Le(s) chinois du Shun-Pao

Pierre Magistry
pierre.magistry@inalco.fr

7 Décembre 2021
Grenoble

Le Corpus

申報

Le Shun-pao, mon «terrain» à Shanghai il y a 150 ans

- Le premier quotidien imprimé en sinogrammes
- 750M de sinogrammes
- 1872 – 1949, un période de grandes mutations
  - fin de l'Empire, début de la République de Chine
  - abandon du chinois classique au profit des langues régionales
  - définition d'une langue nationale unique
- une première une qui ne ressemble pas vraiment à la dernière

申報

第二萬五千二百七十九號

發行人　陳景韓

今日本報貳大張

每份售國幣八萬元
（外埠另加郵費）

地址上海漢口路三〇九號　電報掛號六〇七八號　電話九三二四八號四錄

立院今開秘密會議
討論國家總預算案
物價問題亦將提出檢討

〔本報南京一日電〕立法院長孫科，一日下午五時招待記者，籲請對該院會議機密之保持予以協助。孫氏首聲明招待會旨在根據院會決議，代表立委同人與記者交換該院新聞發布之意見，決無其他用意，請勿誤會。旋解釋該院依法舉行秘密會議，謝絕旁聽，乃為避免將不應公開或未至公開時機之討論事項宣揚院外，致生不良影響。且消息之片斷發布，易滋誤解，危國家安全，社會安定，故不得不仿照一般行政性會議之秘密舉行。六月廿四日臨時會議，涉及軍事機密，應防止洩露予敵人得悉，而所有足以刺戟物價，危害一般人生活及國家財政之消息，自當儘量避免登載。渠以個人之立場認為秘密會議容易引起猜測謠諑，使人心搖惑，應公開為宜，而一國議會在平時舉行秘密會議實屬例外，蓋公開庶政，取信人民，乃民主國家根本原則，但如國家遭遇戰亂時，為國家安全，自不得不舉行秘密會議。孫氏復聲明二日之秘密會議，乃為聽取政府關於下半年預算編製之報告，為使政院報告暢所欲言，俾立院可得充分瞭解，故有此項措施。渠謂：會議雖稱秘密，但會議情形記者當然可能知悉，但請暫放心中，俟有決定後，再行報道。至此，渠以英國前財相道爾頓之辭職為例，指出民主國家議會對預算案之重視。英例預算案在下院提出後即可公諸全國，然我國行憲伊始，立院為昭慎重，乃決定俟有決議時，再對社會公開，此為避免刺戟人心，影響物價而採取之非常時期措施。同時，會中對物價問題，亦將予以檢討，亦盼記者幫忙，協同保守機密。孫氏語末，強調民主國家中新聞界與議會關係密切，我國民主政治之發展，尤賴記者等與立院同人加强聯繫，共同努力。最後並接納記者多人建議，對以後採訪該院新聞，儘量予以可能方便。

〔本報南京一日電〕立法院第一會期第三次秘密會議，定二日晨九時舉行，討論卅七年下半年度中央政府歲入歲出總預算案。普通部門總預算草案及施行條例草案，業經政院附同施政方針，施政計劃與編審辦法送達立院，特別歲入歲出預算草案，亦已另案提出，政院翁院長將偕主計長徐堪，財長王雲五列席，說明有關總預算案編製之各事項。同時，並報告物價飛漲問題，央行總裁俞鴻鈞亦被邀列席，以備質詢。

〔本報南京一日電〕國民黨籍立委，一日下午三時在中央黨部舉行談話會，就政院行將提出之總預算案交換意見，負責主管之翁文灝．張厲生．王雲五．徐堪等均列席，將總預算案提出作詳盡說明，各立委亦提供若干意見，至下午五時許始散。

協定中雖未明文規定，但經濟作局有權聘請參加馬歇爾計劃之國，作關於改革幣制之商談，美國無意干涉該國之內政，故由各國自行從事其所需之幣制革。霍氏稱：經濟合作局之地與投資銀行家同，故該局投資，而受援國復原工作不能適當則合作局可停止其援助，或請國匡正阻礙其發展之局勢。霍氏就簽訂雙邊協定發表正式聲謂：此一計劃之成敗，將視大洋彼岸同志之努力如何而定。

司法院舉行接
王寵惠就
勉僚屬共同

〔本報南京一日電〕司法院長王寵惠，一日晨十時蒞司法，正式就職。前院長居正於十前到院，等候王院長蒞臨後，在院長室內交接鈐記，旋在大堂內，舉行簡單之交接儀式，體職員均參加。行禮如儀後，

擬製造局新刻西學書十三種序

今夫人欲習新奇之技藝必當有秘密之傳通迎機啟悟方能力窮其奧窒精析夫毫芒圖以爲測算之學而測算無不精微也挨其之人指點之益則雖有規矩迹象焉耳矣而焉故有知其所當然而不能知其所以然者則何以爲考究之資孟子有言曰羿之教人矩學者亦必以規矩此物此志也滬上自行之然執事於其間者大抵皆藉西師爲之知其所以然也聞有受其書而伏誦之者則不可不繙譯出之以爲工師之規矩儒生之數年之功費數萬之資現已印行之書則有夫西人技藝之精甲於天下洩造化而通神購之歐洲今既設局製造而又有繙出各書法之原起於積算勾絃圓徑製器所宗推步規約指三書萬物之質有凝有化凝全乎天原化學鑑原二書氣分輕養遇熱而融蒸發機輒汽機必以二書錐鑿雜施詎傷地脈若金石識別開煤要法二書海中船趨帆檣則有若御風要術航海簡法二書地球之下方時則有若地學淺釋一書火攻之毒至今則有若製火藥法一書以上十三種精深微皆戢上稱過者以之作爲工師之規矩儒生有爲加意於富國強兵之道取其所長集其將可以夐絕古今矣豈不美哉豈不美哉

吳門張少卿校書花燭詞　并序

吳門名校書張少卿者系出毘陵間小家之

Objectifs

Histoire assistée par ordinateur

- Améliorer la structuration du corpus
- Faciliter la recherche d'information

Linguistique de corpus

- observer la naissance du chinois moderne dans les données de la presse quotidienne
- caractériser les évolutions (ou ruptures)

Cadre et circonstances

2018-2019, postdoc dans le cadre de l'ERC

- difficultés pour obtenir le corpus
- pas de structure, ni de points de comparaisons
- intérêt à «faire du deep»

2021, reprise du sujet

- comparer à d'autres corpus
- liberté dans les méthodes (avec un retour à Salem, 1991)
- outils et méthodes

Deux moments

2018-2019, postdoc dans le cadre de l'ERC

- Traitement de données massives et brutes
- Modèles de langues et plongements contextualisés

2021, reprise du sujet

- Gros corpus un peu plus structuré
- Ajout de (petits) corpus de référence pour comparaison
  - chinois archaïque / médieval / pré-moderne
  - corpus équilibré de chinois standard (TW)
- Utilisation de l'Analyse Factorielle des Correspondances (AFC)

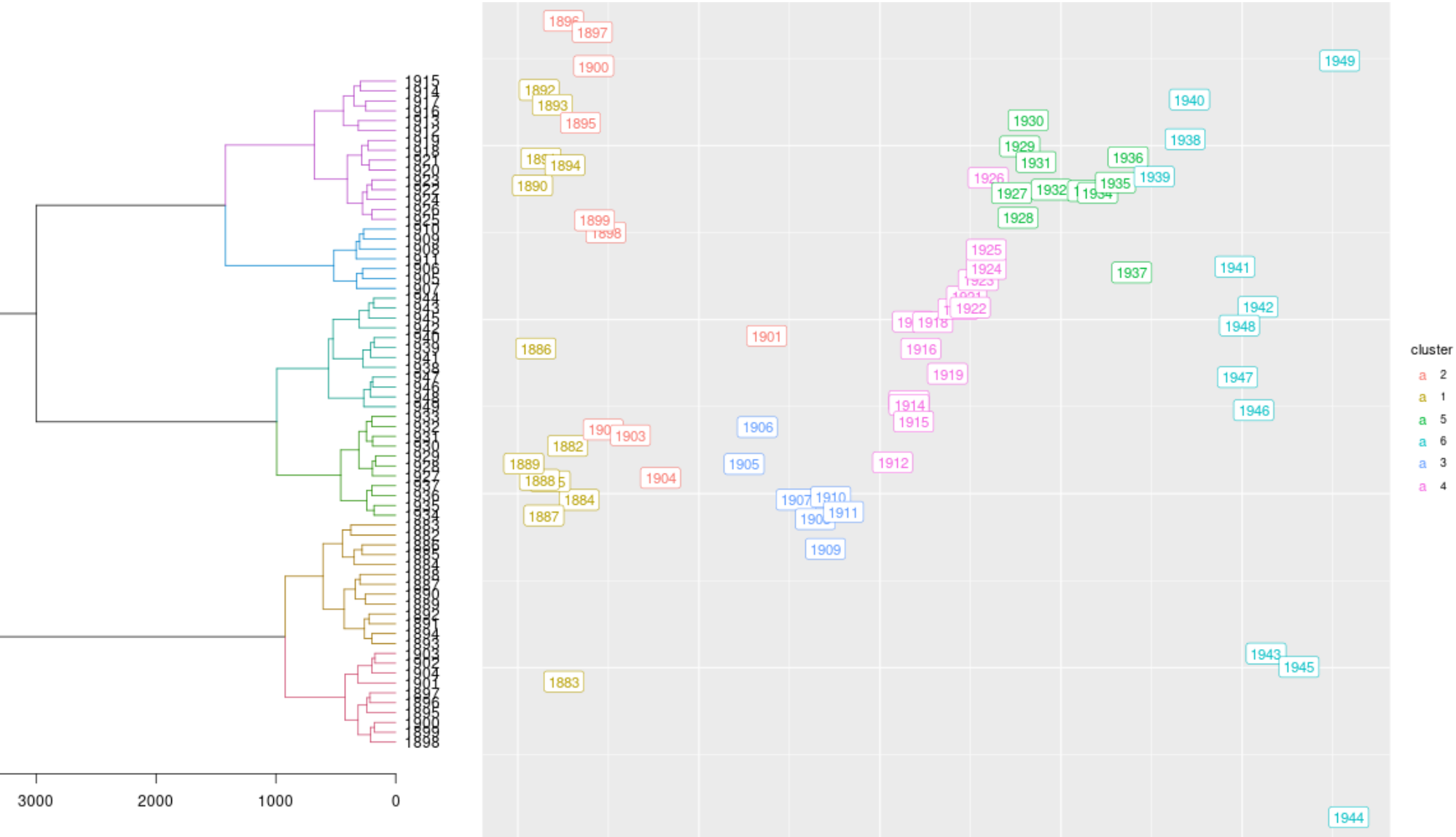
Perplexité de modèles de langue

Objectif

Détecter des périodes, ou des ruptures dans le temps.

Procédure

- diviser le corpus par année
- entraînement de 77 modèles
  - niveau du sinogramme
  - sans ponctuation
- construction d'une matrice de distances entre les années
- MDS et clustering hiérarchique pour visualiser


3000
2000
1000
0
cluster
a 2
a 1
a 5
a 6
a 3
a 4

Résultats

En comparant les sorties de différents algorithmes de clustering.

- 1894 – 1904
- 1905 – 1911
- 1912 – 1922
- 1927 – 1938
- 1938 – 1949

Plongements lexicaux contextualisés

Objectifs

Suivre des compétitions entre formes

Procédure

- sur chacune des 5 périodes précédemment définies
- entraînement d'un Bi-LSTM, à la *flair* (Akbik et al., 2018)
- choix d'une liste d'items en concurrence à échantillonner
- pour chaque occurrence, on obtient un vecteur associé à la forme, dans un contexte et à une date précise.
- projeter et clusteriser ces vecteurs

Character-based representation of words in context

from (Akbik et al., 2018)

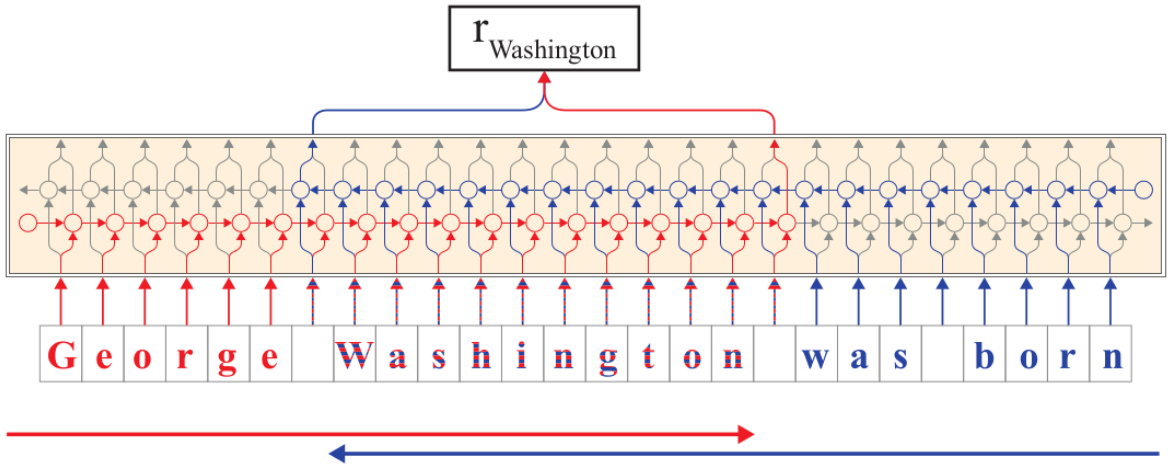


Figure 2: Extraction of a contextual string embedding for a word ("Washington") in a sentential context. From the forward language model (shown in red), we extract the output hidden state after the last character in the word. This hidden state thus contains information propagated from the beginning of the sentence up to this point. From the backward language model (shown in blue), we extract the output hidden state before the first character in the word. It thus contains information propagated from the end of the sentence to this point. Both output hidden states are concatenated to form the final embedding.

list of 字

很狠甚什極

time period

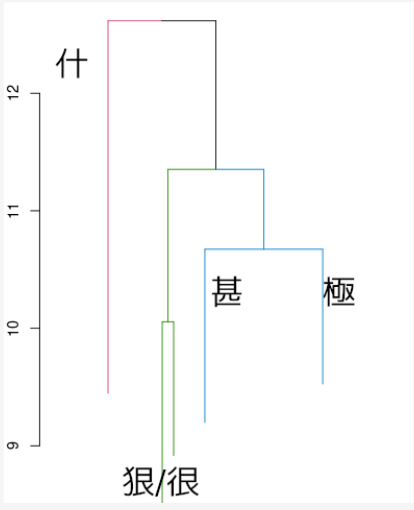
1894-1904

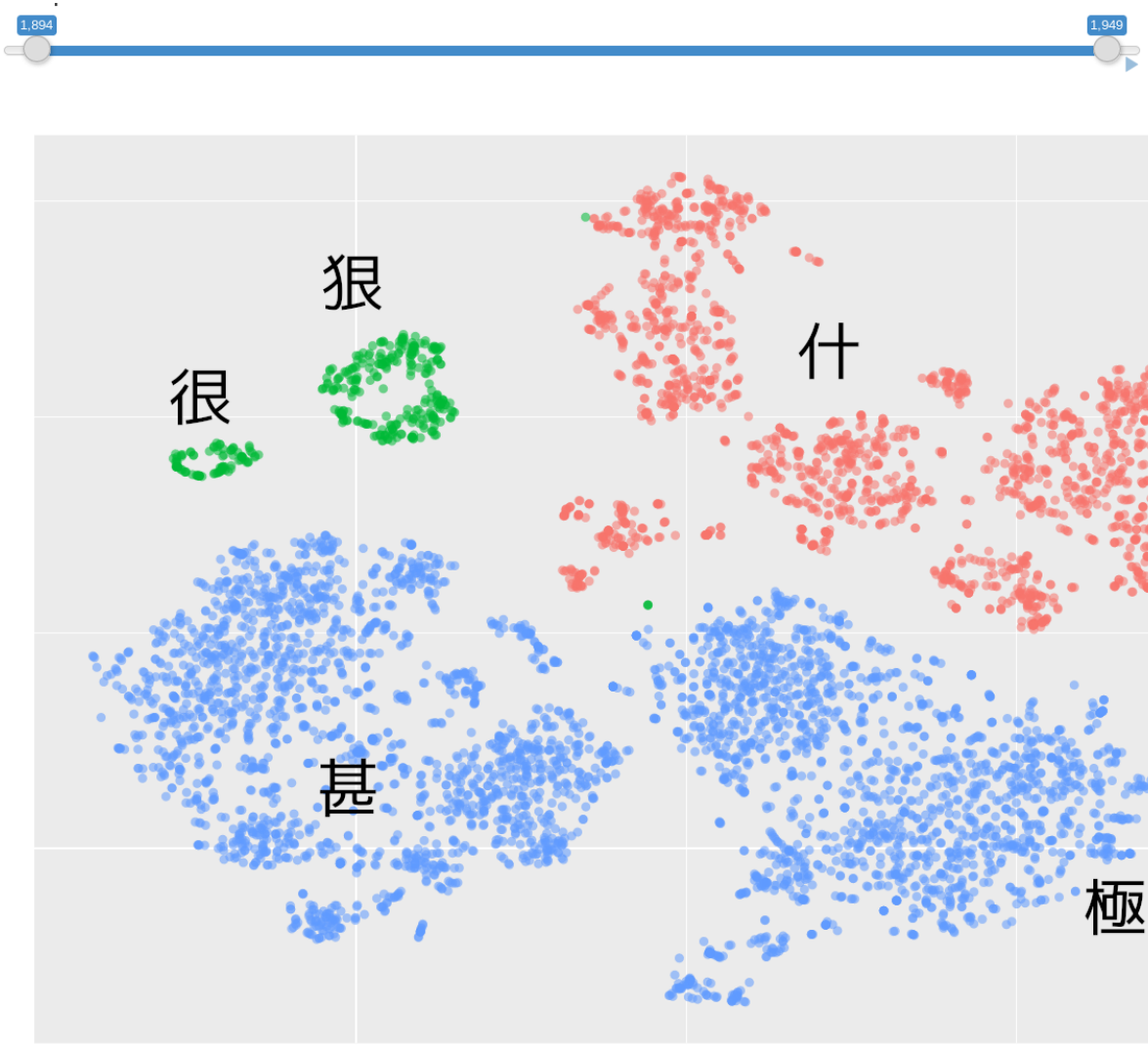
color by

字 cluster

number of clusters

3





Résultats (Magistry, 2019)

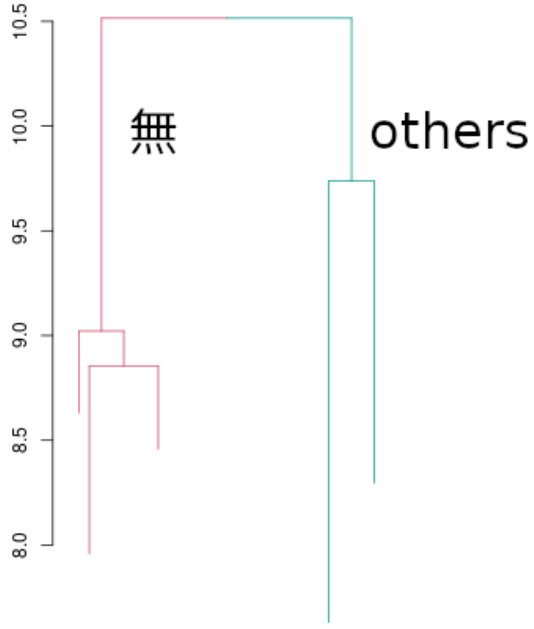
adverbes intensifieurs de verbes d'états

極、很、狠、甚、什

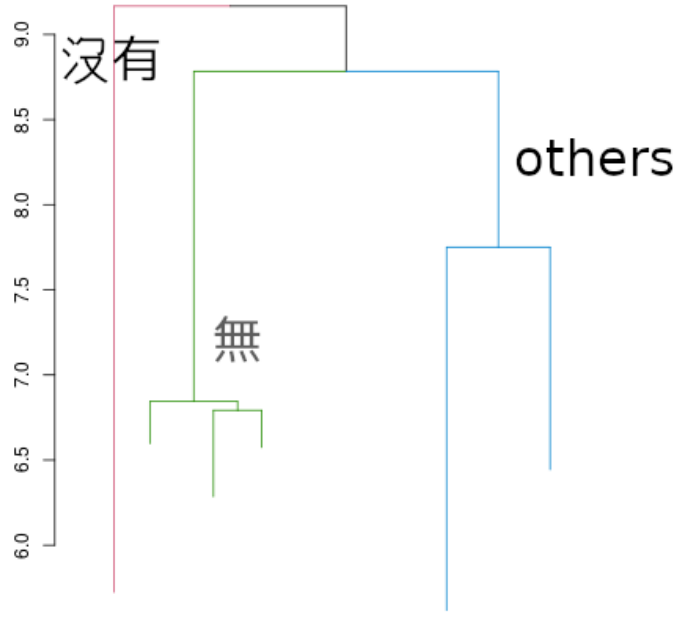
- 甚 et 什 variantes graphiques ? (juste dans 什麼)
- 很 et 狠 variantes graphiques ?
- 什 surtout dans des emprunts au manchu.
- figements de 極 au fil des époques.

négation des existentiels

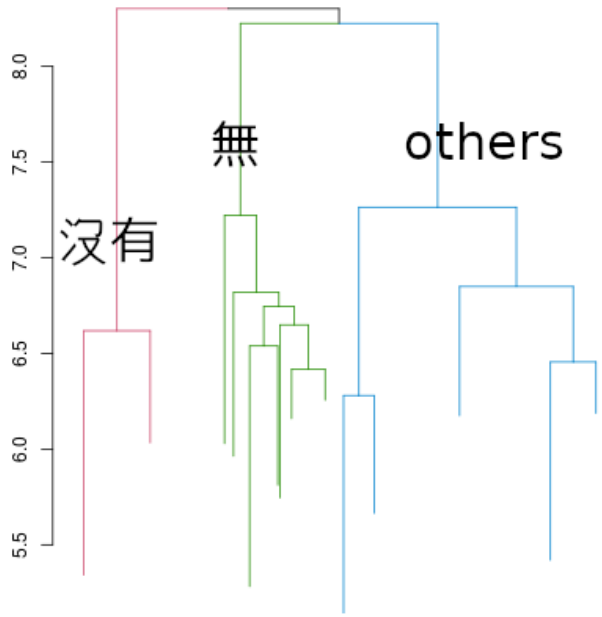
- Coblin (2000) décrit 未曾，不曾 et 沒有 comme équivalents, avec "an apparent freedom of choice"
- En fait 沒有 n'arrive que plus tard et semble distinct, sauf dans sur la fin où c'est 無 (classique) qui s'isole.



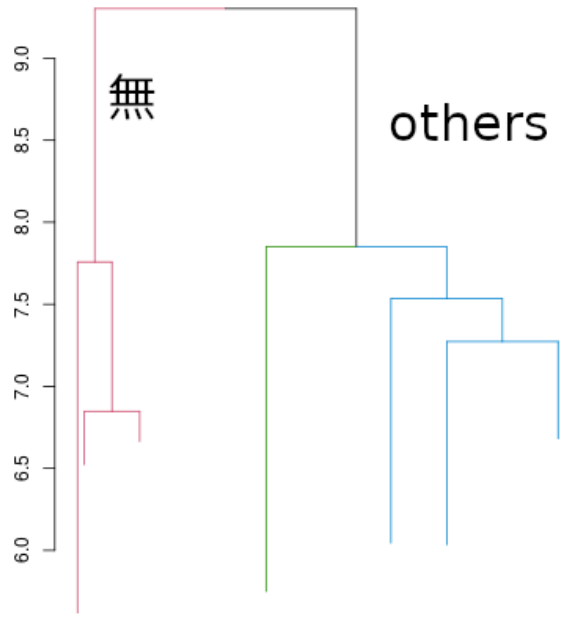
1894 ~ 1904



1905 ~ 1911



1912 ~1922



1939 ~1949

nouveaux corpus, nouvelle hypothèse

corpus historique Academia Sinica

- 上古 48 livres (-8 – +2)
- 中古 36 livres （2 – 6)
- 近世 13 livres (9 – 16)

Sinica Treebank

350k tokens, années 90

- news
- oral
- manuels scolaires
- magazines

Dans le Shun-pao

- colinguisme du début à la fin (selon les rubriques)
- particularité de la première page ?

Analyse des correspondances

(Salem, 1991) Les séries textuelles chronologiques

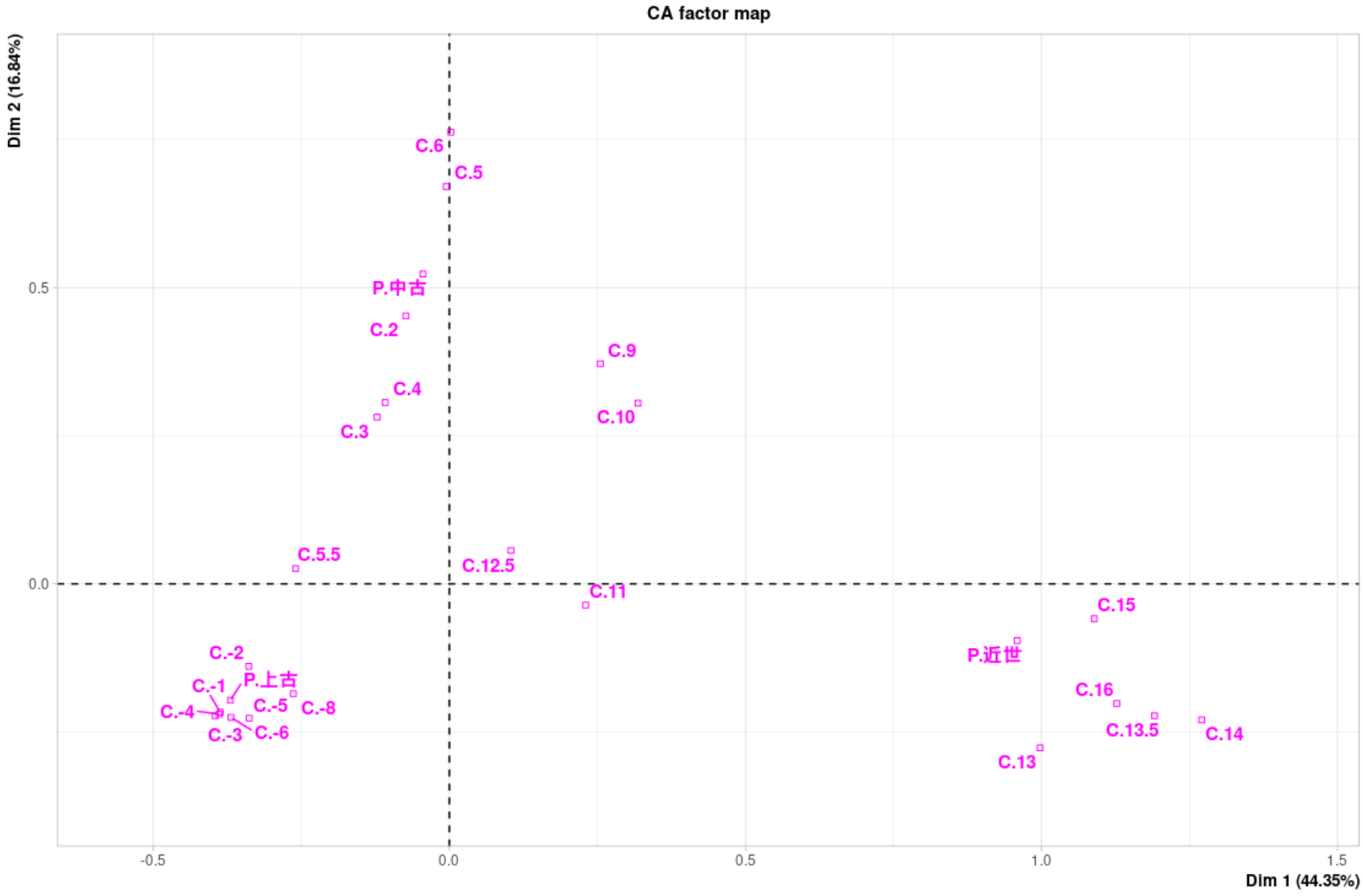
prétraitements

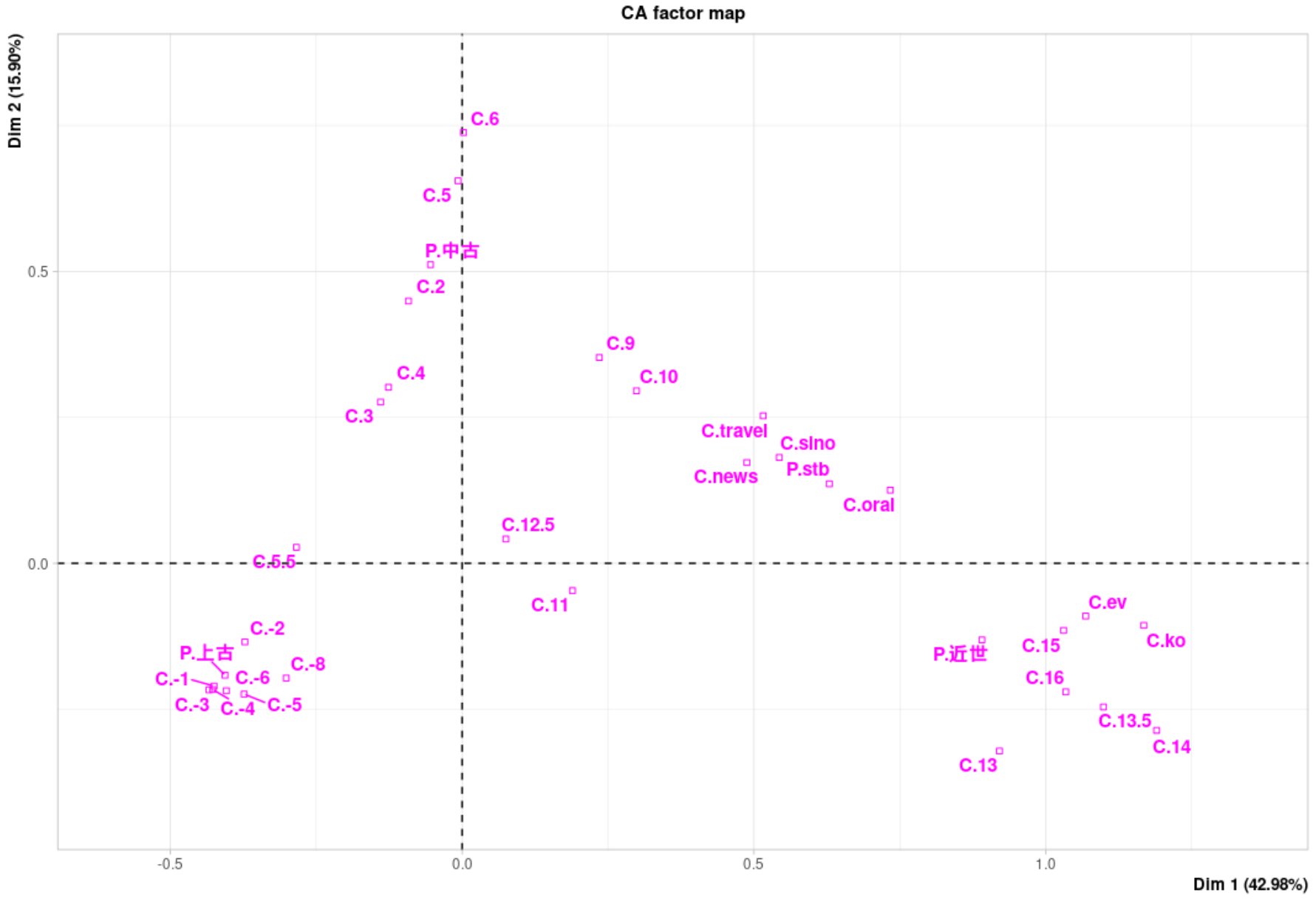
- sélection de mots gramaticaux dans les corpus de références
- compte d'occurrences par document
- du Shun-pao, pour chaque année je sample les premières et les avant-dernières pages (séparément, on obtient donc deux points par an).

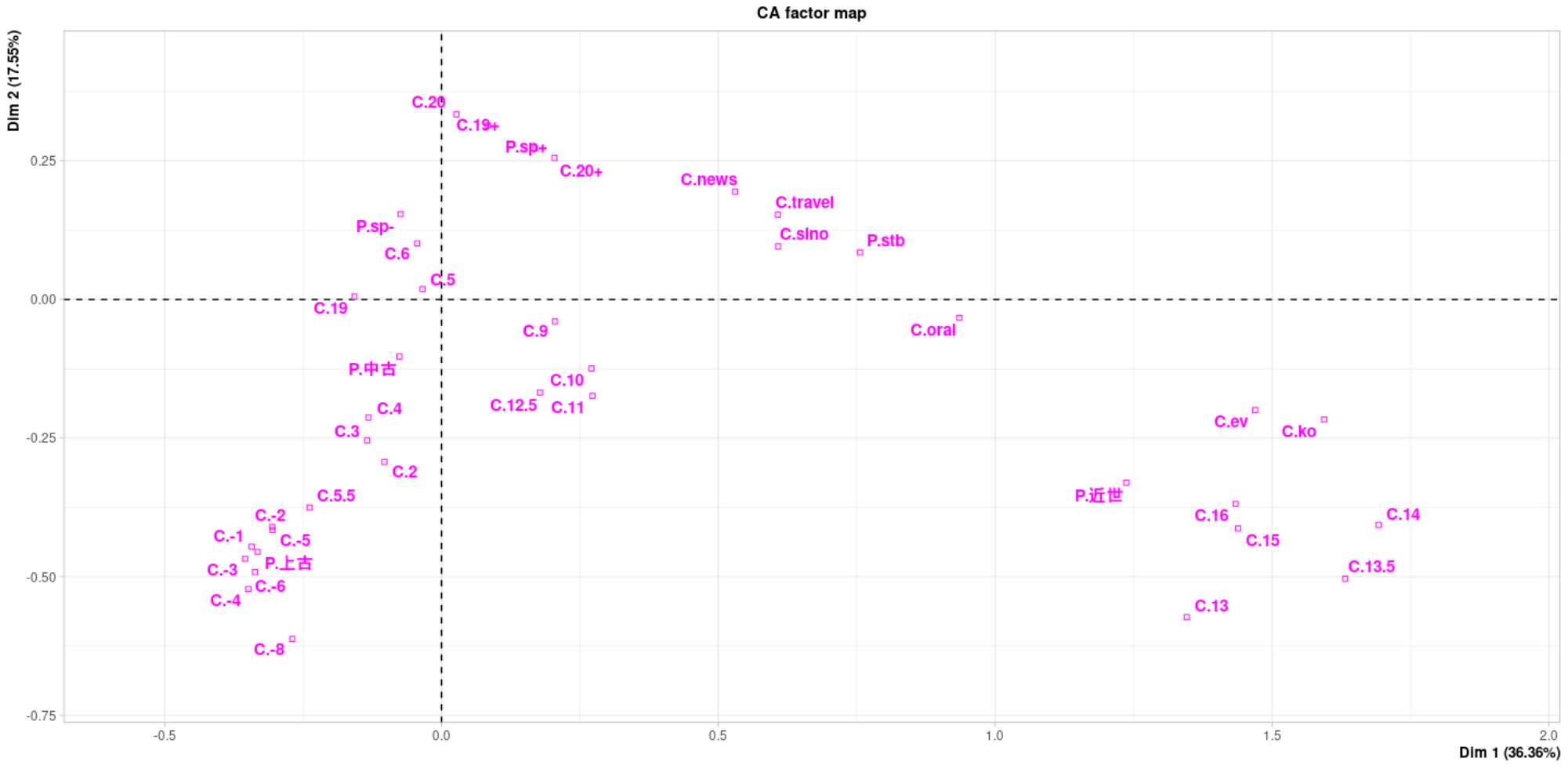
Calculs, avec FactoMineR (Lê, Josse & Husson, 2008)

- AFC
- clustering hierarchique

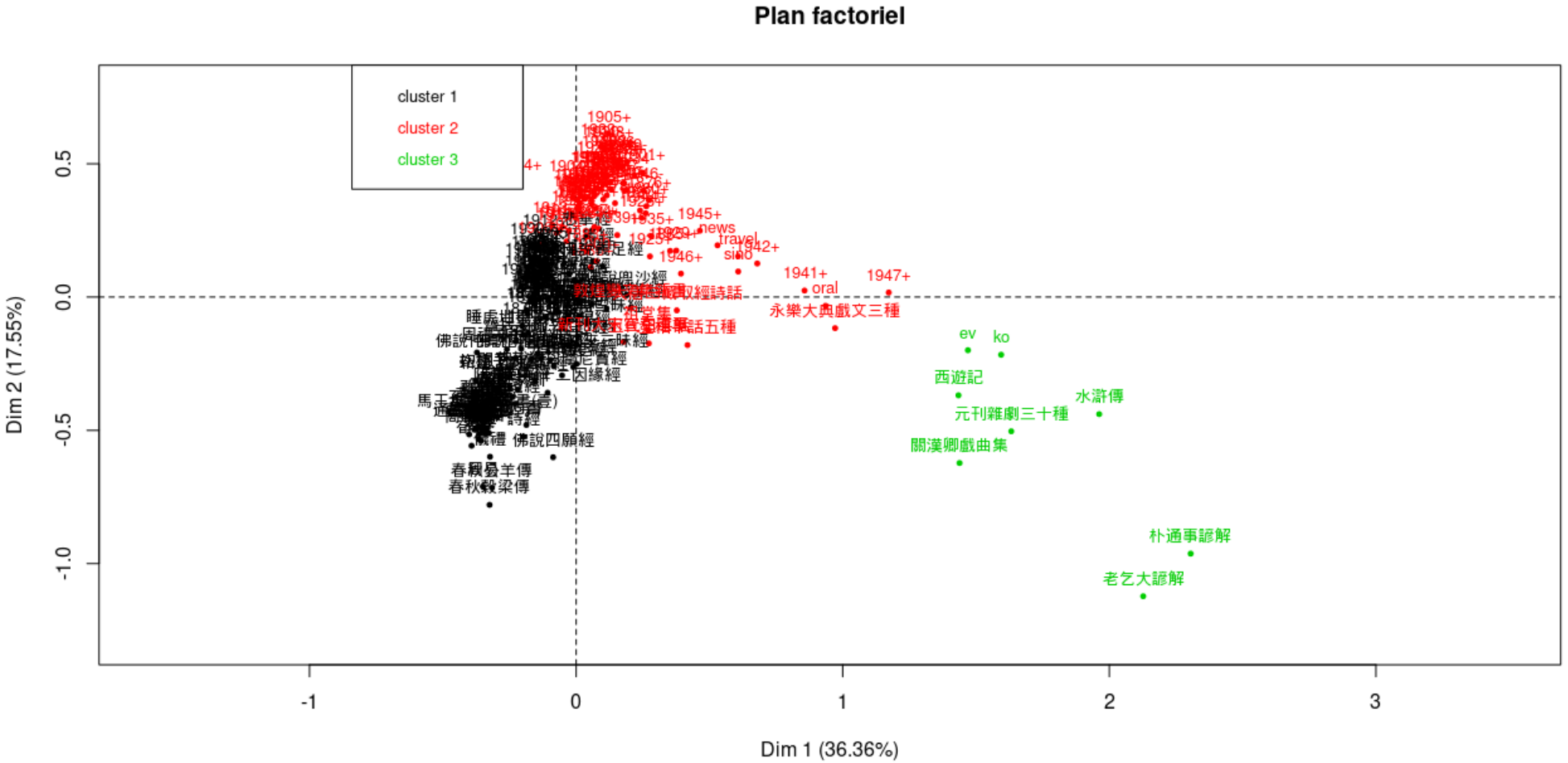

CA factor map
Dim 2 (16.84%)
Dim 1 (44.35%)
1.0
0.5
0.0
-0.5
0
1
了本生死經
悲華經
佛本行集經
佛說兜沙經
妙法蓮華經
道行般若經
大莊嚴論經
阿含口解十二因緣經
佛說般舟三昧經
佛說阿闍世王經
生經
敦煌變文集新書
祖堂集
世說新語
大唐三藏取經詩話
永樂大典戲文三種
新刊大宋宣和遺事
五代史平話
老子
詩經
西遊記
元刊雜劇三十種
水滸傳
關漢卿戲曲集
朴通事諺解
老乞大諺解


CA factor map
Dim 2 (16.84%)
Dim 1 (44.35%)
C.6
C.5
P.中古
C.2
C.9
C.4
C.10
C.3
C.5.5
C.12.5
C.11
C.15
P.近世
C.-2
C.-1
P.上古
C.16
C.-4
C.-5
C.-8
C.13.5
C.14
C.-3
C.-6
C.13
0.5
0.0
-0.5
0.0
0.5
1.0
1.5

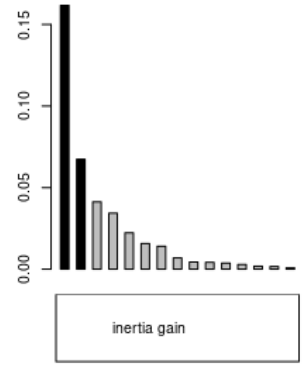

CA factor map
Dim 2 (15.90%)
Dim 1 (42.98%)
0.5
0.0
-0.5
0.0
0.5
1.0
C.6
C.5
P.中古
C.2
C.9
C.4
C.10
C.3
C.travel
C.slno
C.news
P.stb
C.oral
C.12.5
C.5.5
C.11
C.ev
C.-2
C.ko
P.上古
P.近世
C.15
C.-8
C.-1
C.-6
C.16
C.-3
C.-4
C.-5
C.13.5
C.14
C.13

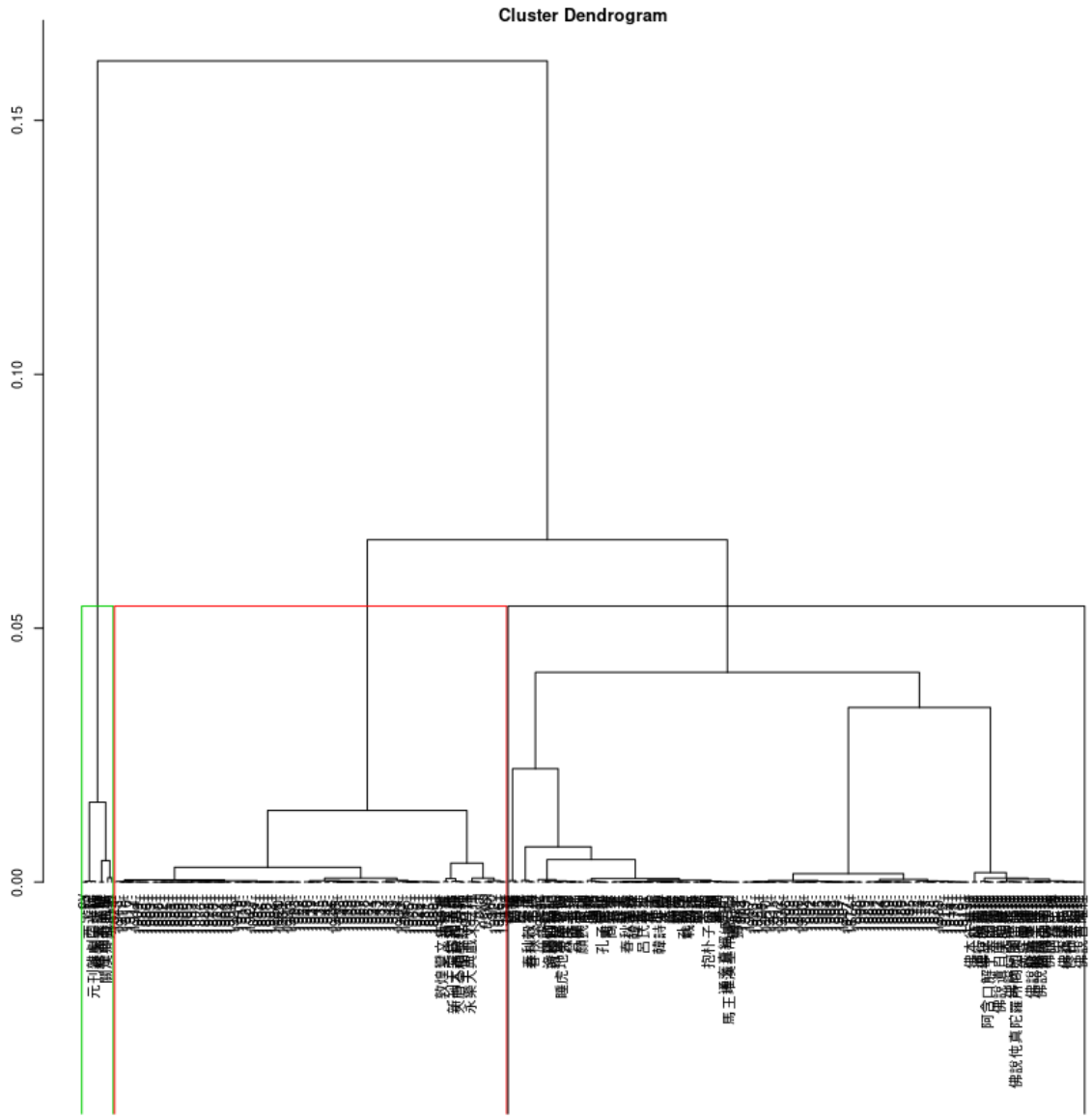

CA factor map
Dim 2 (17.55%)
Dim 1 (36.36%)
C.20
C.19+
P.sp+
C.20+
C.news
C.travel
C.slno
P.stb
P.sp-
C.6
C.5
C.19
C.9
C.oral
P.中古
C.10
C.12.5
C.11
C.4
C.3
C.2
C.ev
C.ko
C.5.5
P.近世
C.-2
C.16
C.14
C.-1
C.-5
C.15
C.-3
P.上古
C.13.5
C.-6
C.-4
C.13
C.-8
0.25
0.00
-0.25
-0.50
-0.75
-0.5
0.0
0.5
1.0
1.5
2.0

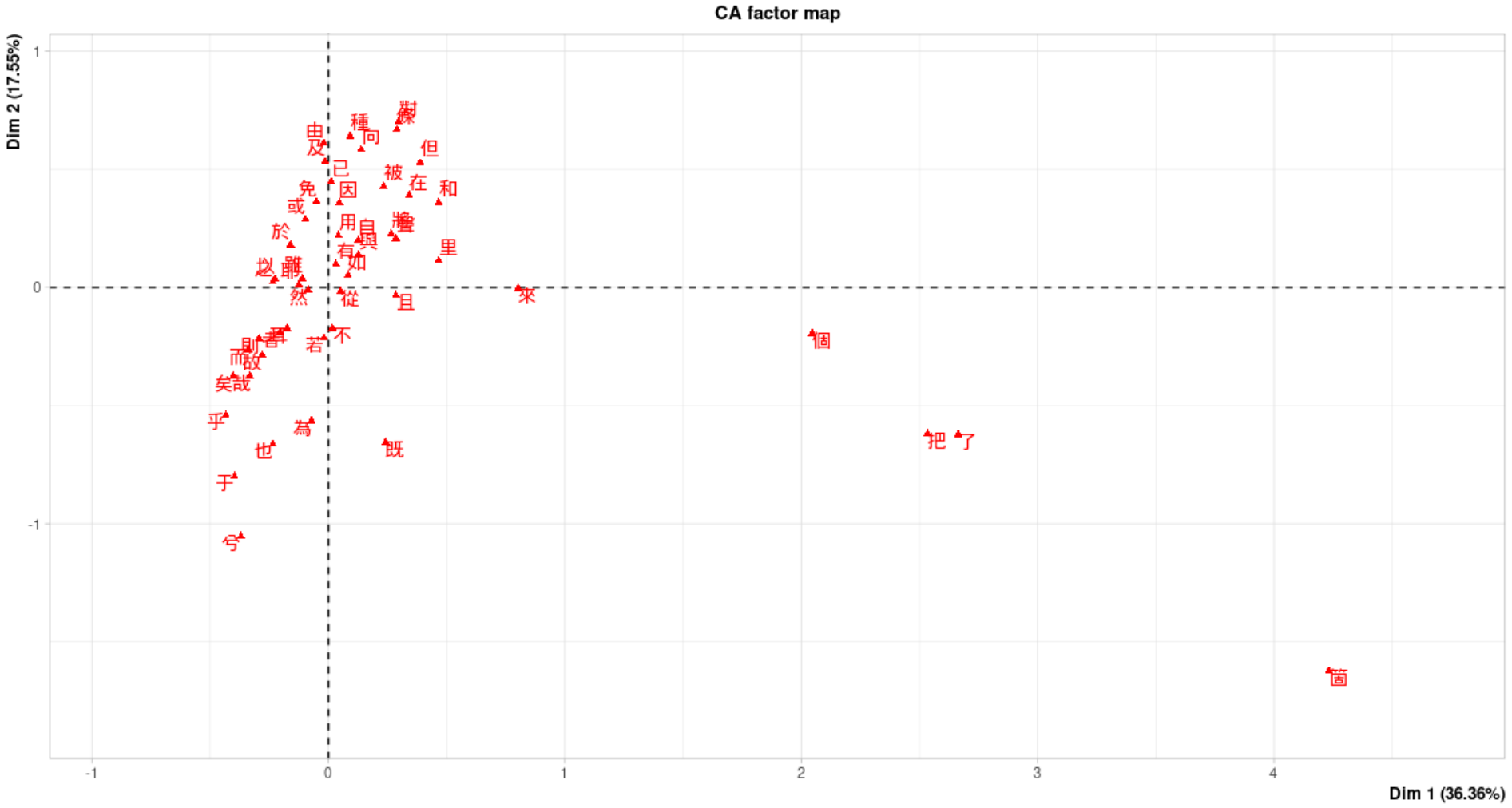

1905+
1948+
1929-
1901+
1924+
1904+
1938-
1937-
1946-
1876+
1878+
1880+
1888+
1926+
1923-
1913
1917-
1914-
1912-
1924-
1939+
1935+
1945+
1930+
1906+
1905-
1883+
news
travel
1925+
1942+
1919+
1886+
1921+
1915+
1946+
sino
1874+
1917+
1908-
1873+
1941+
1947+
1872+
1880-
oral
1875-
1912+
1874-
1873-
1872-
ev
ko
0.6
0.4
0.2
0.0
-0.2
y
0.0
0.5
1.0
1.5
x
color
a -
a +
a stb


Plan factoriel
cluster 1
cluster 2
cluster 3
1905+
1945+
news
travel
sino
1942+
1946+
1941+
1947+
oral
永樂大典戲文三種
ev
ko
西遊記
水滸傳
元刊雜劇三十種
關漢卿戲曲集
朴通事諺解
老乞大諺解
佛說四願經
春秋穀梁傳
Dim 1 (36.36%)
Dim 2 (17.55%)
-1
0
1
2
3
0.5
0.0
-0.5
-1.0

Arbre hiérarchique






CA factor map
Dim 2 (17.55%)
Dim 1 (36.36%)
1
0
-1
-1
0
1
2
3
4
由
及
種
向
但
已
被
在
和
免
或
因
於
用
自
與
里
有
如
雖
然
從
且
來
個
若
不
則
而
故
矣
哉
乎
為
也
既
把
了
于
箇
兮

	Classe 1	Classe 2	Classe 3
P=上古	7.54	-7.11	-1.37
哉	6.8	-6.55	-0.912
乎	6.71	-6.2	-1.63
矣	6.54	-6.09	-1.45
P=中古	6.32	-5.96	-1.09
而	5.87	-5.33	-1.68
也	5.82	-6.23	0.926
C=19	5.44	-5.13	-0.891
者	5.41	-5.34	-0.376
則	5.35	-4.95	-1.3
故	5.32	-4.91	-1.28

	Classe 1	Classe 2	Classe 3
了	-2.86	-0.373	8.97
個	-3.7	0.67	8.46
把	-2.28	-0.743	8.38
來	-1.65	-0.804	6.79
箇	-1.43	-0.946	6.59
P=近世	-4.44	1.07	5.24
里	-1.69	0.196	4.17
且	0.513	-1.98	4.02
聲	-1.03	-0.249	3.55
既	1.78	-2.97	3.21
將	-1.15	0.0549	3.05

Synthèse

Comparaison impossible

Les trois stratégies

- répondent à des questions différentes
- traitent des corpus différents
- ont des définition d'«observable» différentes

Synthèse

Perplexité et clustering

Objectif	Détecter des sous-périodes (des ruptures)
Corpus	Un sous-corpus par année
Observable	chaque modèle de langue

Plongements contextualisés

Objectif	Suivre une compétition entre items dans le temps
Corpus	Un sous-corpus d'une période
Observable	chaque **occurrence** d'un terme ciblé

Analyse factorielle des correspondances

Objectif	Situer des documents les uns par rapport aux autres
Corpus	ensemble de documents, enrichi de points de référence
Observable	chaque document

avantages/inconvénients

Plongements contextualisés

+ niveau occurrence très pratique pour suivre une évolution
+ robuste face aux variantes
+ découverte de n-grammes figés
- besoin d'un gros corpus
- magie noire

Analyse factorielle des correspondances

+ marche sur des petits documents (avec metadonnées)
+ moins boite-noire
+ correspondance entre documents et lexèmes
- sensible aux variantes, aveugle aux n-grammes

Merci !

Questions ? Commentaires ? Conseils ?

pierre.magistry@inalco.fr