



**HAL**  
open science

## 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.

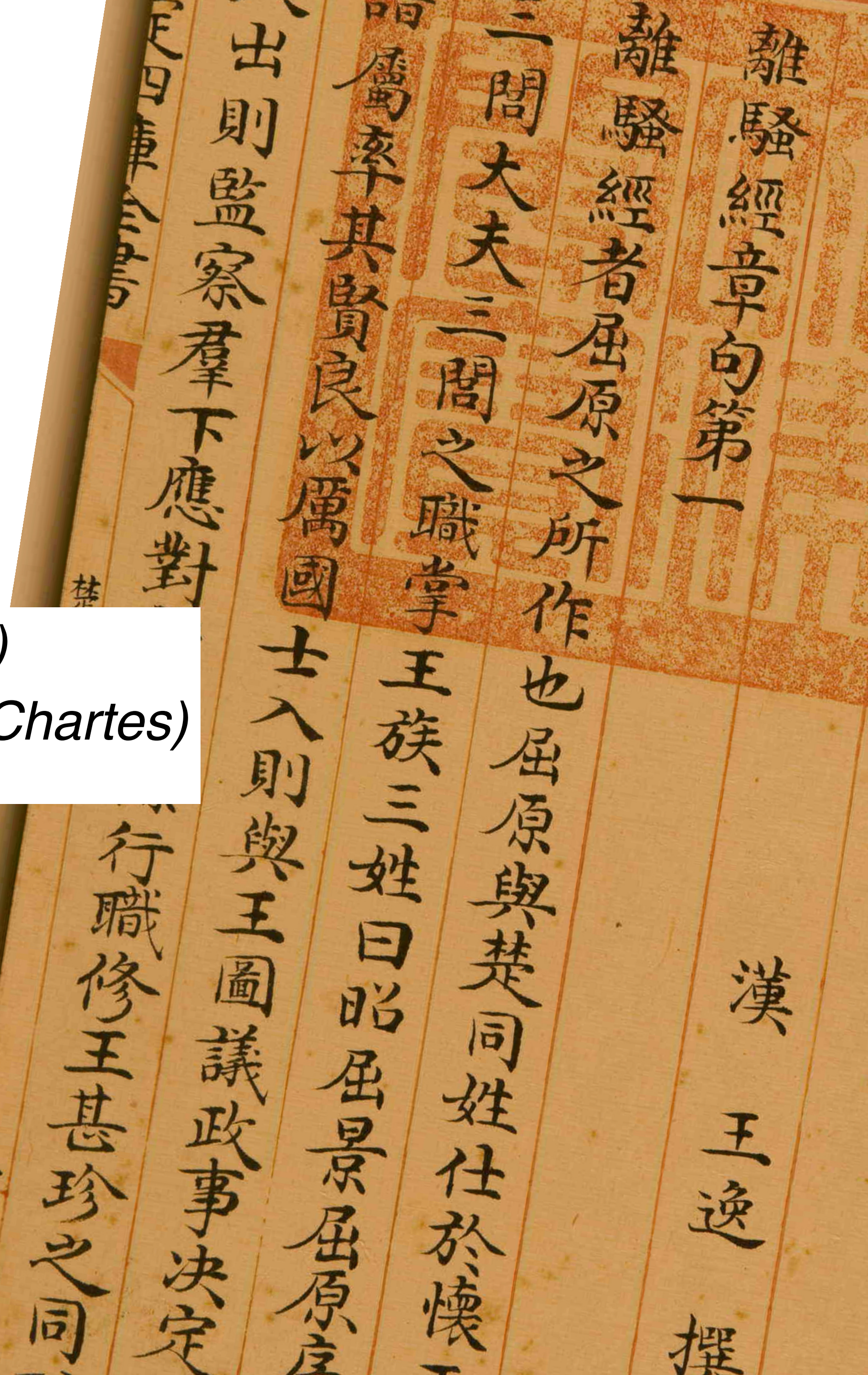


HAL Authorization

# Expérimentations pour l'analyse automatique de sources chinoises anciennes

Marie Bizais-Lillig (*Université de Strasbourg*)

Chahan Vidal-Gorène (*École nationale des Chartes*)

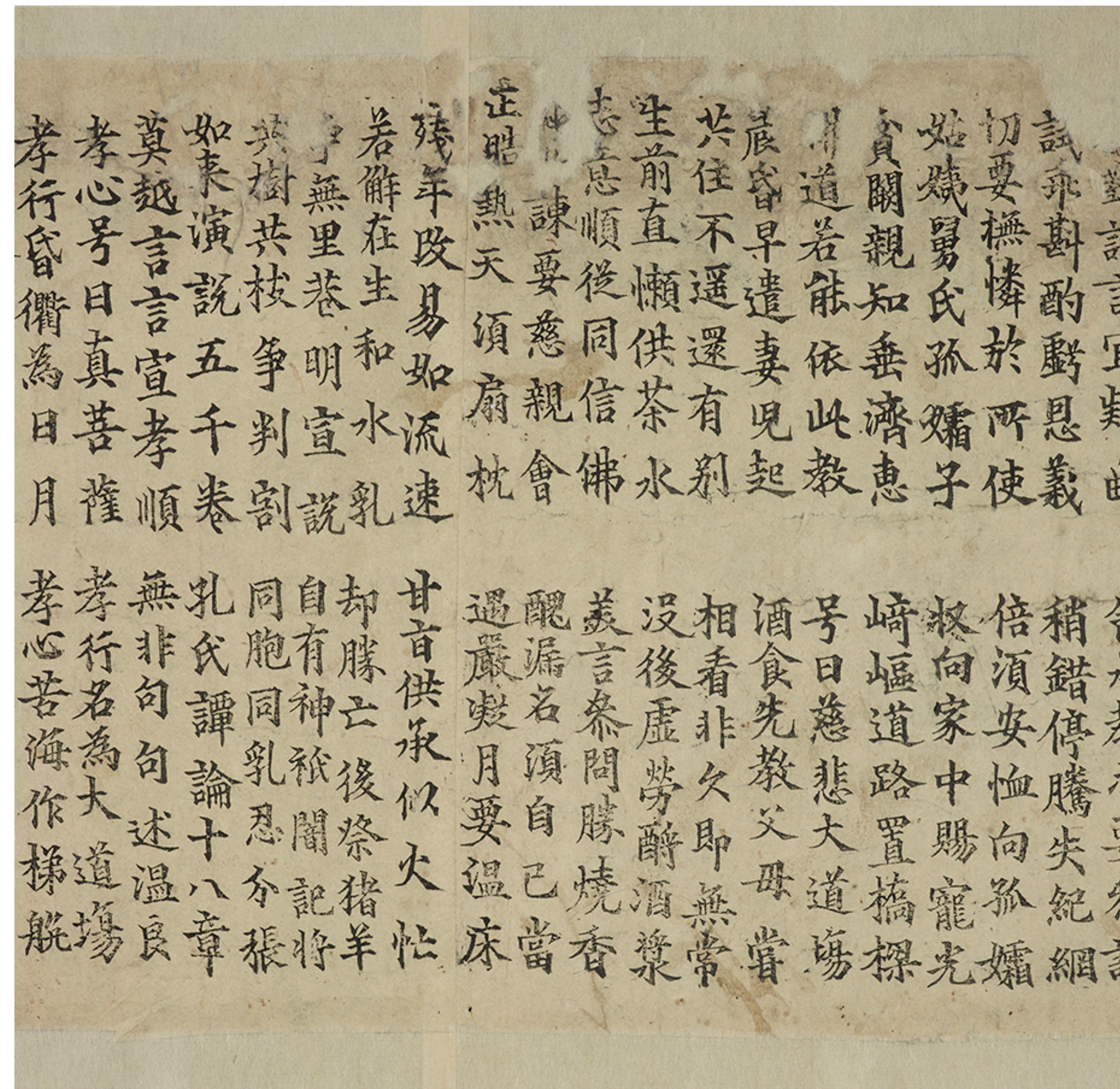


- Projet de fouille de textes chinois, pour l'essentiel du Moyen âge (IIe-Xe s.) : CHI-KNOW-PO
- Nécessite :
  - l'acquisition du corpus dans son intégralité et sans restriction de droits
  - la structuration du corpus (dimension éditoriale)
- Mai-juin 2021 : début d'une collaboration entre
  - des sinologues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

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

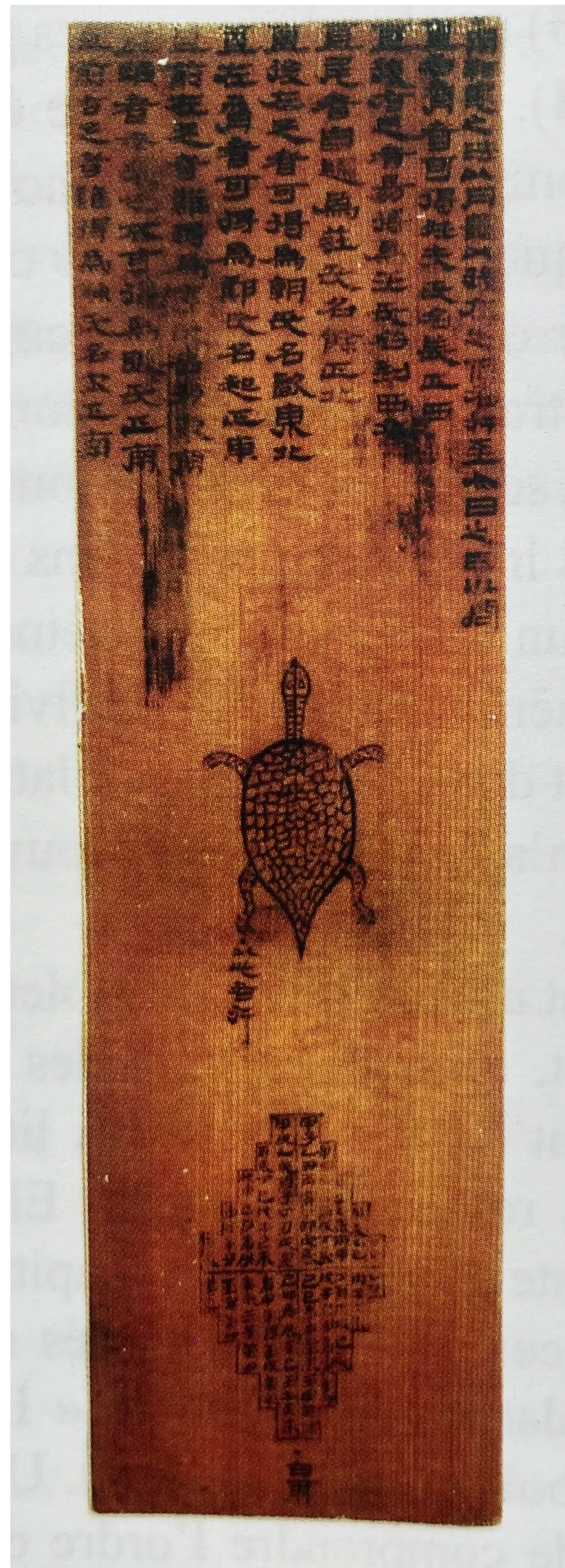
# Spécificités du corpus - des textes xylographiés

- 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)

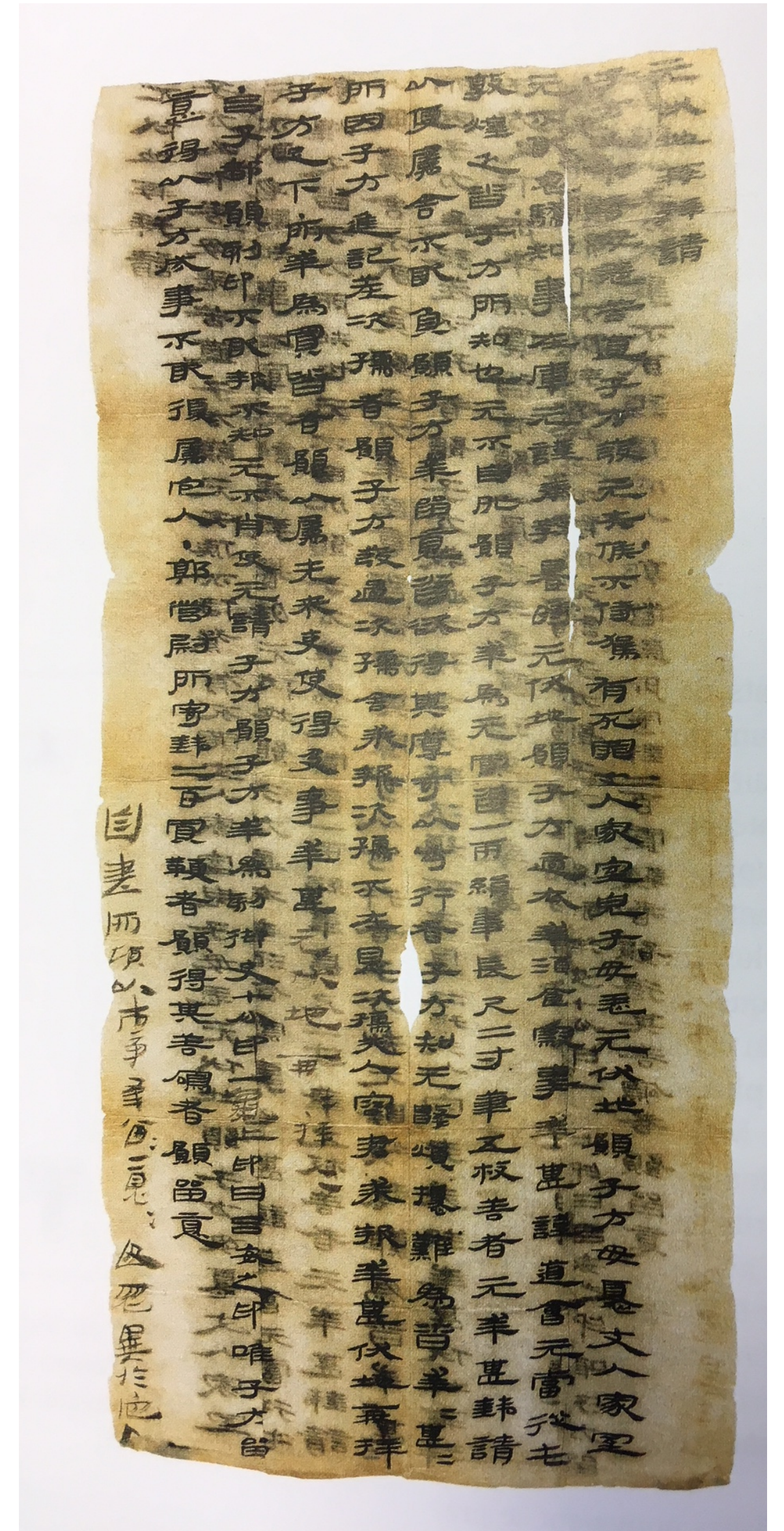
# Spécificités du corpus



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 - 1er 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)

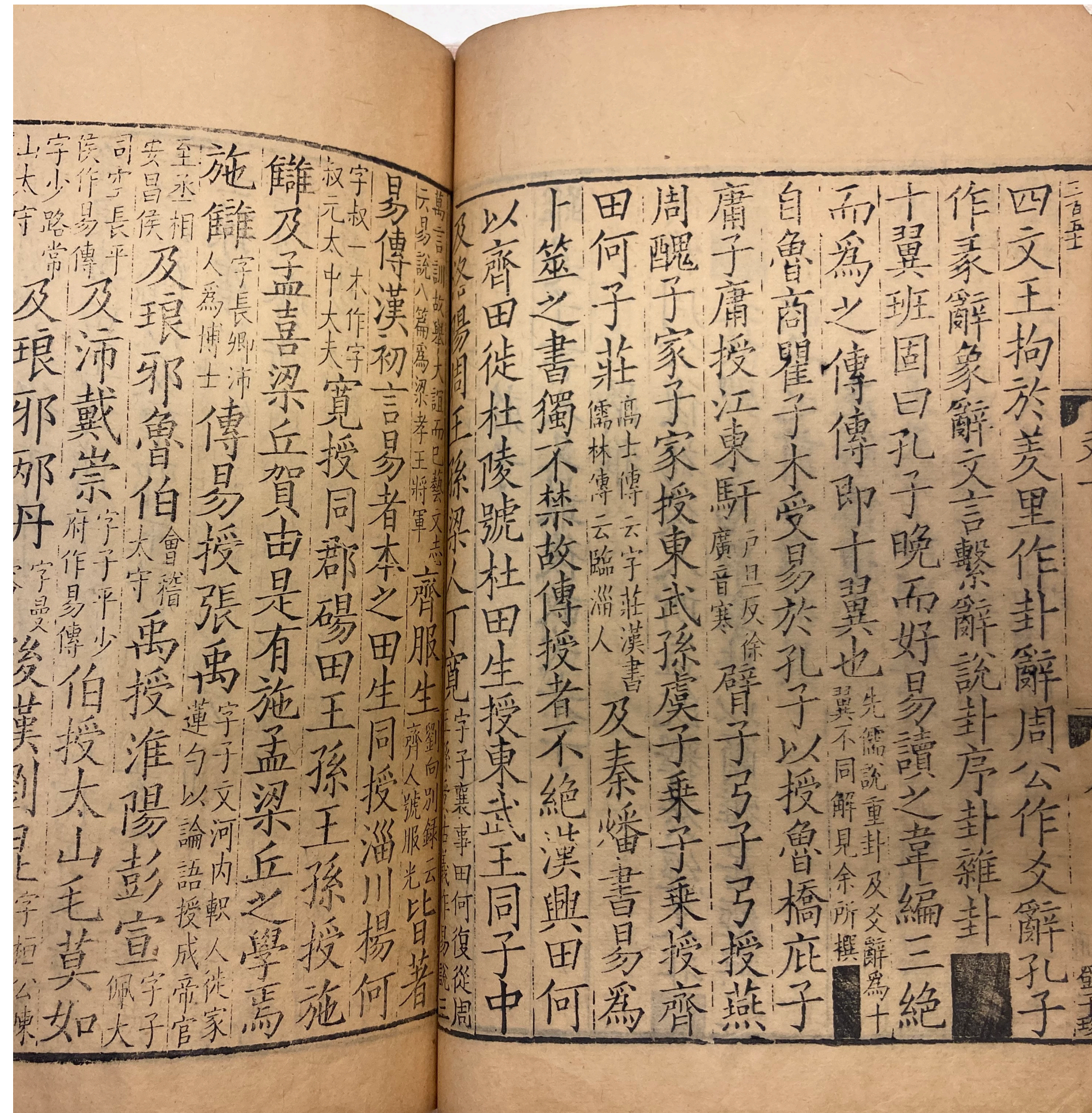


Reliure « accordéon » (BULAC)

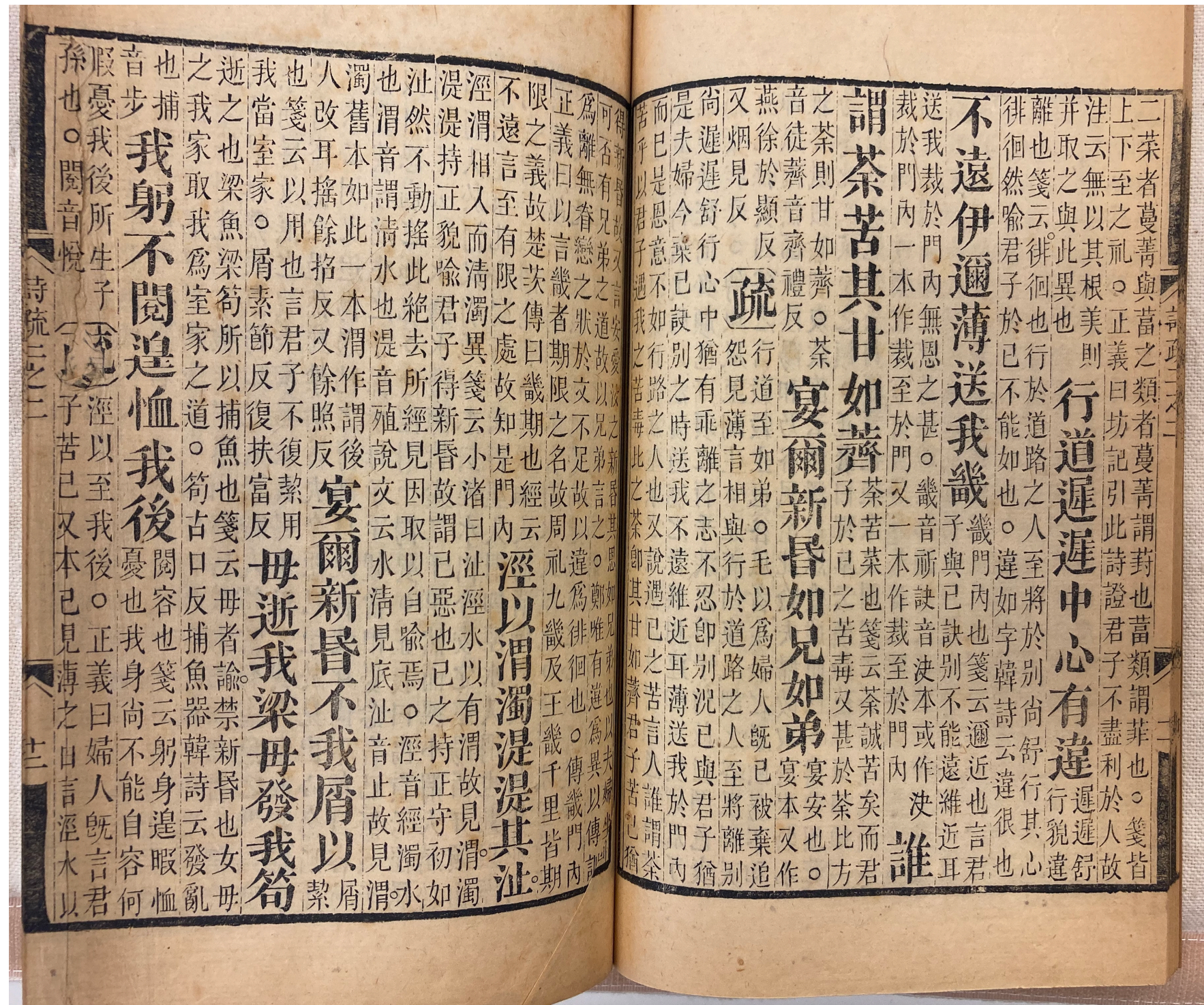
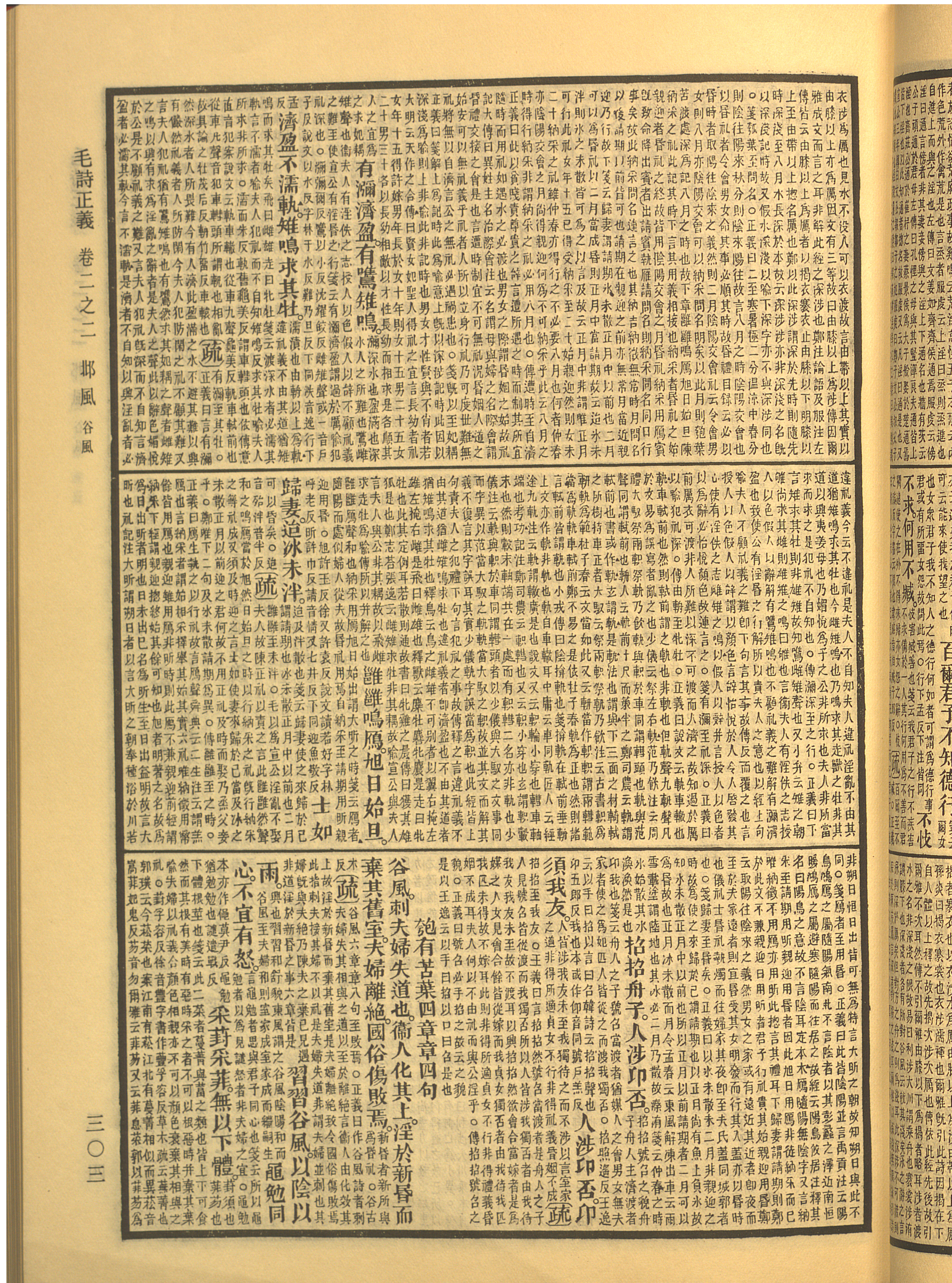


# Spécificités du corpus - l'écriture chinoise

- 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



# Spécificités du corpus - le Classique des Poèmes xylographié



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

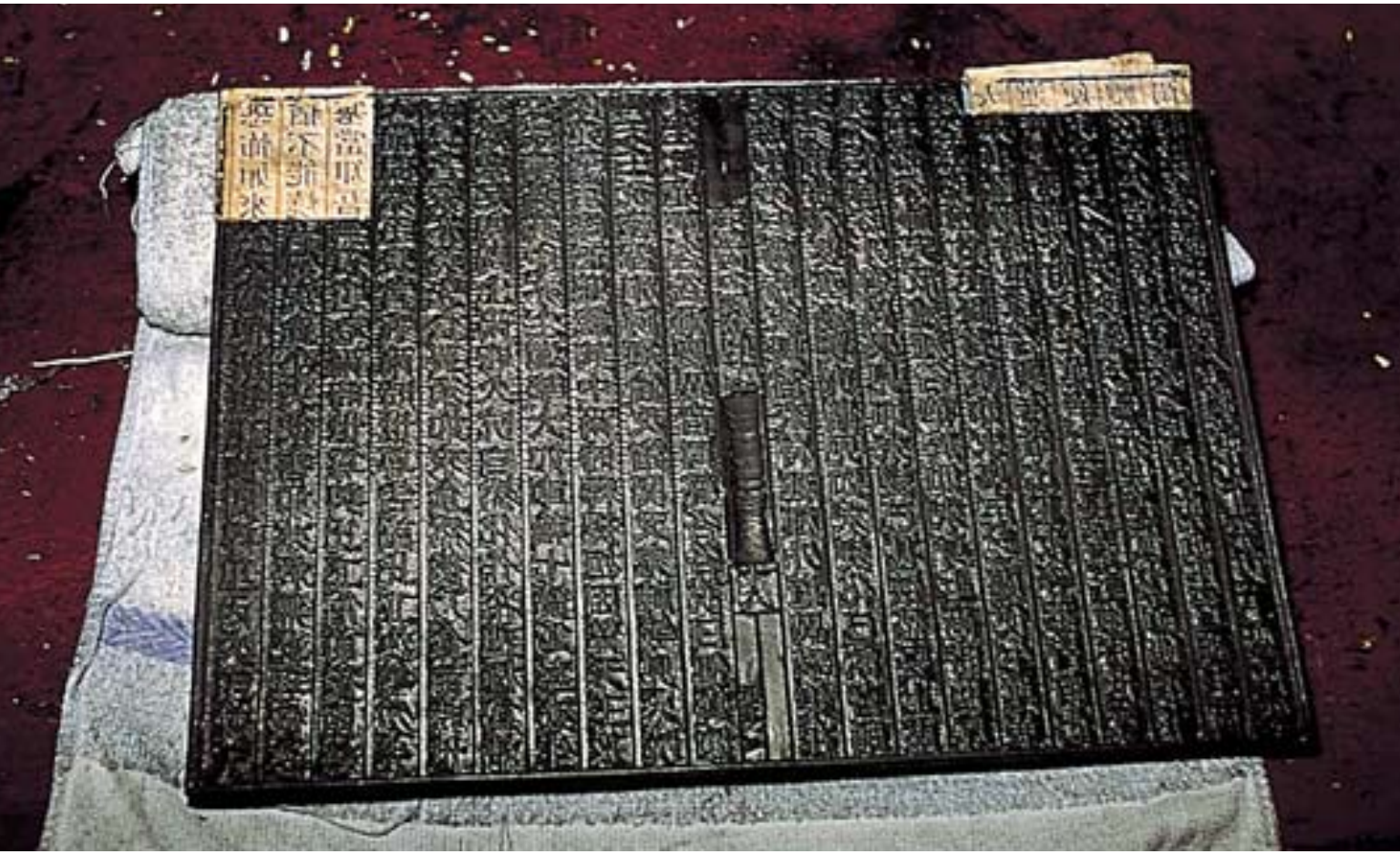


Planche à xylographier (à partir des IXe-Xe s.)

毛詩正義 卷二之一 邶風 谷風

三〇三

Édition xylographiée du Shijing utilisée reproduction par la Shanghai guji chubanshe

# Les premières expérimentations - généralités

Corpus annoté de 50 images pour les expérimentations :

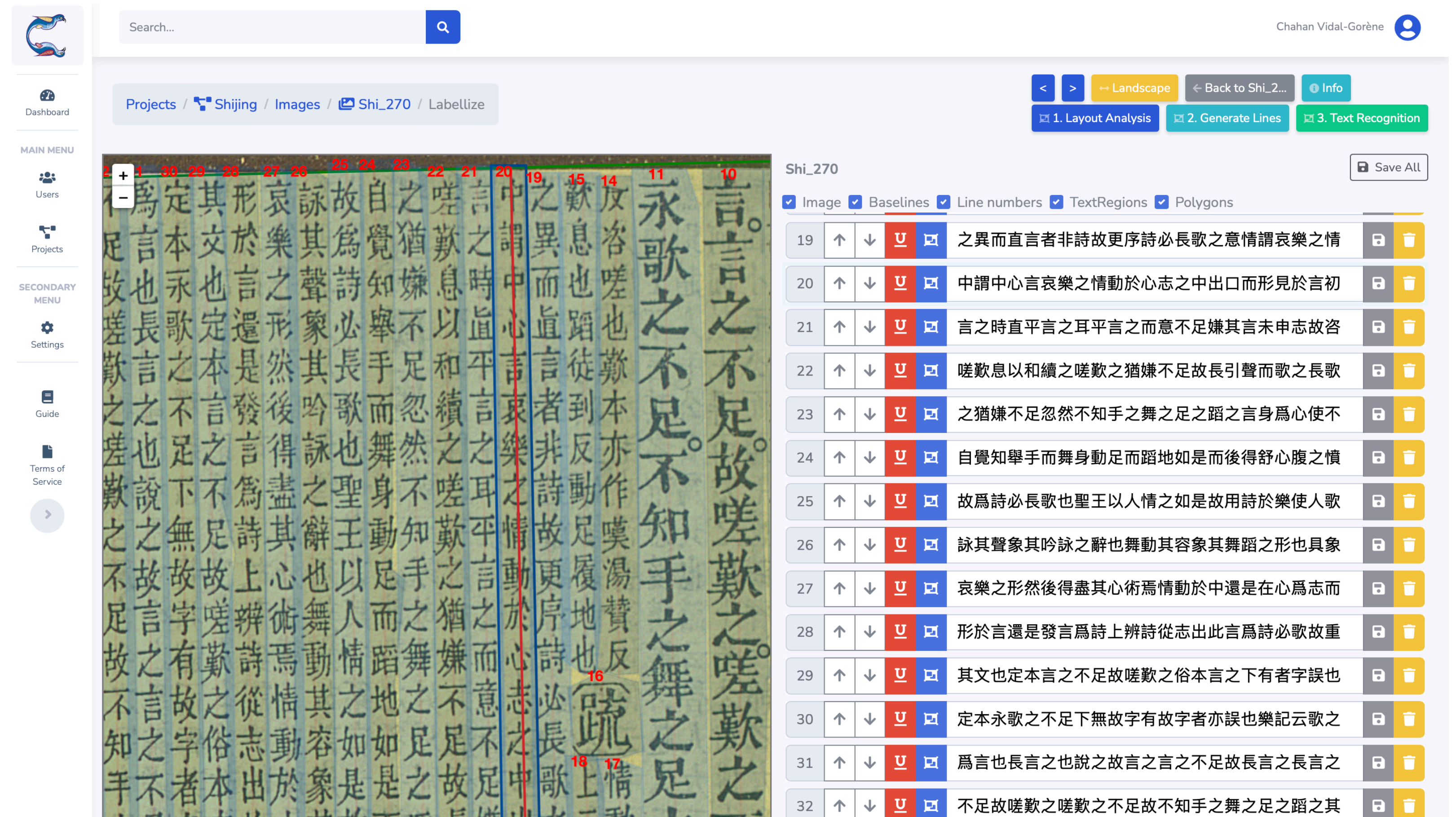
- 3 240 lignes en apprentissage
- 92 234 caractères

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)



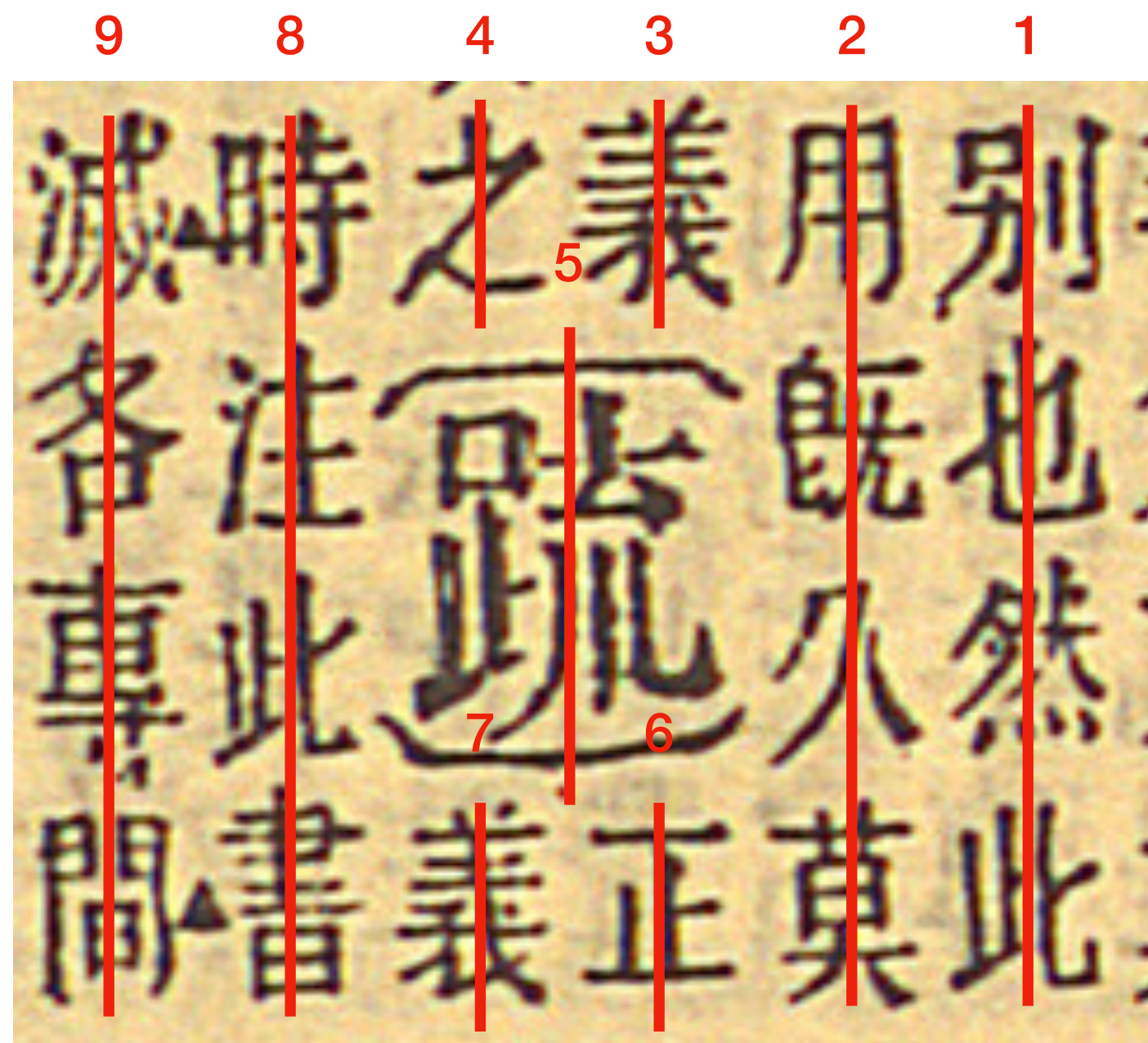
The screenshot displays the Calfa Vision web interface. On the left is a sidebar with navigation options: Dashboard, MAIN MENU (Users, Projects), and SECONDARY MENU (Settings, Guide, Terms of Service). The main area shows a document with Chinese text, with a search bar at the top and a navigation bar below it. The document is annotated with a grid of lines and columns. A table on the right lists the detected lines, with columns for line number, navigation icons, and the text content.

Line number	Navigation icons	Text content
19	↑ ↓ U	之異而直言者非詩故更序詩必長歌之意情謂哀樂之情
20	↑ ↓ U	中謂中心言哀樂之情動於心志之中出口而形見於言初
21	↑ ↓ U	言之時直平言之耳平言之而意不足嫌其言未申志故咨
22	↑ ↓ U	嗟歎息以和續之嗟歎之猶嫌不足故長引聲而歌之長歌
23	↑ ↓ U	之猶嫌不足忽然不知手之舞之足之蹈之言身為心使不
24	↑ ↓ U	自覺知舉手而舞身動足而蹈地如是而後得舒心腹之憤
25	↑ ↓ U	故為詩必長歌也聖王以人情之如是故用詩於樂使人歌
26	↑ ↓ U	詠其聲象其吟詠之辭也舞動其容象其舞蹈之形也具象
27	↑ ↓ U	哀樂之形然後得盡其心術焉情動於中還是在心為志而
28	↑ ↓ U	形於言還是發言為詩上辨詩從志出此言為詩必歌故重
29	↑ ↓ U	其文也定本言之不足故嗟歎之俗本言之下有者字誤也
30	↑ ↓ U	定本永歌之不足下無故字有故字者亦誤也樂記云歌之
31	↑ ↓ U	為言也長言之也說之故言之言之不足故長言之長言之
32	↑ ↓ U	不足故嗟歎之嗟歎之不足故不知手之舞之足之蹈之其

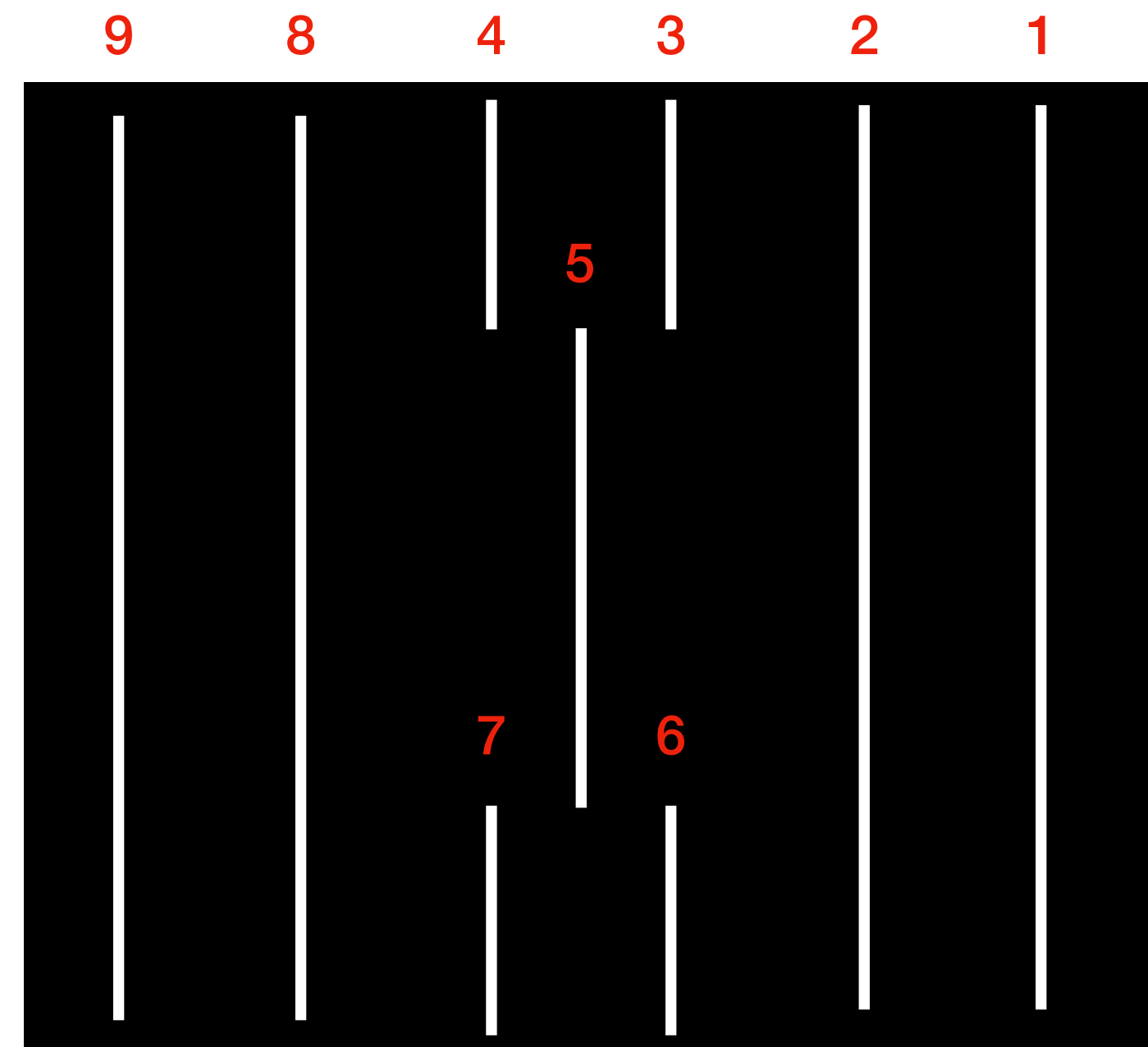
# Les premières expérimentations - l'ordre de lecture

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

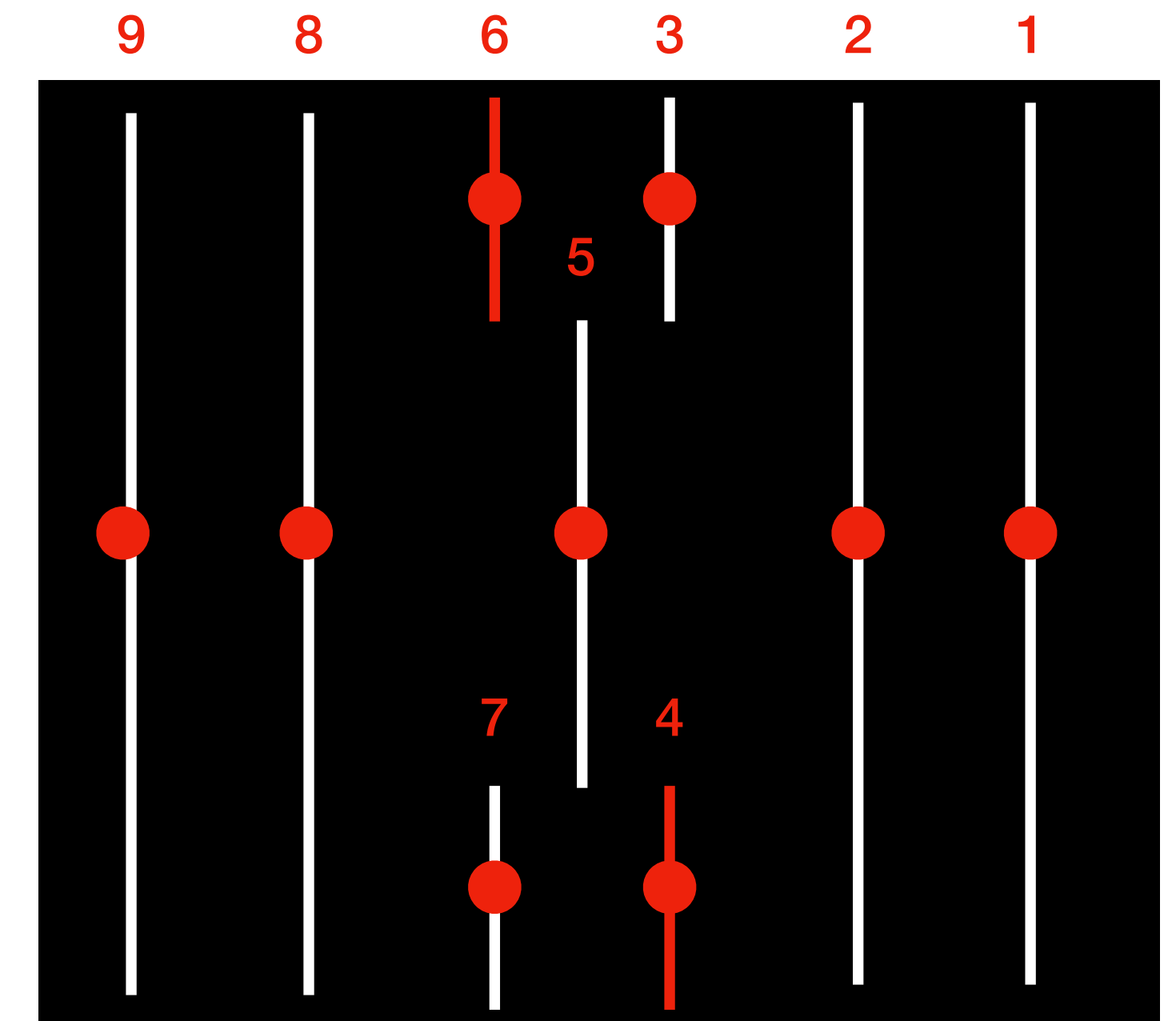
Efficace à **82,37%** sur le dataset, échoue y compris en situation simple



Détection des lignes



Sens de lecture GT

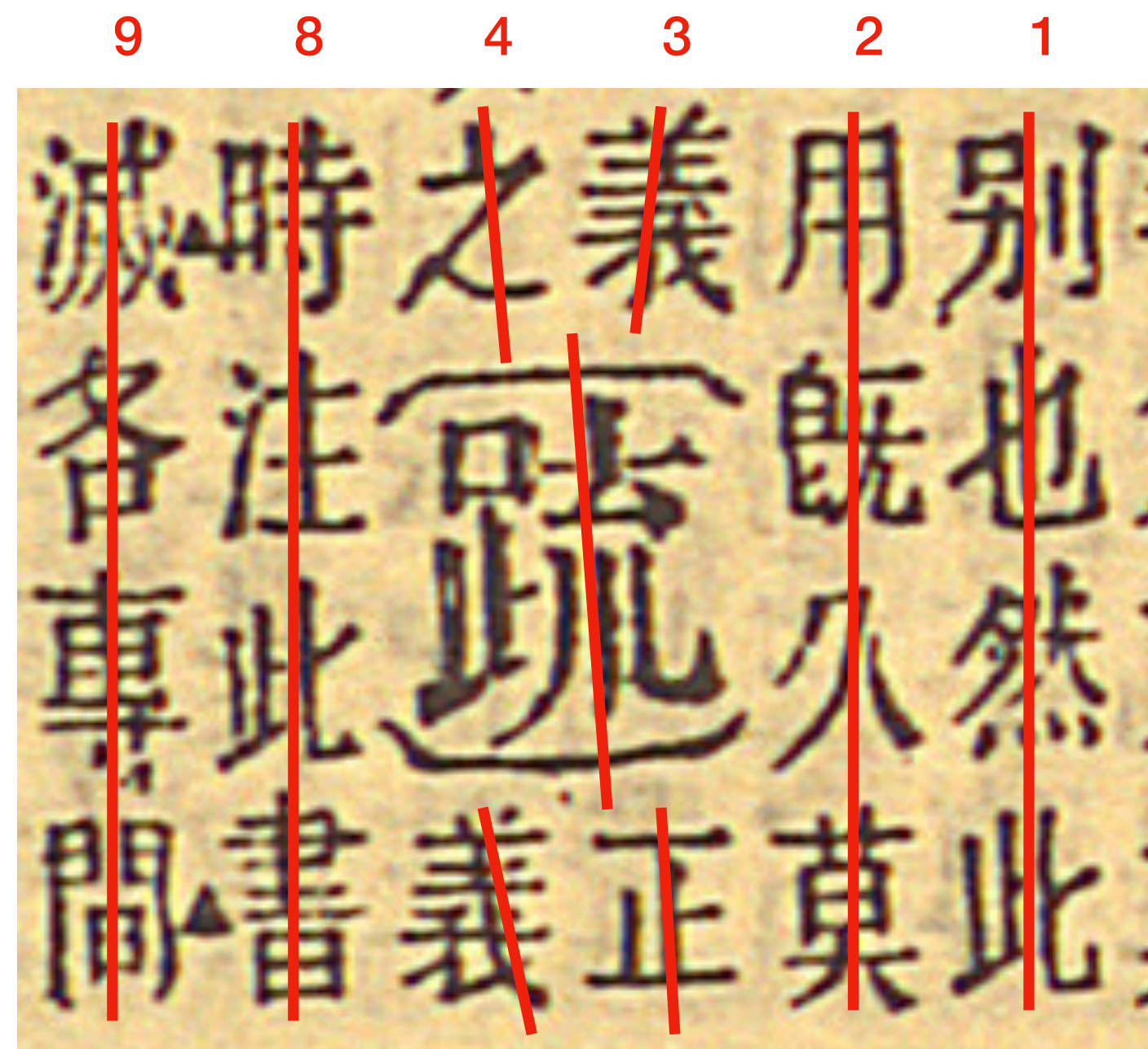


Tri des lignes par centroïdes

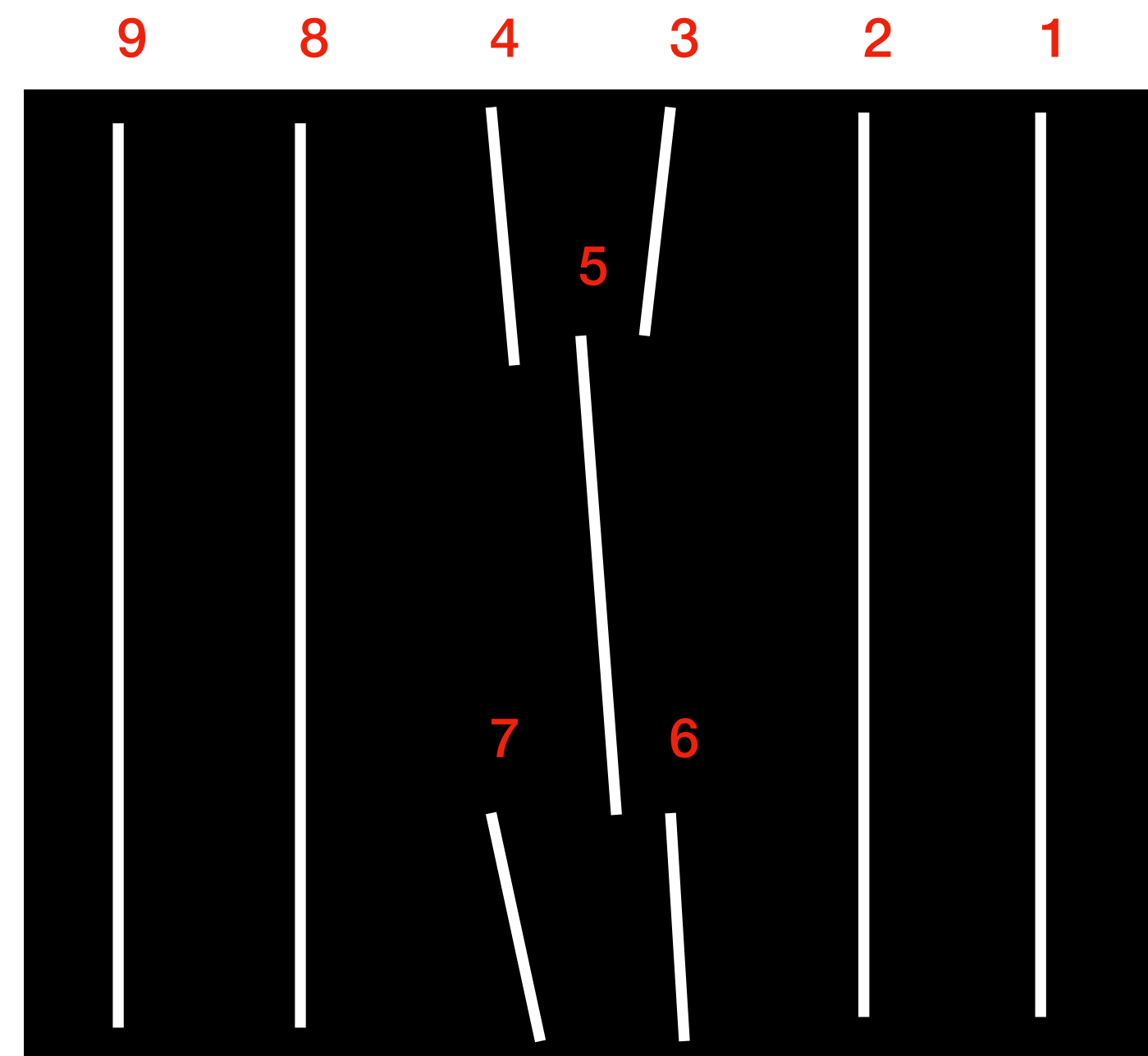
# Les premières expérimentations - l'ordre de lecture

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

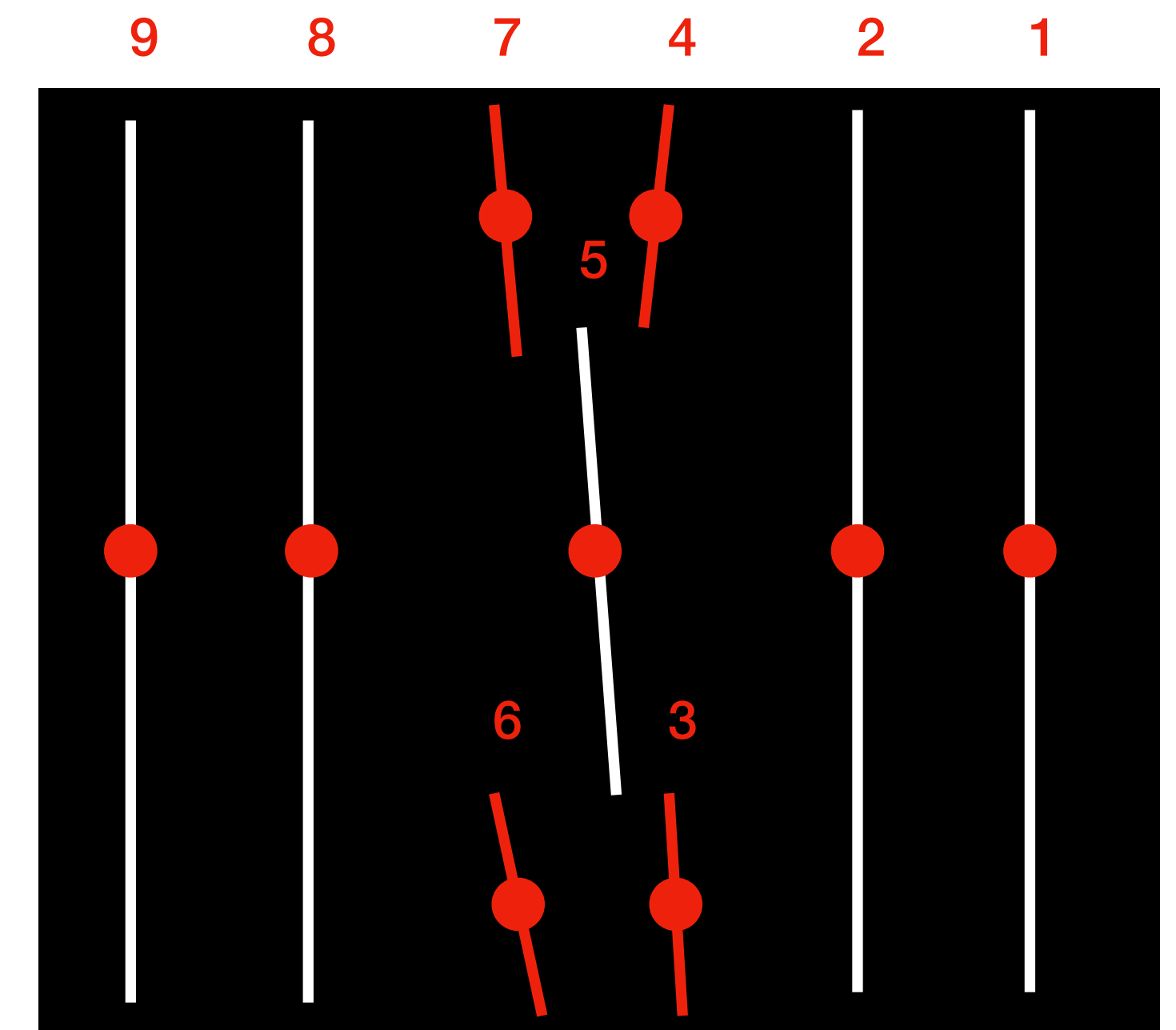
Efficace à **82,37%** sur le dataset



Détection des lignes



Sens de lecture GT



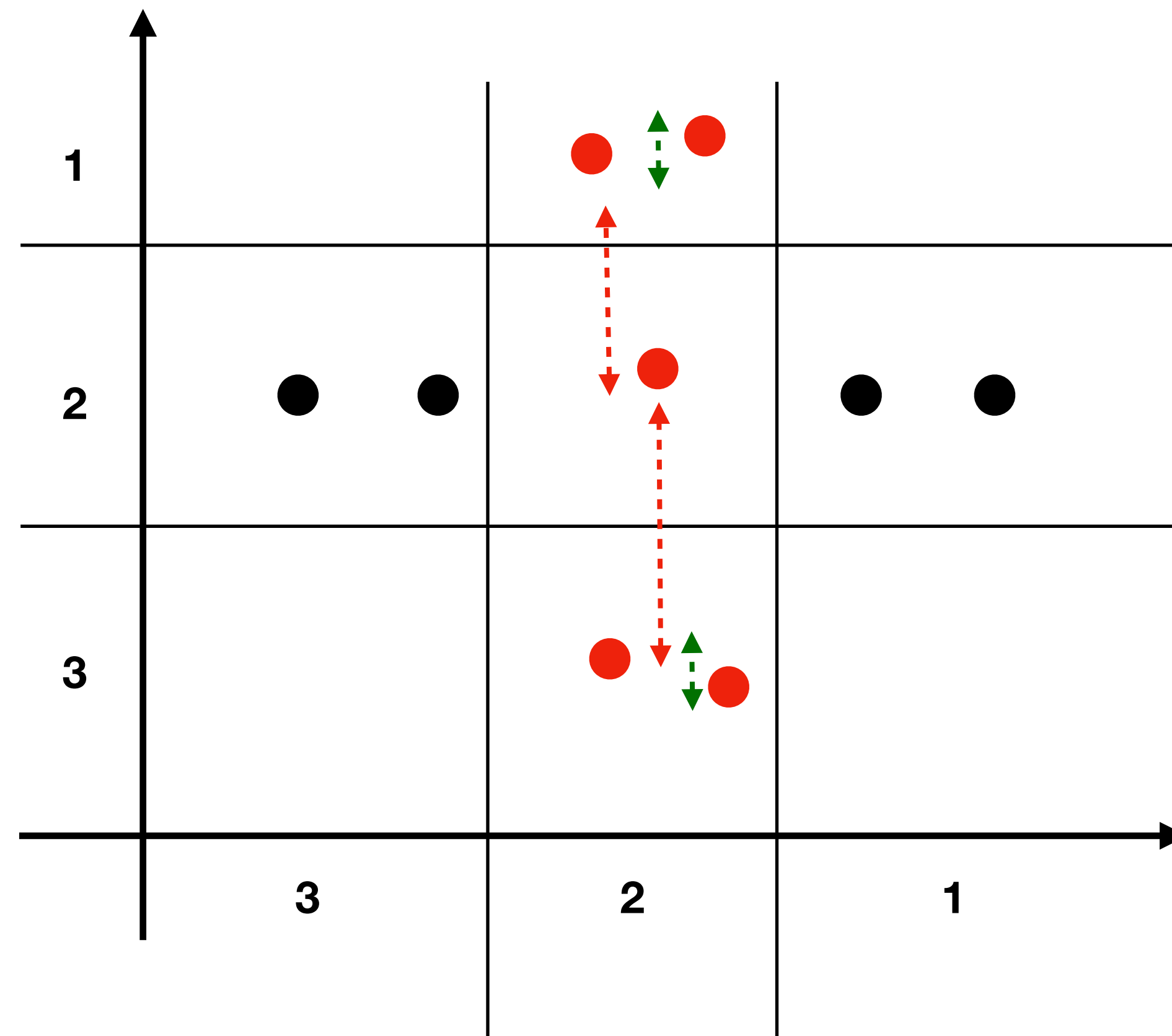
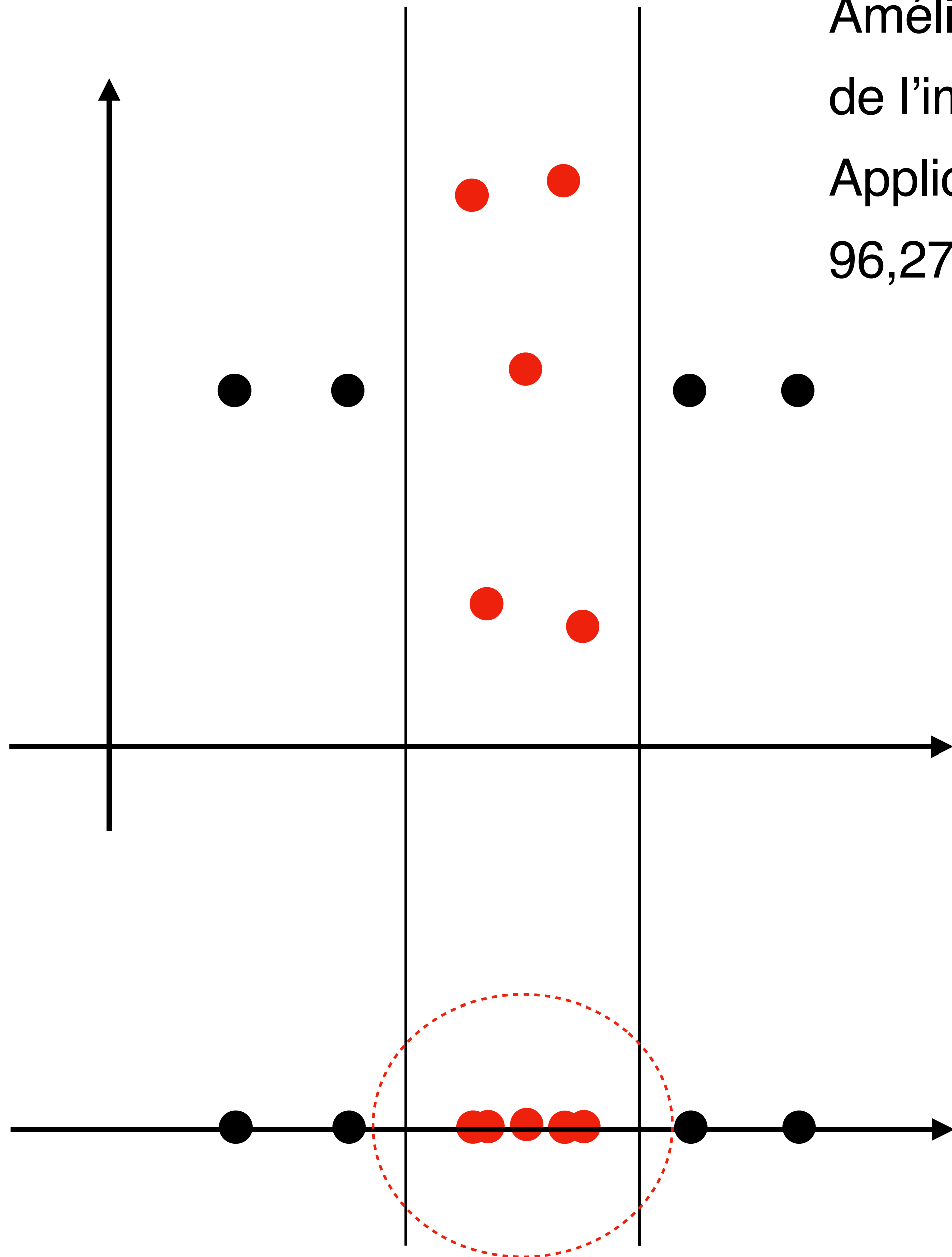
Tri des lignes par centroïdes

# 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%



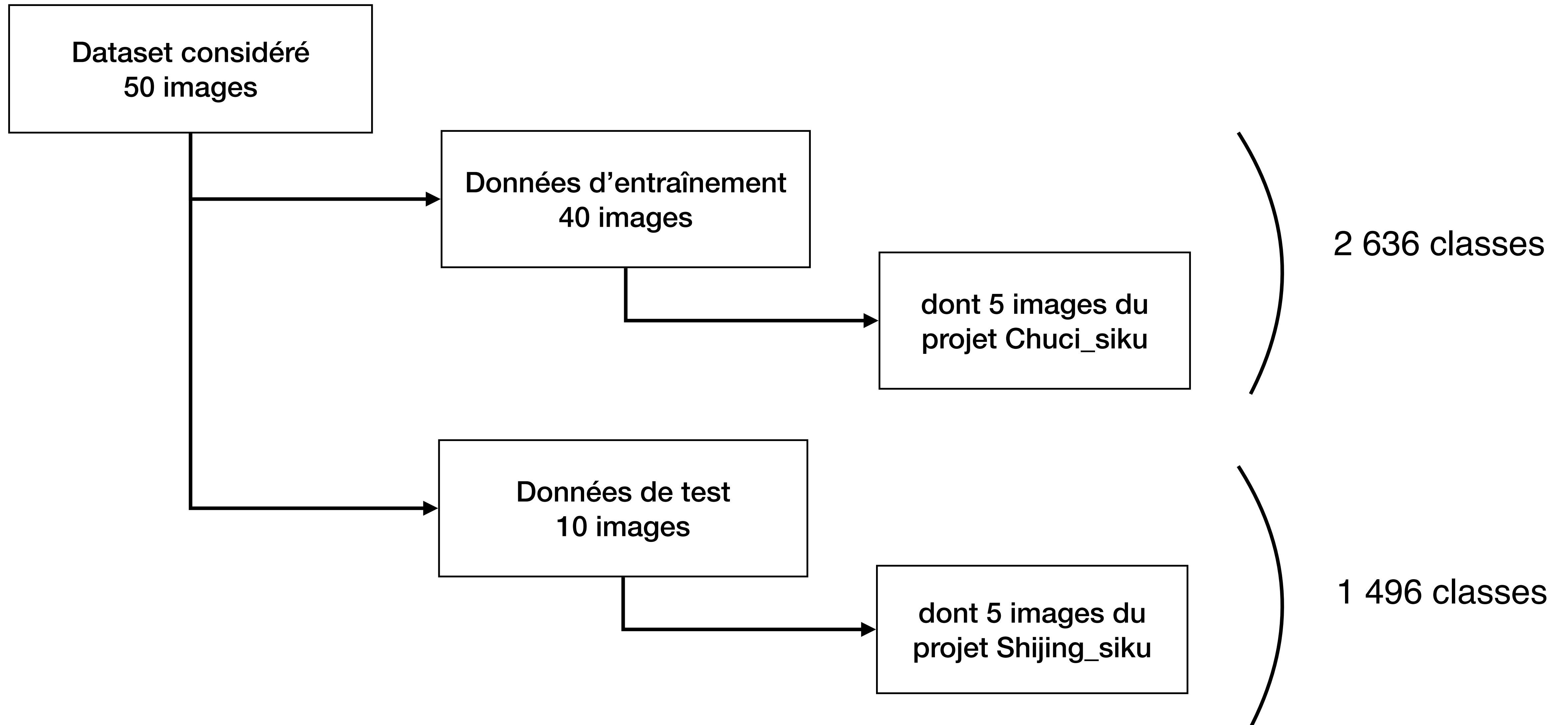
# Les premières expérimentations - l'ordre de lecture

故以莫莫為成就貌也。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為重毛意必不然。傳黃鳥至遠聞。正義曰釋鳥云皇黃鳥舍人曰皇各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鸞一名倉庚一名商庚一名鷺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為灌孫炎曰族叢也是灌為叢木也。箋葛延至遠方。正義曰知葛當延蔓之時搏黍飛鳴亦因以與者以前葛之生長是為因與則此亦宜然也言搏黍往飛集於灌木之時其鳴恒啾啾然其鳴啾啾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閭才美之稱得達遠方。葛之覃兮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。其可采之。莫莫博反。是刈是穫。為絺為綌。服之無斃。漢莫之也。精曰絺。籬曰綌。公侯夫人絺綌。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。獲胡郭反。韓詩云獲淪也音羊灼反。絺取知反葛之精者曰絺綌去逆反。數本亦作獸音亦獸於豔反本亦作厭統都覽反。統織五采如緇狀用縣瑱也絺獲耕反。纓之無綏者從下仰屬於冠經音延冕上覆也朝直遙反下同庶士謂庶人在官者本或作庶人衣。疏葛之至無斃。正義曰言葛之漸延蔓兮所移於既反。在於是谷中生長不已其葉則莫莫然成就葛既成就已可採用故后妃於是刈取之於是獲莫之。莫治已訖。后妃乃緝績之為絺為綌言后妃整治此葛以為絺綌之時志無厭倦是后妃之性貞專也。傳獲莫至其夫。正義曰釋訓云是刈是穫獲莫之也舍人曰是刈刈取之是獲莫治之孫炎曰獲莫以為絺綌以黃之於獲故曰獲莫非訓獲為黃曲禮云為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。厭釋詁文彼數作射音義同自王后織紵以下皆魯語敬姜之言也。統縣瑱之物織五采為之故著箋云人君五色則天子之統五色獨言玄者以玄為尊故舉以言焉。絺者纓之無綏從下而上者也祭義曰天子冕而朱紵諸侯冕而青紵此諸侯當以青為組在冕下仰屬於冠禮註云有笄者屈組為紵垂為飾無笄者纓而結其條是也。經者冕上覆玄表纁裏是也。內子卿之適妻儻二十四年左傳趙姬請以叔隗為內子而已下之是也。大帶者玉藻所云大夫以玄華黃也。以素為帶飾之外以玄內以黃也。大夫命婦成祭服者大夫助祭服玄冕受之於君故大宗伯命婦受服是也。妻所成者自祭之服少牢禮朝服玄冠緇布衣素裳韋昭云祭服玄衣纁裳謂作玄冕之服非也。士妻朝服者作朝於君服亦玄衣纁裳也。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%

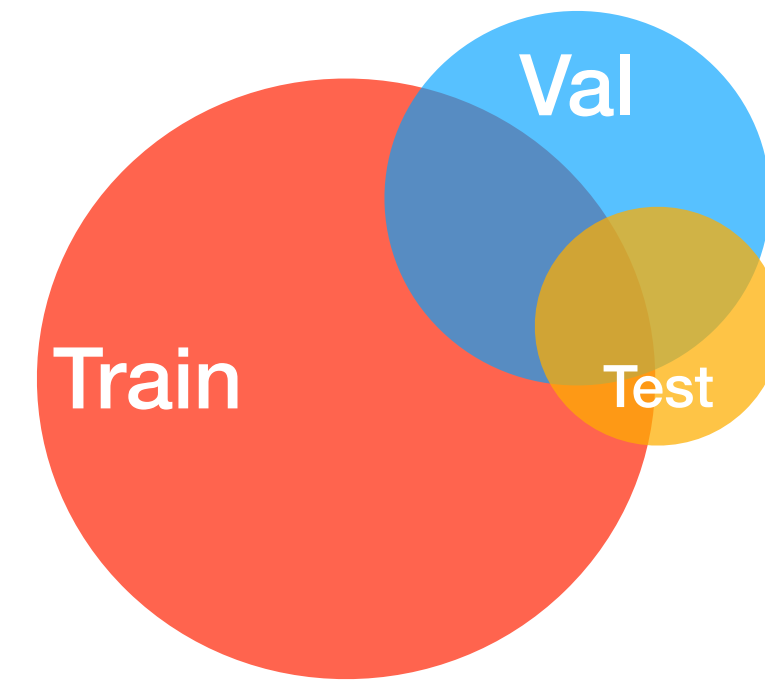
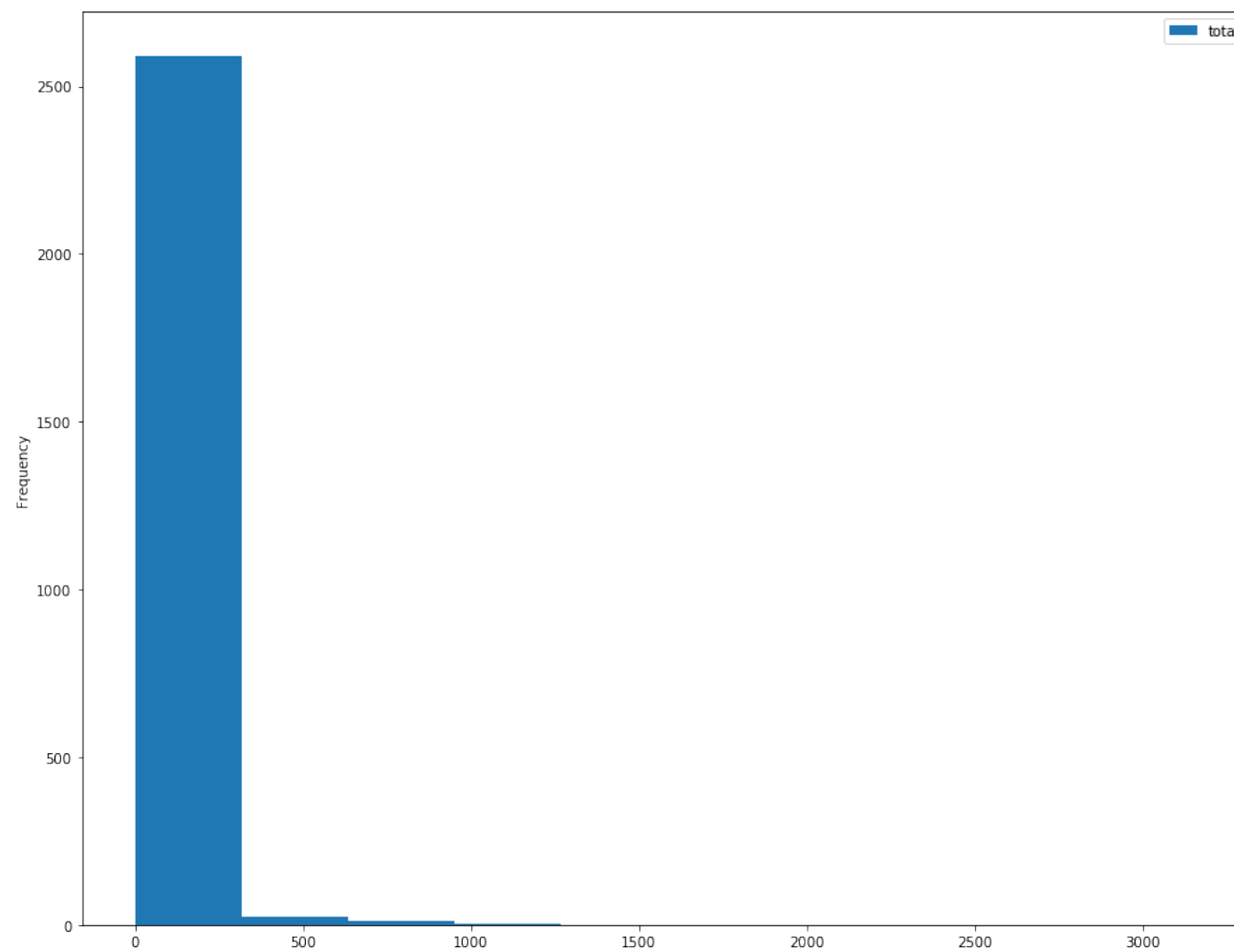
# Les premières expérimentations - la question des classes

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



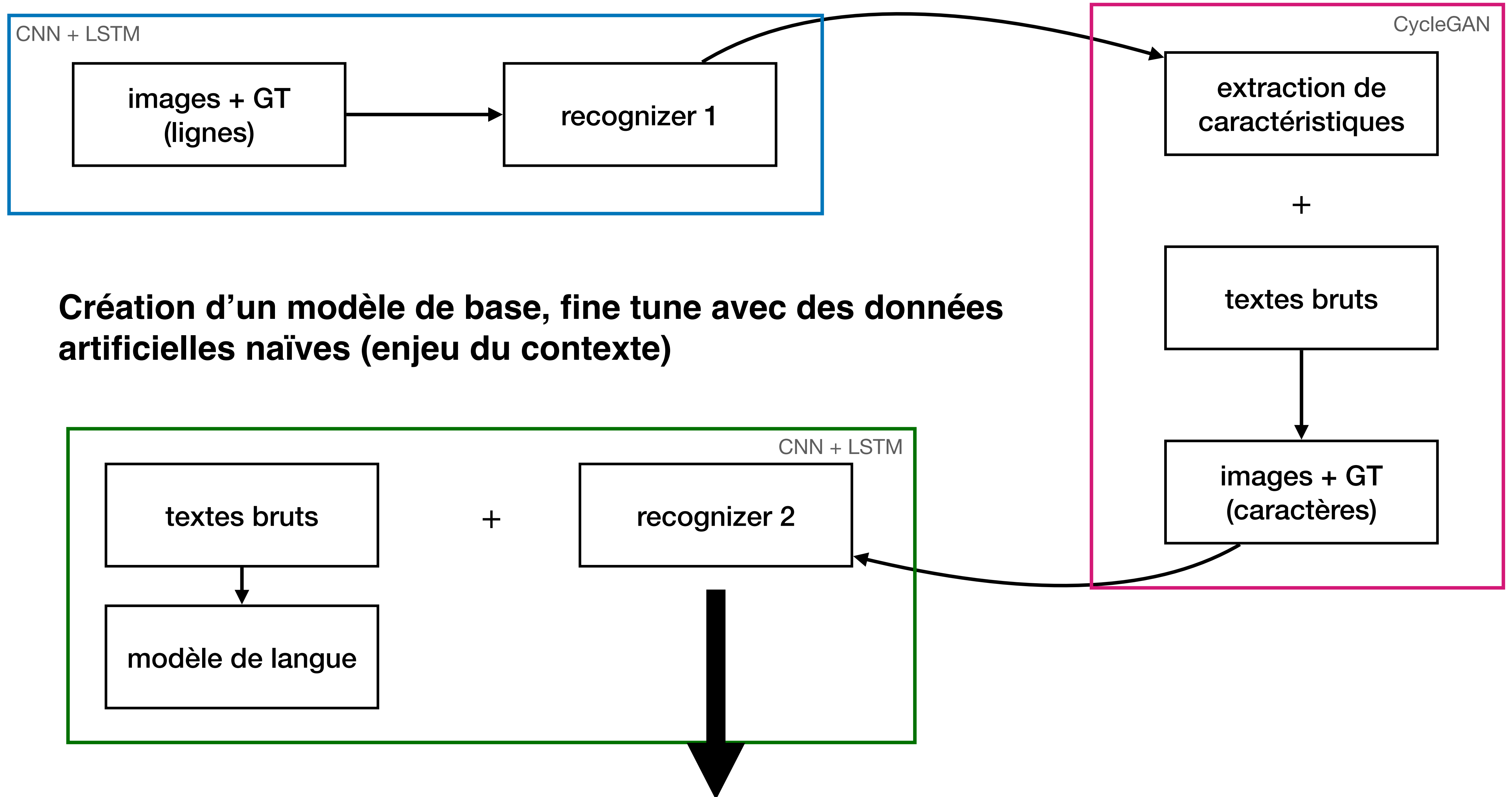
之	3170	芋	1
也	2032	恕	1
以	1463	刃	1
云	1110	訪	1
不	1074	隈	1
言	960	沔	1
者	944	斐	1

**2 095 classes sur 2636 avec un seul échantillon**

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

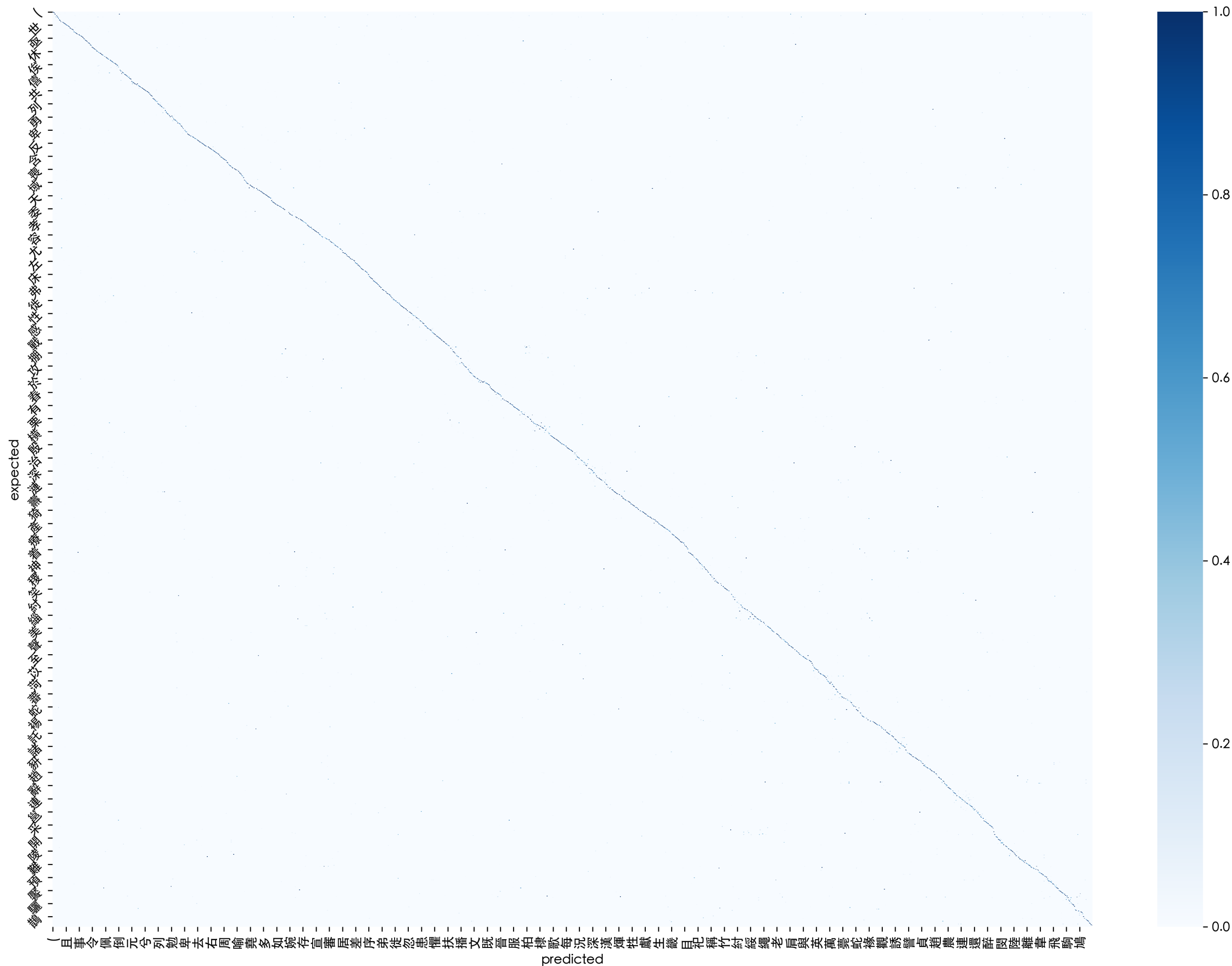
**==> Situation de *one-shot learning* + classes non vues en apprentissage**

# Les premières expérimentations - la question des classes





# Les premières expérimentations - la question des classes



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

Posent problème :

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

為 / 為

茶 / 菜

苻 / 持

已 / 己

千 / 干

旋 / 施

## 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

## Exemples :

爲 為

既 旣

即 卽

總 摠

**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)

## Outil utilisé :

- [vision.calfa.fr](http://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