Le projet CollEx CHI-KNOW-PO CORPUS : acquérir un large corpus pour analyser la circulation des savoirs dans la Chine médiévale

Marie Bizais-Lillig
Université de Strasbourg - USIAS - consortium Distam (Huma-Num, CNRS) - porteuse du projet COLLEX CHI-KNOW-PO

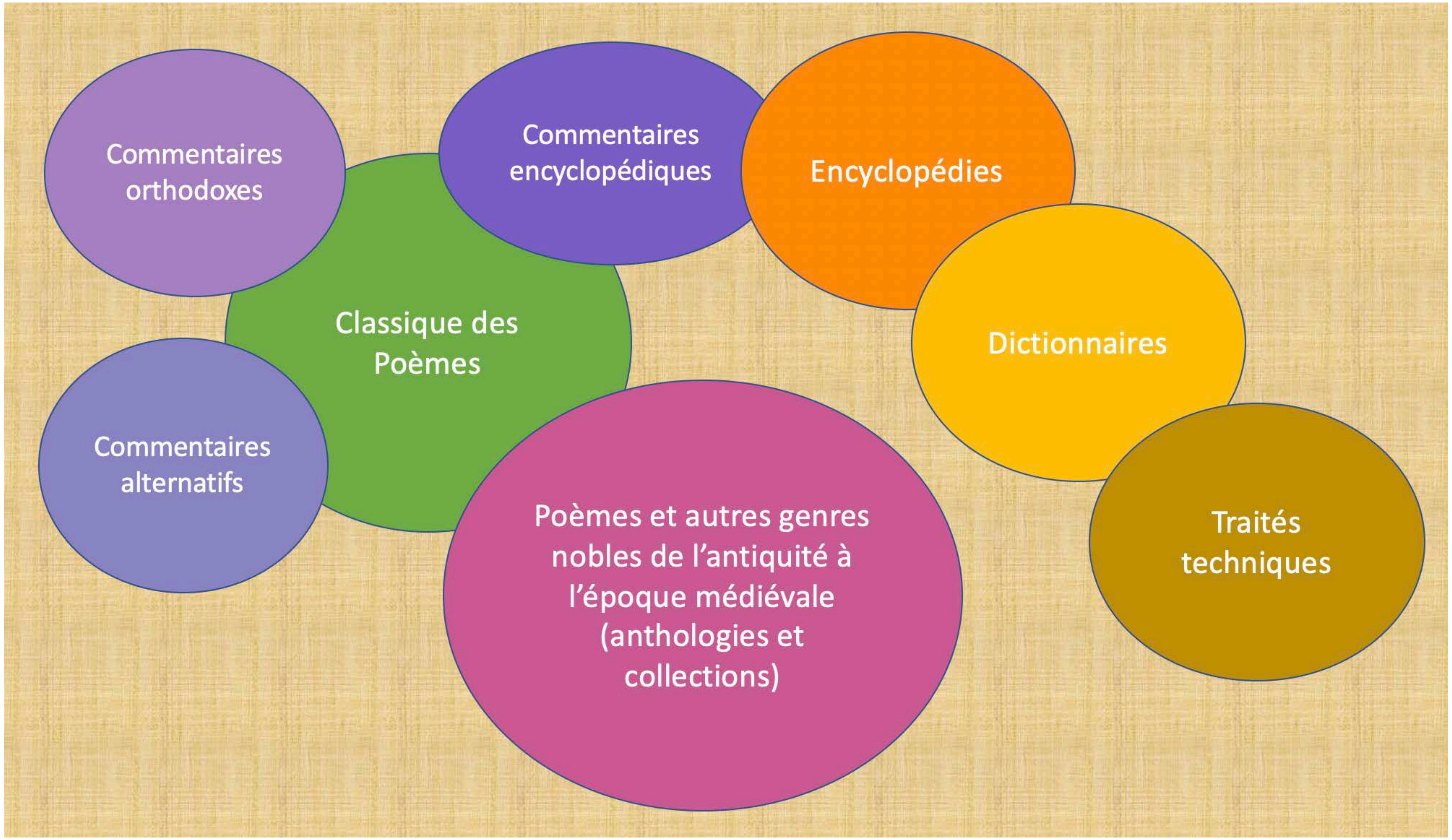
D’un mot

- Dates : juillet 2022 - juin 2024
- Objectifs : numérisation, HTR, édition, archivage, exposition
- Partenaire : Calfa pour l’HTR

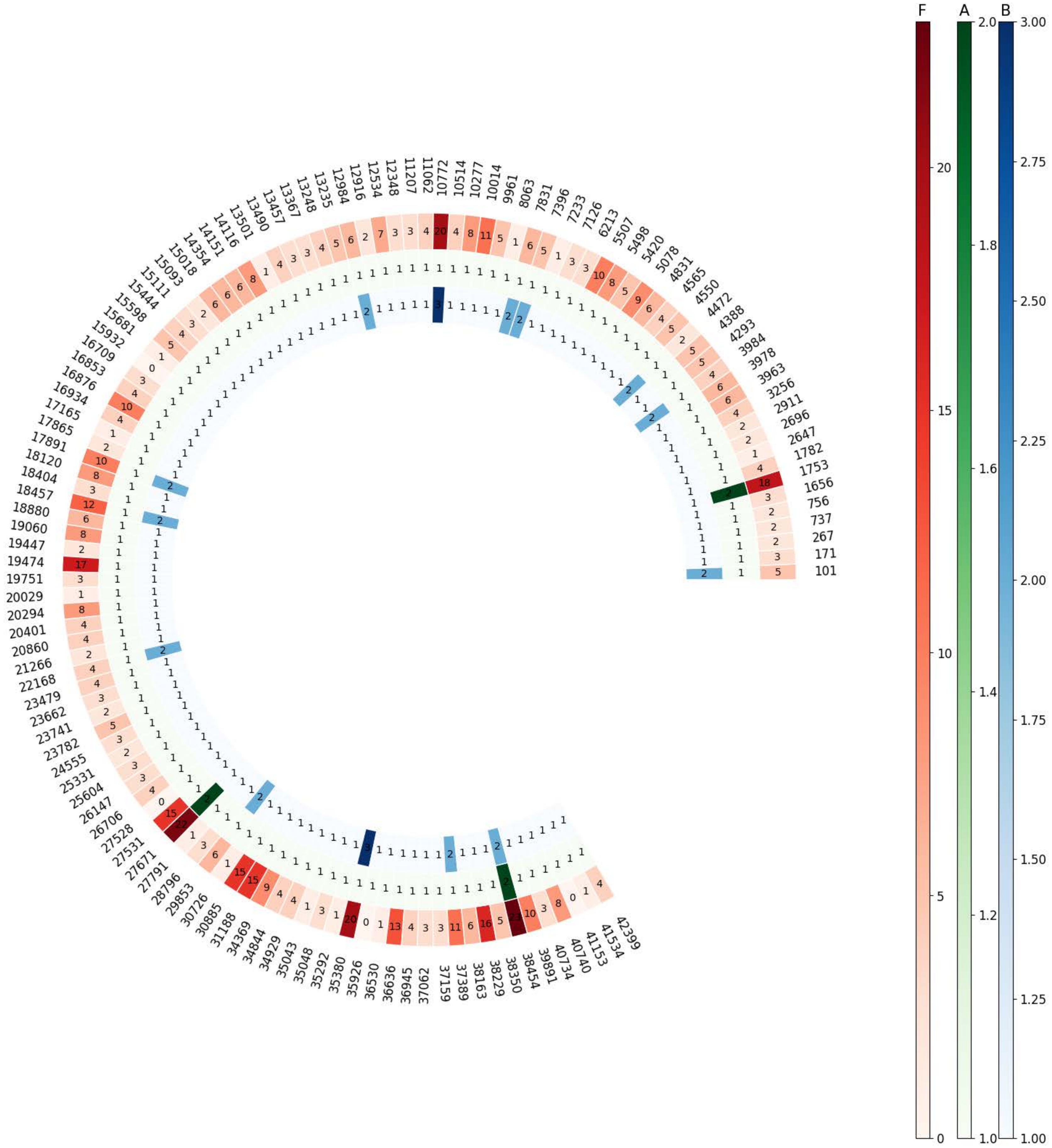
- Co-porteuses : Marie Bizais-Lillig (Université de Strasbourg) et Soline Lau-Suchet (BULAC)
- Collections concernées : BULAC (fonds chinois ancien), BIHEC (Collège de France), BNU (fonds JP et Colette Diény)
- 1 ouvrage de la Bibliothèque d’études chinoises de l’Université de Strasbourg (don du Collège de France) ne sera finalement pas pris en compte.

1. Pourquoi ? Les besoins de la recherche

Cadre : Le projet CHI-KNOW-PO



Explorations : cooccurrences



孟浩然 《夜泊廬江，聞故人在東寺，以詩寄之》

poem 5 of 81, no. 6703 (40 words)

江路經廬阜，松門入虎溪。聞君尋寂樂，清夜宿招提。
石鏡山精怯，禪枝怖鴿棲。一燈如悟道，為照客心迷。

孟浩然 《李少府與楊九再來》

poem 6 of 81, no. 6785 (39 words)

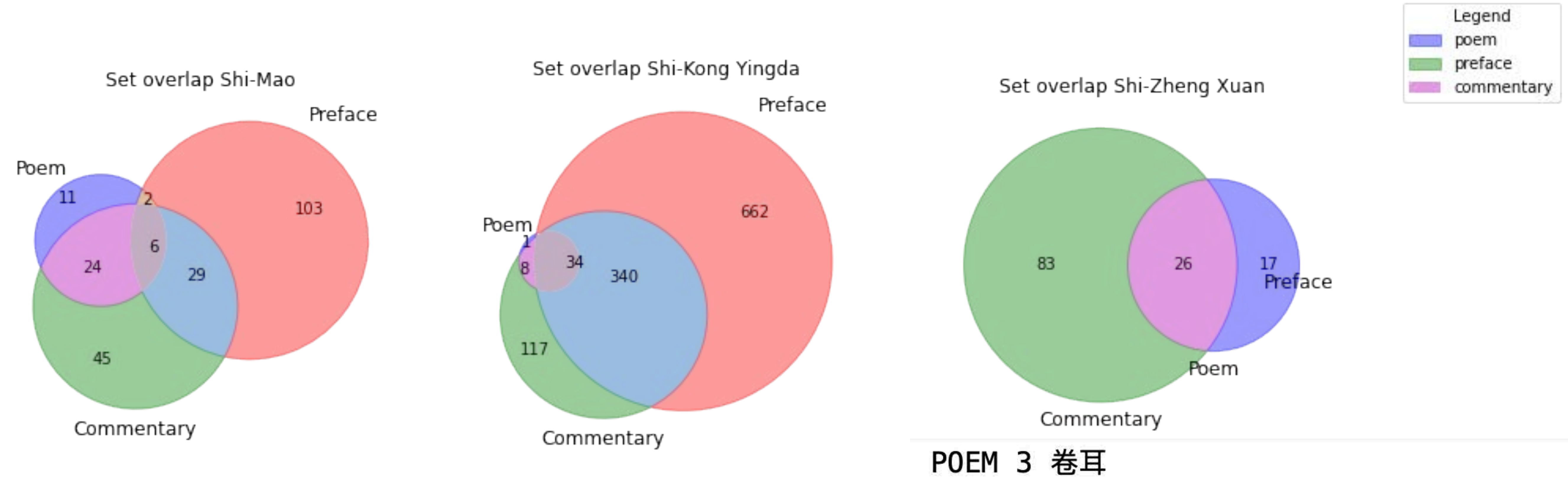
弱歲早登龍，今來喜再逢。如何春月柳，猶憶歲寒松。
煙火臨寒食，笙歌達曙鐘。喧喧鬥雞道，行樂羨朋從。

李白 《永王東巡歌十一首》

poem 7 of 81, no. 7037 (302 words)

永王正月東出師，天子遙分龍虎旗。樓船一舉風波靜，江漢翻為雁鶩池。
三川北虜亂如麻，四海南奔似永嘉。但用東山謝安石，為君談笑靜胡沙。
雷鼓嘈嘈喧武昌，雲旗獵獵過尋陽。秋毫不犯三吳悅，春日遙看五色光。
龍蟠虎踞帝王州，帝子金陵訪古丘。春風試暖昭陽殿，明月還過鳷鵲樓。
二帝巡遊俱未回，五陵松柏使人哀。諸侯不救河南地，更喜賢王遠道來。
丹陽北固是吳關，畫出樓臺雲水間。千岩烽火連滄海，兩岸旌旗繞碧山。
王出三山按五湖，樓船跨海次陪都。戰艦森森羅虎士，征帆一一引龍駒。
長風掛席勢難回，海動山傾古月摧。君看帝子浮江日，何似龍驤出峽來。

Explorations : reprises



POEM 3 卷耳
MAO

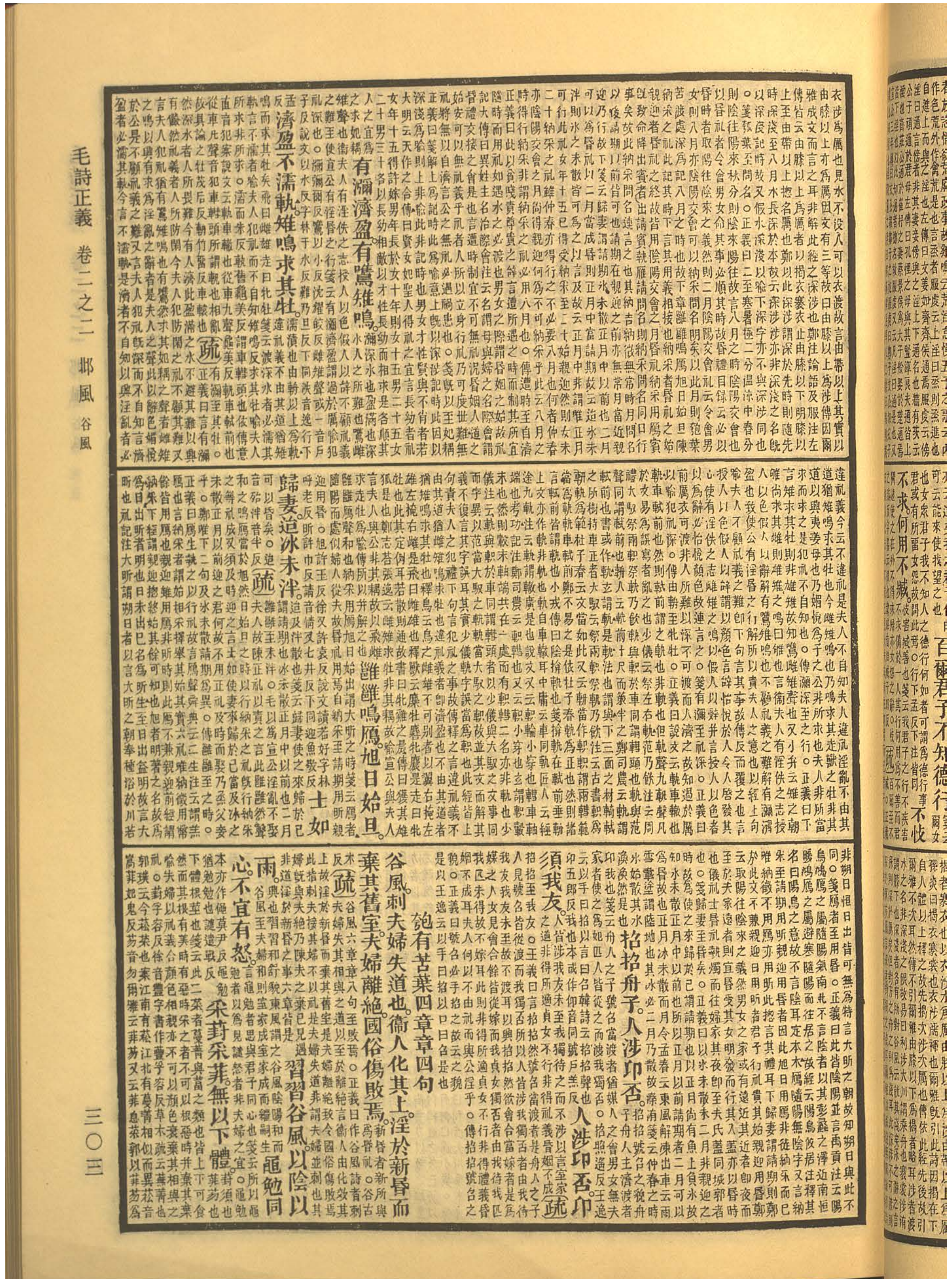
卷耳后妃之志也又當輔佐君子求賢審官知臣下之勤勞內有進賢之志而無險詖私謁之心朝夕思念至於憂勤也憂者之興也采采事采之也卷耳苓耳也頃筐畚屬易盈之器也懷思寘置行列也思君子官賢人置周之列位陟升也崔嵬土山之戴石者虺隤病也姑且也人君黃金罍永長也山脊曰岡玄馬病則黃兕觥角爵也傷思也石山戴土曰砠瘏病也痡亦病也籲憂也

++++++++++

KYD

作卷耳詩者言后妃之志也后妃非直憂在進賢躬率婦道又當輔佐君子其志欲令君子求賢德之人審置於官位復知臣下出使之勤勞欲令君子賞勞之內有進賢人之志唯有德是用而無險詖不正私請用其親戚之心又朝夕思此欲此君子官賢人乃至於憂思而成勤此是后妃之志也言又者繫前之辭雖則異篇而同是一人之事故言又為亞次也輔佐君子總辭也求賢審官至於憂勤皆是輔佐君子之事君子所專后妃志意如然故云后妃之志也險詖者情實不正譽惡為善之辭也私謁者婦人有寵多私薦親戚故厲王以豔妻方煽七子在朝成湯謝過婦謁盛與險詖私謁是婦人之常態聖人猶恐不免后妃能無此心故美之也至於憂勤勤為勞心憂深不已至於勞勤后妃之篤志也至於憂勤即首章上二句是也求賢審官即首章下二句是也經敘倒者敘見后妃求賢而憂勤故先言求賢經主美后妃之志能為此憂勤故先言其憂也言有人事采此卷耳之菜不能滿此頃筐頃筐易盈之器而不能滿者由此人志有所念憂思不在於此故也此采菜之人憂念之深矣以興后妃志在輔佐君子欲其官賢賞勞朝夕思念至於憂勤其憂思深遠亦如采菜之人也此后妃之憂為何事言后妃嗟呼而嘆我思君子官賢人欲令君子置此賢人於彼周之列位以為朝廷臣也我者后妃自我也(ZX S2 E1+E2)下箋云我我使臣我我君此不解者以詩主美后妃故不特言也言彼者后妃主求賢人為此故以周行為彼也不云興也而云憂者之興明有異於餘興也餘興言采菜即取采菜喻言生長即以生長喻此言采菜而取憂為興故特言憂者之興言興取其憂而已不取其采菜也言事采之者言勤事采此菜也此與芣苢俱言采采彼傳云非一辭與此不同者此取憂為興言勤事采菜尚不盈筐言其憂之極故云事采之彼以婦人樂有子明其采者眾故云非一辭其實采采之義同故鄭志答張逸

Fondements : HTR 2021



CALFA

- Calfa
- *Maoshi Zhengyi* 毛詩正義
- Corpus annoté de 50 images pour les expérimentations :

  - 3 240 lignes en apprentissage

  - 92 234 caractères

- Intégration des variantes dans la transcription
- 4 annotateurs : Shueh-Ying Liao, Weihang Wu, Hugo Dubois-Mouro, Marie Bizais-Lillig

Publier les textes, en relation avec des éditions

- Un site Web (Plateforme ESTRADES, avec l'aide de Pikselkraft)
- Donner à voir les images (manifestes IIIF) et les textes structurés (XML dans BaseX)
- Offrir des possibilités de circulation à travers le corpus grâce à des index liés à notre base de données
- Renvoyer vers un espace de correction TACT
- Donner accès aux fichiers sources

Structurer et éditer les textes

- Conservation d’une partie des données du XML-Alto pour l’alignement texte/image

- Structuration des textes issus de l’HTR sur la base des types de régions et colonnes selon un schéma TEI préalablement défini

- Enrichissement des fichiers XML : balisage des noms de personnes et titres (cf. base de données)

```xml
    <div resp="#KYD" type="subcommentary">
     <div source="#sj034 #sj034_com_Mao">
      <head>疏「《匏有苦葉》四章，章四句」至「淫亂」。正義曰：</head>
      <ab>並為淫亂，亦應刺夫人，獨言宣公者，以詩者主為規諫君，故舉君言之，其實亦刺夫人也。故經首章、三章責公不依
之。</ab>
     </div>
     <div source="#sj034_com_ZX">
      <head>箋「夫人謂夷姜」。正義曰：</head>
      <ab>知非宣姜者，以宣姜本適伋子，但為公所要，故有魚網離鴻之刺。此責夫人，云「雉鳴求其牡」，非宣姜之所為，明
     </div>
    </div>
   </div>
   <div type="poem" xml:id="sj034">
    <lg type="stanza" xml:id="sj034s01">
     <lg type="couplet" xml:id="sj034s01e1">
      <l xml:id="sj034s01l1">匏有苦葉，</l>
      <l xml:id="sj034s01l2">濟有深涉。</l>
     </lg>
     <lg type="couplet" xml:id="sj034s01e2">
      <l xml:id="sj034s01l3">深則厲，</l>
      <l xml:id="sj034s01l4">淺則揭。</l>
     </lg>
    </lg>
```

2. La préparation du corpus

Définition du corpus

Belles Lettres

- *Wenxuan* 文選 (commenté par Li Shan (1809), commentée par les 6 ministres (1923)) — BNU
- *Yutai xinyong* 玉臺新詠 (1879) — BIHEC (V XIV 69 (1-8))
- *Quan Tang shi* 全唐詩 (1707) — BIHEC (SB 4002 (1-12) (1-120))
- *Yuefu shiji* 樂府詩集 / Wikisource

Textes techniques

- *Gujin shiwen leiju* 古今事文類聚: (1604) — BIHEC(SB 3705 (1-26))
- *Qimin yaoshu* 齊民要術 (1896) — BIHEC (V I 22 (1-4))
- *Xinzhai shizhong* 心齋十種 (1785/1788) — BIHEC (V I 53 (1))
- *Shennong bencao jing jizhu* 神農本草經集注 / Wikisource

Définition du corpus

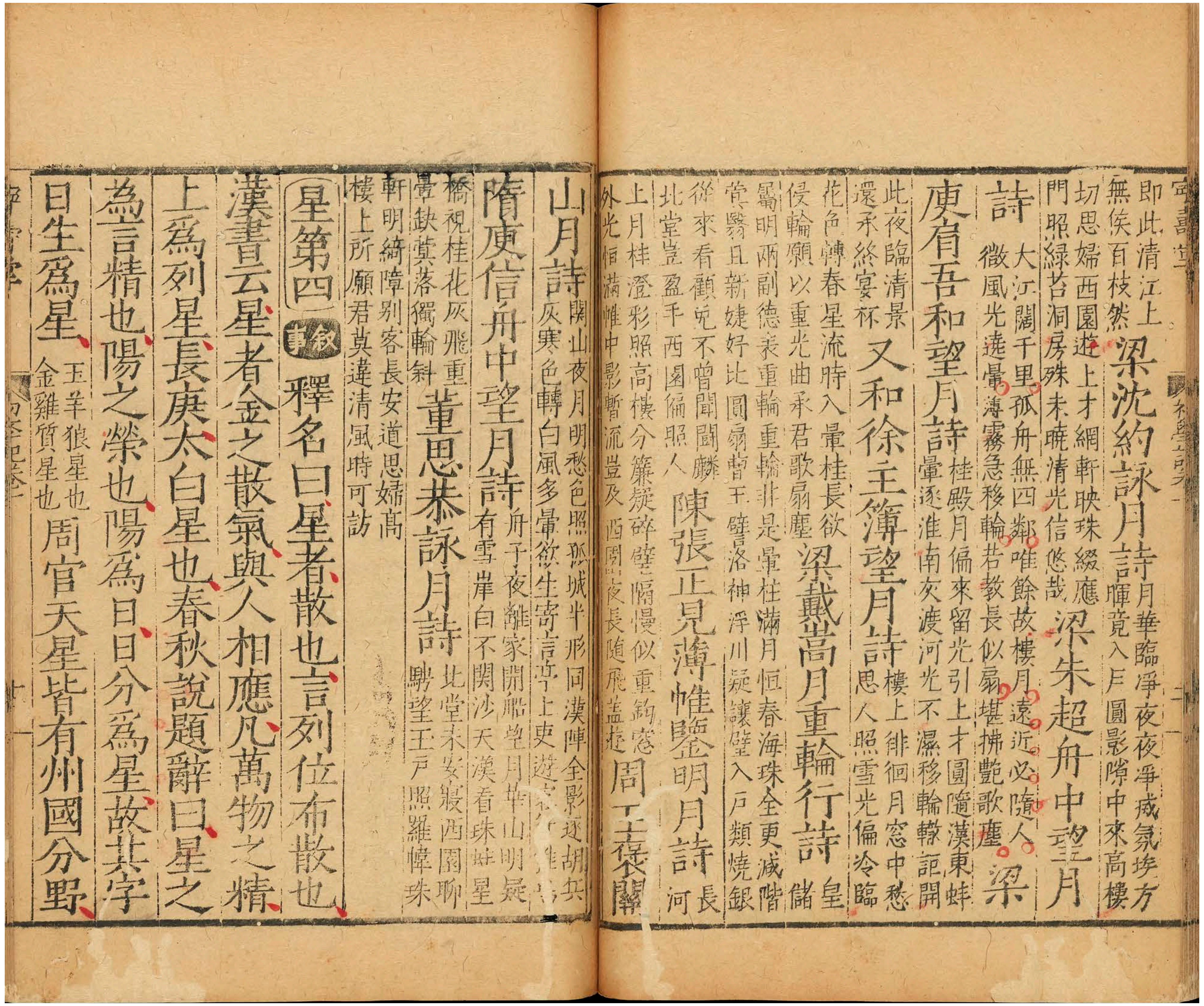
Textes de savoirs

- *Beitang shuchao* 北堂書鈔 (1888) — BULAC (BIULO CHI.1087)
- *Bowu zhi* 博物志 (1875) — BULAC (BIULO CHI.1140)E*rya yintu* 影宋鈔繪圖爾雅 (1801) — BULAC(BIULO CHI.1938(1)-(3))
- *Mao Shi caomu niaoshou chongyu shu* 毛詩草木鳥獸蟲魚疏 (1857) — BIHEC Paris (V I 111 (1) 5)

Textes de savoirs

- *Yiwen leiju* 藝文類聚 (1879) — BIHEC (CIII 5-7 (1-8))
- *Chuxue ji* 初學記 (?) — BIHEC in Paris (SB 3701 (1-2))
- *Zhi bu zu zhai cong shu* 知不足齋叢書 (1921) — BIHEC (F X 2 (1-15) 1-120)
- *Mao Shi zhengyi* 毛詩正義 / Wikisource
- *Shuowen jiezi* 說文解字 / Wikisource

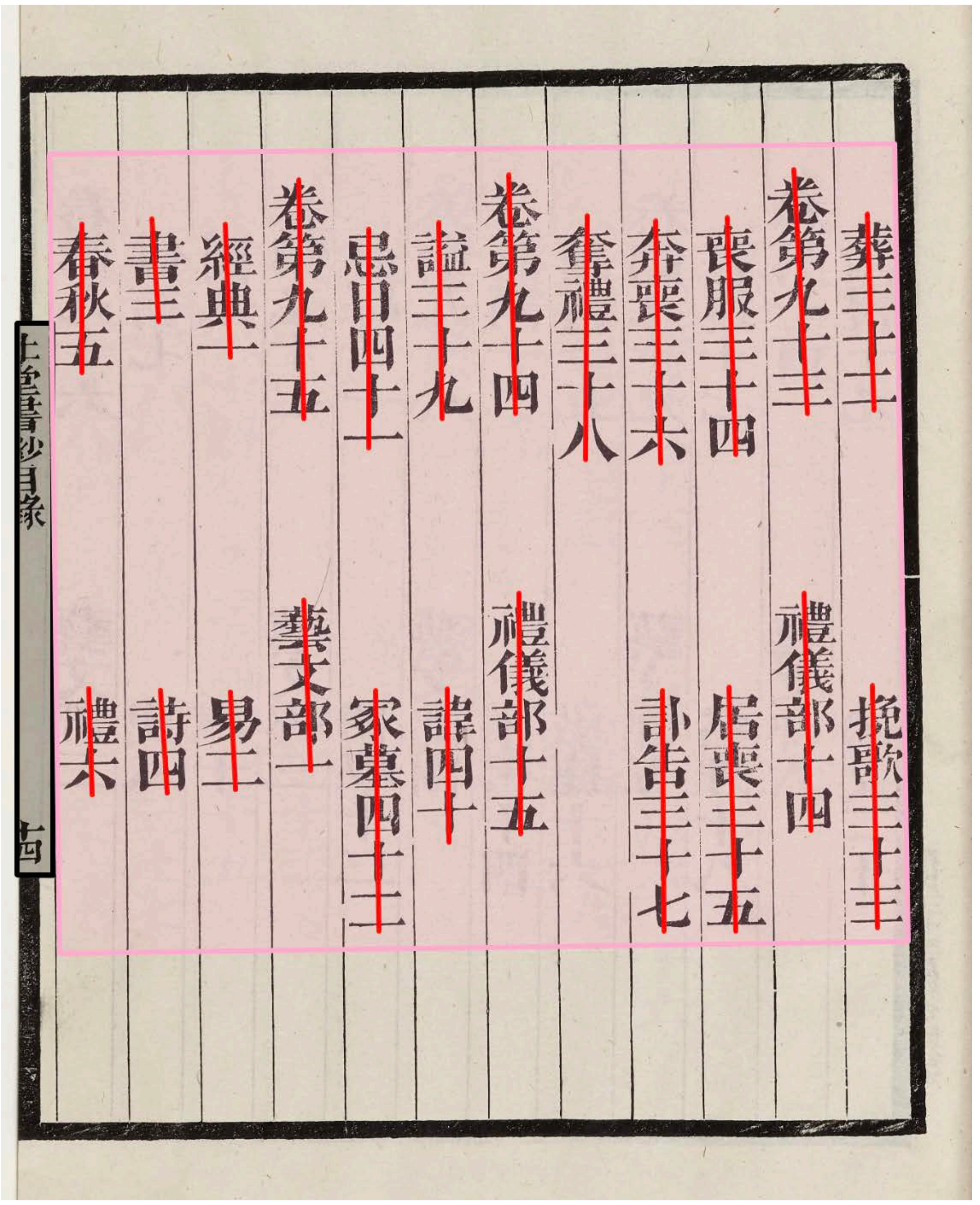
Numérisation du corpus

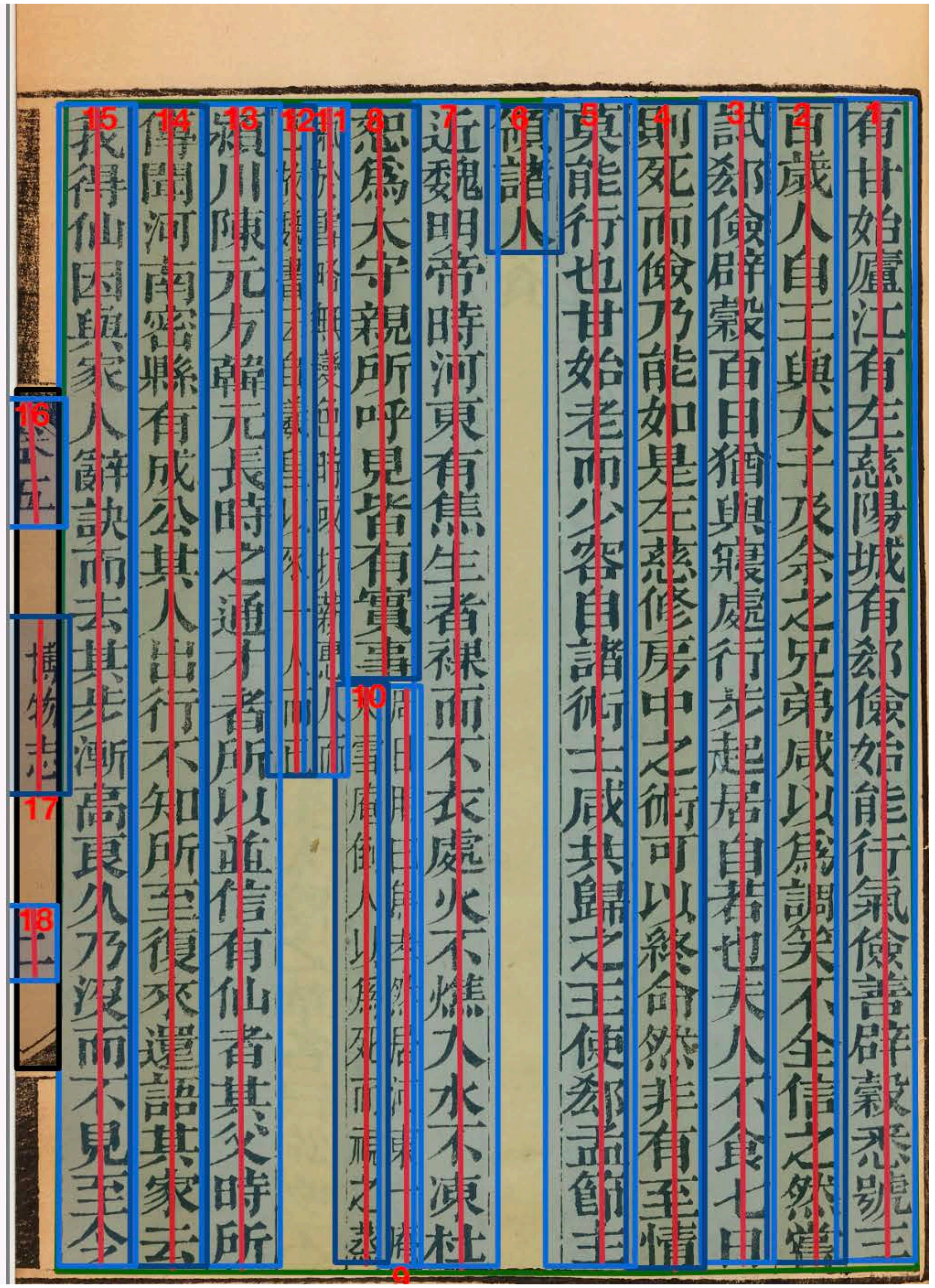


BIHEC-SB3701

- Environ 56 000 pages
- Janvier-Juin 2023
- Modalités différentes selon les institutions de conservation (simple/double page, format de destination)
- Nommage uniformisé
- Une version conservée et exposée par les bibliothèques
- Une version versée dans l'entrepôt Nakala (modalités et métadonnées variées)

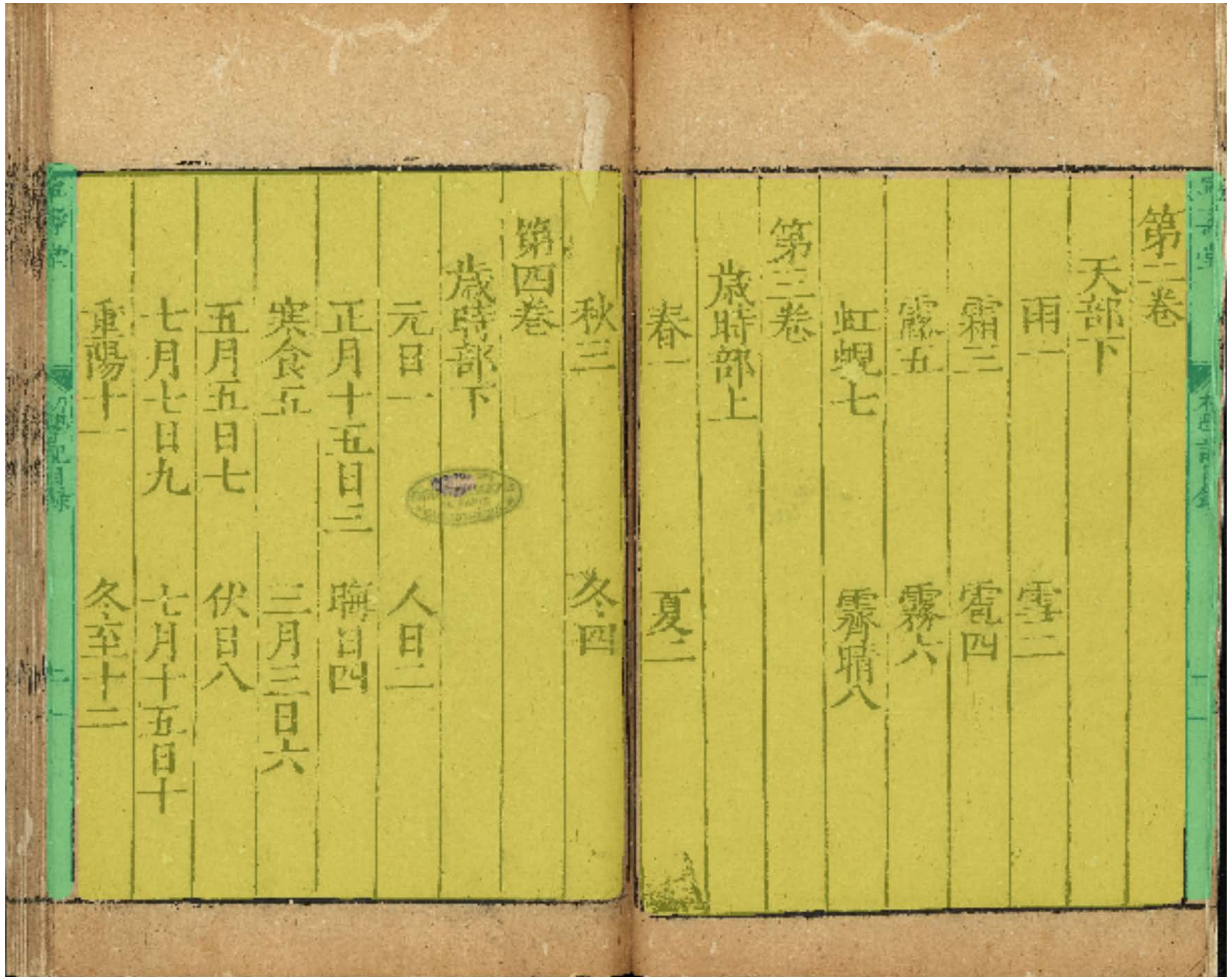
Annoter et transcrire



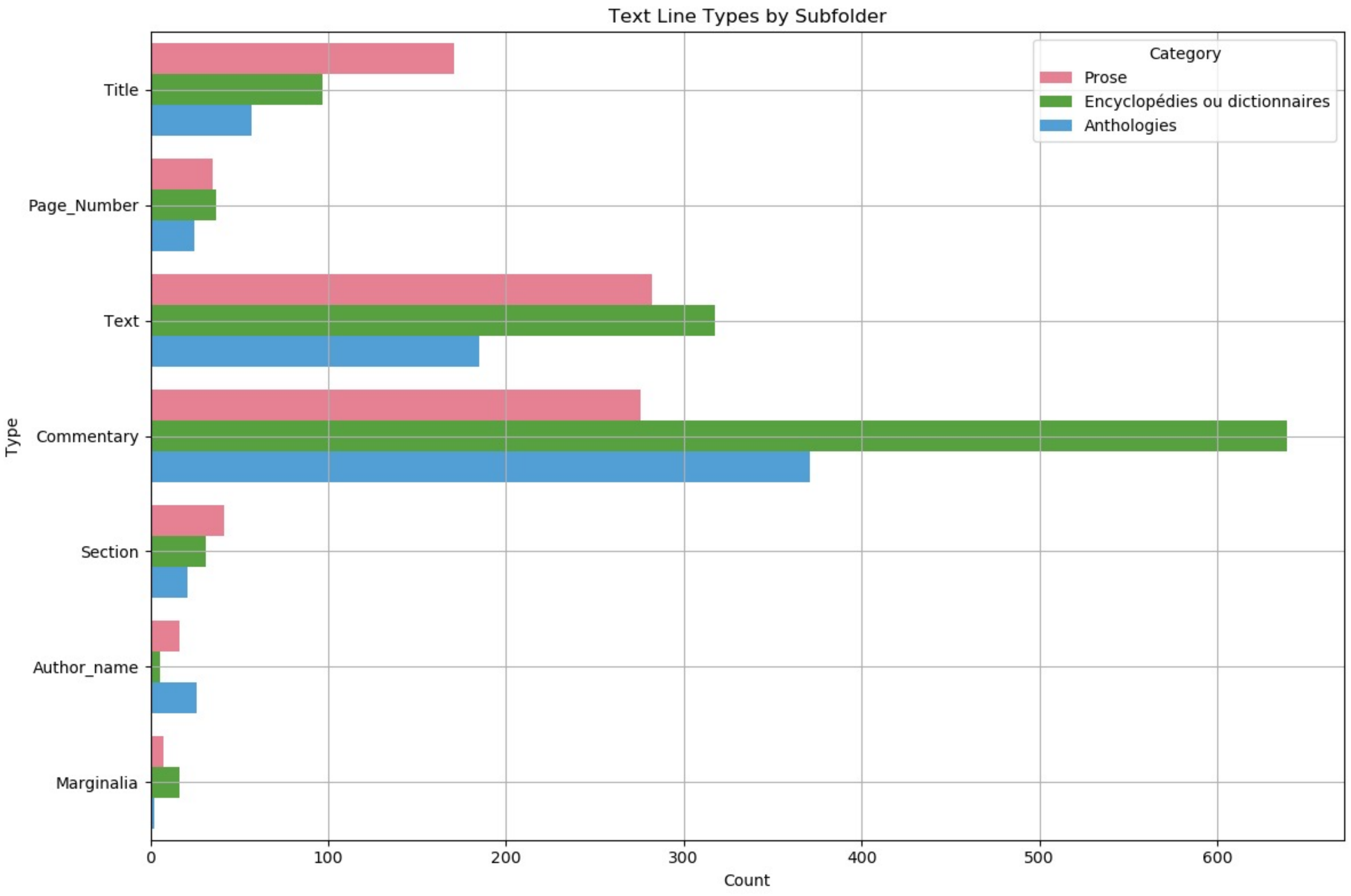


Annotateurs 2023 : Xinmin Hu, Elsa Cuillé, Marie Bizais-Lillig, Ani Tanelian, Anahide Kasparian, Chahan Vidal-Gorène

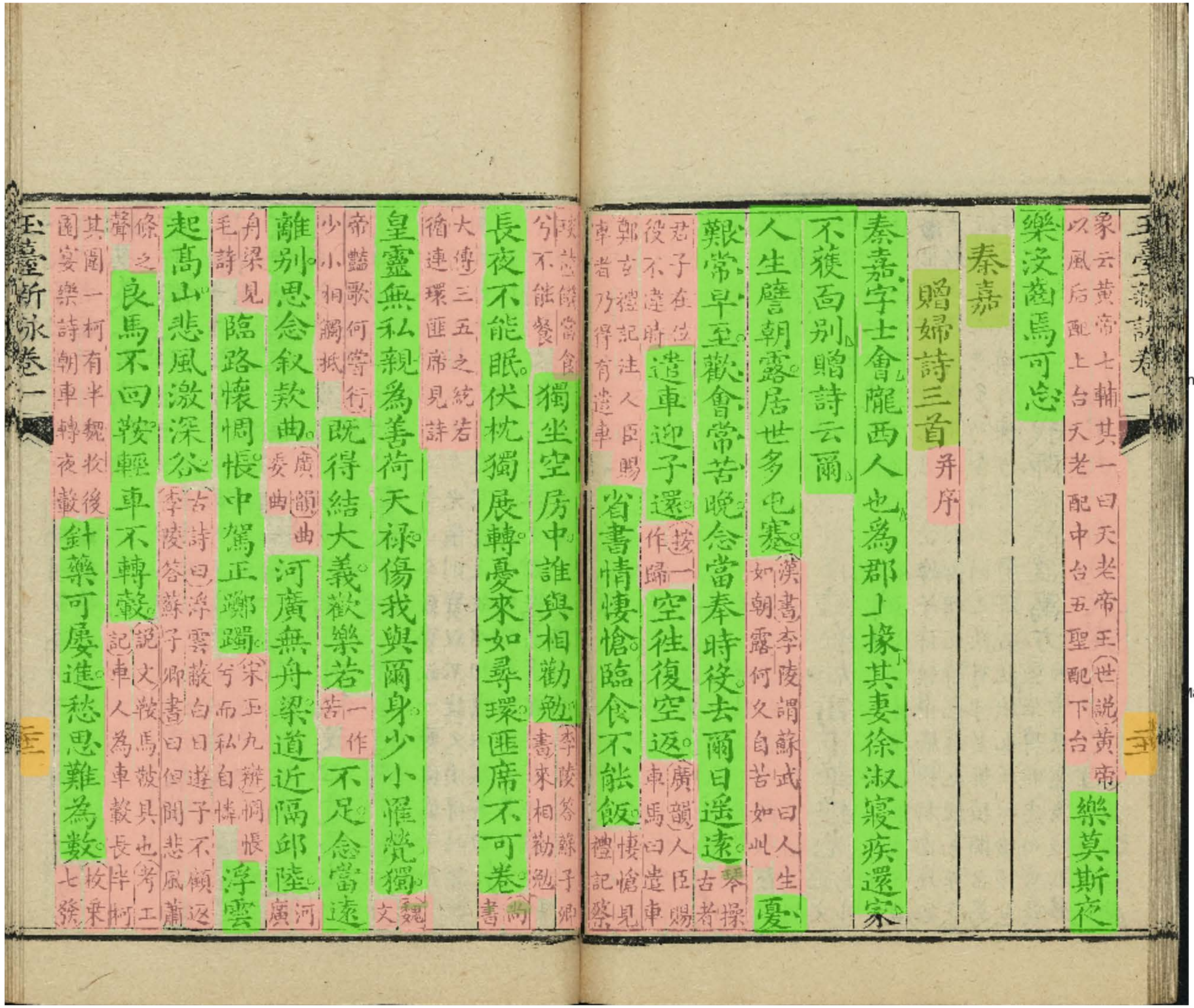
Typage des régions



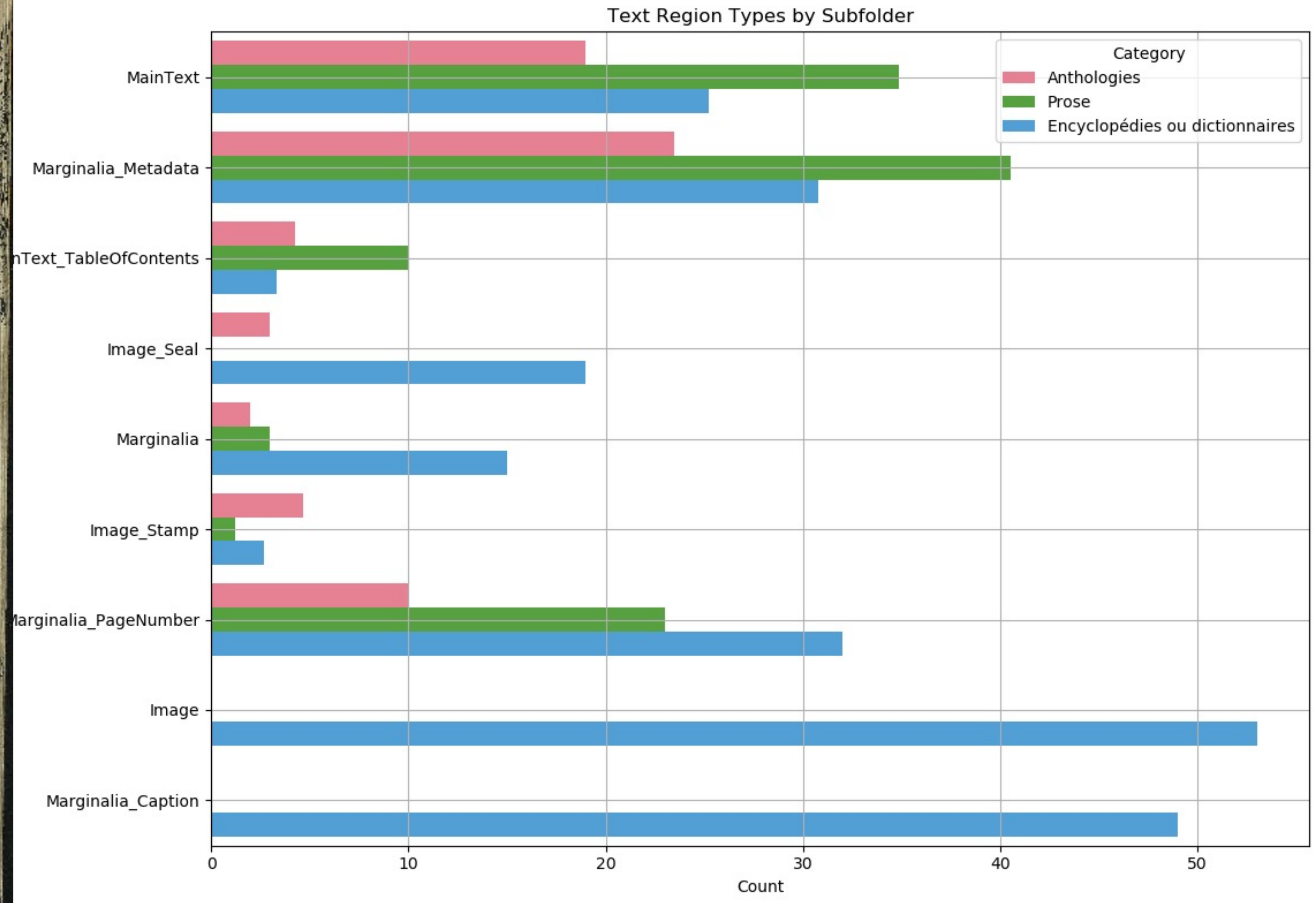
BIHEC-SB3701



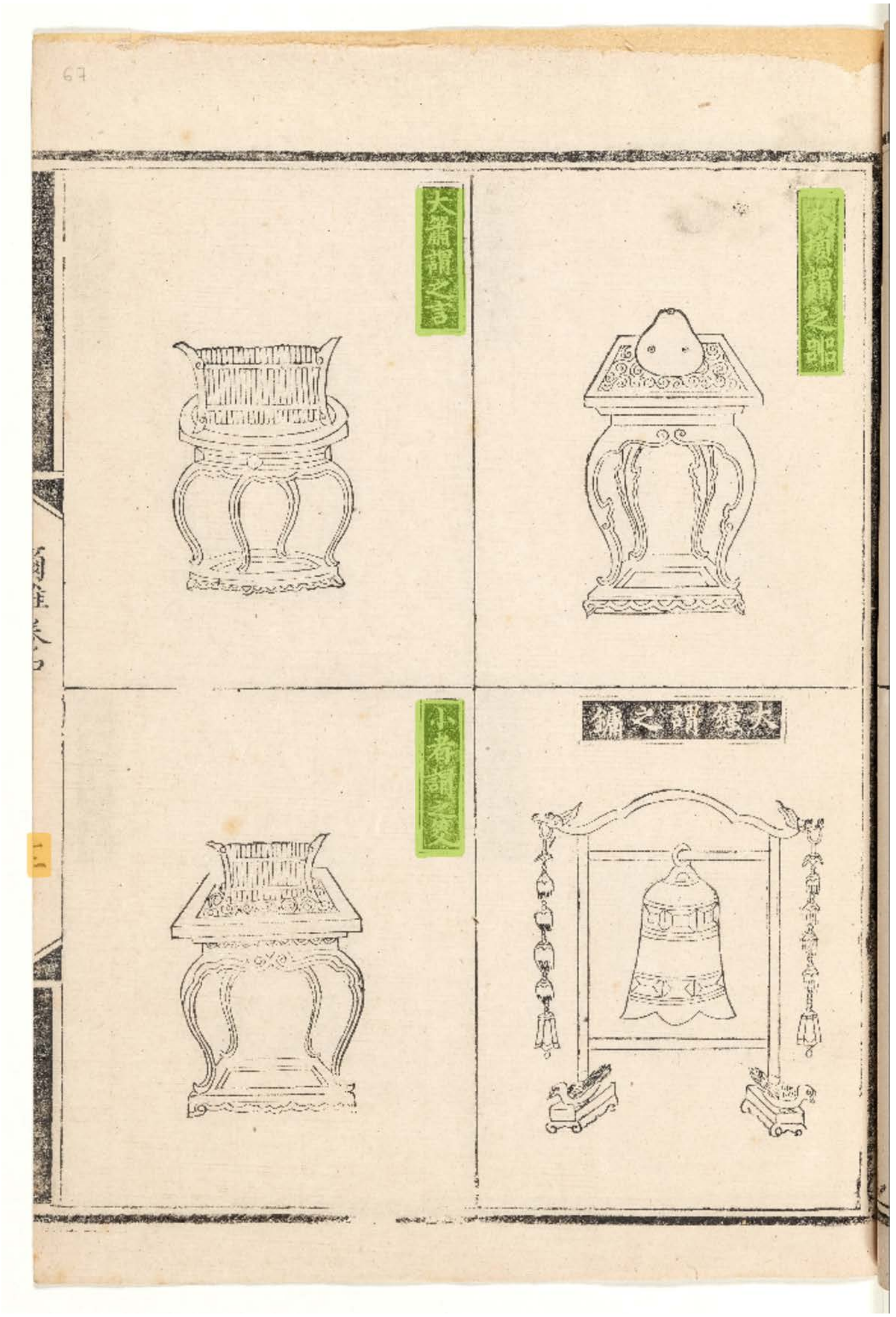
Typage des baselines



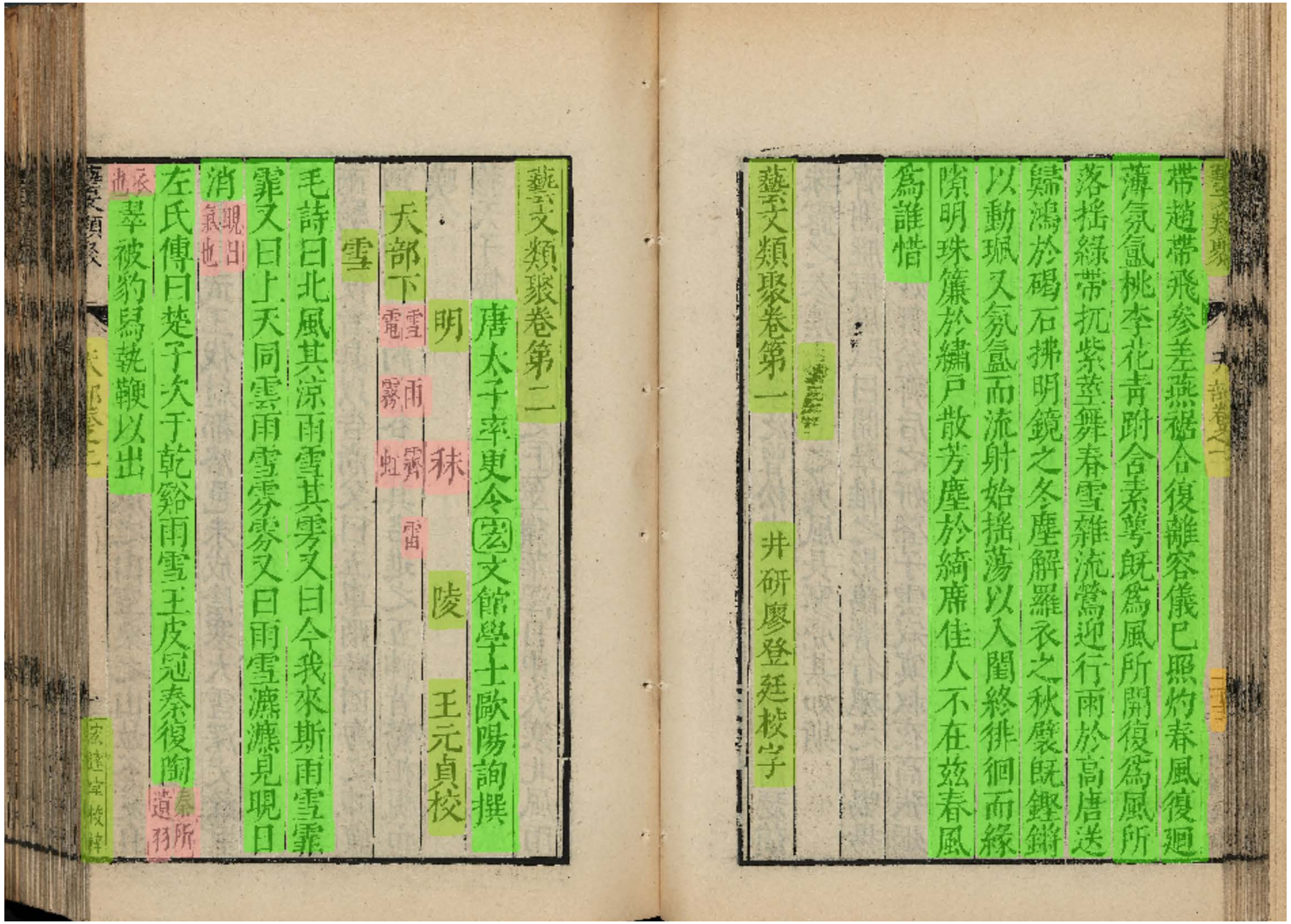
BIHEC-V-XIV-69



Typages



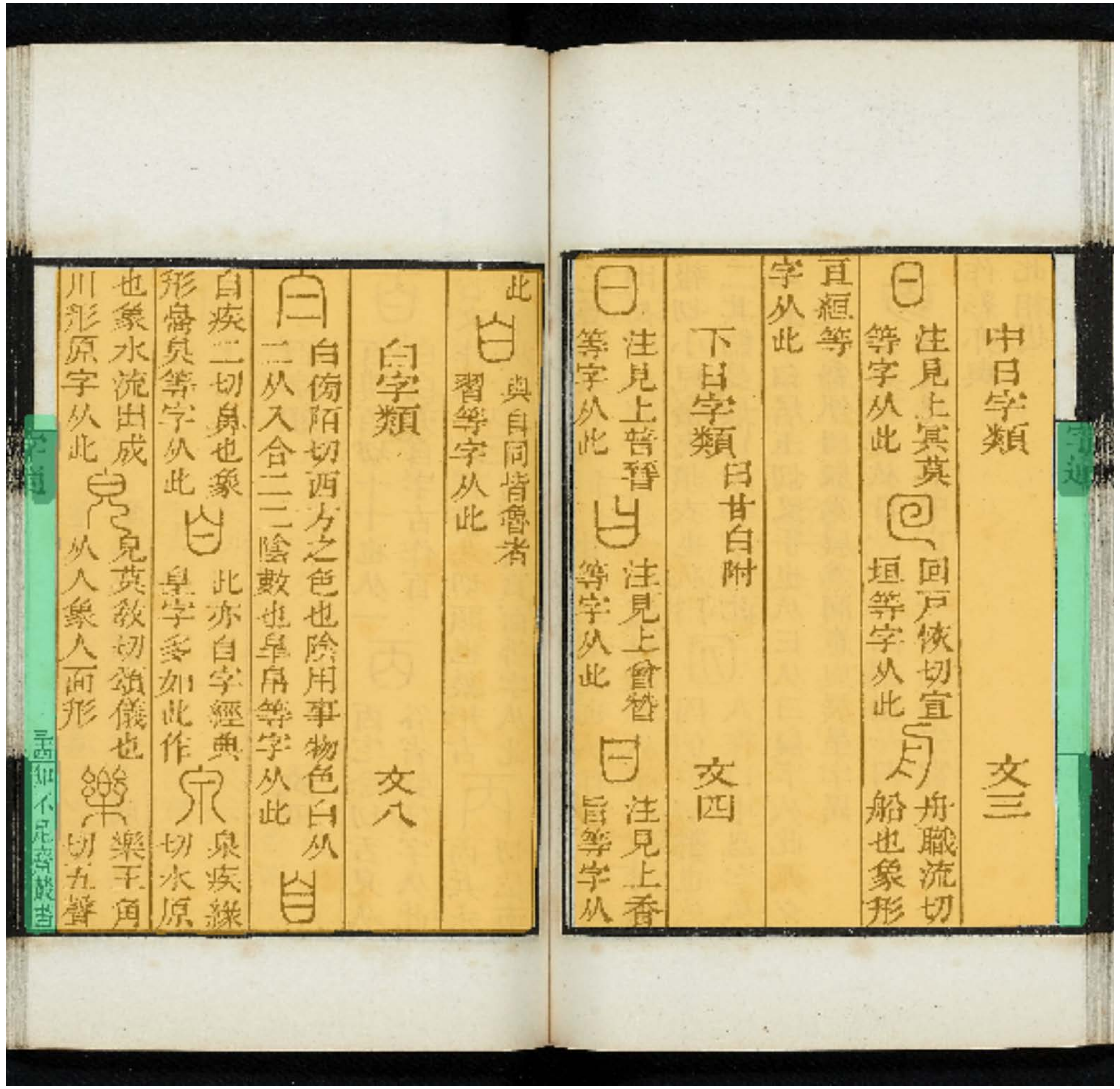
BULAC-BIULO-CHI-1938



BIHEC-C_III_5-7

Transcrire

- Les caractères tabous
- Les graphies historiques
- Les variantes



BIHEC-FX2-27-214

Transcrire

1421 caractères n’existent qu’une seule fois dans la dataset.
Soit 26,6% de classes uniques

- Images annotées : 306
- Régions annotées : 1 175
- Lignes annotées : 12 198
- Glyphes : 97 523
- Glyphes différents : 5 334

10 glyphes les plus fréquents

之	2 115	occurrences
日	1464	occurrences
也	1343	occurrences
不	885	occurrences
一	851	occurrences
以	841	occurrences
十	720	occurrences
有	720	occurrences
而	690	occurrences

10 glyphes les plus rares

佯
?
艕
窶
?
棰
獶
帕
伉
懾

Premiers résultats

Reconnaissance

Pour les régions

Precision = 0.955

Rappel = 0.96

mAP = 0.968 (= mean average precision)

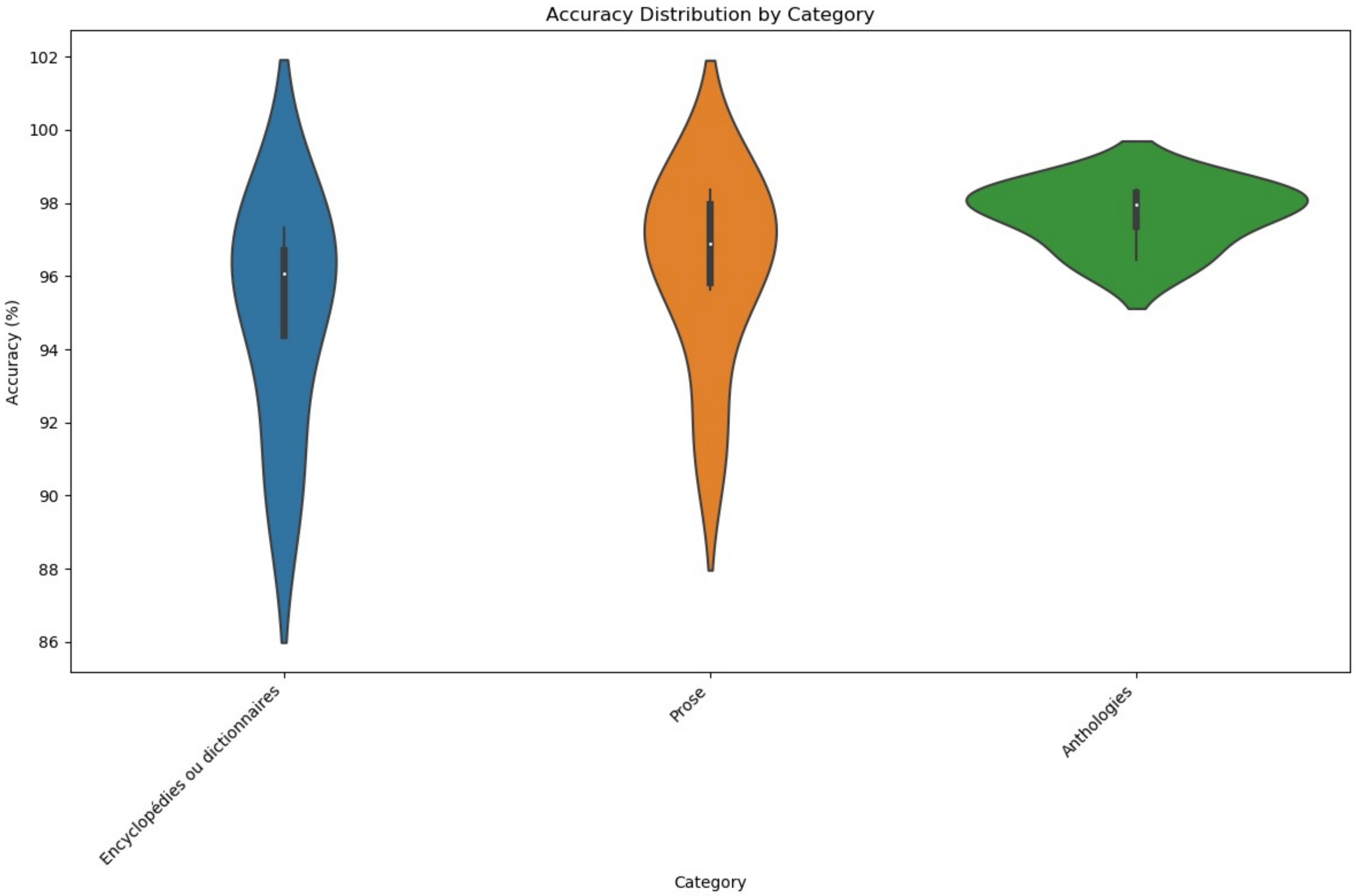
Pour les caractères

CER moyen 2%

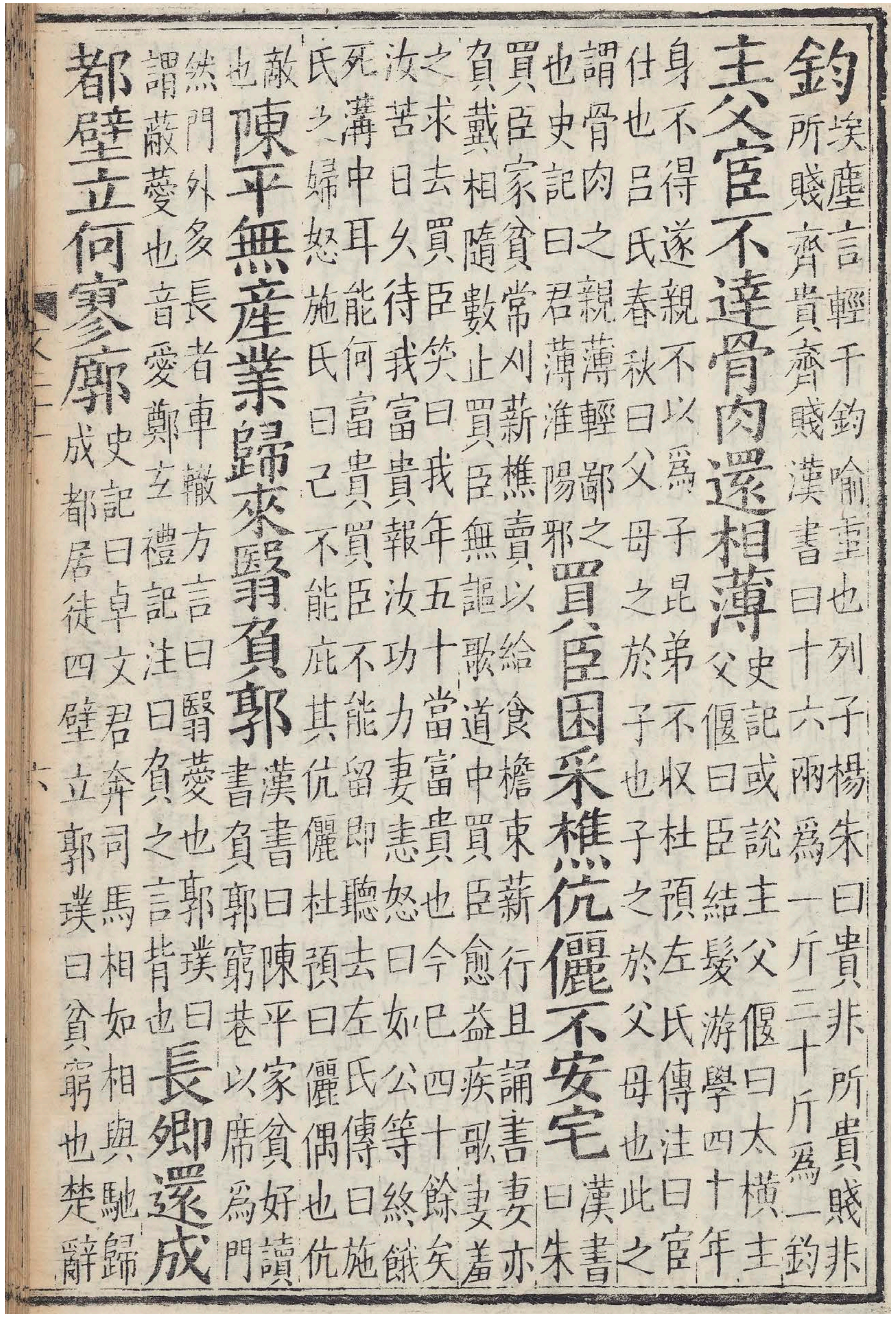
Entre 86% et 100% selon les lignes

Grande variété selon genres

Pour les caractères

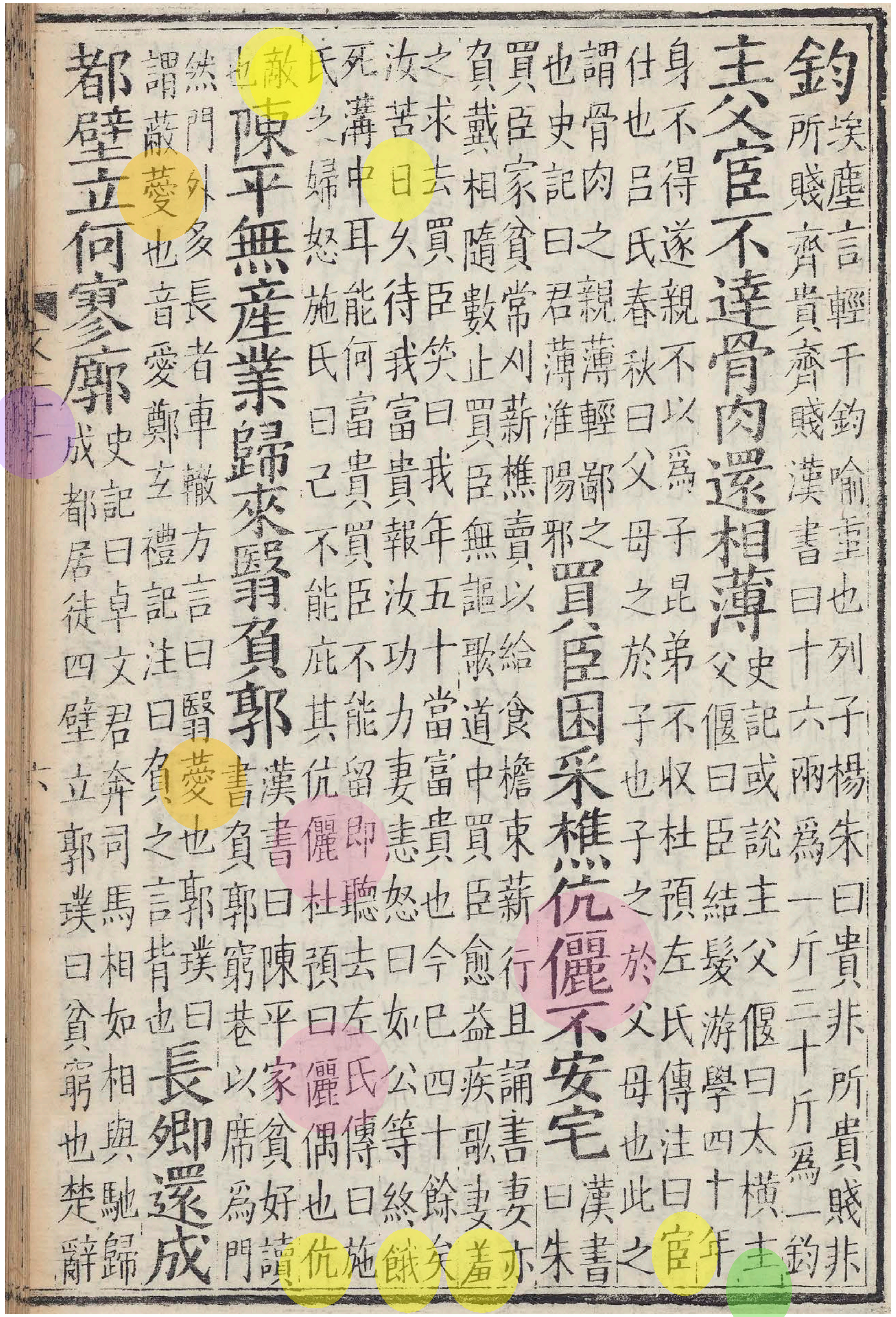


Illustration



鈞
埃塵言輕千鈞喻重也列子楊朱曰貴非所貴賤非
所賤齊貴齊賤漢書曰十六兩為一斤三十斤為一鈞
主父宦不達骨肉還相薄
史記或說主父偃曰太橫主一
父偃曰臣結髮游學四十年
身不得遂親不以為子昆弟不收杜預左氏傳注曰官(宦)
仕也呂氏春秋曰父母之於子也子之於父母也此之
謂骨肉之親薄輕鄙之
也史記曰君薄淮陽邪
買臣困采樵伉儀(儷)不安宅
漢書
曰朱
買臣家貧常刈薪樵賣以給食檐束薪行且誦書妻亦
負戴相隨數止買臣無謳歌道中買臣愈益疾歌妻養(羞)
之求去買臣笑曰我年五十當富貴也今已四十餘矣
汝苦曰(日)久待我富貴報汝功力妻恚怒曰如公等終餘(餓)
死溝中耳能何富貴買臣不能留即聽去左氏傳曰施
氏之婦怒施氏曰己不能庇其伉懼(儷)杜預曰曜(儷)偶也位(伉)
敢(敵)
也
陳平無產業歸來翳負郭
漢書曰陳平家貧好讀
書負郭窮巷以席為門
然門外多長者車轍方言曰翳憂(薆)也郭璞曰
謂蔽變(薆)也音愛鄭玄禮記注曰負之言背也
長卿還成
都壁立何寥廓
史記曰卓文君奔司馬相如相與馳歸
成都居徒四壁立郭璞曰貧窮也楚辭
文二二(十一)
六

Illustration



鈞
埃塵言輕千鈞喻重也列子楊朱曰貴非所貴賤非
所賤齊貴齊賤漢書曰十六兩為一斤三十斤為一鈞
主父宦不達骨肉還相薄
史記或說主父偃曰太横主一
父偃曰臣結髮游學四十年
身不得遂親不以為子昆弟不収杜預左氏傳注曰官(宦)
仕也呂氏春秋曰父母之於子也子之於父母也此之
謂骨肉之親薄輕鄙之
也史記曰君薄淮陽邪
買臣困采樵伉儀(儷)不安宅
漢書
曰朱
買臣家貧常刈薪樵賣以給食檐束薪行且誦書妻亦
負戴相隨數止買臣無謳歌道中買臣愈益疾歌妻養(羞)
之求去買臣笑曰我年五十當富貴也今已四十餘矣
汝苦曰(日)久待我富貴報汝功力妻恚怒曰如公等終餘(餓)
死溝中耳能何富貴買臣不能留即聽去左氏傳曰施
氏之婦怒施氏曰己不能庇其伉懼(儷)杜預曰曜(儷)偶也位(伉)
敢(敵)
也
陳平無產業歸來翳負郭
漢書曰陳平家貧好讀
書負郭窮巷以席為門
然門外多長者車轍方言曰翳憂(蔓)也郭璞曰
謂蔽變(蔓)也音愛鄭玄禮記注曰負之言背也
長卿還成
都壁立何寥廓
史記曰卓文君奔司馬相如相與馳歸
成都居徒四壁立郭璞曰貧窮也楚辭
文二二(十一)
六

En cours et à venir

Partager / Mettre à disposition

- Les modèles et données d'entraînement de l'HTR (Github de Calfa + Zenodo) avec signalement (HTR United)
- Les images (entrepôt Nakala)
- Les fichiers en XML-TEI (entrepôt Nakala)
- La base de données (bio-bibliographique et lexicale)
- Sous licence CC BY-NC 4.0

CollEx-Persée

USIAS

Merci.

UR 1340 Groupe d'études orientales, slaves et néo-helléniques GEO
Université de Strasbourg

UAR2999 distam

bizais@unistra.fr
https://gitlab.huma-num.fr/chi-know-po
https://github.com/calfa-co/chi-know-po

Les soutiens

Travail d’exploration fondé sur

- Une base de données :
  - lexique
  - base temporelle
  - titres
  - personnes
- Des outils d’exploration associés
- Contributeurs :
  - Tilman Schalmey
  - Shueh-Ying Liao
  - Marie Bizais-Lillig

- Des scripts (Python) d’exploration :
  - cooccurrences
  - reprises ou échos
- Des scripts associés de visualisation
- Contributeurs :
  - Mariana Zorkina
  - Tilman Schalmey
  - Xinmin Hu
  - (Marie Bizais-Lillig)

- Un corpus :
  - 4 schémas modèles XML-TEI
  - odd + RelaxNG
  - fichiers image + manifestes IIIF
- Contributeurs :
  - Ilaine Wang
  - Shueh-Ying Liao
  - Xinmin Hu
  - Marie Bizais-Lillig

Inviter à la correction des textes

TACT, plateforme de Transcription et d'Annotation de Corpus Textuels

TACT est une plateforme de transcription et d'annotation collaborative qui accueille des projets de recherche et permet à chacun d'apporter sa contribution ou de solliciter la communauté de la plateforme.

Vous voulez participer à un projet ? Vous pouvez transcrire et annoter des textes originaux, relire et vérifier des transcriptions faites par d'autres internautes, consultez ce manuel du contributeur et allez explorer les projets !

Vous souhaitez déposer un projet sur TACT ? N'hésitez pas à nous en faire la demande via ce formulaire. Vous deviendrez gestionnaire de votre projet TACT, vous pourrez en définir les modalités et gérer les contributions. Nous vous invitons à lire ce manuel du gestionnaire qui vous donnera toutes les informations nécessaires ainsi que la marche à suivre. Bonne lecture !

Vous pouvez aussi nous contacter à cette adresse : tact {arobase} univ-grenoble-alpes.fr

Dépôt/lien sur TACT

Correction des erreurs de l’HTR

Correction de la structuration automatisée

Correction/enrichissement de la détection automatisée des entités nommées