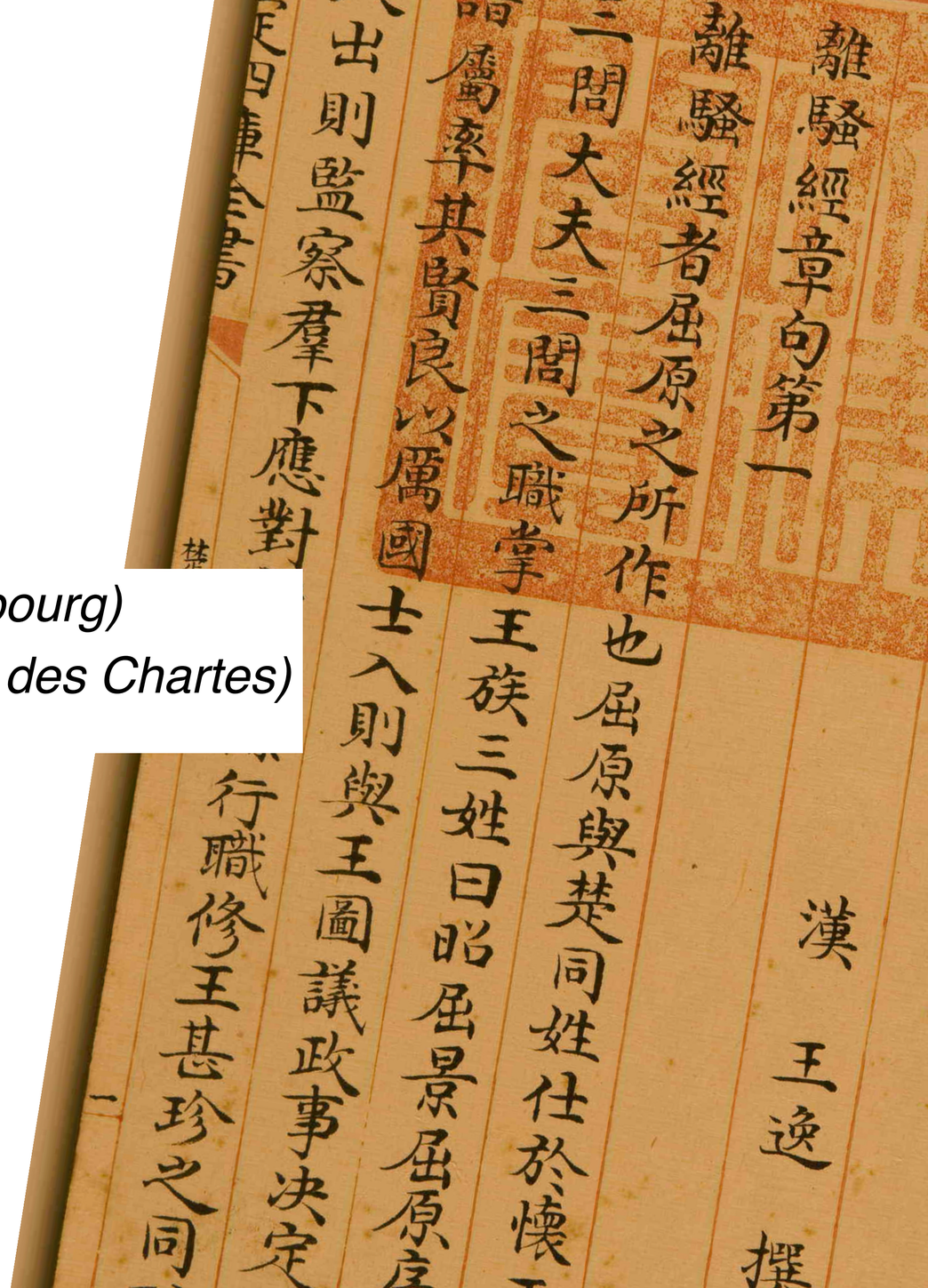
Expérimentations pour l'analyse automatique de sources chinoises anciennes

Marie Bizais-Lillig *(Université de Strasbourg)*
Chahan Vidal-Gorène *(École nationale des Chartes)*



Cadre des expérimentations

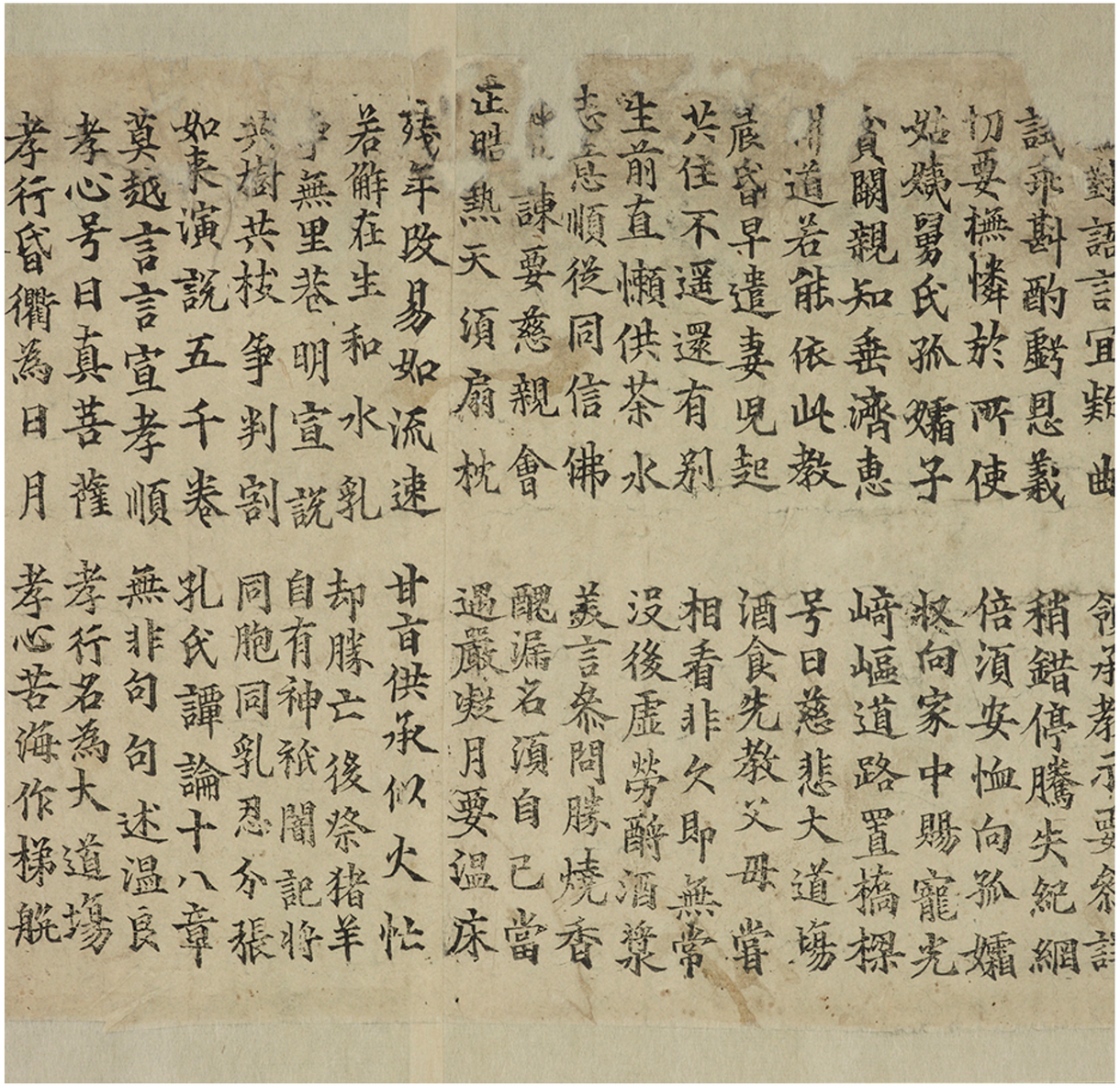
- Projet de fouille de textes chinois, pour l'essentiel du Moyen âge (IIIe-Xe s.) : CHI-KNOW-PO
- Nécessite :
    - l'acquisition du corpus dans son intégralité et sans restriction de droits
    - la structuration du corpus (dimension éditoriale)
- Mai-juin 2021 : début d'une collaboration entre
    - des sinologues (de la Chine impériale)
    - l'équipe de Calfa pour l'HTR
- Nos craintes : les limites que poserait le nombre de classes

Plan de l’intervention

- Spécificités du corpus
- Les premières expérimentations
- La question des classes

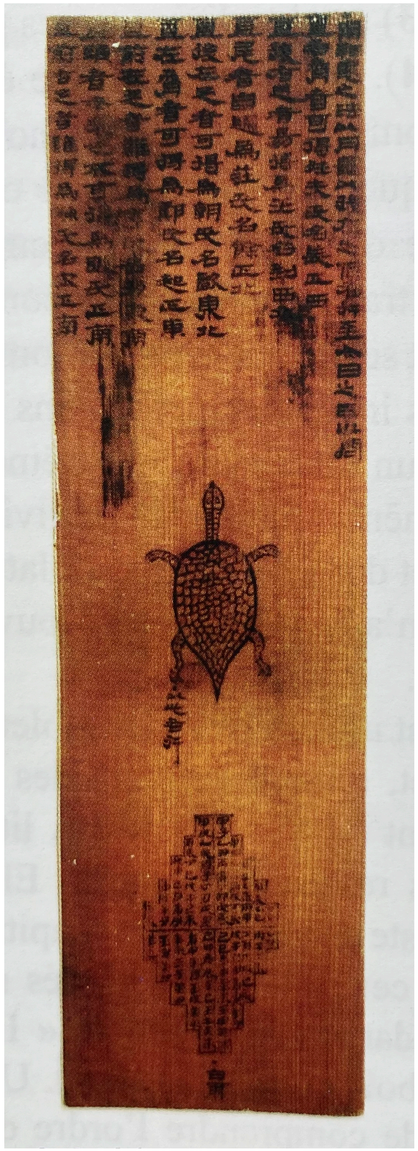
Spécificités du corpus - des textes xylographiés

- Des textes établis entre 100 AE et 1000 NE
- Support : papier
- La xylographie au service de la conservation et de la dissémination
- Xylographie *versus* caractères mobiles



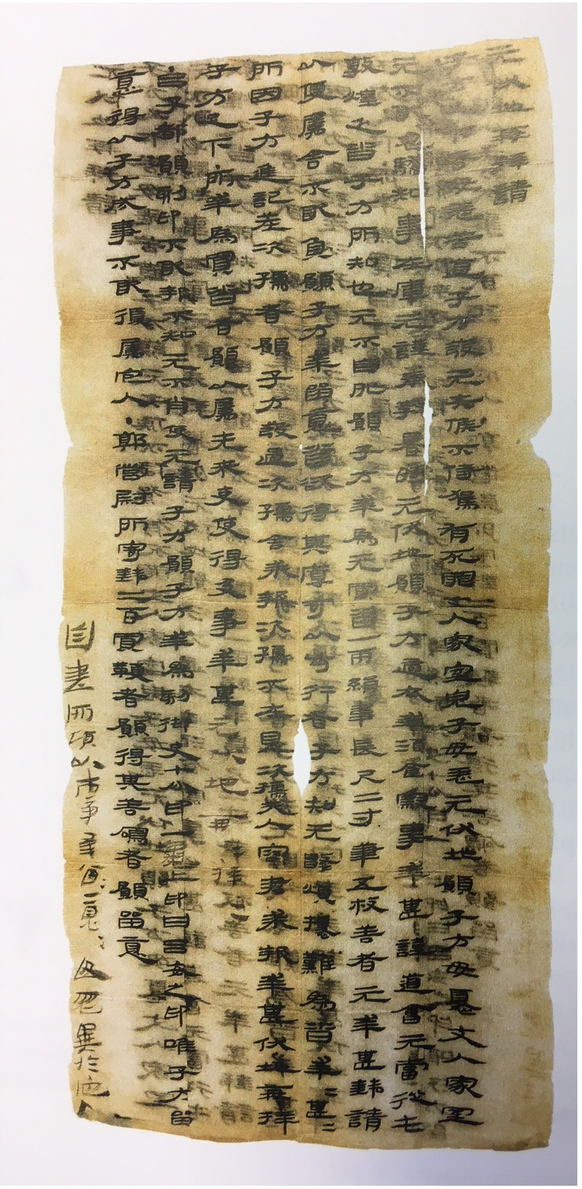
Impression xylographique (Dunhuang - British Library - OR8210)

Spécificités du corpus



plaquette de bois (*ca*. 10 NE - Musée de Lianyungang)
source : *La Fabrique du lisible*

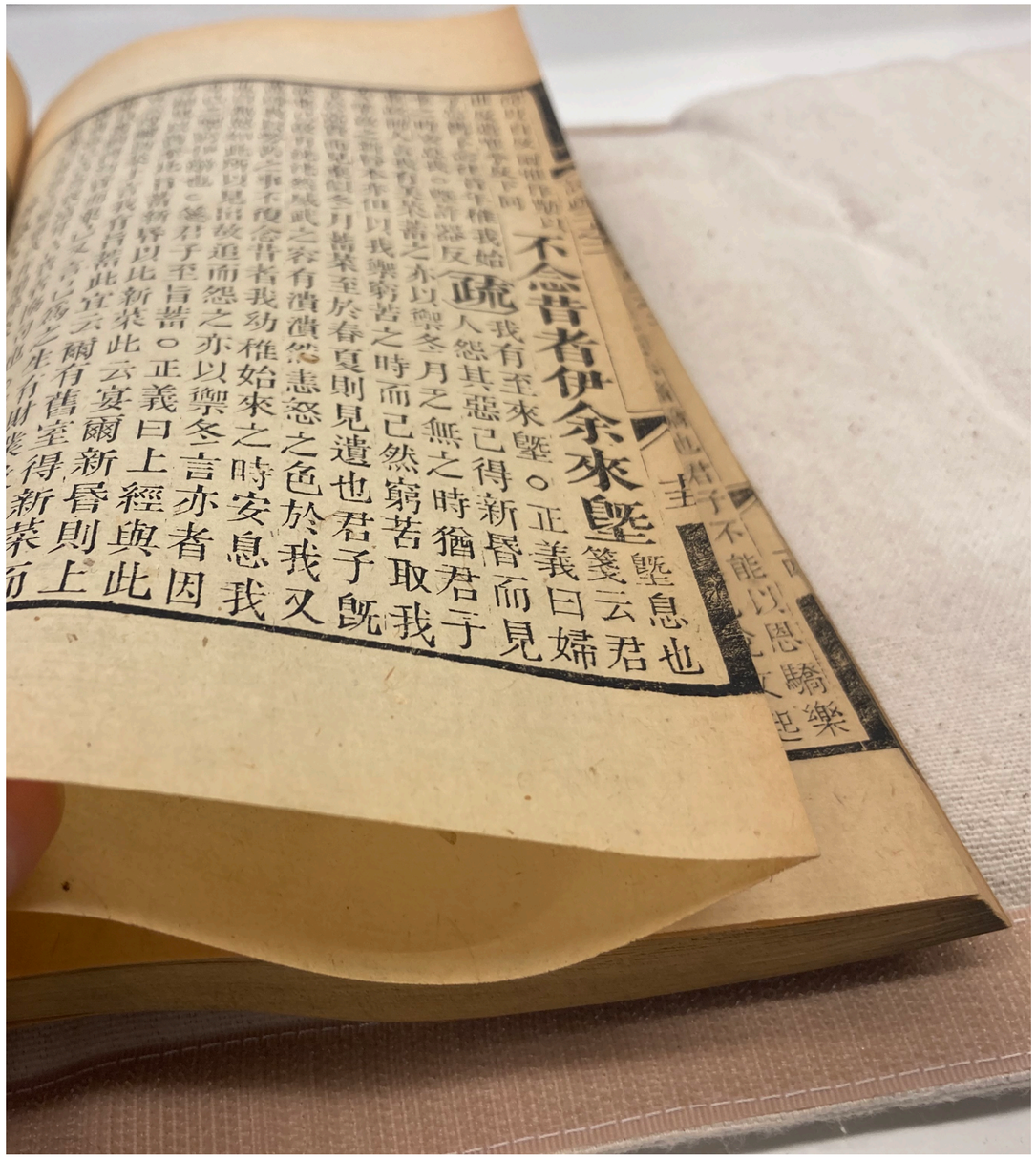
lamelles de bambou (*Shijing* 詩經 - ca. 250 AE - Université de l'Anhui)
source : 安徽大学藏战国竹简（一）



soie (lettre - Ier s. AE - Dunhuang)
source : *La Fabrique du lisible*

Spécificités du corpus - mise en page

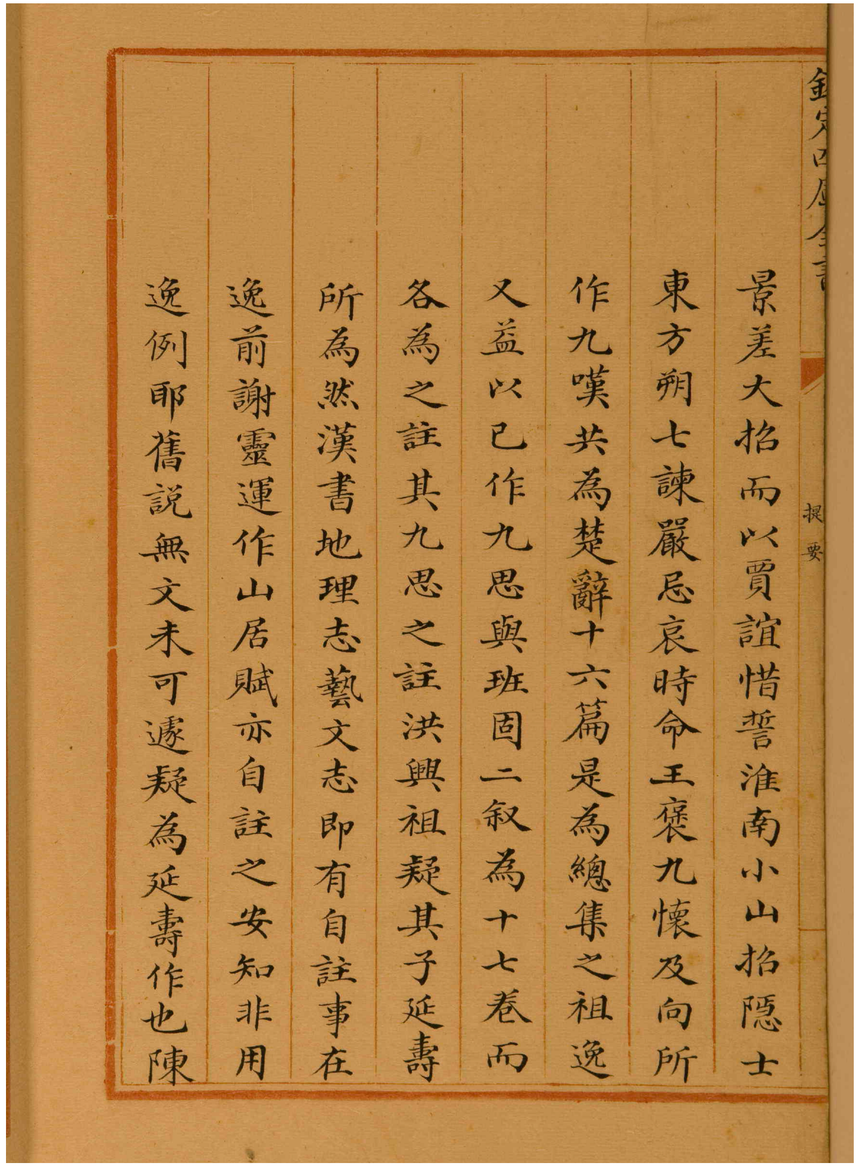
- Des textes transmis de pair avec un ou plusieurs commentaires
- Mise en page : entrelacement (à partir du IIe s. NE)
- Impression sur feuillets longs
  - pliés en accordéon
  - reproduits avec 2, 4 ou 6 double-pages par page (densité)



Reliure « accordéon » (BULAC)

Spécificités du corpus

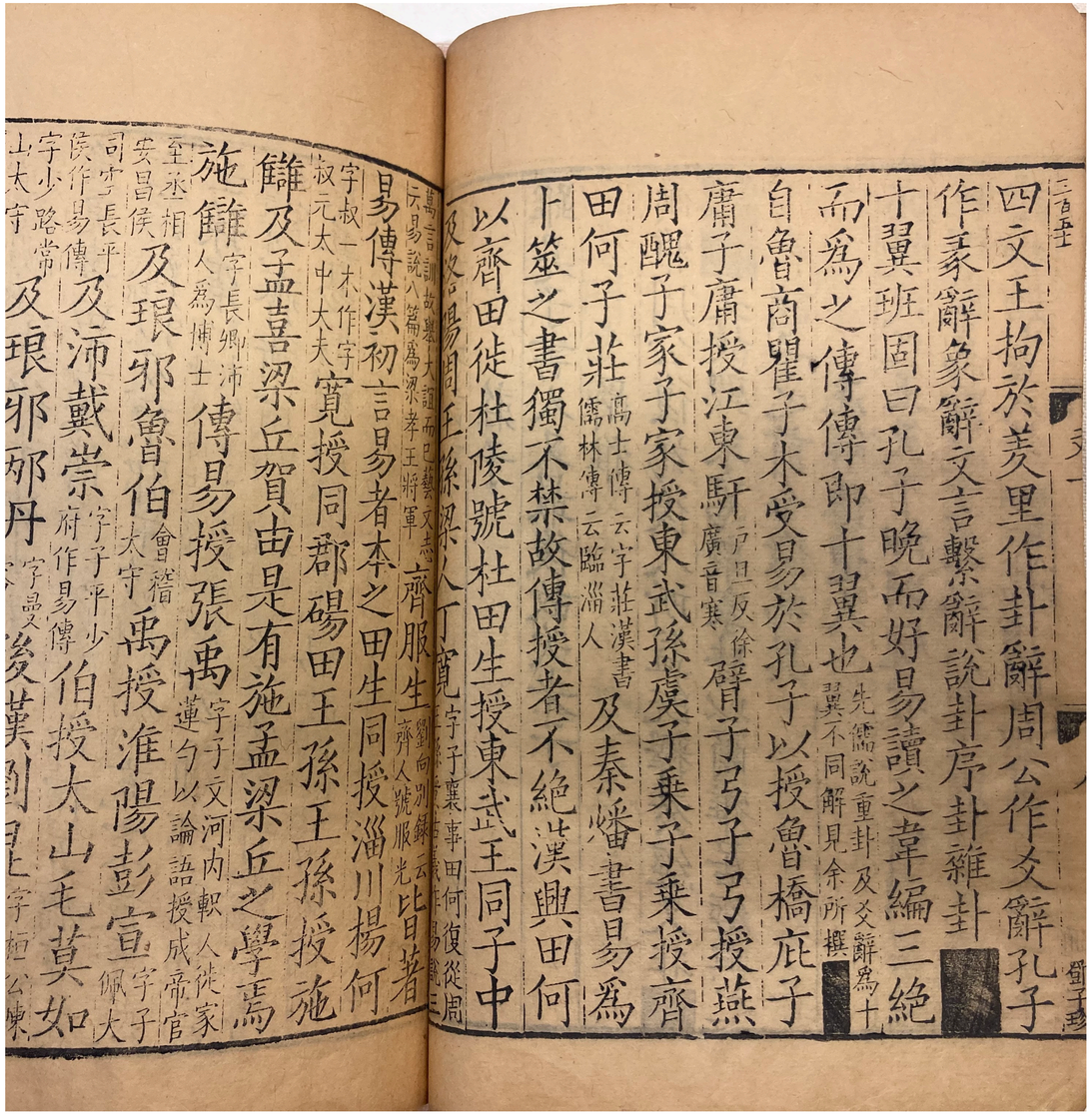
Entrelacement, découpage
爾雅 édition xylographiée illustrée (1801)
BULAC CHI1938(2)



Édition impériale manuscrite
楚辭 dans le *Siku*

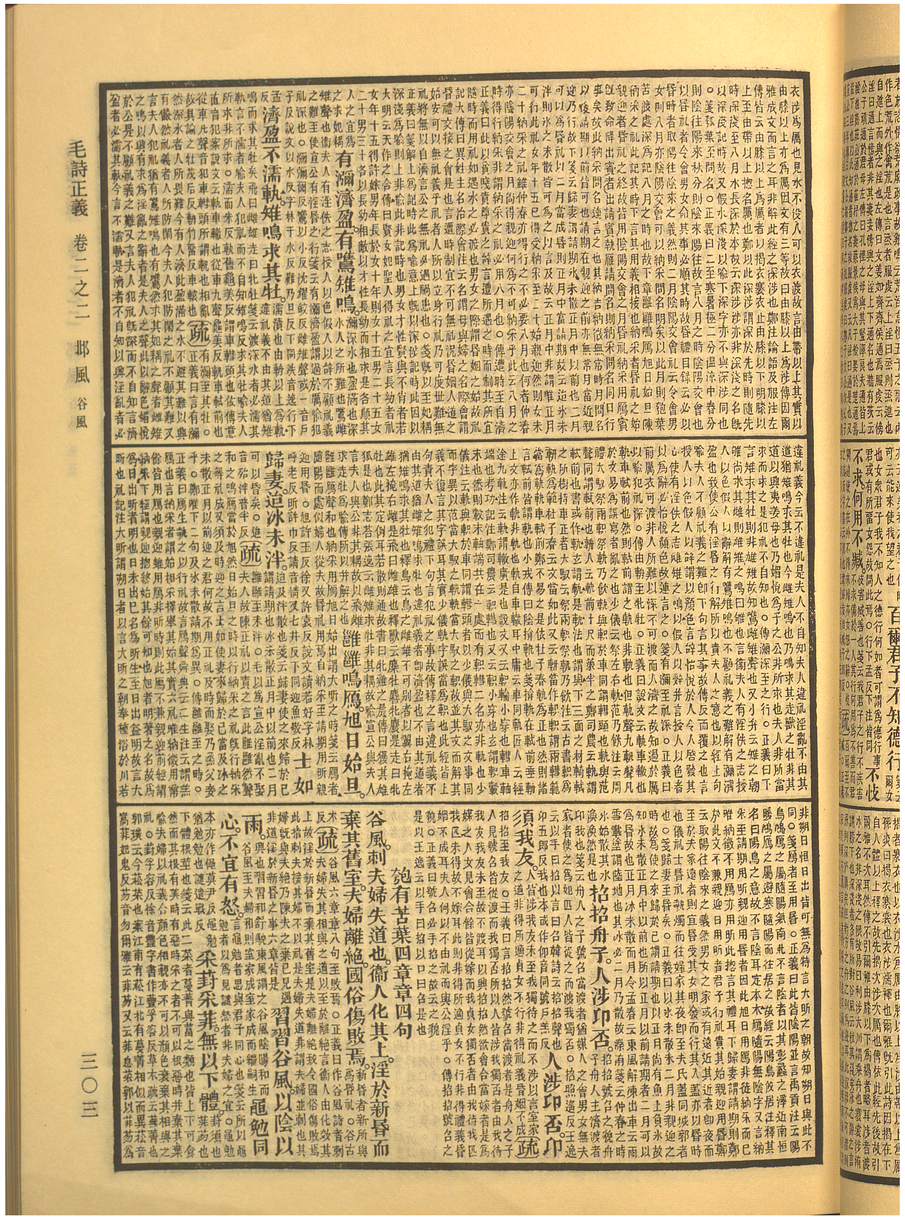
Spécificités du corpus - l’écriture chinoise

- Un lexique s’appuyant sur plus de 54 000 caractères (glyphes) différents
- Dans les textes anciens, la variété des caractères est grande
- Corpus de départ : > 7000 glyphes distincts
- En outre : variétés de styles / graphies (avec des équivalents unicodes distincts)
- Des caractères tabous sous chaque empereur

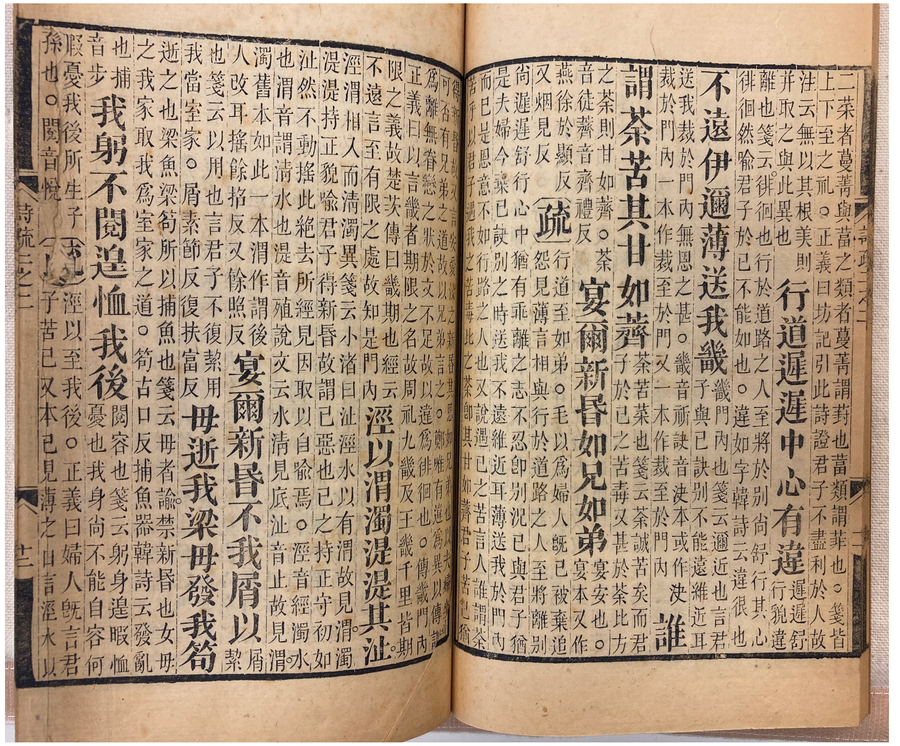


經典釋文 édition de 1800
BULAC CHI567

Spécificités du corpus - le *Classique des Poèmes* xylographié



Édition xylographiée du *Shijing* utilisée
reproduction par la Shanghai guji chubanshe



Édition XIIe, impression
1815 (BULAC-CHI518(3))

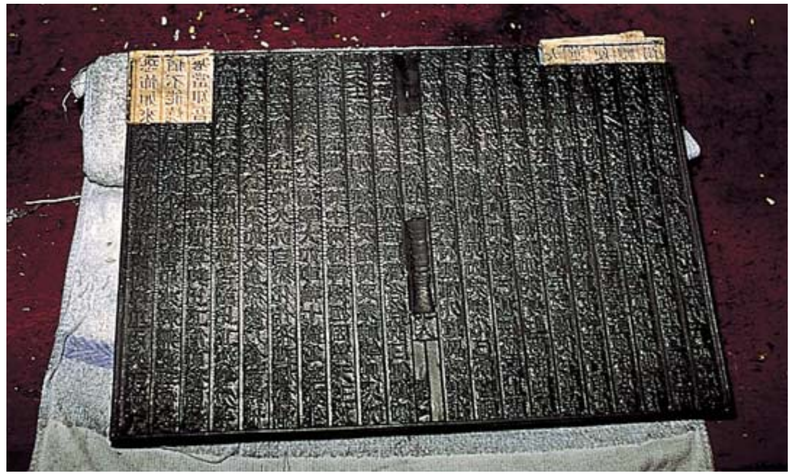


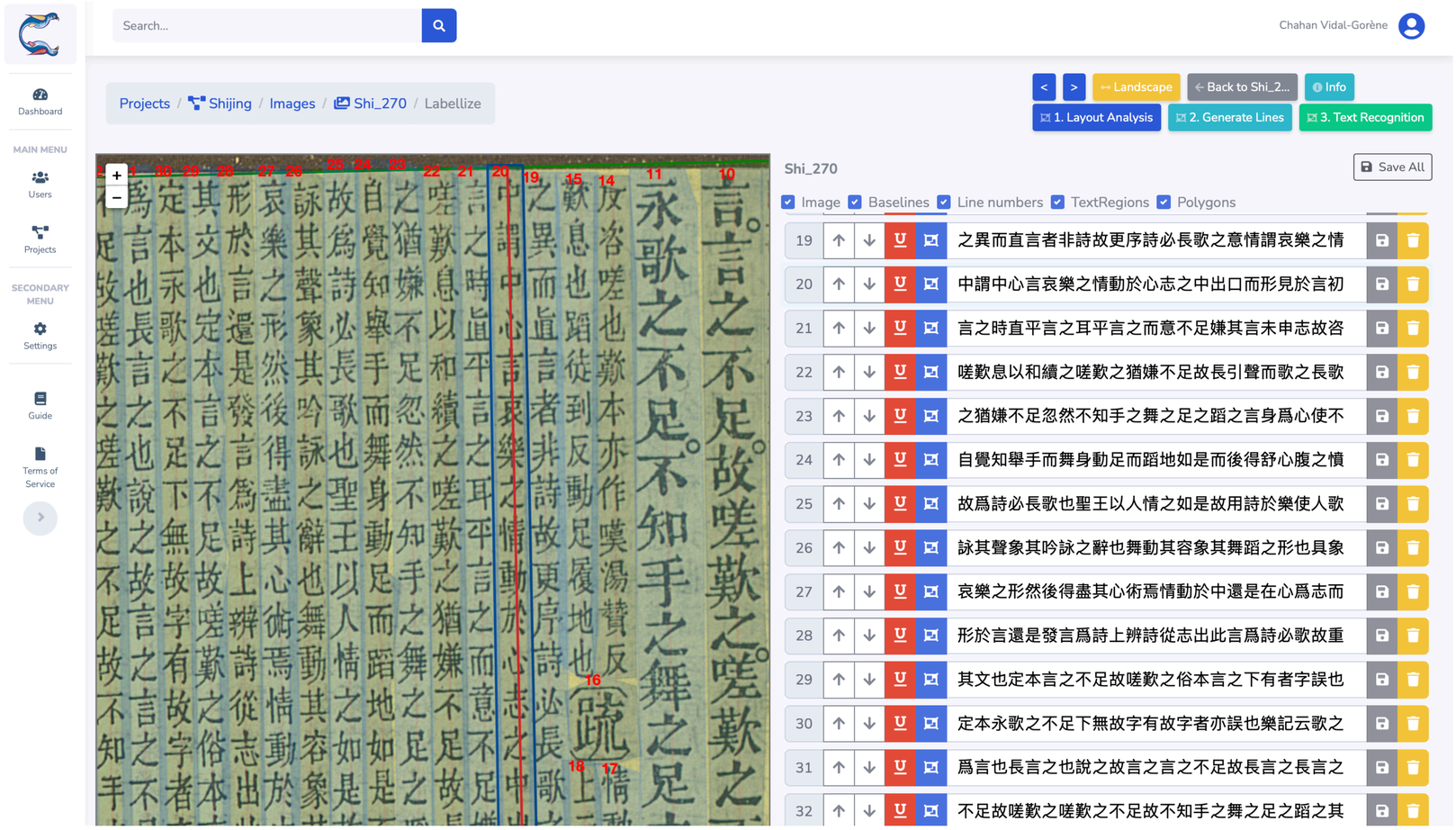
Planche à xylographier
(à partir des IXe-Xe s.)

Les premières expérimentations - généralités

Corpus annoté de 50 images pour les expérimentations :

- 3 240 lignes en apprentissage
- 92 234 caractères

Approche par « baseline » initialement favorisée pour faciliter l'échange de ces données



Approche itérative sur la plateforme Calfa Vision : analyse automatique puis vérification manuelle

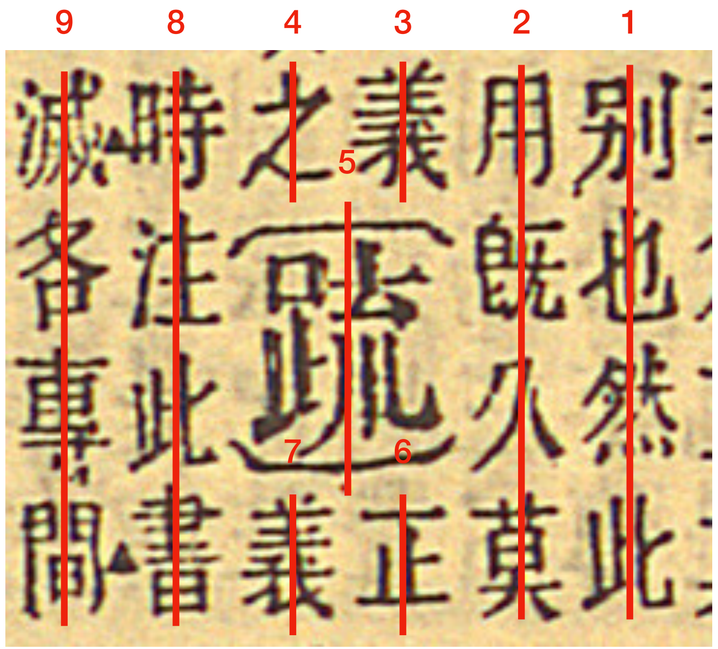
=> https://vision.calfa.fr

=> spécialisation rapide de la plateforme sur la détection de la mise en page (régions et lignes)

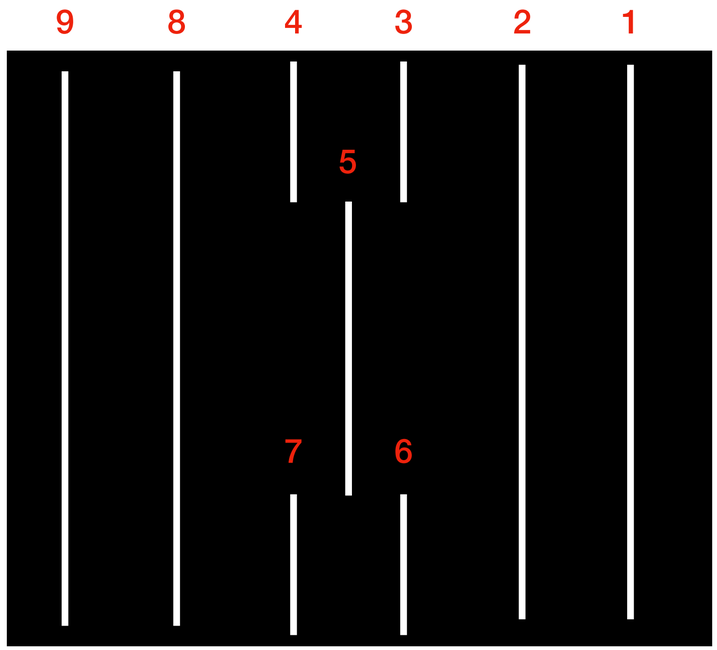
Les premières expérimentations - l’ordre de lecture

Approche de base : tri des lignes avec le centroïde, solution de traitement d’image basique

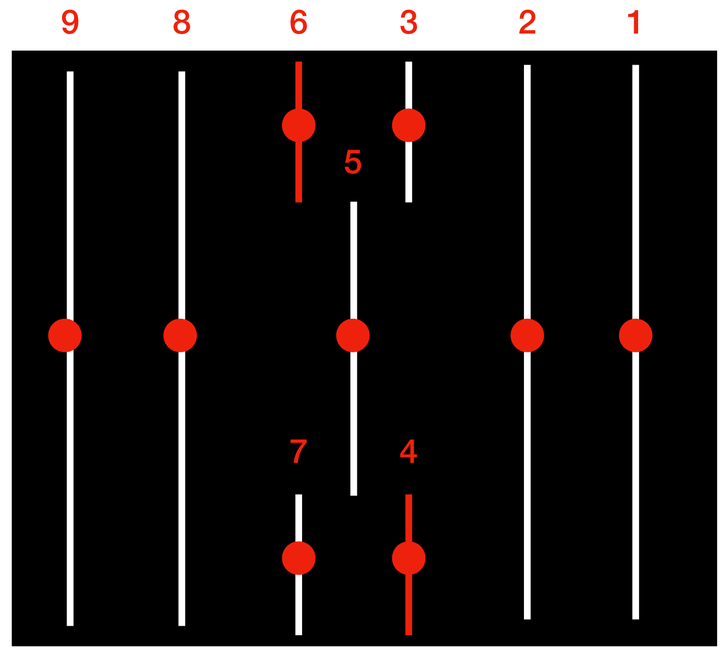
Efficace à **82,37**% sur le dataset, échoue y compris en situation simple



Détection des lignes



Sens de lecture GT

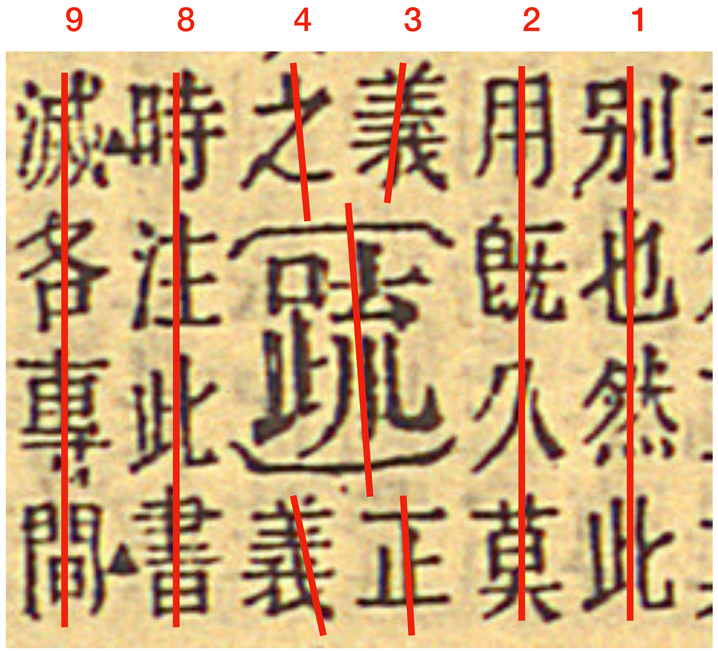


Tri des lignes par centroïdes

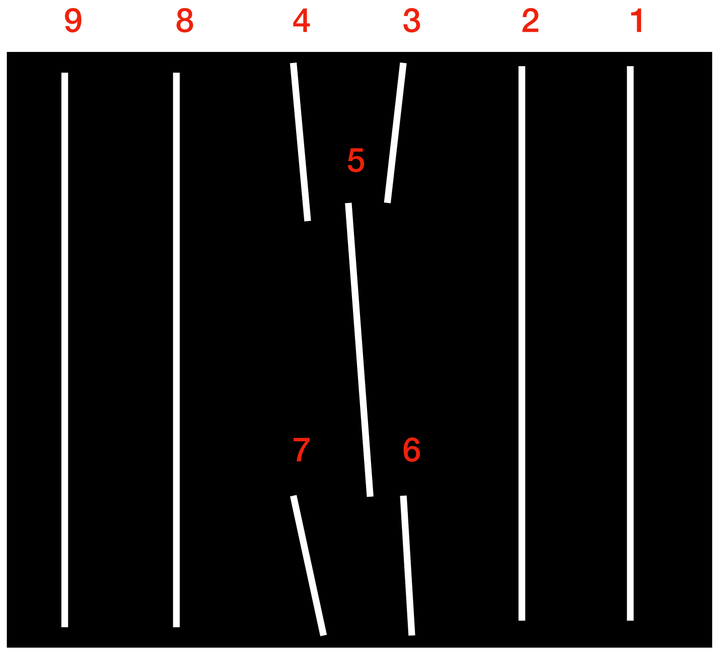
Les premières expérimentations - l’ordre de lecture

Approche de base : tri des lignes avec le centroïde, solution de traitement d’image basique

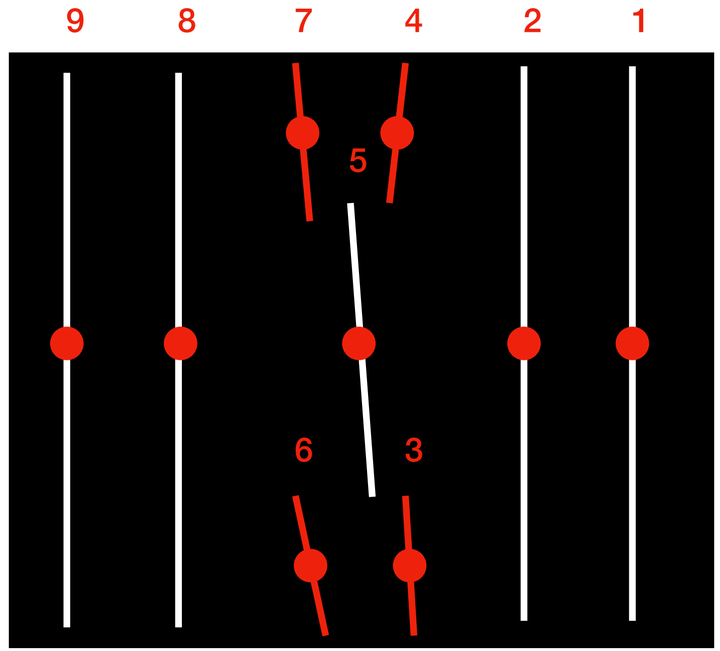
Efficace à **82,37**% sur le dataset



Détection des lignes

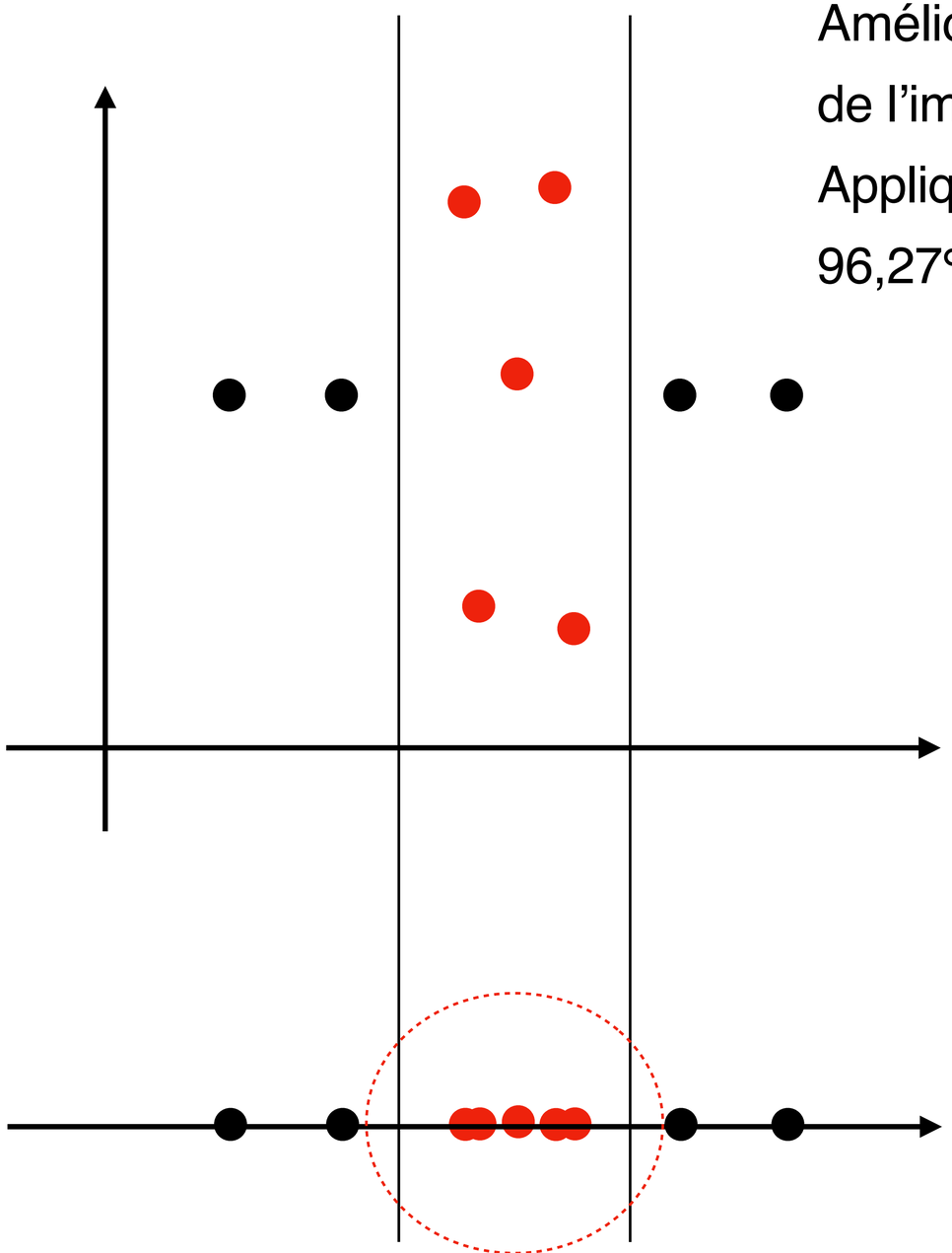


Sens de lecture GT



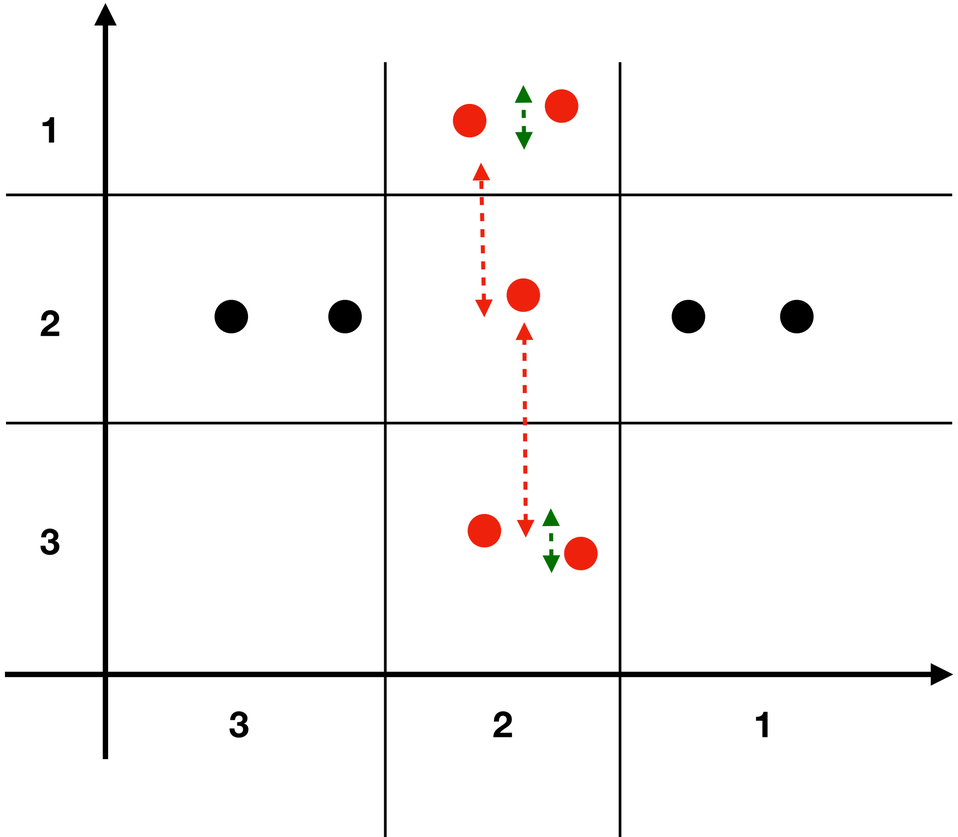
Tri des lignes par centroïdes

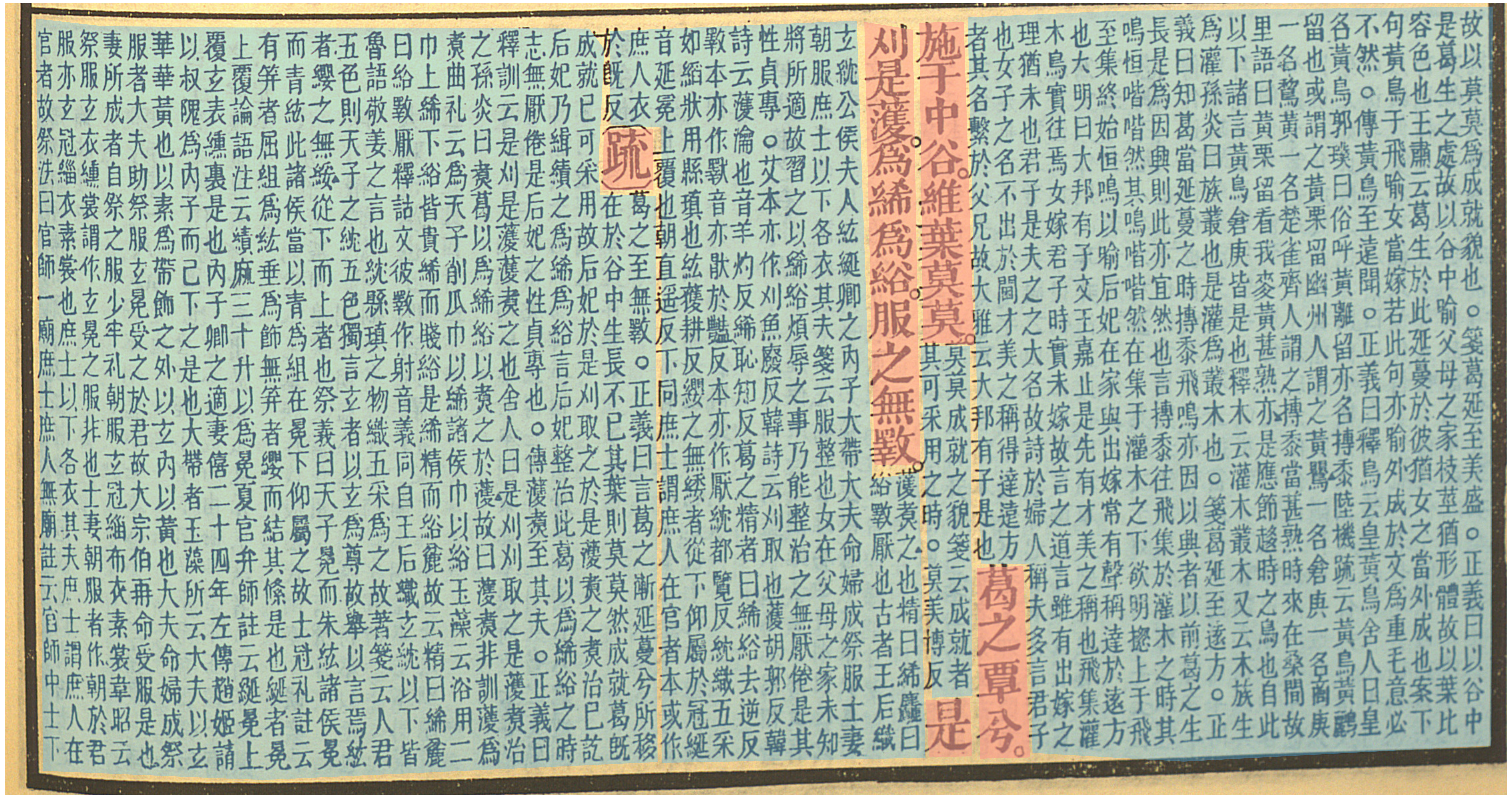
Les premières expérimentations - l’ordre de lecture



Amélioration du tri avec une solution de traitement de l’image : projection des centroïdes et identification d’un cluster : 87,88%

Appliquée directement aux bounding-box vs baseline : de 91,31% à 96,27%

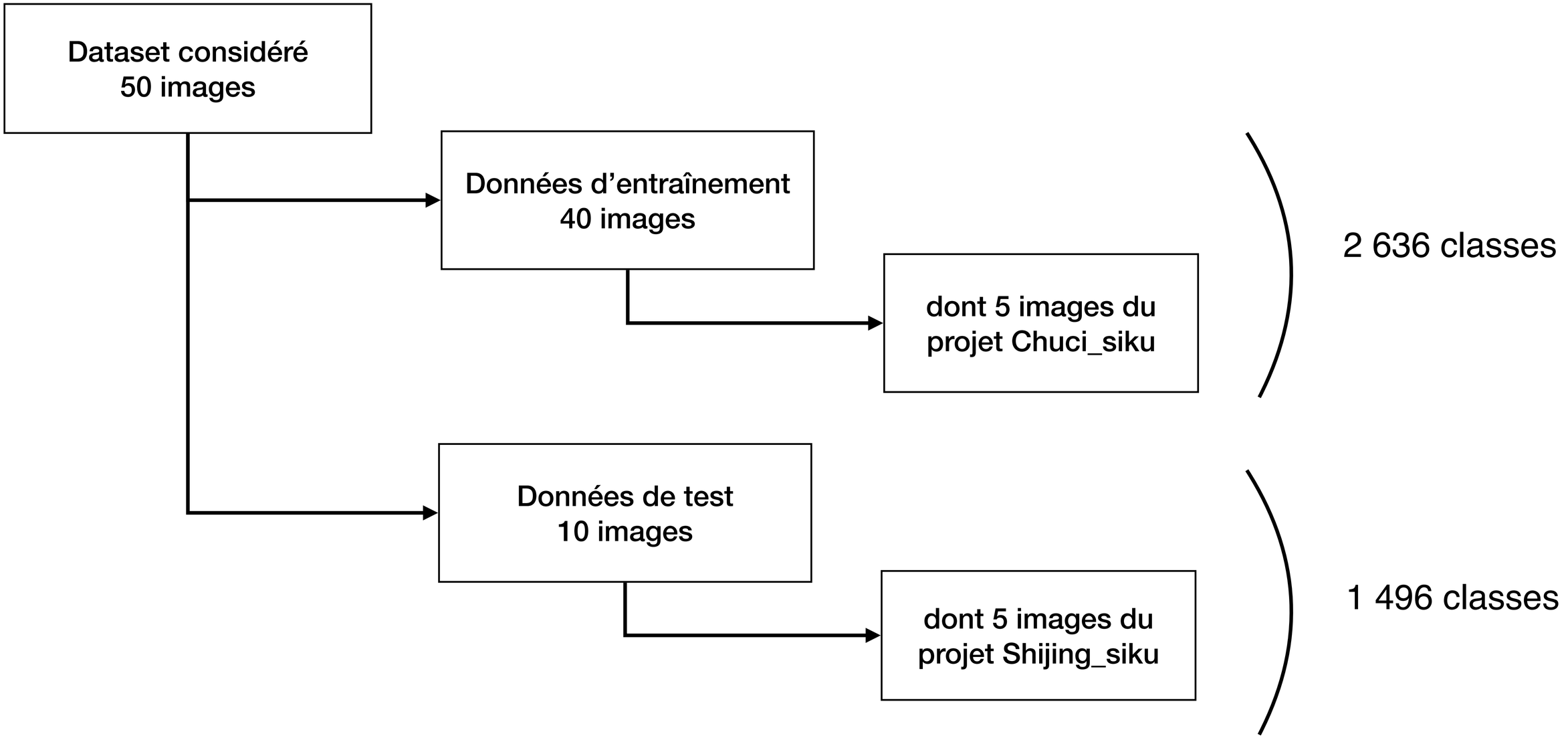




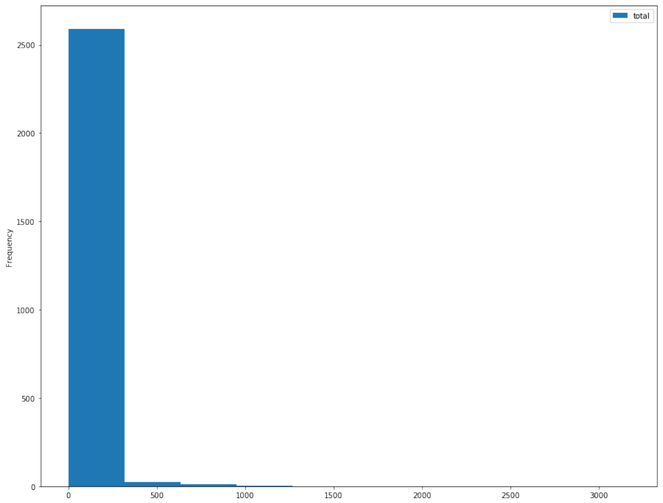
Solution en cours d’expérimentation : pré-classification sémantique au sein d’une région de texte puis tri indépendamment des lignes pour chaque zone identifiée. Abandon des baselines. Score obtenu sur un échantillon manuellement annoté, sans apprentissage : 99,67%

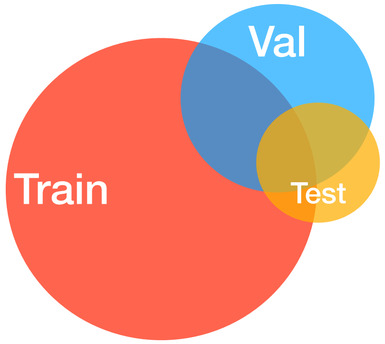
Les premières expérimentations - la question des classes

Création d’un modèle de zéro, pas de données disponibles pour un *fine-tuning*



Les premières expérimentations - la question des classes





caractères les plus fréquents / les moins fréquents

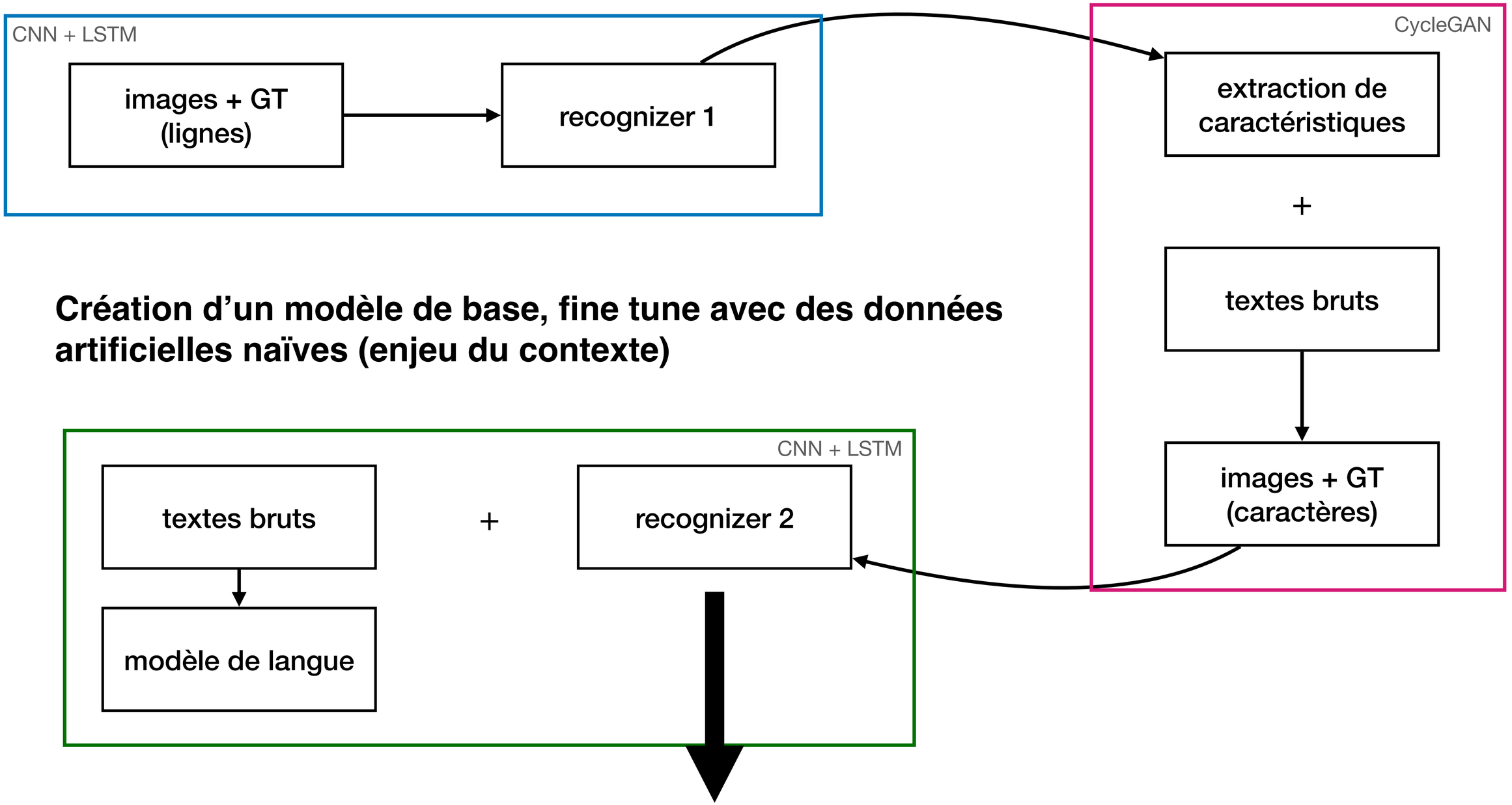
之	3170			芋	1
也	2032			恕	1
以	1463			刃	1
云	1110		…	訪	1
不	1074			隈	1
言	960			沔	1
者	944			斐	1

2 095 classes sur 2636 avec un seul échantillon

Seulement 1269 classes communes en moyenne entre les ensembles d’apprentissage et de test

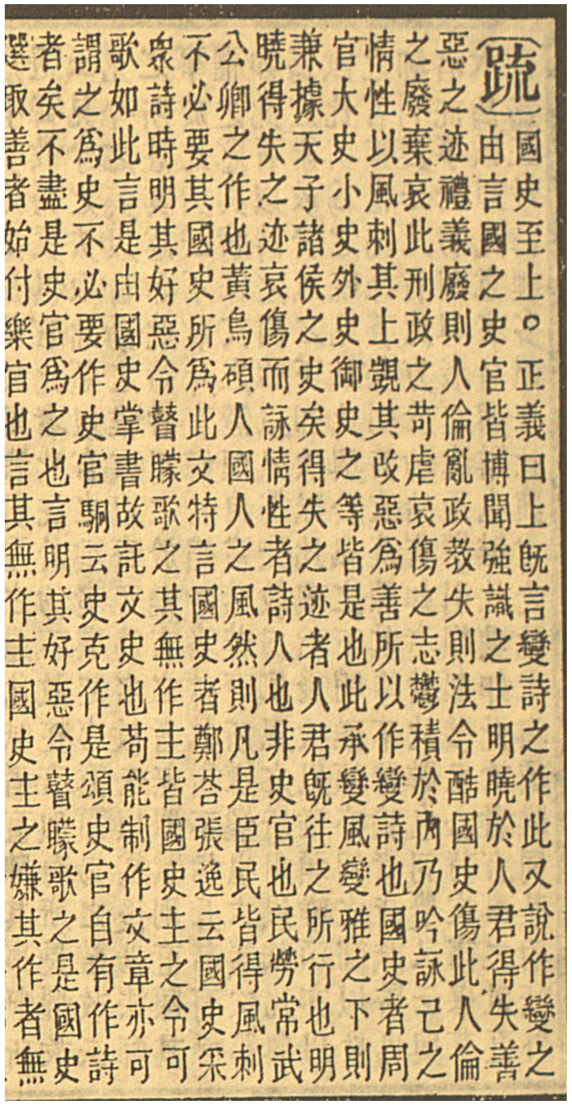
==> Situation de *one-shot learning* + classes non vues en apprentissage

Les premières expérimentations - la question des classes



Création d'un modèle de base, fine tune avec des données artificielles naïves (enjeu du contexte)

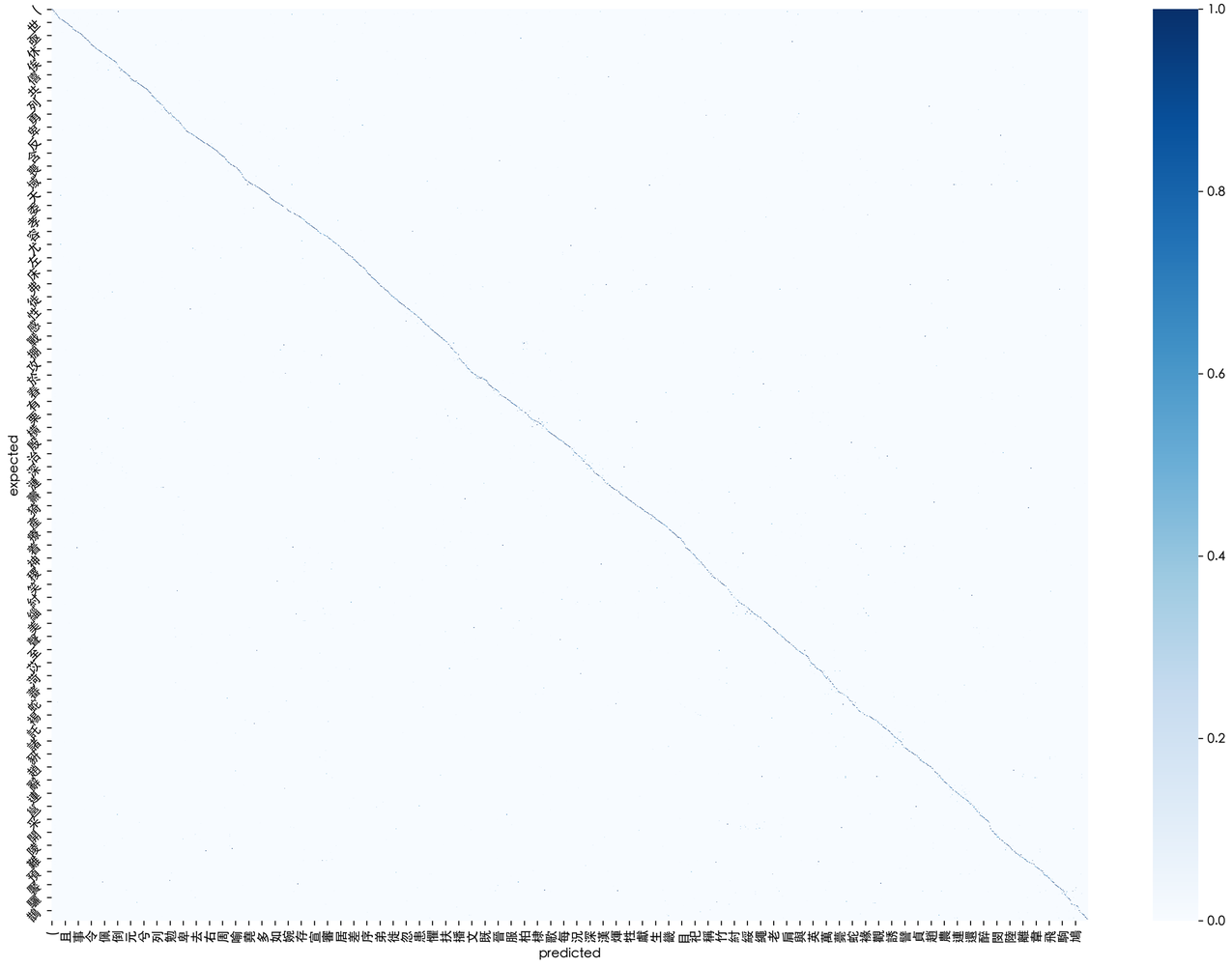
Les premières expérimentations - la question des classes



二國史至上*正義曰上既言變詩之作此又說作變之
一由言國之史官皆博聞強識之士明曉於人君得失善
二惡之述禮義廢則人衛亂政教失則法令酷國史傷此人儉
二一之廢棄哀此別政之青唐哀傷之志蘋積於內乃除詠己二
二情性以風刺其上觀其改惡爲善所以作變詩也國史者周
官大史小史外史衛史之等皆是也此承變風變雅之下則
二兼據天子諸侯之史矣得失之述者人君既往之所行也明
時得失之述哀傷而詠備性者詩人也非史官也民勞常並
二公卿之作也黃鳥碩人國人之風然則凡是臣民皆得風刺
二不必要其國史所為此文特言國史者鄭荅張逸云國史采
二眾詩時明其好惡令誓勝歌之其無作主皆國史主之令可
二歌如此言是由國史掌書故託文史也苟能制作文章亦可
二謂之爲史不必要作史官明云史克作是頌史官自有作蓋
一者矣不盡是史官爲之也言明其好惡令譬勝歌之是國史
二還取善者始竹樂官也言其無作主國史主之嫌其作者無

竹 : glyphe erroné
明 : glyphe vu une seule fois en apprentissage
制 : glyphe non vu en apprentissage

Les premières expérimentations - la question des classes



Expérimentation	Accuracy
Sur les glyphes inconnus	**86,21 %**
Sur les glyphes connus (y compris *one shot learning*)	**94,49 %**
Données brutes	**91,5 %**
Données pondérées	**93,09 %**

Posent problème :

- les débuts de colonnes
- la ponctuation excentrée des colonnes
- certains caractères :

爲 / 為

荼 / 菜

苻 / 持

已 / 己

千 / 干

旋 / 施

La question des classes - le choix de la normalisation

Exemples :

爲　為

既　旣

即　卽

總　揔

Considérations pratiques :

- ressources humaines (et financières) limitées
- densité du corpus, fatigabilité de l'œil et de l'esprit des annotateurs
- hétérogénéité des transcriptions

Considérations philologiques :

- corpus constitué d'éditions de périodes différentes (caractères tabous différents)
- plusieurs mains, des simplifications qui ne sont pas uniformes au fil des pages

Considérations liées à l'étape 1 du projet :

- l'objectif n'est pas d'étudier la réception au XVIIIe s. d'un texte antique (les « erreurs » seraient alors signifiantes et précieuses)
- pas d'approche diplomatique (édition choisie = la plus commune)
- structuration en vue d'une fouille de texte pour chercher des phénomènes d'emprunts d'un genre textuel à un autre

Conclusion et perspectives

Layout : jusqu'à 96% de bon tri des lignes mais résultats peu robustes - impact négatif de la « baseline »

HTR : 93,09 d'accuracy

- Consolider ces premiers essais dans le cadre du COLLEX-PERSÉE CHI-KNOW-PO-CORPUS
  - Corpus diversifié (mise en page, lexique)
  - Nouveaux développements en perspective pour améliorer les performances concernant l'ordre des colonnes
  - Enrichir la reconnaissance des types de textes (texte *versus* commentaires, titres, *etc.*)
- Partager ces modèles et données d'entraînement pour qu'ils servent aux chercheurs travaillant sur des textes anciens de l'Asie extrême-orientale (Chine, Corée, Japon, Vietnam)

Pour finir

Outil utilisé :

- vision.calfa.fr

Publication des données et modèles à venir :

- Nakala
- Github (et référencement dans HTR United)

Remerciements :

- COLLEX-PERSÉE
- DISTAM (Digital Studies: Africa, Asia and the Middle East)
- USIAS