

Analyse Textuelle avec R

Package TextoMineR

Types de corpus – Méthodes - Exemples

Séminaire RUSS-INED - Paris, 21 janvier 2016

Mónica Bécue Bertaut (monica.becue@upc.edu)

Universitat Politècnica de Catalunya

Annie Morin (annirisa@orange.fr)

Université de Rennes I

1. Introduction
2. Corpus que l'on peut analyser
3. Exemples utilisés
4. Codification
5. Méthodes statistiques dans TextomineR
6. Types de résultats au travers des exemples
7. Conclusions

1. Introduction

Le package **TextoMineR** a été conçu pour :

- offrir l'accès à des méthodes d'analyse textuelle avancées
- tout en conservant un point de vue proche de celui offert par l'analyse des correspondances (en suivant les travaux de Benzécri et Lebart), complétée par plusieurs méthodes

Traits caractéristiques

- Introduire dans l'analyse du corpus l'information complémentaire connue sur les documents (variables contextuelles).
- Intérêt porté aux corpus chronologiques
- Analyse d'un seul texte rhétorique comme, par exemple, un réquisitoire. Ce texte peut être découpé de façon automatique en « parties lexicalement homogènes »

Auteurs: Mónica Bécue Bertaut; Daria Hernández; Belchin Kostov; Josep Anton Sánchez-Espigares; Ramón Álvarez-Esteban

1. Introduction

Textual Statistics



Documentation for package 'TextoMineR' version 1.1

- [DESCRIPTION file.](#)

Help Pages

CharDocWord	Characteristic Documents and Words (CharDocWord)
dataBiblio	dataBiblio (data)
dataOpen.question	dataOpen.question (data)
dataSpeech	dataSpeech (data)
DocVarTable	Documents by Variables Table (DocVarTable)
DocWordTable	Documents by Words Table (DocWordTable)
MacroBiblio	Analysis of Bibliography (MacroBiblio)
MacroCaHcpc	Correspondence Analysis and Hierarchical Clustering (MacroCaHcpc)
MacroTxChrono	Chronological Corpus (MacroTxChrono)

Page WEB en construction: tutorial

Livre: Analyse Textuelle avec R, par Mónica Bécue-Bertaut, Annie Morin et Fionn Murtagh, éditions PUR

2. Types de corpus

La Bible

Les textes classiques

Discours politiques

Articles de presse

Entrevues non directives

Questions ouvertes

Recherche documentaire

Bibliographie d'articles scientifiques, de patentes: veille
technologique

Lettres de réclamation

Interrogation automatique de bases de données textuelles

Organisation de bases de données textuelles

Scripts de films ou de séries de télévision

Un script de film

Un réquisitoire de procureur

Exemples traités (principalement, types de résultats) :

- Questions ouvertes dans les enquêtes par questionnaire

Utilisation par différentes méthodologies de l'information contextuelle

- Base bibliographique

Corpus chronologique

- Corpus de sentences judiciaires du Tribunal Suprême Espagnol

Dater et caractériser les changements

- Réquisitoire dans un procès pour assassinat

Dévoiler la construction de ce texte et de l'argumentation

4. Codification

Enquête Aspiration internationale”

- Qu'est-ce qui est le plus important pour vous dans la vie?
- Quelles sont les autres choses très importantes pour vous
- Que pensez-vous de la culture de votre pays

Life-Fr

Culture-Fr

1009 répondants échantillon français

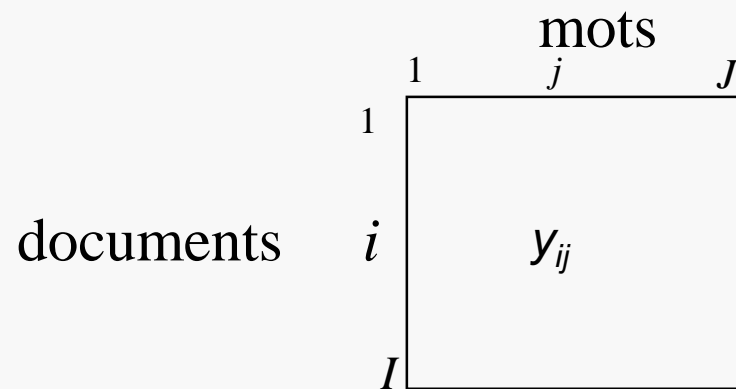
KIDEN	Sexe	Age_class	Age	Education	Niveau de Votre niv	Votre niv	Les gens s	La tranqui	On aura pl	Age-class	Sexe_Age	Sexe_edu	ageeduc	Important	Relance	Culture_pays
391	Femme	>70	71	faible	égal	Per-égal	Fut-BCPP	moins he	diminuer	moins de	plus de 55 Fr_F>55	F-Educfail	>55-Edufa	la santé.	ne pas ma ce	serait repartir à zéro sur des bases saines. éducation à l'école et en
313	Femme	55-59	59	faible	un peu m	Per-Beauc	Fut-Un pe	pareil	diminuer	moins de	plus de 55 Fr_F>55	F-Educfail	>55-Edufa	c'est de fa lire.	voya g	garder le patrimoine : les pierres, la nature. les beauxarts.
314	Femme	50-54	50	faible	un peu m	Per- plus r	Fut-peu p	moins he	sans cgt	moins de	de 31 à 55 Fr_F31-55	F-Educfail	31-55-Edu	la santé p	une bonni que	tout marche aussi bien dans l administration que dans tout. j'en ai
315	Femme	55-59	55	moyen	un peu m	Per- plus r	Fut-égal	moins he	diminuer	plus de lit	de 31 à 55 Fr_F31-55	F-Educmo	31-55-Edu	la famille,	vivre avec là	j ne sais pas. certainement qu'il faut développer la culture. je ne sa
937	Homme	40-44	42	moyen	un peu m	Per-Beauc	Fut-égal	moins he	diminuer	moins de	de 31 à 55 Fr_H31-55	H-Educmc	31-55-Edu	se sentir t	avoir de l'	elle est insuffisante par manque de temps.
277	Femme	50-54	50	faible	beaucoup	Per-Beauc	Fut-égal	moins he	diminuer	moins de	de 31 à 55 Fr_F31-55	F-Educfail	31-55-Edu	la santé.	le travail.	l'argent.
274	Homme	30-34	30	moyen	un peu m	Per- plus r	Fut-Un pe	moins he	diminuer	moins de	30ans-et-r Fr_H<=30	H-Educmc	30-Edumo	une vie ca	la vie de famille.	un bon niveau de vie.
275	Homme	65-70	67	faible	un peu m	Per-Un pe	Fut-égal	moins he	diminuer	moins de	plus de 55 Fr_H>55	H-Educfail	>55-Edufa	d'avoir du travail et	pour les jeunes il y en a besoin.	si les jeunes n'y mettent pas les mains
276	Femme	20-24	24	faible	un peu m	Per-Un pe	Fut-Bien r	pareil	diminuer	plus de lit	30ans-et-r Fr_F<=30	F-Educfail	30-Edufail	une bonni	avoir un b	musique. peinture. sculpture.
929	Homme	40-44	42	superieur	beaucoup	Per-égal	Fut-peu p	pareil	diminuer	la même l	de 31 à 55 Fr_H31-55	H-Educscu	31-55-Edu	la santé.	vivre normalement.	avoir surtout du travail avec du temps de libre si possible et
270	Femme	35-39	35	faible	égal	Per-égal	Fut-Un pe	plus heureux	la même l	de 31 à 55 Fr_F31-55	F-Educfail	31-55-Edu	de l'argen	la santé.	du travail.	
569	Femme	18-19	19	moyen	un peu m	Per-égal	Fut-peu p	moins he	diminuer	moins de	30ans-et-r Fr_F<=30	F-Educmo	30-Edumo	le weeker	les vacanc	les hautes études. la littérature.

Une ligne= un document

4. Codification

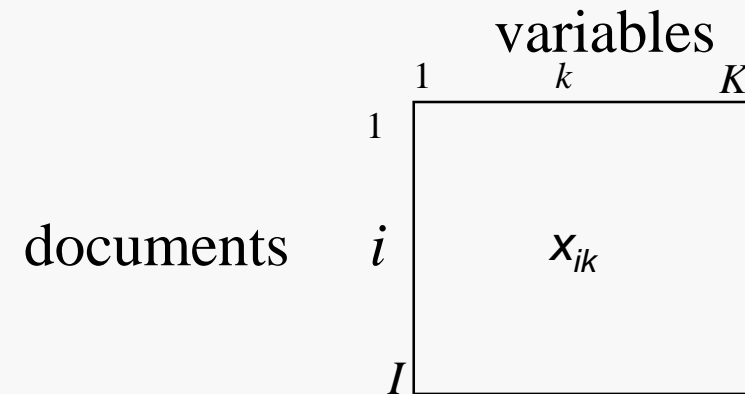
4.1 cas le plus usuel

Tableau lexical (entier)

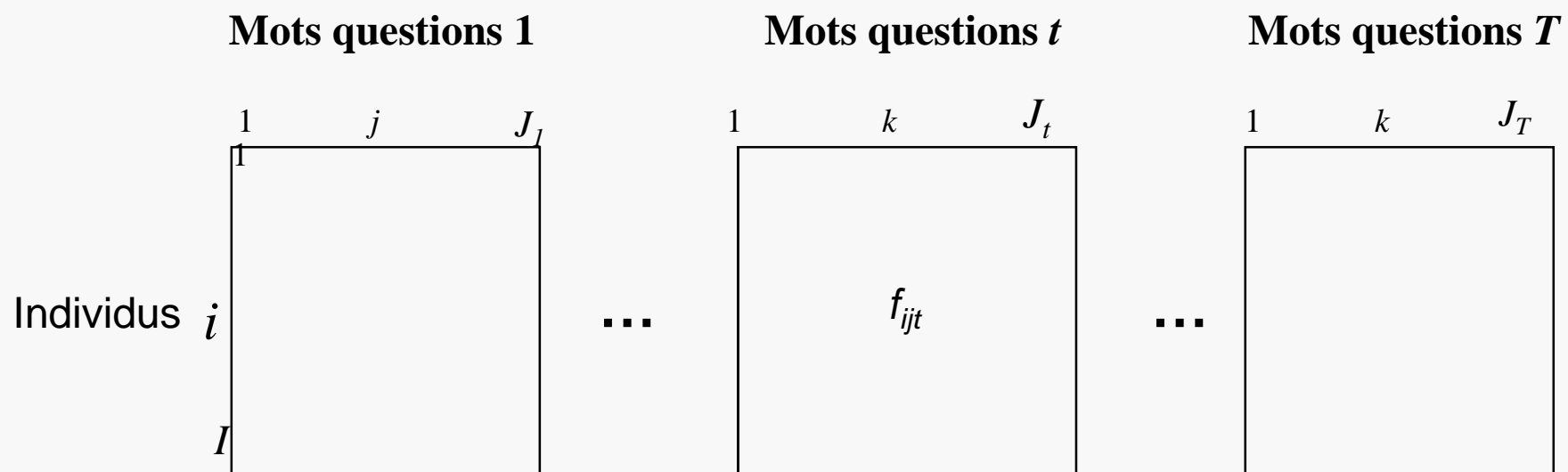


Fonction
DocWordTable

Tableau contextuel



Fonction
DocVarTable



Fonction
MxDocWordTable

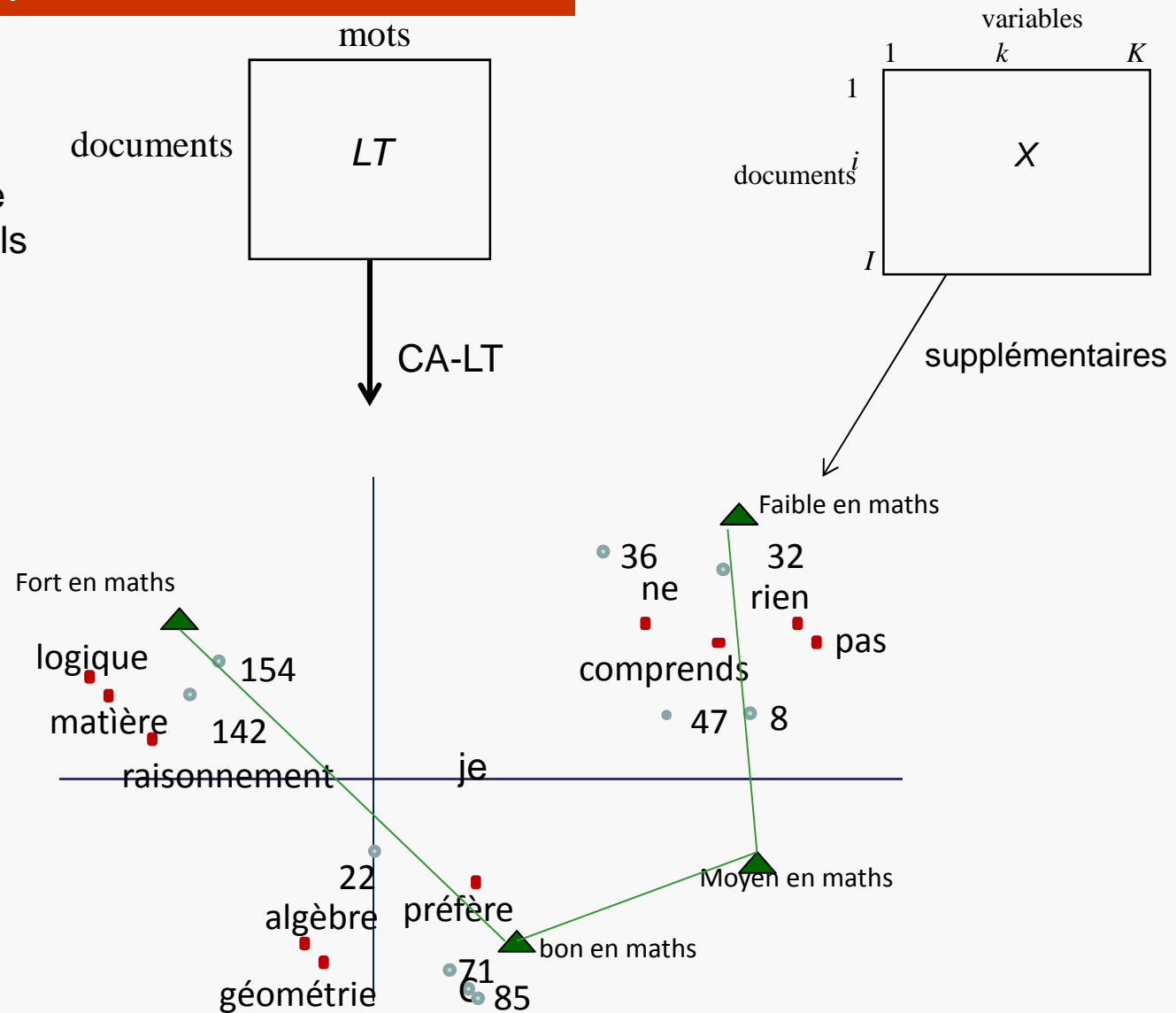
5. Principales méthodes

5.1 . CA-LT

Fonction

TxCA

Seuils de fréquence
Elimination mots-outils



Fonction

TxCA

Seuils de fréquence
Elimination mots-outils
Choix de la variable
d'agrégation

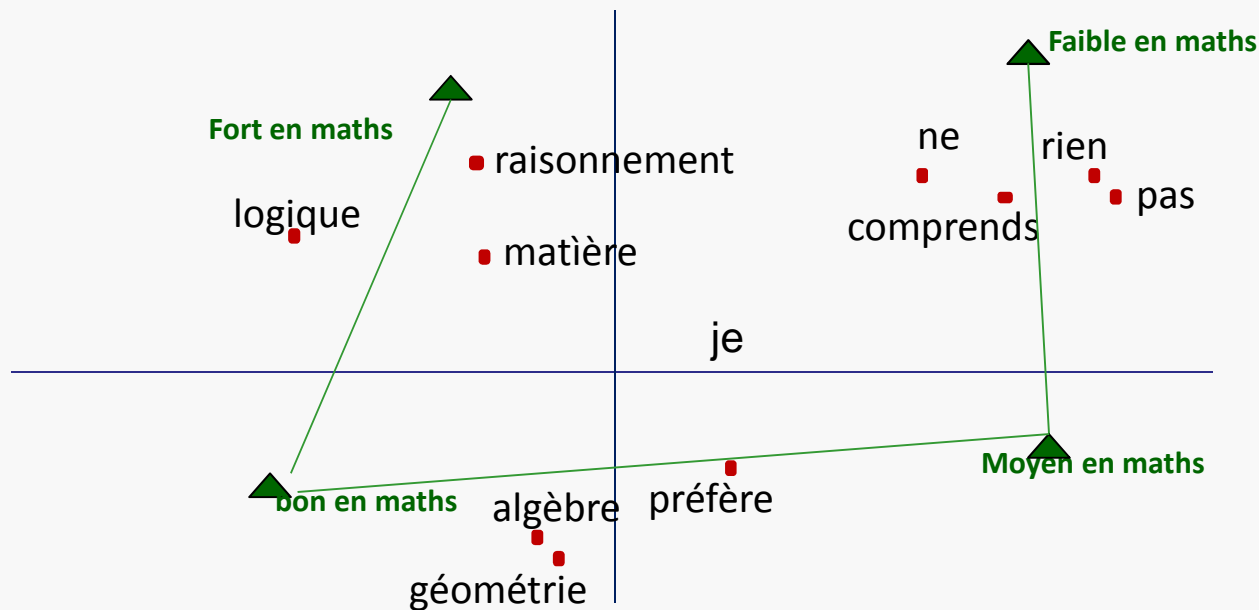
Categories

mots

ALT

étudie l'association entre les
mots et les catégories

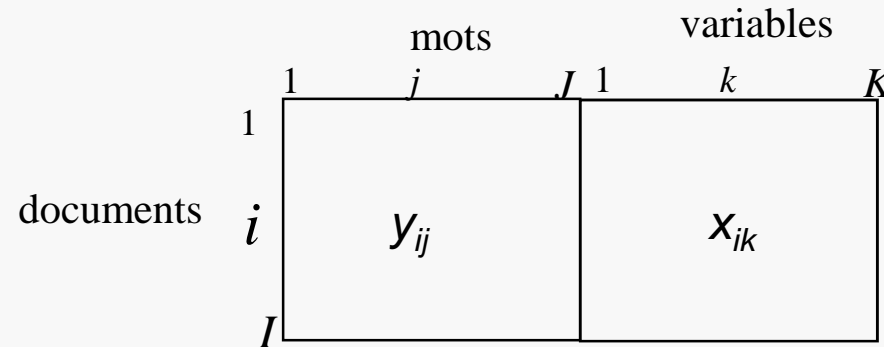
CA-ALT



Fonction

TxCaGalt

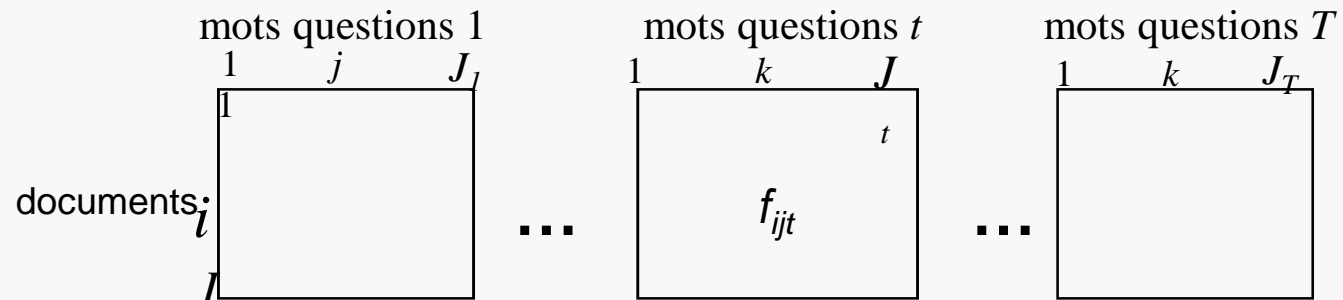
Extension de CA-ALT
à plusieurs variables,
quantitatives ou
qualitatives



Fonction

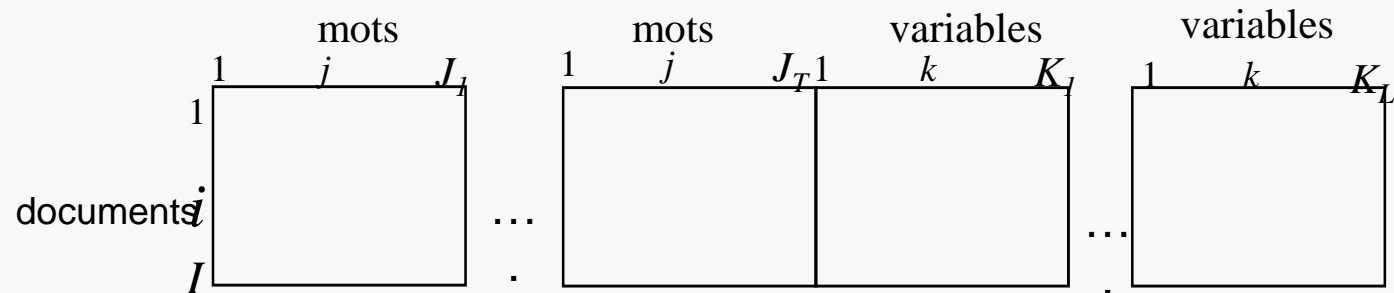
TxMFACT

Extension de CA à
plusieurs tableaux de
contingence



ou

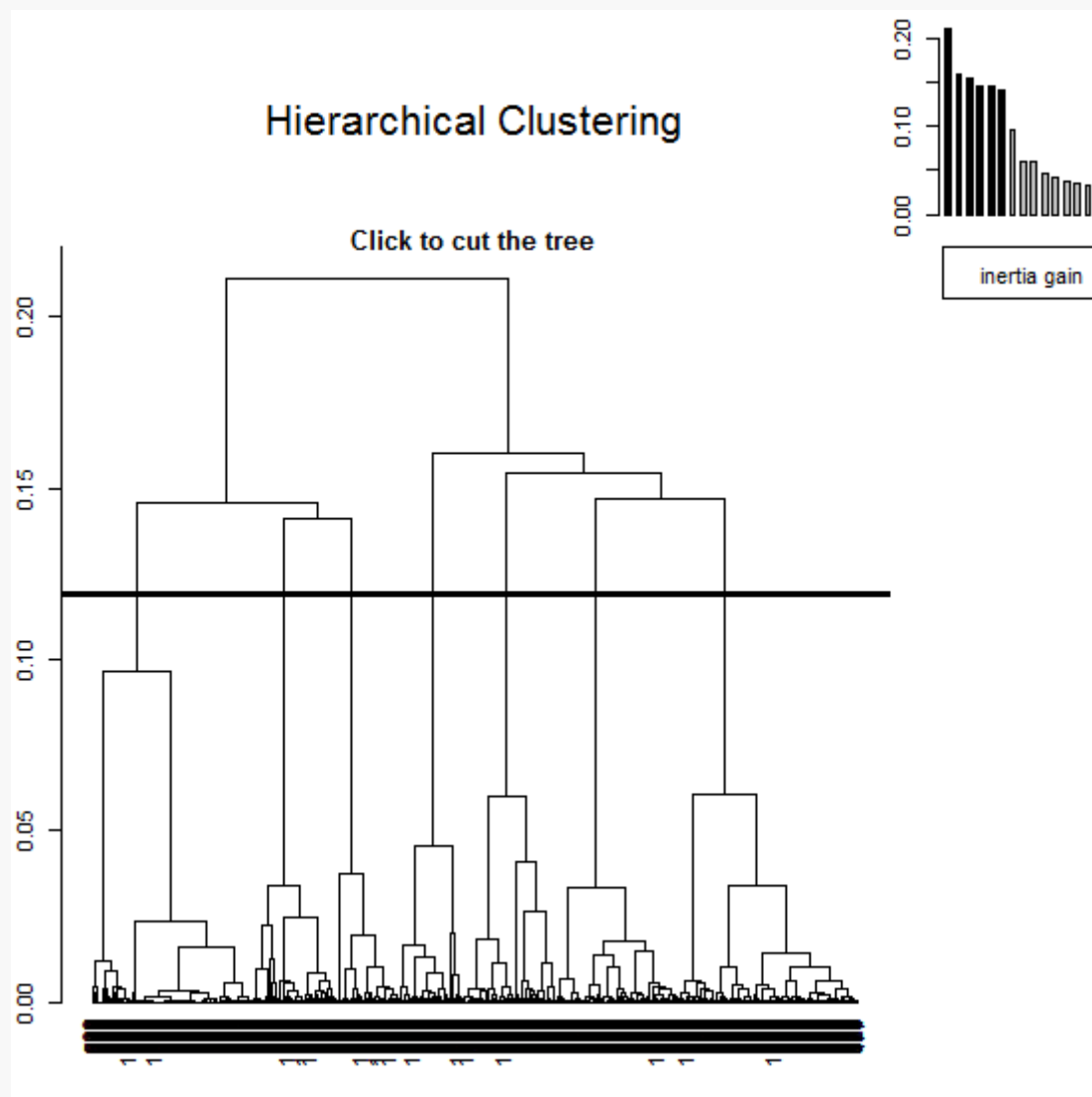
à des tableaux
hétérogènes



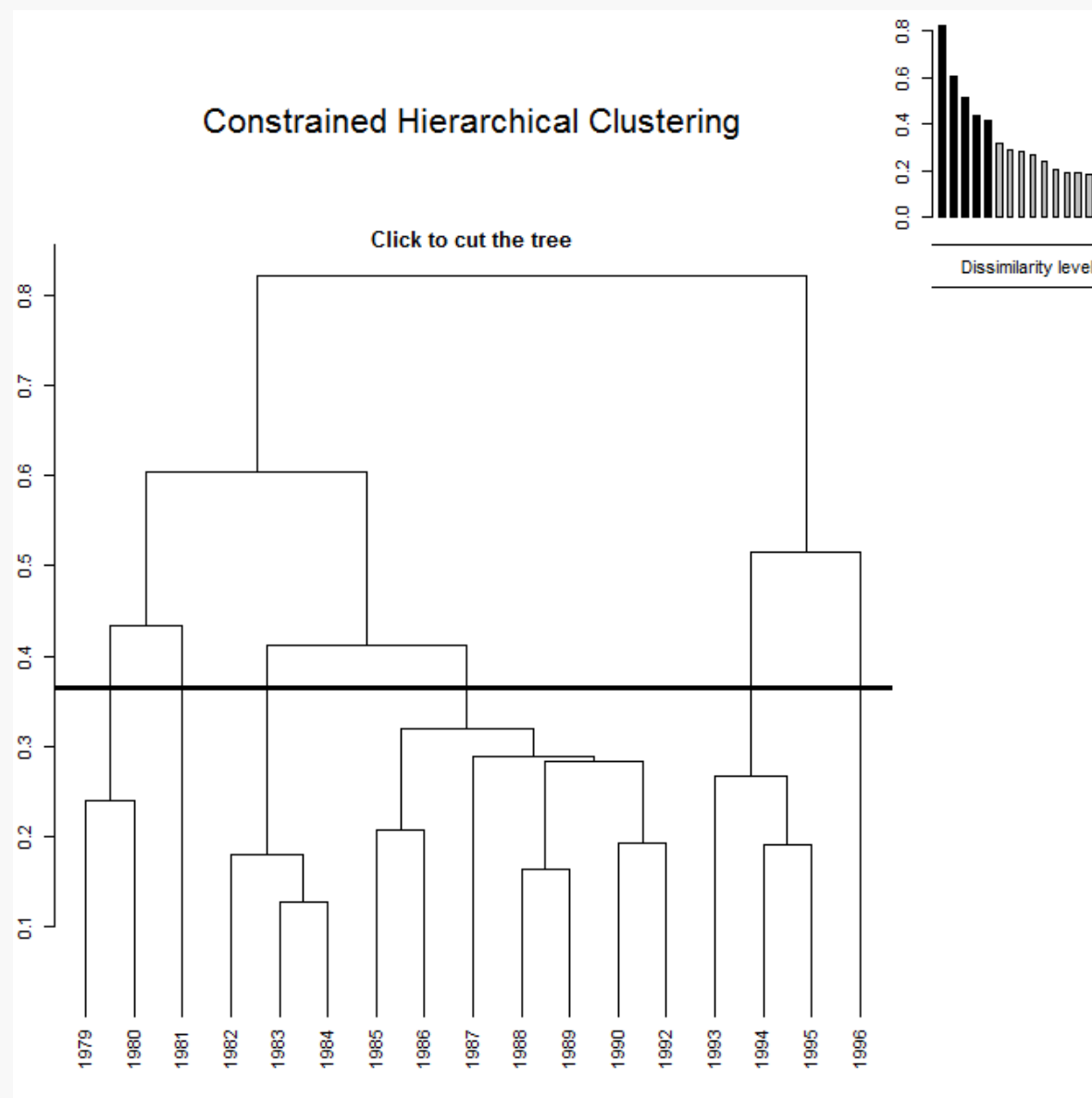
Fonction

HCPC

De FactoMineR



Fonction TxCHCPC



Fonction

CharDocWord

Catégories “femmes entre 36 et 45 ans”

Mots caractéristiques

- 1 *enfants*
- 2 *famille*
- 3 *mon*

Réponses caractéristiques

- 1 *mes enfants, ma famille.*
- 2 *mon mari, mes enfants*
- 3 *que mes enfants aient du travail*

6. Exemples

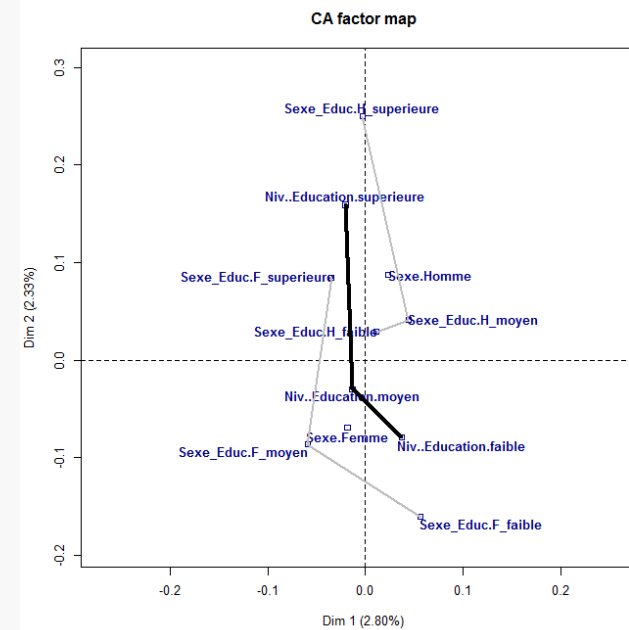
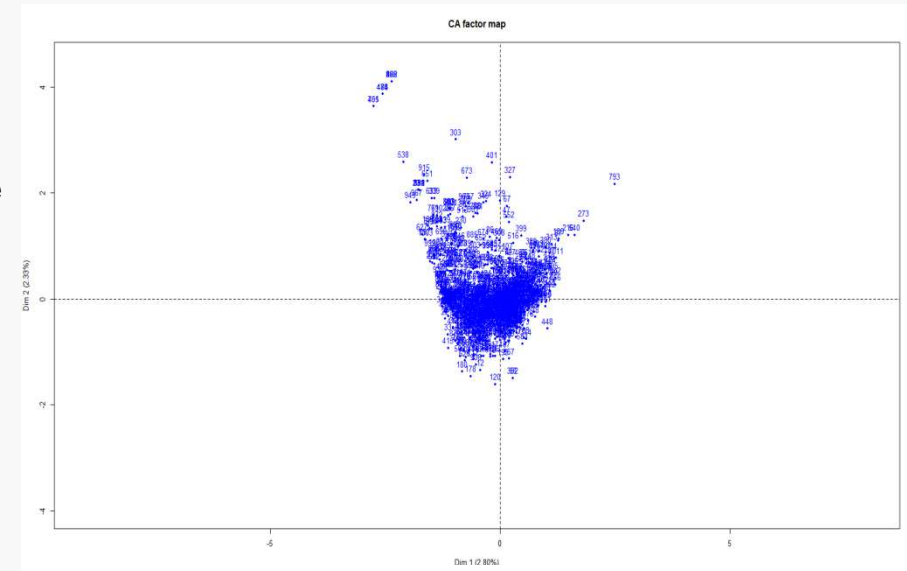
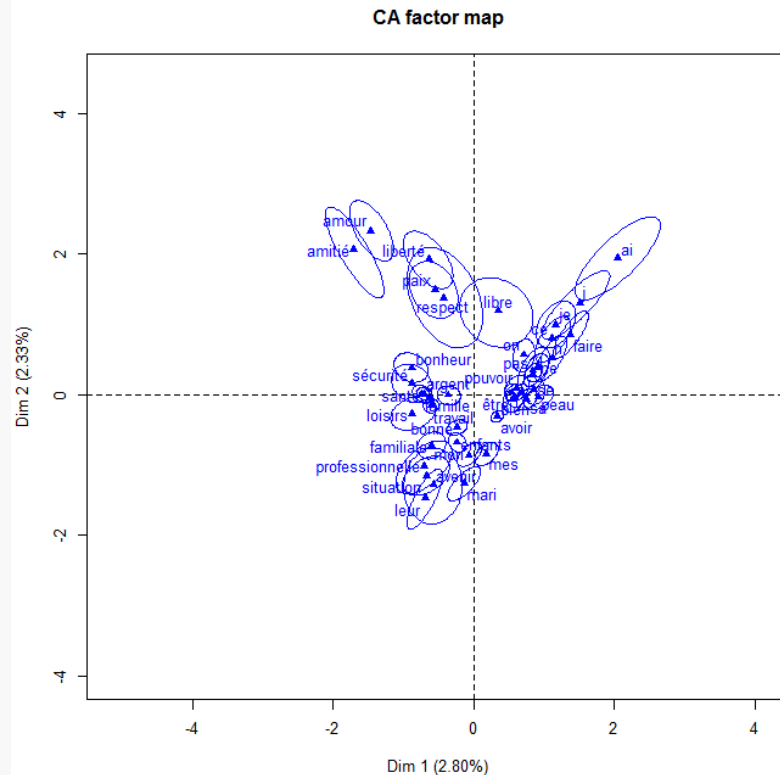
6.1 Questions ouvertes

CA_LT + Classification hiérarchique

stopmots<-

```
c("l","la","le","les","un","une","d","de","des","du","que",
"que","rien","autre","c","est","tout")
```

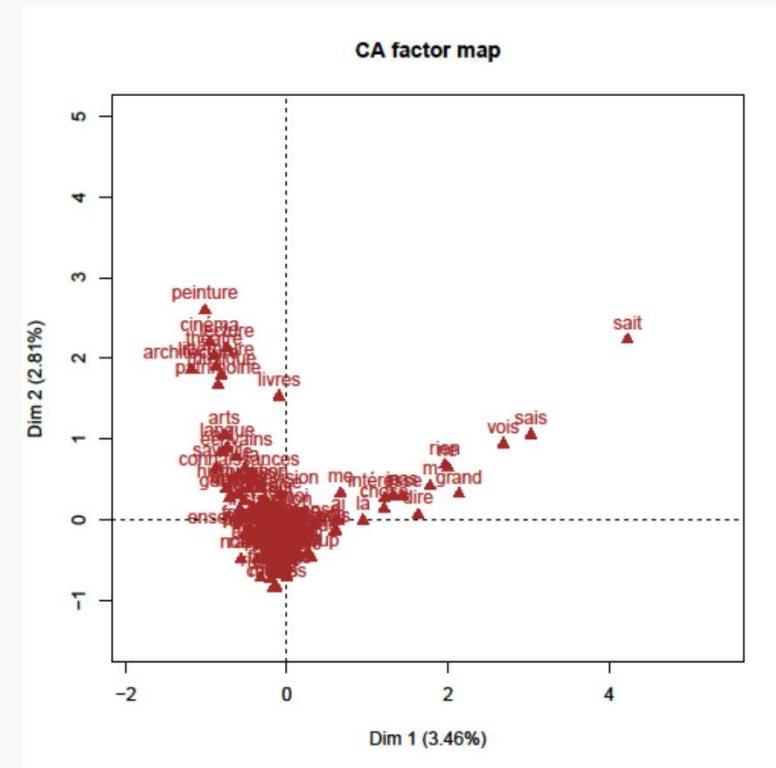
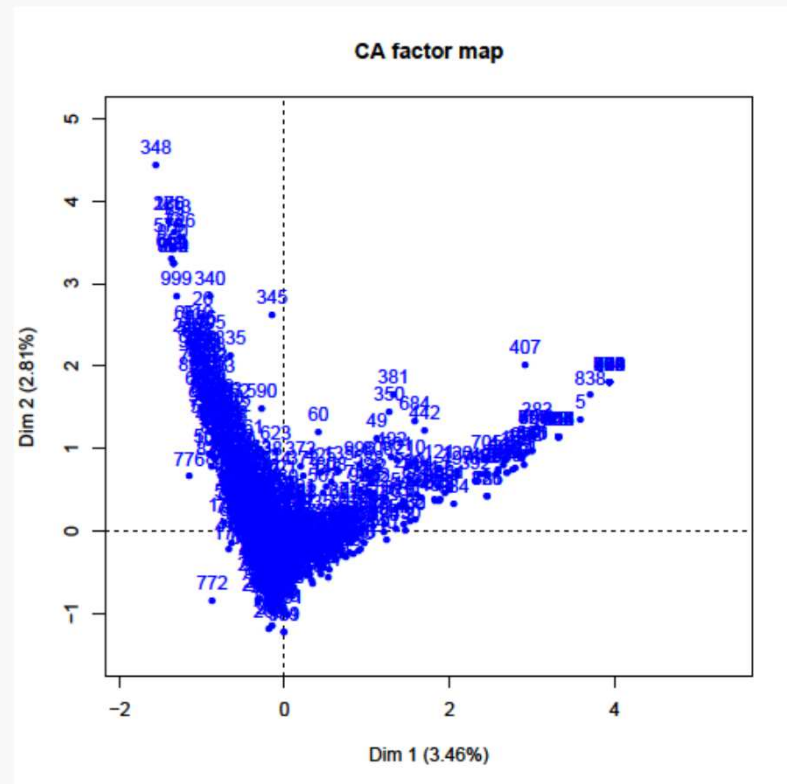
	eigenvalue	percent. of variance	cumulative percent. of variance
dim 1	0.39	2.80	2.80
dim 2	0.32	2.33	5.13
dim 3	0.31	2.28	7.41
dim 4	0.30	2.20	9.61
dim 5	0.29	2.07	11.68



6. Exemples

6.1 Questions ouvertes

Question sur la culture: on garde tous les mots-outils

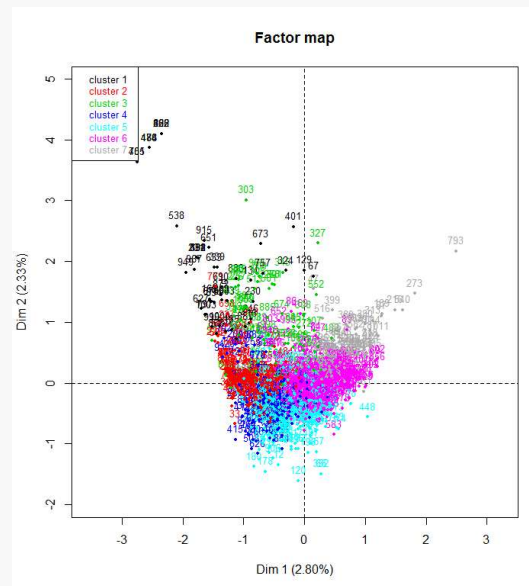
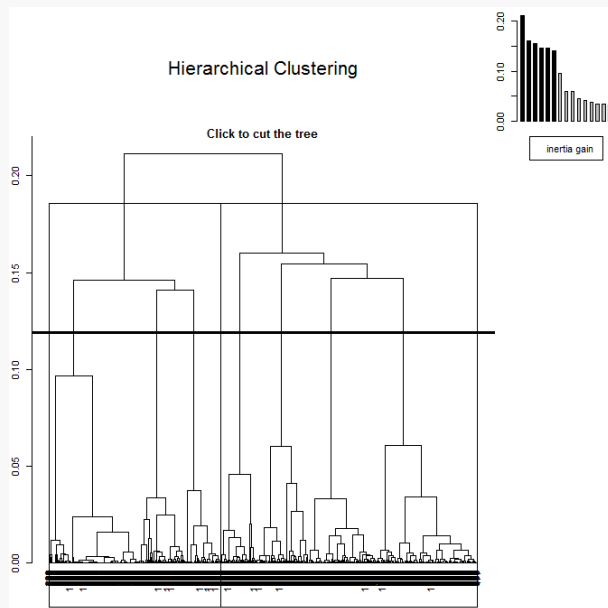


Metaclés- Metadocs

`$DIM1``$DIM1$`Metakeys+```[1] "faire" "pas" "ce"``$DIM1$`Metadocs+```[1] "793" "273" "87" "216" "131" "540" ...``$DIM1$`Metakeys-```[1] "santé" "amour" "famille" "amitié" "travail" "bonheur"``$DIM1$`Metadocs-```[1] "538" "894" "967" "184" "476" "488" "651" "893" "690" "958"
"999" "930"``[13] "633" "627" "427"``$DIM2``$DIM2$`Metakeys+```[1] "amour" "amitié" "liberté" "paix" "ai"``$DIM2$`Metadocs+```[1] "793" "184" "476" "488" "273" "651"...`

6. Exemples

6.1 Questions ouvertes



effectifs:

	1	2	3	4	5	6	7
effectifs:	50	224	67	105	187	283	87

> carac\$category\$ "7"

Important

793 faire ce que j'ai envie de faire le jour où j'ai envie de le faire.
 273 pouvoir réussir ce que j'ai envie d'entreprendre, ce que j'ai décidé d'entreprendre.

Relance

793 c'est tout le reste est secondaire.
 273 pour moi c'est un tout, changer de profession. j'ai quatre enfants.

Cla/Mod	Mod/Cla	Global	p.value	v.test
Sexe=Homme		11.134904	59.77011	46.56032
Sexe_Educ=H_superieure		14.678899	18.39080	10.86740
Sexe_Educ=F_moyen		5.426357	16.09195	25.72283
Sexe=Femme		6.529851	40.22989	53.43968

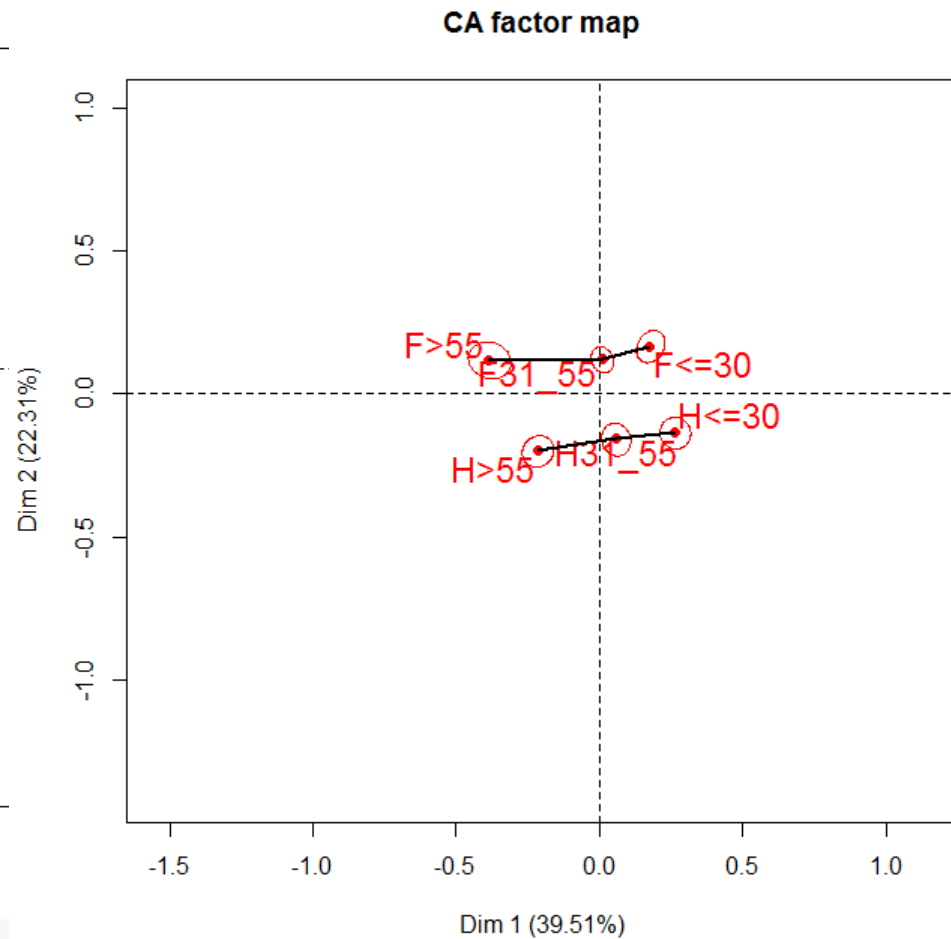
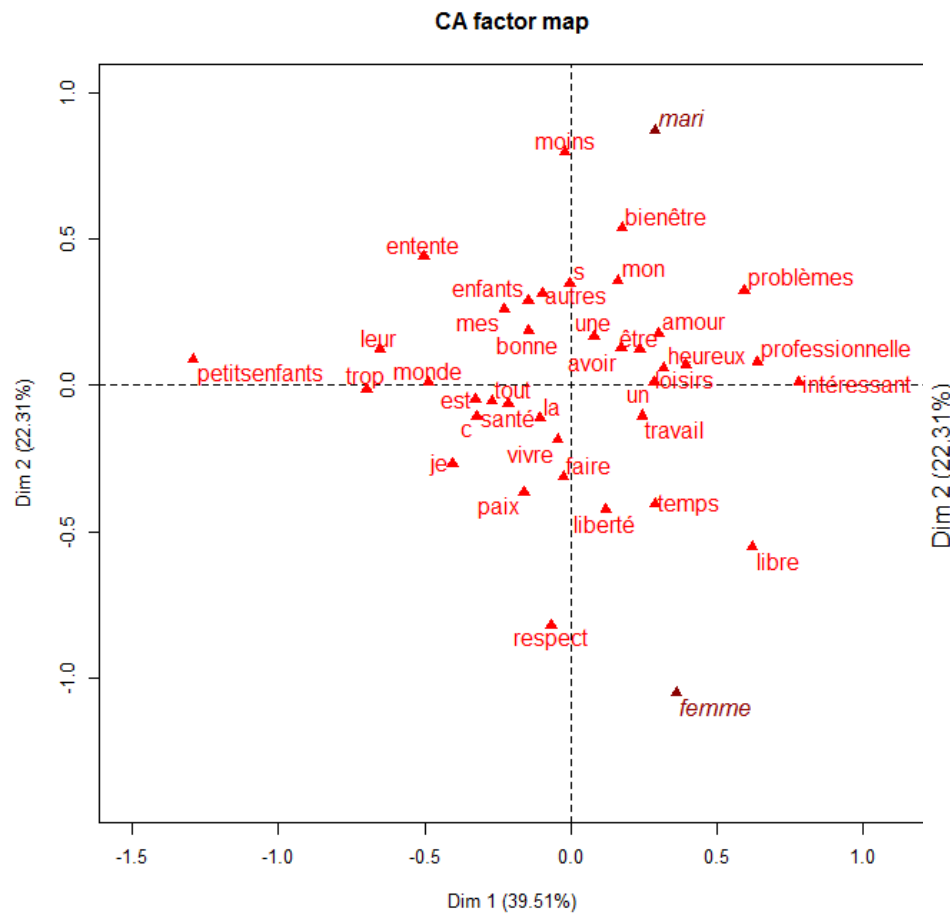
6. Exemples

6.1 Questions ouvertes

CA_ALT

Associations entre mots et catégories

Ordre des catégories conservé?



Mots et réponses caractéristiques

```
$`H<=30`$Words$Over_represented_word
```

[1] "heureux"	"libre"	"un"	"intéressant"
[5] "dans"	"travail"	"être"	"maison"
[9] "femme"	"professionnelle"	"faire"	"réussite"
[13] "ce"	"bon"	"famille"	"vie"
[17] "vivre"	"temps"	"gens"	"le"

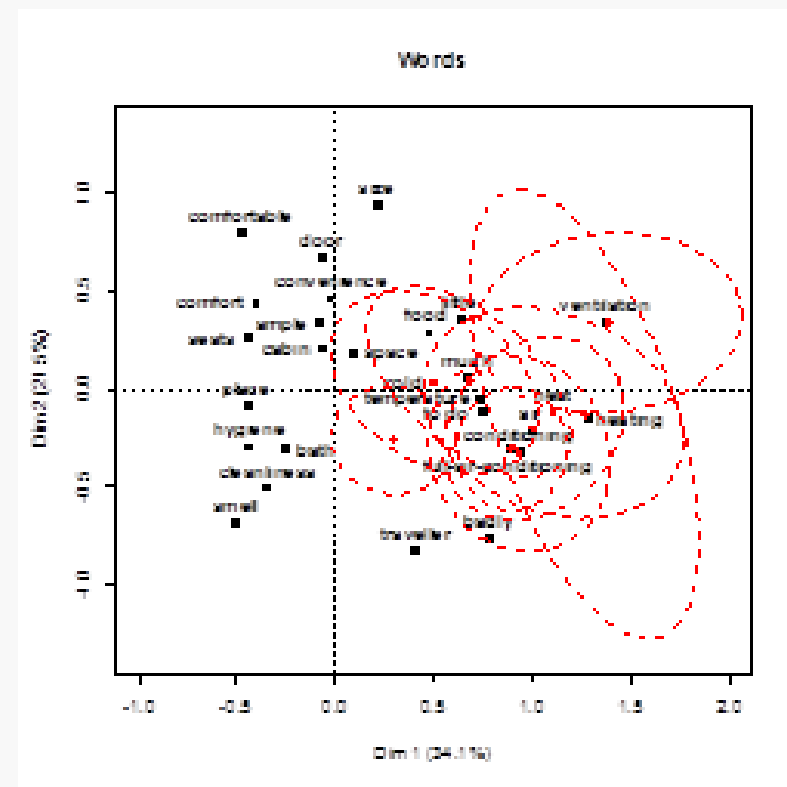
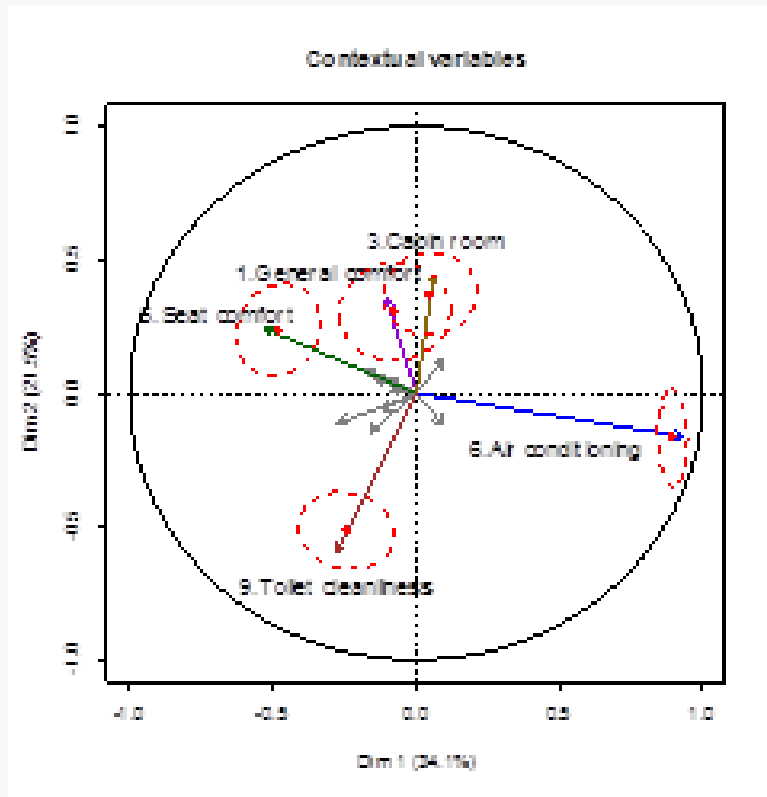
```
$`H<=30`$ Documents$Document_content
```

```
[1] "être heureux.."
[2] "l'ambition..vivre heureux."
[3] "d'être heureux..heureux dans son travail, heureux dans sa
famille, se sentir bien dans sa peau, ne pas être malade."
[4] "être heureux..avoir un bon travail. réussite professionnelle et
familiale."
[5] "être créatif..ne jamais êtes inoccupé."
```

6. Exemples

6.1 Questions ouvertes

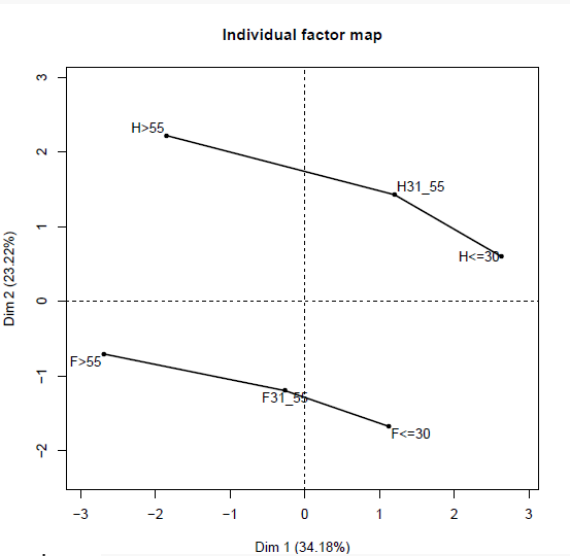
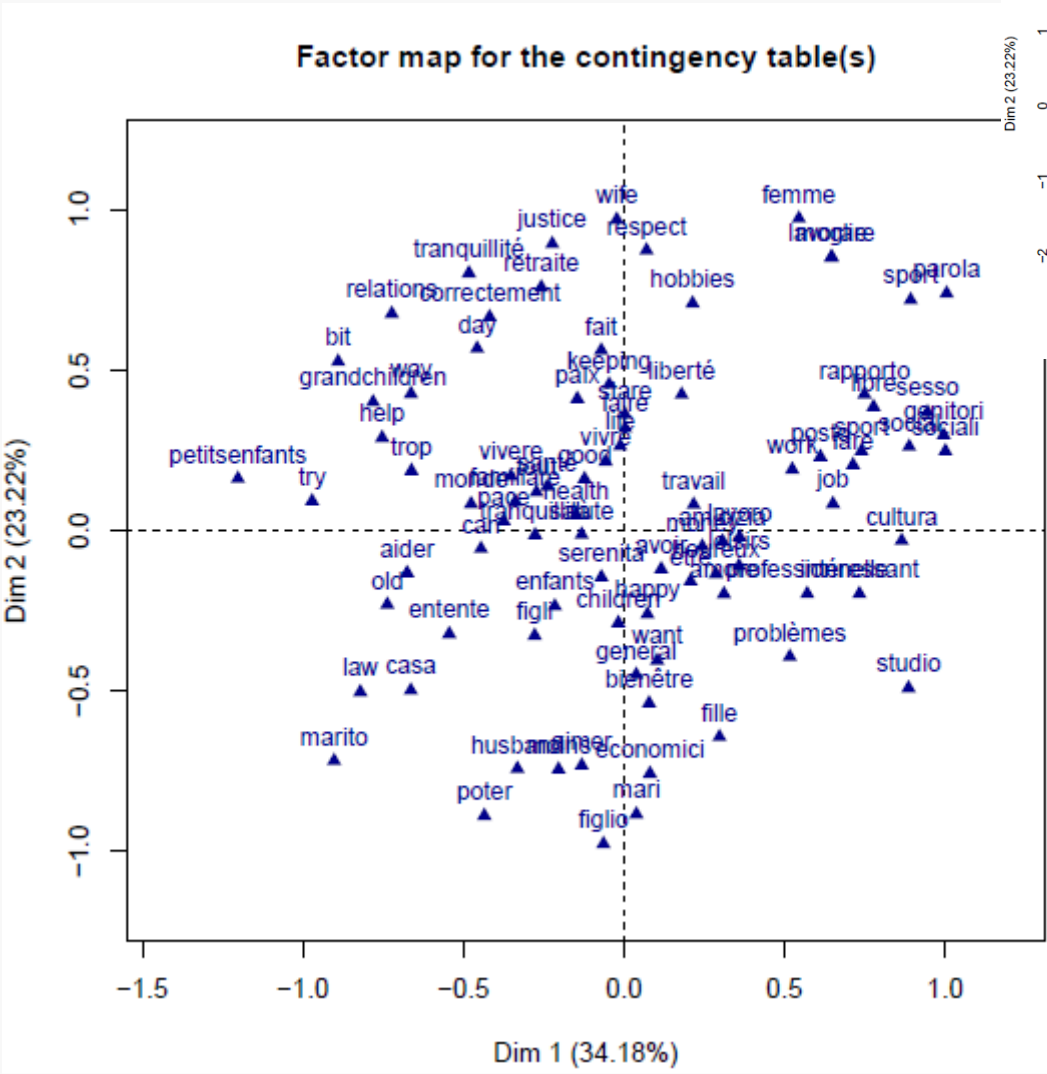
CA-GALT: analyse projetée; on peut étudier l'association entre mots et variables



6. Exemples

6.1 Questions ouvertes

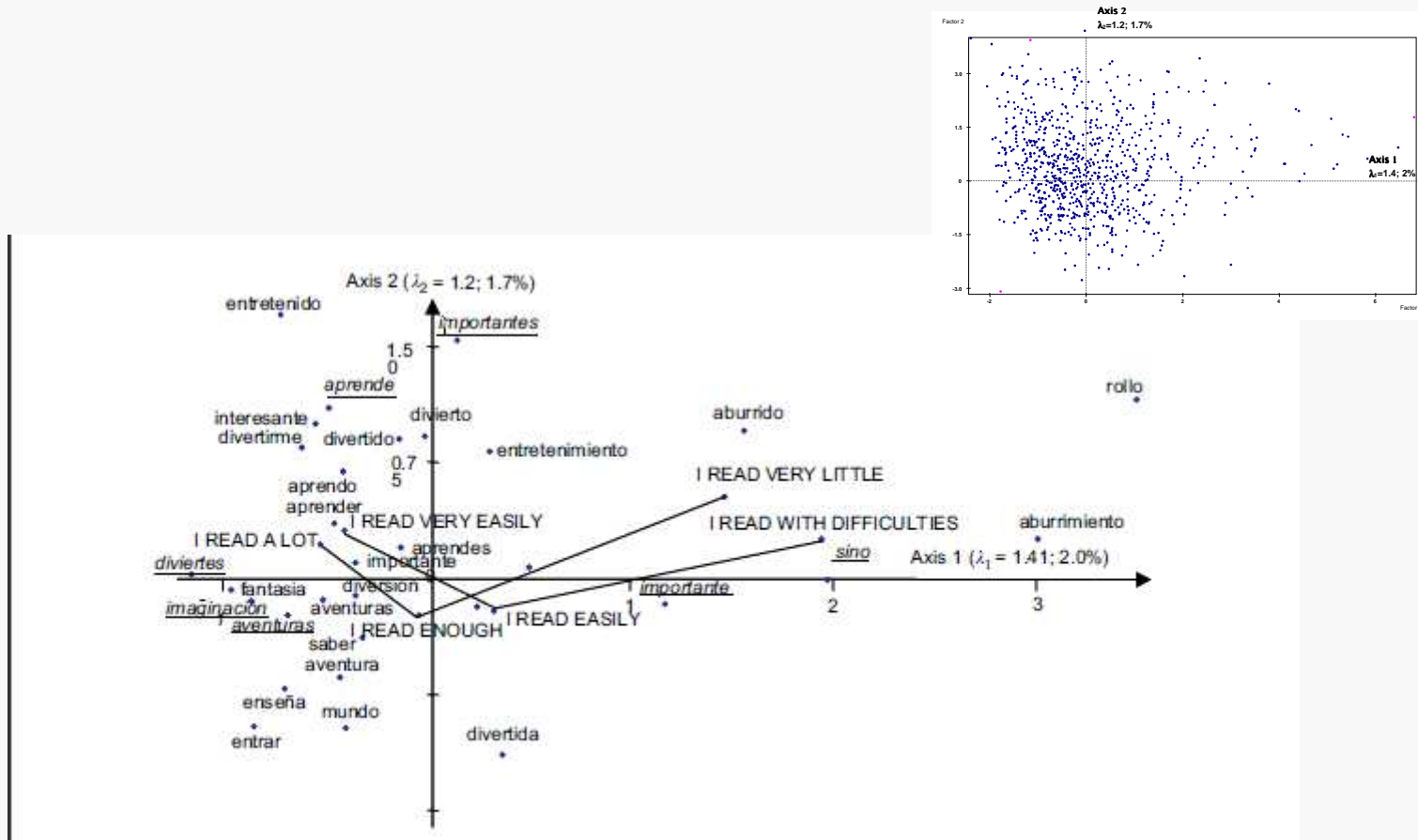
MFACT: analyse de corpus multilingues



6. Exemples

6.1 Questions ouvertes

TxMFACT: analyse simultanée de questions ouvertes et fermées



6. Examples

6.3 Corpus de sentences judiciaires

Number of documents

4595

Corpus size

484233

Vocabulary size

16918

Glossary of the 50 most frequent words

	Frequency	N.Documents
de	37293	4586
la	23603	4275
que	17106	4075
el	16401	4100
en	15162	4160
y	12133	3925
del	10599	3687
a	9417	3730
por	7711	3475
se	7004	3329
los	6582	3138
artículo	4765	2638
no	4714	2797
...		

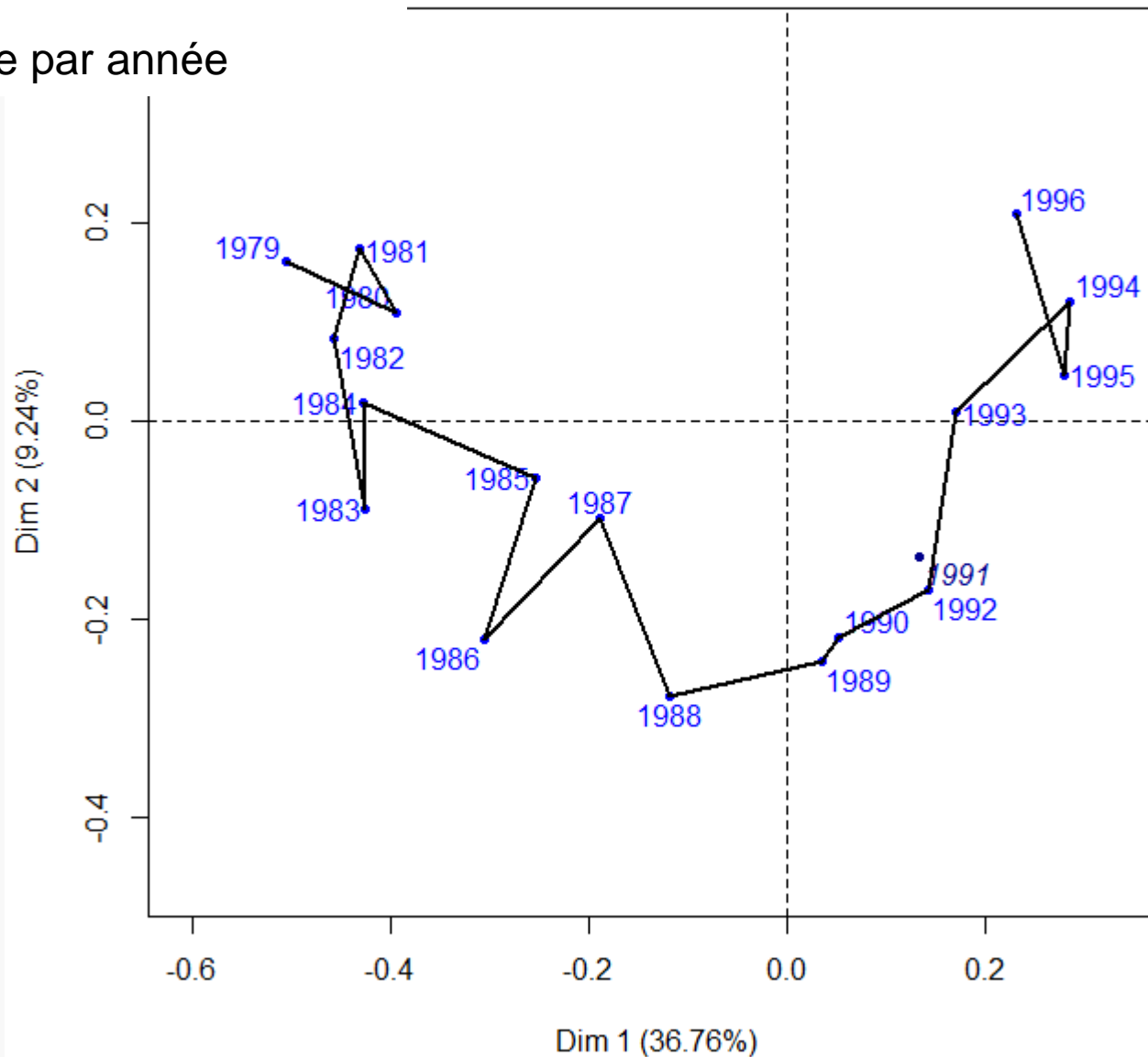
6. Exemples

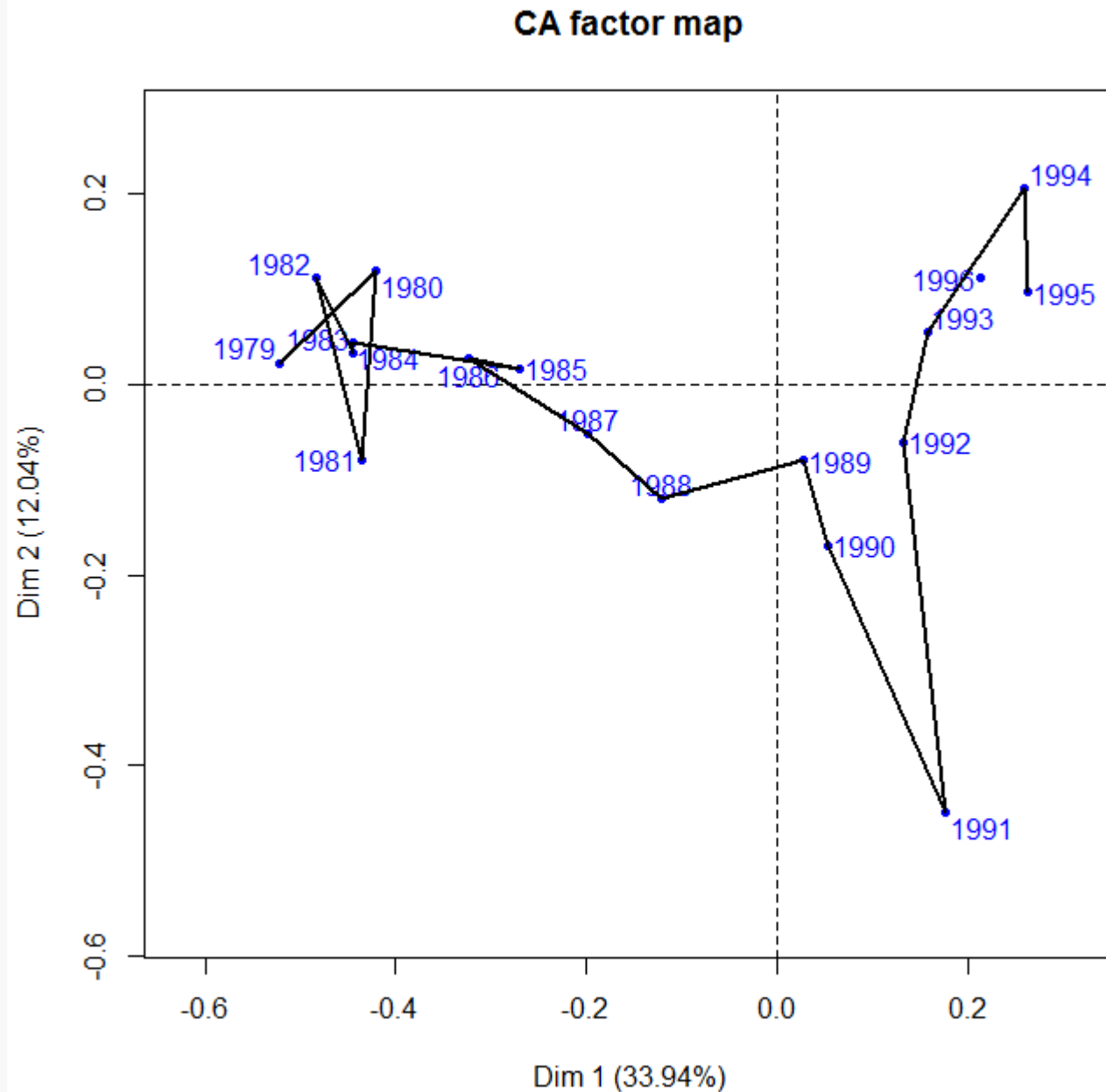
6.3 Corpus de sentences judiciaires

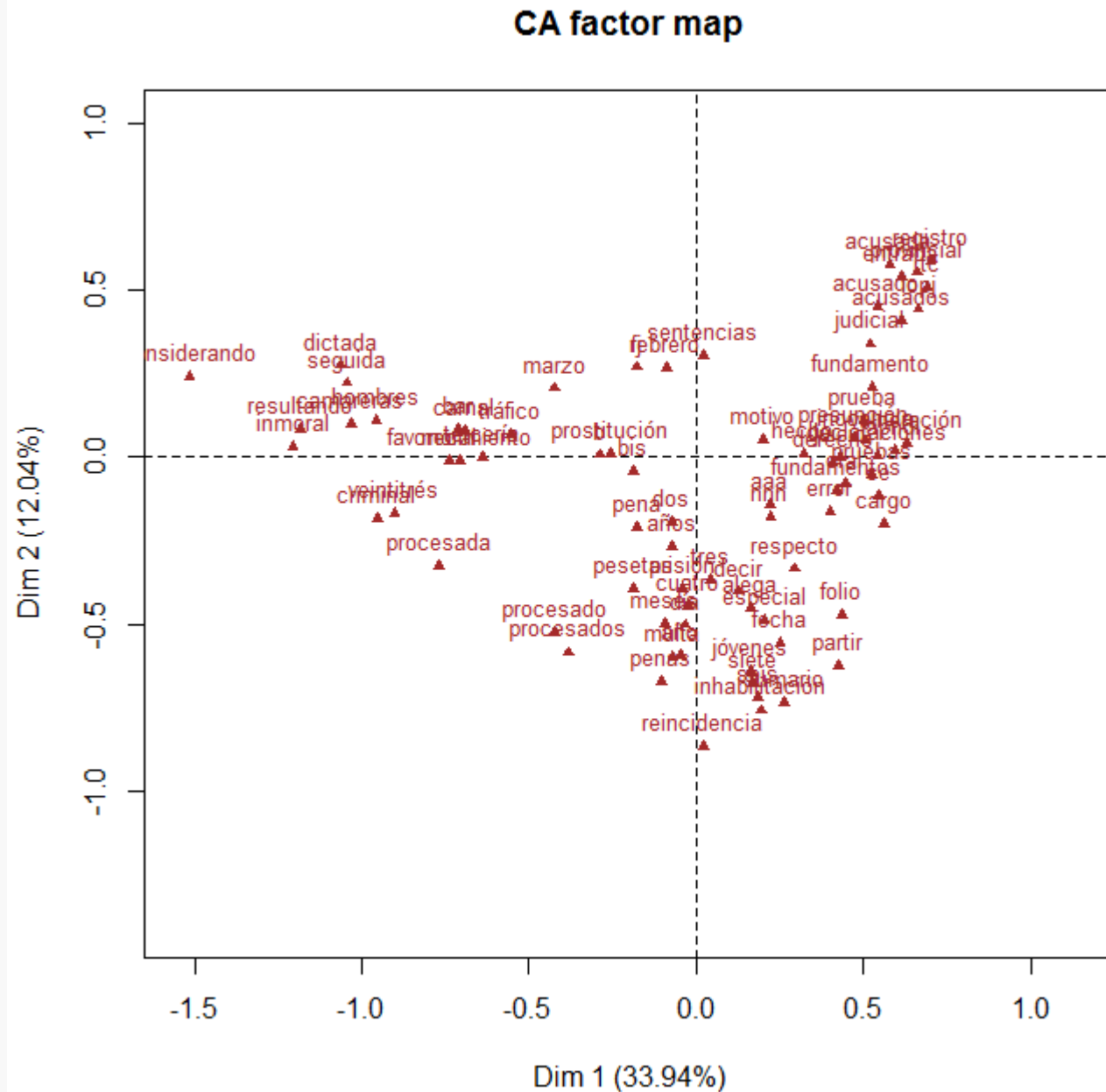
Elimination des mots outils
(liste de tm): 783 mots gardés

CA agrégée par année

CA factor map

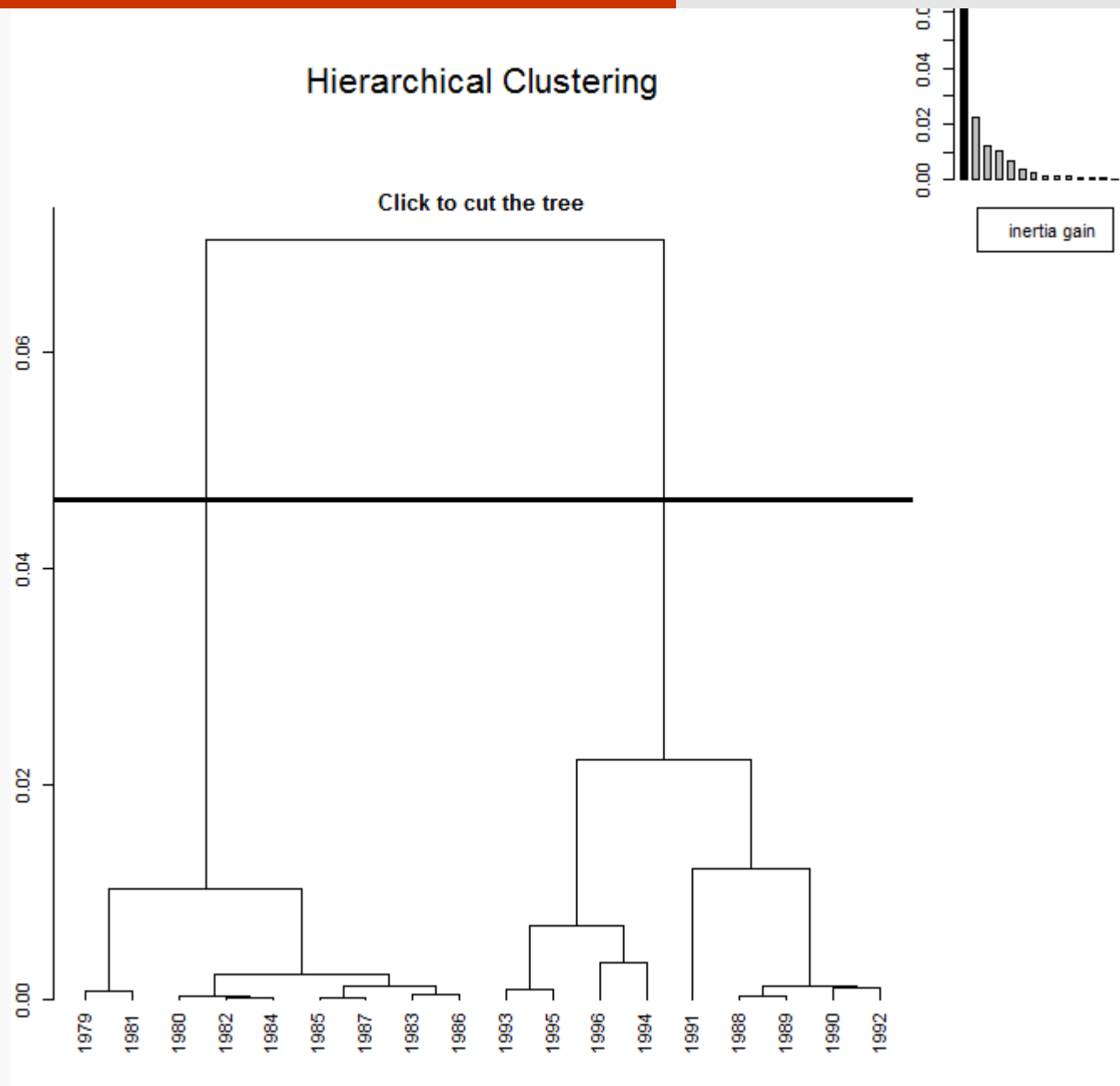




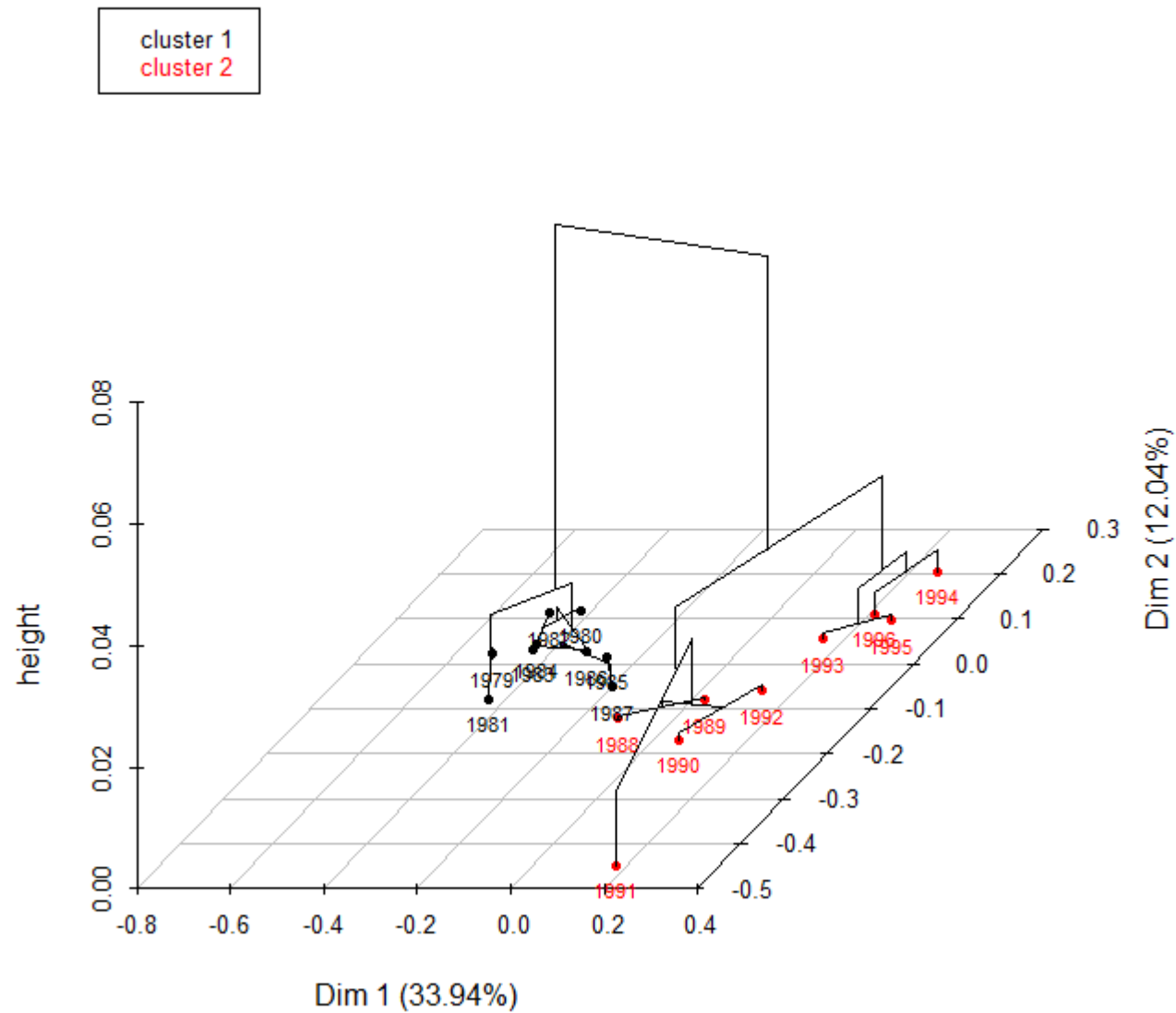


6. Exemples

6.3 Corpus de sentences judiciaires



Hierarchical clustering on the factor map



6. Exemples

6.3 Corpus de sentences judiciaires

Avant 1988: mots sur-représentés

Substantifs: camareras, prostitución, considerandos, hombres, bar, procesada, tercería, procesado, tráfico, favorecimiento, decreto, auxilio, precepto, número, corrupción, precio, modalidades, escándalo, texto, mujeres, entrega, resumen, ponente, clientes, señor, jurisdicción, prostitutas, cooperación, casación, consistencia, omisiones, facilitación, inciso, lacra, resolución, reservado, proxenetismo, huella, delito, improcedencia, corriente, estrago, sujetos, causa, reservado, camarera, lucro, mérito, índole, urgencia, empleo, cantidades, pudor, peligrosidad, matrimonio, comercio, consumiciones, trato, procesados, origen, grado, yacimiento, pisos, moralidad, apartados, vicio, mitad

Adjetifs: carnal, seguida, criminal, dictada, inmoral, interpuesto, siendo, moral, mentado, legal, empleadas, refundido, marginal, excelentísimo, impugnada, señalada, explotado, relativos, carnales, fáctica, citado, locativa, lícita, impúdicas, delictiva, ética, incardinada, ajenas, perniciosos, venal, mismos, encaminadas, activos, colectiva, organizada, buenas, prostituidas, general, inferiores, viciosa, social, defensiva, acuartelada, terminante

Verbes: considerando, resultando, cohabitar, desestimar, comprende, careciendo, revisado, tipifica, solicitaban, explotar, declara, facilitar, quedan, anula

A partir de 1988: mots sur-représentés

Substantifs: prueba, derecho, declaraciones, inocencia, acusado, presunción, juicio, vulneración, fundamentos, pruebas, cargo, motivo, error, hecho, acusados, registro, fundamento, documentos, valoración, constitución, principio, folio, instrucción, libertad, folios, apreciación, defensa, acta, violación, vía, víctima, fiscal, juzgado, credibilidad, testigos, entrada, denuncia, apoyo, ministerio, testimonio, convicción, testigo, juez, agresión, detención, intermediación, acusada, garantías, existencia, informe, autor, manifestaciones, declaración, documento, proceso, respuesta, base, jurisprudencia, policía, diligencias, relación, inadmisión, intervención, contradicción, secretario, experiencia, letrado, pub, igualdad, impugnación, violencia, atestado, validez, sección, médico, inhabilitación, asistencia

Adjectifs: oral, probatoria, judicial, constitucional, fundamental, casacional, desestimado, documental, prestadas, plenario, provincial, efectiva, procesal, acusatorio, correlativo, especial,

Verbes: ha, véanse, confrontar, dijo, existe, debe, aduce, dice, podido, cabe

Mots chronologiques: de nombreux mots caractérisent
1979-1985; 1979-1987; 1979-1988; 1985-1996; 1986-1996; 1987-1996

	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
C.P.																		
G.S.																		
H.P.																		
G.L.C.																		
L.B.																		
R.L.																		
C.M.P.																		
D.P.																		
G.M.																		
H.A.L.																		
V.M.																		
M.M.																		
D.V.R.																		
Mo.M.																		
S.N.																		
G.A.																		
R.V.																		
B.Z.																		
M.F.C.																		
M.P.																		
P.L.																		
M.P.R.																		
C.P.F.																		

Corpus composé d'en seul texte: un réquisitoire de procureur dans un procès pour assassinat. Réquisitoire prononcé à la fin d'un procès pour assassinat à l' «Audience de Barcelone » (en espagnol). Durée 1h15

Objectif: établir la stratégie discursive du procureur

- Texte de **N**=10400 occurrences
- **V**=1800 mots distincts.
- **V'**=302 words répétés au moins 5 fois **N'**=8031 occurrences sont conservées
- Segmentation initiale en 191 phrases

a. Segmentation du réquisitoire en parties lexicalement homogènes

- Le réquisitoire est considéré comme une série multi-mots

b. Codage du réquisitoire en un tableau

- Identification des discontinuités dans le discours

- Nous avons recours à une classification chronologique (CC)

Distance:

- La distance entre les phrases: distance du chi-2
- Distance entre groupes de phrases: maximum linkage

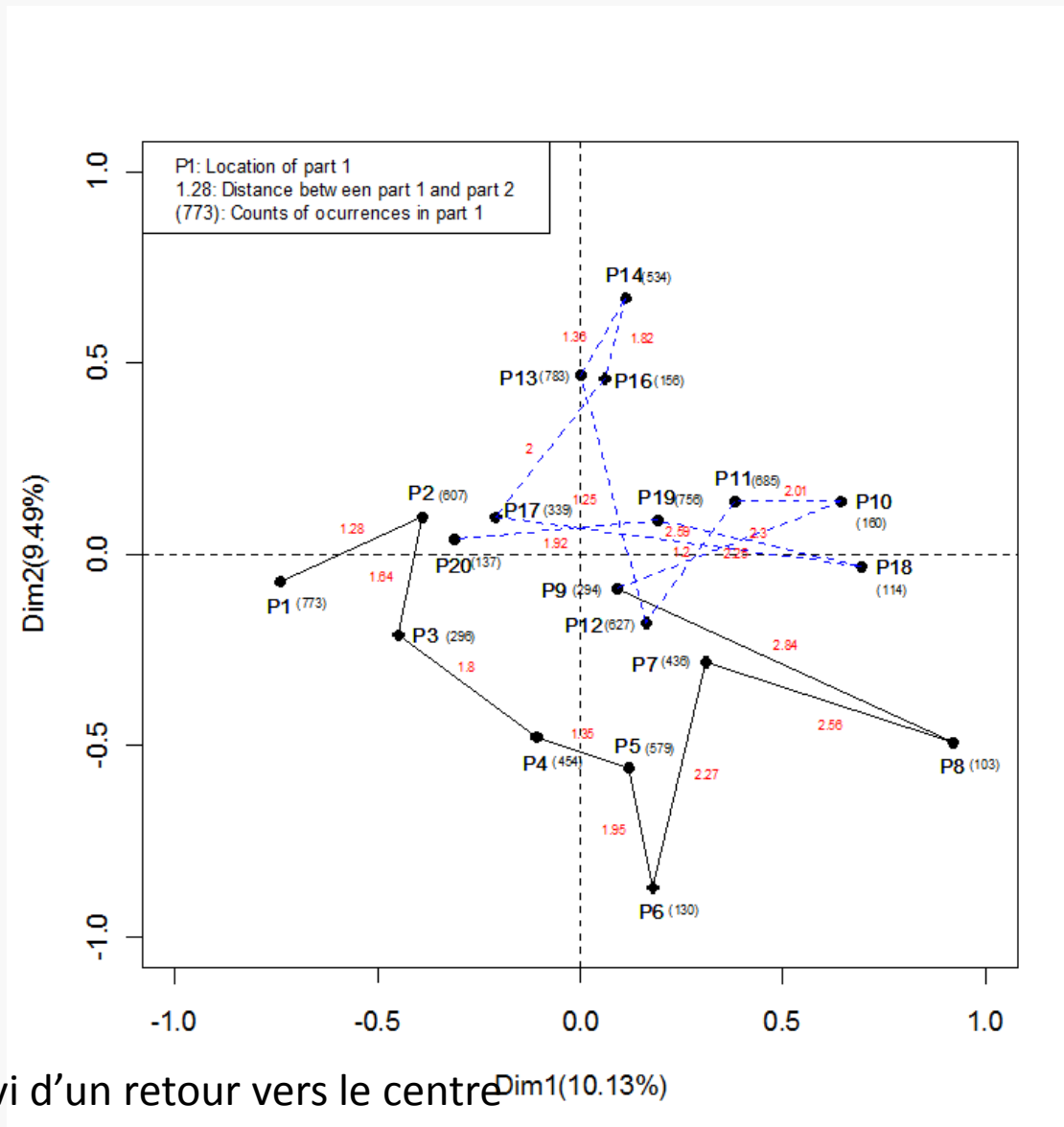
Test avant la fusion de deux groupes de phrases

- $\alpha = 0,1$ est choisi menant à 20 parties

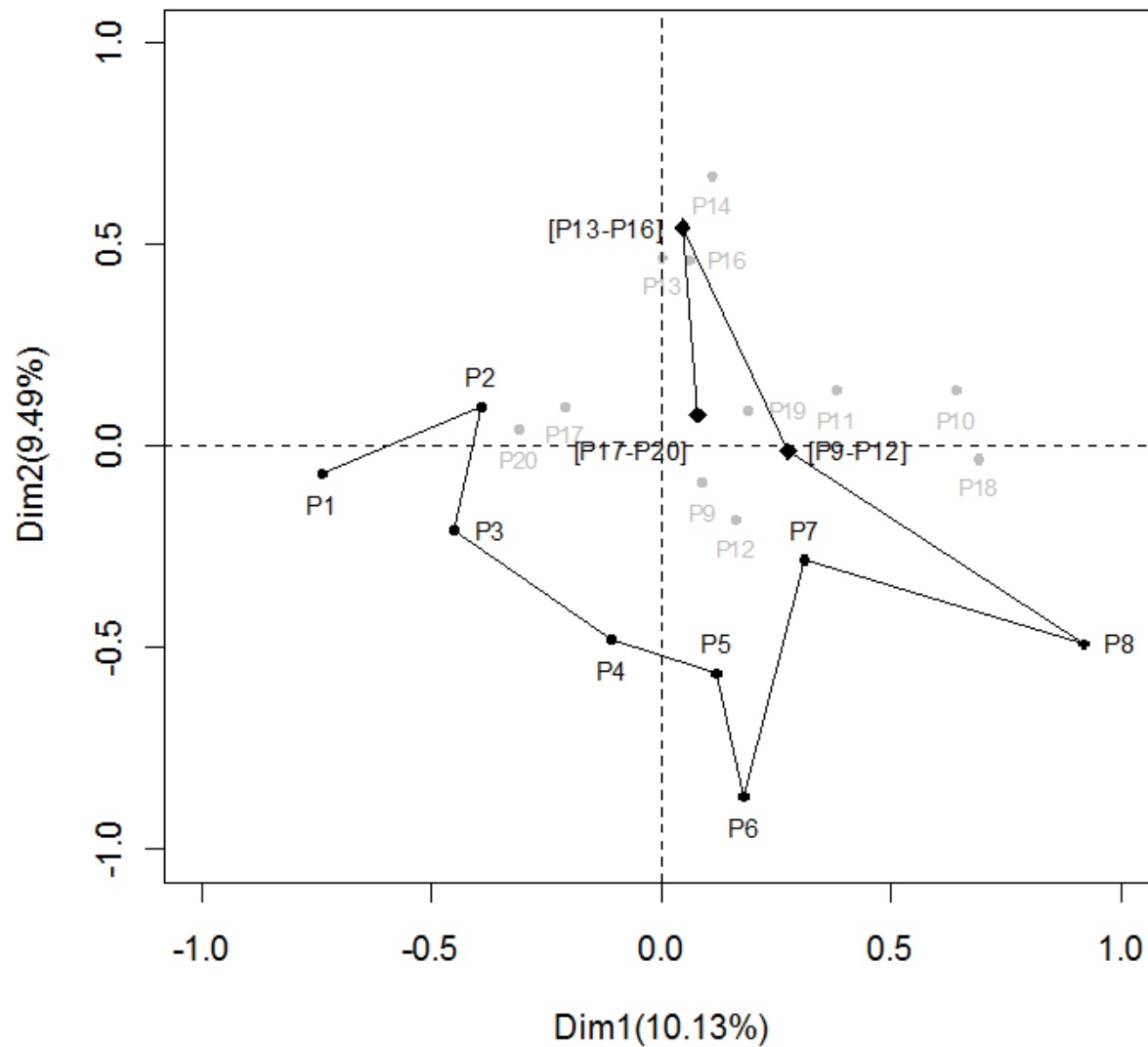
b. Codage du réquisitoire en un tableau

	w1	w2	w3	w4	w5	w302
P1	2	0	1	0	0	0
....	...								
P20	0	0	0	0	1	3

c. Trajectoire des parties révélée par l'AC

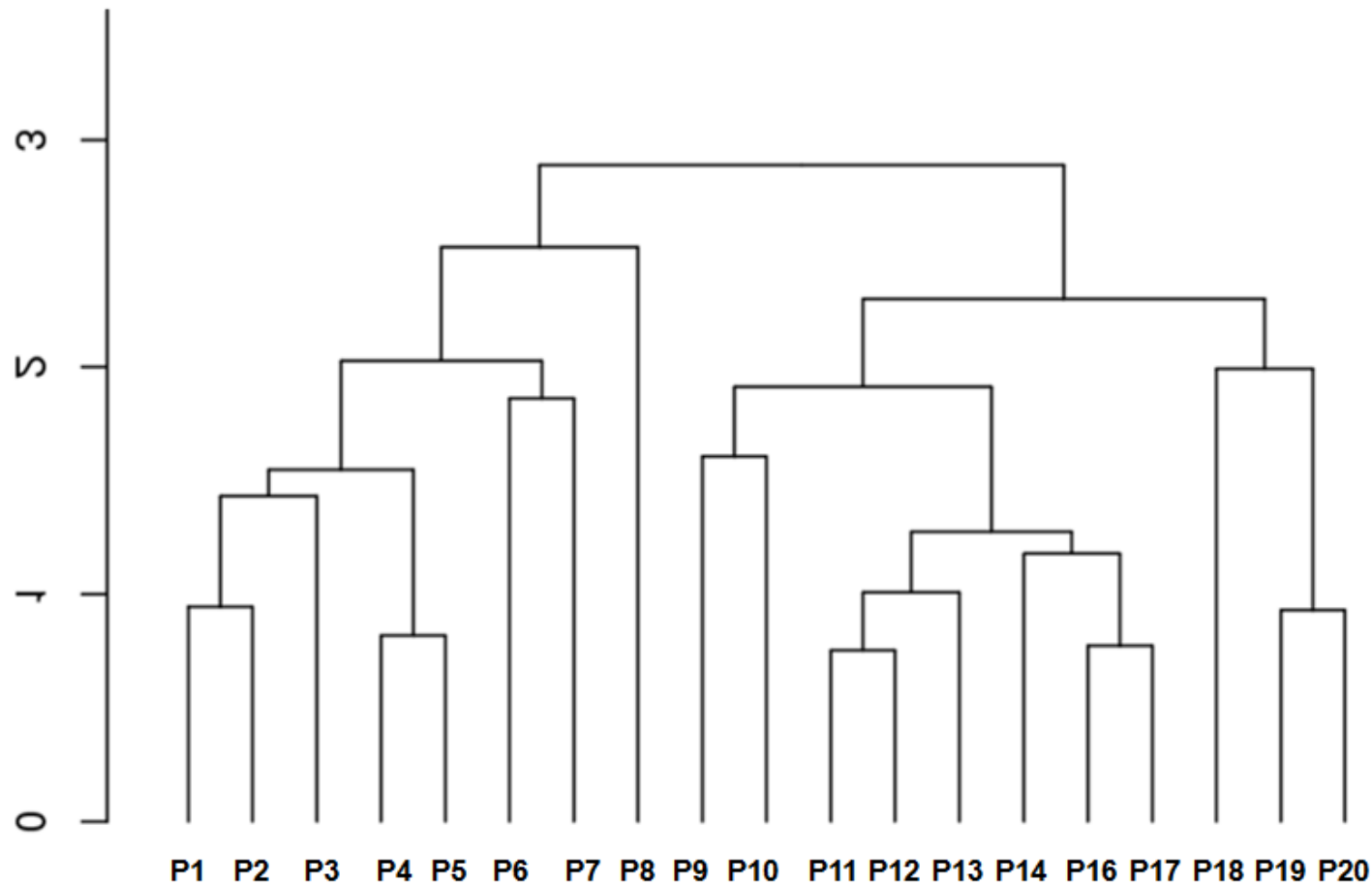


Effet Guttman suivi d'un retour vers le centre

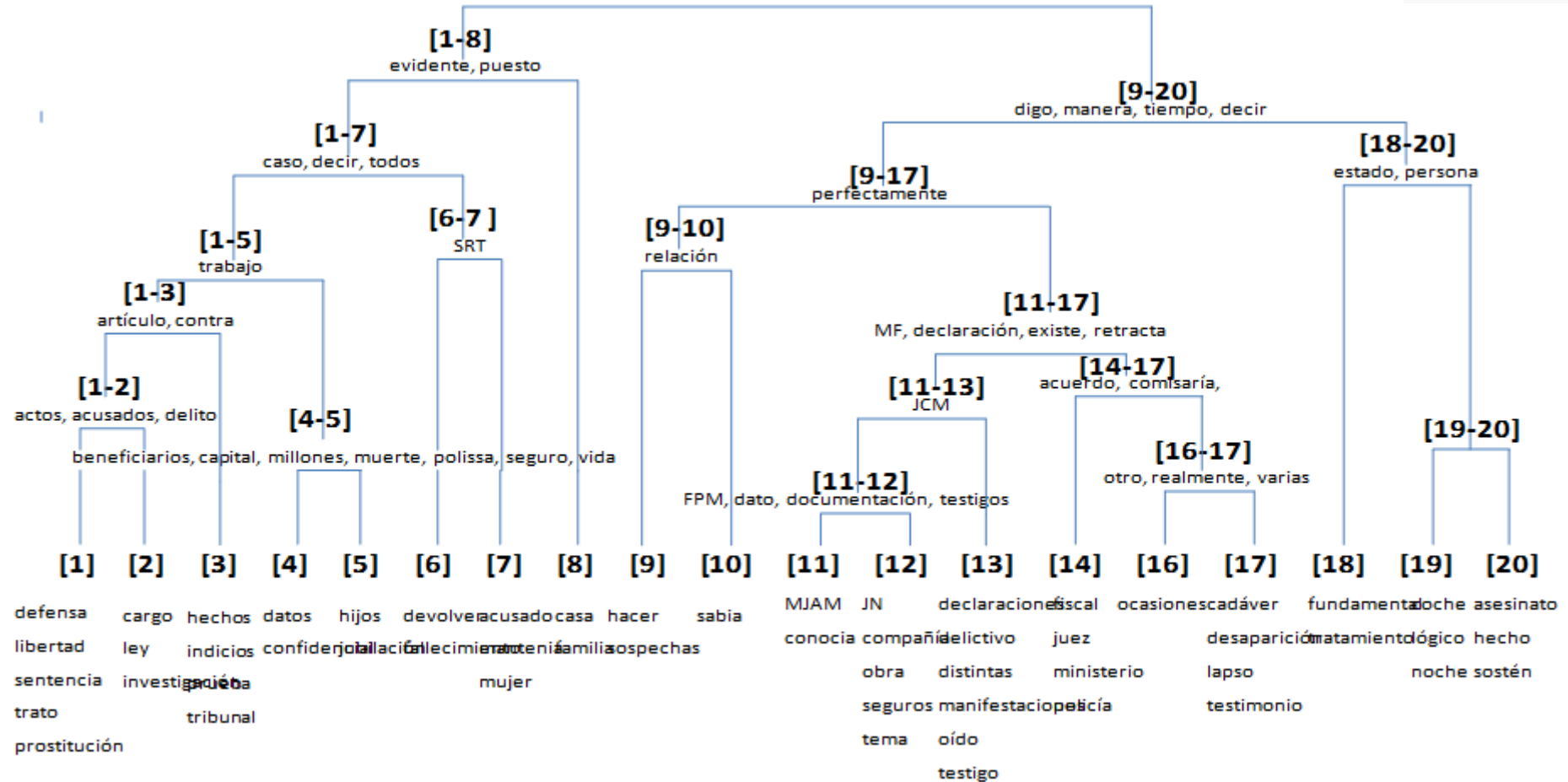


d. Classification hiérarchique avec contrainte de contiguité

L'arbre montre l'organisation hiérarchique du réquisitoire



e. La hiérarchie étiquetée montre le flux des arguments



Caractérisation des deux principales séquences

Sequence [1-8]

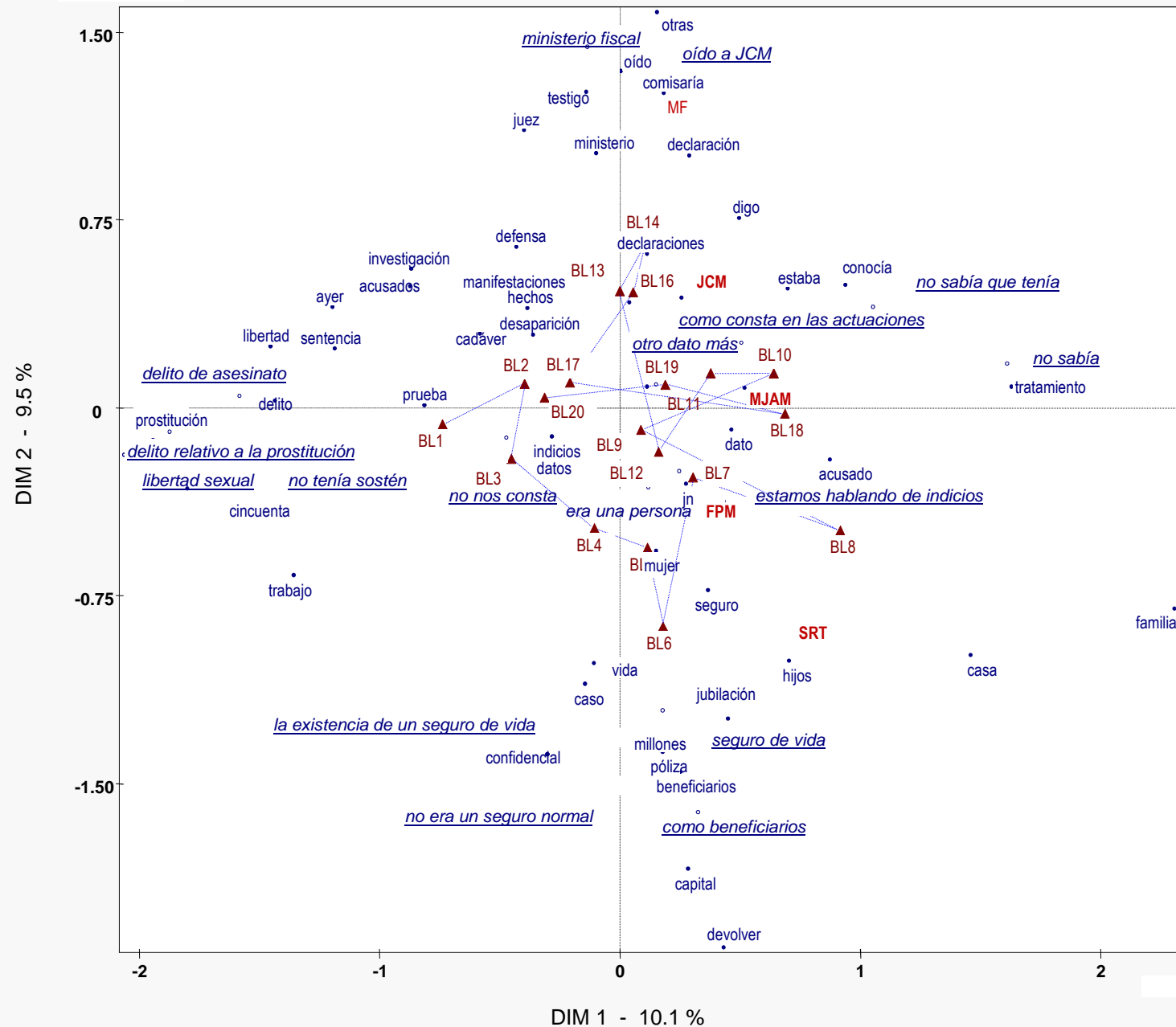
Objective information: Delito (crime), vida (life), caso (case), prostitución (prostitution), capital (capital), SRT (wife of the main defendant), acto (act), cincuenta (fifty), hijos (children), beneficiarios (beneficiaries), consta (we know), millones (millions), jubilación (retirement), años (years), prueba de cargo (incriminating evidence), confidencial (confidential), trabajo (work), devolver (to give back), relativo (relative)

Sequence [9-20]

Speculative argumentation: MF (witness), declaración (statement), policía (police), JCM (accomplice), conocía (knew), persona (person), cuatro (four), dato (data), declaraciones (statements), meses (months), tiempo (time), MJAM (victim), estaba (was), sido (has been), digo (I say), JN , hizo (did/made), dicho (said), había (had), estado (condition/ state)

6. Exemples

6.4 Analyse d'un discours rhétorique



6. Exemples

L'argumentation le long de la trajectoire

P1: Location of part 1
1.28: Distance between part 1 and part 2
(773): Counts of occurrences in part 1

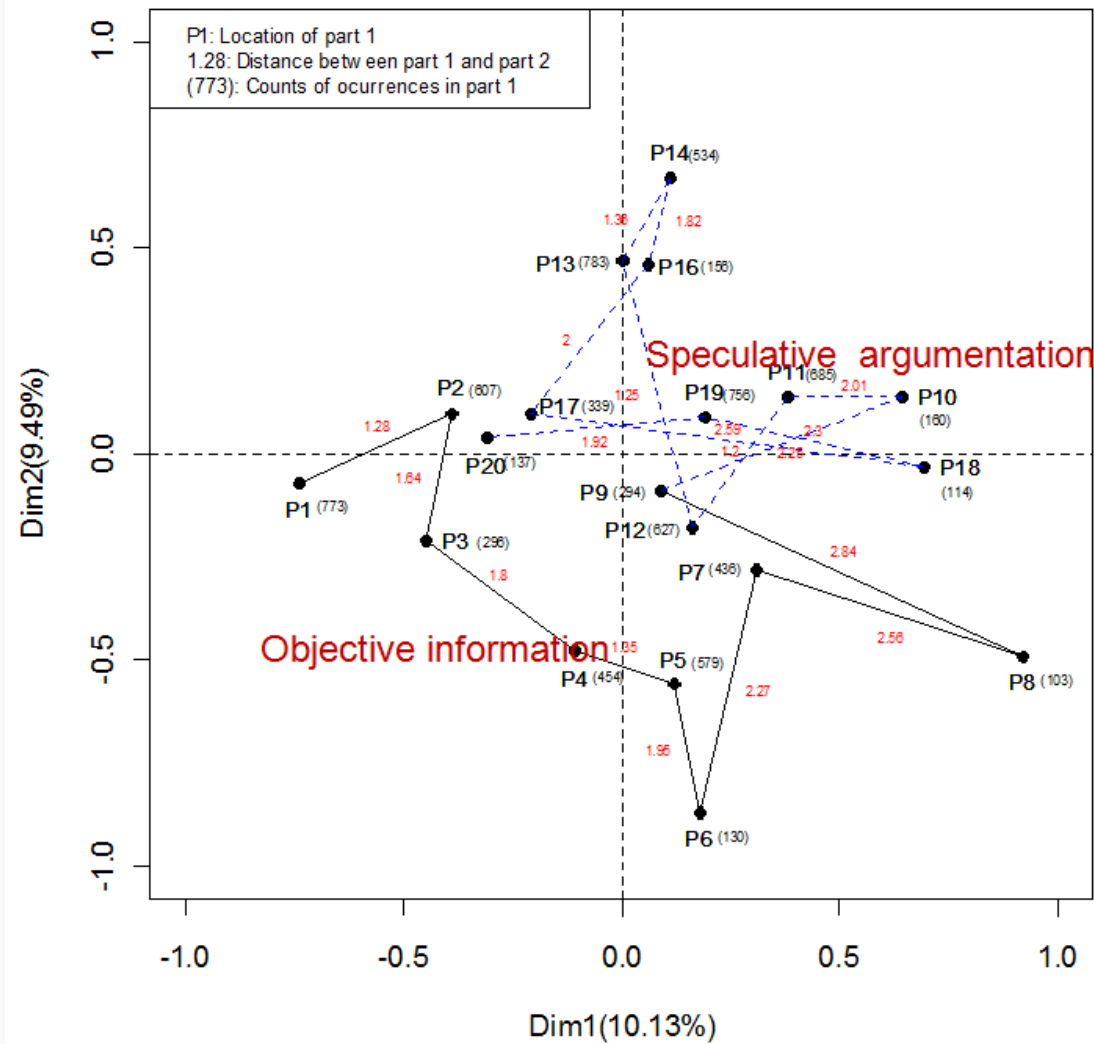
Speculative argumentation

Objective information

Dim2(9.49%)

Dim1(10.13%)

P1 (773) P2 (607) P3 (296) P4 (454) P5 (579) P6 (130) P7 (436) P8 (103) P9 (294) P10 (160) P11 (688) P12 (627) P13 (783) P14 (534) P16 (156) P17 (339) P18 (114) P19 (756) P20 (137)



6. Exemples

Si vous voulez connaître le verdict final: Les accusés ont été condamnés pour meurtre par l'Audience de Barcelone, puis par le Tribunal Suprême

En conclusion de cette étude

- L'analyse des correspondances permet d'accéder à la forme rhétorique d'un texte
- La classification hiérarchique reflète l'organisation hiérarchique du texte et la façon dont les arguments sont imbriqués
- La sélection automatique de mots caractéristiques découvre le sens du texte
- On peut partir d'un découpage automatique initial

6. Evolution du package

- Améliorer la partie graphique
- Utiliser knitr pour proposer des sorties pré-organisées
- Fonction de lemmatisation qui utilise TreeTagger (première version réalisée)
- Autres méthodes:
 - Identifier la fonction des mots
 - Mettre en relief le modèle évolutif d'un corpus chronologique
 - Comparaison de trajectoires

- Benzécri, J.P., 1981. Pratique de l'analyse des données, Vol. 3, Linguistique & Lexicologie. Dunod, Paris.
- Kuyumcuyan, A. (1999), Hétérogénéité textuelle : l'exemple de la fable, *Cahiers de Linguistique Française* 151-179.
- Bécue-Bertaut M., Pagès, J. Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. *Advances in Data Analysis and Classification ADAC*, 9(2), 125-142, 2015.
- Bécue-Bertaut M., Pagès, J., Kostov B. Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. *SORT* 38, 285-302, 2014.
- Bécue-bertaut, M., Kostov, B., Morin A., Naro G. Rhetorical Strategy in Forensic Closing Speeches. *Multidimensional Statistics-Based Methodology. Journal of Classification*, 31, pp- 85-106, 2014.
- Lebart, L., Salem, A., Berry, L., 1998. Exploring textual data. Kluwer, Dordrecht. .