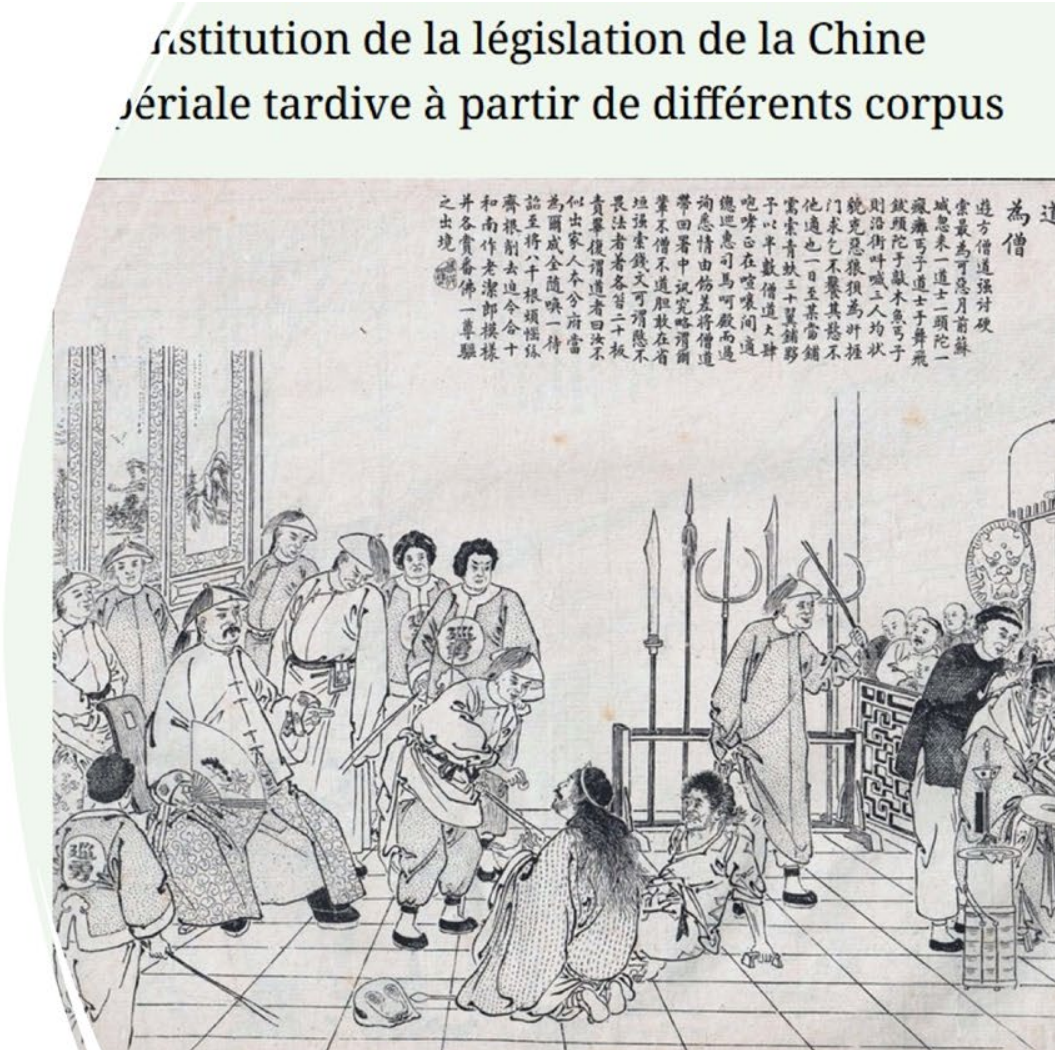
Le projet CollEx COREL (Code relationnel) : numérisation, OCR et bases de données au service de l'analyse de sources juridiques chinoises

Colin BRISSON (EPHE/CRCAO**)**
Frédéric CONSTANT (Université Côte d'Azur)

Objectifs du projet

- Création d'un site Internet de dépôt de d'affichage et sources juridiques chinoises

- Reconstitution de l'ensemble de la legislation chinoise à partir de plusieurs corpus



- legal handbooks :
  for magistrate, for legal advisers, for pettifoggers.

Resources on other Asian legal traditions

- Japan
- Korea
- Mongolia
- Vietnam

Maps

- Exile maps
- Lingchi sentences
- Regional special law

Ming and Qing codes with translations

- Methodology
- Glossary

Ming Code

- overview
- Da Ming lü jijie fuli 1610 /all

Qing Code

- overview
- Da Qing lü jijie fuli 1646 /all
- Xingbu xianxing zeli 1680 /all
- Da Qing lüli 1740 /all
- Da Qing lüli 1871 /all
- Huidian shili-Xingbu part 1899 /all
- Da Qing lüli-Duli cunyi 1905 /all

Virtual Synoptical Code

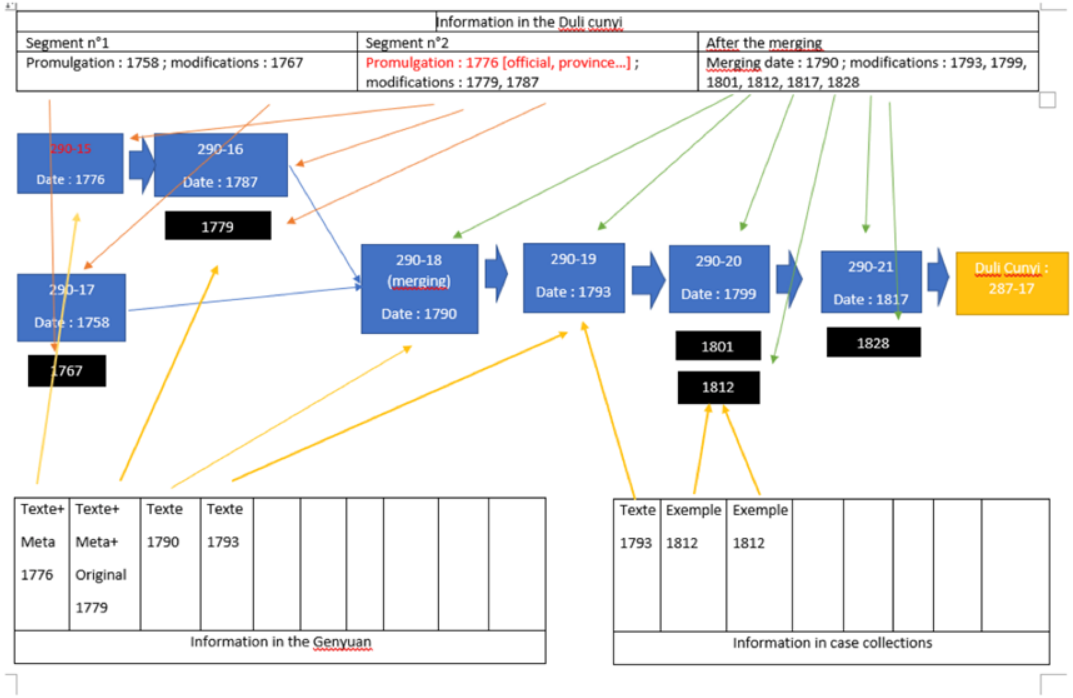
ID	Title	Year	Author	Topic
389	(Chongke) Xiyuan lu 重刻洗冤錄 [A New Carving of The Washing Away of Wrongs]		Song Ci 宋慈	4.1 Magistrates handbooks: General
335	(Jingxin haimu)Huaji miaopan(驚心駭目)滑稽妙判		Fu Waishi 芙外史(ed.)	4.1 Magistrates handbooks: General
334	(Jingxin haimu)Laoli pipan daquan (驚心駭目)老吏批判大全		Fan Zengxiang 樊增祥 etc.	4.2 Magistrates handbooks: Handbooks for legal experts
394	(Lüli guan jiaozheng) Xiyuan lu 律例館校正洗冤錄 [Washing Away the Wrongs, Edited by the Bureau of the Code]		Anonymous	4.2 Magistrates handbooks: Handbooks for legal experts
395	(Lüli guan jiaozheng) Xiyuan lu 律例館校正洗冤錄 [Washing Away the Wrongs, Edited by the Bureau of the Code]		Anonymous	4.2 Magistrates handbooks: Handbooks for legal experts
491	(Qinban) Zhouxian shiyi 欽頒州縣事宜 [Advice for Magistrates Published by Imperial Order]		Tian Wenjing 田文鏡	3.2 Regulations collections: local regulations
523	(Wen Jinghan xiansheng) Zili yan (文靜涵先生) 自歷言 [Words from One's Own Progress]		Wenhai 文海 (Jinghan 靜涵)	4.1 Magistrates handbooks: General
470	(Wusuo Liu xiansheng) Juguan shuijing 勿所劉先生居官水鏡 [Mr. Liu Wusuo's Lucid Mirror of The Office Holder]		Liu Shijun 劉時俊	4.1 Magistrates handbooks: General
361	(Xinjuan) Fajia tou danhan 新鐫法家透膽寒 [Lawyer's Piercing Gallbladder Awe, A new edition]		Buxiang zi 補相子	4.3 Magistrates handbooks: Handbooks for Pettifoggers (songshi 訟師)
463	(Xue Wenqing gong) Congzheng mingyan 薛文清公從政名言 [Xue Xuan's Famous Sayings on Government Service]		Xue Xuan 薛瑄	4.1 Magistrates handbooks: General
464	(Yuzhi) Guanzhen 御製官箴 [Imperially Composed Admonitions to Officials]		emperor Xuanzong of the Ming (Zhu Zhanji 朱瞻基)	4.1 Magistrates handbooks: General
156	(Zengding) Xingbu shuotie 增訂刑部說帖 [Memoranda from the Ministry of Justice]		comp. Guoying 國英, from Jilin, et al.	2.1 Judicial cases: general casebooks
222	A chinese-english dictionary	1912	Herbert A. Giles	
86	An Wu qinshen xigao 按吳親審檄稿 [Draft Opinions from Cases Personally Tried as Regional Inspector of the Wu Region]	0	Qi Biaojia 祁彪佳	2.2 Judicial cases: Local casebooks

Présentation des sources

- Deux sources principales :
  - Les lois principals (*lü* 律)
  - Les lois secondaires (*tiaoli* 條例)

- Code des Qing *Da Qing lüli* 大清律例 (1740)
- Commentaire du code
  - Doutes persistants sur le code *Du li cunyi* 讀例存疑 (1905)
- Deux compilations
  - Les origines du code des Qing *Da Qing lüli genyuan* 大清律例根源 (1871)
  - Les institutions réunies des Grands Qing *Da Qing huidian* 大清會典 (1899)

Généalogie d'un article



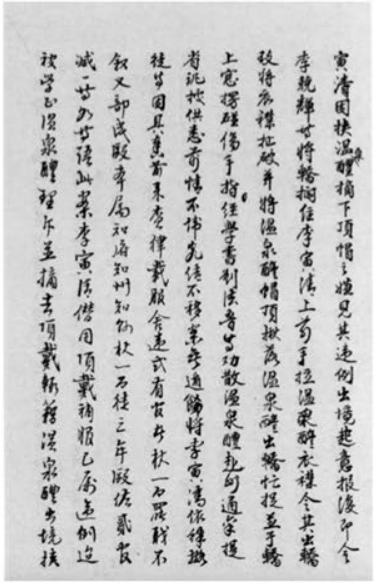
Freizo/Data Futures

num: 282 盜田野穀麥 number: 280, (related code: 271), (related duli-cunyi 271), (related huidian 282),
p. 0: num: number: 280-1,
p. 1: num: 282-1-1 number: 280-2, date 1646, (related huidian 282-1-1),
p. 2: num: number: 280-3,
p. 3: num: number: 280-4,
p. 4: num: 282-9-7 number: 280-5, date 1725, (related huidian 282-9-7),
p. 5: num: number: 280-6,
p. 5: num: number: 280-7,
p. 6: num: number: 280-8,
p. 7: num: number: 280-9,
p. 7: num: number: 280-10,
p. 8: num: number: 280-11,
p. 9: num: 282-4 number: 280-12, date 1731, (related duli-cunyi 271-2), (related huidian 282-4),
p. 10: num: number: 280-13,
p. 11: num: 282-9-8 number: 280-14, date 1725, (related huidian 282-9-8),
p. 13: num: 282-9-9 number: 280-15, date 1740, (related duli-cunyi 271-10), (related huidian 282-9-9),
p. 14: num: 282-9-10 number: 280-16, date 1725, (related huidian 282-9-10),
p. 15: num: 282-9-11 number: 280-17, date 1732;1740, (related huidian 282-9-11),
p. 15: num: 282-3-1 number: 280-18, date 1725, (related huidian 282-3-1),
p. 16: num: 282-3-2 number: 280-19, date 1740, (related duli-cunyi 271-3), (related huidian 282-3-2),
p. 17: num: number: 280-20,
p. 17: num: 282-5 number: 280-21, date 1745, (related duli-cunyi 271-9), (related huidian 282-5),
p. 17: num: number: 280-22,
p. 18: num: number: 280-23,
p. 18: num: 282-6 number: 280-24, date 1747, (related duli-cunyi 271-4), (related huidian 282-6),
p. 18: num: 282-8-1 number: 280-25, date 1745, (related huidian 282-8-1),
p. 19: num: 282-7 number: 280-26, date 1747, (related duli-cunyi 271-5), (related huidian 282-7),
p. 20: num: 282-9-1 number: 280-27, date 1756, (related huidian 282-9-1),
p. 21: num: number: 280-28,
p. 24: num: 282-9-3 number: 280-29, date 1759, (related huidian 282-9-3),
p. 25: num: number: 280-30,
p. 25: num: number: 280-31,
p. 26: num: 282-9-12 number: 280-32, date 1738;1767, (related huidian 282-9-12),

Enrichir les relations. Les codes et les jugements.

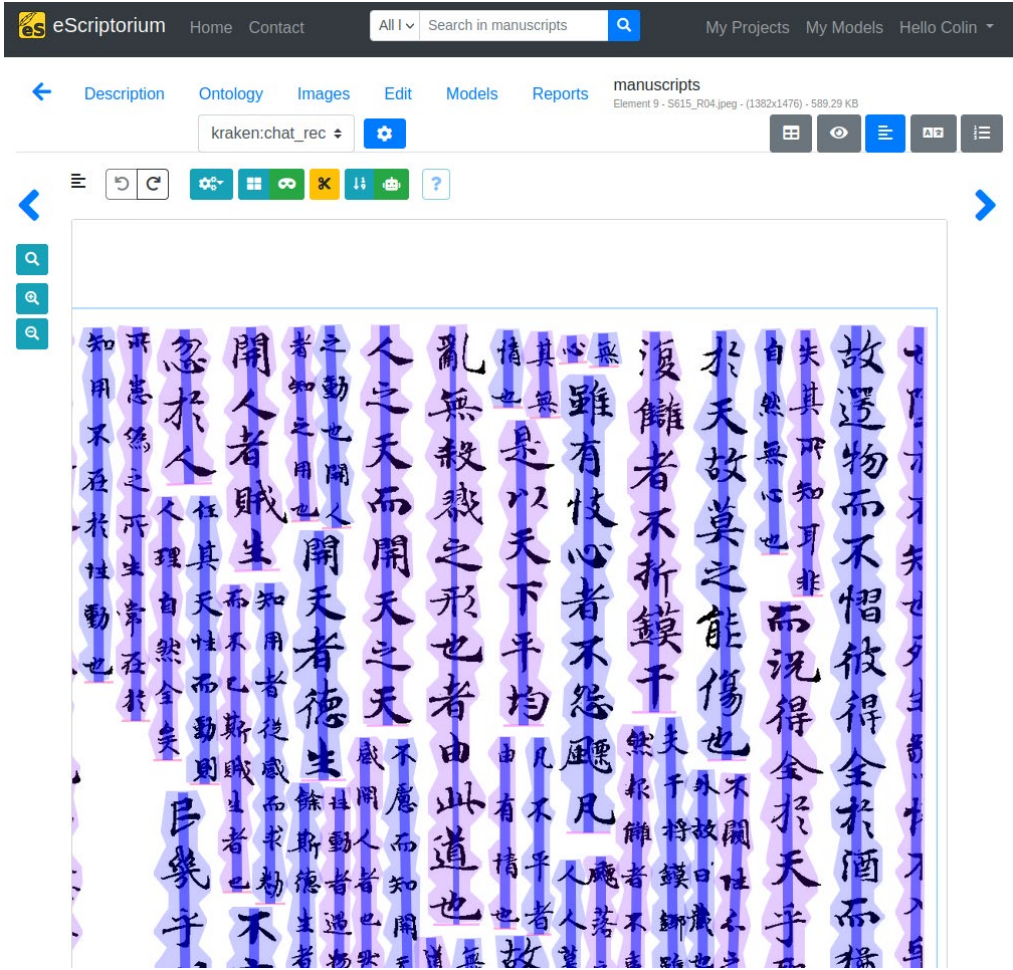
Relations bidirectionnelles

- Les lois additionnelles peuvent avoir leur origine dans des jugements
- Les jugements appliquent des lois



Transcription automatique

1.1 Vérité de terrain pour la segmentation



E

escriptorium

Project ID: 26940117

2,483 Commits **28** Branches **134** Tags **891.4 MiB** Project Storage

A project providing digital recognition of handwritten documents using machine learning techniques.

https://gitlab.com/scripta/escriptorium

1.2 Vérité de terrain pour la reconnaissance

https://github.com/kanripo

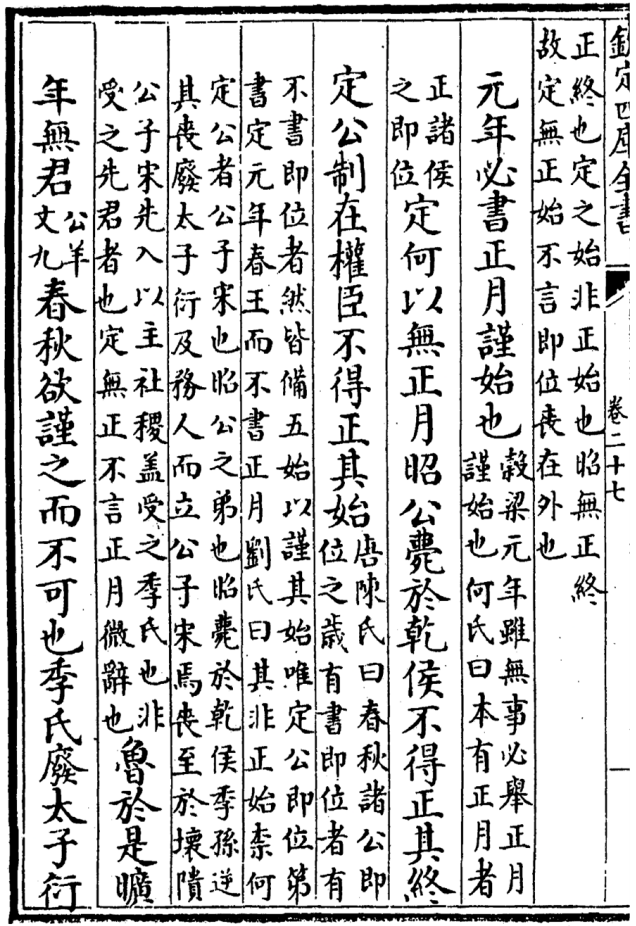
1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées

https://github.com/kanripo

1.2 Vérité de terrain pour la reconnaissance

3 millions de lines alignées


欽定四庫全書 卷二十七

正終也定之始非正始也昭無正終故定無正始不言即位喪在外也

元年必書正月謹始也穀梁元年雖無事必舉正月謹始也何氏曰本有正月者正諸侯之即位定何以無正月昭公薨於乾侯不得正其終定公制在權臣不得正其始唐陳氏曰春秋諸公即位之歲有書即位者有不書即位者然皆備五始以謹其始唯定公即位第書定元年春王而不書正月劉氏曰其非正始奈何定公者公子宋也昭公之弟也昭薨於乾侯季孫逆其喪廢太子衍及務人而立公子宋焉喪至於壞隤公子宋先入以主社稷蓋受之季氏也非受之先君者也定無正不言正月微辭也魯於是曠

年無君公羊文九春秋欲謹之而不可也季氏廢太子衍

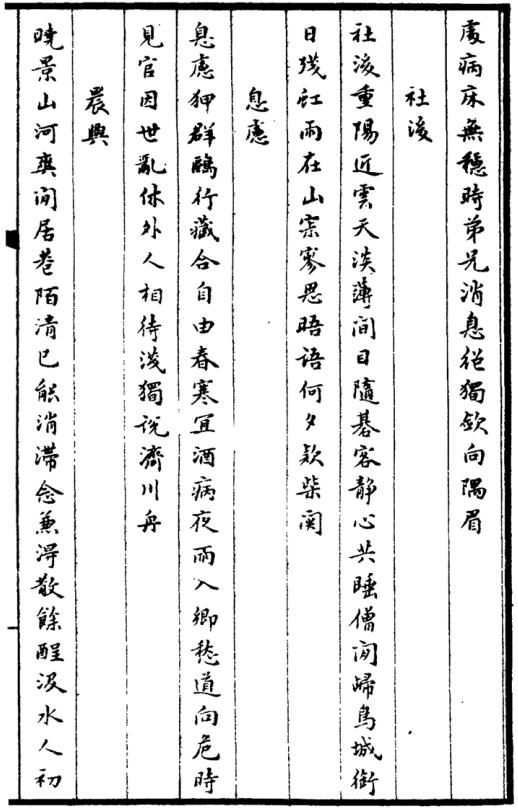
1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées


文理支脉其来龍者地之根源所自本也又取其
勢如龍之来蜿蜒活健也勢之大者厚德載物次
則廣濶坦平委蛇坡陀巉岐雀鬼之狀也支者勢
之分也又外則路之所通内則脉之所貫也脉者
真陽生意流行之迹也穴者地氣山勢来龍支脉
真陽生意之妙畢聚于此凝結不滯活動不流之
窠也此穴之能福于人者真陽生意凝結不滯活
動不流之澤也穴者竅眼也針穴灸穴非竅眼曰
筋曰骨曰肉不曰穴葵穴非竅眼曰土曰石曰泉

1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées



1.2 Vérité de terrain pour la reconnaissance

- 內 (U+5185) → 內 (U+5167)

https://github.com/kanripo

1.2 Vérité de terrain pour la reconnaissance

- 内 (U+5185) → 內 (U+5167)
- 黃 (U+9EC4) → 黃 (U+9EC3)

https://github.com/kanripo

1.2 Vérité de terrain pour la reconnaissance

- 內 (U+5185) → 內 (U+5167)
- 黃 (U+9EC4) → 黃 (U+9EC3)
- 德 (U+5FB7) → 德 (U+5FB3)

https://github.com/kanripo

1.2 Vérité de terrain pour la reconnaissance

- 內 (U+5185) → 內 (U+5167)
- 黃 (U+9EC4) → 黃 (U+9EC3)
- 德 (U+5FB7) → 德 (U+5FB3)

…

Plus de 4 000 glyphes normalisés

https://github.com/kanripo

1.2 Vérité de terrain pour la reconnaissance

Chinese Historical documents Automatic Transcription (CHAT) models

https://github.com/colibrisson/CHAT_models

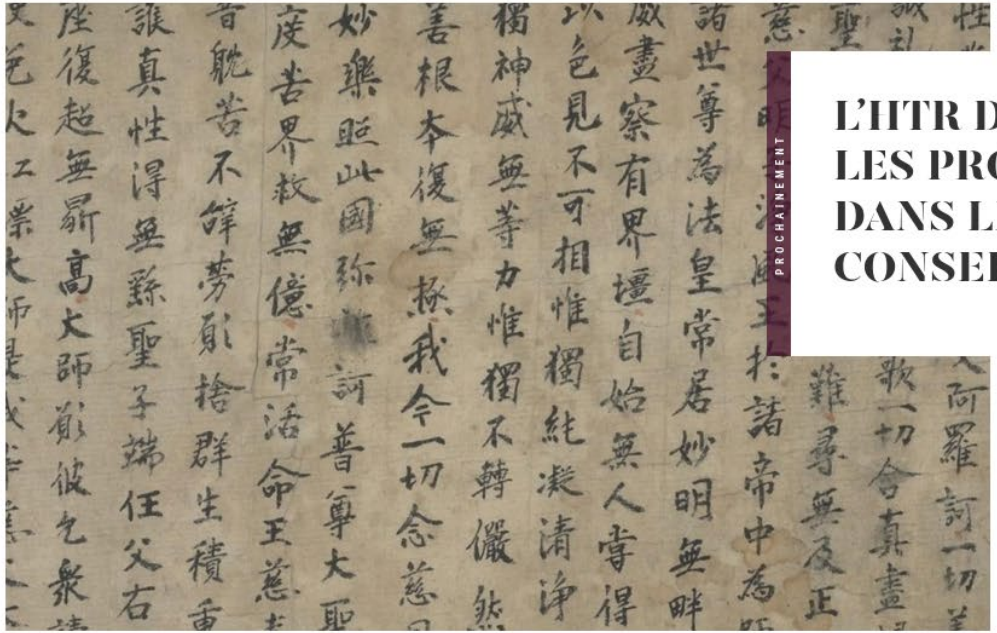
1.2 Vérité de terrain pour la reconnaissance

```
Loading model /home/colibri/datasets/sbck/hot/aligned/models/
Evaluating /home/colibri/datasets/sbck/hot/aligned/models/tes
Evaluating ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 9993
=== report  ===

1699210 Characters
14442   Errors
99.15%  Accuracy

207     Insertions
189     Deletions
14046   Substitutions

Count   Missed  %Right
1697864 13600   99.20%  Han
433     121     72.06%  Common
912     531     41.78%  Unknown
1       1       0.00%   Ethiopic

Errors  Correct-Generated
520     { 已 } - { 巳 }
298     { 巳 } - { 已 }
117     { 於 } - { 扵 }
115     { 己 } - { 已 }
97      { 毋 } - { 母 }
82      { 掌 } - { 蒙 }
71      { 髮 } - { 髮 }
63      { 日 } - { 曰 }
51      { 扵 } - { 於 }
48      { 世 } - { 丗 }
47      { 傅 } - { 傳 }
46      { 丗 } - { 世 }
42      { 己 } - { 巳 }
40      { 博 } - { 博 }
40      { 曰 } - { 日 }
39      { 補 } - { 補 }
39      { 戍 } - { 戌 }
```

Chinese Historical documents Automatic Transcription (CHAT) models

https://github.com/colibrisson/CHAT_models

1.2 Vérité de terrain pour la reconnaissance



Pelliot chinois 3847 - - BnF, département des Manuscrits

PROCHAINEMENT

L'HTR DES LANGUES PEU DOTÉES DANS LES PROGRAMMES DE RECHERCHE ET DANS LES ÉTABLISSEMENTS DE CONSERVATION FRANÇAIS

COLLOQUES

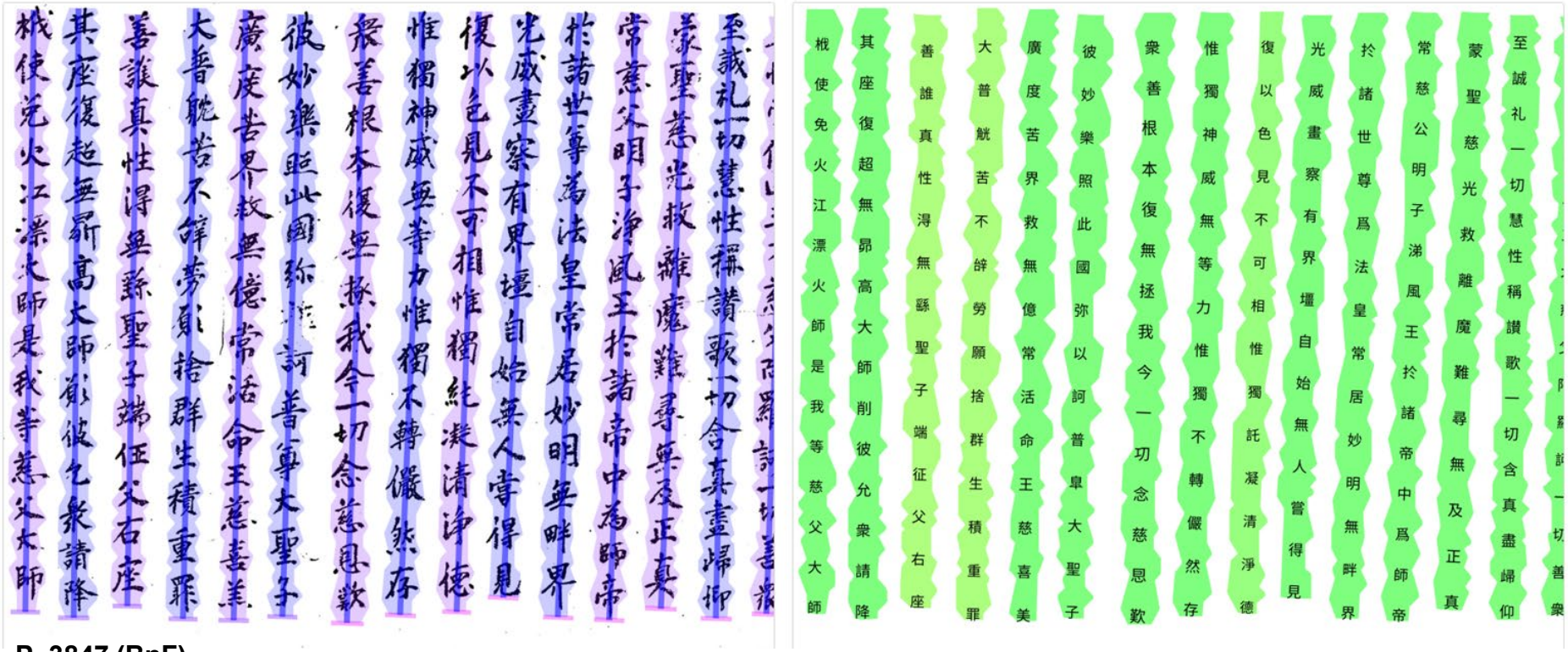
14 fév. 2024

9 h - 17 h

Richelieu

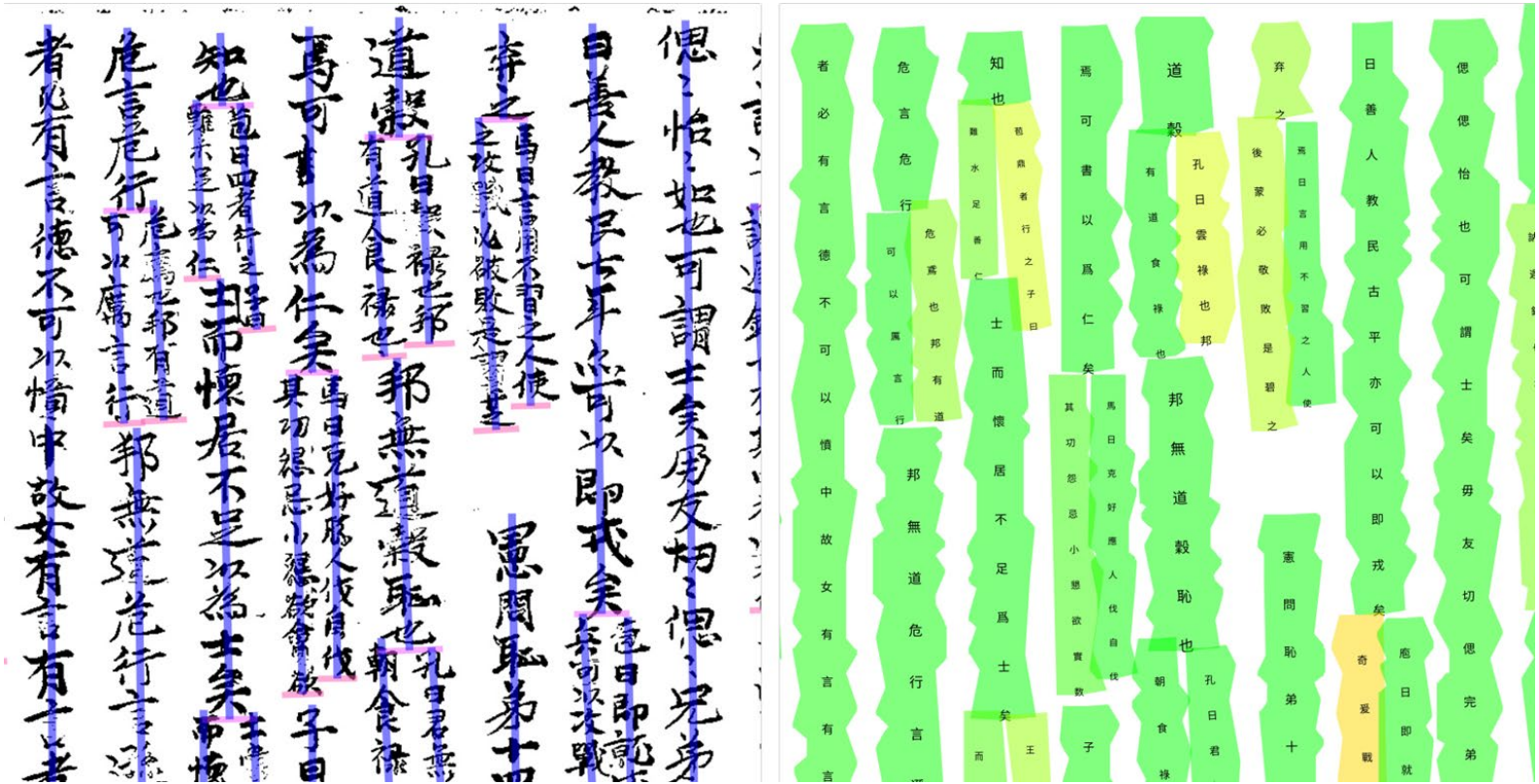
Salle des conférences

1.2 Vérité de terrain pour la reconnaissance

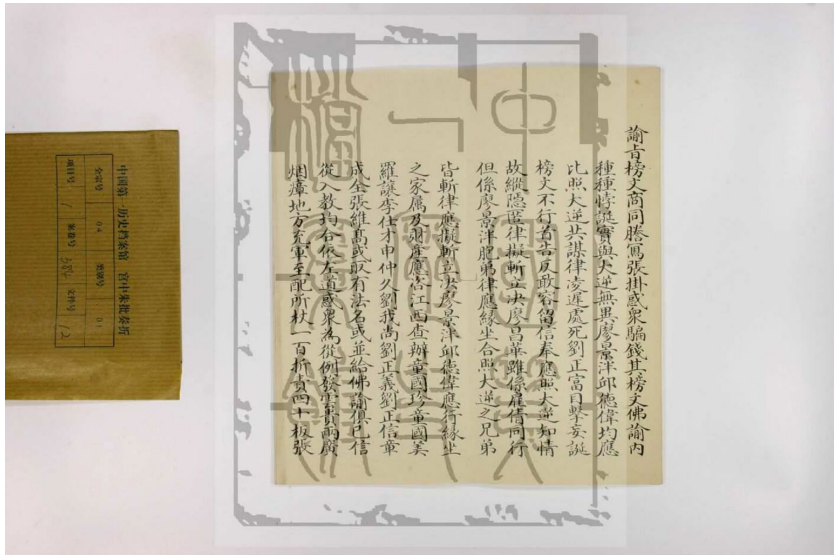


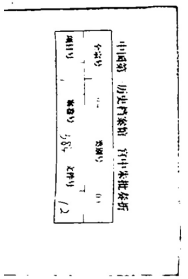
P. 3847 (BnF)

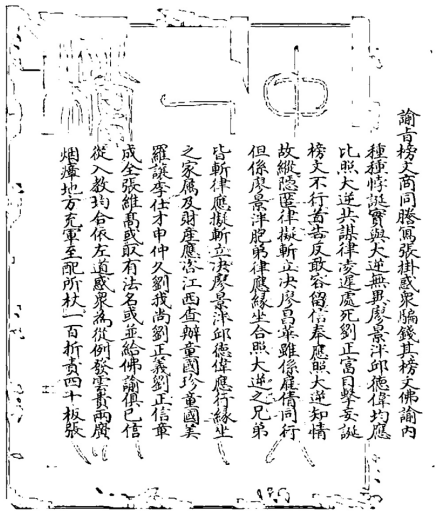
1.2 Vérité de terrain pour la reconnaissance



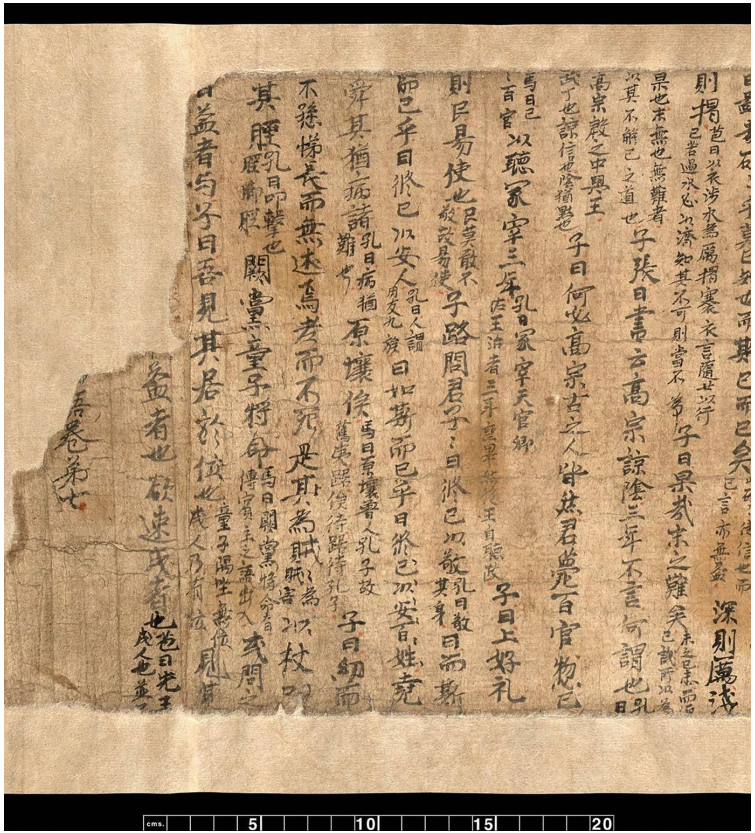
1.3 Binarisation

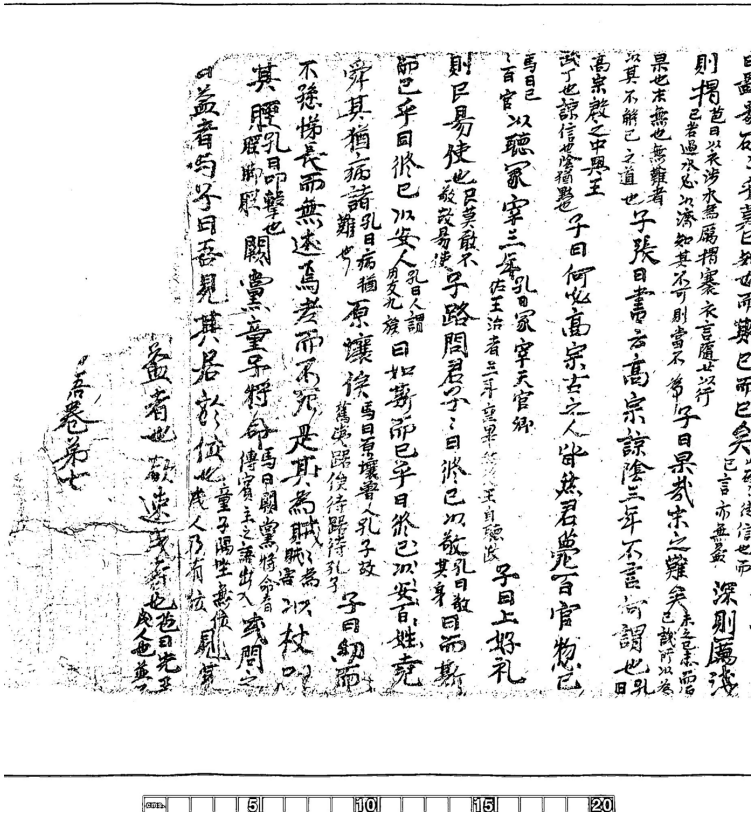






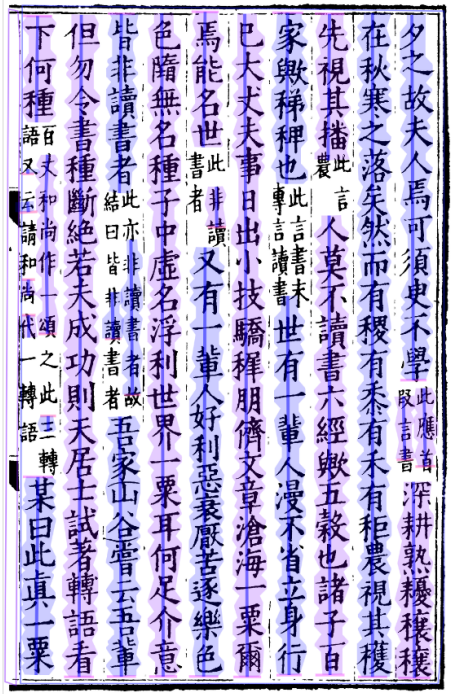
1.3 Binarisation

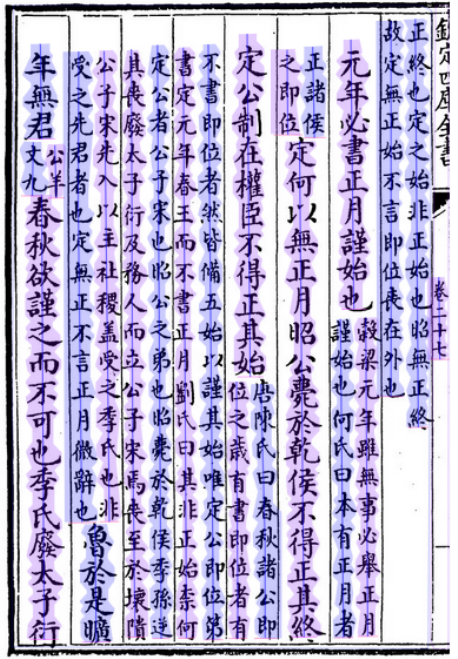




1.4 Limitations

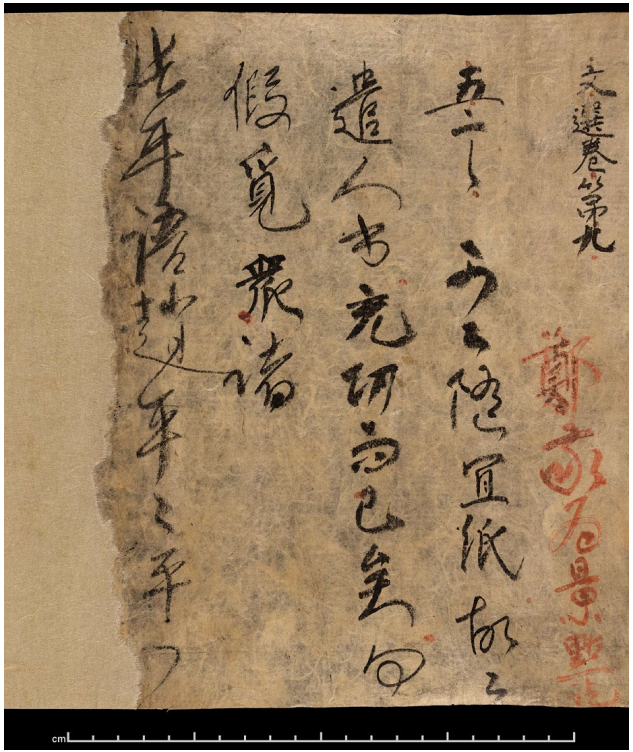
- segmentation





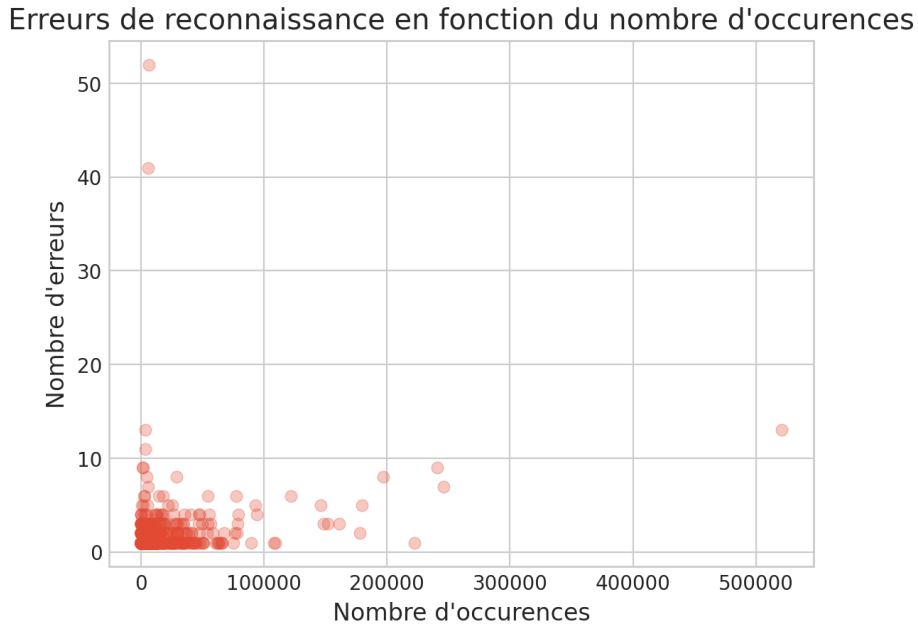
1.4 Limitations

- segmentation
- reconnaissance des styles fortement cursifs



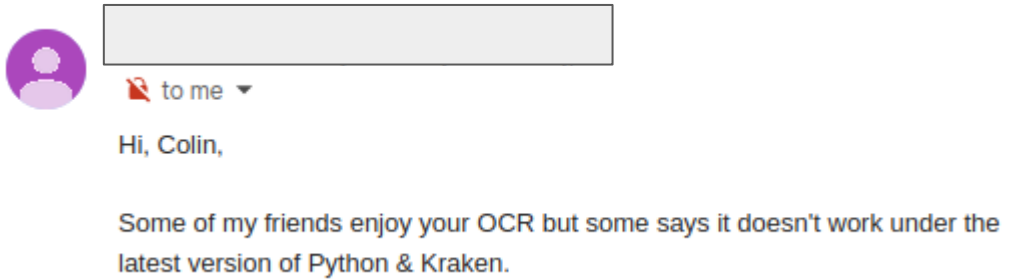
1.4 Limitations

- segmentation
- reconnaissance des styles fortement cursifs
- reconnaissance des caractères rares



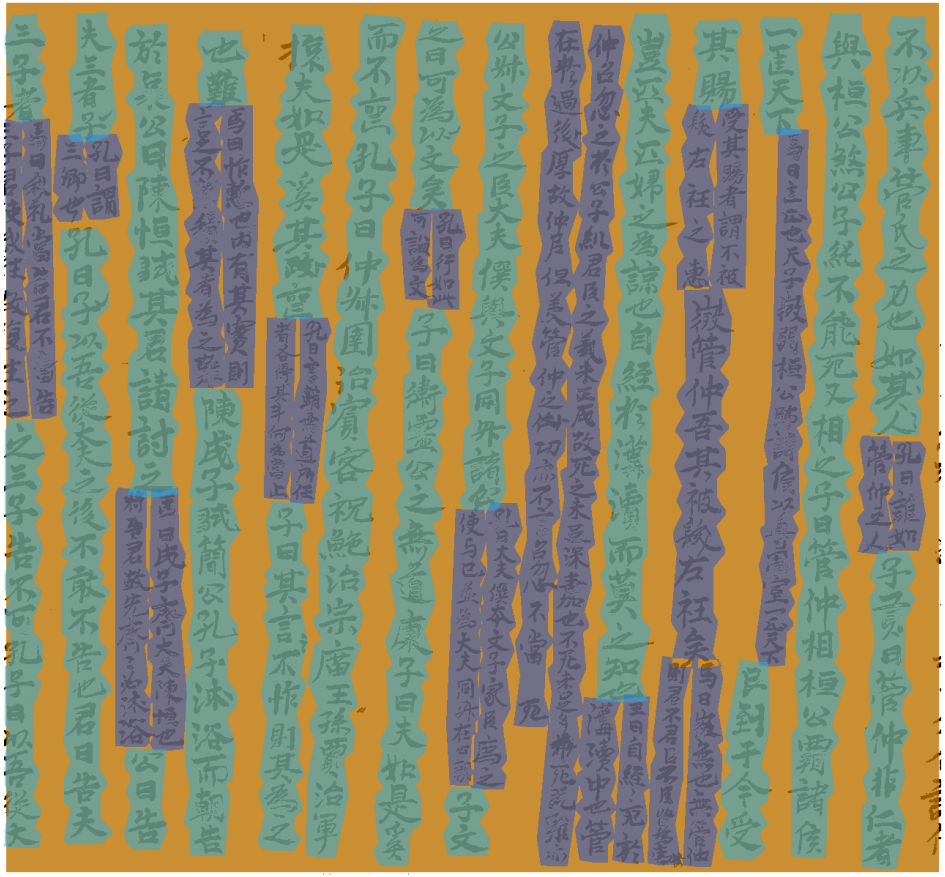
1.4 Limitations

- segmentation
- reconnaissance des styles fortement cursifs
- reconnaissance des caractères rares
- diffusion



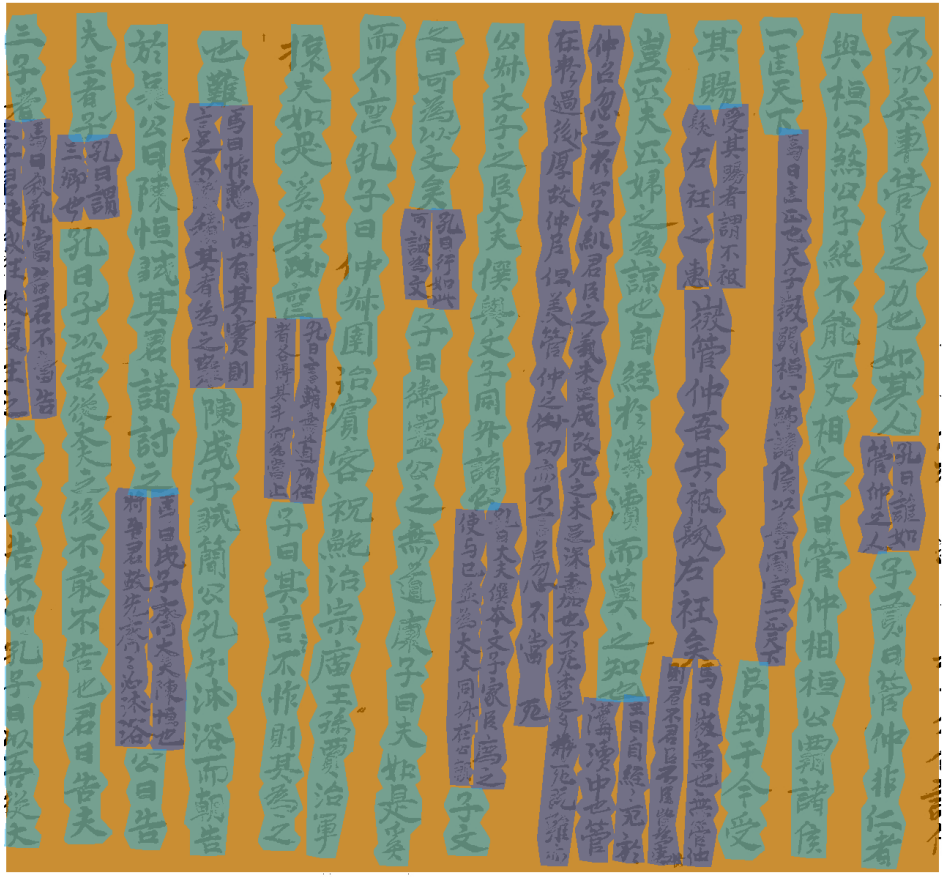
2. Coming soon

- amélioration de la segmentation



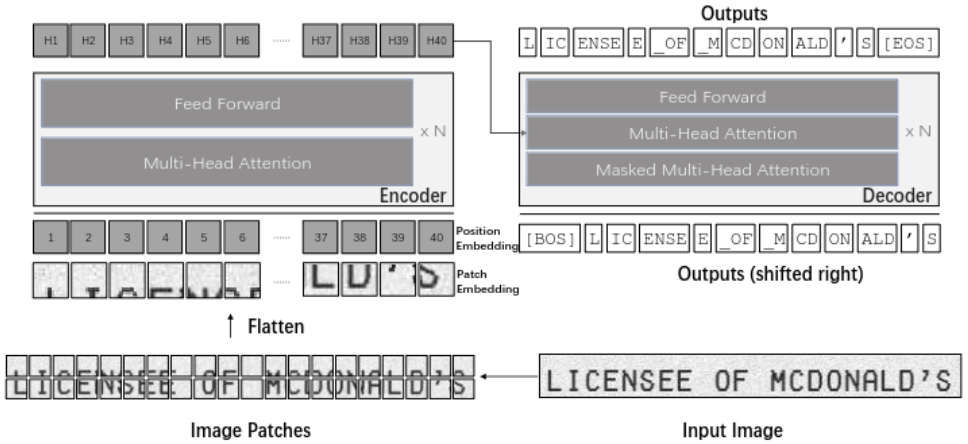
2. Coming soon

- amélioration de la segmentation
- alignement massif de nouvelles données



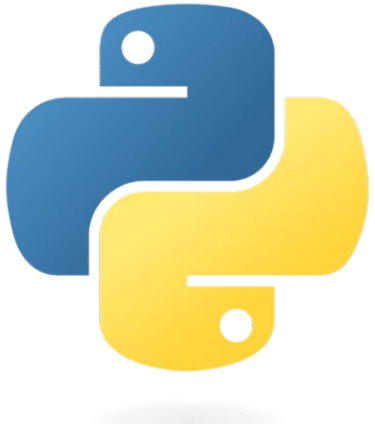
2. Coming soon

- amélioration de la segmentation
- alignement massif de nouvelles données
- architecture Tr-OCR



2. Coming soon

- amélioration de la segmentation
- alignement massif de nouvelles données
- architecture Tr-OCR
- package



Merci !