

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 législation 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

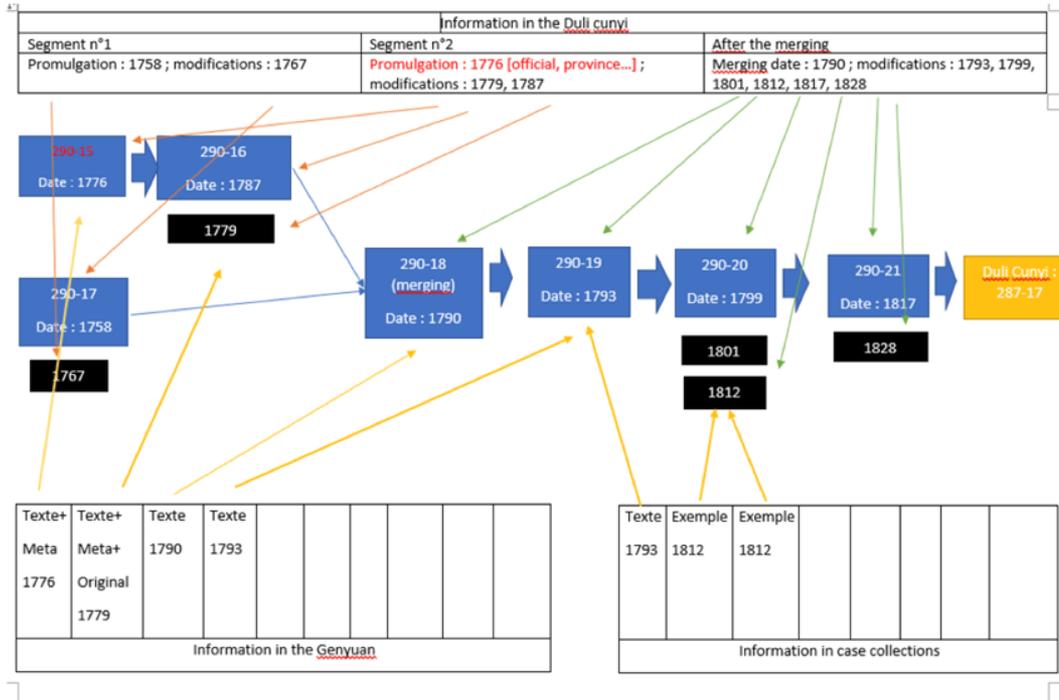
Virtual Sumptuary Code

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)Huaaji miaopan(驚心駭目)滑稽妙判		Fu Waishi 美外史(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) 自屢言 [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 新鶴法家透膽寒 [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 按吳親審機稿 [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 principaux (*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

Type

Title (zh)

Title (zh-Latn)

Sub-statute 282-9-7

Related (h) 282-9-9,o

Related (d)

Related (c)

Related (g) *280-5

Ghost

Date 1725

© Annuler

卷五十一。百根者杖六十。每百根加一等。

古者擊獲不論珠數多
凡三千里旗人銷去旗
領打珠之號騎校並總
處。一。

入領票砍伐木植如有
林數多寡定罪砍至數

續纂

令展轉扳指違者忝究治罪

竊盜

路背

捕人員 家收藏

一 在熱河承德府所屬地方偷挖金銀礦砂
論人數砂數多寡爲首俱枷號三箇月係
人發雲貴兩廣極邊烟瘴充軍以足四千
爲限係蒙古人發四省驛站當差爲從係
人枷號三箇月解回內地杖一百徒三年

大野律例長原 刊正 嚴監中 空 盜田野

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

Transcription automatique

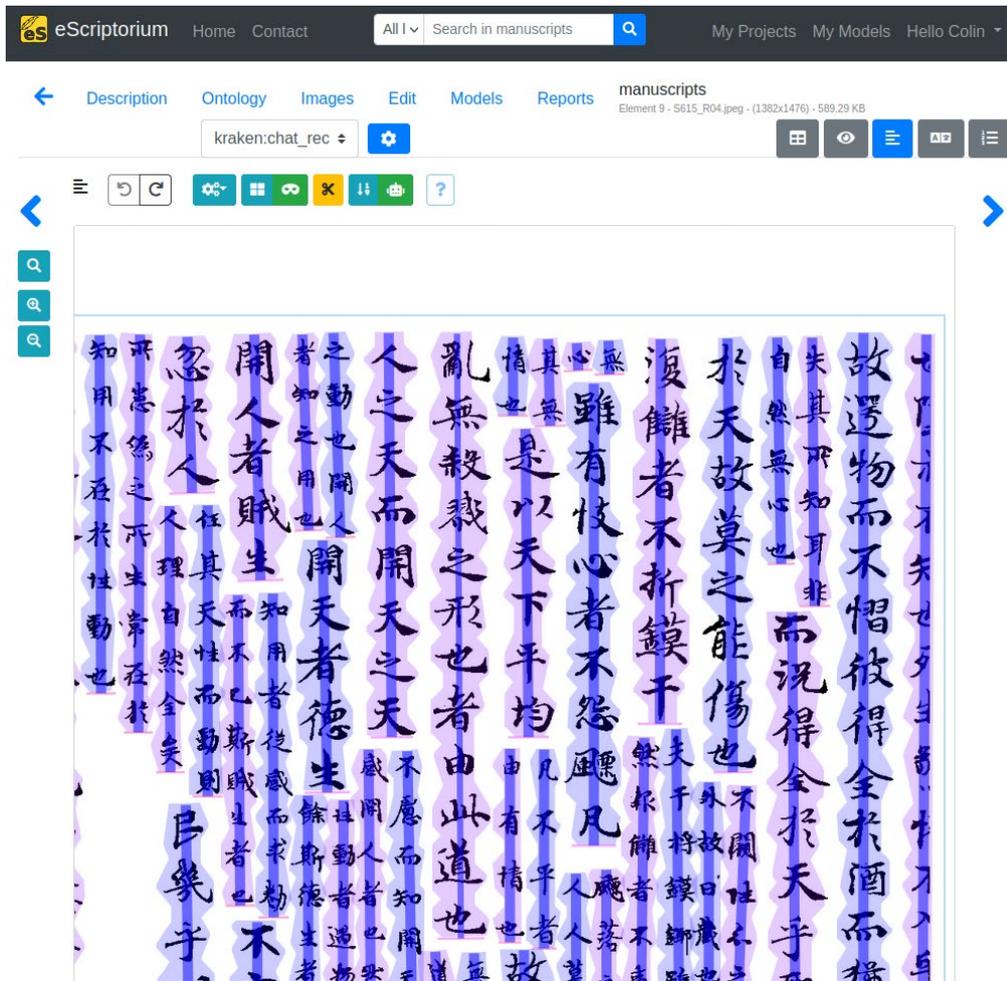
1.1 Vérité de terrain pour la segmentation

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>



The screenshot displays the eScriptorium web application interface. At the top, there is a navigation bar with the eScriptorium logo, 'Home', 'Contact', a search bar containing 'All I', and a user profile 'Hello Colin'. Below this is a secondary navigation bar with tabs for 'Description', 'Ontology', 'Images', 'Edit', 'Models', and 'Reports'. The main content area shows a document titled 'kraken:chat_rec' with various icons for editing and viewing. The document itself is a page of handwritten Chinese text in cursive script. The text is segmented into vertical columns, with some characters highlighted in blue and others in purple. The segmentation appears to be based on individual characters or small groups of characters, demonstrating the application's ability to recognize and segment complex handwritten text.

1.2 Vérité de terrain pour la reconnaissance



 **Kanseki Repository 漢籍リポジトリ**
Comprehensive collection of premodern Chinese texts. Licensed as CC BY SA 4.0.
🔍 42 followers 📍 Kyoto, Japan 🔗 <https://www.kanripo.org>

<https://github.com/kanripo>

1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées



<https://github.com/kanripo>

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柰何

定公者公子宋也昭公之弟也昭薨於乾侯季孫逆

其喪廢太子行及務人而立公子宋焉喪至於壞墮

公子宋先入以主社稷蓋受之季氏也非魯於是曠

年無君文公羊春秋欲謹之而不可也季氏廢太子行

1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées

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嶮峻崔嵬之状也支者勢
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
窟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
筋曰骨曰肉不曰穴葬穴非竅眼曰土曰石曰泉

1.2 Vérité de terrain pour la reconnaissance

3 millions de lignes alignées

度病床無穩時弟兄消息從獨欽向陽眉

社後

社後重陽近雲天淡薄間日隨暮客靜心共睡僧
河歸鳥城街日殘紅雨在山宗寥思晤語何夕款柴
關

息慮

息慮狎群鷗竹藏合自由春寒宜酒病夜雨入
鄉愁道向危時見官因世亂休外人相待淺
獨說濟川舟

農興

曉景山河英閑居巷陌清已能消滯念兼得
散餘醒汲水人初

1.2 Vérité de terrain pour la reconnaissance

- 内 (U+5185) → 内 (U+5167)



<https://github.com/kanripo>

1.2 Vérité de terrain pour la reconnaissance

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



<https://github.com/kanripo>

1.2 Vérité de terrain pour la reconnaissance

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



<https://github.com/kanripo>

1.2 Vérité de terrain pour la reconnaissance

- 内 (U+5185) → 内 (U+5167)
- 黄 (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

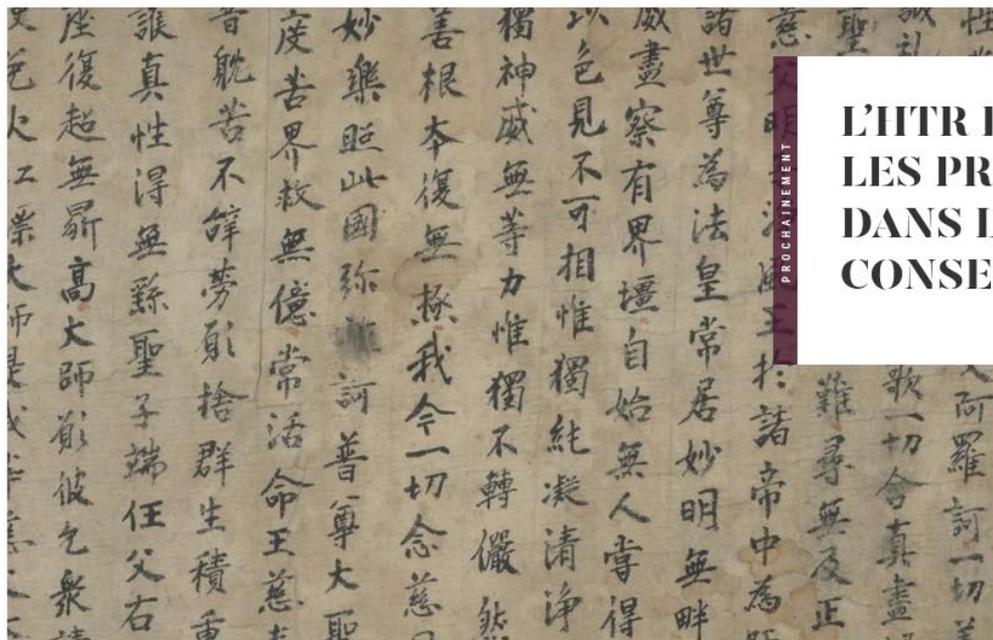
Errors Correct-Generated
520 { 巳 } - { 巳 }
298 { 巳 } - { 巳 }
117 { 於 } - { 於 }
115 { 己 } - { 己 }
97 { 母 } - { 母 }
82 { 蒙 } - { 蒙 }
71 { 髮 } - { 髮 }
63 { 日 } - { 日 }
51 { 於 } - { 於 }
48 { 世 } - { 世 }
47 { 傳 } - { 傳 }
46 { 世 } - { 世 }
42 { 己 } - { 己 }
40 { 博 } - { 博 }
40 { 日 } - { 日 }
39 { 補 } - { 補 }
39 { 成 } - { 成 }
```



Chinese Historical documents Automatic Transcription (CHAT) models

https://github.com/colibrisson/CHAT_models

1.2 Vérité de terrain pour la reconnaissance



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

COLLOQUES

14 fév. 2024

9h - 17h

Richelieu

Salle des conférences

1.2 Vérité de terrain pour la reconnaissance

至誠禮一切慧性稱讚歌一切含真盡歸仰
蒙聖慈光救離魔難尋無及正真
常慈父明子淨風王於諸帝中為師帝
於諸世尊為法皇常居妙明無畔界
光威盡察有界壇自始無人嘗得見
復以色見不可相惟獨託凝清淨德
惟獨神威無等力惟獨不轉儼然存
衆善根本復無拯我今一切念慈恩歎
彼妙樂照此國紛訶普尊大聖子
廣度苦界救無億常活命王慈喜美
大普脫苦不辭勞勇捨群生積重罪
善護真性得無絲聖子端伍父石座
其座復超無罪高大師前彼乞衆請降
棧使免火江漂火師是我等慈父大師

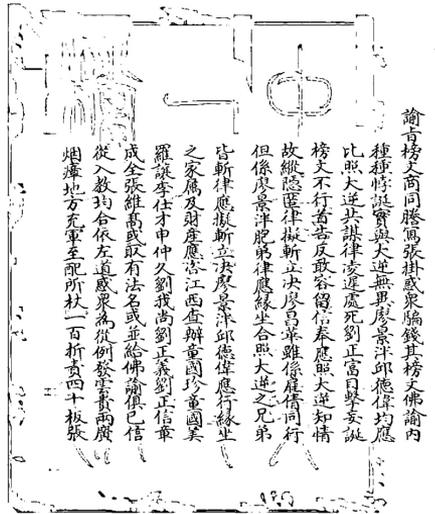
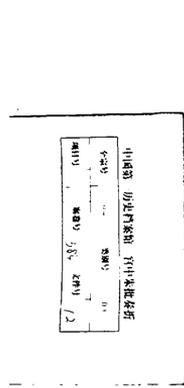
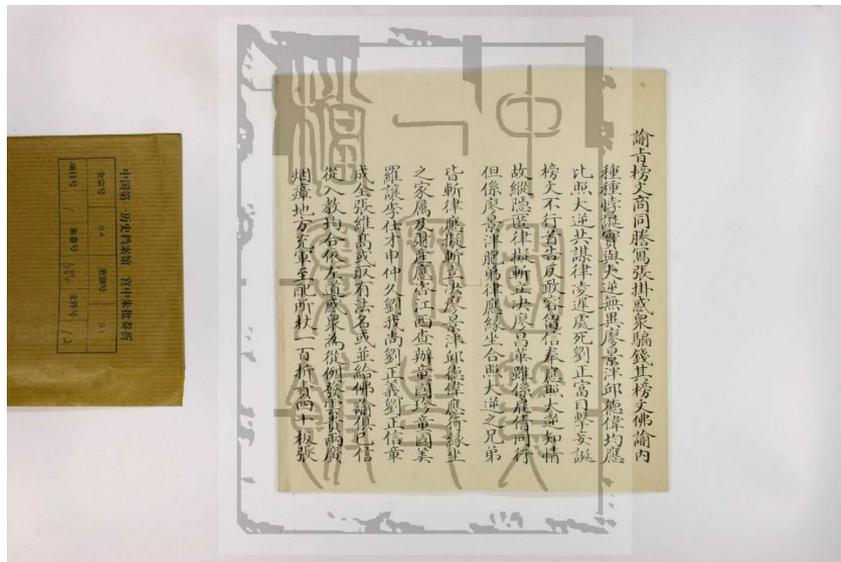
至誠禮一切慧性稱讚歌一切含真盡歸仰
蒙聖慈光救離魔難尋無及正真
常慈公明子涕風王於諸帝中為師帝
於諸世尊為法皇常居妙明無畔界
光威盡察有界壇自始無人嘗得見
復以色見不可相惟獨託凝清淨德
惟獨神威無等力惟獨不轉儼然存
衆善根本復無拯我今一切念慈恩歎
彼妙樂照此國彌以詞普尊大聖子
廣度苦界救無億常活命王慈喜美
大普脫苦不辭勞勇捨群生積重罪
善誰真性得無絲聖子端征父石座
其座復超無罪高大師前彼允衆請降
棧使免火江漂火師是我等慈父大師

1.2 Vérité de terrain pour la reconnaissance

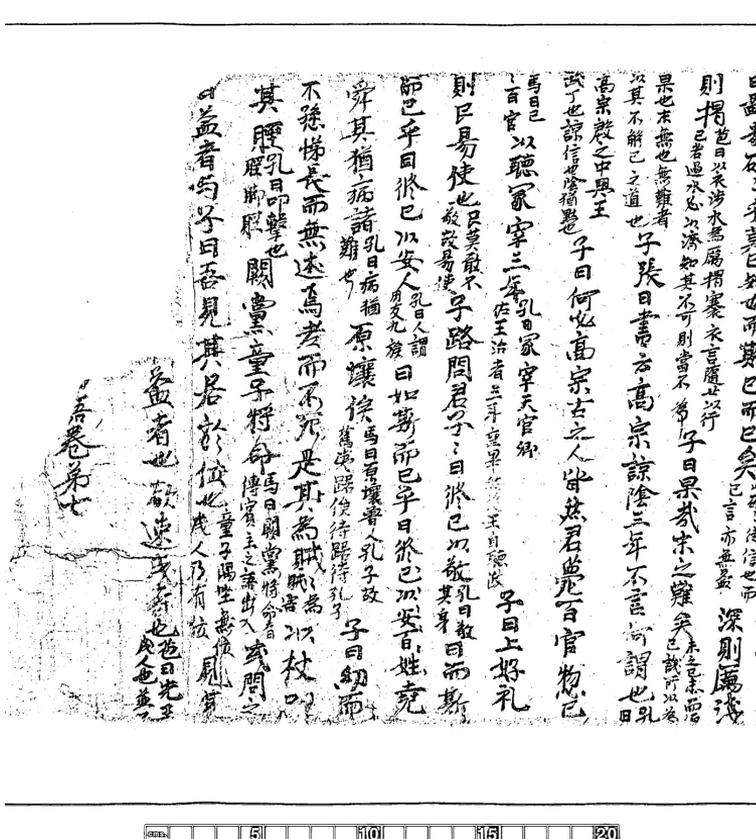
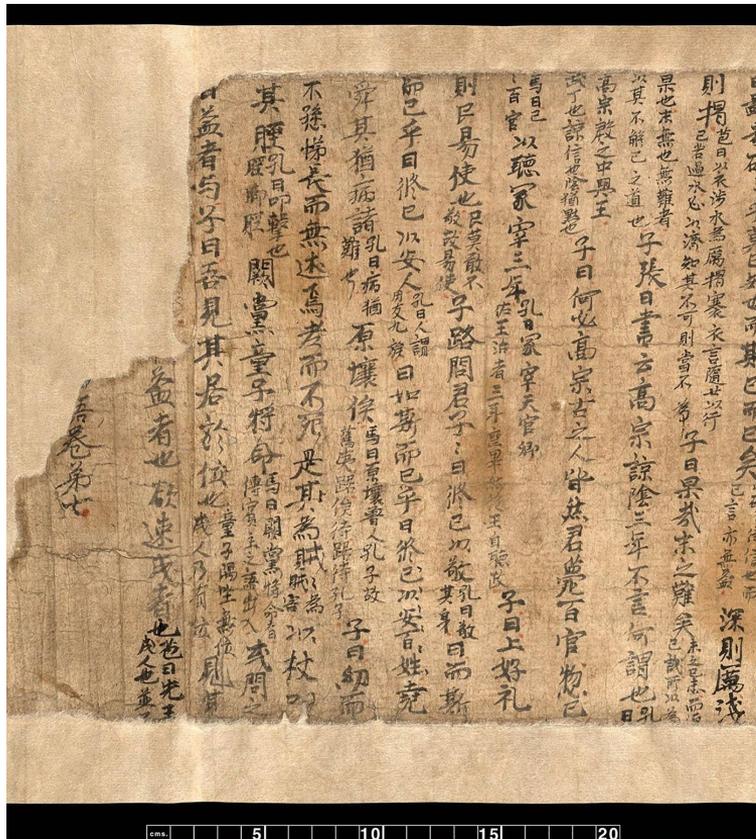
德之怡如也可謂士矣朋友切悃兄弟
 日善人教民古平亦可以即戎矣
 棄之馬曰言用不習之人使
 道穀孔曰雲祿也邦邦無道穀恥也孔曰君無
 知也難不足以為仁邦無道穀恥也朝食祿
 焉可書以為仁矣馬曰克好勝人使自伐子曰
 危言危行危言也邦有有道邦無道穀恥也
 者必有言德不可以慎中故女有言有言

者必有言德不可以慎中故女有言有言
 危言危行危言也邦有有道邦無道穀恥也
 知也難不足以為仁邦無道穀恥也朝食祿
 焉可書以為仁矣馬曰克好勝人使自伐子曰
 道穀孔曰雲祿也邦邦無道穀恥也孔曰君無
 棄之馬曰言用不習之人使
 日善人教民古平亦可以即戎矣
 德之怡如也可謂士矣朋友切悃兄弟

1.3 Binarisation



1.3 Binarisation



1.4 Limitations

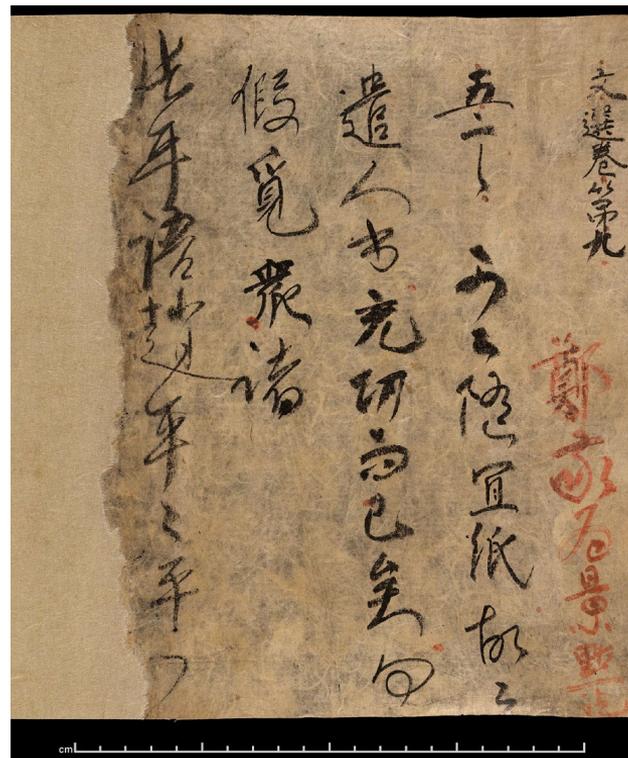
- segmentation

正終也。昭無正終。故定無正始。不言即位。表在外也。
 元年必書正月。謹始也。穀梁元年。雖無事。必舉正月。正諸侯。定何以無正月。昭公薨於乾侯。不得正其終。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。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。具喪廢太子。行及務人。而昭公子宋馬喪。至於壞墮。公子宋先入。以主社稷。蓋受之季氏也。非魯於先君者。也。定無正。不言正月。微辭也。
 年無君。公羊文九。春秋欲謹之。而不可也。季氏廢太子。行

夕之故。夫人焉可須臾不學。此應首。深耕熟耰。穰穰在秋。寒之落矣。然而有稷。有黍。有禾。有秬。農視其穫。先視其播。此言人莫不讀書。六經。歟。五穀也。諸子百家。歟。稱稗也。此言書末。專言讀書。世有一輩人。漫不省立身。行已。大丈夫事。日出小技。驕釋朋儕。文章滄海一粟。爾焉能名世。此非讀又有一輩人。好利惡衰。厭苦逐樂。色色。隨無名種子。中虛名。浮利。世界一粟耳。何足介意。此非讀書者。結曰。皆非讀書者。吾家山谷。嘗云。吾輩但勿令書種斷絕。若夫成功。則天居士。試著轉語。看。下何種。百丈和尚作一頌。云。轉語。某曰。此真一粟。

1.4 Limitations

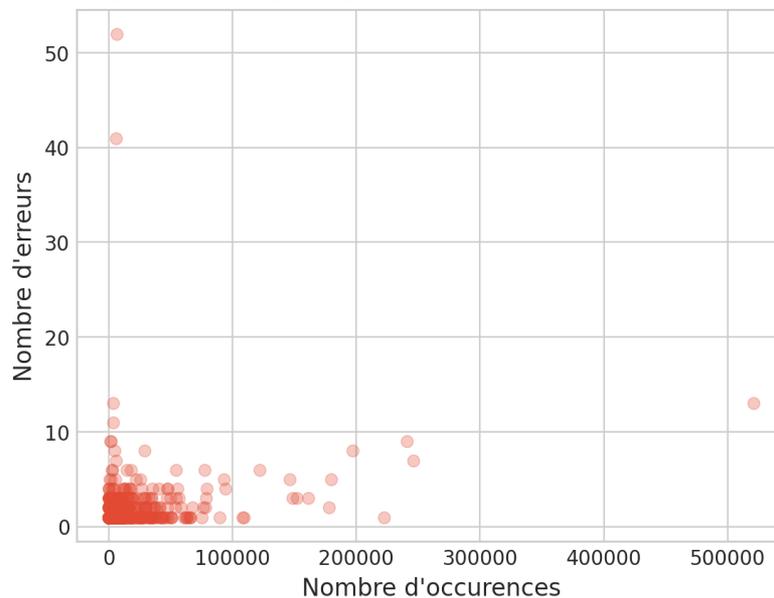
- segmentation
- reconnaissance des styles fortement cursifs



1.4 Limitations

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

Erreurs de reconnaissance en fonction du nombre d'occurrences



1.4 Limitations

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



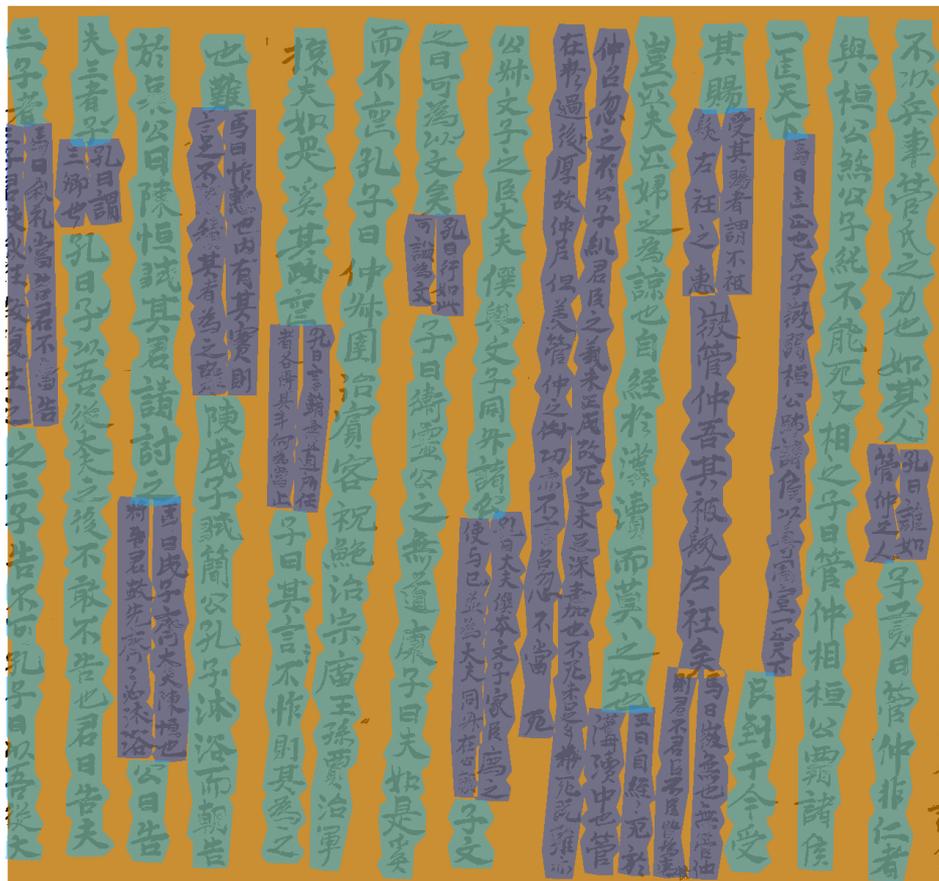
 to me ▾

Hi, Colin,

Some of my friends enjoy your OCR but some says it doesn't work under the latest version of Python & Kraken.

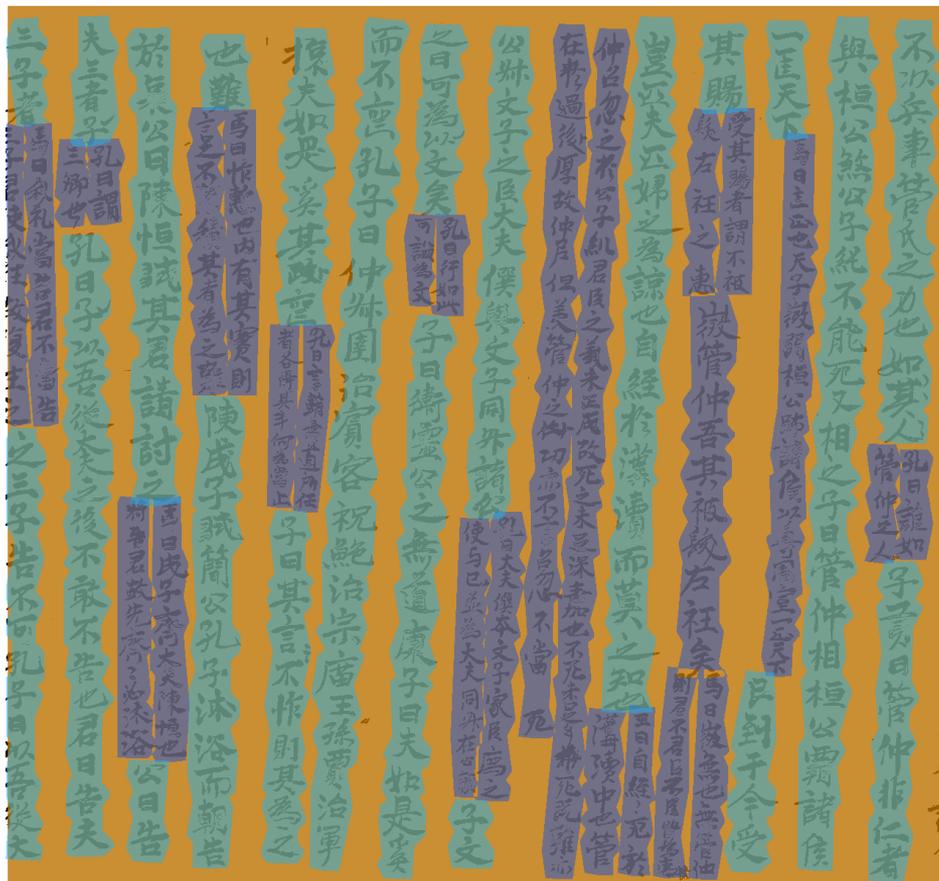
2. Coming soon

- amélioration de la segmentation



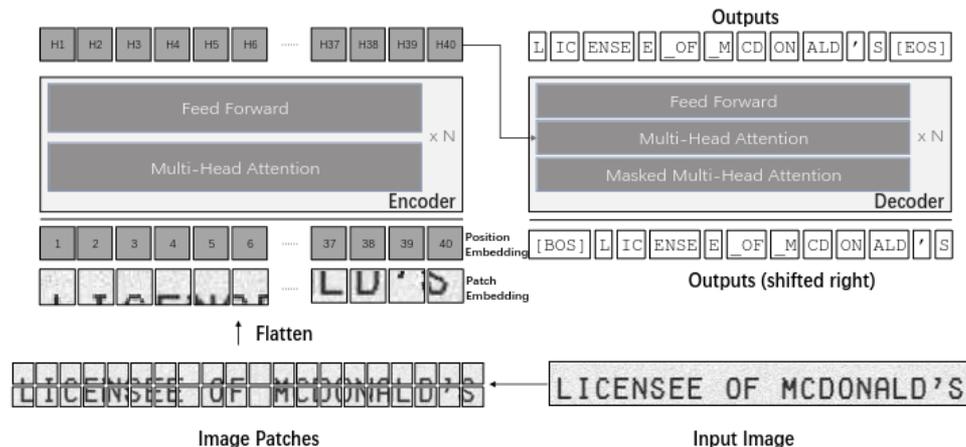
2. Coming soon

- amélioration de la segmentation
- alignement massif de nouvelles données



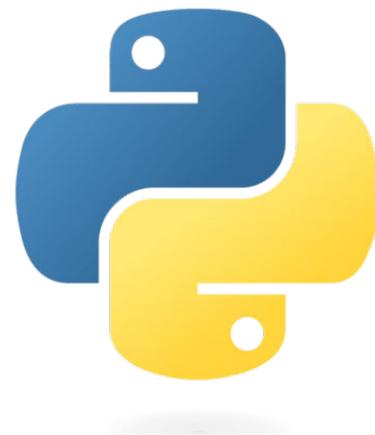
2. Coming soon

- 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 !