Experiments for automatic treatment of ancient Chinese sources

Marie Bizais-Lillig, Chahan Vidal-Gorène

▶ **To cite this version:**

Marie Bizais-Lillig, Chahan Vidal-Gorène. Experiments for automatic treatment of ancient Chinese sources. Documents anciens et reconnaissance automatique des écritures manuscrites, Ariane Pinche; Jean-Baptiste Camps, Jun 2022, Paris, France. hal-03792246

HAL Id: hal-03792246
https://hal.science/hal-03792246v1

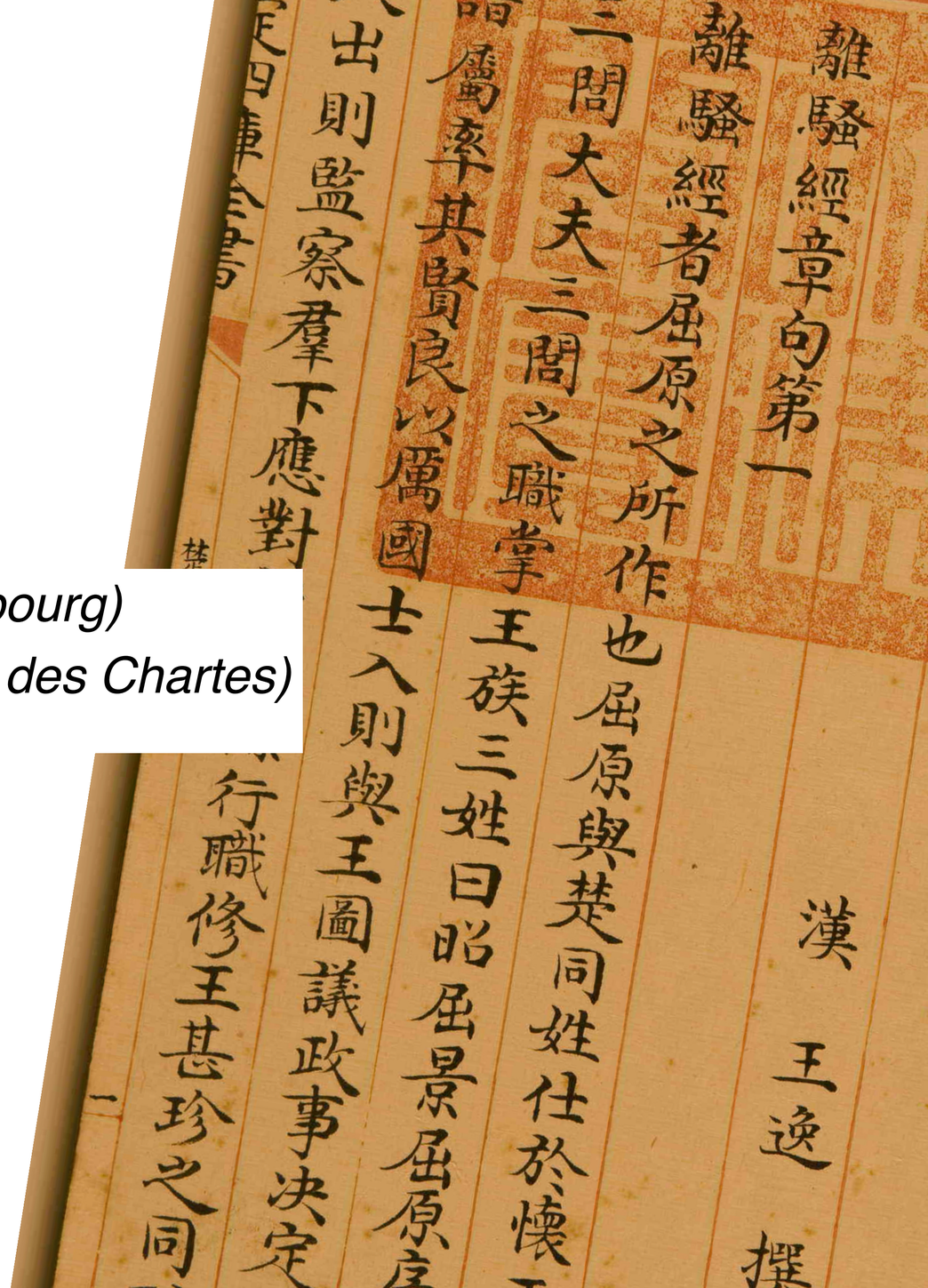
Submitted on 16 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

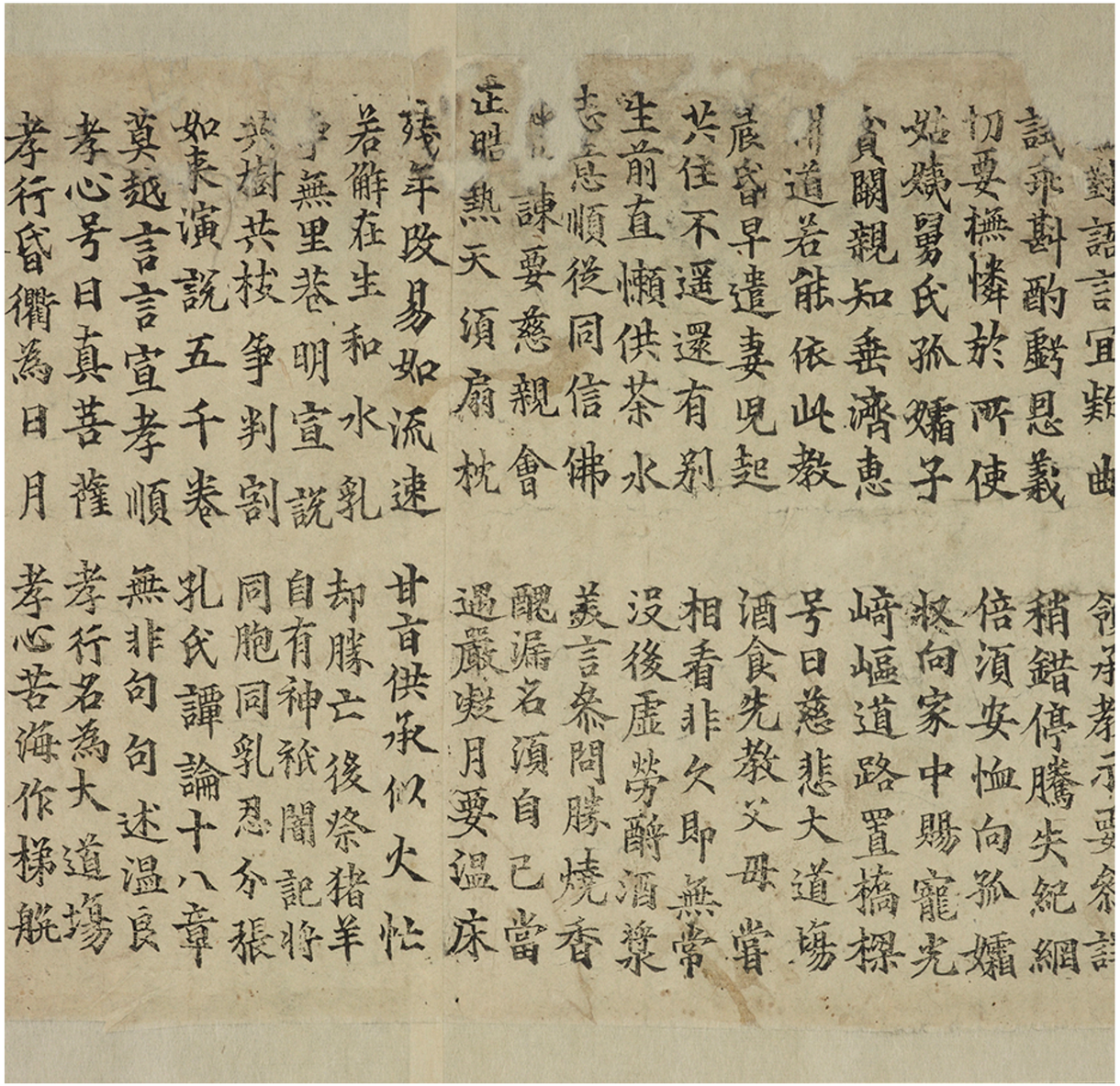
- Projet de fouille de textes chinois, pour l'essentiel du Moyen âge (IIIe-Xe s.) : CHI-KNOW-PO
- Nécessite :
    - l'acquisition du corpus dans son intégralité et sans restriction de droits
    - la structuration du corpus (dimension éditoriale)
- Mai-juin 2021 : début d'une collaboration entre
    - 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

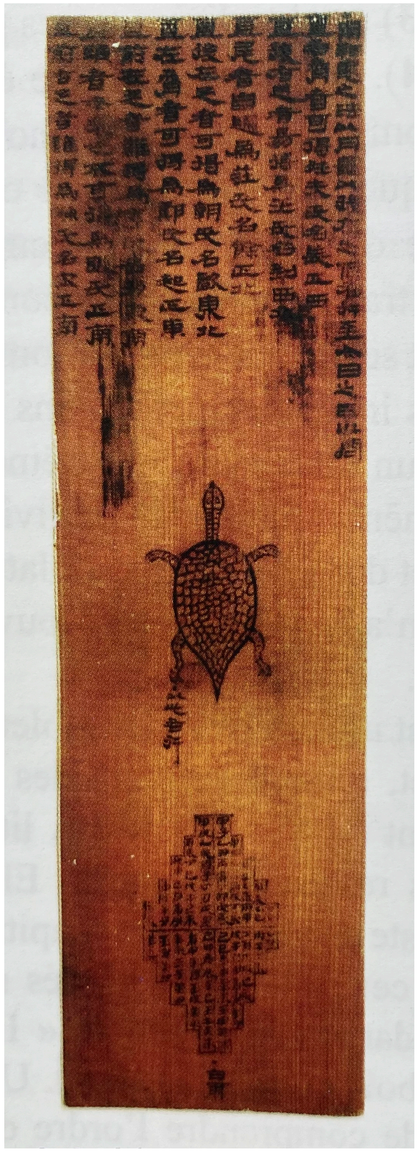
Spécificités du corpus - des textes xylographiés

- 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



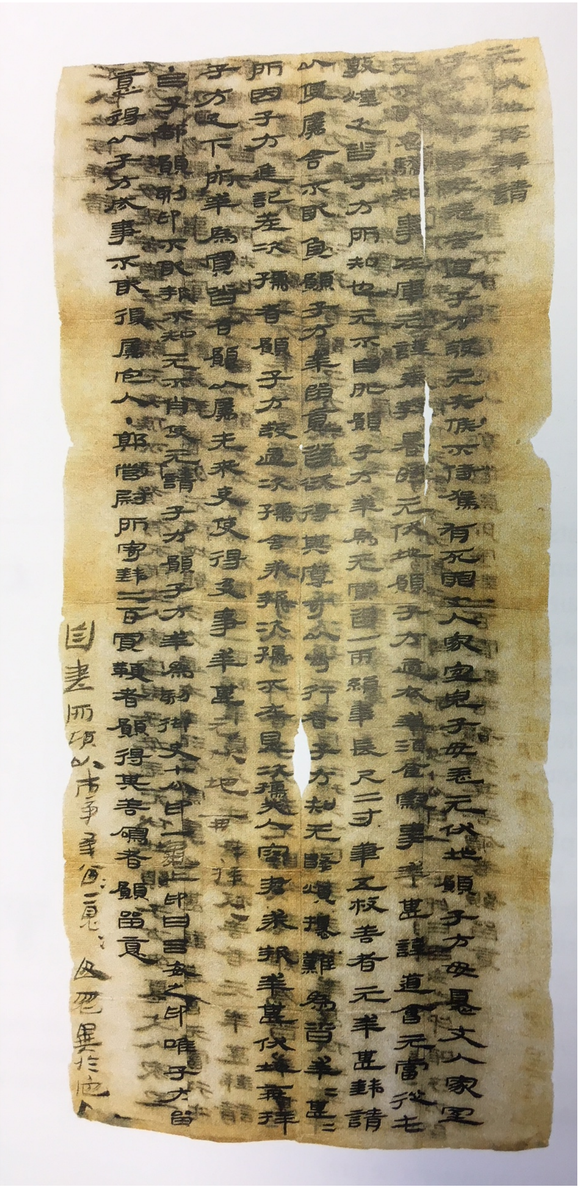
Impression xylographique (Dunhuang - British Library - OR8210)

Spécificités du corpus



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

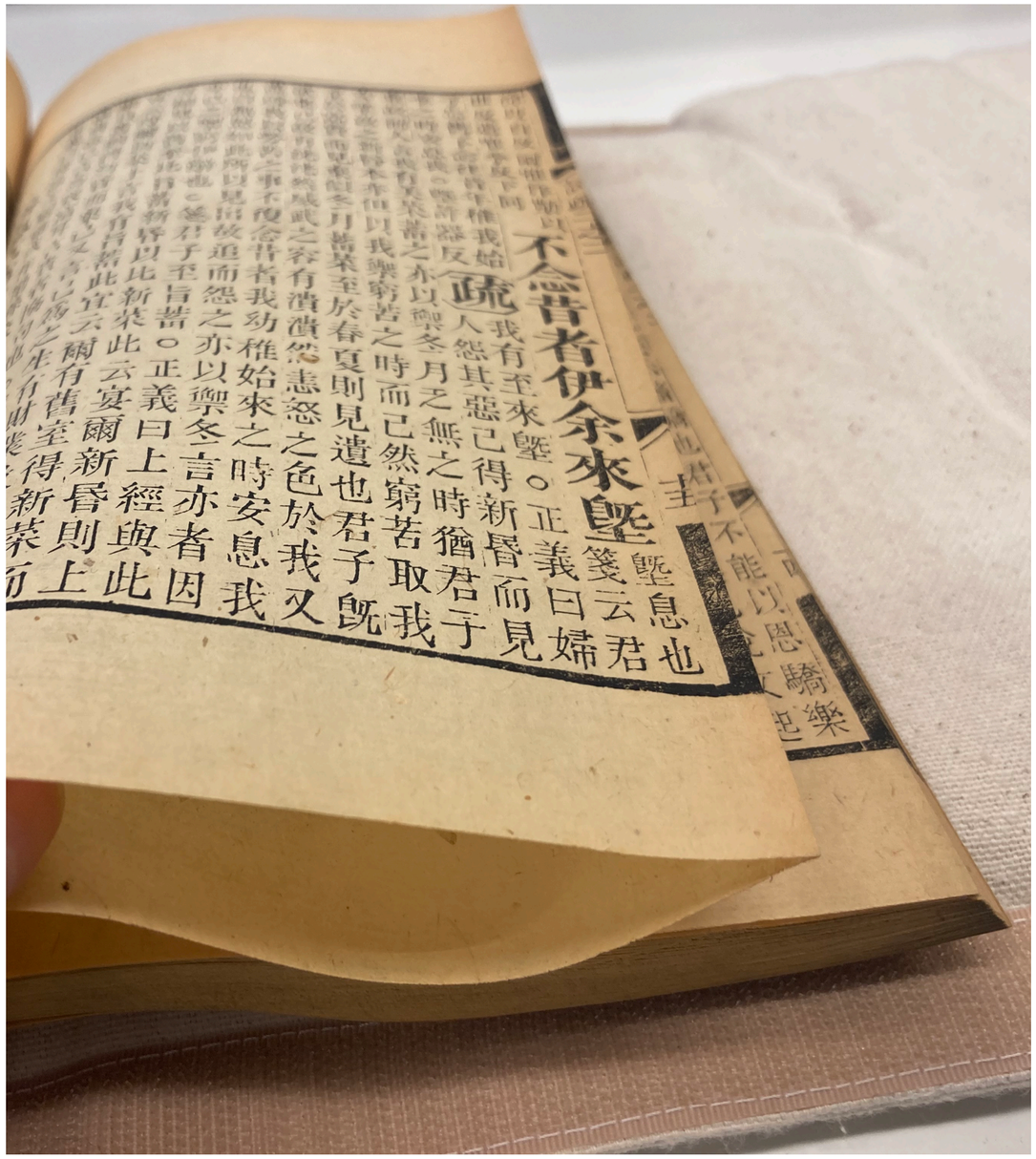
lamelles de bambou (*Shijing* 詩經 - ca. 250 AE - Université de l'Anhui)
source : 安徽大学藏战国竹简（一）



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

Spécificités du corpus - mise en page

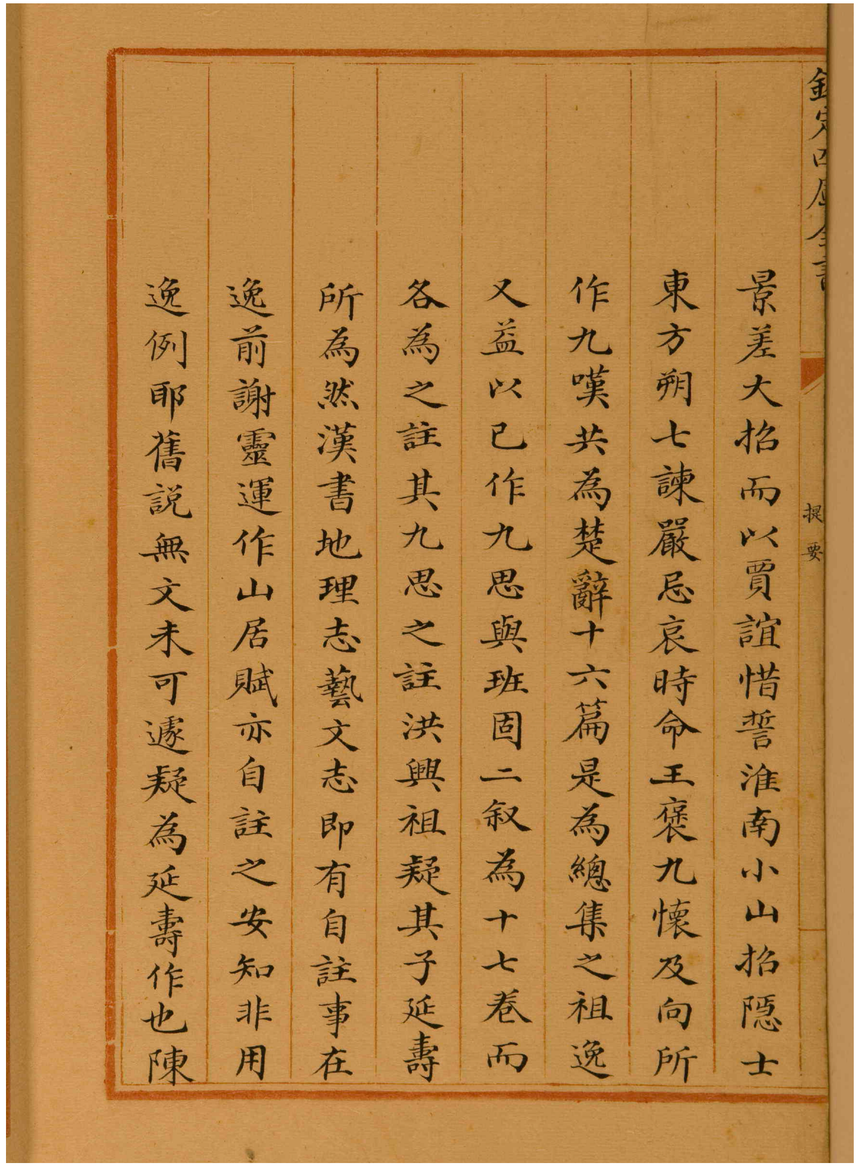
- Des textes transmis de pair avec un ou plusieurs commentaires
- Mise en page : entrelacement (à partir du IIe s. NE)
- Impression sur feuillets longs
  - pliés en accordéon
  - reproduits avec 2, 4 ou 6 double-pages par page (densité)



Reliure « accordéon » (BULAC)

Spécificités du corpus

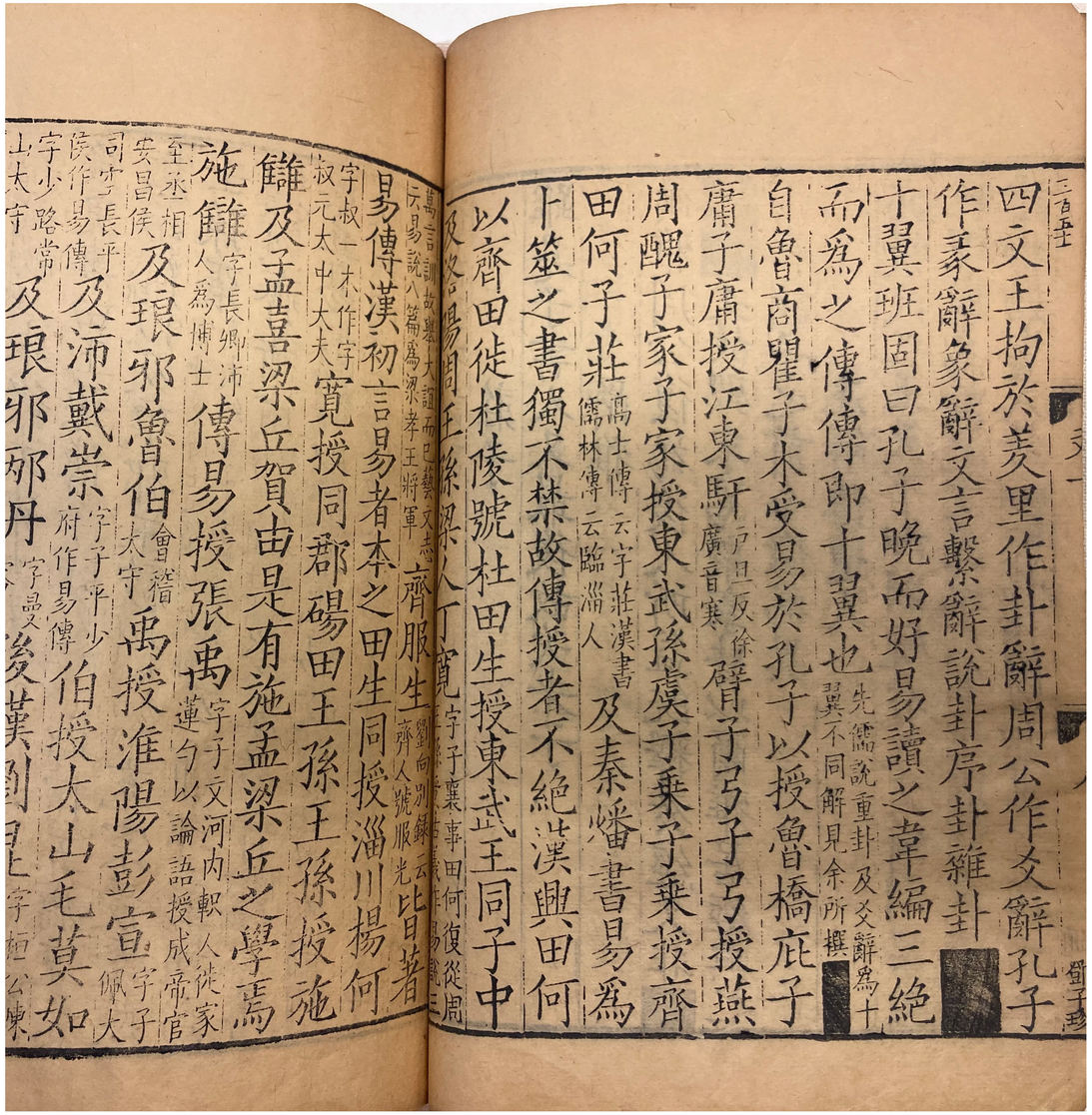
Entrelacement, découpage
爾雅 édition xylographiée illustrée (1801)
BULAC CHI1938(2)



Édition impériale manuscrite
楚辭 dans le *Siku*

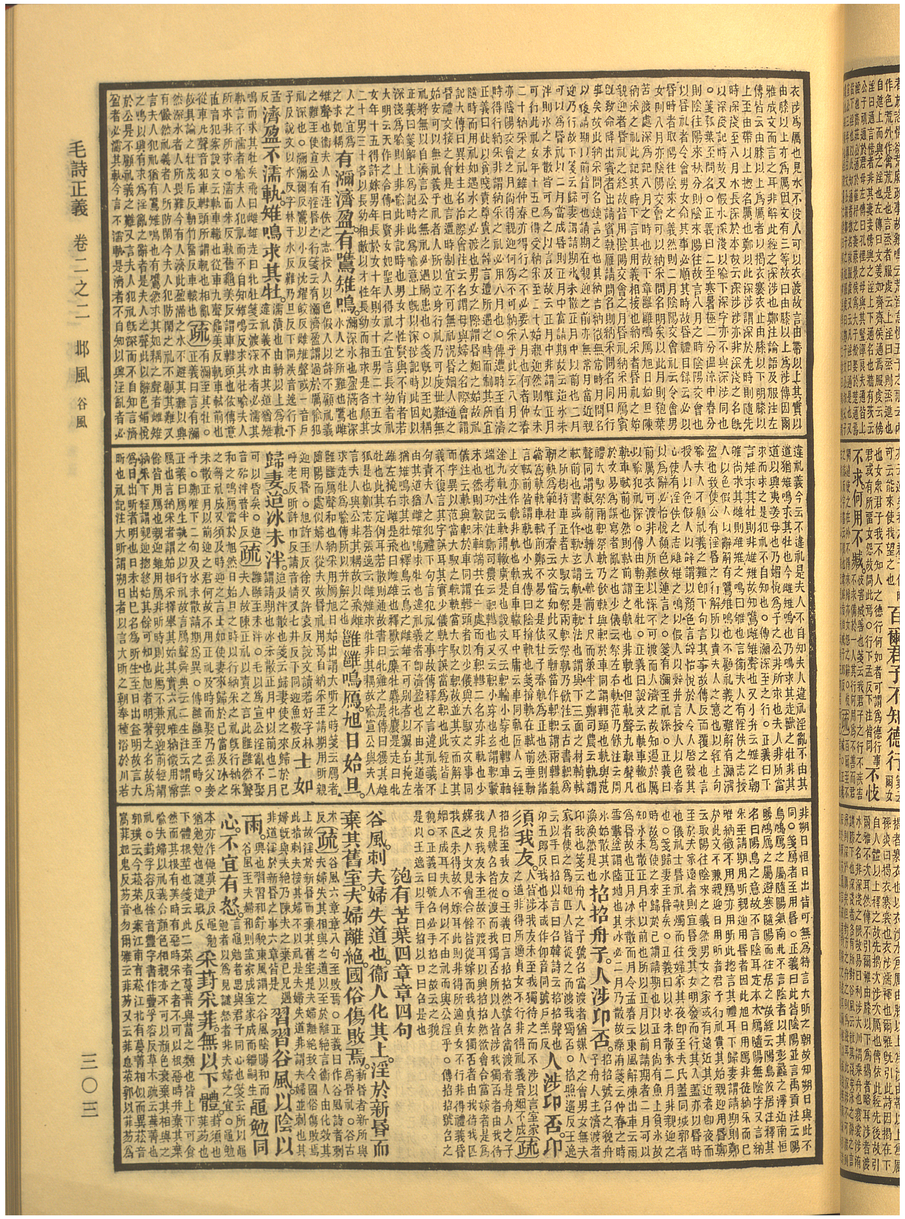
Spécificités du corpus - l’écriture chinoise

- 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

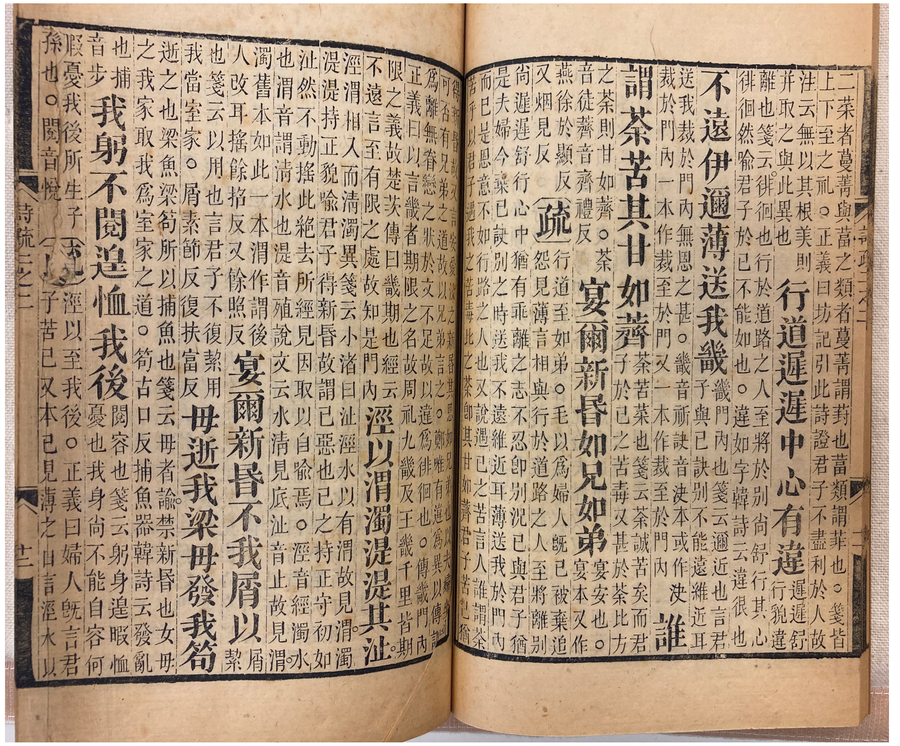


經典釋文 édition de 1800
BULAC CHI567

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))

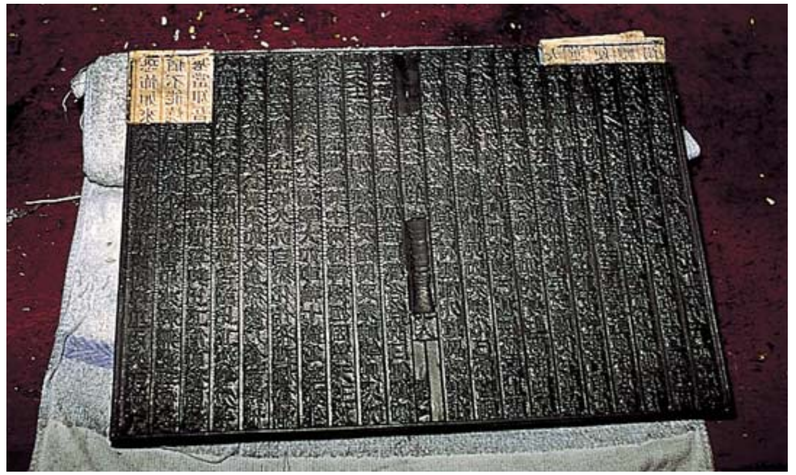


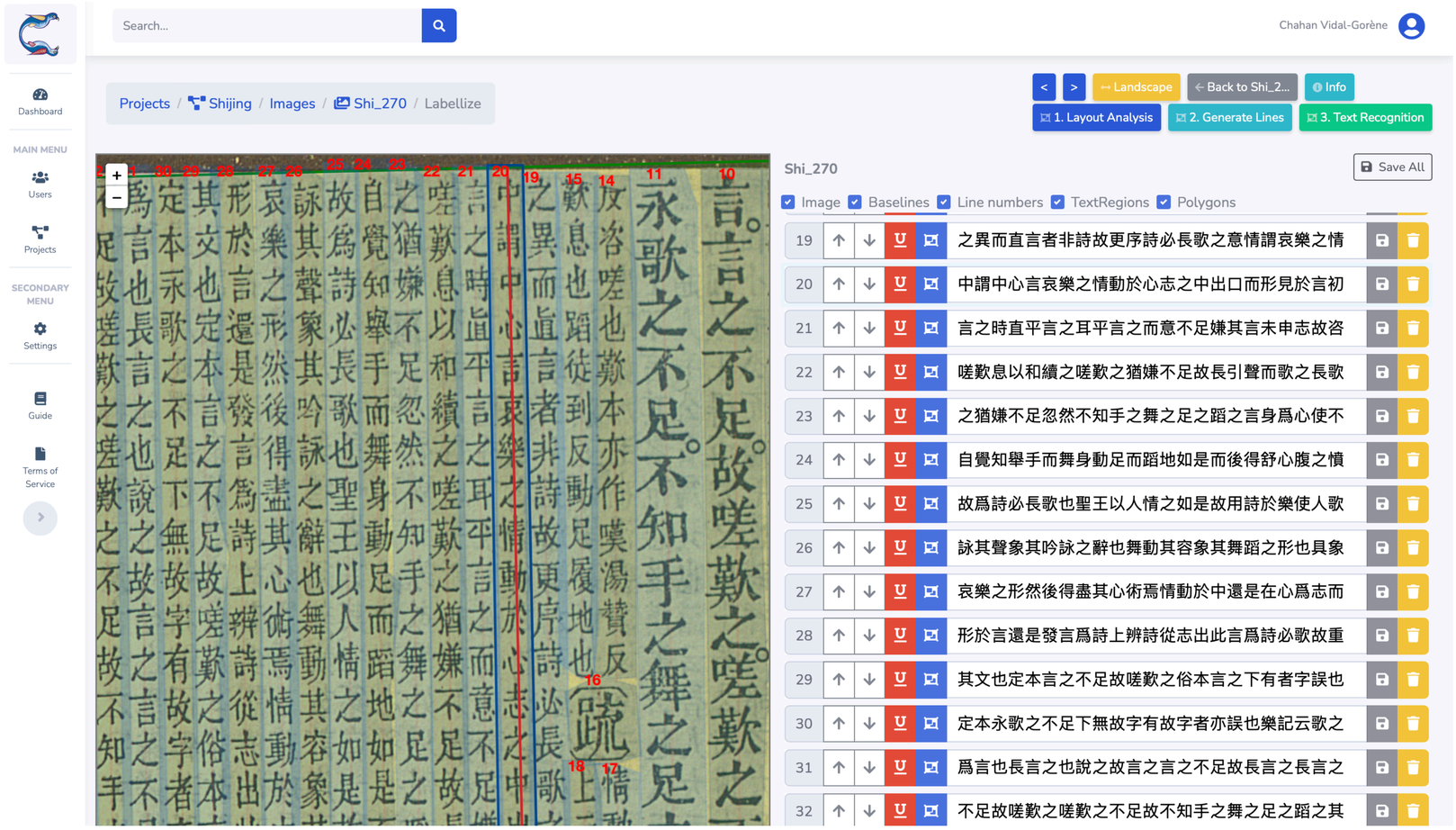
Planche à xylographier
(à partir des IXe-Xe s.)

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

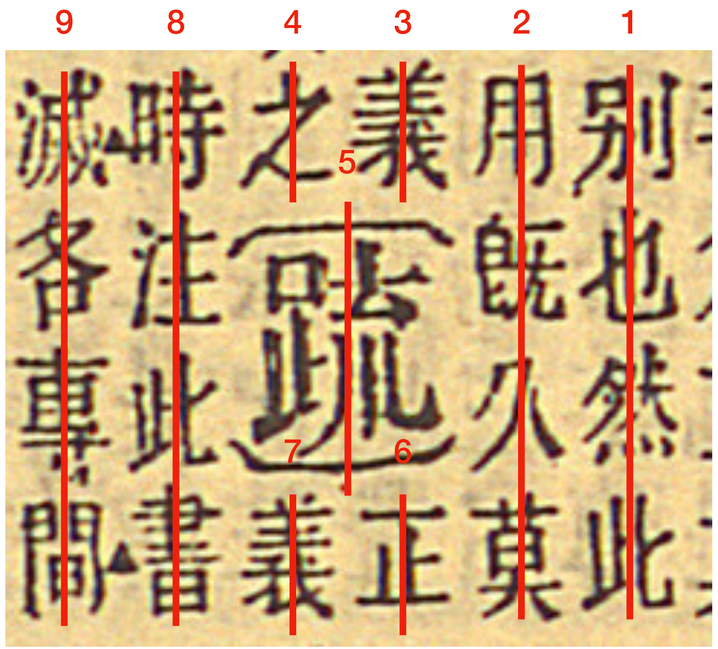
=> https://vision.calfa.fr

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

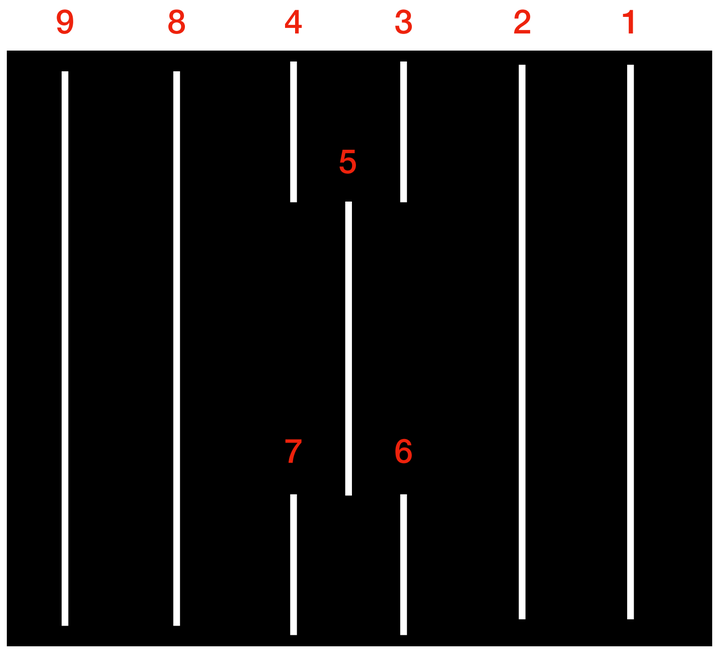
Les premières expérimentations - l’ordre de lecture

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

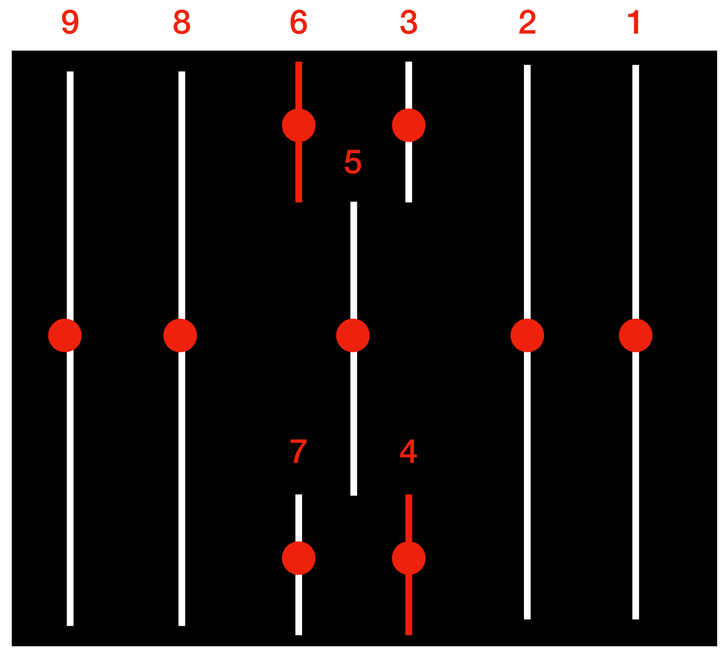
Efficace à **82,37**% sur le dataset, échoue y compris en situation simple



Détection des lignes



Sens de lecture GT

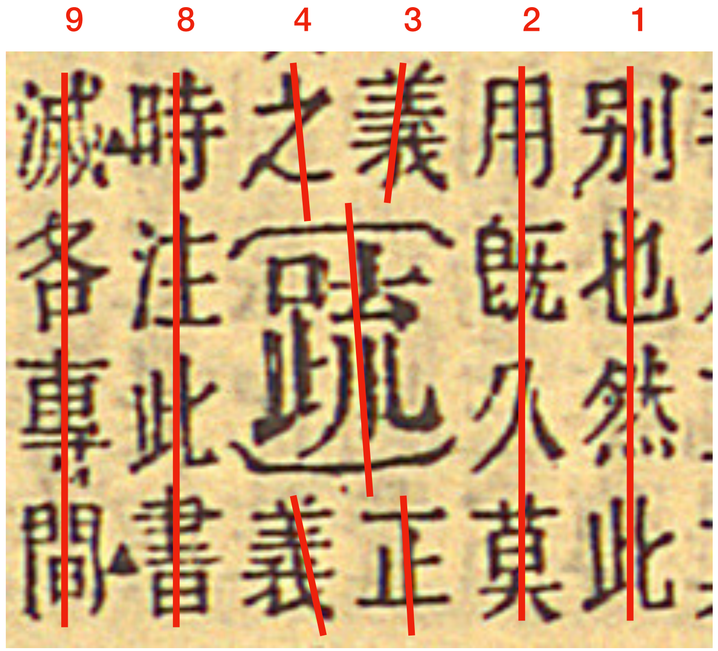


Tri des lignes par centroïdes

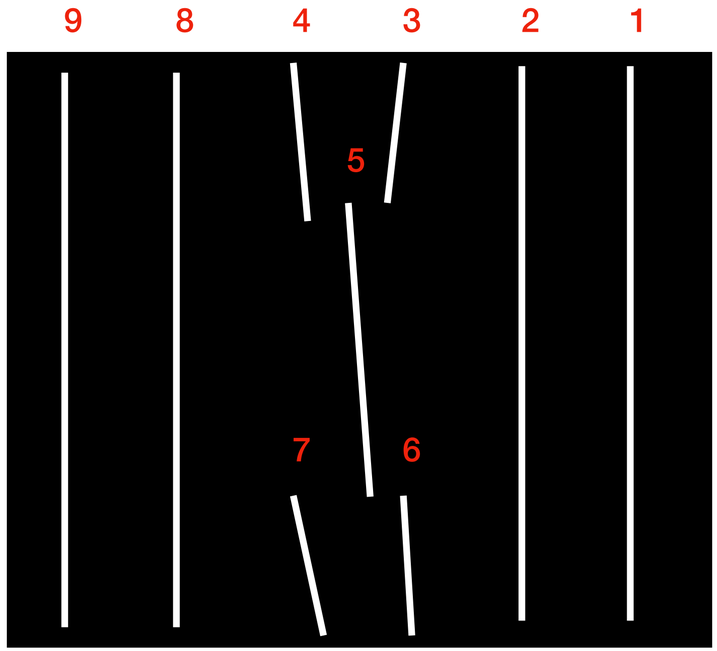
Les premières expérimentations - l’ordre de lecture

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

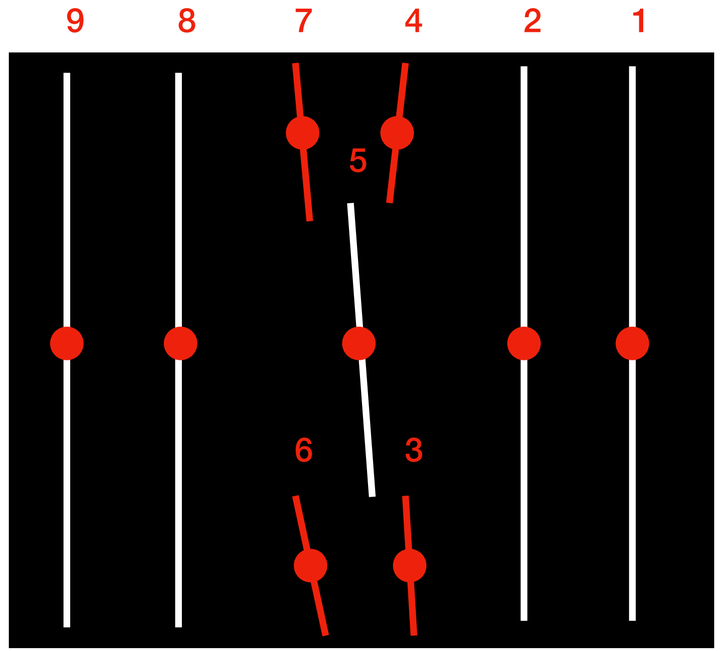
Efficace à **82,37**% sur le dataset



Détection des lignes

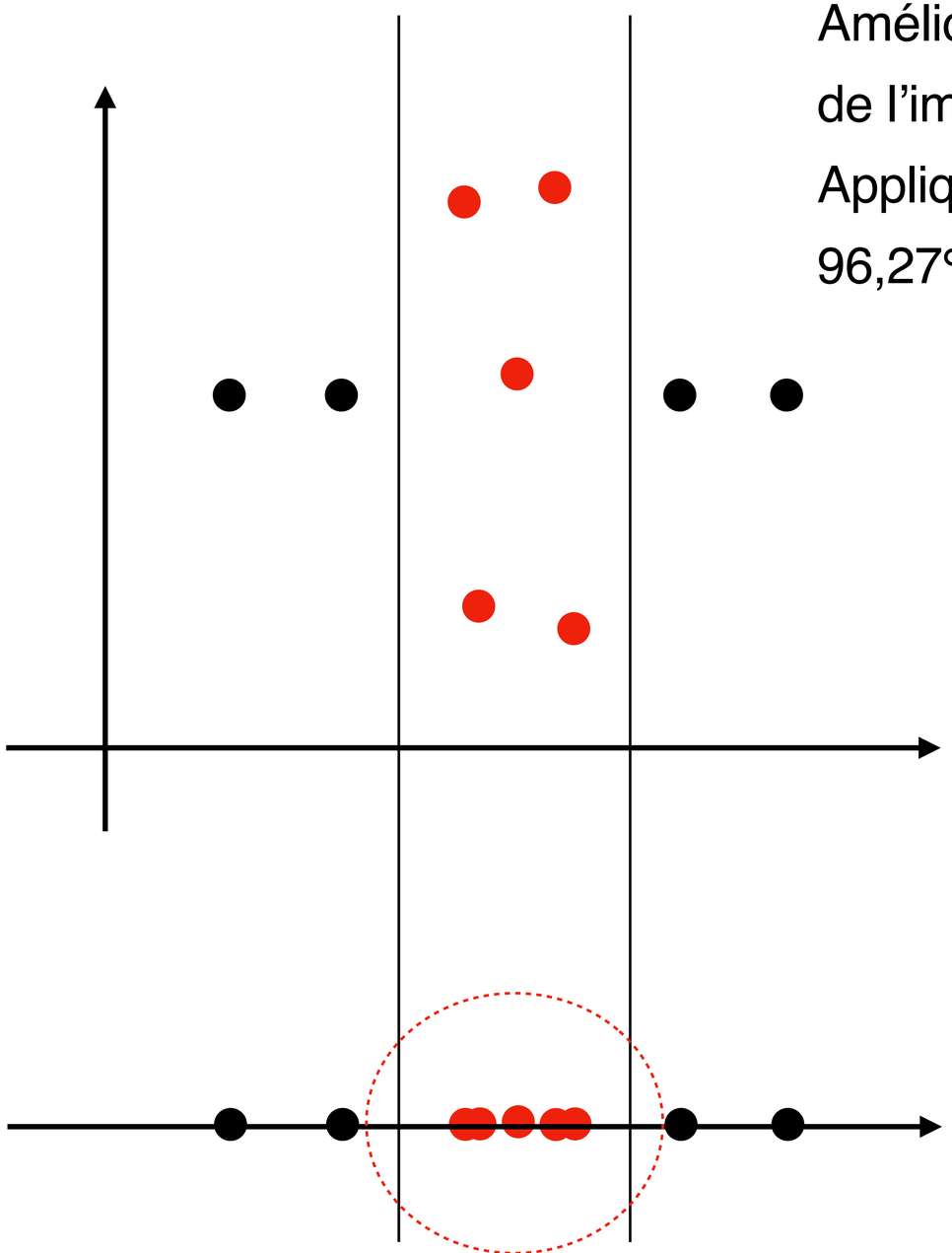


Sens de lecture GT



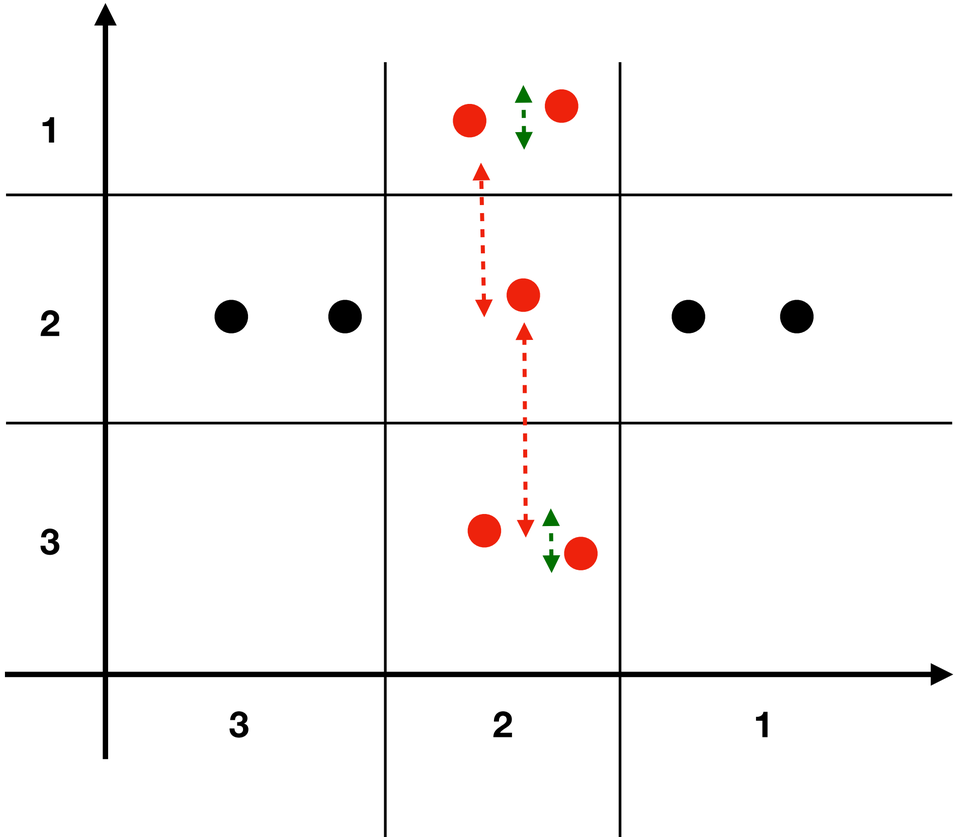
Tri des lignes par centroïdes

Les premières expérimentations - l’ordre de lecture

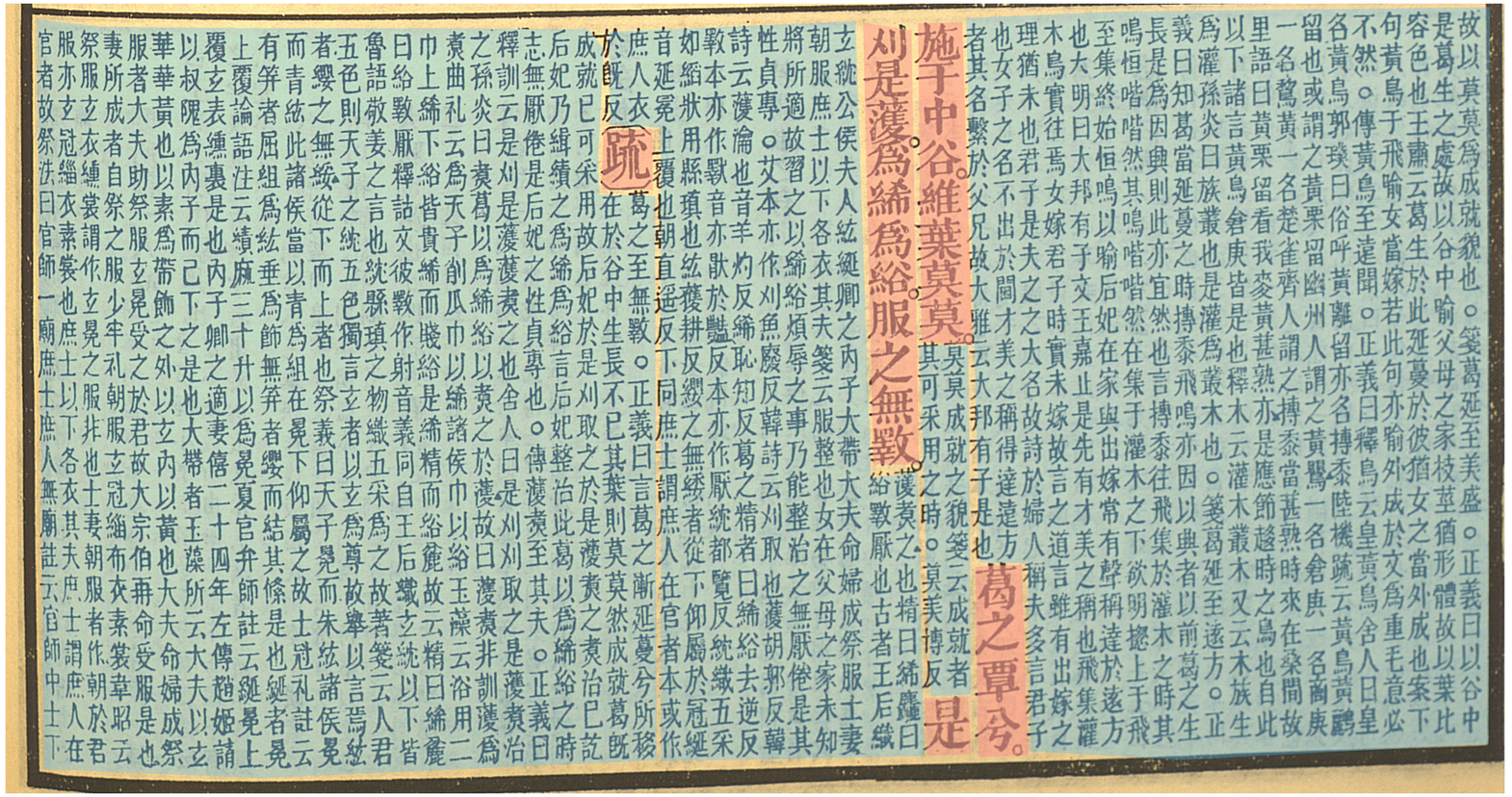


Amélioration du tri avec une solution de traitement de l’image : projection des centroïdes et identification d’un cluster : 87,88%

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



Les premières expérimentations - l'ordre de lecture



故以莫莫爲成就貌也。箋葛延至美盛。正義曰以谷中是葛生之處故以谷中喻父母之家枝莖猶形體故以葉比容色也王肅云葛生於此延蔓於彼猶女之當外成也案下句黃鳥于飛喻女當嫁若此句亦喻外成於文爲重毛意必不然。傳黃鳥至達聞。正義曰釋鳥云皇黃鳥舍人曰皇名黃鳥郭璞曰俗呼黃離留亦名搏黍陸機疏云黃鳥黃鸝留也或謂之黃栗留幽州人謂之黃鸝一名倉庚一名鶩黃一名楚雀齊人謂之搏黍當甚熟時來在桑間故里語曰黃栗留看我麥黃甚熟亦是應節趨時之鳥也自此以下諸言黃鳥倉庚皆是也釋木云灌木叢木又云木族生爲灌孫炎曰族叢也是灌爲叢木也。箋葛延至遠方。正義曰知葛當延蔓之時搏黍飛鳴亦因以興者以前葛之生長是爲因興則此亦宜然也言搏黍往飛集於灌木之時其鳴恒喈喈然其鳴喈喈然在集于灌木之下欲明摠上于飛至集終始恒鳴以喻后妃在家與出嫁常有聲稱達於遠方也大明曰大邦有子文王嘉止是先有才美之稱也飛集灌木鳥實往焉女嫁君子時實未嫁故言之道言雖有出嫁之理猶未也君子是夫之之大名故詩於婦人稱夫多言君子也女子之名不出於閨才美之稱得達遠方者其名繫於父兄故大雅云大邦有子是也。葛之覃兮。

施于中谷。維葉莫莫。莫莫成就之貌箋云成就者其可采用之時。莫美博反是刈是濩。爲絺爲綌。服之無斁。濩煮之也精曰絺麤曰綌斁厭也古者王后織玄統公侯夫人紘綖卿之內子大帶大夫命婦成祭服士妻朝服庶士以下各衣其夫箋云服整也女在父母之家未知將所適故習之以絺綌煩辱之事乃能整治之無厭倦是其性貞專。艾本亦作刈魚廢反韓詩云刈取也濩胡郭反韓詩云蒦倫也音羊灼反絺恥知反葛之精者曰絺綌去逆反斁本亦作厭於豔反本亦作厭統都覽反統織五采如綢狀用縣緜也紘覆耕反纊之無緌者從下仰屬於冠纓音延冕上覆也朝直遥反下同庶士謂庶人在官者本或作庶人衣

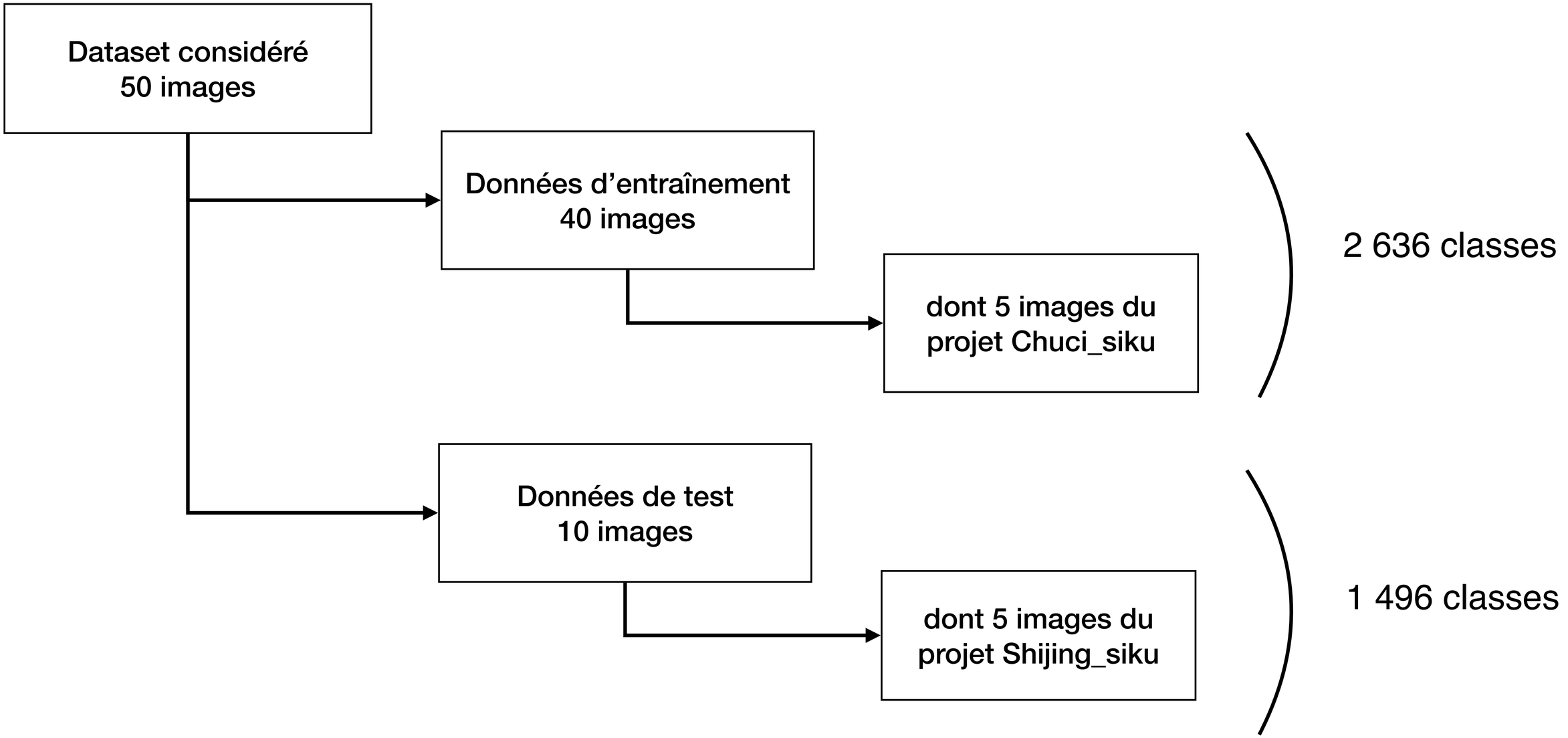
疏

葛之至無斁。正義曰言葛之漸延蔓兮所移於既反在於谷中生長不已其葉則莫莫然成就葛既成就已可采用故后妃於是濩煮之煮治已訖后妃乃緝績之爲絺爲綌言后妃整治此葛以爲絺綌之時志無厭倦是后妃之性貞專也。傳濩煮至其夫。正義曰釋訓云是刈是濩蒦煮之也舍人曰是刈刈取之是濩煮治之孫炎曰煮葛以爲絺綌以煮之於濩故曰濩煮非訓濩爲煮曲礼云爲天子削瓜巾以絺諸侯巾以綌玉藻云浴用二巾上絺下綌皆貴絺而賤綌是絺精而綌麤故云精曰絺麤曰綌斁厭釋詁文彼斁作射音義同自王后織玄統以下皆魯語敬姜之言也統緜瑱之物織五采爲之故著箋云人君五色則天子之統五色獨言玄者以玄爲尊故舉以言焉紘者纓之無緌從下而上者也祭義曰天子冕而朱紘諸侯冕而青紘此諸侯當以青爲組在冕下仰屬之故士冠礼註云有笄者屈組爲紘垂爲飾無笄者纓而結其絛是也延者冕上覆論語注云績麻三十升以爲冕夏官弁師註云延冕上覆玄表纁裏是也內子卿之適妻僖二十四年左傳趙姬請以叔隗爲內子而己下之是也大帶者玉藻所云大夫以玄華華黃也以素爲帶飾之外以玄內以黃也大夫命婦成祭服亦玄冠緇衣素裳也庶士以下各衣其夫庶士謂庶人在官者故祭法曰官師一廟庶士庶人無廟註云官師中士下

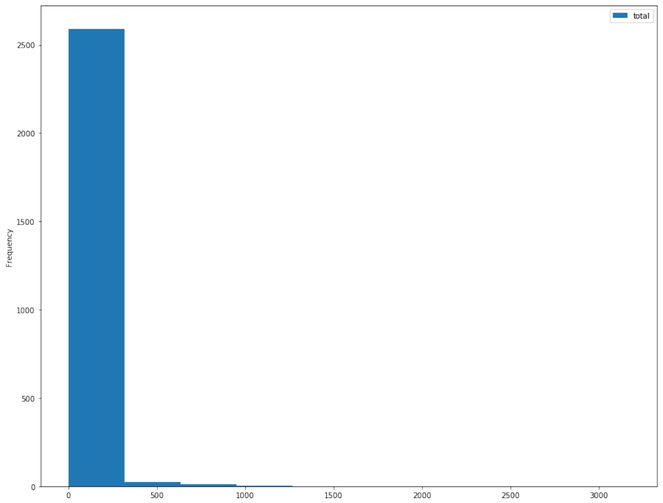
Solution en cours d'expérimentation : pré-classification sémantique au sein d'une région de texte puis tri indépendamment des lignes pour chaque zone identifiée. Abandon des baselines. Score obtenu sur un échantillon manuellement annoté, sans apprentissage : 99,67%

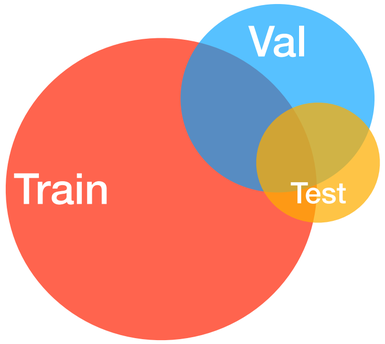
Les premières expérimentations - la question des classes

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

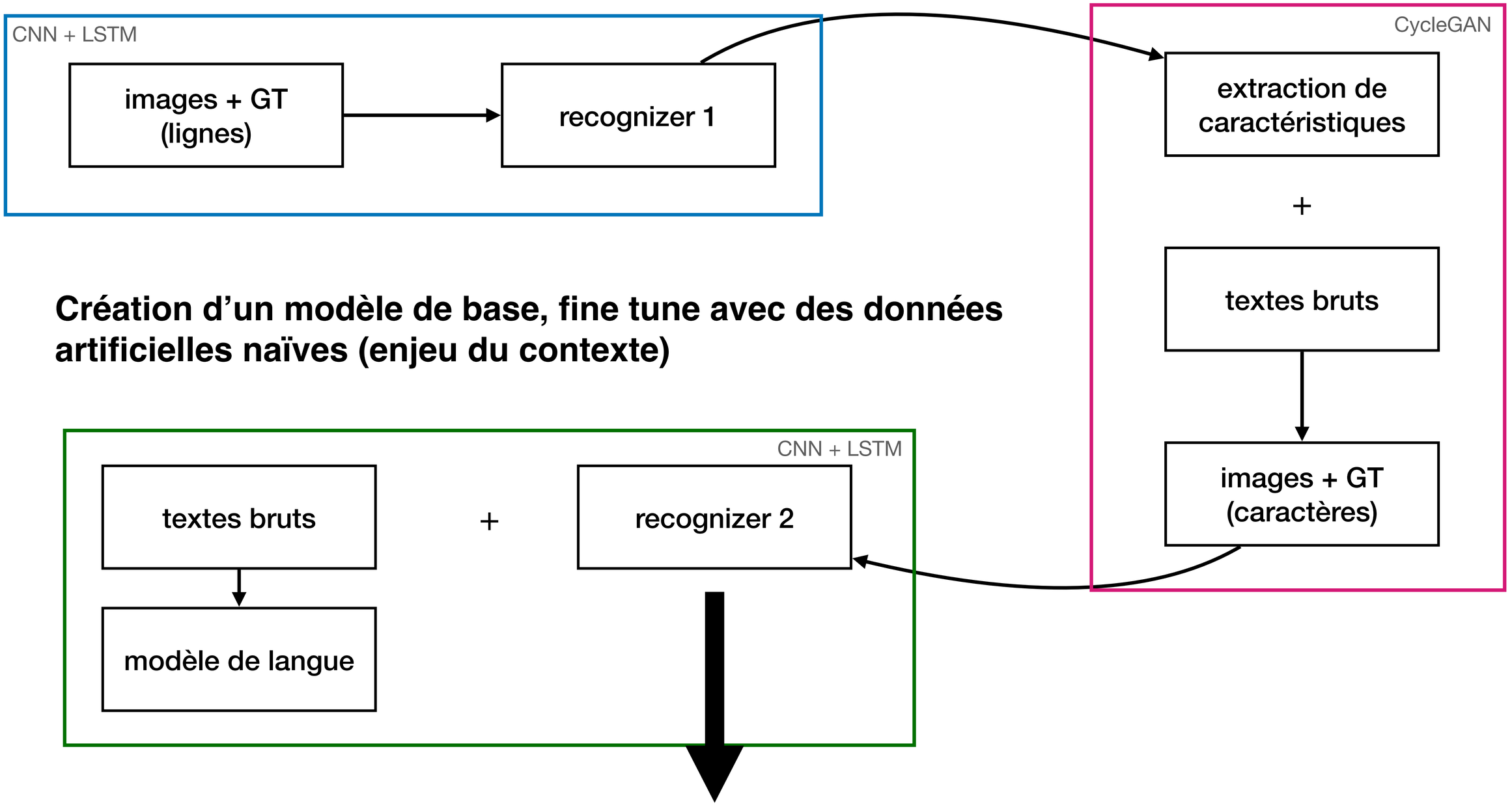
之	3170			芋	1
也	2032			恕	1
以	1463			刃	1
云	1110		…	訪	1
不	1074			隈	1
言	960			沔	1
者	944			斐	1

2 095 classes sur 2636 avec un seul échantillon

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

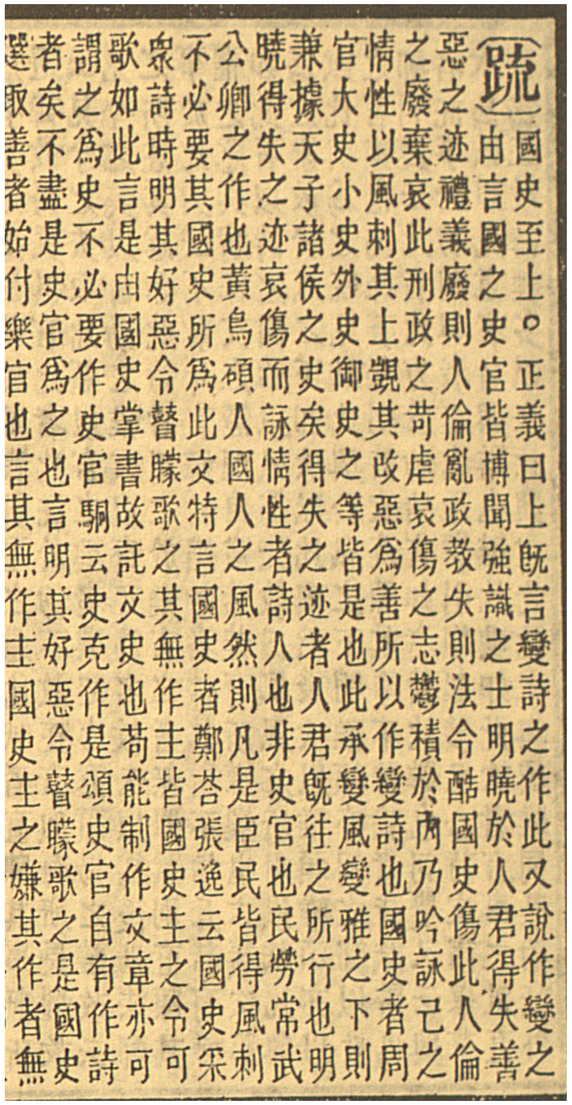
==> Situation de *one-shot learning* + classes non vues en apprentissage

Les premières expérimentations - la question des classes



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

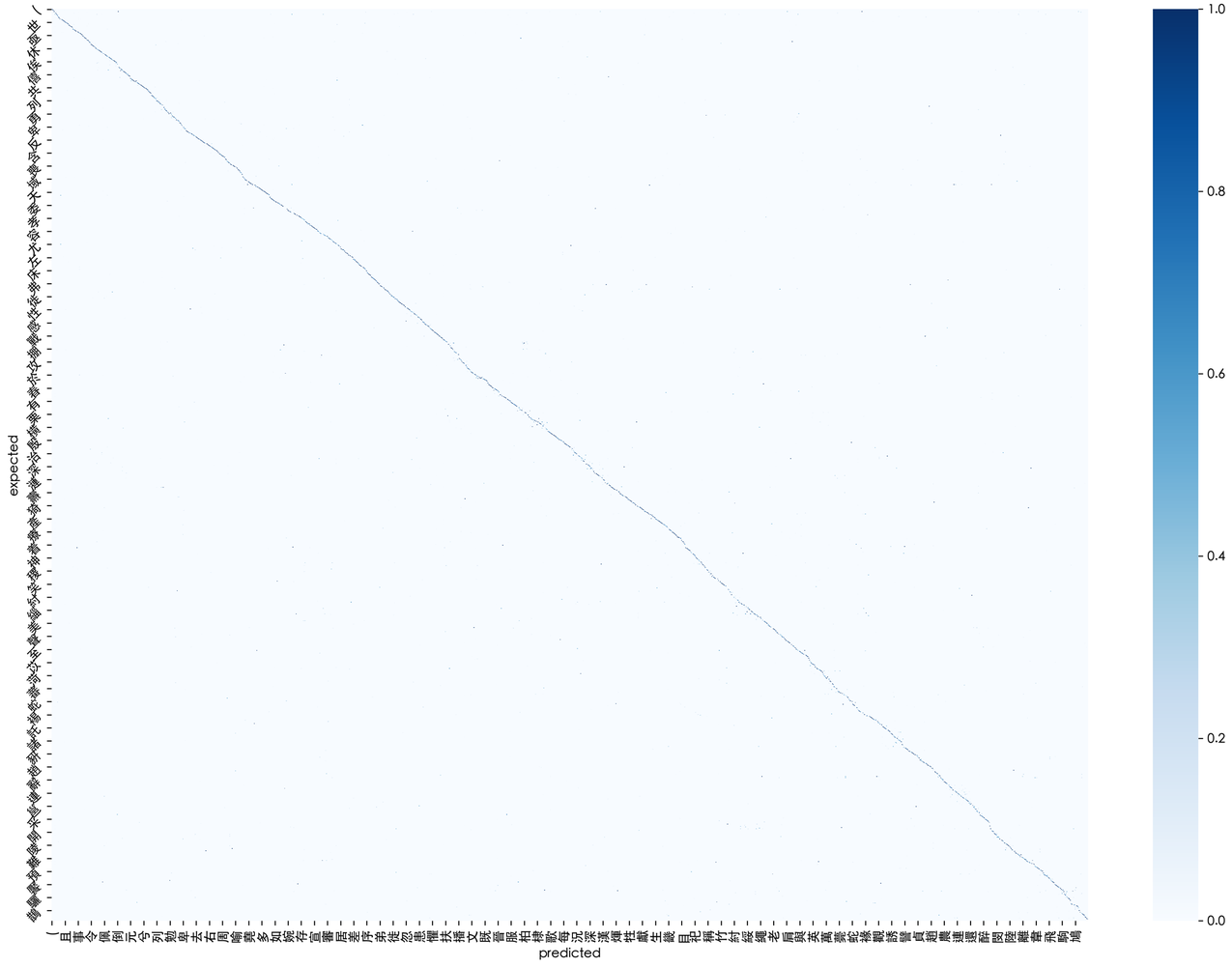
Les premières expérimentations - la question des classes



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