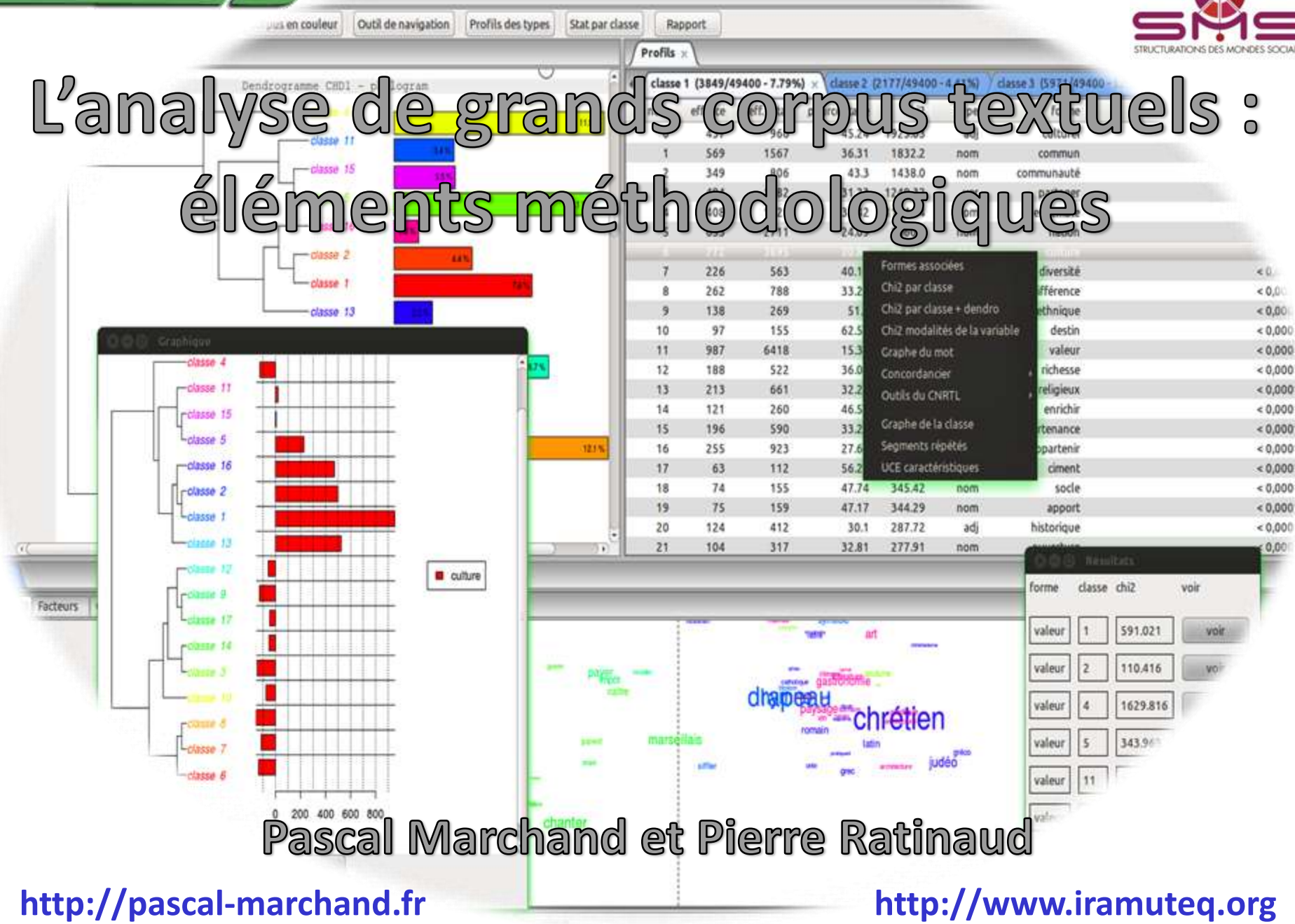
L'analyse de grands corpus textuels : éléments méthodologiques



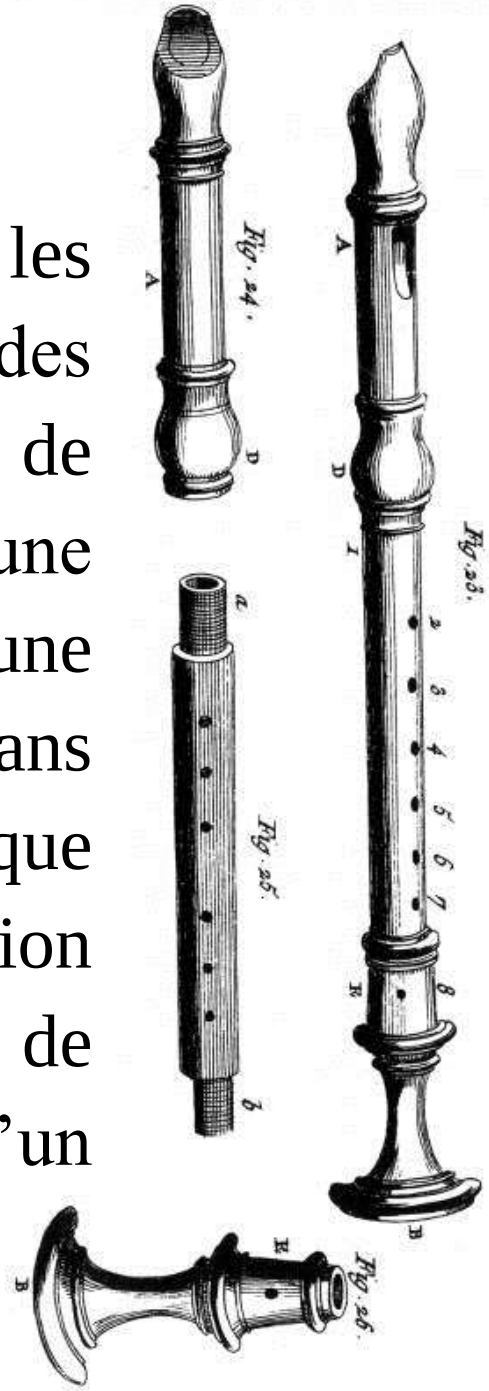
Pascal Marchand et Pierre Ratinaud

http://pascal-marchand.fr

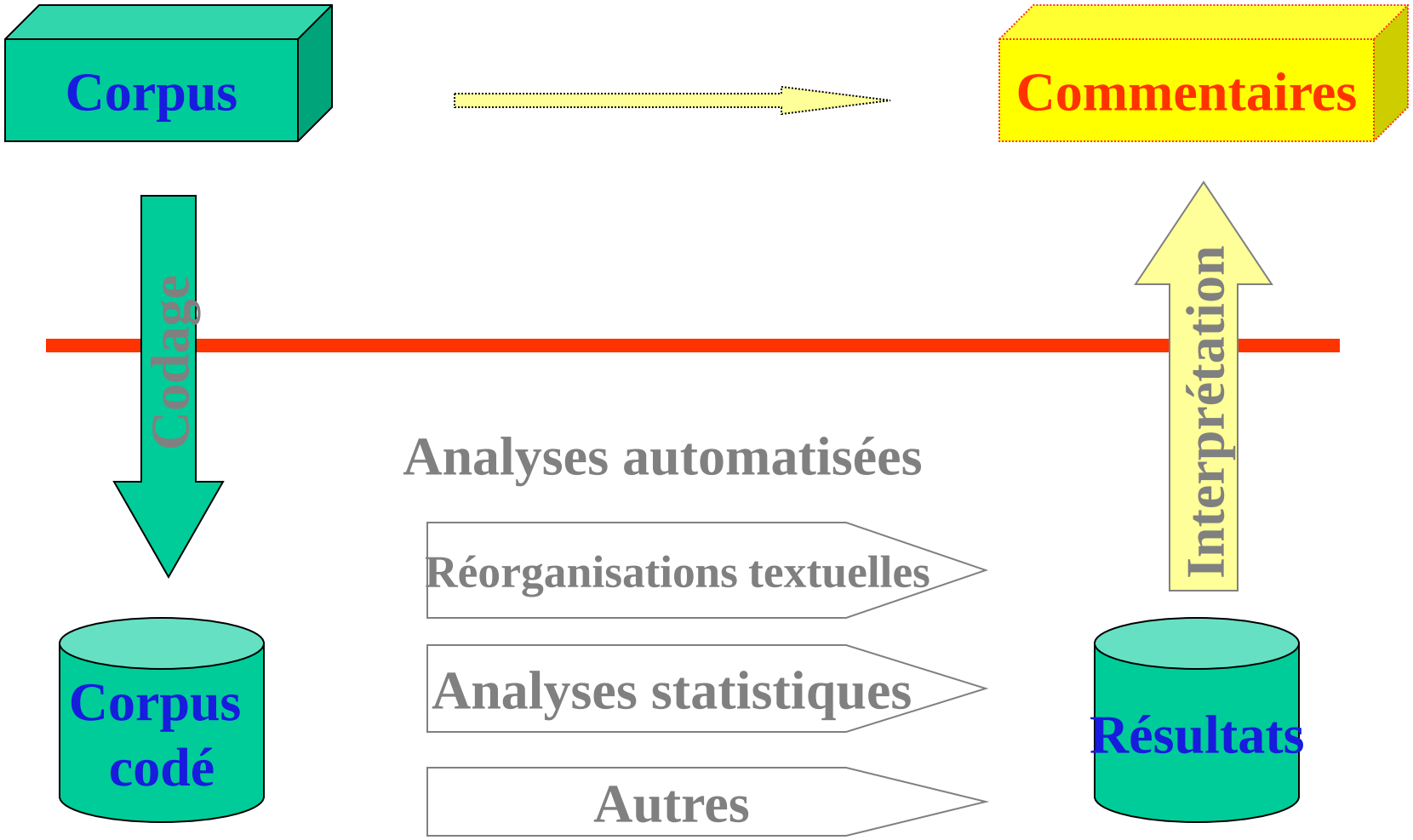
http://www.iramuteq.org

Lebart & Salem (1994)

« Supposons (...) que l'on étudie les histogrammes des longueurs d'ondes correspondant aux couleurs d'un tableau de Rembrandt (pour chacun des pixels d'une reproduction). Il va de soi que l'on utilise une fraction dérisoire de l'information contenue dans l'image d'origine. Il est cependant possible que la forme de l'histogramme (ou d'une fonction plus élaborée des mêmes mesures et données de base) permette de distinguer un Rembrandt d'un Rubens ou d'un Van Dyck » (p.21).



L'interprétation en ADT



Merci à André Salem

Quelques logiciels de lexicométrie

- **Alceste** ➢ M. Reinert (http://www.image-zafar.com)
- **DtmVic** ➢ L. Lebart (http://lebart.org)
- **Hyperbase** ➢ E. Brunet (http://ancilla.unice.fr/~brunet/pub/hyperbase.html)
- **Lexico 3** ➢ A. Salem (http://lexico3.no-ip.org)
- **SPAD** ➢ Decisia / M. Bécue (http://www.spad.eu)
- **Sphinx** *Lexica* ➢ Y. Baulac (http://www.lesphinx-developpement.fr)
- ***Taltac*** ➢ S. Bolasco (http://www.taltac.it/it/index.shtml)
- ***WebLex/TXM*** ➢ S. Heiden (http://textometrie.ens-lyon.fr/)

- **IRAMuTeQ** ➢ P. Ratinaud (Win, Mac, Linux) (http://www.iramuteq.org)

Lexicometrica: http://www.cavi.univ-paris3.fr/lexicometrica/

Quelques définitions

- Les questions que se donne la statistique lexicale sont les suivantes : « quels sont les textes les plus semblables en ce qui concerne le vocabulaire et la fréquence des *formes* utilisées ? Quelles sont les *formes* qui caractérisent chaque texte, par leur présence ou leur absence ? »

  (Lebart & Salem, 1994, p.135).

- **Tableau lexical** (*formes* ***** *textes*)
- La lexicométrie regroupe “ toute une série de méthodes qui permettent d’opérer des ré-organisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire à partir d’une segmentation ”

  (Salem, 1986)

Premiers ministres (1959-2007)

- La constitution de la Ve République - Article 49:
  - Le Premier Ministre, après délibération du Conseil des ministres, engage devant l'Assemblée nationale la responsabilité du Gouvernement sur son programme, ou éventuellement sur une déclaration de politique générale.
  - L'Assemblée nationale met en cause la responsabilité du Gouvernement par le vote d'une motion de censure …/…

Premiers ministres (1959-2012)

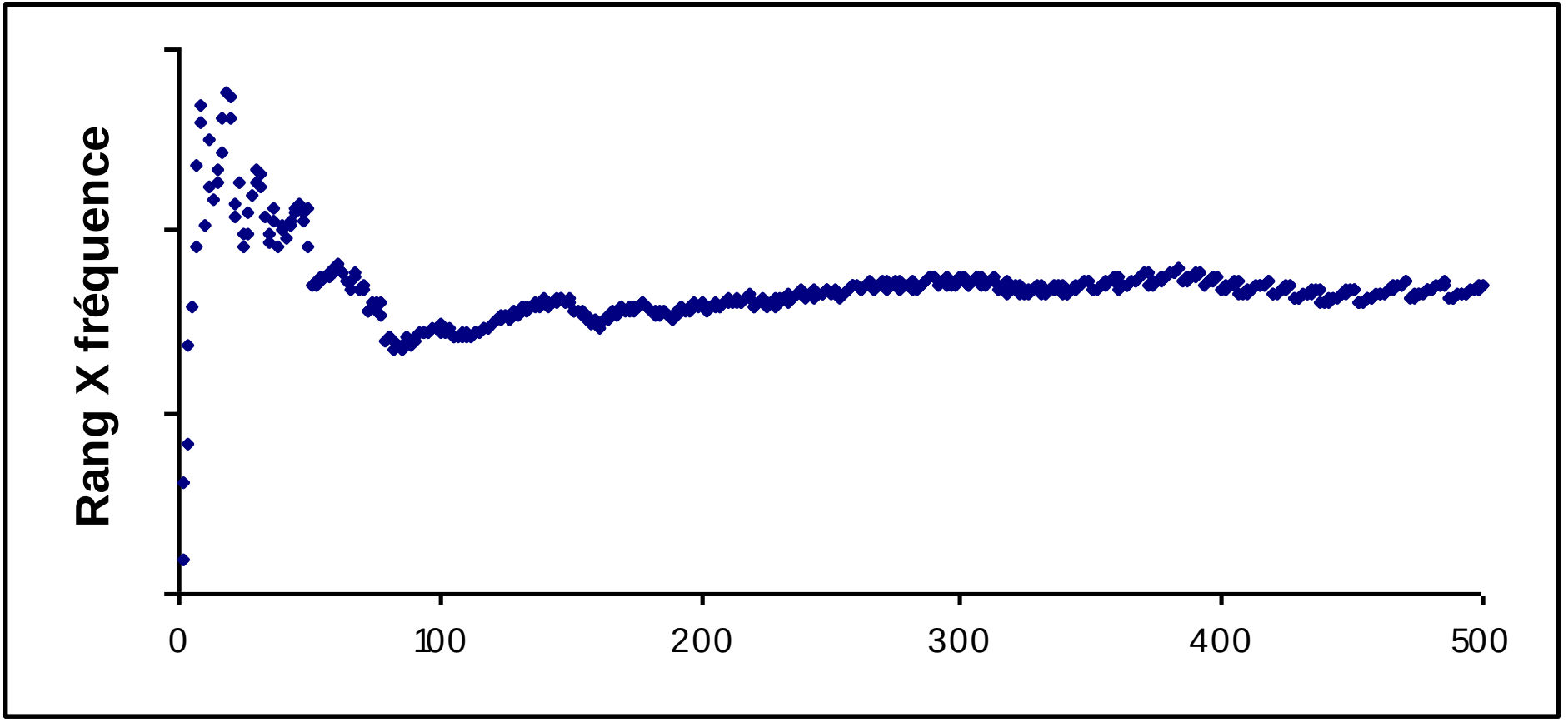
Michel Debré (08/01/59) ; **Georges Pompidou** (26/04/62); **Maurice Couve de Murville** (17/07/68) ; **Jacques Chaban-Delmas** (16/09/69) ; **Pierre Messmer** (03/10/72) ; **Jacques Chirac** (05/06/74) ; **Raymond Barre** (05/10/76 ; 28/04/77 ; 20/04/78) ; **Pierre Mauroy** (08/07/81; 20/09/81; 06/04/83 ; 19/04/84) ; **Laurent Fabius** (24/07/84) ; **Jacques Chirac** (09/04/86 ; 07/04/87 ; 03/12/87) ; **Michel Rocard** (29/06/88 ; 16/01/91) ; **Edith Cresson** (22/05/91 ; 07/02/92) ; **Pierre Bérégovoy** (08/04/92 ; 25/11/92) ; **Edouard Balladur** (08/04/93 ; 15/12/93) ; **Alain Juppé** (23/05/95 ; *15/11/95* ; 02/10/96) ; **Lionel Jospin** (19/06/97); **Jean-Pierre Raffarin** (03/07/02; 02/07/03 ; 05/04/04) ; **Dominique de Villepin** (08/06/05 ; 21/02/06 ; 16/05/06) ; **François Fillon** (03/07/07 ; 24/11/10) ; **Jean-Marc Ayrault** (03/07/12)

Analyse lexicale: 1. Segmentation

- Une suite de caractères bornée par deux caractères délimiteurs est une **occurrence** (*word-tokens*). Deux suites identiques constituent deux occurrences d'une même **forme graphique** (*word-type*).
- Délimiteurs: espace, retour à la ligne, [(« ,.;?:!’/-_ »)]
  - Le tiret / trait d’union / moins / parenthèse
  - L’apostrophe
    - e muet (*c’, d’, j’, jusqu’, lorsqu’, qu’, m’, n’, quoiqu’, presqu’, puisqu’*, etc.)
    - autre voyelle (*ç’* pour ça, *l’* pour le/la, *s’* pour se/si, *t’* pour te/tu, etc.).
    - *aujourd’hui* ou *prud’hommes*

12528	de	1195	c	530	sera	341	ai	233	développement
8324	la	1188	je	528	doit	323	travail	231	économie
6211	l	1183	ne	527	aussi	310	entre	229	deux
5815	et	1127	par	509	ont	306	si	227	enfin
5217	les	1117	ce	494	français	297	économique	226	encore
4908	le	1074	sur	479	y	290	aujourd	226	temps
4631	à	985	qu	462	j	290	hui	222	ensemble
4435	des	908	france	453	etat	288	dont	221	vie
3832	d	855	s	447	sans	283	sociale	220	société
3051	est	838	aux	434	ou	282	on	219	depuis
2982	en	838	n	425	comme	280	seront	216	ceux
2799	que	816	nos	422	ces	278	monde	215	donc
2441	une	810	gouvernement	422	tout	278	république	210	toutes
2425	nous	803	avec	421	son	266	fait	209	soit
2273	qui	744	mais	413	avons	265	loi	208	droit
2142	un	711	elle	410	ses	265	où	208	sécurité
2060	pour	697	cette	409	même	264	contre	207	ainsi
2024	du	695	vous	406	été	263	leurs	206	elles
1977	dans	693	politique	400	faire	262	action	206	moyens
1809	il	667	se	390	ils	256	europe	203	cet
1410	au	651	être	386	faut	243	effort	202	autres
1393	notre	647	sont	375	entreprises	241	peut	202	cela
1368	plus	633	leur	362	emploi	236	nationale	199	mesures
1275	pas	603	pays	346	bien	235	avenir	197	jeunes
1214	a	533	tous	342	sa	235	président	195	croissance

Linguistique quantitative



• Représentation graphique de la loi de Zipf: Produit *rang x fréquence* des 500 premières formes lexicales du corpus « discours de politique générale »

Formes initiales / réduites

Lemmatisation

- Reconnaître les chaînes de caractères communes : deux formes se succédant dans un index alphabétique sont potentiellement liées par une racine commune (*jeune*, *jeunes*, *jeunesse* = *jeune*+).
- Mais des formes très proches ne peuvent être regroupées (*grand*, *gras*, *grave* ≠ *gra*+) ;
- Définir un critère permettant de décider de leur regroupement : on peut, par exemple, construire une liste des suffixes grammaticaux usuels (SHRDLU: Winograd, 1972).

+a	+at	+er	+i	+ir	+it	+re	+u
+able	+ates	+era	+ible	+ira	+ite	+resse	+ude
+ablement	+ateur	+erai	+ice	+irai	+ites	+rez	+ue
+ace	+atif	+eraient	+icien	+iraient	+itif	+rice	+ueuse
+ade	+ation	+erais	+icien	+irais	+ition	+rie	+ueusement
+age	+atique	+erait	+icienne	+irait	+itive	+riez	+ueux
+ai	+ative	+eras	+icienne	+iras	+itude	+ron	+umes
+aie	+atre	+ere	+ide	+irent	+lure	+rons	+ur
+aient	+atrice	+erent	+idement	+irez	+ment	+ront	+ure
+aire	+aux	+eresse	+ie	+iriez	+mental	+s	+urent
+ais	+cale	+erez	+iel	+irions	+mentaux	+se	+us
+aise	+cite	+erie	+ielle	+irons	+mment	+sement	+use
+aison	+d	+eriez	+ien	+iront	+nt	+ssant	+usses
+ait	+dre	+erions	+ienne	+is	+oir	+sse	+ussiez
+al	+e	+eron	+ier	+isant	+oire	+ssement	+ussions
+ale	+eau	+erons	+iere	+isante	+on	+ssent	+ut
+ames	+eaux	+eront	+ieusement	+ise	+ons	+t	+utes
+amment	+ee	+es	+iez	+isme	+ont	+te	+ux
+ance	+een	+esque	+if	+ison	+orat	+teur	+vre
+ant	+eenne	+esse	+ille	+issage	+osite	+tif	+x
+ante	+elle	+et	+iment	+issaient	+pre	+tion	
+ard	+ement	+ete	+imes	+issais	+que	+tique	
+as	+emental	+ette	+in	+issait	+r	+tive	
+asse	+ementaux	+eur	+ion	+issant	+ra	+tre	
+assent	+emment	+euse	+ions	+issante	+rai	+trice	
+asses	+ence	+eusement	+ique	+isse	+raient	+tte	
+assez	+ent	+eux		+issement	+rais	+tude	
+assiez	+ente	+ez		+issent	+rait		
+assions				+isses	+ras		
				+issez			
				+issiez			
				+issions			
				+issons			
				+iste			

Formes initiales / réduites

- ✓Formes dont la flexion entraîne une modification morphologique: *culpabilité* et *coupable*
- ✓Dictionnaire à étiquettes.
  - ✓TreeTagger - a language independent part-of-speech tagger

    http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
  - ✓Cordial Analyseur

    http://www.synapse-fr.com/
  - ✓Lexique 3 (Paris 5)

    http://www.lexique.org/
- ✓Normes de saisie (Labbé, 1990)

  http://halshs.archives-ouvertes.fr/docs/00/43/71/50/PDF/LabbeNormes.pdf

Principes des grammaires formelles

Table:

Catégorie	= Nom
Nombre	= Singulier
Genre	= Féminin
Lex	= table

ferma:

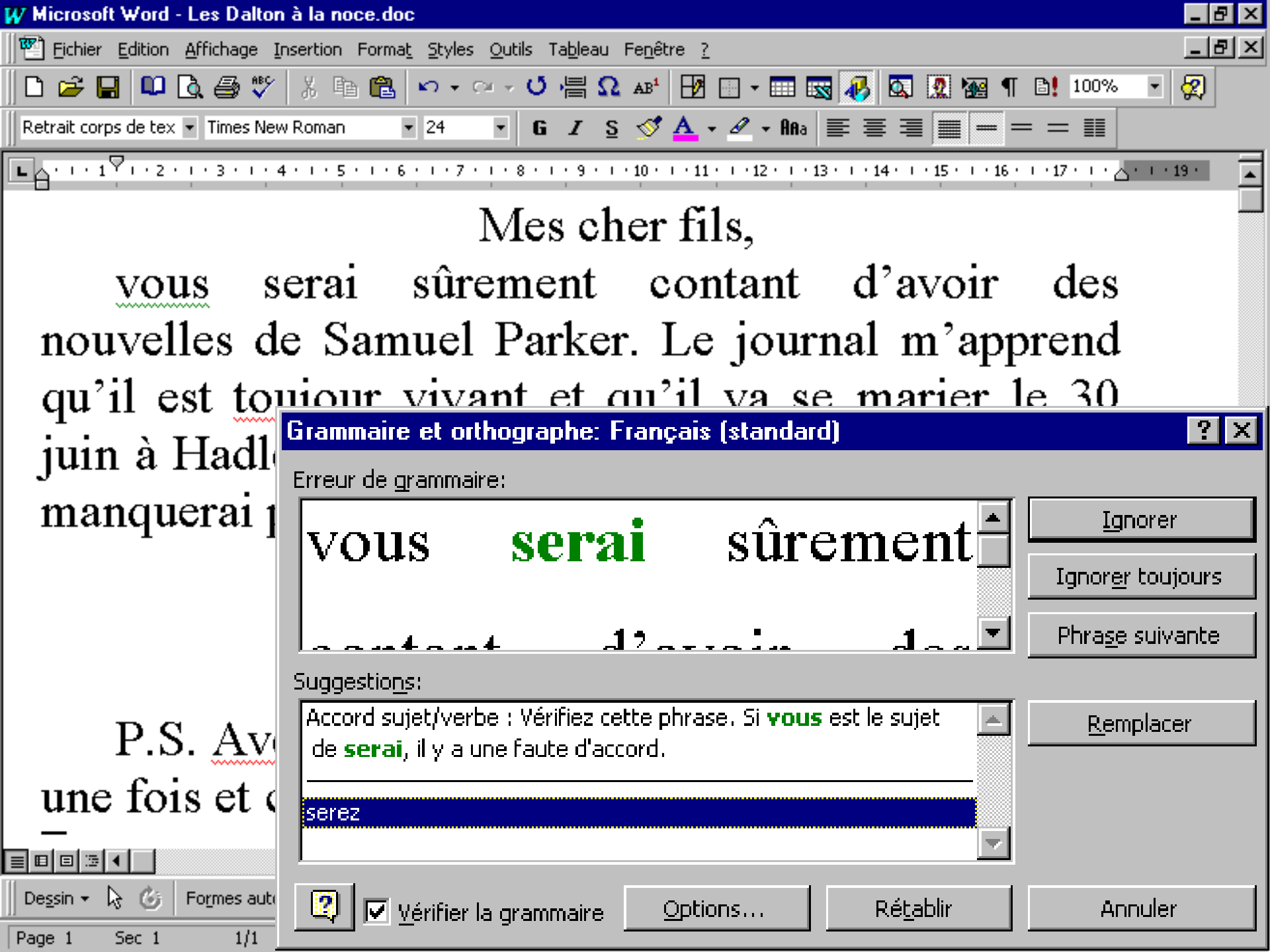
Catégorie	= Verbe
Temps	= Passé simple
Type	= Action
Racine	= fermer
Lex	= ferma

Le bureau:

Cat	= GN	
Forme	= (DET NOM)	
DET	=	Cat = Article
		Type = Défini
		Nombre = Sing.
		Genre = Masculin
		Lex = le
NOM	=	Cat = NOM
		Nombre = Sing.
		Genre = Masculin
		Lex = bureau

Une proposition de type GN + GV + GN peut alors être formalisée par les règles grammaticales suivantes (où “ * ” indique que l’élément qui suit est facultatif ou répété un nombre quelconque de fois) :

$$
\left[\begin{array}{l}
\text{Cat} = \text{Phrase} \\
\text{Forme} = (\text{GN GV GN}) \\
\text{GN} = \left[\begin{array}{l}
\text{Cat} = \text{GN} \\
\text{Forme} = (\text{DET *ADJ NOM *ADJ}) \\
\text{DET} = [\text{Cat} = \text{DET.}] \\
\text{NOM} = [\text{Cat} = \text{NOM}] \\
\text{ADJ} = [\text{Cat} = \text{ADJ}]
\end{array}\right] \\
\left[\begin{array}{l}
\text{Cat} = \text{GN} \\
\text{Forme} = \text{PRON.} \\
\text{PRON} = [\text{Cat} = \text{PRON}]
\end{array}\right] \\
\text{GV} = \left[\begin{array}{l}
\text{Cat} = \text{VERBE} \\
\text{Forme} = (\text{VERBE}) \\
\text{VERBE} = [\text{Cat} = \text{VERBE}]
\end{array}\right]
\end{array}\right]
$$


Microsoft Word - Les Dalton à la noce.doc
Fichier Edition Affichage Insertion Format Styles Outils Tableau Fenêtre ?
Retrait corps de tex
Times New Roman
24
Mes cher fils,
vous serai sûrement contant d'avoir des nouvelles de Samuel Parker. Le journal m'apprend qu'il est toujour vivant et qu'il va se marier le 30 juin à Hadl
manquerai
P.S. Av
une fois et
Grammaire et orthographe: Français (standard)
Erreur de grammaire:
vous serai sûrement
Ignorer
Ignorer toujours
Phrase suivante
Suggestions:
Accord sujet/verbe : Vérifiez cette phrase. Si vous est le sujet de serai, il y a une faute d'accord.
serez
Remplacer
Vérifier la grammaire
Options...
Rétablir
Annuler
Dessin
Page 1 Sec 1 1/1

Regroupements (SR)

- **Locutions :** “aujourd’hui”, “Etats-Unis”, “peut-être”, “point de vue”, “lutte des classes”, “sécurité sociale”, ou “week-end”.
  - En corpus : Segments répétés ➔ Salem

Segment	Fréq.	Segment	Fréq.
président de la république	98	sécurité sociale	22
projet de loi	51	service public	22
il y a	47	en ce qui concerne	21
partenaires sociaux	46	en matière de	21
mesdames et messieurs les députés	43	en même temps	21
en faveur	42	en sorte que	21
en matière	42	mise en place	21
en place	39	économie française	20
dans le cadre	35	commerce extérieur	20
parce qu	34	formation professionnelle	19
mise en oeuvre	29	union européenne	19
mettre en oeuvre	27	assemblée nationale	19
collectivités locales	27	bien entendu	18
secteur public	26	temps partiel	18
en sorte	25	protection sociale	18
bien sûr	25	construction européenne	18

Regroupements (SR)

- **Locutions :** “aujourd’hui”, “Etats-Unis”, “peut-être”, “point de vue”, “lutte des classes”, “sécurité sociale”, ou “week-end”.
  - En corpus : Segments répétés ➔ Salem
  - En langue : Dictionnaires ➔ Silberztein (1993, 1998)
- **Expressions figées :** Gross (1982): *prendre la poudre d’escampette* = fuir
- **La définition des unités d’analyse amène à sortir du champ strictement statistique pour envisager de repérer les statuts syntaxiques (et sémantiques) des formes et l’usage de la langue plutôt que la seule distribution des formes graphiques… Et ça, c’est une autre histoire !**

de	13006	*pre*	**par**	1147	*pre*	**ou**	449	*con*	**savoir**	312	*ver_sup*
la	8806	*art_def*	**sur**	1133	*pre*	**ces**	449	*adj_dem*	**république**	307	*nom*
être	6744	*aux*	**qu**	965	*con*	**falloir**	444	*ver_sup*	**aujourd_hui**	302	*adv_sup*
l	6474	*art_def*	**faire**	956	*ver_sup*	**vouloir**	440	*ver_sup*	**année**	297	*nom*
et	6100	*con*	**france**	948	*nr*	**comme**	438	*con*	**on**	294	*pro_per*
les	5481	*art_def*	**s**	892	*pro_per*	**tout**	433	*pro_ind*	**moyen**	292	*nom*
le	5134	*art_def*	**nos**	885	*adj_pos*	**son**	432	*adj_pos*	**européen**	290	*adj*
à	4684	*pre*	**pouvoir**	884	*ver_sup*	**public**	432	*adj*	**dont**	290	*pro_rel*
des	4569	*art_ind*	**aux**	881	*art_def*	**ses**	431	*adj_pos*	**demander**	277	*ver*
d	3779	*pre*	**avec**	863	*pre*	**ils**	411	*pro_per*	**économie**	276	*nom*
avoir	3271	*aux*	**gouvernement**	849	*nom*	**économique**	399	*adj*	**monde**	276	*nom*
en	3091	*pre*	**politique**	811	*nom*	**premier**	396	*adj*	**contre**	273	*pre*
que	2818	*pro_rel*	**mais**	779	*con*	**national**	390	*adj*	**société**	270	*nom*
nous	2587	*pro_per*	**cette**	722	*adj_dem*	**mettre**	383	*ver*	**où**	270	*pro_rel*
une	2534	*art_ind*	**elle**	709	*pro_per*	**prendre**	379	*ver*	**leurs**	270	*adj_pos*
qui	2390	*pro_rel*	**français**	702	*adj*	**monsieur**	379	*nom_sup*	**europe**	270	*nr*
un	2239	*art_ind*	**se**	695	*pro_per*	**bien**	376	*nom_sup*	**mesure**	266	*nom*
pour	2135	*pre*	**social**	694	*adj*	**travail**	374	*nom*	**agir**	266	*ver*
ne	2135	*adv_sup*	**vous**	686	*pro_per*	**nouveau**	374	*adj*	**an**	263	*nom_sup*
du	2111	*art_def*	**leur**	666	*pro_per*	**sa**	357	*adj_pos*	**président**	258	*nom*
dans	2069	*pre*	**pays**	638	*nom*	**permettre**	355	*ver*	**aller**	257	*ver*
il	1857	*pro_per*	**aussi**	552	*adv_sup*	**effort**	351	*nom*	**objectif**	251	*nom*
notre	1487	*adj_pos*	**tous**	546	*pro_ind*	**même**	347	*pro_ind*	**jeune**	245	*adj*
plus	1417	*adv_sup*	**emploi**	543	*nom*	**action**	328	*nom*	**engager**	245	*ver*
au	1386	*art_def*	**entreprise**	525	*nom*	**si**	320	*con*	**projet**	244	*nom*
pas	1360	*adv_sup*	**grand**	517	*adj*	**loi**	318	*nom*	**avenir**	244	*nom*
je	1271	*pro_per*	**y**	503	*pro_per*	**entre**	318	*pre*	**temps**	243	*nom*
devoir	1244	*ver_sup*	**j**	493	*pro_per*	**droit**	318	*nom*	**service**	241	*nom*
c	1177	*pro_dem*	**etat**	476	*nr*	**donner**	313	*ver*	**réforme**	241	*nom*
ce	1170	*pro_dem*	**sans**	467	*pre*	**dire**	313	*ver_sup*	**assurer**	241	*ver*

Analyse lexicale : 2. partition

- La statistique mesure des différences
  - Echantillonnage
  - Groupes (indépendants, appariés)
- Hypothèses
  - Approche hypothético-déductive
  - Variables construites
  - Approche inductive
- Recherche improbable d'une homogène diversité

Globale

nombre de parties : 37

nombre d'occurrences : 239802

nombre de formes : 7455

moyenne d'occurrences par forme : 32.17

nombre d'hapax : 2669 (1.11% des occurrences - 35.80% des formes)

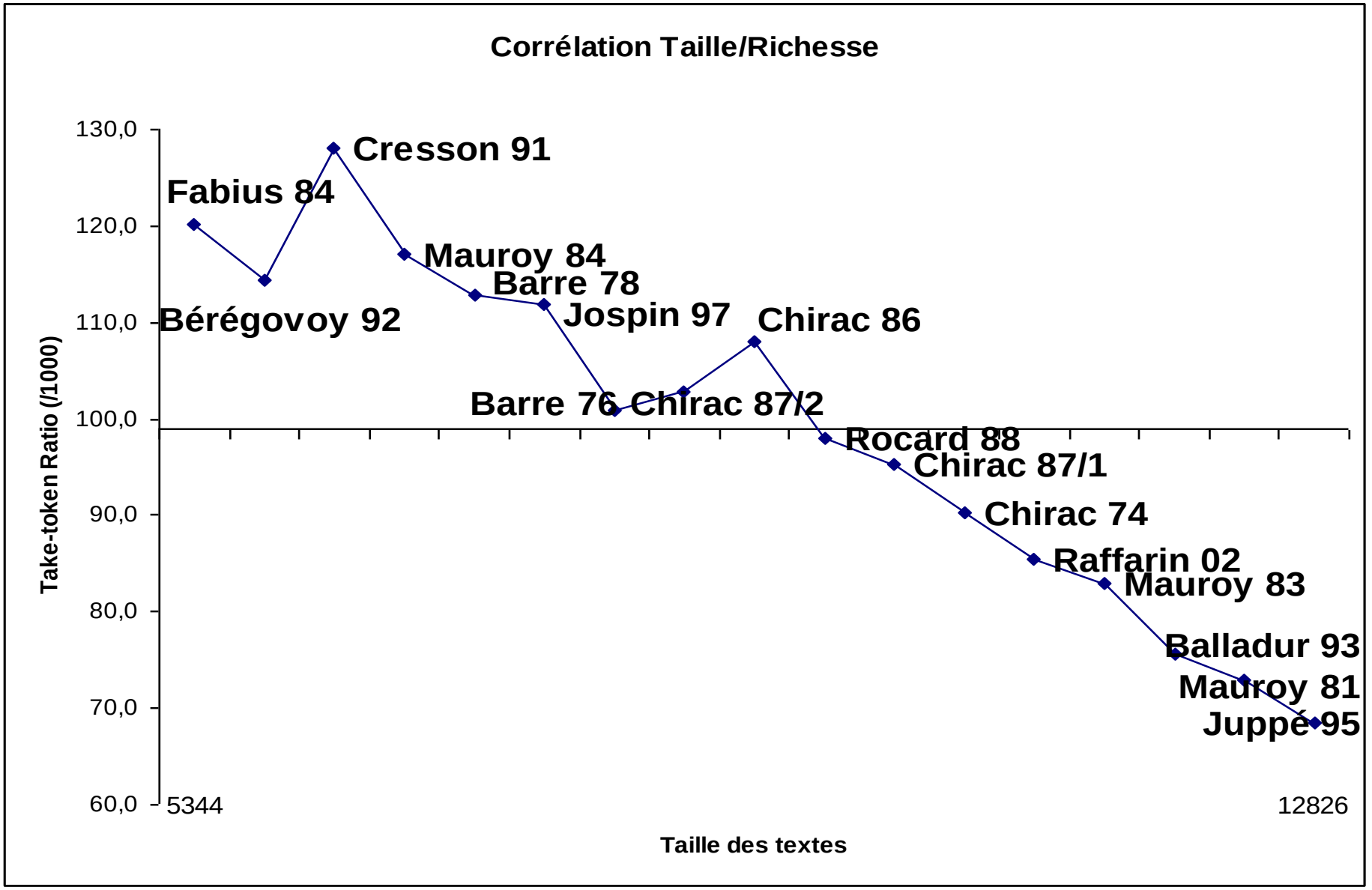
moyenne d'occurrences par partie: 6481.14

	Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	1	159debré	8810	2017	1188	442	de
✓	2	162pompidou	4395	1333	875	204	de
✓	3	168couve	4182	1176	706	181	de
✓	4	169chaban	1804	742	517	95	de
✓	5	172messmer	7545	2061	1344	424	de
✓	6	174chirac	8519	1997	1224	509	de
✓	7	176barre	7013	1802	1135	388	de
✓	8	177barre	4121	1229	821	183	de
✓	9	178barre	6048	1636	1007	354	de
✓	10	181amauroy	11652	2648	1590	662	de
✓	11	181bmauroy	8711	2168	1333	477	de
✓	12	183mauroy	9277	2112	1229	563	de
✓	13	184amauroy	5425	1507	916	287	de
✓	14	184bfabius	5351	1477	953	224	de
✓	15	186chirac	9416	2415	1511	569	de
✓	16	187achirac	7856	1840	1109	415	de
✓	17	187bchirac	6687	1837	1160	357	de
✓	18	188rocard	7610	2091	1366	407	de
✓	19	191arocard	1794	631	427	71	la
✓	20	191cresson	5265	1532	983	291	de
✓	21	192acresson	4266	1381	925	226	de
✓	22	192bbérégovoy	5484	1556	1006	227	de
✓	23	192cbérégovoy	2570	801	521	122	de
✓	24	193aballadur	11531	2411	1406	658	de
✓	25	193bballadur	7011	1736	1053	356	de
✓	26	195ajuppé	13063	2689	1495	787	de
✓	27	195bjuppé	4769	1324	840	322	de
✓	28	196juppé	5058	1303	787	243	de
✓	29	197jospin	6715	1794	1105	368	de
✓	30	202raffarin	9217	2087	1216	511	de
✓	31	203raffarin	3247	971	608	171	de
✓	32	204raffarin	3717	1040	640	212	de
✓	33	205villepin	6595	1656	971	394	de
✓	34	206avillepin	4833	1366	871	247	de
✓	35	206bvillepin	3527	1027	673	175	de
✓	36	207fillon	8148	2294	1546	409	de

Richesse du vocabulaire

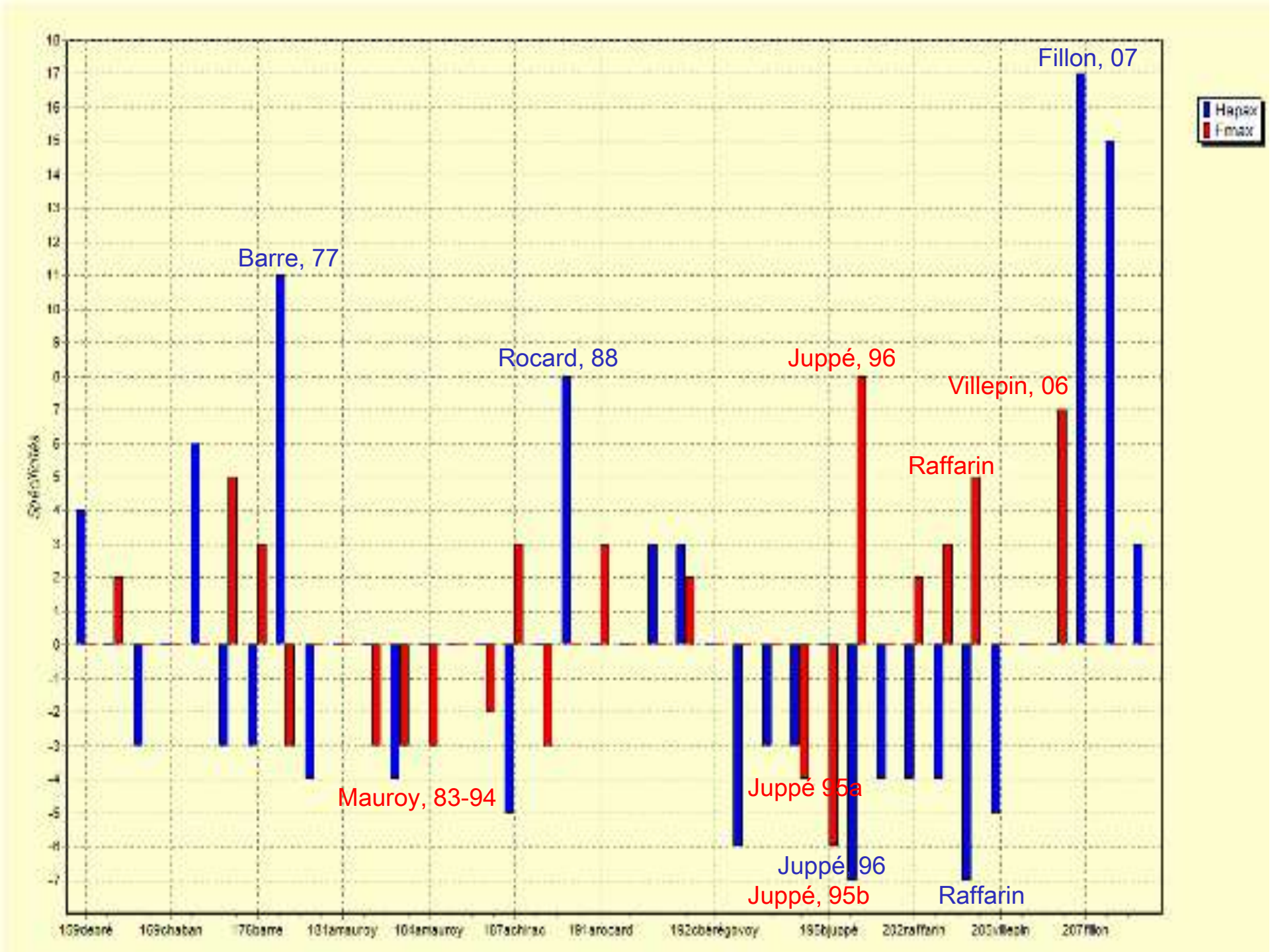
- Type-Token Ratio : Richesse max. = 1, lorsque la réponse ne contient que des mots différents.
    - Sensible à la taille

Richesse du vocabulaire



Richesse du vocabulaire

- Hapax
    - Rapport hapax / formes
- Fmax
    - Rapport Fmax / formes


Fillon, 07
Barre, 77
Rocard, 88
Juppé, 96
Villepin, 06
Raffarin
Mauroy, 83-94
Juppé 95a
Juppé, 96
Juppé, 95b
Raffarin
Hepax
Fmax
109debré
109chaban
76barre
101armauroy
104armauroy
107achirac
191arocard
192cbérégovoy
193bjuppé
202raffarin
207fillon

Lexico3 - [Sections découpées d'après le délimiteur: ! . ? § - vue n°1]

Fichier Traitement Fenêtre

Lexico 3

Navigation | Rapport | Dictionnaire

Sélectionnez une couleur :

Recherche :

Formes (ordre lexicométrique)	Fréquence
voulions	1
voulus	1
voyages	1
voyait	1
vraisemblablement	1
vulgarisation	1
vulnérable	1
vus	1
washington	1
weil	1
williamsburg	1
xavier	1
xème	1
xénophobe	1
xénophobes	1
xénophobie	1
xième	1
xvième	1
xxx	1
yves	1
yvette	1
zapatero	1
zélande	1
zep	1
!	100
"	398
'	15527
(	11
)	11
,	14191
-	2238
.	10645
/	39
:	900
;	406
?	289
[	1
]	1
§	3772

12981 formes

9 Partition : d seuillage 1000+ Spécifs

Feuille 1

d 206avillepin 13550 13600 13650 13700 13750

d 206bvillepin 13800 13850 13900

d 207fillon 13950 14000 14050 14100 14150 14200 14250 14300 14350

d 207_sarkozy 14400 14450 14500 14550 14600 14650 14700 14750 14800

Section :

Occurrence :

Rapport Effacer

et j ' ai en mémoire ce sous - **officier** français qui , il y a quelques années , m ' expliquait dans un **village** constamment **bombardé** de **somalie** , que la première tâche de sa compagnie **consistait** à reconstruire la **maternité** détruite .

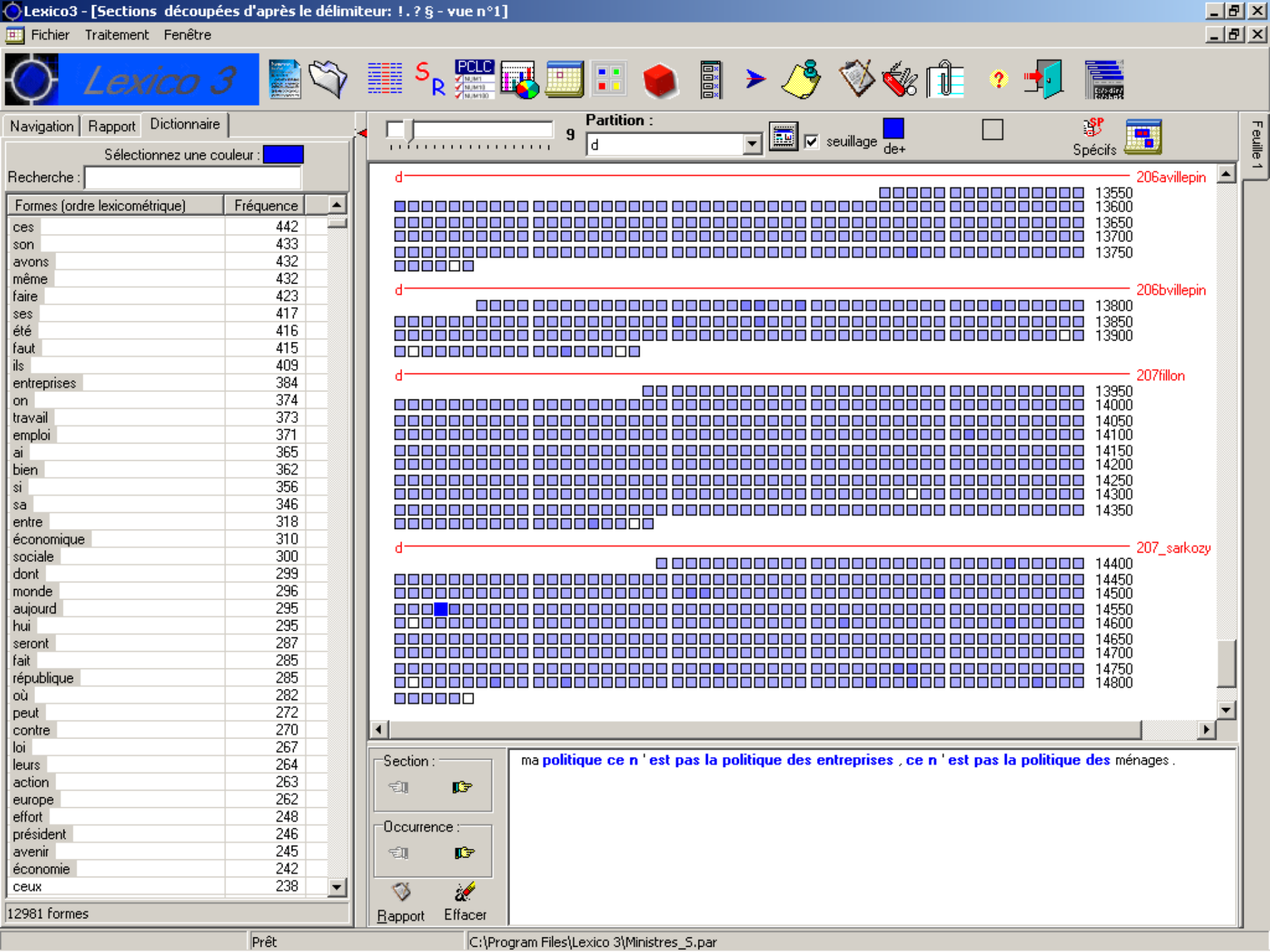
Prêt C:\Program Files\Lexico 3\Ministres_S.par

Les hapax de F.Fillon

- un vocabulaire « recherché » :
  - (…) une *opportunité* de se *détacher* des *postures* idéologiques et des réflexes *claniques*.
  - L'immense *cohorte* de nos savants, *biologistes*, *mathématiciens*, philosophes, juristes, *historiens* qui *firent* notre rayonnement ne doit pas s'arrêter au seuil d'un siècle, où, précisément, le pouvoir de la matière *grise* dessinera notre avenir.
- Des événements d'actualité (narration) :
  - La France est grande lorsqu'elle défend, à travers la libération d'*Ingrid Bétancourt* et des *infirmières bulgares* injustement *condamnées*, les droits *inaliénables* de tout être humain.
  - Et j'ai en mémoire ce *sous-officier* français, qui, il y a quelques années, m'expliquait que dans un *village* constamment *bombardé* de *Somalie*, la première tâche de sa compagnie *consista* à reconstruire la *maternité* détruite.
- « Storytelling » (Salmon, 2007)?

Banalité

- Fmax (100 premiers rangs de l'index) :
  - Mots outils
  - *France (862), gouvernement (790), politique (673), pays (589), français (478), État (440), entreprises (367), emploi (354), travail (303), économique (291), aujourd'hui (289), sociale (271), monde (268), République (259), action (258), loi (257), Europe (242), effort (233), avenir (232), développement (230), économie (230), nationale (228) ...*


Lexico3 - [Sections découpées d'après le délimiteur: ! . ? § - vue n°1]
Fichier Traitement Fenêtre
Lexico 3
Navigation Rapport Dictionnaire
Sélectionnez une couleur :
Recherche :
Formes (ordre lexicométrique) Fréquence
ces 442
son 433
avons 432
même 432
faire 423
ses 417
été 416
faut 415
ils 409
entreprises 384
on 374
travail 373
emploi 371
ai 365
bien 362
si 356
sa 346
entre 318
économique 310
sociale 300
dont 299
monde 296
aujourd 295
hui 295
seront 287
fait 285
république 285
où 282
peut 272
contre 270
loi 267
leurs 264
action 263
europe 262
effort 248
président 246
avenir 245
économie 242
ceux 238
12981 formes
Partition :
d
seuillage
de+
Spécifs
Feuille 1
d 206avillepin
13550
13600
13650
13700
13750
d 206bvillepin
13800
13850
13900
d 207fillon
13950
14000
14050
14100
14150
14200
14250
14300
14350
d 207_sarkozy
14400
14450
14500
14550
14600
14650
14700
14750
14800
Section :
Occurrence :
Rapport
Effacer
ma politique ce n ' est pas la politique des entreprises , ce n ' est pas la politique des ménages .
Prêt
C:\Program Files\Lexico 3\Ministres_S.par

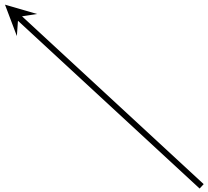
Les Fmax de N.Sarkozy

- ma politique ce n'est pas la politique des entreprises, ce n'est pas la politique des ménages.
- ma politique c'est la politique pour tout le monde.
- l'économie, ce n'est pas que de la statistique.
- l'économie, ce n'est pas que de l'arithmétique.
- n'est - ce pas ce que les français attendent de nous ?
- on ne peut pas être le seul pays à faire comme nous faisons aujourd'hui.
- si ce n'est pas important, si ce n'est pas utile, nous ne le ferons pas.
- parce que si l'on veut réconcilier les français avec l'entreprise, avec l'économie, avec le marché, il ne faut pas accepter les excès, il ne faut pas accepter les dérapages, il ne faut pas défendre qui l'indéfendable, il ne faut pas excuser l'inexcusable.
- la politique que le gouvernement va conduire sera une politique pour tous les français.

Tableau lexical *formes * parties*

Parties (variable-s / unités de contexte):
Partition (Considérations « verticales »)

Lexique:
(segmentation / lemmatisation)
Considérations « horizontales »



•Nombre d'occurrences
•Présence / absence

Formatage (*var_mod)

**** *var1_mod1 *var2_mod3

texte texte ...

**** *var1_mod2 *var2_mod1

texte texte ...

Les variables étoilées (et les thématiques) ne doivent pas contenir d'espaces ou de caractères spéciaux. Elles ne doivent contenir que des caractères parmi a-z, A-Z, 1-9 et des tirets bas (_).

Formatage (thématiques)

**** *var1_mod1 *var2_mod3

-*thematique_1

texte texte ...

-*thematique_2

texte texte ...

**** *var1_mod2 *var2_mod1

-*thematique_1

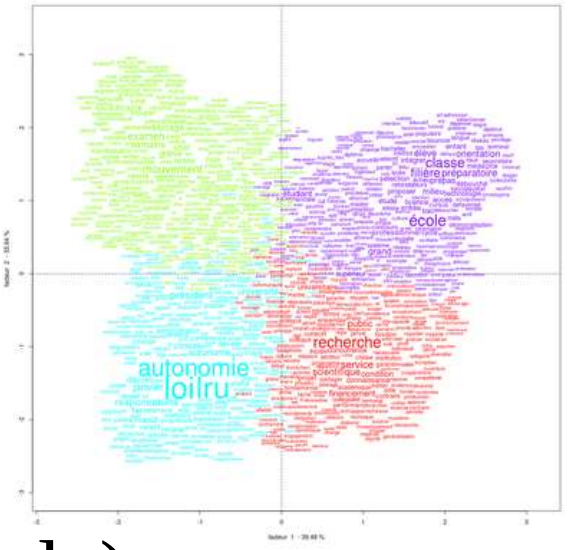
texte texte ...

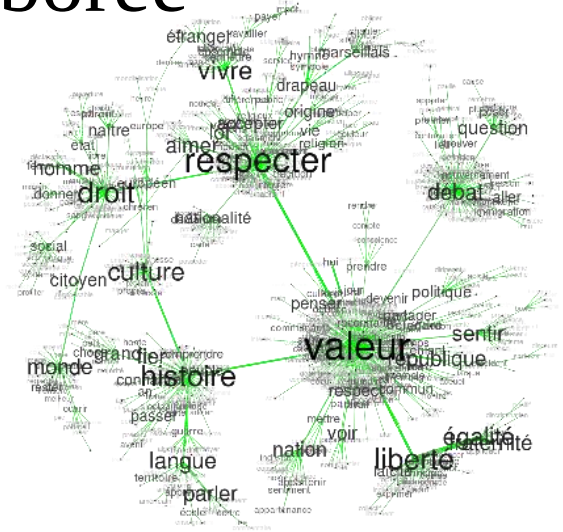
-*thematique_2

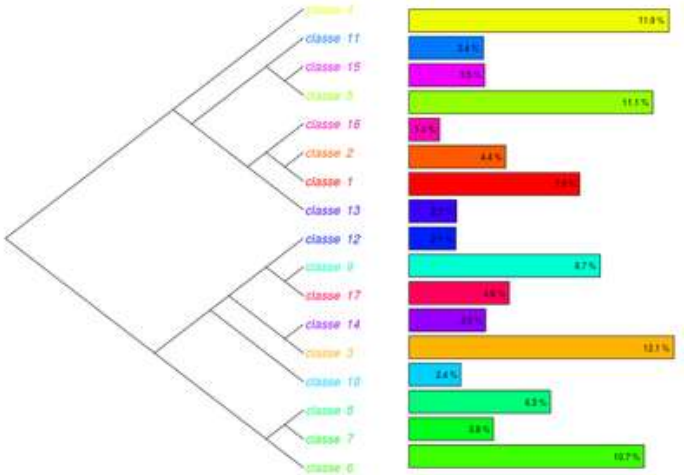
texte texte ...

Analyse lexicale : 3 statistiques

- *Formes* et réponses caractéristiques, **spécificités** (profil)
- Méthodes factorielles
- Classification automatique
- Calculs de distance (connexion lexicale) et représentation arborée
- Cooccurrences et similitudes



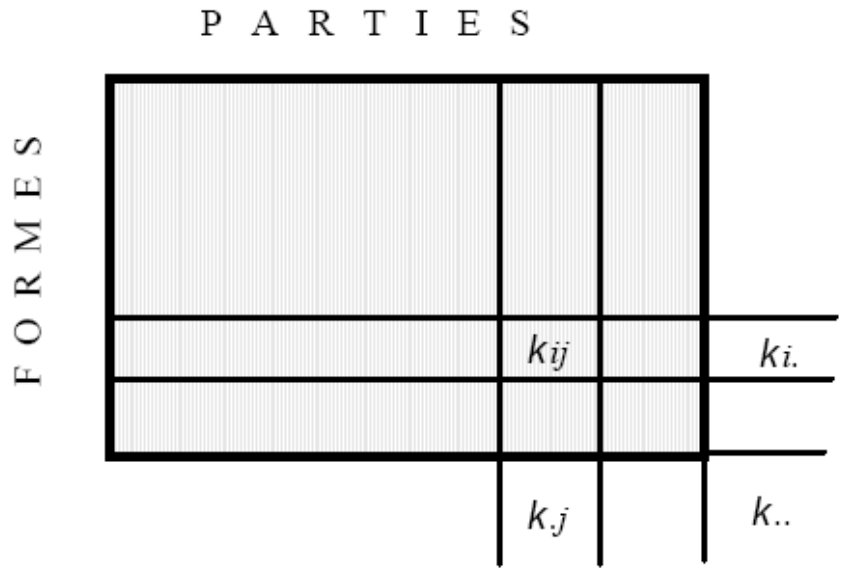




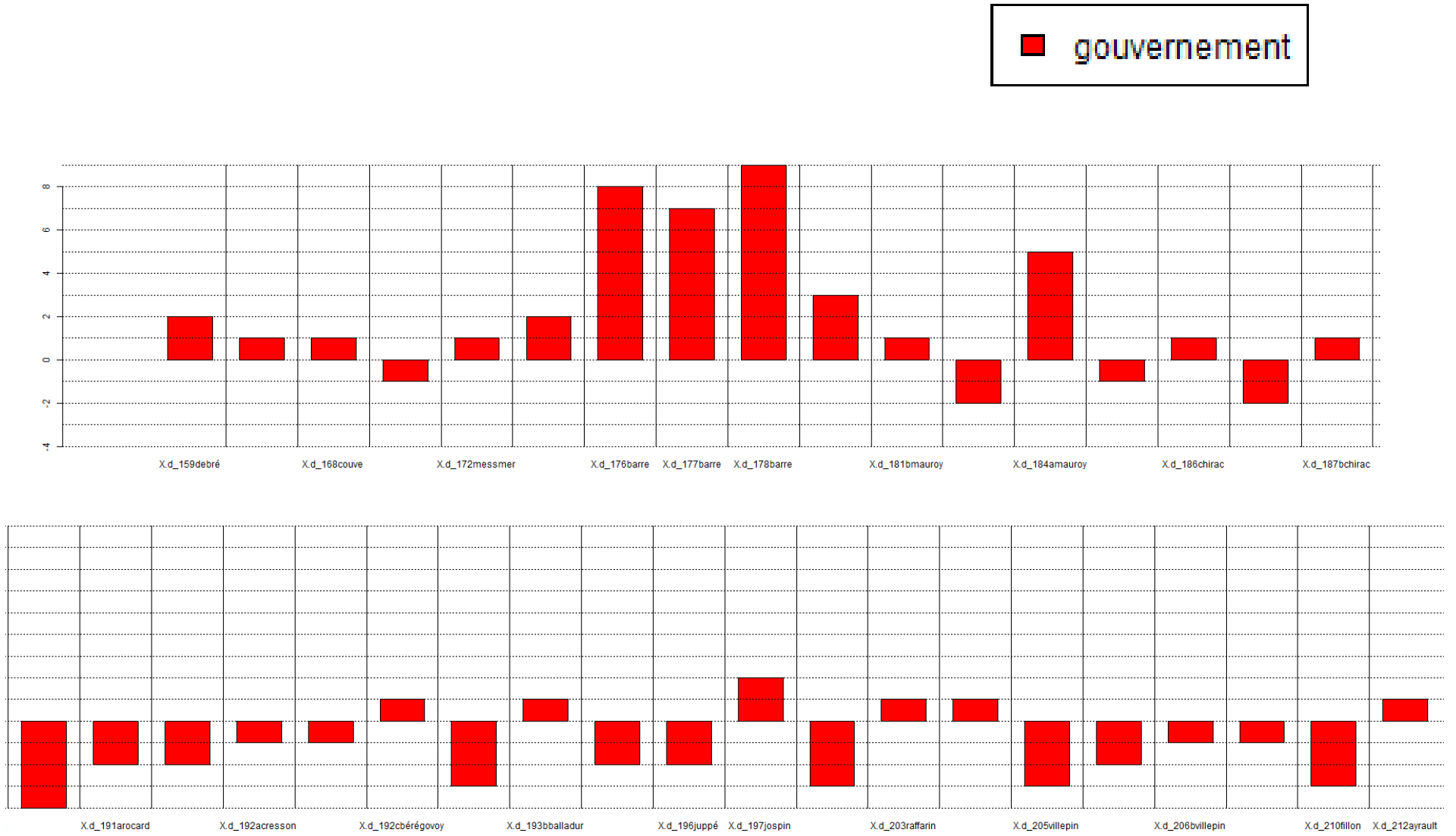
Les spécificités lexicales

- Si l'on considère une *forme* lexicale particulière dans un corpus, les occurrences de cette *forme* peuvent se distribuer:
  - de façon équilibrée dans toutes les *parties* (hasard)
  - ou certaines *parties* peuvent révéler une fréquence de cette *forme* plus élevée que d'autres (écart au hasard).
- A ce calcul, qui fait intervenir la comparaison d'une distribution observée à une distribution équilibrée (ou « théorique »), est associé une probabilité (« Modèle hypergéométrique », Lafon, 1984).

Les spécificités lexicales



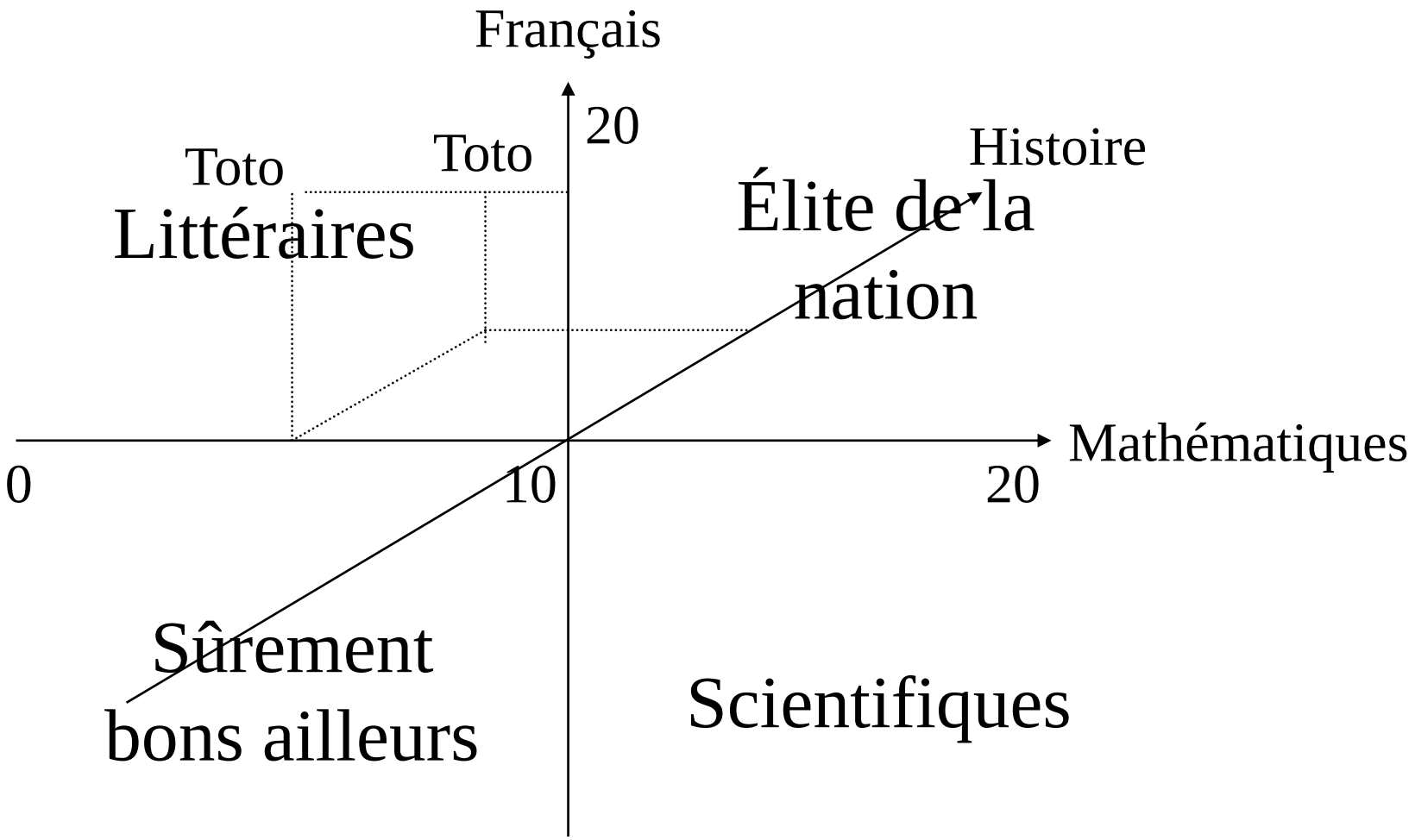
$k_{..}$	taille du corpus
$k_{i.}$	fréquence de la forme dans le corpus
k_{ij}	fréquence de la forme dans la partie
$k_{.j}$	taille de la partie


gouvernement
8
6
4
2
0
-2
-4
X.d_159debré
X.d_168couve
X.d_172messmer
X.d_176barre
X.d_177barre
X.d_178barre
X.d_181bmauroy
X.d_184amauroy
X.d_186chirac
X.d_187bchirac
X.d_191arocard
X.d_192acresson
X.d_192cbérégovoy
X.d_193bballadur
X.d_196juppé
X.d_197jospin
X.d_203raffarin
X.d_205villepin
X.d_206bvillepin
X.d_210fillon
X.d_212ayrault

N. Sarkozy, le 20 juin 2007

- **Spécificités**:
  - **Positives**: « **on** » (91 fois ; *on ne peut pas* : 13 fois ; *on va* : 7 fois), souvent associé à la négation (28 fois dont 13 fois : *on ne peut pas*); « **je** » (150 fois et « j' » : 27 fois ; *je veux prendre mes responsabilités* : 14 fois ; *je vais* : 6 fois); **négation** (*ce n'est pas* : 23 fois) ***; Premier ministre; croissance; travail*** (50 fois: *tag?*) en cooccurrence spécifique avec ***politique***.
    - *« Politique monétaire, politique budgétaire, je ne les jugerai que par rapport à un seul critère : cela récompense le travail ou cela dévalorise le travail. Tout ce qui récompense le travail sera choisi, tout ce qui dévalorise le travail sera écarté ».*
  - **Négatives**: « **nous** » (*notre pays*; *notre* : 20 fois ; *nos* : 10 fois), ***gouvernement, loi, projet, solidarité, république***.

Analyse des correspondances



Les relations entre variables créent de la redondance qui peut être exploitée pour réduire les données à quelques facteurs exprimant l'essentiel de ces relations (*compression avec perte*).

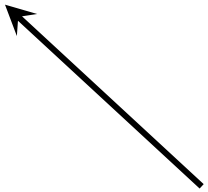
Tableau lexical (Premiers ministres)

Modalité de variable
(partition en 37 locuteurs)

Lexique :

- ✓Segmentation
- ✓Reconnaissance
- ✓Lemmatisation
- ✓Statuts statistiques

Nombre d'occurrences:	239802
Nombre de formes:	7455
Hapax:	2659



•Nombre d'occurrences

L'analyse factorielle des correspondances

L'AFC s'effectue sur des variables nominales (qualitatives).
La corrélation entre les points-lignes et entre les points-colonnes est calculée par la distance du chi².

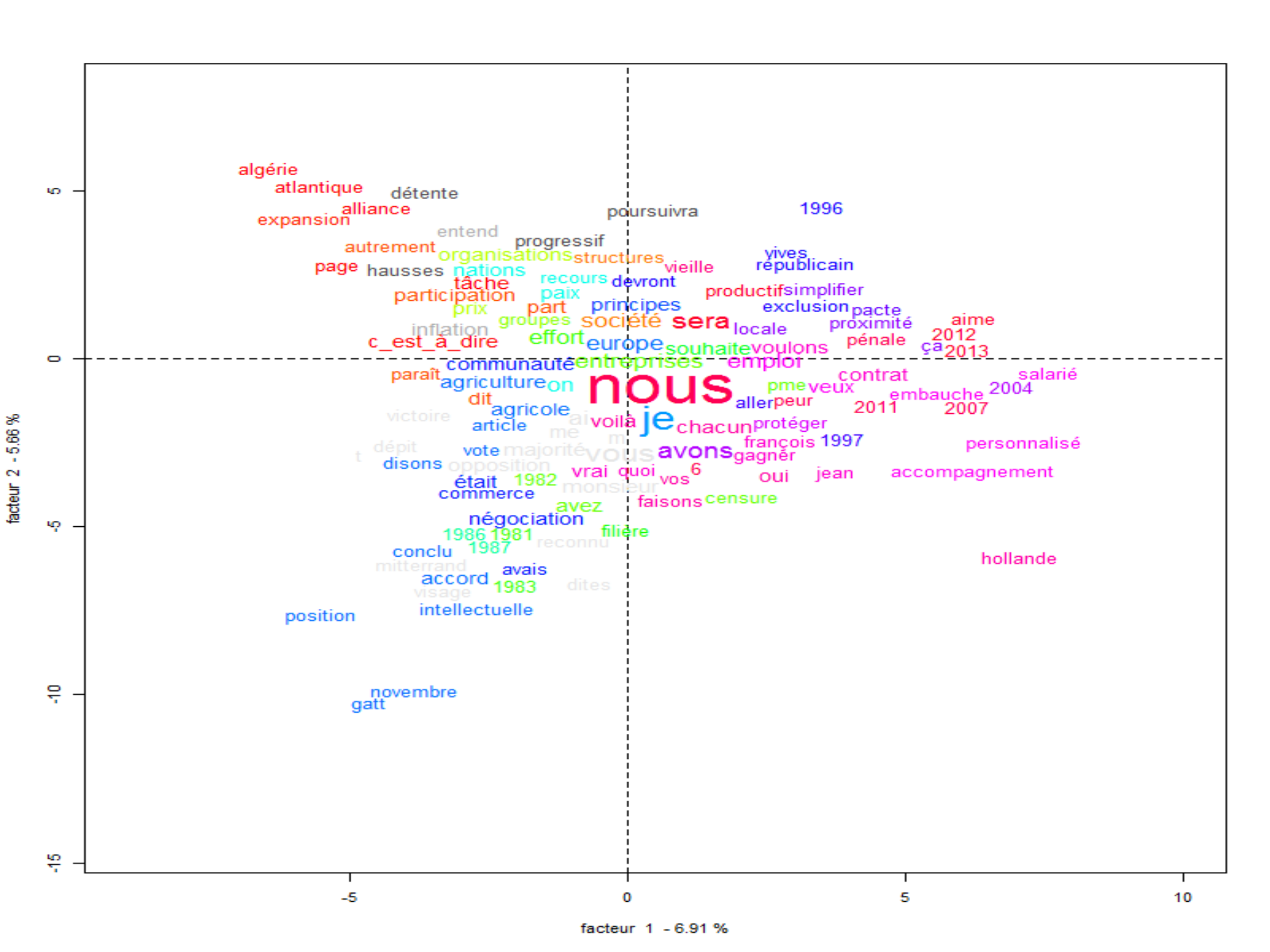
Distance entre points-lignes

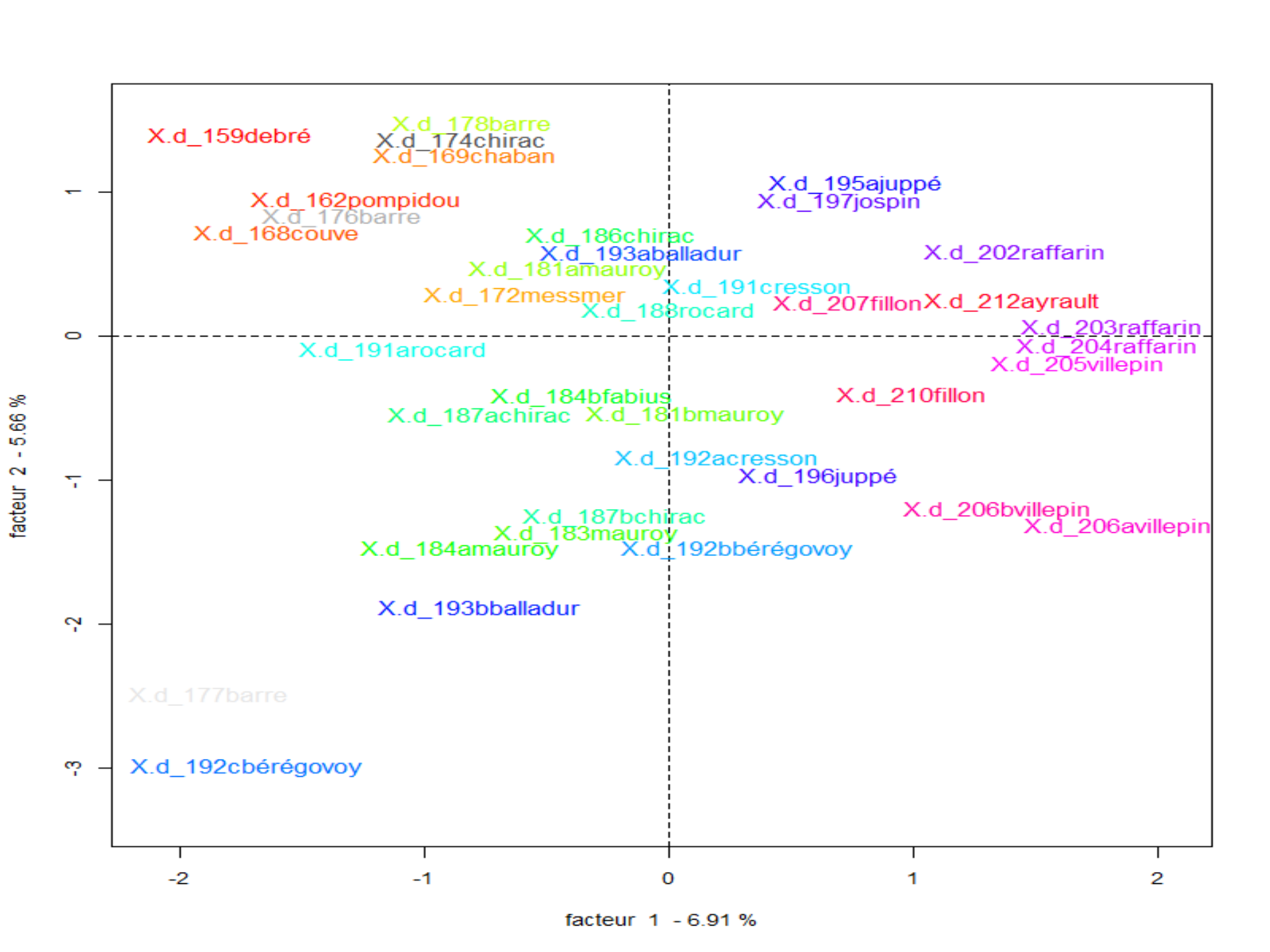
$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{1}{f_{.j}} \right) \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2$$

Distances entre points-colonnes

$$d^2(j,j') = \sum_{i=1}^{n} \left(\frac{1}{f_{i.}} \right) \left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}} \right)^2$$

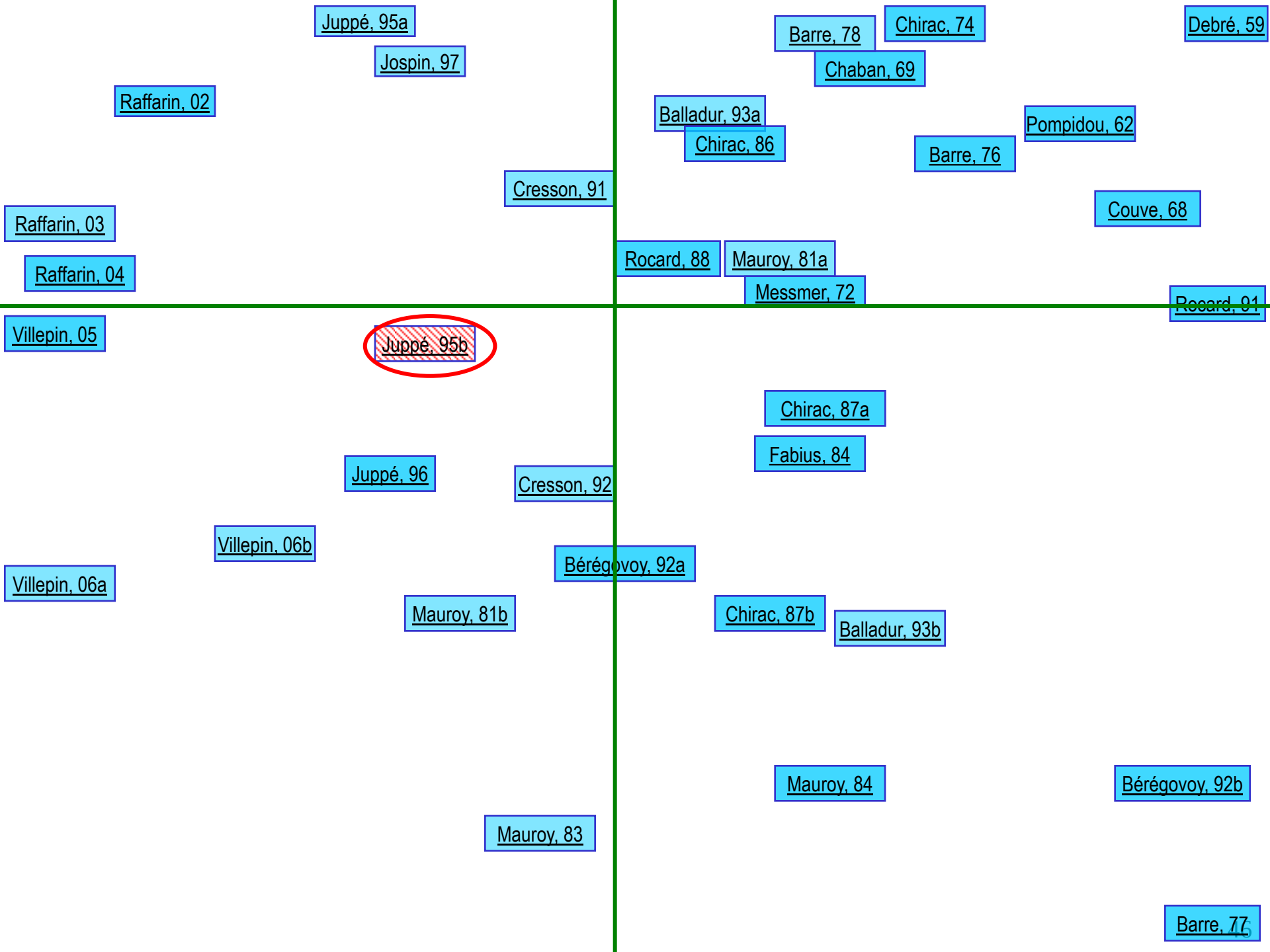
Le principe repose sur la décomposition du tableau de données: On montre qu'un tel tableau peut être décomposé en d'autres tableaux autant de fois que le plus petit des nombres de ses lignes et de ses colonnes.
Il y a donc autant de facteurs que le plus petit des nombres de ses lignes et de ses colonnes

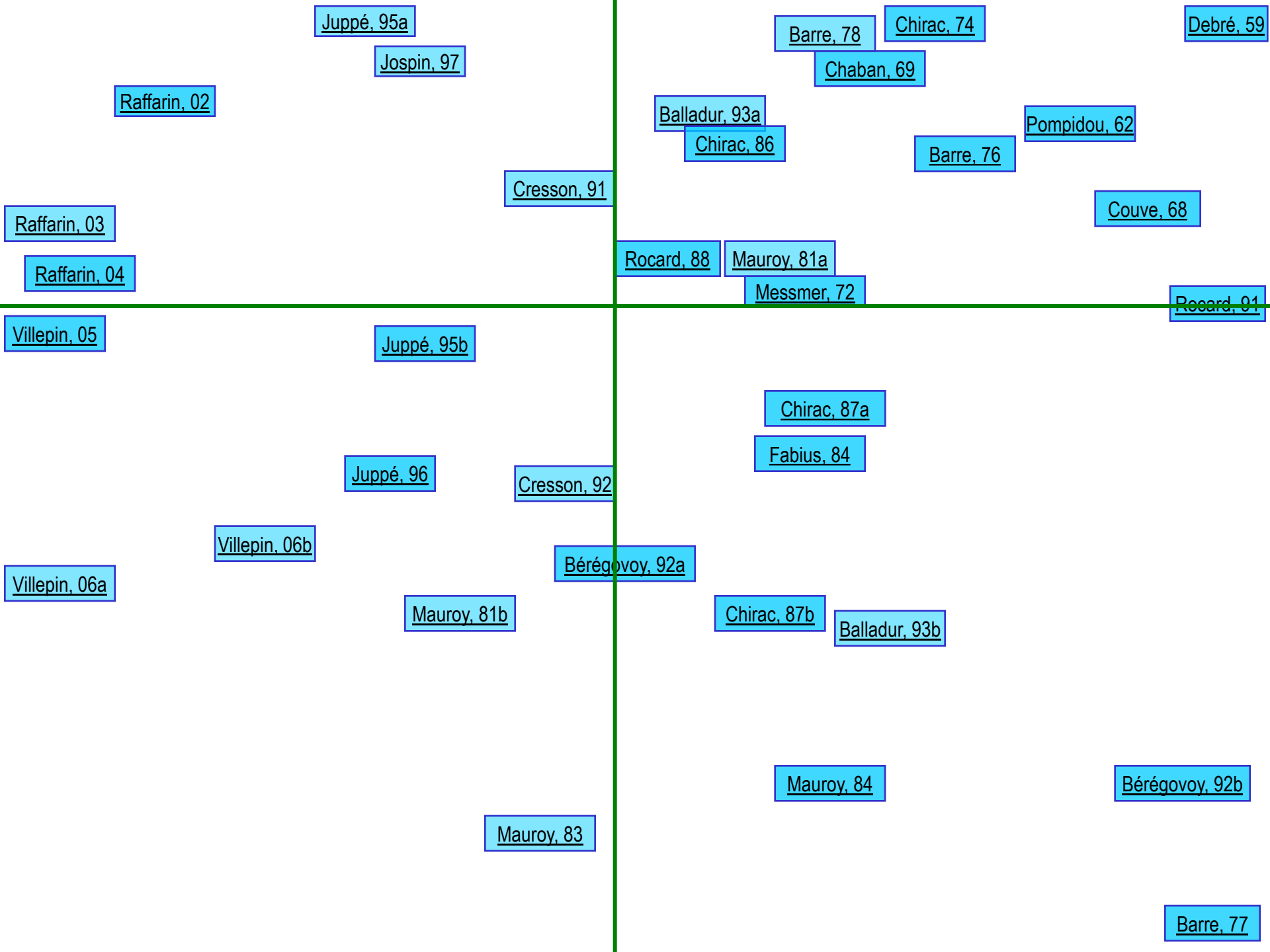

algérie
atlantique
détente
alliance
expansion
poursuivra
1996
entend
progressif
autrement
organisations
structures
vieille
yves
page
hausses
nations
recours
devront
républicain
tâche
participation
paix
productif
simplifier
prix
part
principes
exclusion
pacte
inflation
groupes
société
sera
locale
proximité
aime
c_est_à_dire
effort
europe
souhaite
voulons
pénale
ça
2012
2013
communauté
entreprises
emploi
paraît
agriculture
on
nous
pme
veux
contrat
salarié
dit
aller
peur
embauche
2004
agricole
victoire
article
voilà
je
chacun
protéger
2011
2007
vote
majorité
vous
avons
françois
1997
gagner
personnalisé
dépit
disons
opposition
vrai
quoi
vos
6
oui
jean
accompagnement
était
1982
monsieur
commerce
faisons
censure
avez
négociation
filière
1986
1981
1987
reconnu
conclu
hollande
mitterrand
avais
accord
1983
dites
visage
intellectuelle
position
novembre
gatt
facteur 2 - 5.66 %
facteur 1 - 6.91 %
5
0
-5
-10
-15
-5
0
5
10


X.d_159debré
X.d_178barre
X.d_174chirac
X.d_169chaban
X.d_195ajuppé
X.d_197jospin
X.d_162pompidou
X.d_176barre
X.d_168couve
X.d_186chirac
X.d_193aballadur
X.d_202raffarin
X.d_181amauroy
X.d_191cresson
X.d_172messmer
X.d_207fillon
X.d_212ayrault
X.d_188rocard
X.d_203raffarin
X.d_204raffarin
X.d_191arocard
X.d_205villepin
X.d_184bfabius
X.d_210fillon
X.d_187achirac
X.d_181bmauroy
X.d_192acresson
X.d_196juppé
X.d_206bvillepin
X.d_187bchirac
X.d_206avillepin
X.d_183mauroy
X.d_184amauroy
X.d_192bbérégovoy
X.d_193bballadur
X.d_177barre
X.d_192cbérégovoy
facteur 2 - 5.66 %
facteur 1 - 6.91 %
-2
-1
0
1
2
-3
-2
-1
0
1

Quelques considérations interprétatives

- La Recherche improbable d'une homogène diversité
  - Lexico3


Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Juppé, 95b
Juppé, 96
Mauroy, 81b
Mauroy, 83
Cresson, 91
Cresson, 92
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Bérégovoy, 92a
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77


Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06a
Villepin, 06b
Juppé, 96
Juppé, 95b
Mauroy, 81b
Mauroy, 83
Cresson, 91
Cresson, 92
Bérégovoy, 92a
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77

Quelques considérations interprétatives

- Décrire les axes par les contributions

Juppé, 95a

Jospin, 97

Raffarin, 02

Chirac, 87b

Balladur, 93b

Mauroy, 84

Bérégovoy, 92b

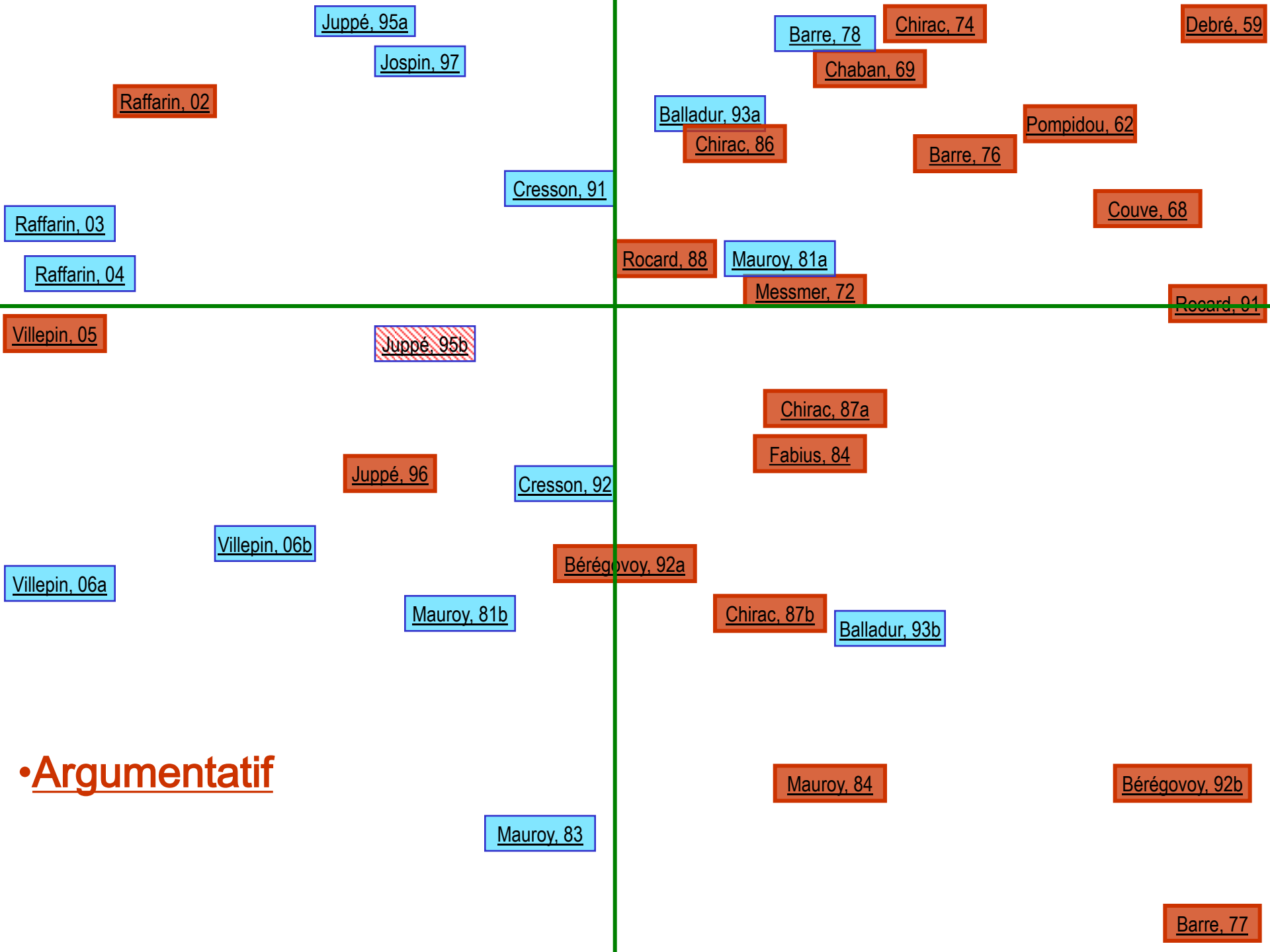
Barre, 77

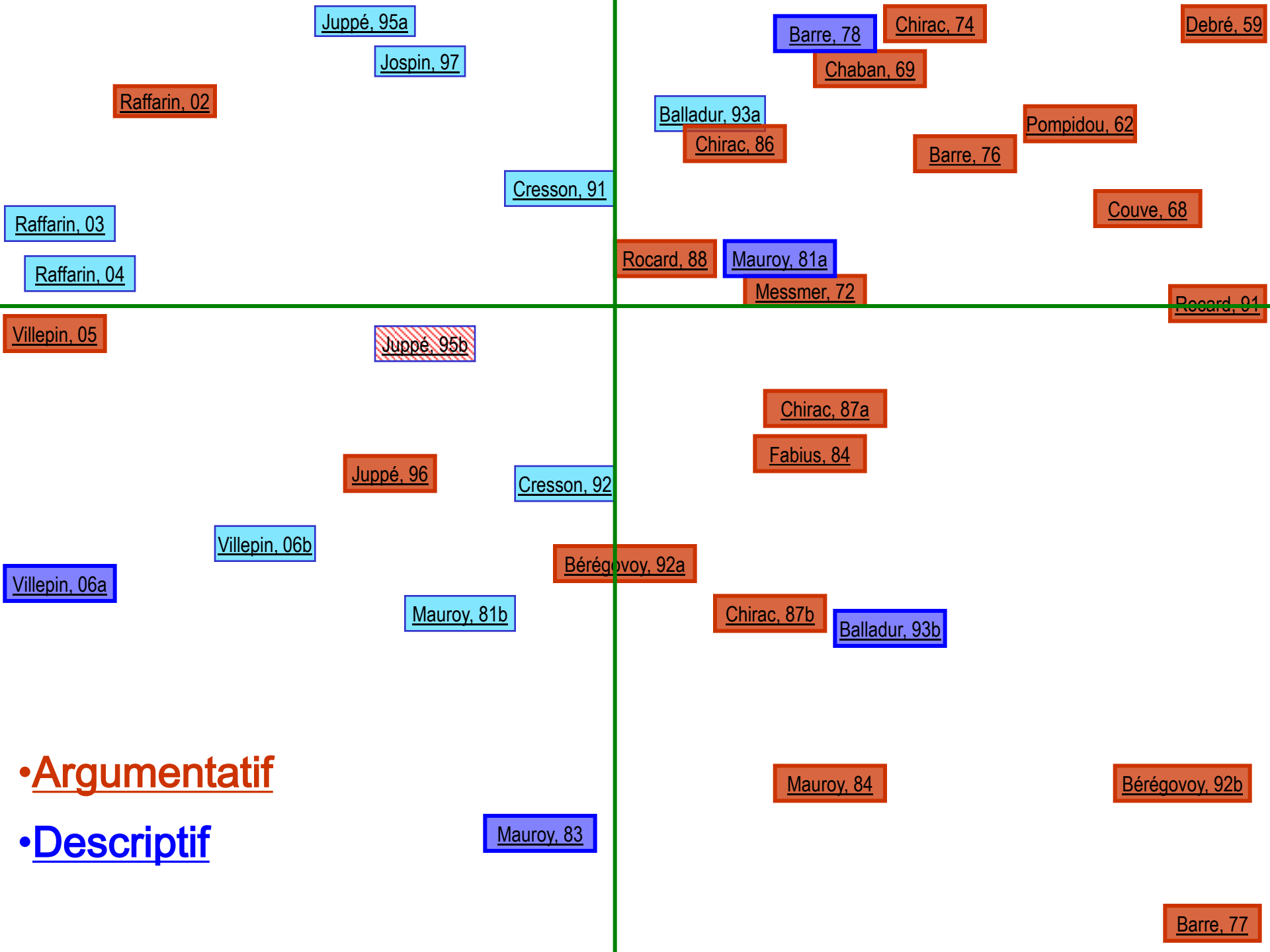
Contributions « négatives » :
Personnalisé, salarié, parcours, euros, accompagnement, aime, immobilisme, mondialisation, embauche, ça, attentes, rupture, confrontés, enfant, discriminations, précarité, métiers, soins, proximité, valoriser, dépense, cohérence, atouts…

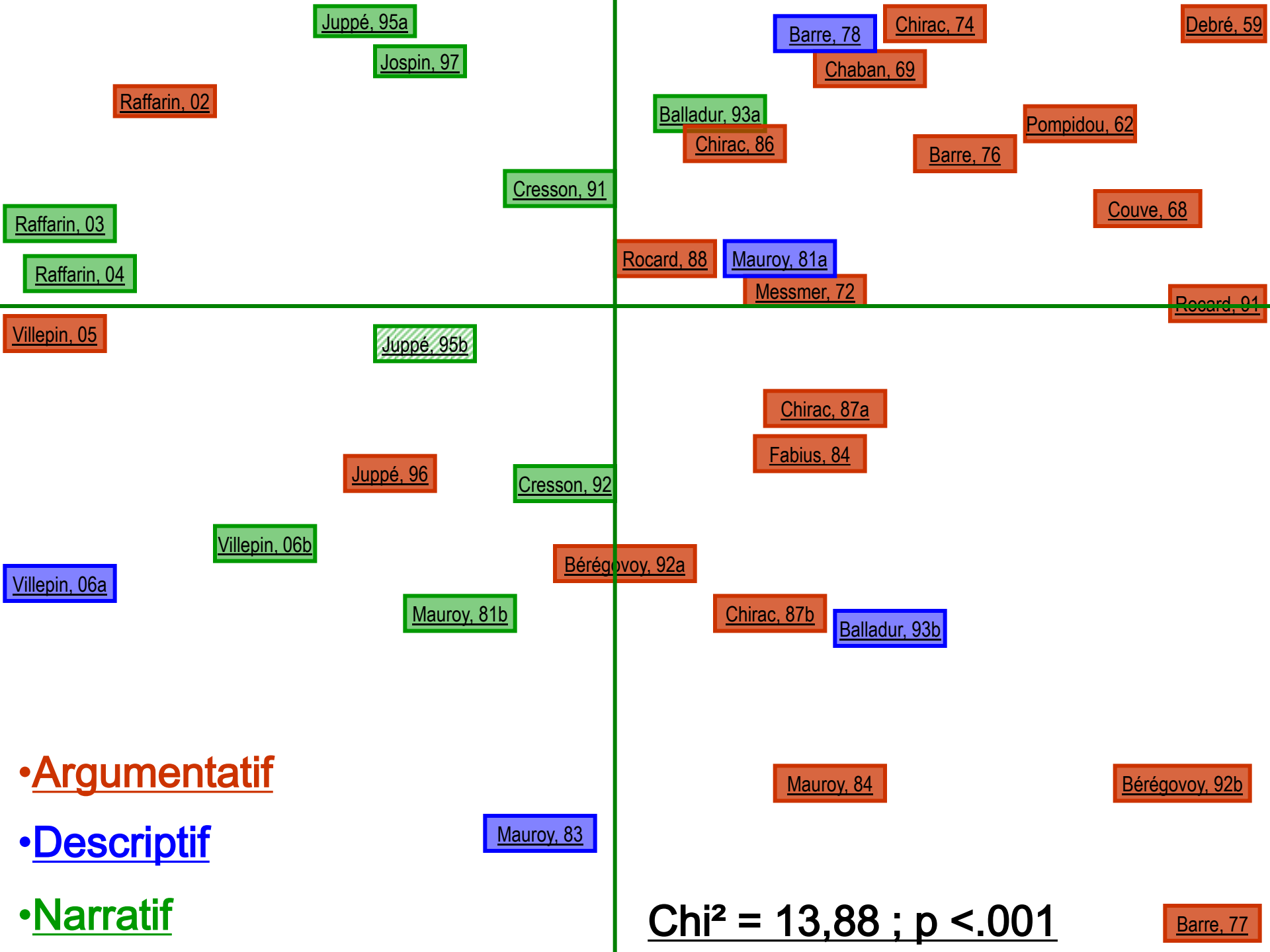
Contributions « positives » :
Algérie, expansion, atlantique, position, page, résolu, méthodes, compréhension, collaboration, réactions, GATT, alliance, misère, évoqué, établissement, foi, isolement, autrement, diverses, approuver, dépit, cependant, détente, nul, hausses, régler, exposer, revendications…

Quelques considérations interprétatives

- Décrire les axes par le style (*Tropes*)

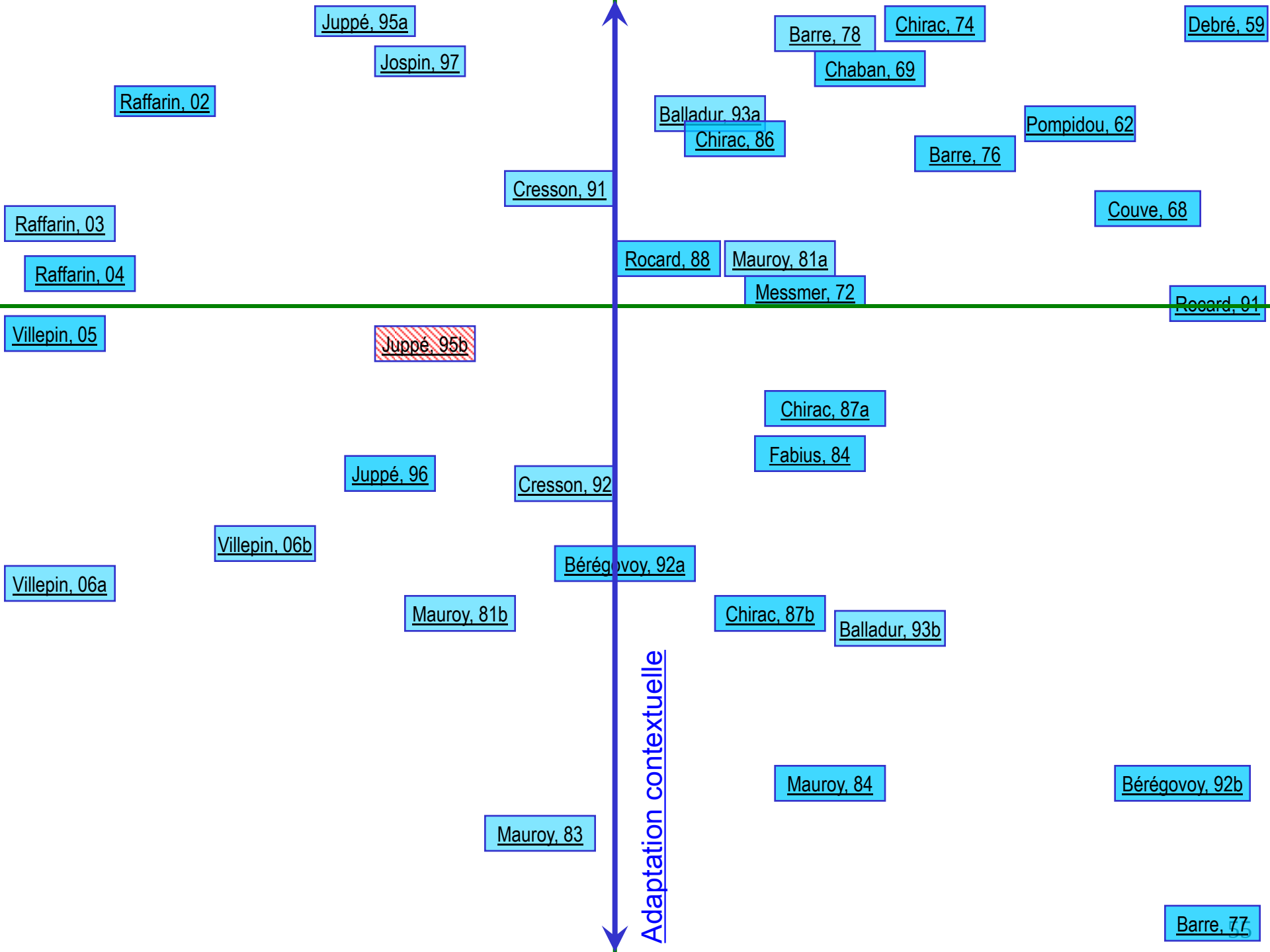

Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Mauroy, 81b
Mauroy, 83
Cresson, 91
Juppé, 95b
Juppé, 96
Cresson, 92
Bérégovoy, 92a
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77
•Argumentatif

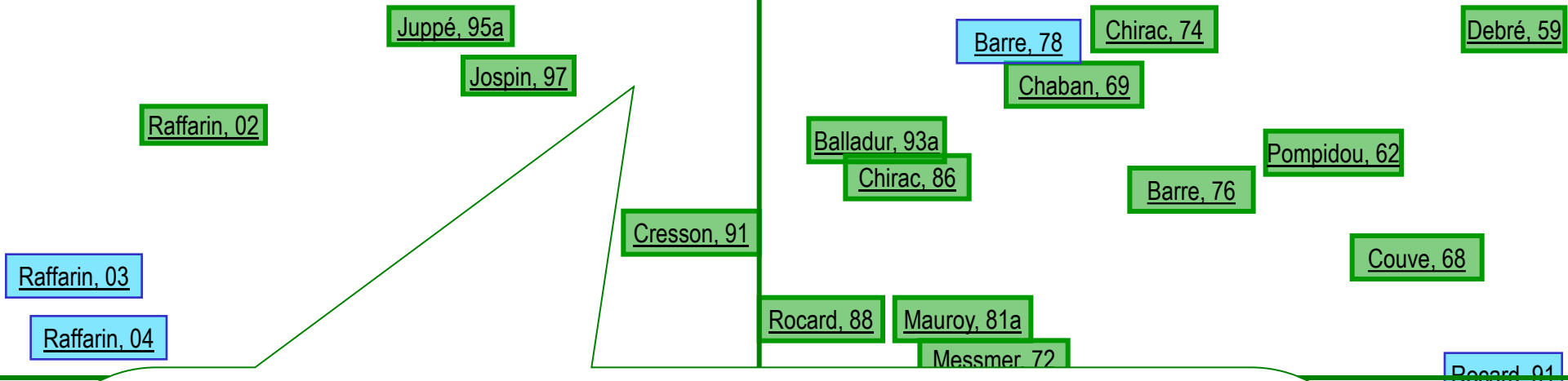

Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Juppé, 95b
Juppé, 96
Cresson, 91
Cresson, 92
Mauroy, 81b
Mauroy, 83
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Bérégovoy, 92a
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77
•Argumentatif
•Descriptif


Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Juppé, 95b
Juppé, 96
Cresson, 91
Cresson, 92
Mauroy, 81b
Mauroy, 83
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Bérégovoy, 92a
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77
•Argumentatif
•Descriptif
•Narratif
Chi² = 13,88 ; p <.001

Quelques considérations interprétatives

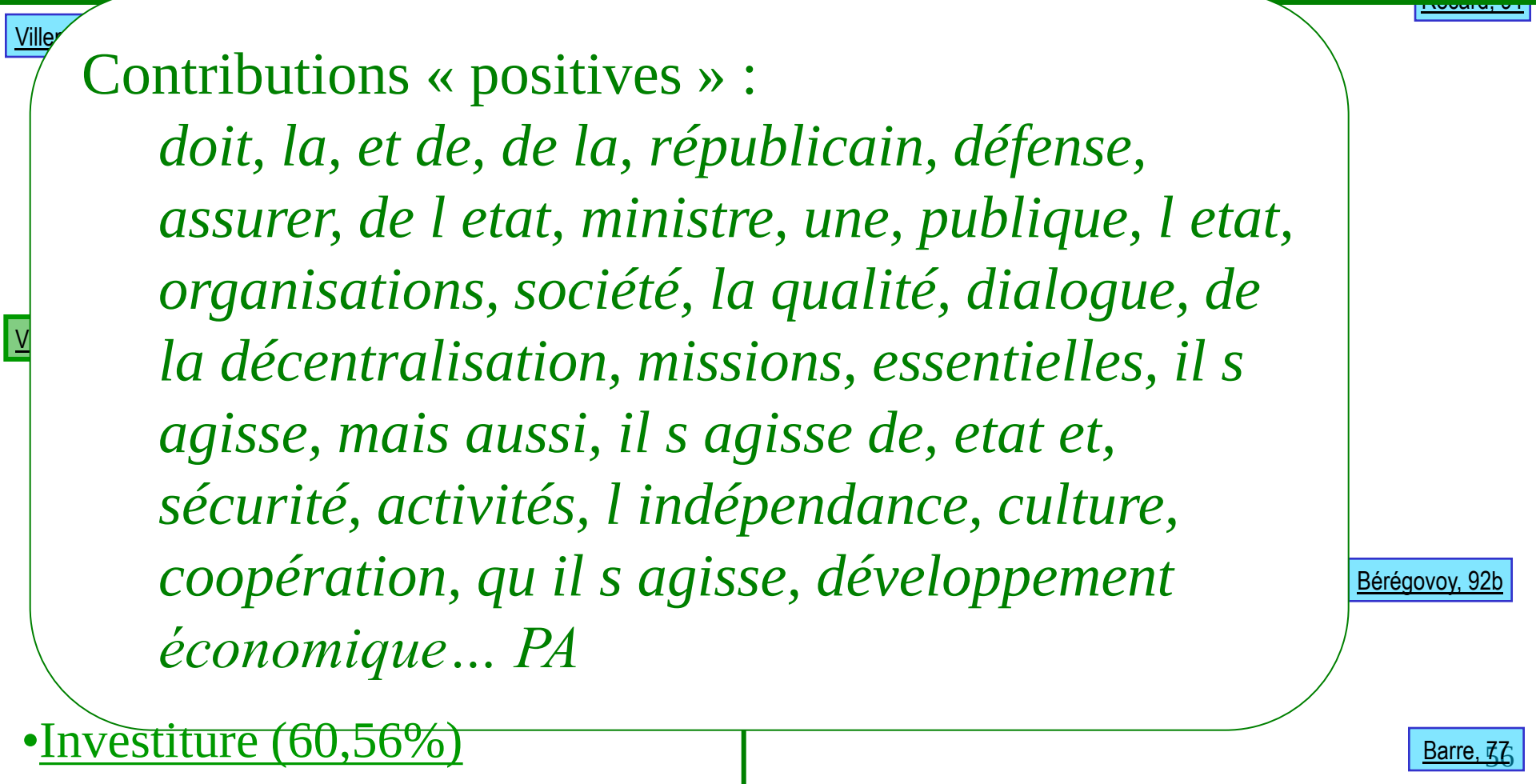
- Décrire les axes par le contexte


Juppé, 95a
Jospin, 97
Raffarin, 02
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Cresson, 91
Couve, 68
Raffarin, 03
Raffarin, 04
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Villepin, 05
Juppé, 95b
Chirac, 87a
Fabius, 84
Juppé, 96
Cresson, 92
Villepin, 06b
Bérégovoy, 92a
Villepin, 06a
Mauroy, 81b
Chirac, 87b
Balladur, 93b
Adaptation contextuelle
Mauroy, 84
Bérégovoy, 92b
Mauroy, 83
Barre, 77

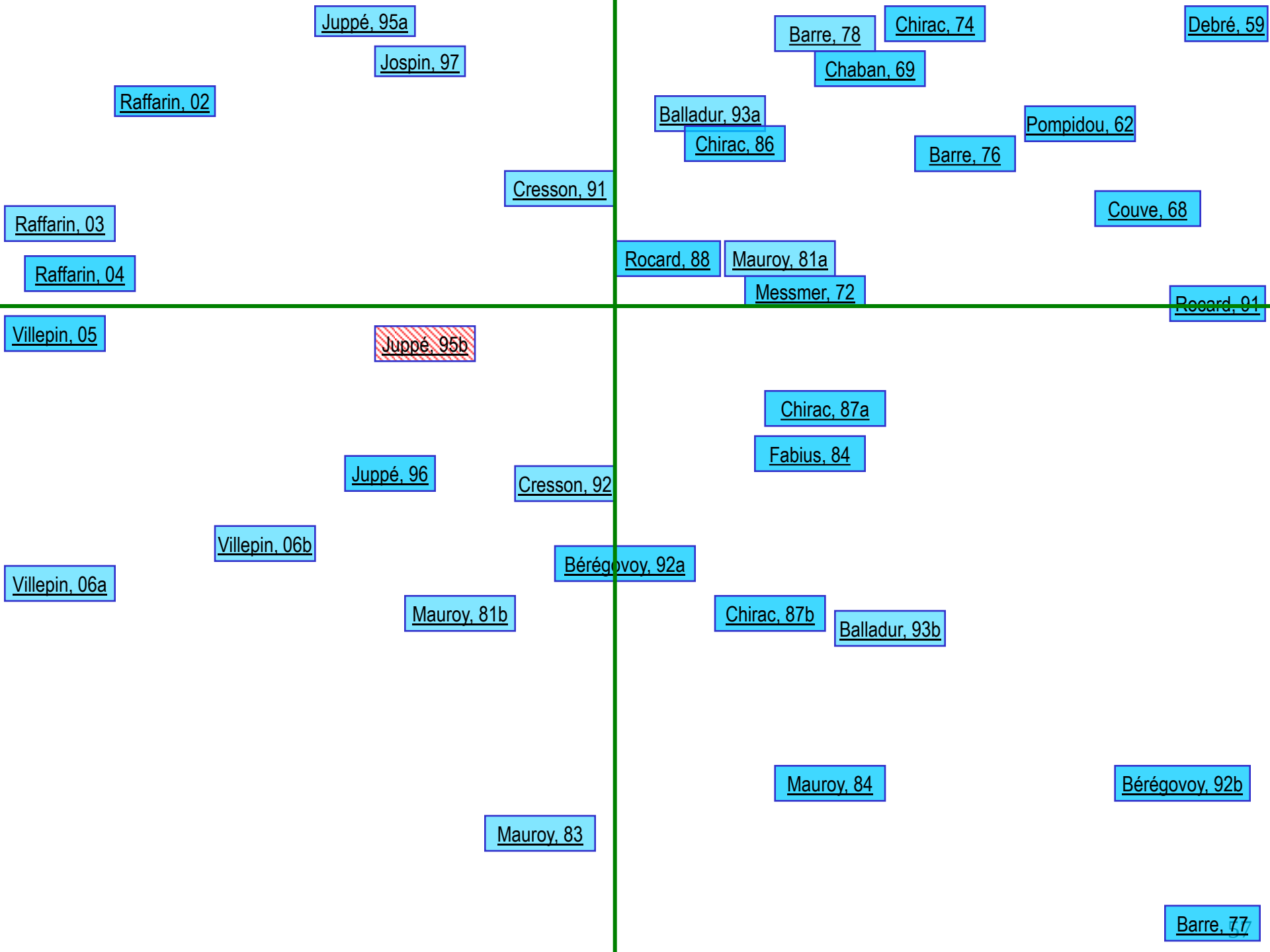


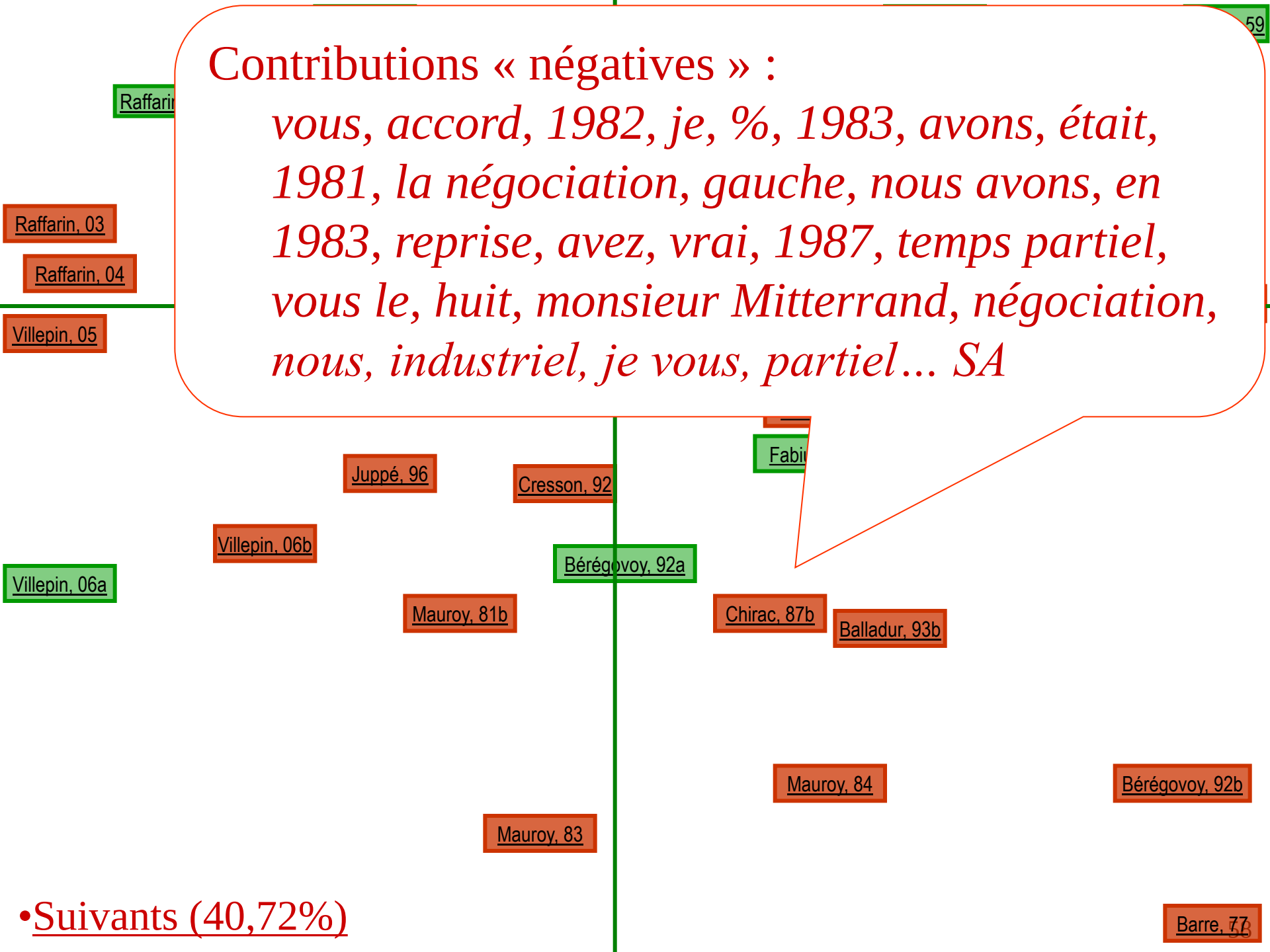
Contributions « positives » :

doit, la, et de, de la, républicain, défense, assurer, de l etat, ministre, une, publique, l etat, organisations, société, la qualité, dialogue, de la décentralisation, missions, essentielles, il s agisse, mais aussi, il s agisse de, etat et, sécurité, activités, l indépendance, culture, coopération, qu il s agisse, développement économique… PA



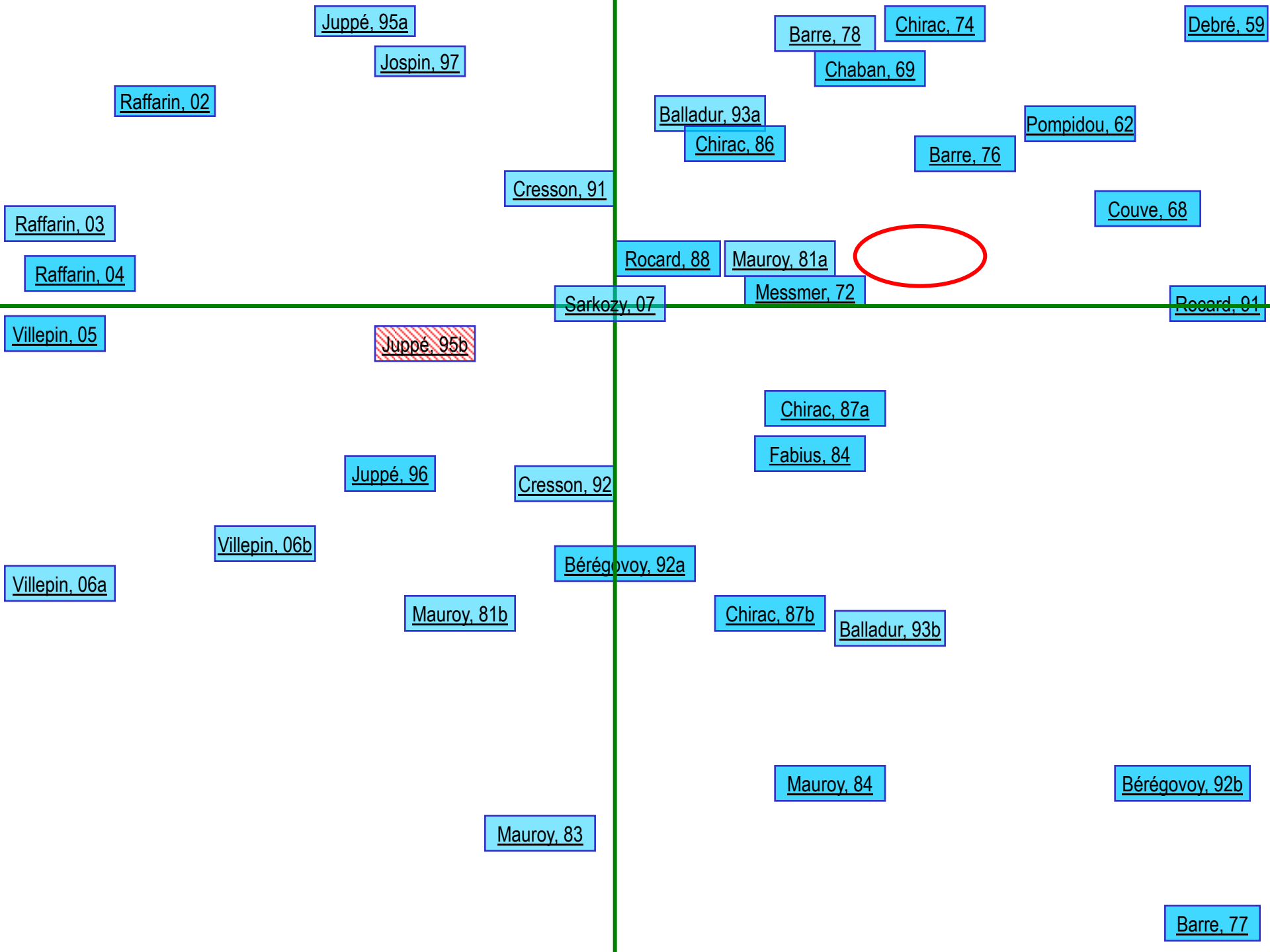
- Investiture (60,56%)

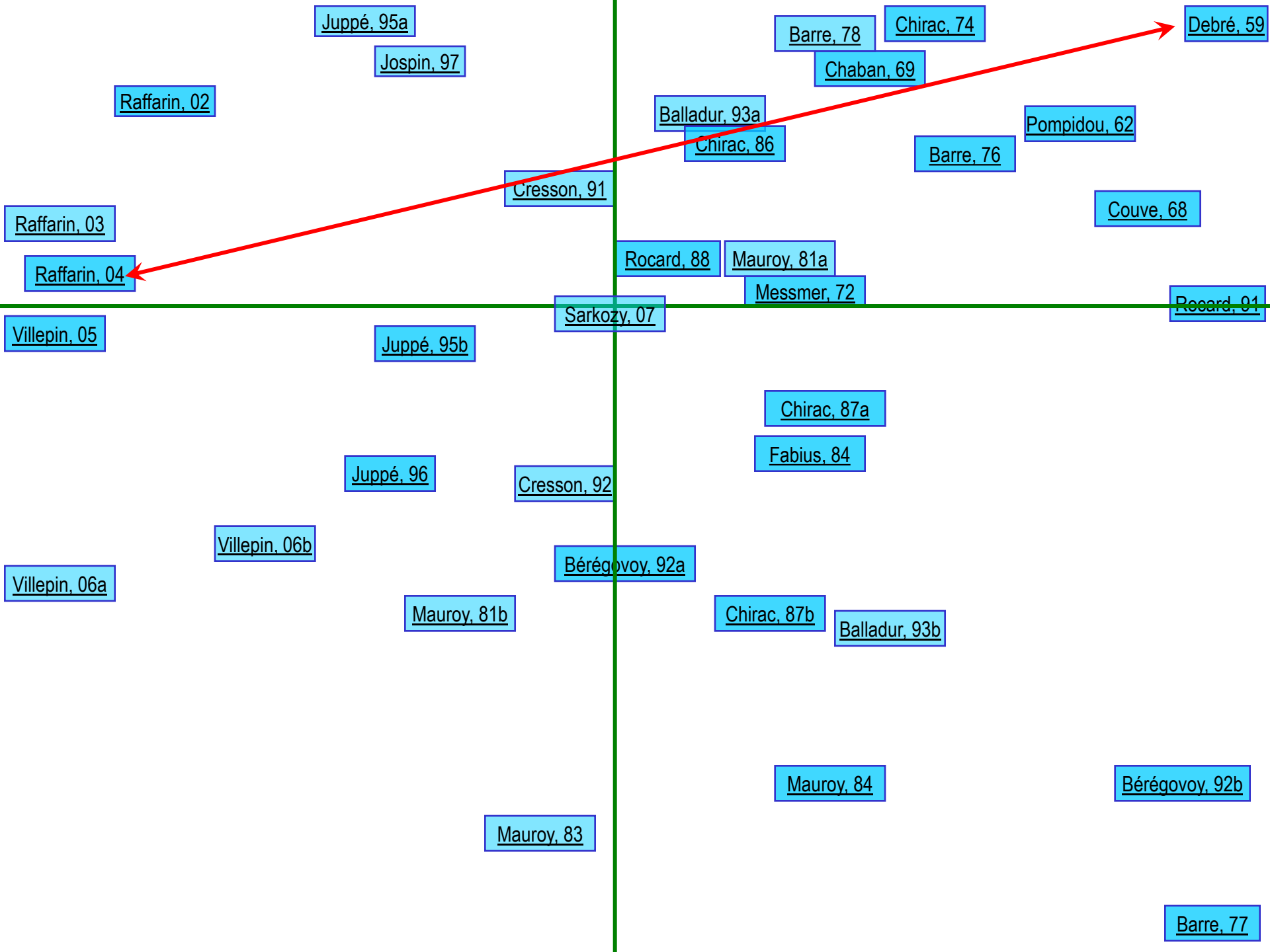

Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Juppé, 95b
Juppé, 96
Mauroy, 81b
Mauroy, 83
Cresson, 91
Cresson, 92
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Bérégovoy, 92a
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77


Contributions « négatives » :
vous, accord, 1982, je, %, 1983, avons, était, 1981, la négociation, gauche, nous avons, en 1983, reprise, avez, vrai, 1987, temps partiel, vous le, huit, monsieur Mitterrand, négociation, nous, industriel, je vous, partiel… SA
59
Raffari
Raffarin, 03
Raffarin, 04
Villepin, 05
Fabi
Juppé, 96
Cresson, 92
Villepin, 06b
Bérégovoy, 92a
Villepin, 06a
Mauroy, 81b
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Mauroy, 83
Barre, 77
•Suivants (40,72%)

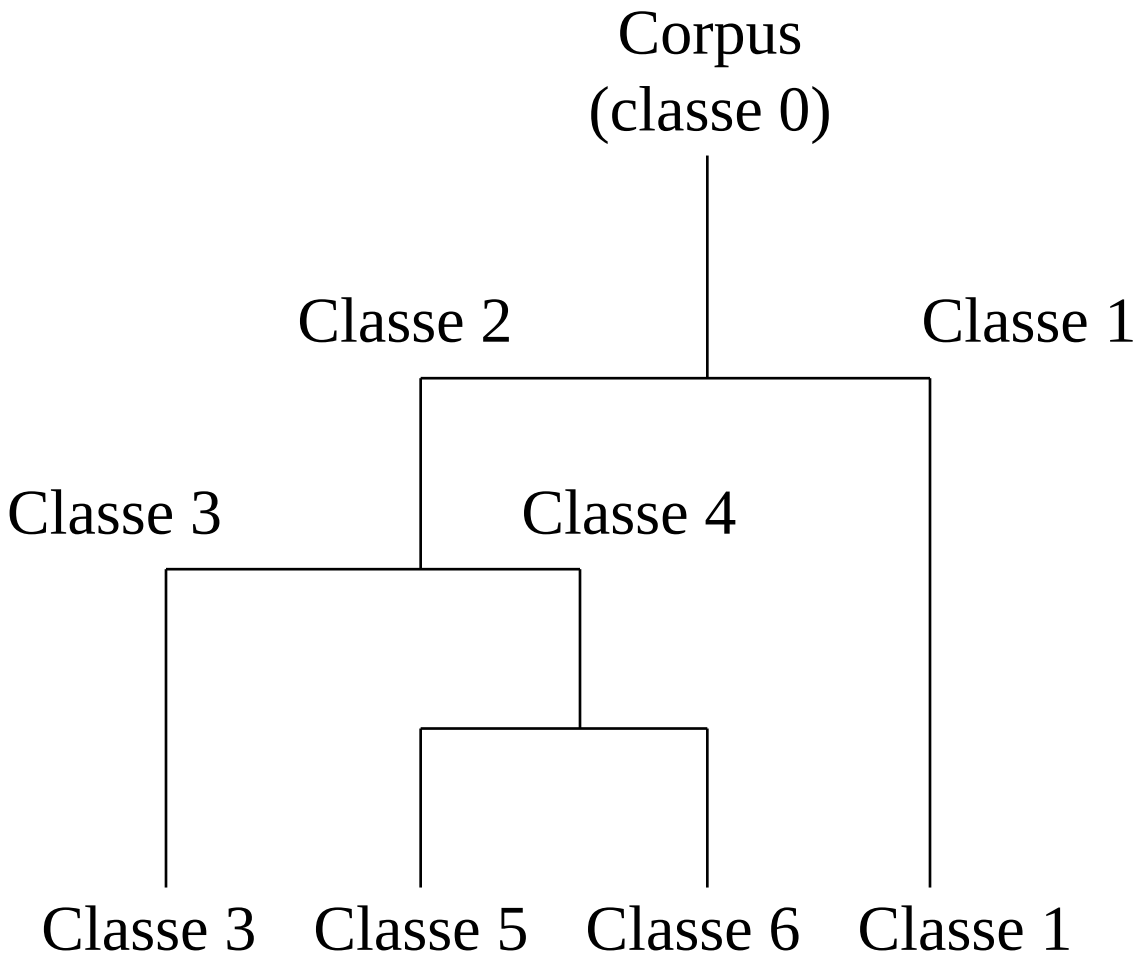
Quelques considérations interprétatives

- Retour sur l'improbable homogène diversité
  - L'effet Sarkozy (« feuille de route », 2007)…


Juppé, 95a
Jospin, 97
Raffarin, 02
Raffarin, 03
Raffarin, 04
Villepin, 05
Villepin, 06b
Villepin, 06a
Juppé, 95b
Juppé, 96
Mauroy, 81b
Mauroy, 83
Cresson, 91
Cresson, 92
Sarkozy, 07
Bérégovoy, 92a
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Couve, 68
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Chirac, 87a
Fabius, 84
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Barre, 77


Juppé, 95a
Jospin, 97
Raffarin, 02
Barre, 78
Chirac, 74
Debré, 59
Chaban, 69
Balladur, 93a
Chirac, 86
Pompidou, 62
Barre, 76
Cresson, 91
Couve, 68
Raffarin, 03
Raffarin, 04
Rocard, 88
Mauroy, 81a
Messmer, 72
Rocard, 91
Sarkozy, 07
Villepin, 05
Juppé, 95b
Chirac, 87a
Fabius, 84
Juppé, 96
Cresson, 92
Villepin, 06b
Bérégovoy, 92a
Villepin, 06a
Mauroy, 81b
Chirac, 87b
Balladur, 93b
Mauroy, 84
Bérégovoy, 92b
Mauroy, 83
Barre, 77

La classification lexicale



Pour vous, qu'est-ce qu'être français - Grand débat sur l'identité nationale - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://www.debatidentitenationale.fr/participation/pour-vous-qu-est-ce-qu-etre/

Pour vous, qu'est-ce qu'être français ...

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Grand débat sur l'identité nationale

Liberté
Egalité
Fraternité

Le grand débat | Actualités | Ils prennent position | Participer au forum | Vidéothèque | Vu d'ailleurs | Textes de référence | Espace presse

Accueil > Pour vous, qu'est-ce qu'être français

Pour vous, qu'est-ce qu'être français

Le site de débat sur l'identité nationale s'enrichit et met à votre disposition un forum.
Vous pouvez apporter vos analyses et vos propositions sur les sujets qui vous semblent les plus importants, et réagir dans le même temps aux contributions des autres internautes.
Nous continuerons à vous proposer de nouveaux thèmes de discussion en fonction de vos messages.
Le moteur de recherche ci-dessous vous permet d'accéder à des contributions qui vous permettront d'entrer dans le débat.
Merci à tous pour vos analyses et vos témoignages. Bonne lecture et bonnes contributions.

Postez votre contribution

Pour vous qu'est-ce qu'être Français ?

Vos suggestions autour du débat

Votre nom / pseudo : rechercher

diclofenac
02/12/09 à 18:19

Bonjour,
Je m'appelle Mohamed et suis comme on se plait à le dire, issu de l'immigration.

Oui l'immigration de mon papa venu en France dans les années 70 pour travailler et retourner par la suite dans ses montagnes Marocaines.

Ce que papa a oublié, c'est qu'il a fini par aimer la France, aimer sa justice, sa générosité, sa beauté et a cru bon de s'y installer définitivement et d'y fonder une famille.

Maman est alors arrivée quittant à son tour ses montagnes Marocaines pour rejoindre son bien aimé.

« Pour vous qu'est-ce qu'être Français ? »

Postez votre contribution

DERNIERES VIDEOS

LES GRANDS THEMES DU DEBAT

Terminé

Corpus

- Importation (manuelle) des contributions consultables sur le site du *Ministère de l'Immigration, de l'intégration, de l'identité nationale et du développement solidaire:* http://www.debatidentitenationale.fr/participation/pour-vous-qu-est-ce-qu-etre/
- 18.240 contributions du 2/11/09 à 7:44 au 2/12/09 à 18:19
- Taille moyenne de 103 mots par contribution (de 1 à près de 5000 mots)

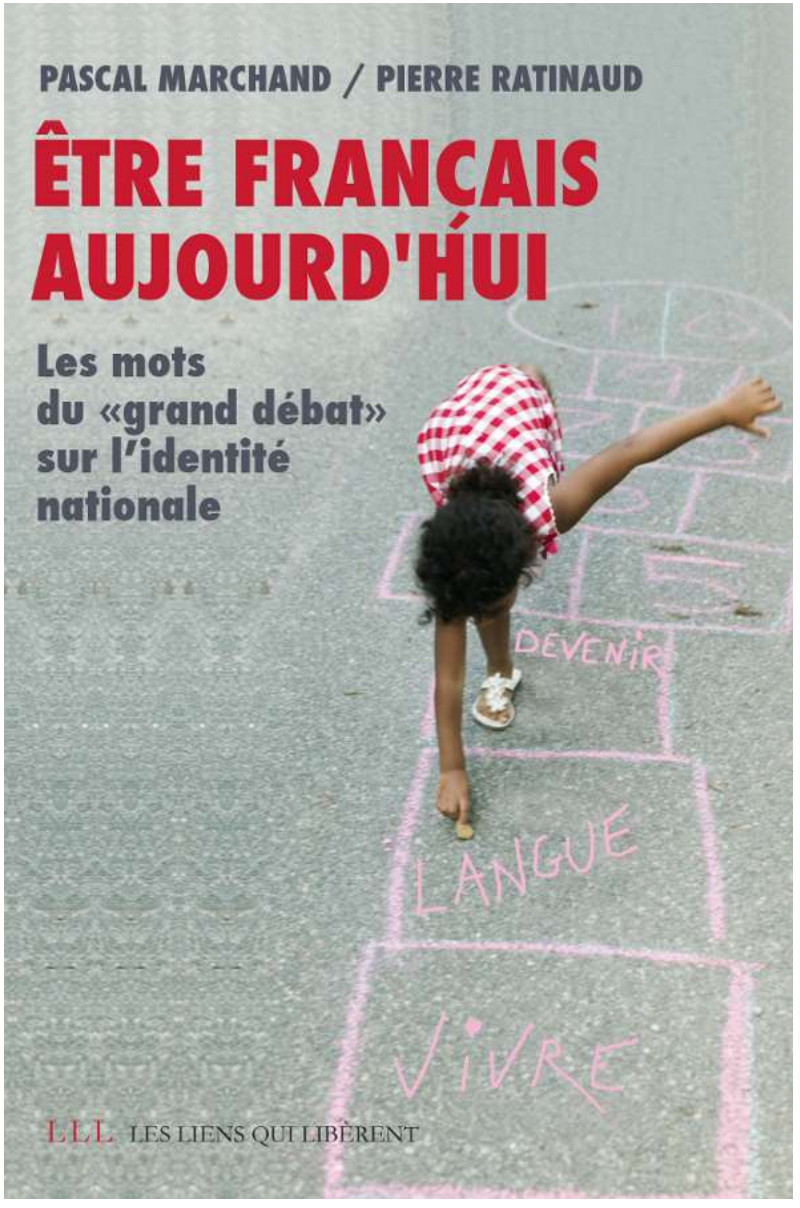


Tableau lexical

Partition (contributions ou paragraphes)

Lexique

- **Segmentation**
- **Reconnaissance**
- **Lemmatisation**

Nombre d'occurrences:
Nombre de formes:
Nombre de lemmes :

Normalisation de la saisie:

- **caractères (accentuation, majuscules)**
  - ne/né
- **frappe et orthographe**
  - assez peu d'écriture « SMS » et d'abréviations (bcp, bjr, cpdt, ds, gd, grd, ke, ki, kes, *ps*, qd, qq, qlq, pb, tjs, ts, tt…)
  - *idenbtité, idendite, idendité, idenditées, idenité, identite, identité, idéntité, identitéde, identitee, identitée, identitèe, identitées, identiter, identites, identités, identitéss, identitié, identitier, identitité, indenté, indentite, indentité, indentités…*
  - *Burqa*
  - l'égalité / légalité

CDH (Reinert, 1983)

Lexique

- **Segmentation**
- **Reconnaissance**
- **Lemmatisation**
- **Statut statistique**

Partition (articles ou segments de textes)

0	0	1	1	0	1
0	0	1	1	0	1
0	0	1	1	0	1
1	0	0	1	1	0
0	0	1	1	0	1
1	1	0	0	1	0
1	1	0	0	1	0
0	0	1	1	0	1
0	0	1	1	0	1
1	1	0	0	1	0
1	1	0	0	1	0
1	1	0	0	0	1
0	0	1	1	0	1

CDH (Reinert, 1983)

Lexique

- **Segmentation**
- **Reconnaissance**
- **Lemmatisation**
- **Statut statistique**

Partition (articles ou segments de textes)

0	0	1	1	0	1
0	0	1	1	0	1
0	0	1	1	0	1
1	1	1	1	1	0
0	0	1	1	0	1
1	1	0	0	1	0
1	1	0	0	1	0
0	0	1	1	0	1
0	0	1	1	0	1
1	1	0	0	1	0
1	0	0	1	1	0
1	1	0	0	0	1
0	0	1	1	0	1

Classification lexicale (méthode ALCESTE)

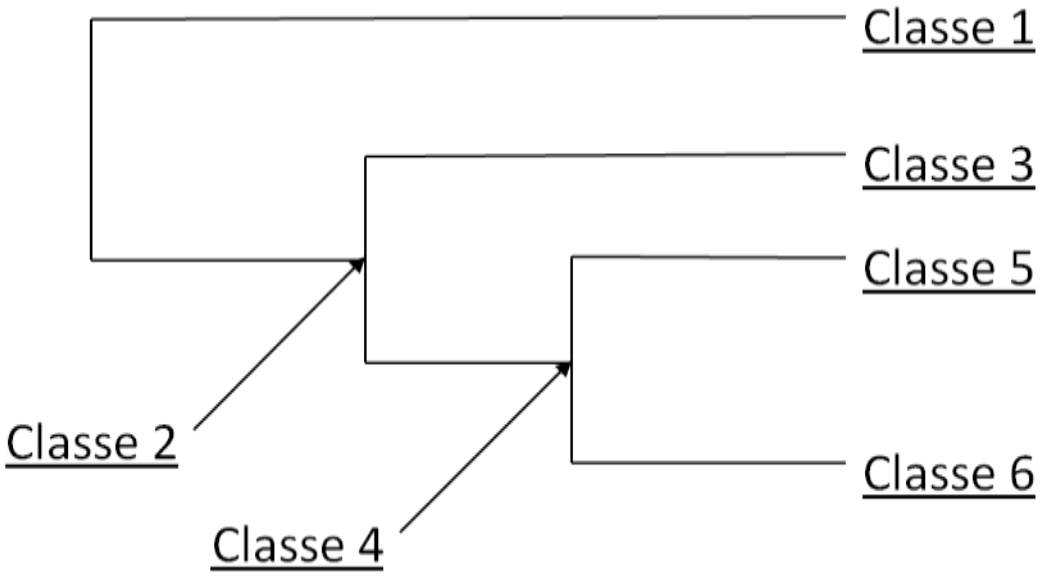
id	a	b	c	d	e	f	g	h	i
1	1	1	1	1	0	0	0	0	0
6	1	1	1	1	0	0	0	0	1
8	1	0	1	0	1	0	0	0	0
4	1	0	1	0	1	0	0	0	1
9	0	0	1	0	1	0	1	0	1
3	0	0	1	0	1	0	1	0	0
5	0	0	1	0	1	0	1	0	0
10	0	0	1	0	1	0	1	0	0
2	0	0	0	0	1	1	1	1	1
7	0	0	0	0	1	1	1	1	0

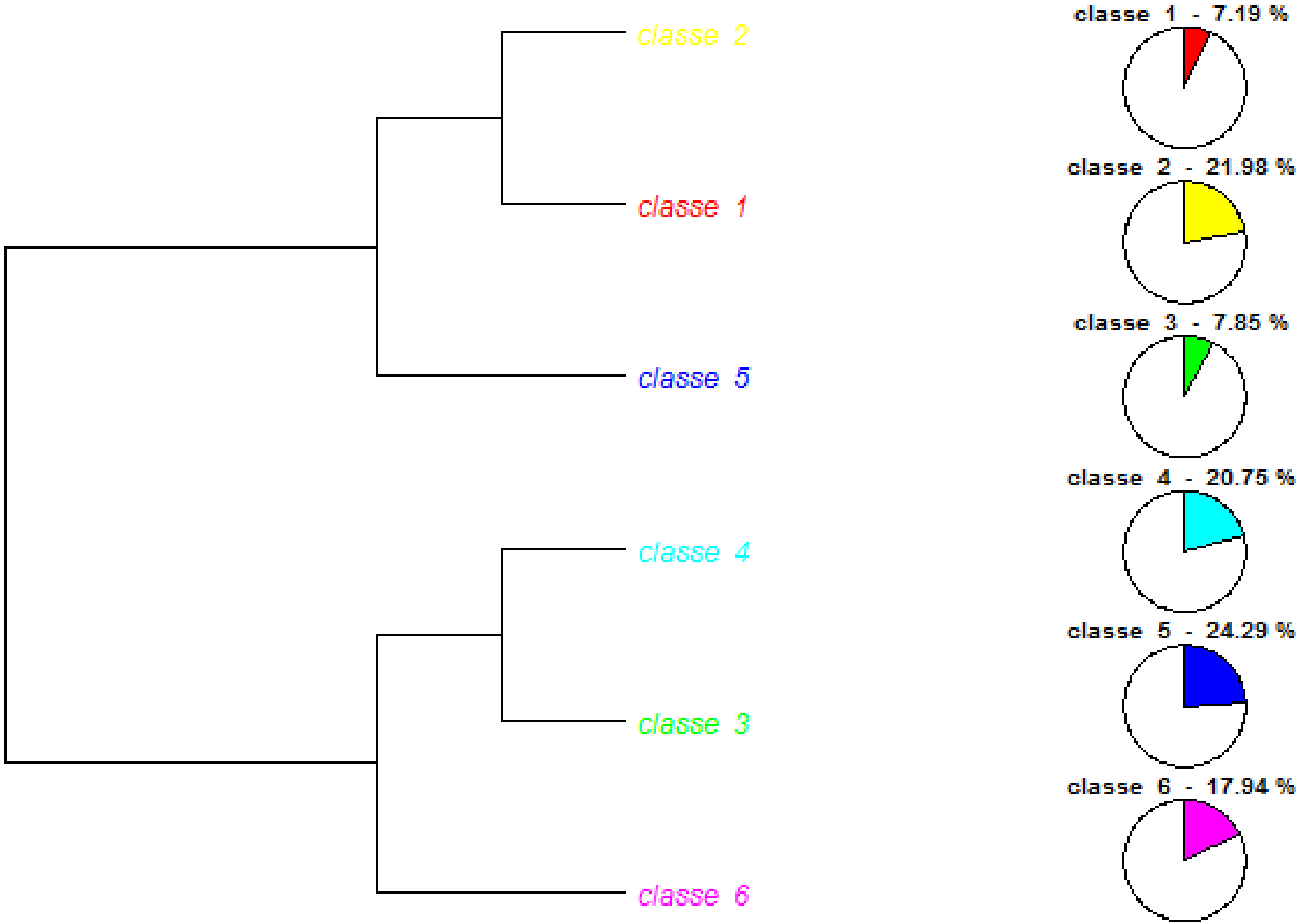
Classe 1

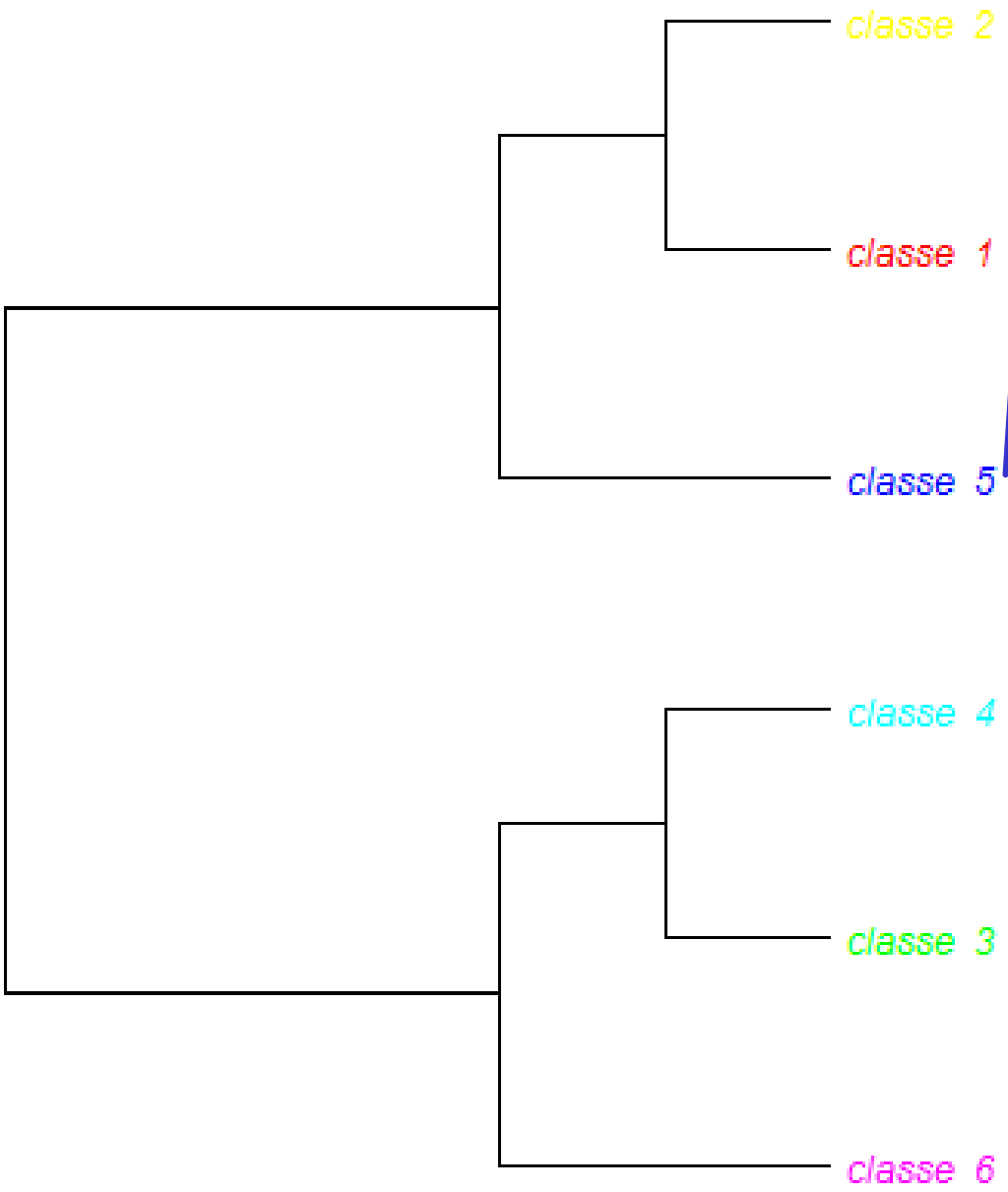
Classe 6

Classe 5

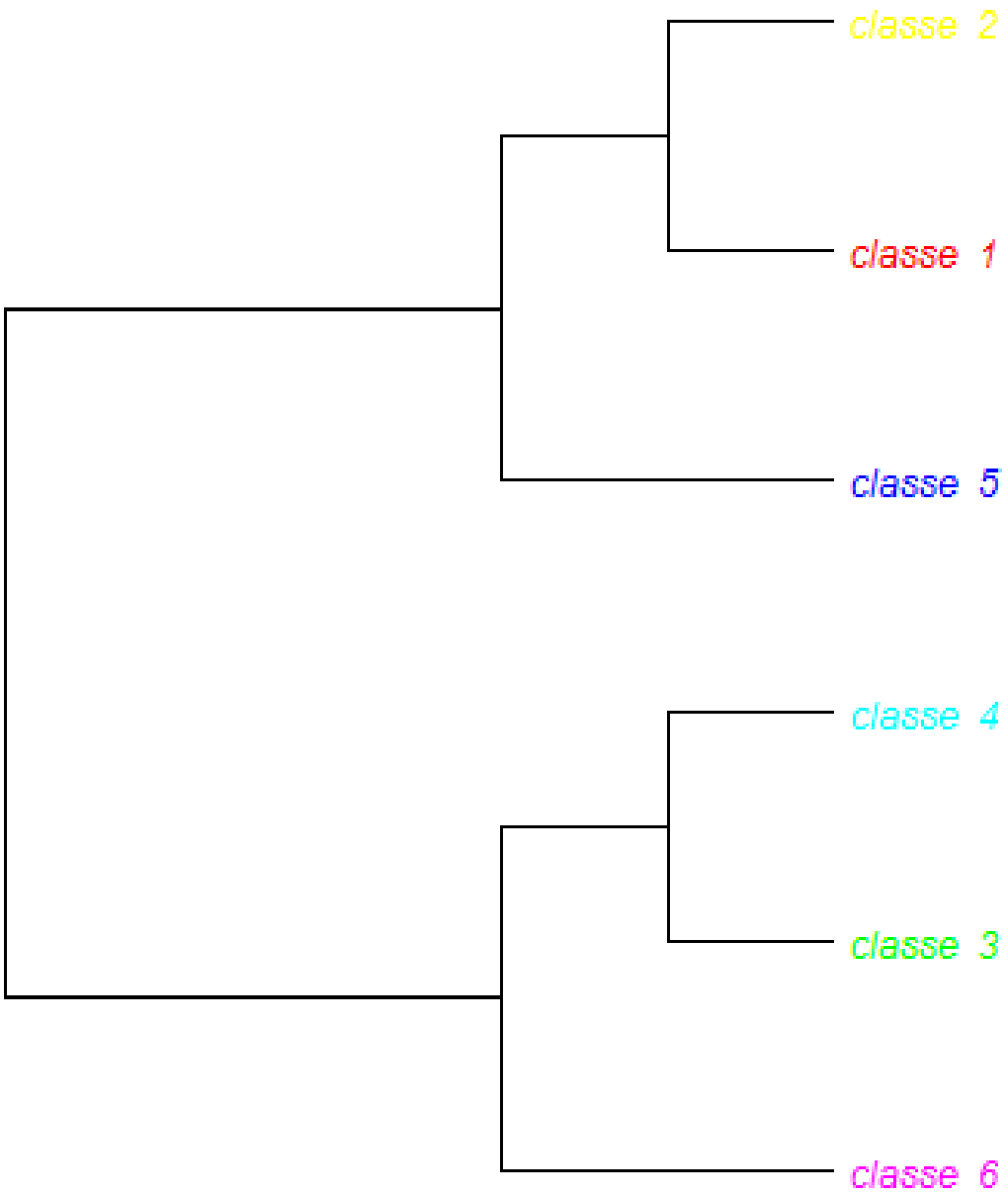
Classe 3



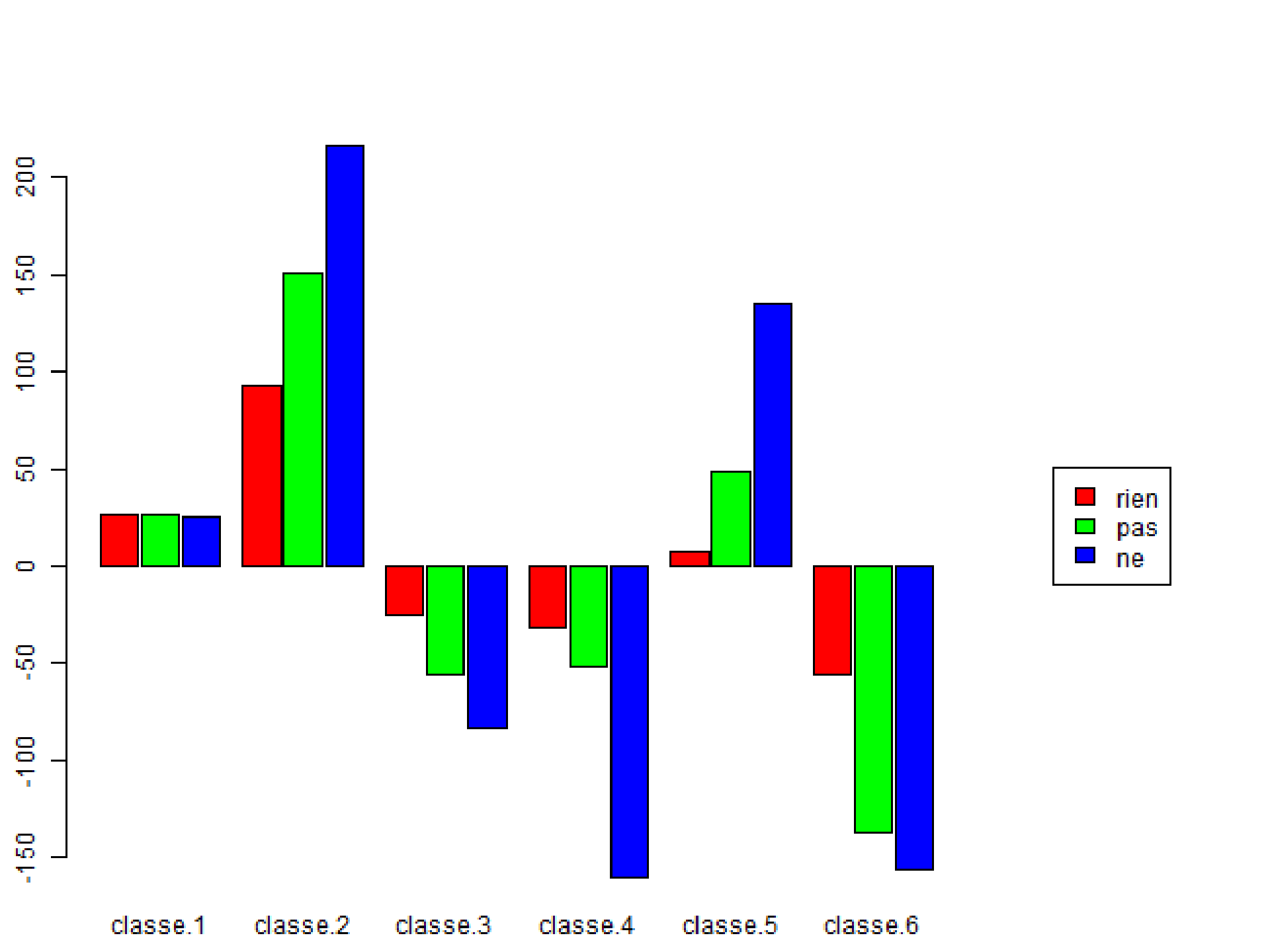

classe 2
classe 1
classe 5
classe 4
classe 3
classe 6
classe 1 - 7.19 %
classe 2 - 21.98 %
classe 3 - 7.85 %
classe 4 - 20.75 %
classe 5 - 24.29 %
classe 6 - 17.94 %

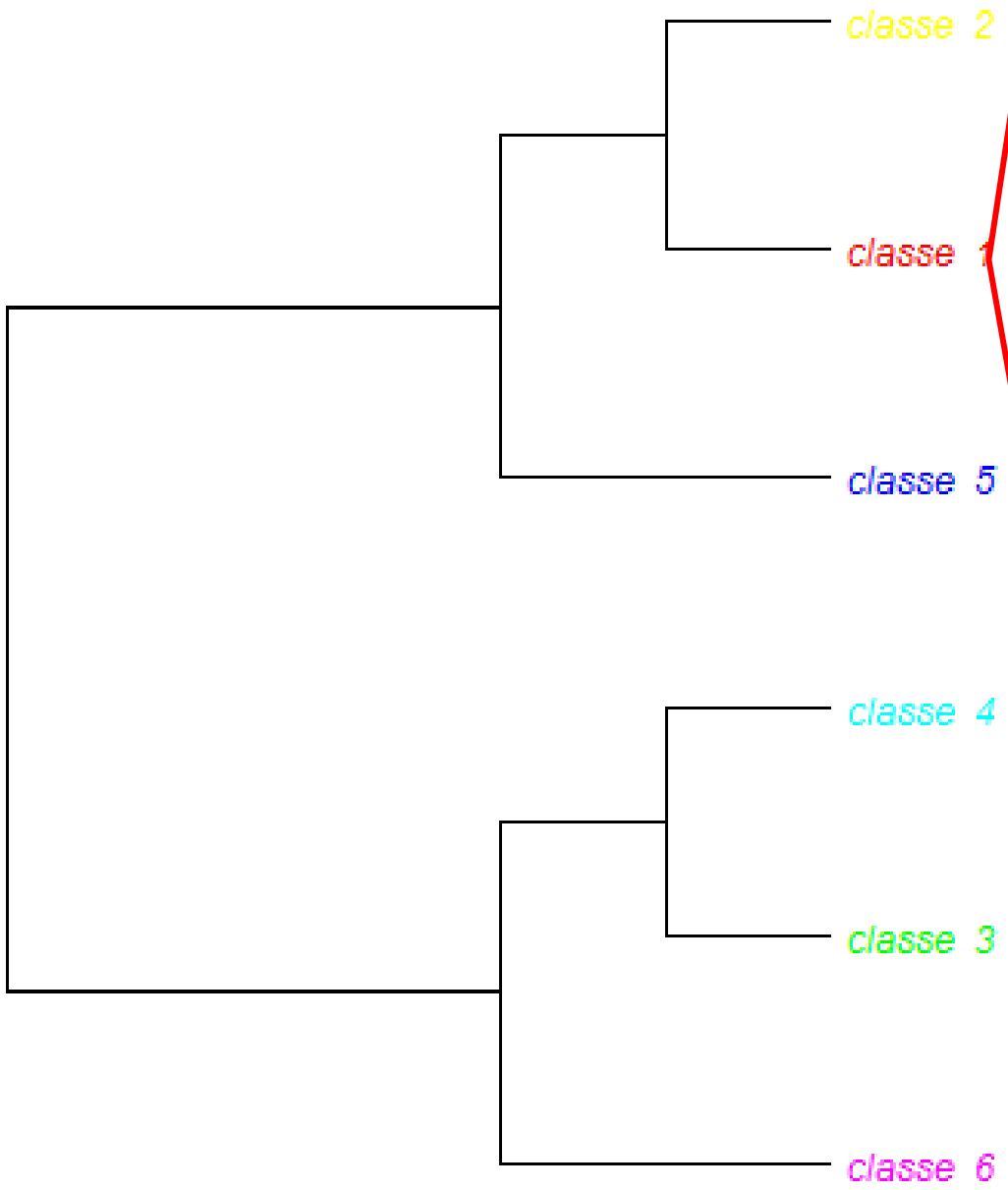


identité, définir, national, européen, nation, débat, définition, question, humain, mondialisation, individu, sentiment, sembler, siècle, commun, notion, terme, ensemble, culturel, groupe, exister, peuple, frontière, concept, projet, seul, appartenance, construire, dépasser, constituer, collectif, diversité, humanité, présent, propre, individuel, sentir, vision, politique, plutôt, est_ce, communauté, identitaire, nationalisme, long, construction, idée, monde, aujourd_hui, population…



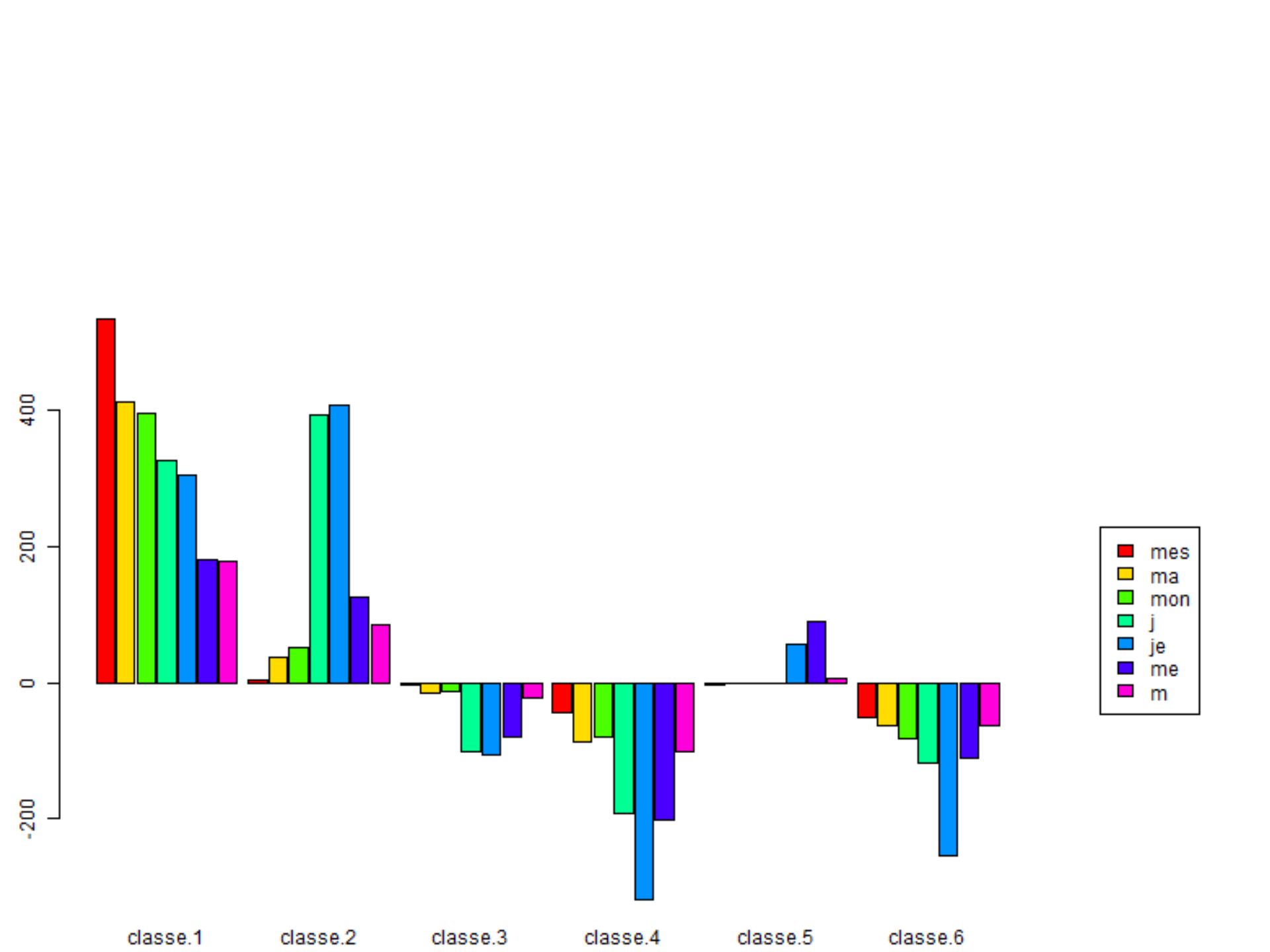
honte, voir, débat, gouvernement, gens, besson, question, aller, droite, argent, payer, poser, renvoyer, gauche, président, vrai, chômage, lancer, équipe, peur, main, faire, couper, jeune, laisser, foot, match, très, problème, ministre, vraiment, jour, chose, genre, expulser, raciste, gagner, penser, guerre, chanson, espérer, manger, rendre, trouver, écouter, compte, demander, chanter, étranger, prendre...

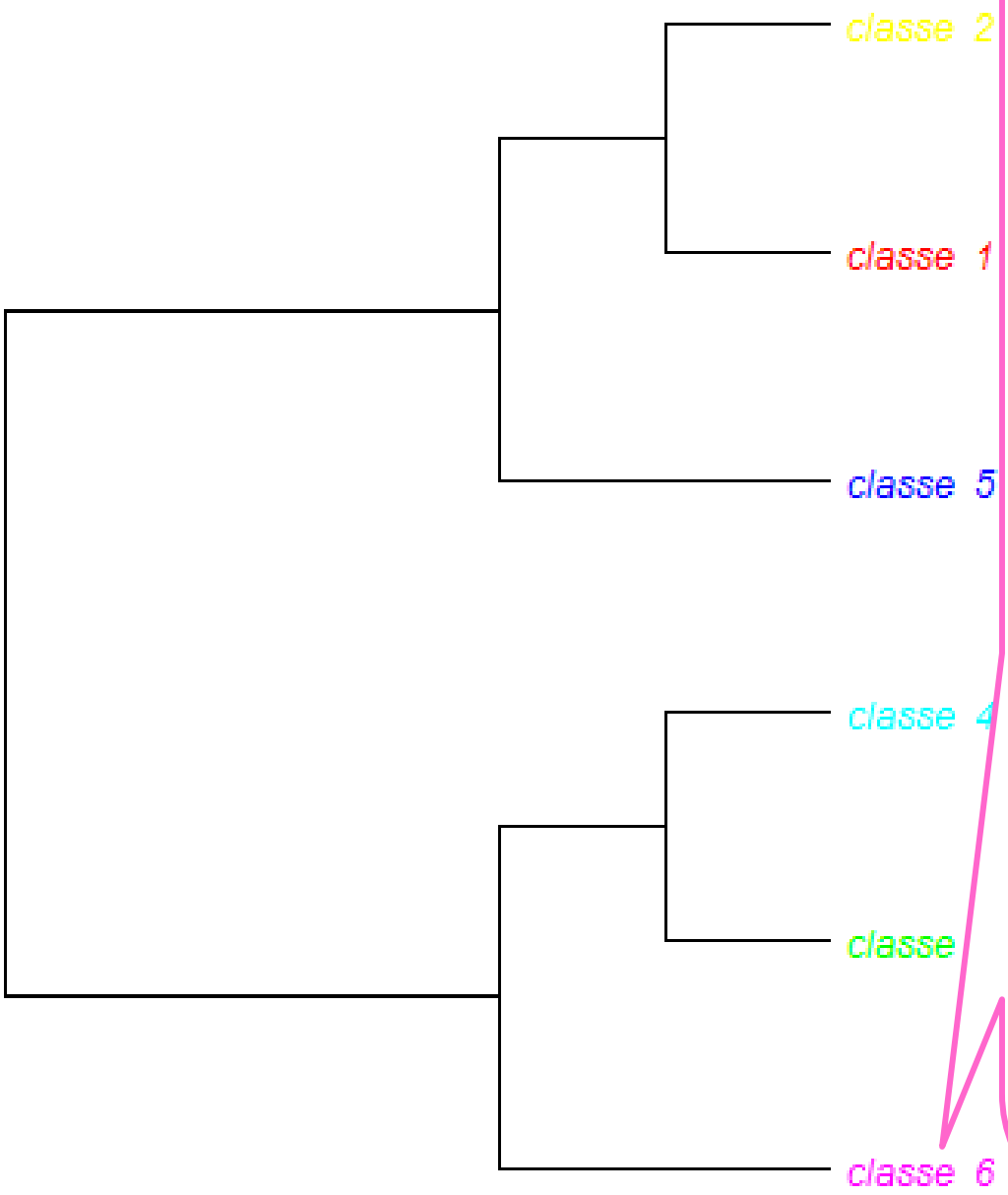

200
150
100
50
0
-50
-100
-150
classe.1
classe.2
classe.3
classe.4
classe.5
classe.6
rien
pas
ne



naître, père, parent, italien, nationalité, mère, espagnol, passeport, polonais, allemand, hasard, algérien, fils, militaire, marier, carte, fille, naissance, double, naturalisation, an, anglais, mariage, demander, immigré, obtenir, enfant, prouver, grandir, arrière, pur, choisir, famille, sang, sol…

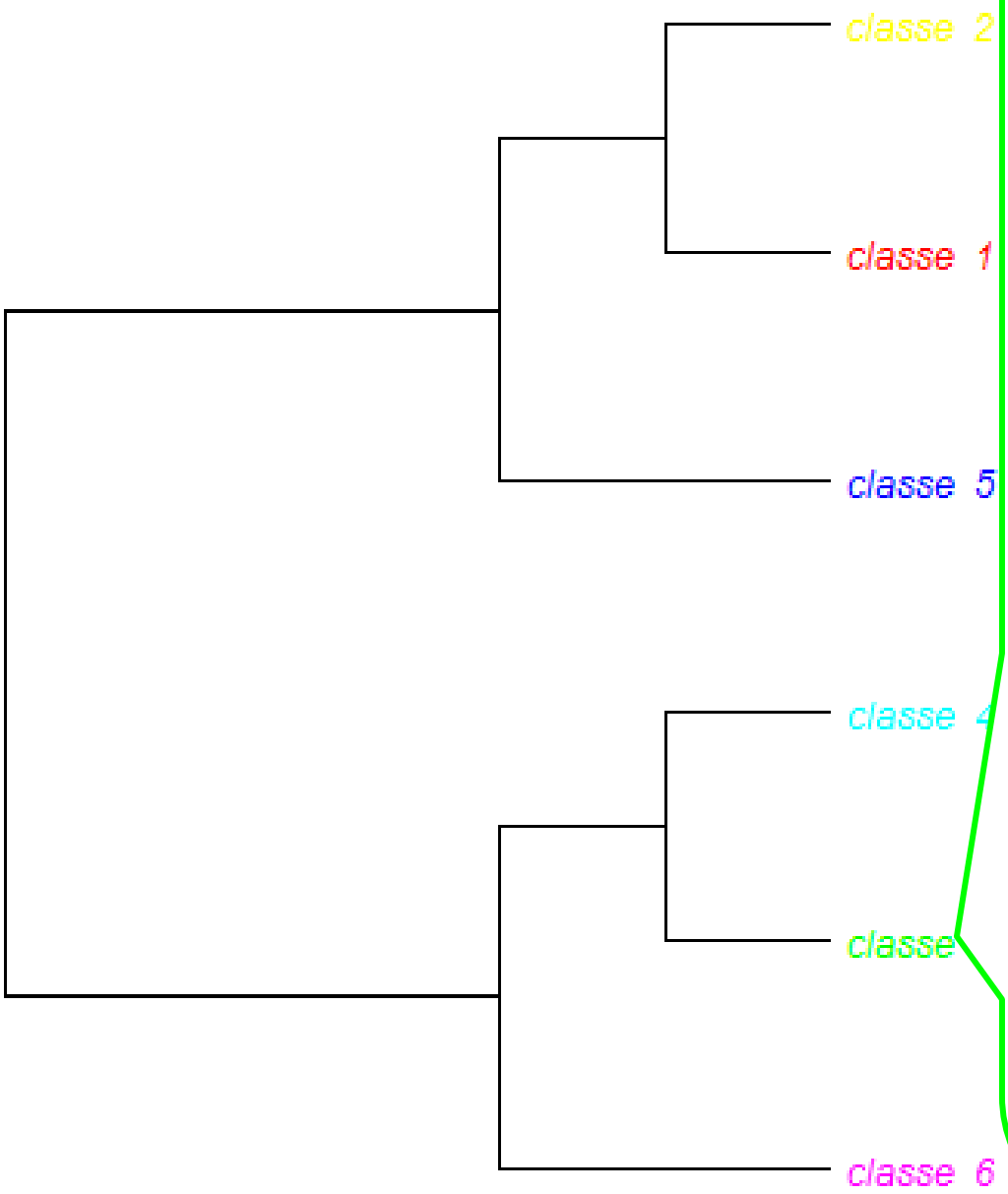
… mon petit fils est né en France, il est donc français jus soli. Il a comme tout le monde 4 grands parents : un grand-père corse et un autre italien, une grand-mère franco-espagnole et une autre camerounaise.


400
200
0
-200
classe.1
classe.2
classe.3
classe.4
classe.5
classe.6
mes
ma
mon
j
je
me
m



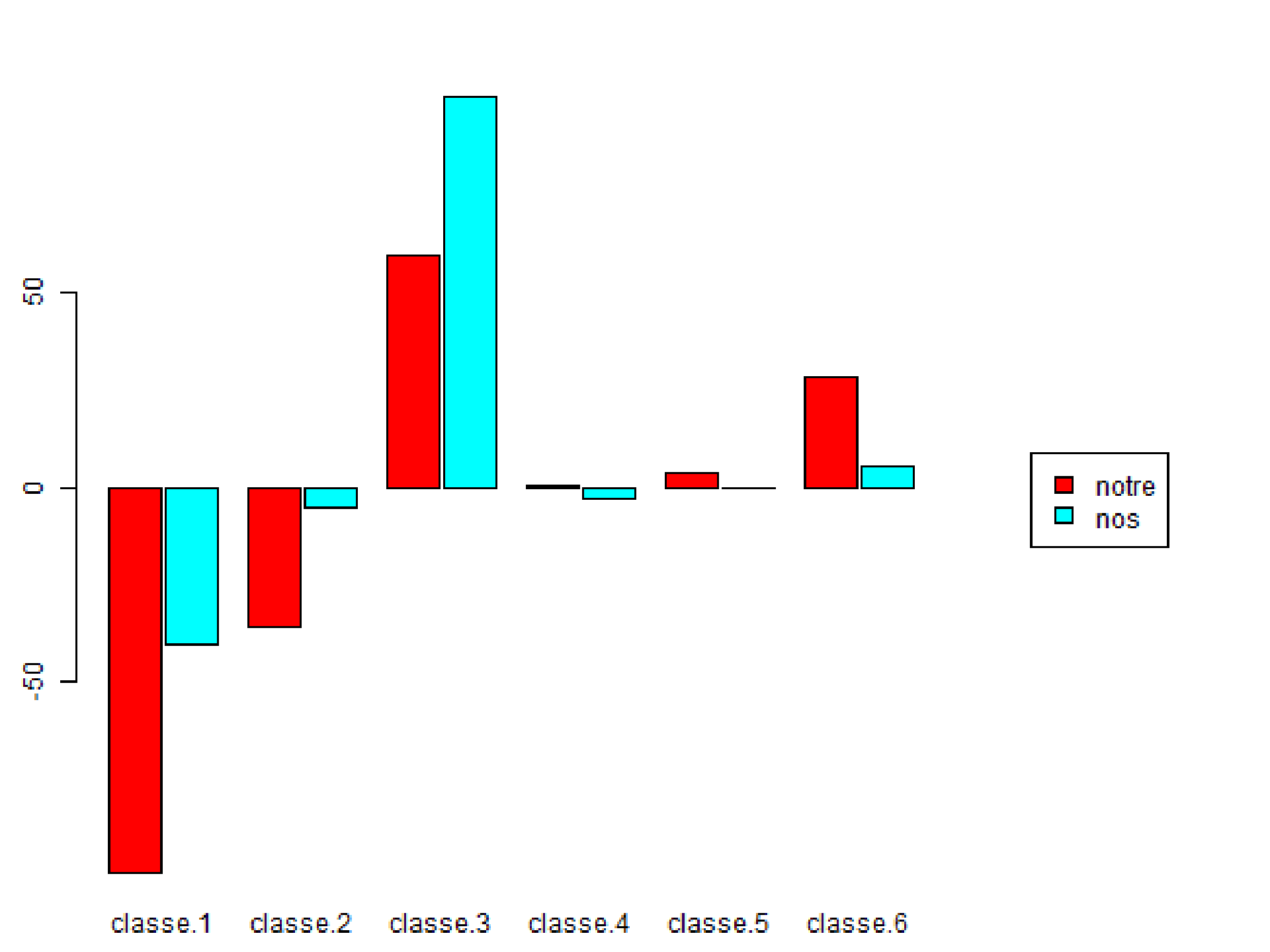
égalité, liberté, fraternité, laïcité, valeur, devise, droit, homme, principe, république, respect, solidarité, expression, fondamental, démocratie, adhérer, républicain, séparation, égal, social, justice, déclaration, tolérance, défendre, femme…

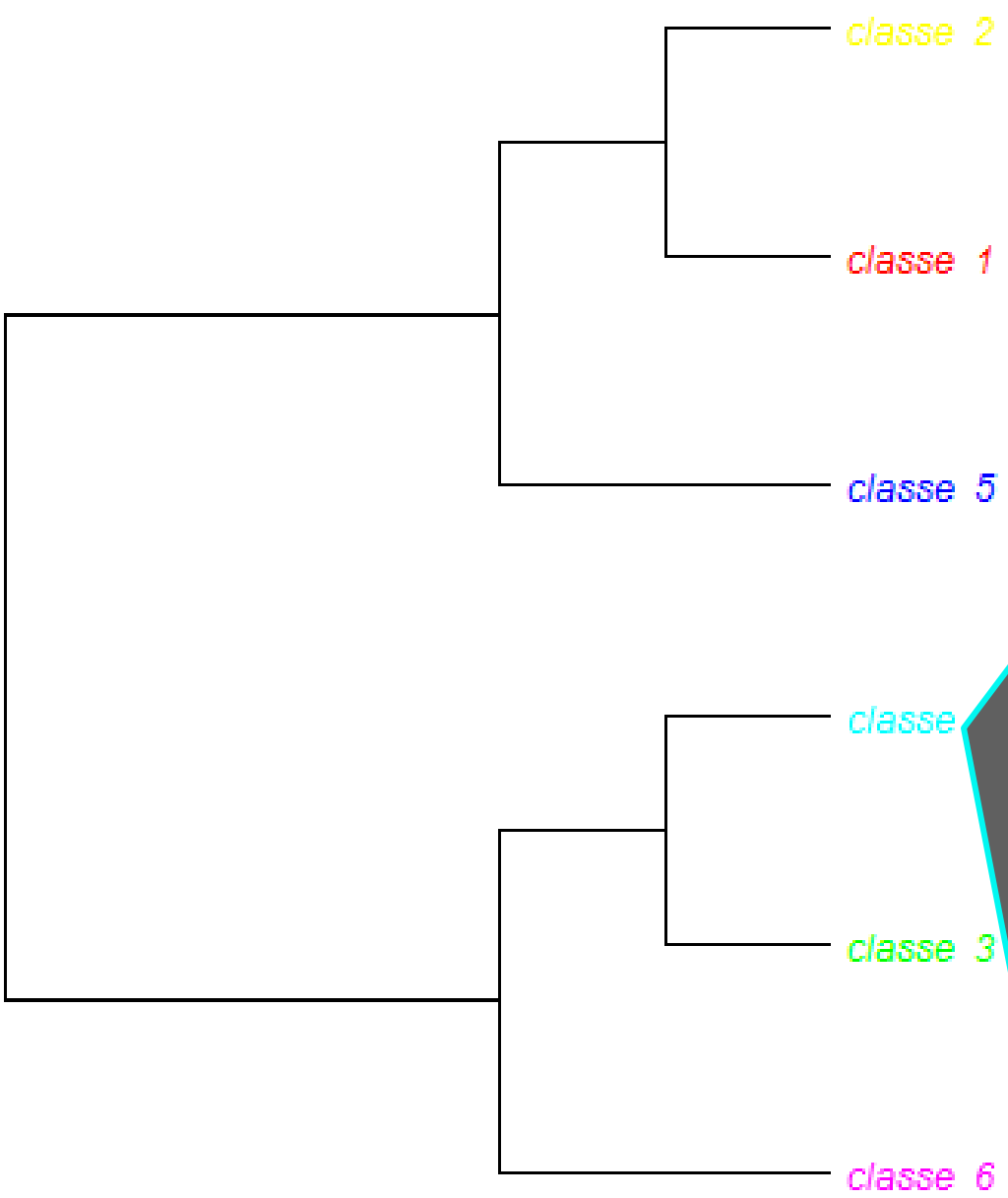
Être français c'est donner une place essentielle à certaines valeurs : laïcité, respect des droits de l'homme, protection sociale. Ces valeurs sont des déclinaisons de la devise républicaine : liberté, égalité, fraternité.



histoire, chrétien, culture, racine, langue, tradition, patrimoine, transmettre, héritage, paysage, ancêtre, gastronomie, judéo_chrétien, art, civilisation, millénaire, aimer, préserver, fier, façonner, attachement, spécificité, église, …

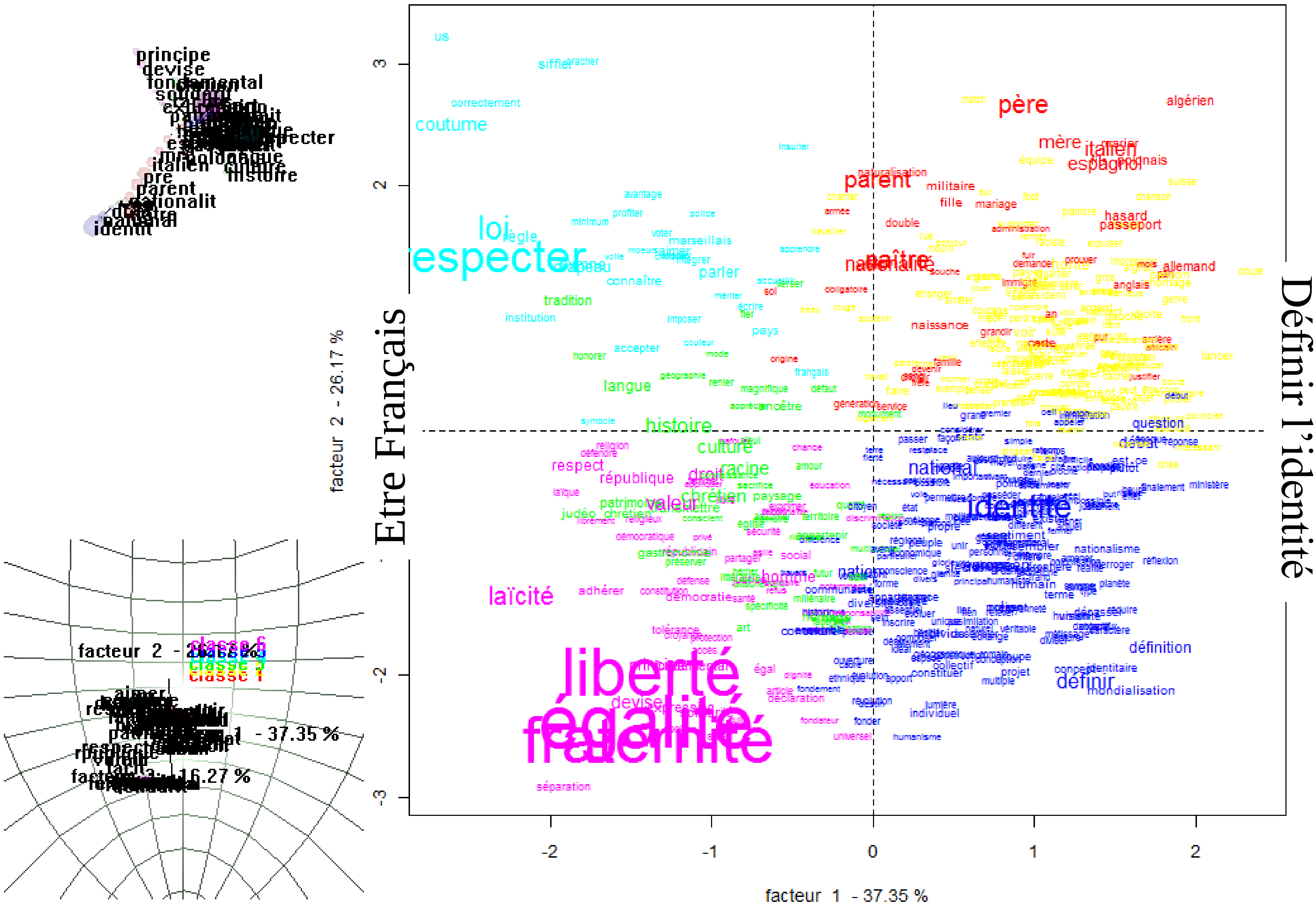
être français c'est partager une histoire, une mémoire. C'est avoir en commun un riche legs de souvenirs (Renan) car la nation comme l'individu est l'aboutissement d'un long passé d'efforts, de sacrifices et dévouements.


50
0
-50
classe.1
classe.2
classe.3
classe.4
classe.5
classe.6
notre
nos



respecter, loi, coutume, drapeau, parler, règle, hymne, siffler, aimer, us, marseillais, connaître, accepter, correctement, institution, langue, intégrer, profiter, tradition, cracher, voter, fier, écrire, chanter,

… cela passe par l'amour de ses symboles : la marseillaise, le drapeau, la loi, les institutions, mais aussi par des actions concrètes comme ne pas siffler son hymne dans les stades, comme ne pas accepter des critiques injustifiées sur son pays…



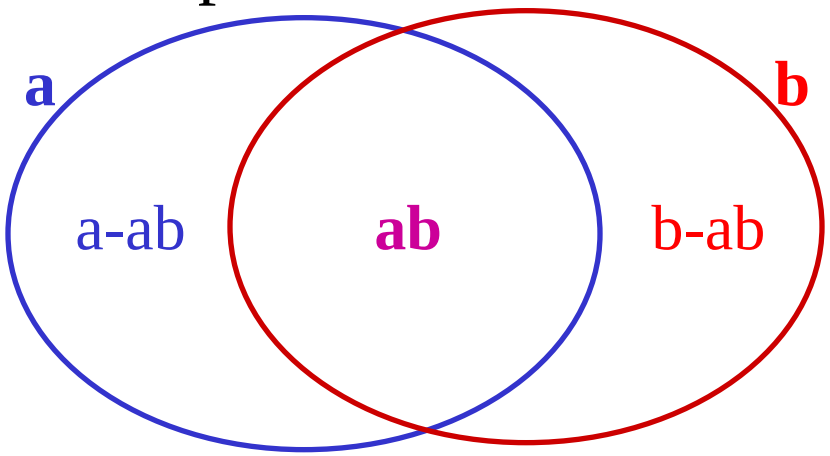
Etre Français
Définir l'identité
facteur 1 - 37.35 %
facteur 2 - 26.17 %
respecter
loi
coutume
père
parent
mère
italien
espagnol
algérien
naître
nationalité
identité
national
liberté
égalité
fraternité
laïcité
histoire
culture
langue
république
valeur
chrétien
respect
définition
définir
question

La distance lexicale

- Distances entre les grandes villes de France: demi-matrice traduisant une certaine idée de la carte de France.
- Peut-on également calculer des distances entre des textes et représenter graphiquement une carte textuelle ?
- **Connexion lexicale** (Muller, 1977 ; Brunet, 1988):
- où:

$$d = \left(\frac{(a\text{-}ab)}{a} + \frac{(b\text{-}ab)}{b} \right)$$

  - ab = partie commune aux vocabulaires a et b
  - a-ab et b-ab = parties privatives de a et b



Di Giacomo (EJSP, 1980)

•Associations libres d'étudiants à propos de neuf mots inducteurs : *agl (*Assemblée Générale des étudiants de Louvain) , *cadre*, *comité*, *droite*, *étudiants*, *gauche*, *grève*, *pouvoir* et *travailleur*.

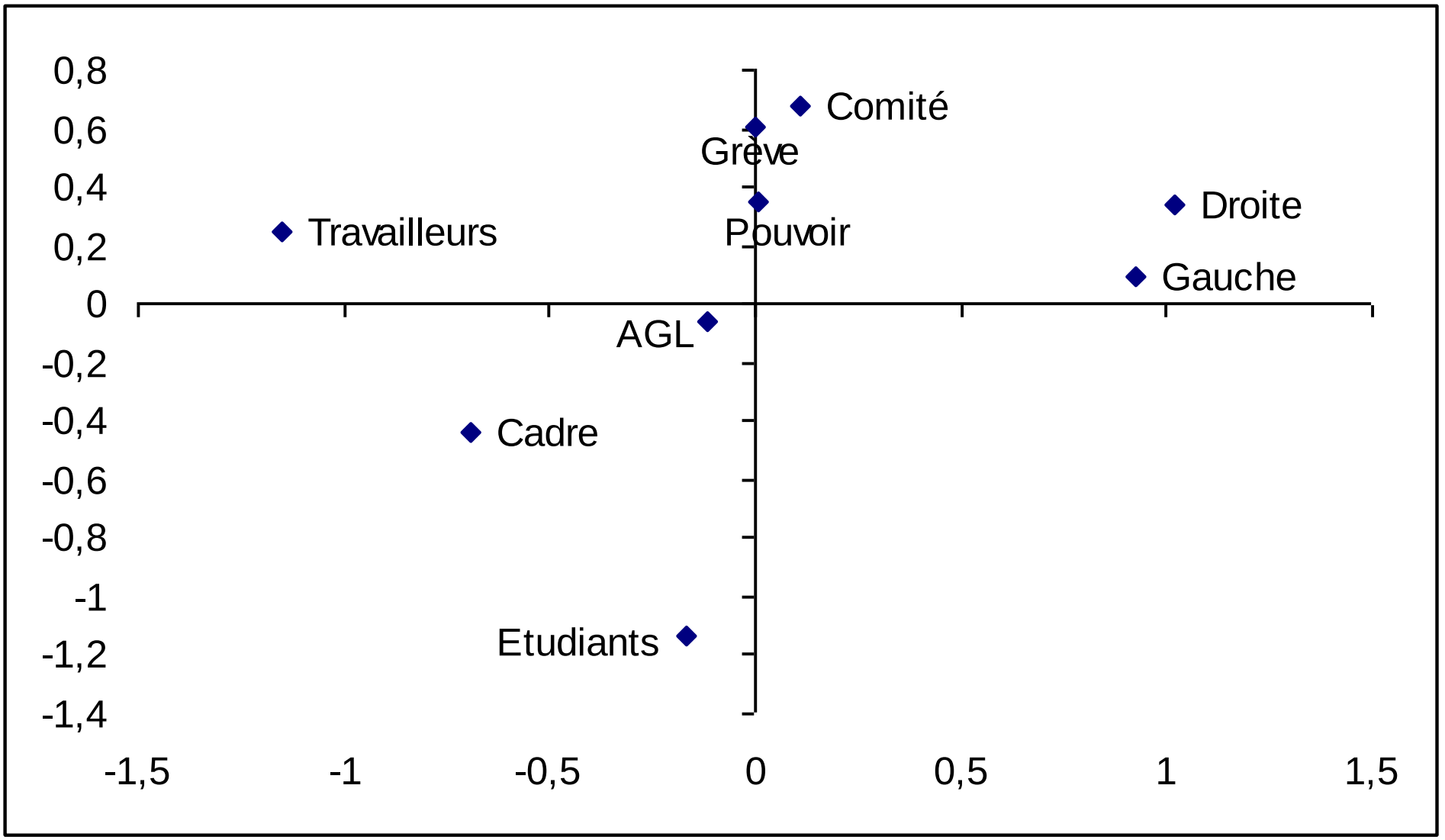
•Calcul de la proximité des lexiques obtenus, en comparant les distances entre chacune des 36 paires de lexiques

•Indice d'association de Ellegard (Di Giacomo, 1980) :

$$r^n = \frac{\text{nombre de mots communs}}{\sqrt{L1 \times L2}}$$

où L1 est le nombre de mots du 1er lexique, et L2, le nombre de mots du 2ème lexique. Cet indice varie de 0 à 1 et permet de repérer les lexiques les plus semblables, ou les plus éloignés.

•AFC

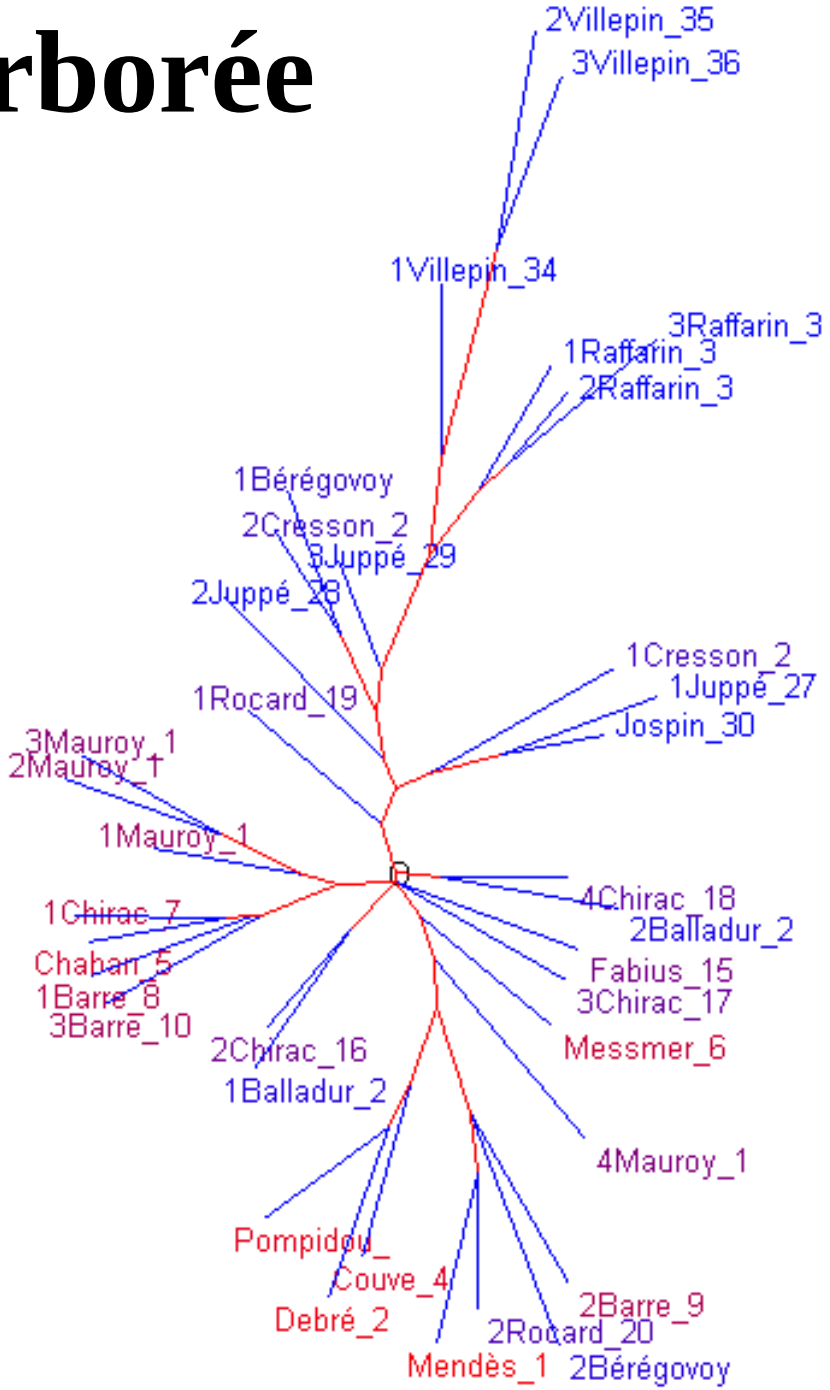


•Les représentations sociales des termes utilisés semblent s'organiser sur la base de l'appartenance ou non à un univers « politique ».

•Les étudiants se situent eux-mêmes dans l'apolitique qu'ils différencient ensuite suivant la proximité que le terme entretient avec eux ».

Représentation arborée

- La méthode Luong permet de représenter sous forme d'arbres la matrice des distances intertextuelles



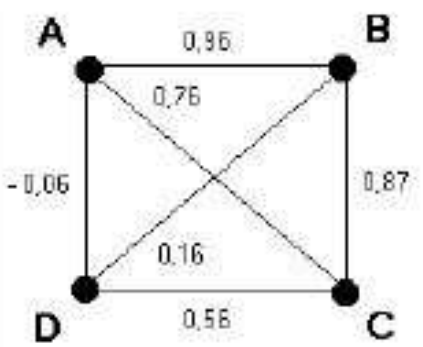
Cooccurrences et similitudes (graphes)

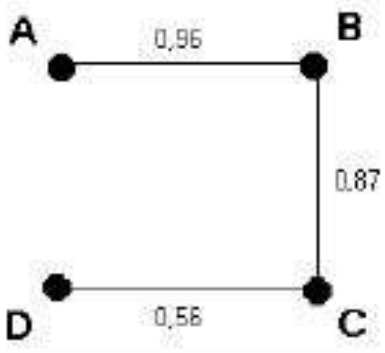
- Méthode(s) graphique(s) pour l'étude des relations entre les parties d'un ensemble.
- Les matrices :
  - Distances (euclidiennes, maximum, Jaccard, Manhattan…)
  - Similitude (cooccurrences, Jaccard, simple matching, phi…)
- L'ADS est classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête.
  - Flament, 1962 ; Flament, 1981 ; Vergès & Bouriche, 2001.

L’analyse de similitude (ADS)

- étudier la proximité et les relations entre les éléments d’un ensemble, généralement sous forme d’*arbres maximum* :
  - le nombre de liens entre deux items évoluant « comme le carré du nombre de sommets » (Flament & Rouquette, 2003 : 88), l’ADS cherche à réduire le nombre de ces liens pour aboutir à « un graphe connexe et sans cycle » (Degenne & Vergès, 1973 : 473).
  - L’« arbre maximum ») est créé par les arêtes les plus fortes du graphique. C’est l’arbre le plus simple que l’on peut obtenir, mais c’est aussi le plus lourd (en termes d’information).
  - On considère toutes les « cliques » possibles (ex. ABCA, BCDB) et on élimine les liens les plus faibles (ex. entre A et C et entre B et D).

Arbre de similitude



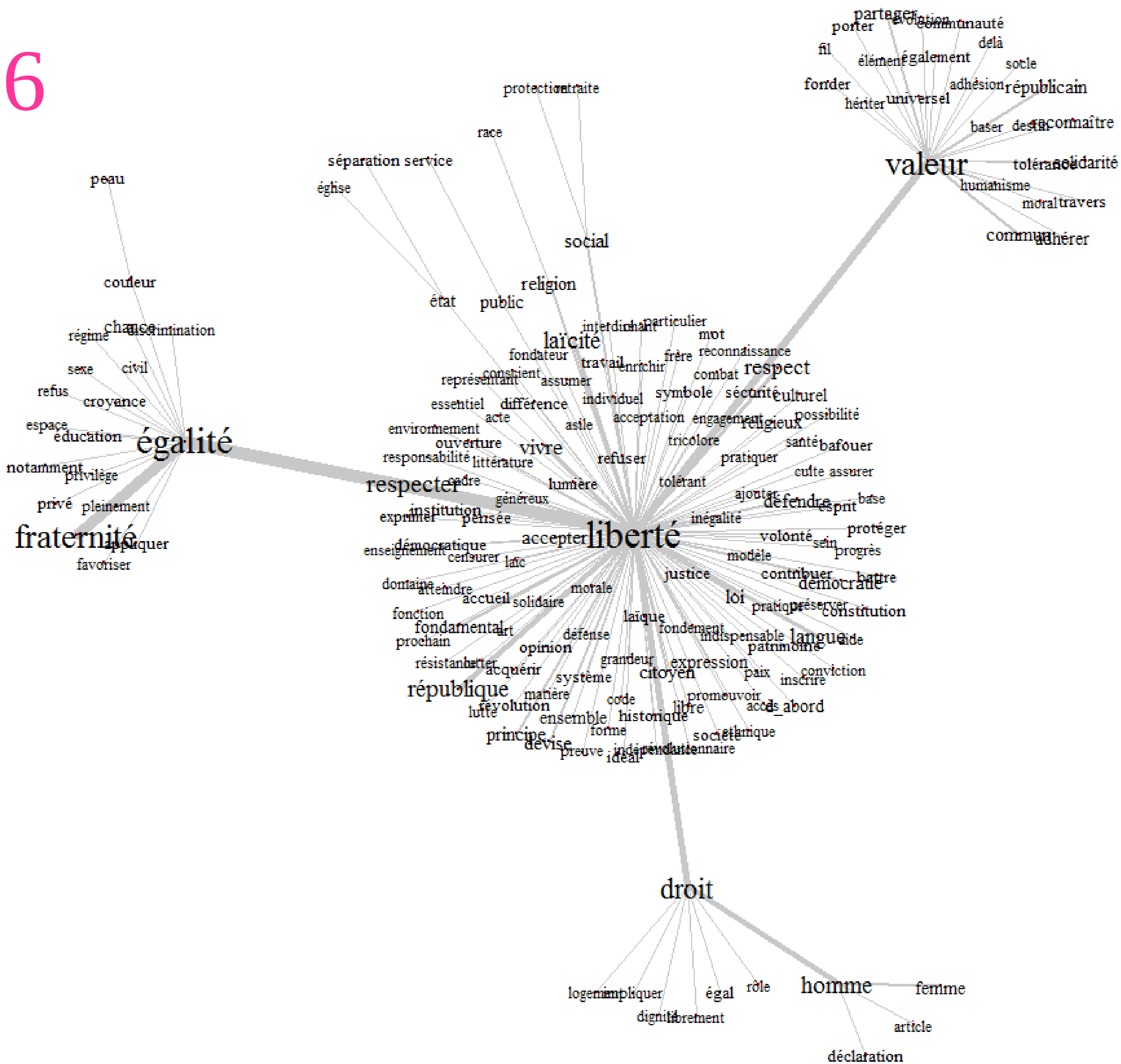


Arbre maximum

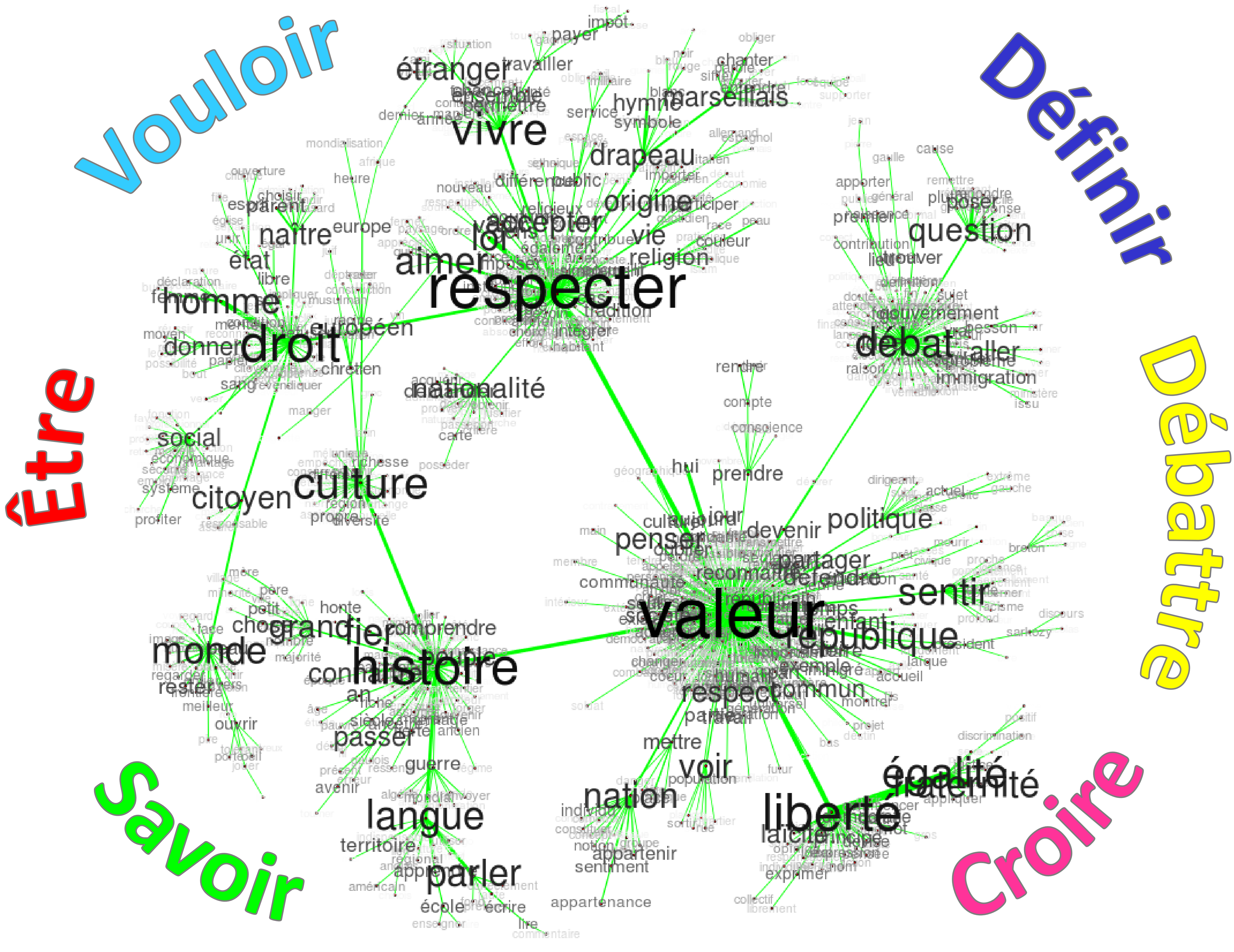
L’analyse de similitude (ADS)

- L’ADS d’une matrice textuelle a été intégrée au logiciel *IRaMuTeQ* (P.Ratinaud)
- Elle permet de décrire des classes lexicales, des profils de spécificités ou des corpus entiers.
- Matrices *formes pleines * segments de textes*.
- Librairies de *R*:
  - igraph (Csardi & Nepusz, 2006)
  - Proxy (Meyer & Buchta, 2012)

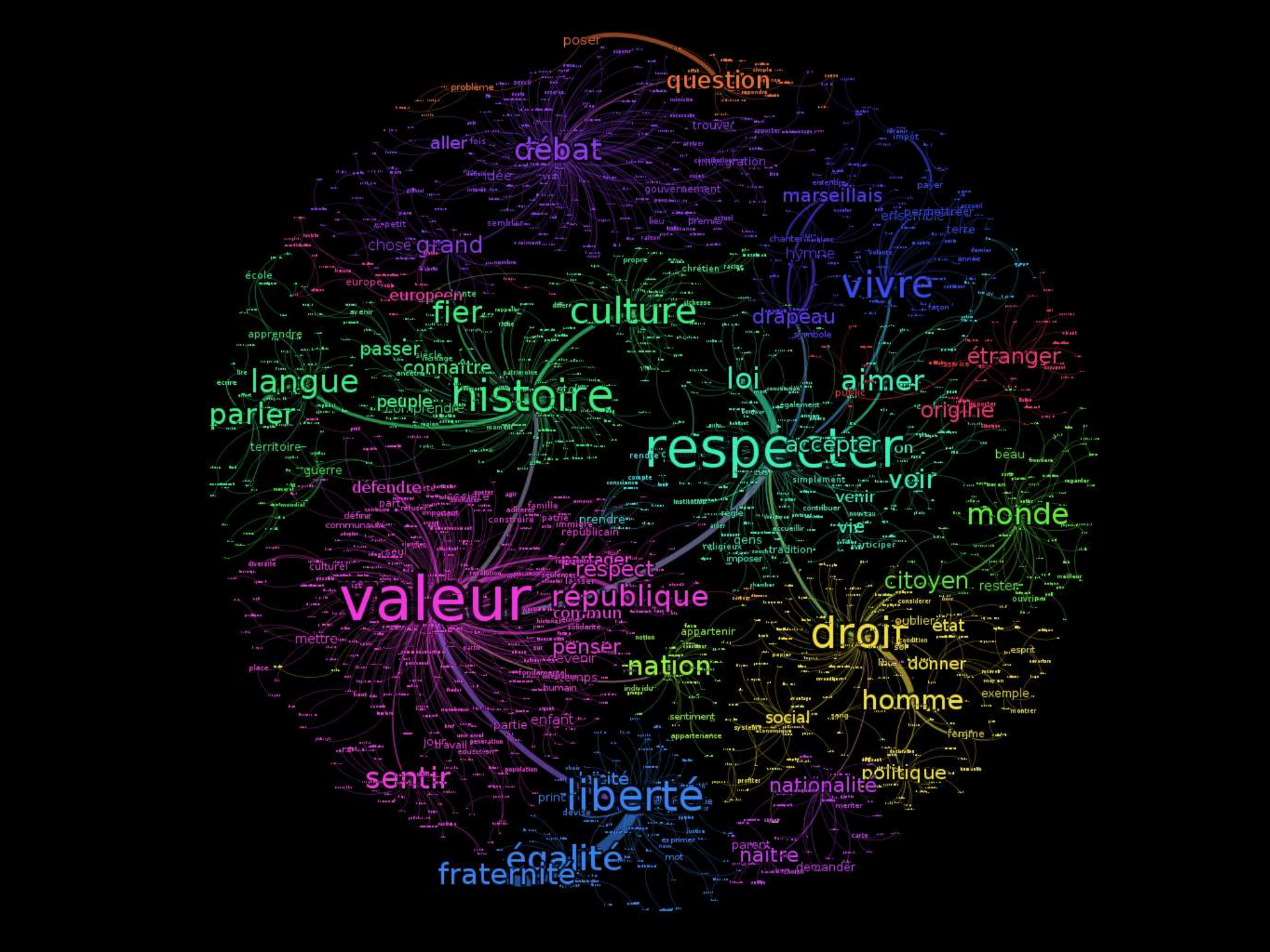
Classe 6

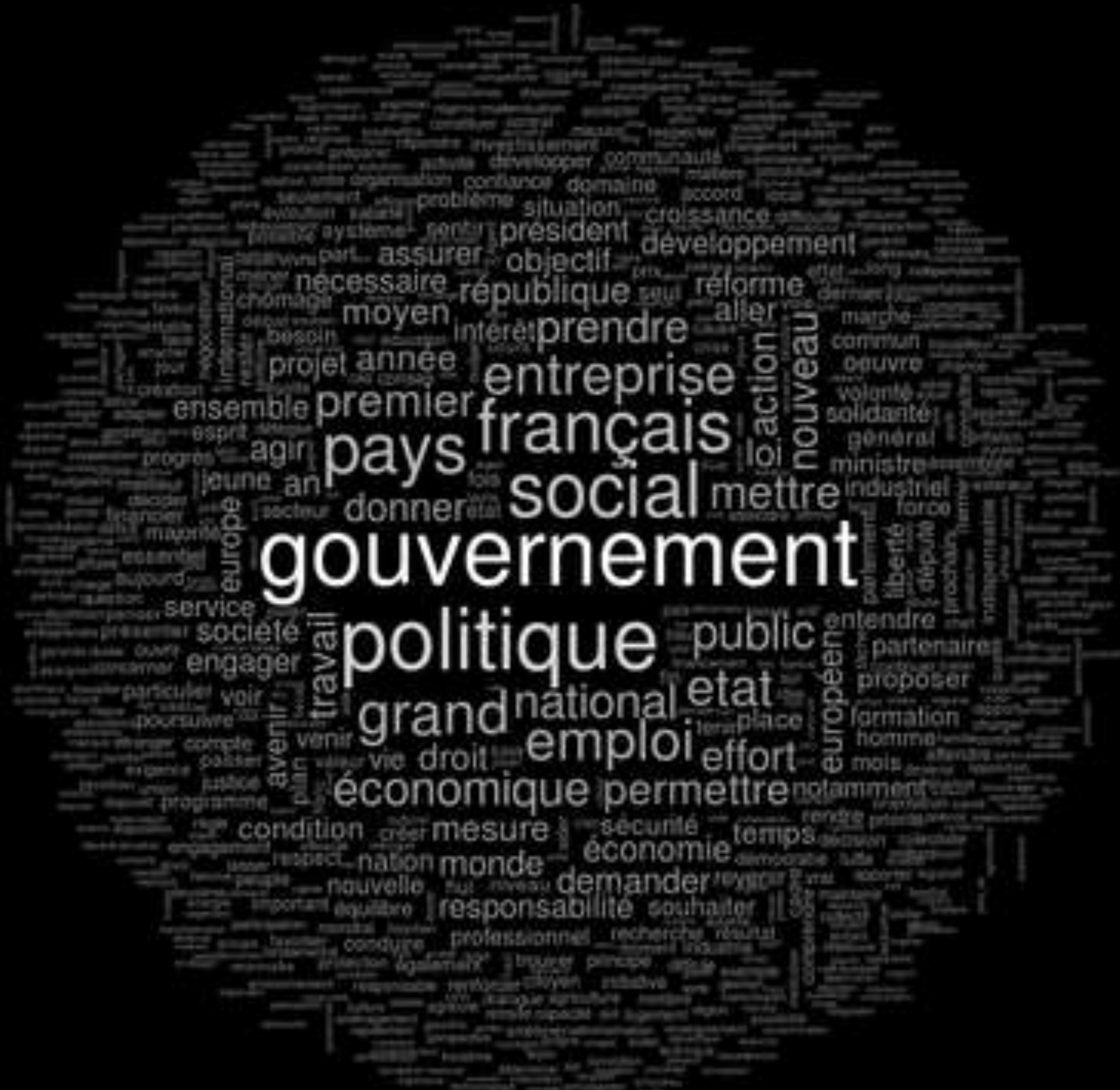


liberté
égalité
fraternité
valeur
droit
homme
femme
article
déclaration
égal
rôle
dignité
librement
impliquer
peau
couleur
chance
discrimination
régime
sexe
civil
refus
croyance
espace
éducation
notamment
privilège
privé
pleinement
appliquer
favoriser
partager
porter
communauté
fil
élément
également
delà
socle
fonder
adhésion
républicain
hériter
universel
baser
reconnaître
solidarité
humanisme
moral
travers
adhérer
protection
retraite
race
séparation
service
église
social
religion
état
public
laïcité
fondateur
travail
enrichir
frère
mot
particulier
reconnaissance
combat
respect
assumer
individuel
symbole
sécurité
culturel
essentiel
différence
acte
acceptation
asile
possibilité
environnement
religieux
ouverture
vivre
tricolore
santé
bafouer
responsabilité
littérature
refuser
pratiquer
culte
assurer
respecter
cadre
lumière
tolérant
généreux
ajouter
défendre
base
esprit
institution
exprimer
pensée
inégalité
protéger
accepter
volonté
sein
démocratique
progrès
censurer
laïc
modèle
justice
contribuer
démocratie
battre
domaine
atteindre
morale
loi
accueil
solidaire
préserver
constitution
fonction
laïque
fondement
fondamental
art
défense
indispensable
langue
aide
prochain
opinion
patrimoine
acquérir
grandeur
expression
conviction
citoyen
paix
inscrire
système
république
matière
code
promouvoir
lutte
révolution
libre
ensemble
historique
société
principe
forme
devise
preuve
idéal



Vouloir
Définir
Être
Débattre
Savoir
Croire
valeur
respecter
droit
histoire
liberté
culture
vivre
drapeau
origine
question
débat
politique
république
sentir
langue
parler
nation
voir
monde
homme
naître
aimer
étranger
marseillais
nationalité
citoyen
social
immigration
égalité
penser
passer
guerre
respect
commun
européen
donner
prendre
mettre
territoire
école
écrire
appartenance
appartenir
sentiment
exprimer
religion
tradition
hymne
symbole
payer
impôt
travailler
mondialisation
europe
chrétien
richesse
diversité
communauté
reconnaître
exemple
accueil
laïque
discours
sarkozy
breton
devenir
partager
comprendre
fier
honte
grand
ouvrir
meilleur
avenir
régional
américain
enseigner
lire
commentaire

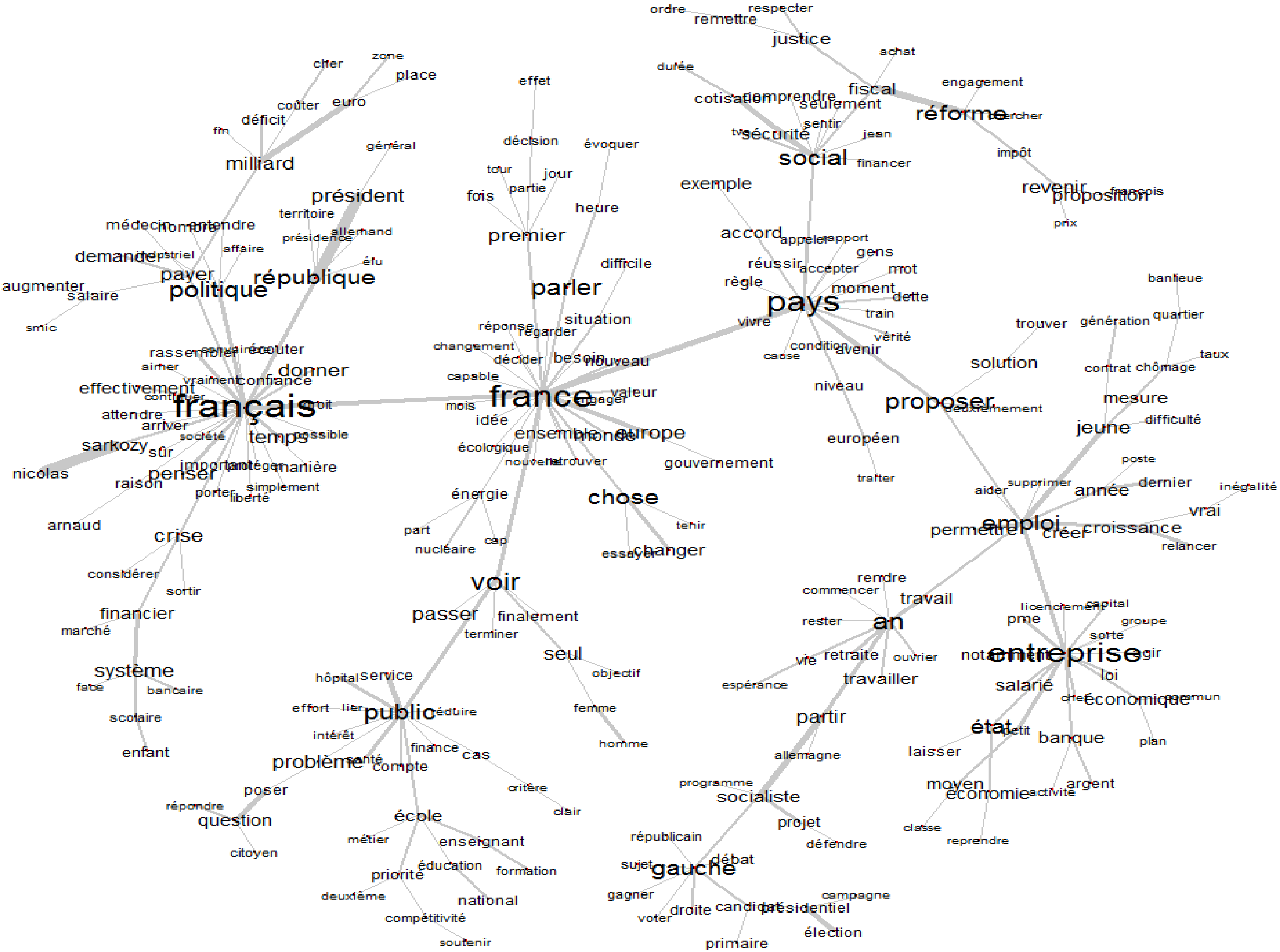

poser
question
problème
débat
aller
trouver
idée
gouvernement
marseillais
permettre
ensemble
terre
chose
grand
hymne
chrétien
école
europe
européen
vivre
fier
culture
drapeau
apprendre
passer
connaître
étranger
langue
histoire
loi
aimer
peuple
parler
origine
respecter
accepter
territoire
guerre
beau
défendre
voir
venir
monde
prendre
vie
gens
tradition
respect
citoyen
valeur
république
commun
droit
état
penser
nation
donner
mettre
homme
exemple
social
enfant
partie
sentir
liberté
nationalité
politique
égalité
naître
fraternité
demander


gouvernement
politique
français
social
pays
entreprise
grand
emploi
national
etat
public
économique
permettre
prendre
république
premier
mettre
donner
effort
nouveau
action
loi
travail
européen
développement
président
objectif
réforme
croissance
situation
assurer
nécessaire
moyen
année
projet
ensemble
agir
jeune
an
mesure
monde
économie
demander
responsabilité
condition
nation
temps
sécurité
droit
vie
avenir
europe
société
engager
service
entendre
partenaire
proposer
formation
notamment
place
aller
intérêt
ministre
industriel
solidarité
volonté
nouvelle
souhaiter
professionnel

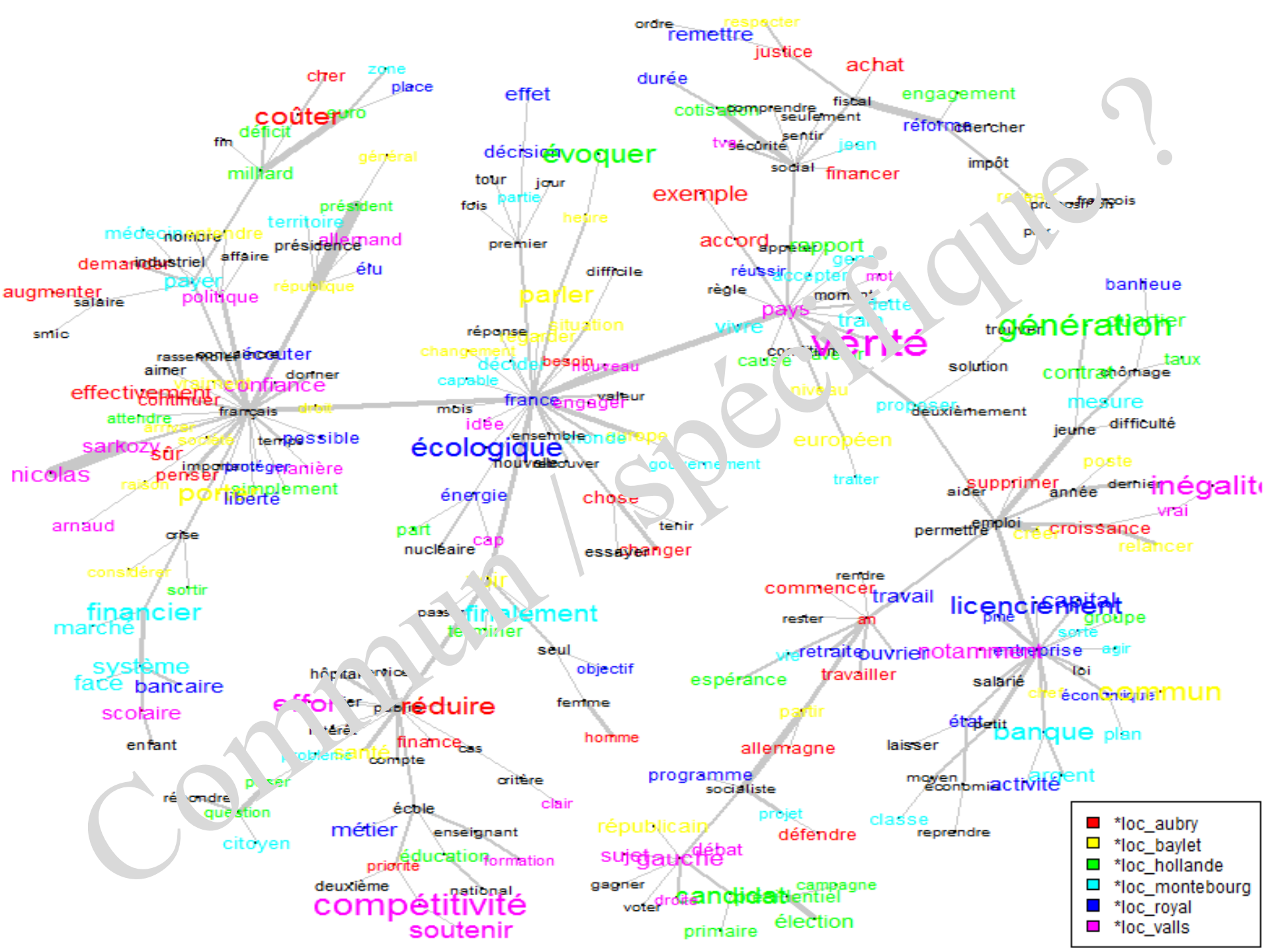
L'ADS du corpus « PS »

- Deux tours:
  - Trois débats entre six candidats : Martine Aubry, Jean-Michel Baylet, François Hollande, Arnaud Montebourg, Ségolène Royal et Manuel Valls ;
  - Les deux finalistes - Martine Aubry et François Hollande - se sont affrontés le 16 octobre 2011
- Retranscription manuelle (assistée par reconnaissance vocale)

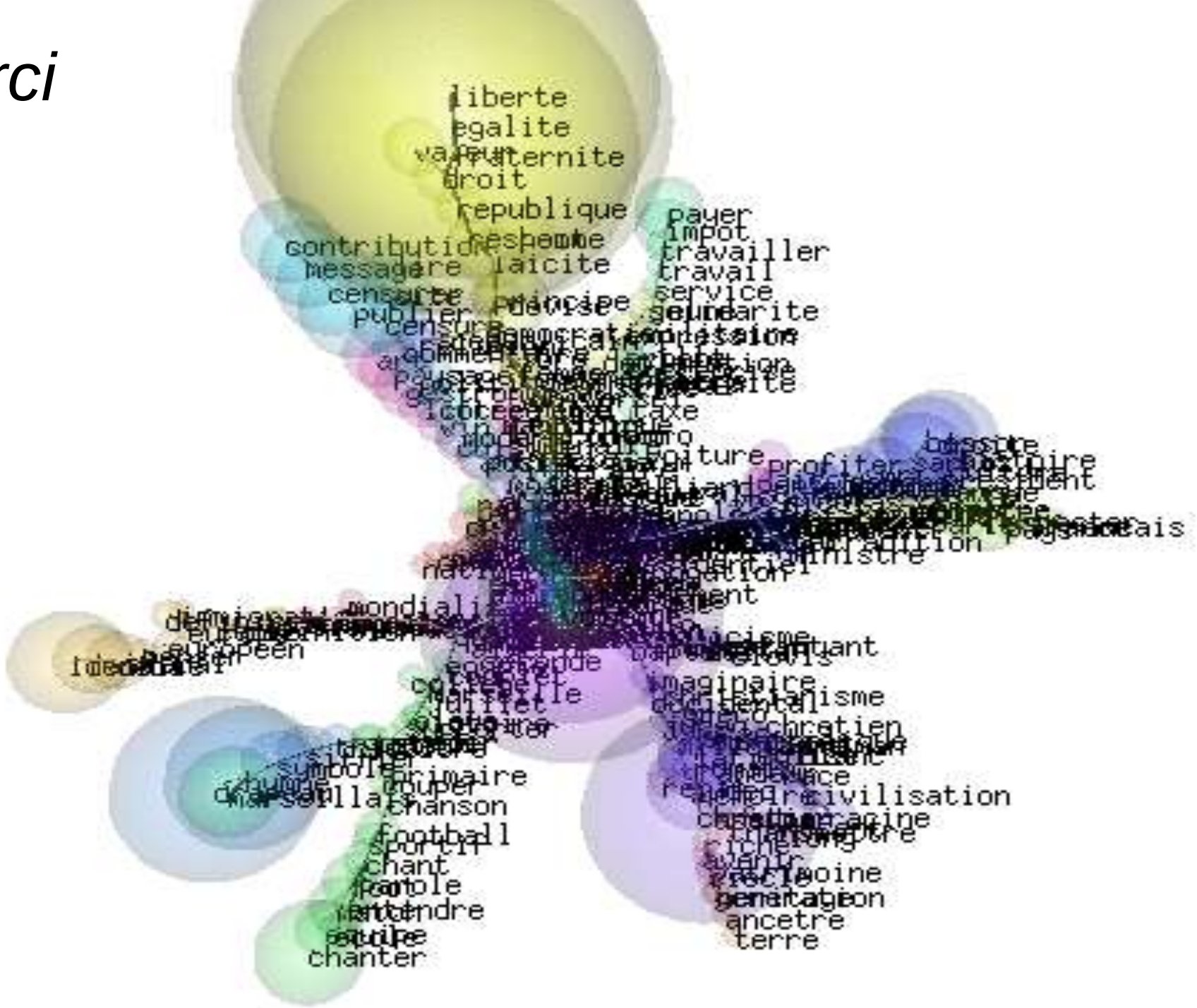
nombre d'uci :	295 (tours de parole)
nombre d'occurrences :	71913
nombre de formes :	5265
moyenne d'occurrences par forme :	18.96
nombre d'hapax :	1472 (2.05% des occurrences - 27.96% des formes)
moyenne d'occurrences par uci :	243.77



ordre
respecter
remettre
justice
achat
durée
fiscal
engagement
cotisation
comprendre
seulement
réforme
chercher
sentir
sécurité
jean
social
financer
impôt
revenir
proposition
françois
prix
exemple
accord
appeler
rapport
gens
réussir
accepter
mot
règle
moment
dette
pays
train
vivre
vérité
condition
avenir
cause
niveau
européen
traiter
proposer
deuxièmement
trouver
génération
banlieue
quartier
taux
solution
contrat
chômage
mesure
jeune
difficulté
poste
supprimer
aider
année
dernier
inégalité
vrai
permettre
emploi
créer
croissance
relancer
effet
décision
évoquer
tour
jour
partie
fois
heure
premier
difficile
parler
situation
réponse
regarder
changement
décider
besoin
nouveau
capable
france
valeur
engager
mois
idée
ensemble
monde
europe
écologique
nouveau
retrouver
gouvernement
chose
tenir
énergie
part
nucléaire
cap
essayer
changer
voir
passer
finalement
terminer
seul
objectif
femme
homme
zone
citer
place
coûter
euro
déficit
fin
milliard
général
président
territoire
médecin
nombre
entendre
allemand
présidence
affaire
demander
industriel
élu
payer
république
augmenter
politique
salaire
smic
rassembler
écouter
aimer
donner
vraiment
confiance
effectivement
contribuer
français
droit
attendre
arriver
temps
possible
société
sarkozy
sûr
nicolas
important
protéger
manière
penser
raison
simplement
porter
liberté
arnaud
crise
considérer
sortir
financier
marché
système
faire
bancaire
scolaire
enfant
hôpital
service
effort
lier
public
réduire
intérêt
finance
cas
problème
santé
compte
poser
critère
répondre
clair
question
citoyen
école
métier
enseignant
éducation
formation
priorité
deuxième
national
compétitivité
soutenir
rendre
commencer
travail
rester
an
vie
retraite
ouvrier
espérance
travailler
partir
allemagne
programme
socialiste
projet
républicain
défendre
sujet
débat
gauche
gagner
droite
candidat
présidentiel
campagne
voter
primaire
élection
licenciement
capital
pme
groupe
sorte
notamment
entreprise
agir
loi
salarié
chef
économique
commun
état
petit
banque
plan
laisser
moyen
argent
économie
activité
classe
reprendre


Commun / spécifique ?
*loc_aubry
*loc_baylet
*loc_hollande
*loc_montebourg
*loc_royal
*loc_valls

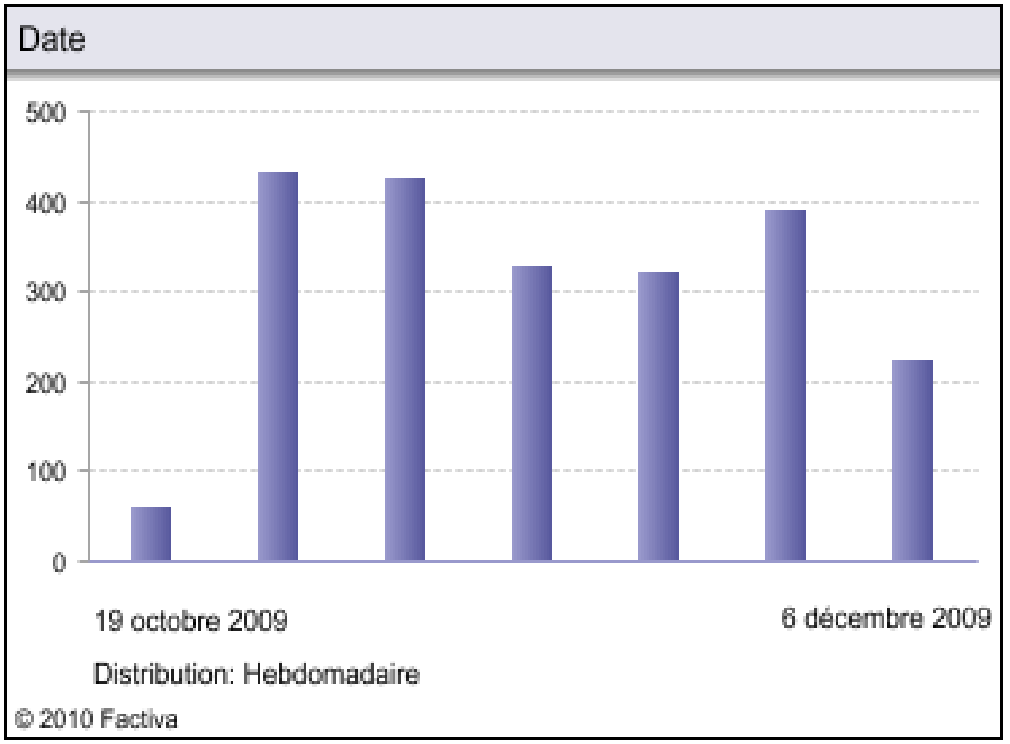
Merci


liberte
egalite
fraternite
droit
republique
payer
impot
travailler
travail
service
solidarite
contribution
messagerie
laicite
publier
censure
voiture
profiter
mondial
europeen
imaginaire
chretien
civilisation
symbole
primaire
chanson
football
sport
chant
chanter
ancetre
terre

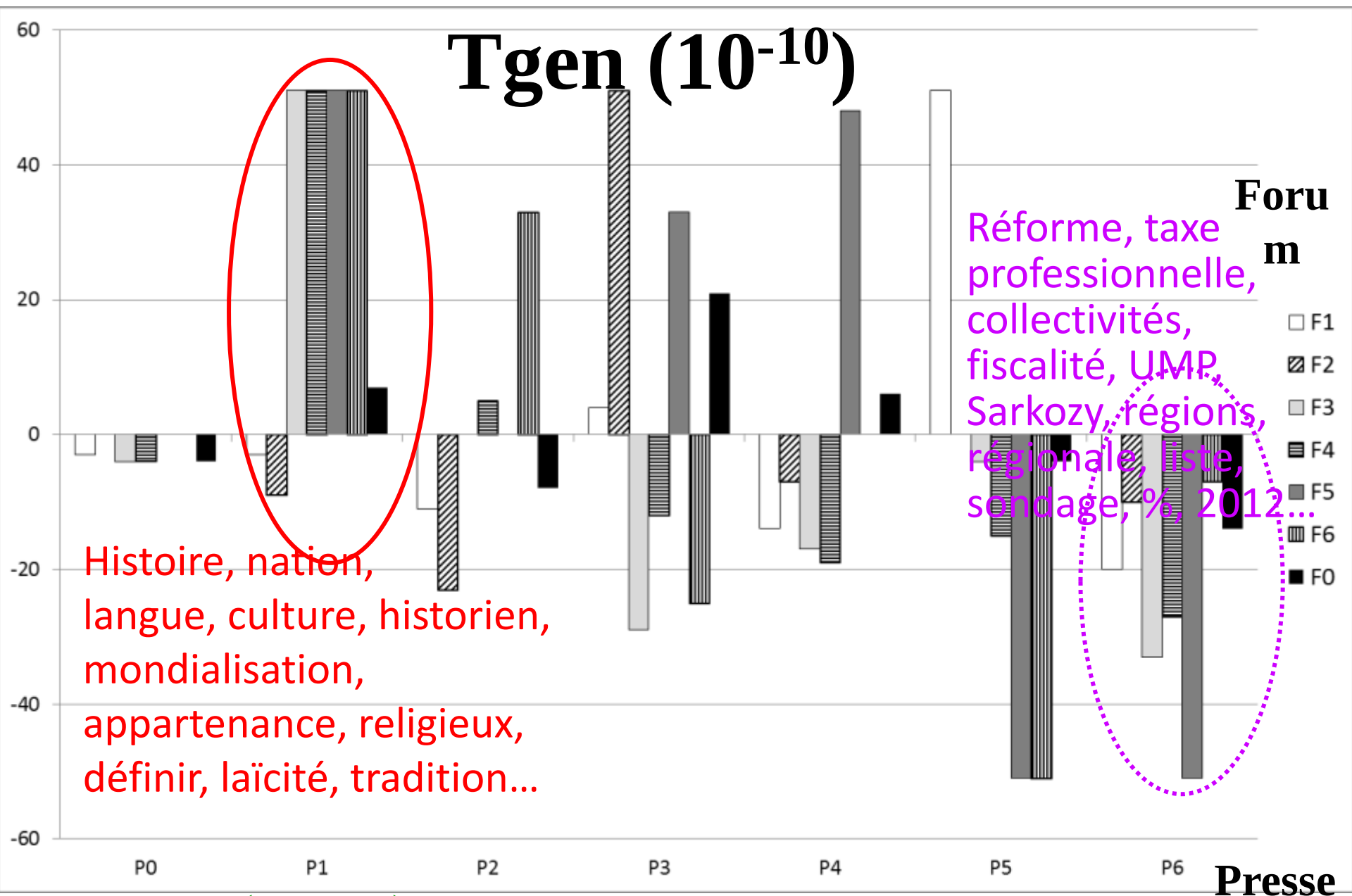
Et les médias ?

- 25 octobre – 2 décembre 2009 (fin du débat dans sa forme originale).

- *Factiva*:
  - « identité nationale »
  - journaux en langue française
    - 2127 articles
  - Presse française
  - Contrôle d'homogénéité
    - 1436 articles



AFP	292	20,30%	Presse Océan	34	2,40%
Libération	111	7,70%	Charente Libre	26	1,90%
l'Humanité	97	6,70%	Valeurs actuelles	21	1,50%
Le Monde	91	6,30%	La Tribune	19	1,30%
Le Figaro	83	5,80%	Le Point	18	1,30%
Sud Ouest	79	5,50%	bakchich info	15	1,00%
Midi Libre	65	4,50%	L'Humanité Dimanche	14	1,00%
L'Est Républicain	61	4,20%	Centre Presse Les Journaux du Midi	11	0,80%
Le Progrès	58	4,00%	Courrier International	10	0,70%
la Croix	56	3,90%	RMC Info Talk Sport	10	0,70%
le Parisien	47	3,30%	Aujourd'hui en France	8	0,60%
Marianne 2	43	3,00%	L'Express	8	0,60%
Reuters France	40	2,80%	Dimanche Ouest France	2	0,10%
L'Indépendant Les Journaux du Midi	39	2,70%	L'Expansion	2	0,10%
Les Echos	37	2,60%	Alternatives Economiques	1	0,10%
Ouest France	37	2,60%	BFM	1	0,10%



- Investiture (60,56%)

(d'après Marty, Marchand & Ratinaud, *BMS*, 2012)