L'analyse du discours assistée par ordinateur: la méthode ALCESTE et le logiciel IRAMUTEQ

Emmanuel Marty

Université Grenoble Alpes

Avec la contribution de **Pascal Marchand** et **Pierre Ratinaud**

Université Toulouse 3

Pour analyser le discours:

- La statistique n’est pas la seule approche possible ...

☞Tout est possible et est question de choix épistémologiques, d’indicateurs et/ou de nature des corpus.

☞La statistique ne peut pas tout faire: il faut prévoir ce qu’on lui demandera : HYPOTHESES

☞Ce n’est pas la statistique qui garantit la qualité d’une recherche, mais le protocole.

Pourquoi l'ADAO en SHS ?

- **Analyse de discours (enquêtes, analyse sur archives): pourquoi, dans l'univers des mots possibles, ceux-là ont-ils été choisis ?**
- **Lien avec la psychologie sociale, la linguistique, les sciences de l'information et de la communication, les sciences politiques…**

→Définir les mots récurrents, leurs fonctions, leurs relations, leurs utilisations pour reconstruire du sens.

Quelques définitions

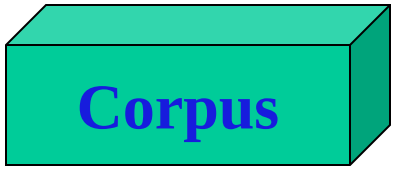
- Les questions que se donne la statistique lexicale sont les suivantes : « quels sont les textes les plus semblables en ce qui concerne le vocabulaire et la fréquence des *formes* utilisées ? Quelles sont les *formes* qui caractérisent chaque texte, par leur présence ou leur absence ? »

  (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)

L'interprétation en ADT



Constitution (normes de saisie):

Interprétation:
- concordances, cooccurrences et distributions de formes lexicales brutes / réduites ou de segments
-profils de parties (UC ou variables extra-textuelles)

Tableau lexical
- partition
  - -Variable(s)
  - -Unités contexte

Analyse
-AFC, classification
-Spécificités

André Salem (Paris, SFDS le 26/04/01)

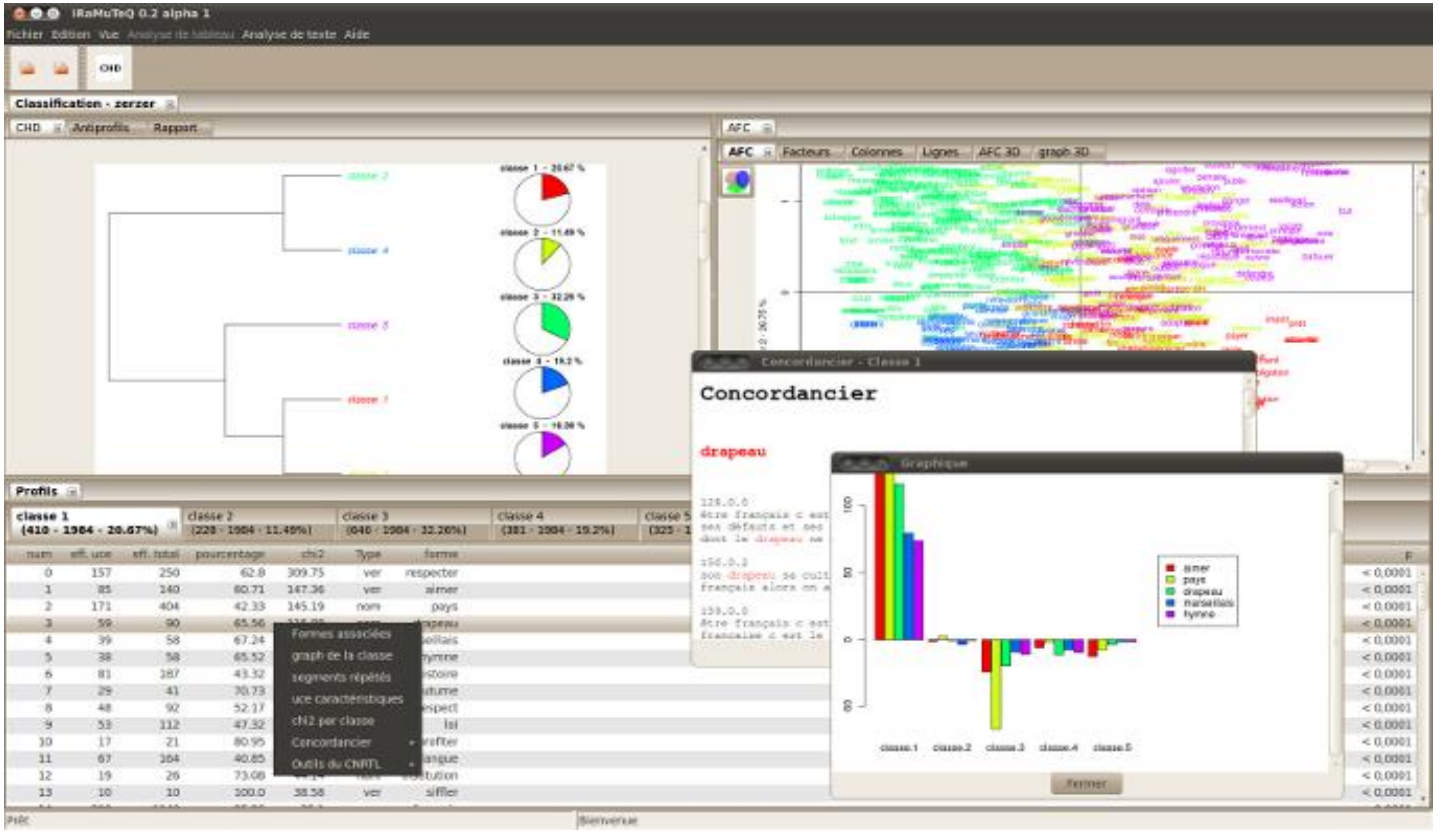
Quelques logiciels de lexicométrie

- **Alceste**
  - M. Reinert (http://www.image-zafar.com)
- **Lexico 3**
  - A. Salem (http://lexico3.no-ip.org)
- **Sphinx** *Lexica*
  - Y. Baulac (http://www.lesphinx-developpement.fr)
- **Hyperbase**
  - E. Brunet (http://ancilla.unice.fr/)
- **TXM**
  - S. Heiden (http://textometrie.ens-lyon.fr/)
- **IRAMuTeQ**
  - P. Ratinaud (http://repere.no-ip.org/Members/logiciel/iramuteq)

IRaMuTeQ (Pierre Ratinaud)

Logiciel libre et gratuit développé sur la base de logiciels libres:
- *Python* http://www.python.org
- R (R Development Core Team, 2009): http://r-project.org
- Lexique 3 (New, Pallier & Ferrand, 2005): http://lexique.org

Reproduit notamment la méthode *ALCESTE* (CDH) (Ratinaud & Dejean, 2009)



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* (INTEX: Silberztein, 1993)

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

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* = *jeune*+).
- Mais des formes très proches ne doivent pas forcément 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 (programme SHRDLU de Winograd, 1972 ; logiciel Alceste).

+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 « DELAF » du Laboratoire d'Automatique Documentaire et Linguistique (Université de Paris 7). Cf. Gross, 1975, 1986 ; Gross et Senellart, 1998.

- ✓TreeTagger - a language independent part-of-speech tagger

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

- ✓Lemmatiseur Lexique 3 (New, Pallier & Ferrand, 2005)

Regroupements (SR)

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

Analyse lexicale : 2. partition

- La statistique mesure des différences
  - Comparaison de modalités de variables
  - Echantillonnage

- Hypothèses
  - Approche hypothético-déductive avec variables pré-codées
  - Approche inductive et reconstructions de variables a posteriori après réorganisation de la matière textuelle

Tableau lexical *formes * parties*

En colonne, les parties (caractérisées par des variables et modalités)

En ligne, les formes: liste des mots du lexique issus de la segmentation et lemmatisation

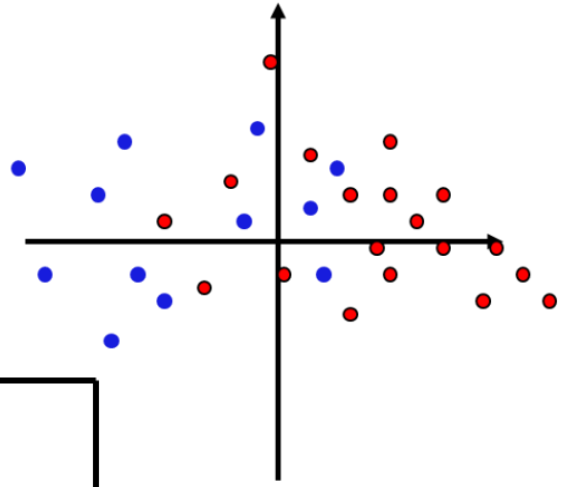
•Nombre d’occurrences

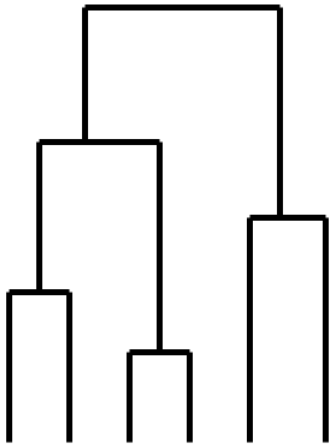
Tableau lexical *formes * parties*

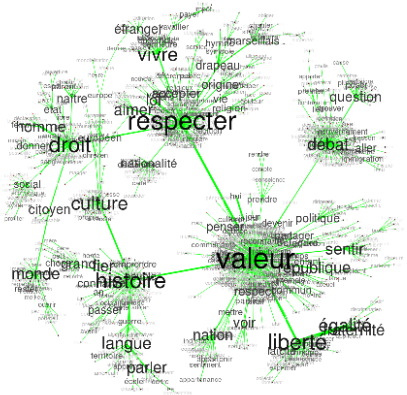
	LHumanité	LaCroix	LeFigaro	LeMonde	LePoint	Libération	NouvelObs
yougoslavie	1	4	3	2	0	1	0
considérable	13	10	21	4	3	8	1
controversé	5	6	14	5	3	3	0
sgen	4	2	0	5	0	1	0
garraud	7	3	0	1	0	0	0
originalité	2	9	7	2	2	2	0
chine	10	10	24	5	3	8	1
naturel	18	25	34	12	10	7	1
controverse	5	15	16	8	0	4	0
mener	81	68	80	53	17	46	2
projection	9	10	2	4	0	1	0
sensibilisatio	5	4	2	1	0	0	0
constitutif	2	7	5	2	0	1	0
lnder	0	14	0	0	0	0	0
commentair	15	25	16	16	3	4	0
radicalisatio	4	10	14	9	0	0	0
rythme	8	21	5	5	5	1	0
bienvenu	3	2	1	1	1	2	0
cependant	37	86	66	29	12	18	0
tisser	4	8	2	1	0	3	1
souci	17	51	20	19	6	6	1
prétendu	7	3	4	5	0	0	0
défiler	6	10	18	14	5	12	0
peur	41	85	85	44	12	38	5

Analyse lexicale 3 : statistiques

- *Formes* et réponses caractéristiques, ou **spécificités** (profil)
- Méthodes factorielles
- Classification automatique
- Cooccurrences et similitudes



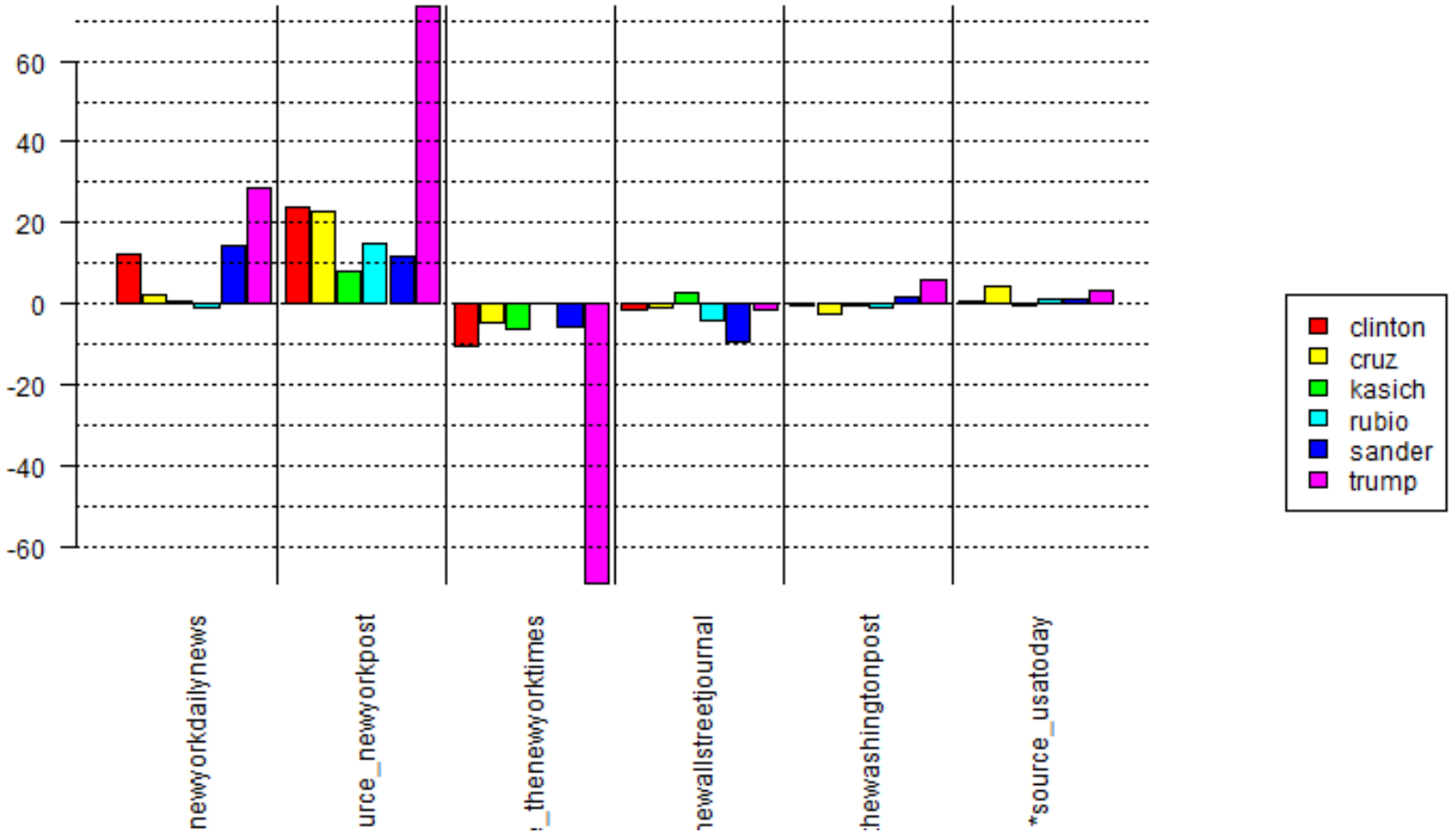




3.1. Les spécificités lexicales

- 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).

3.1. Les spécificités lexicales



3.2. La classification hiérarchique descendante

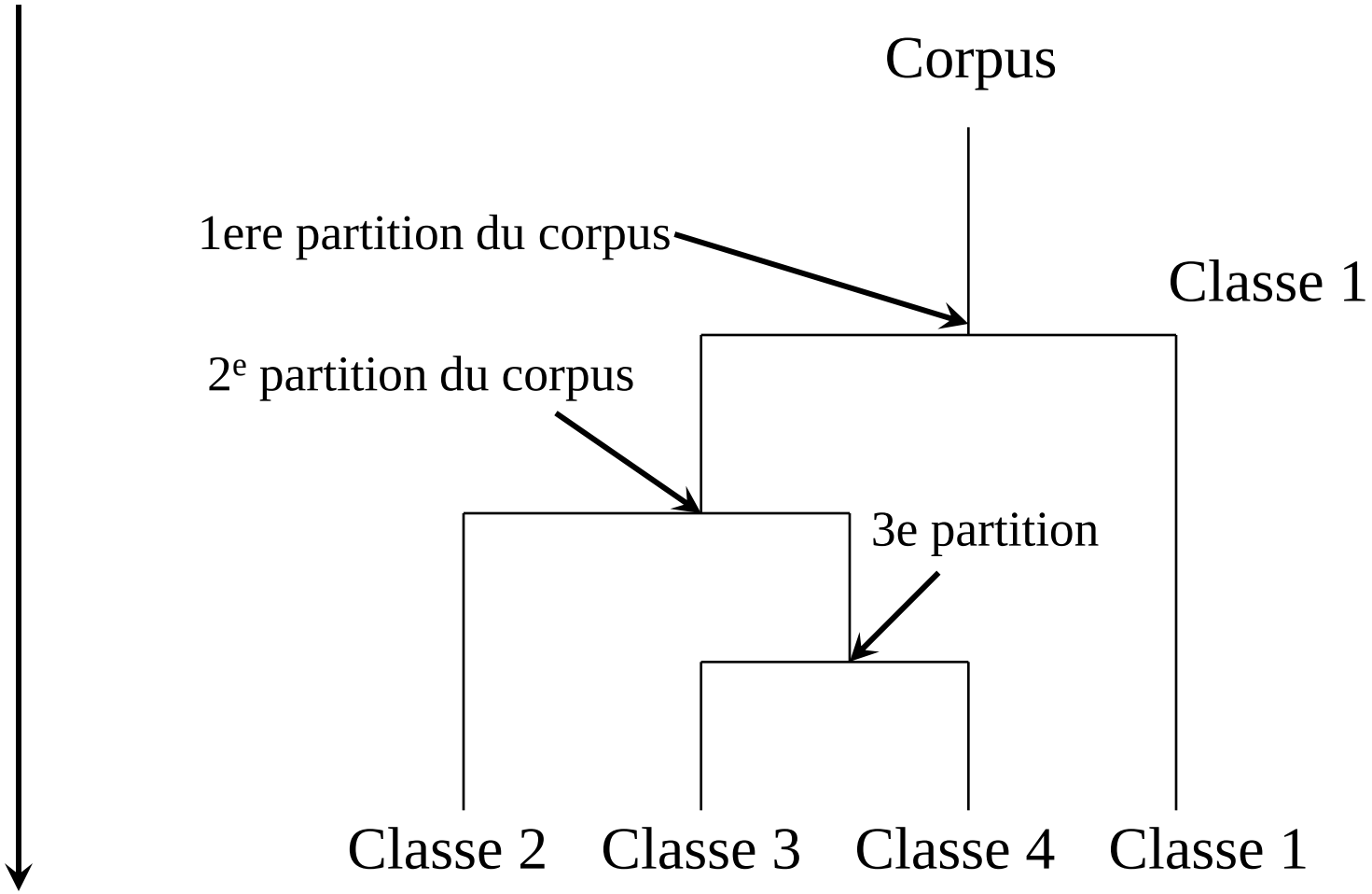


Tableau lexical

Parties = segments de texte, dits « unités de contexte »

Segment 1

Formes du lexique :

« Je ne veux pas être de la génération qui aura reporté le poids d'une dette excessive sur ses enfants et petits-enfants. Mon gouvernement sera celui de la responsabilité devant la jeunesse. Nous pourrions utiliser la situation que nous avons trouvée pour justifier des renoncements. Et bien non, nous ne renonçons à rien. Cette majorité n'a pas été élue pour trouver des excuses, mais des solutions. Je veux dire aux français la vérité. je veux leur dire ce que nous ferons. Je veux qu'ils puissent être juges à chaque instant des chemins que nous empruntons. »

Segment 2

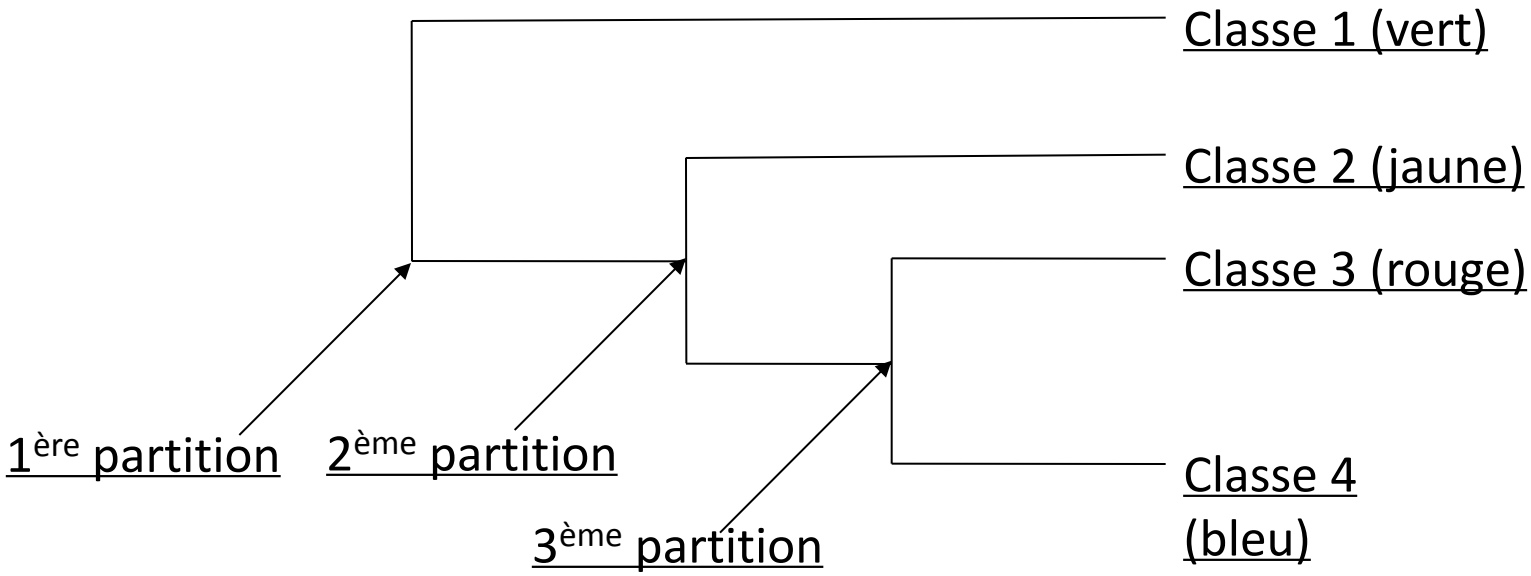
•Présence / absence

Extrait du discours de J-M Ayrault, 2012

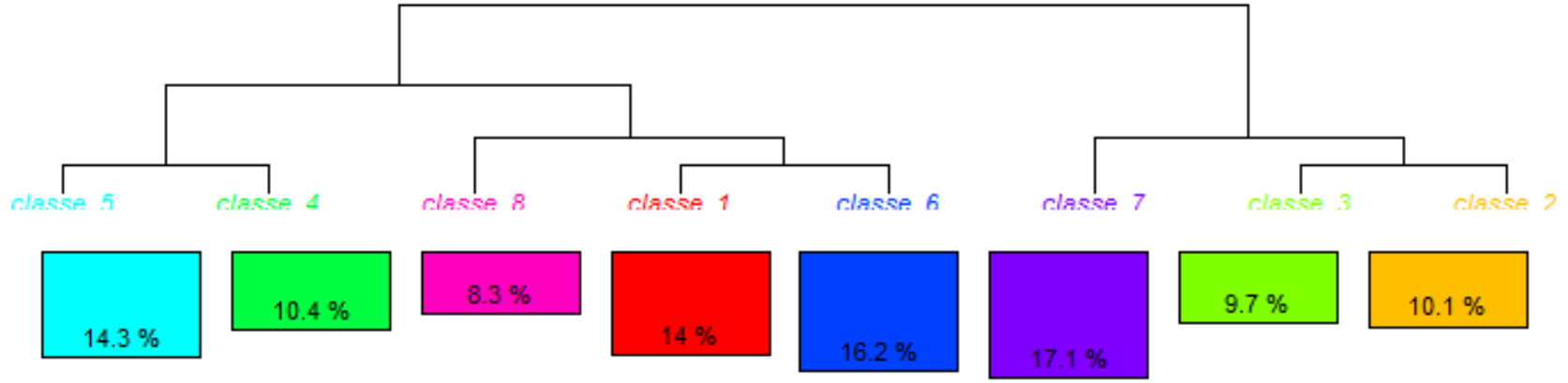
Segment 3

Classification lexicale (méthode Reinert)

Formes/ UC	a	b	c	d	e	f	g	h	i
france	1	0	1	0	1	0	1	0	0
travail	1	1	0	1	1	1	0	1	1
nation	1	0	1	0	1	0	1	0	0
politique	0	1	1	0	0	0	1	1	1
engagement	0	0	1	0	0	0	1	0	0
pouvoir	0	1	1	0	0	0	1	1	1
social	0	0	0	1	0	1	0	0	0
projet	0	1	0	1	0	1	0	1	1
partenaires	0	0	0	1	0	1	0	0	0



Dendrogramme et profils de classe



film
image
télévision
exposition
art
the
musée
esthétique
vidéo
série
cinéma
récit
spectateur
jeu
visiteur
fiction
réception
personnage
scène
of
festival
téléphone
émission
photographique
visuel
visite

presse
journaliste
journalisme
genre
journalistique
journal
médiatique
féminin
média
femme
discours
magazine
écrire
siècle
révolution
événement
opinion
propos
actualité
libération
lectorat
électoral
sexuel
orange
république
présidentiel
profane
affaire

territoire
territorial
local
développemen
durable
mondialisation
touristique
région
pays
tourisme
société
mondial
montagne
collectivité
patrimoine
habitant
urbain
occidental
économique
tic
citoyenneté
ville
intelligence
etat
échelon
gouvernance
nord

communica
approcher
socio
théorique
notion
dimension
social
humain
communicatior
concept
sémiotique
organisationnel
cognition
pragmatique
signification
changement
sentir
culture
théorie
organisation
questionner
anthropologie
problématique
processus
penser
appréhender

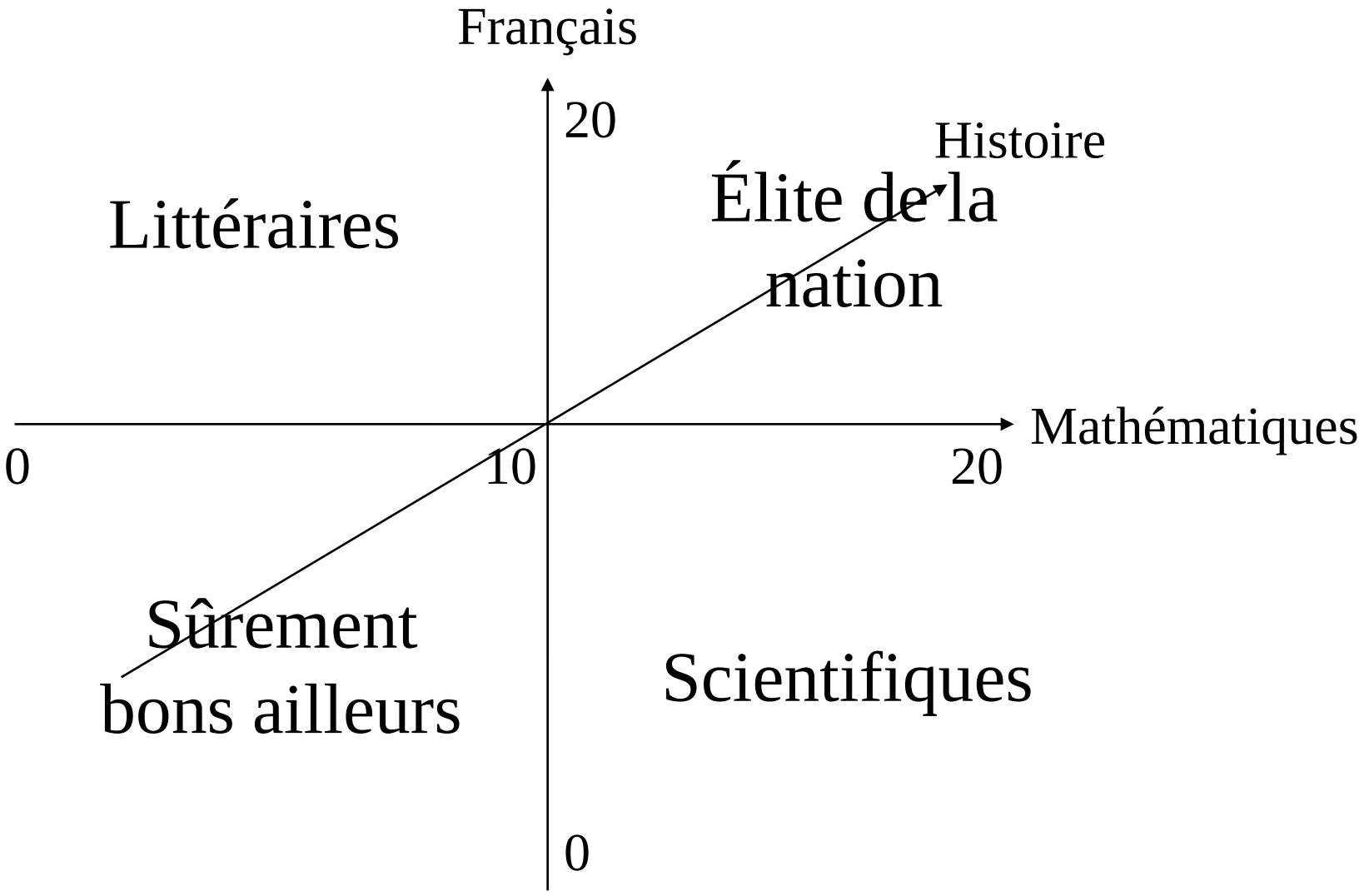
public
espace
action
individu
débat
collectif
norme
discussion
forme
acteur
citoyen
participation
dispositif
communicant
social
tension
logique
soin
organisation
prendre
sphère
négociation
activité
échange
participer
engager
santé

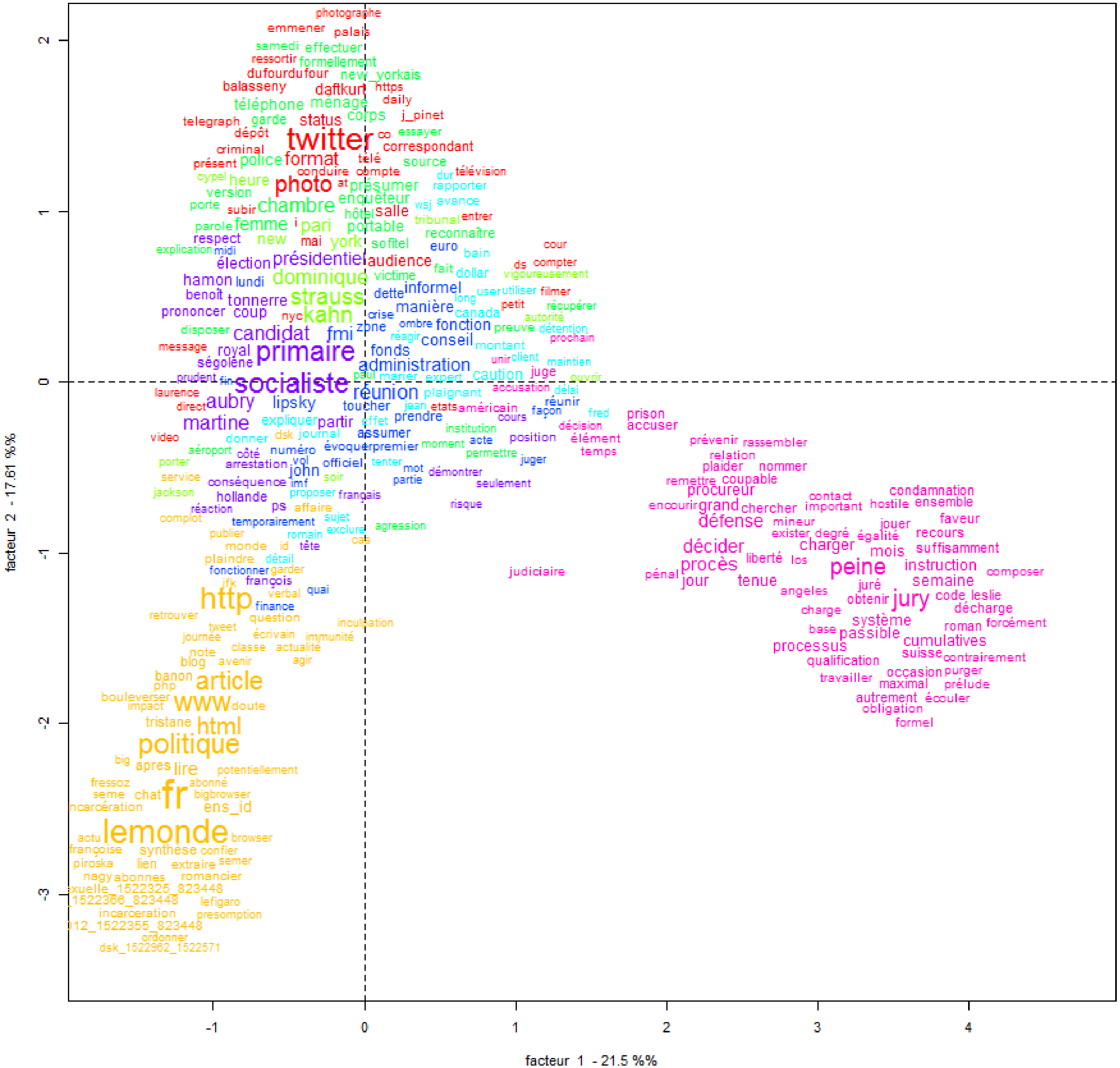
archive
scientifique
libre
université
bibliothèque
accès
chercheur
publication
ouvert
revue
universitaire
dépôt
édition
électronique
cnrs
enseignement
open
enseignant
archivage
publier
formation
gris
documentation
france
institut

méthode
présenter
analyse
résultat
corpus
statistique
utiliser
classification
évaluation
page
donnée
web
recherche
méthodologie
obtenir
moteur
proposition
exploratoire
base
cartographie
indicateur
article
approcher
bibliométriques
critère
calcul
extraction
proposer

utilisateur
système
document
information
connaissance
outil
apprentissage
améliorer
base
donnée
besoin
modélisation
annotation
faciliter
gestion
application
traitement
sémantique
métadonnées
informatique
utilisation
xml
structure
automatique
interface
opérationnel
exploitation

3.3. Analyse des correspondances

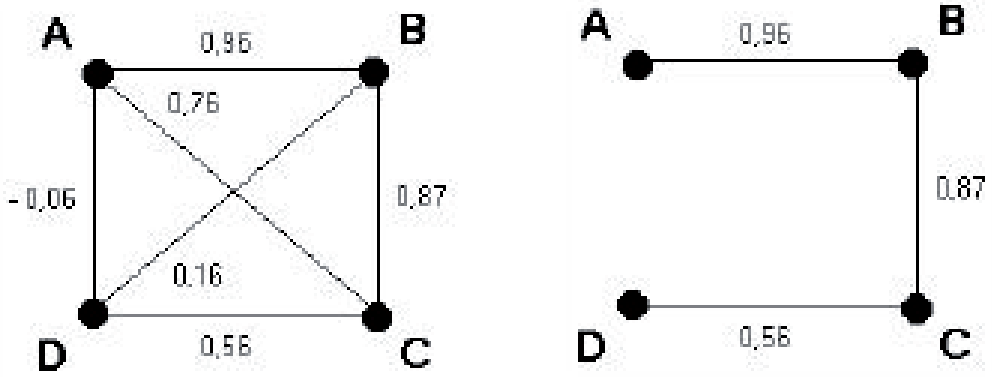




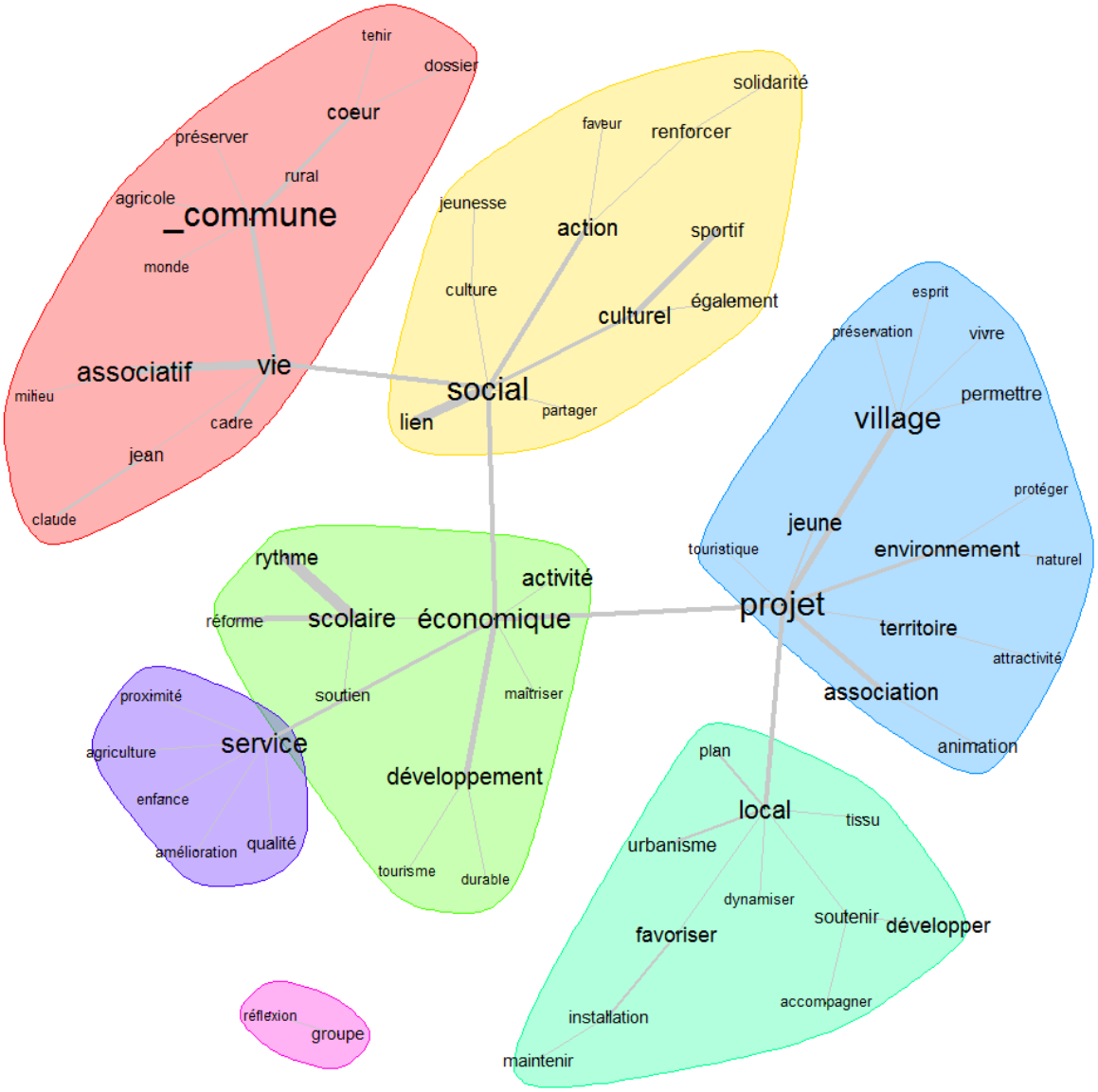
facteur 1 - 21.5 %%
facteur 2 - 17.61 %%
photographe
emmener
palais
samedi
effectuer
ressortir
formellement
dufourdufour
new_yorkais
balasseny
daftkurt
https
téléphone
ménage
daily
telegraph
garde
status
corps
j_pinet
dépôt
twitter
essayer
criminal
correspondant
présent
police
format
télé
source
cypel
conduire
compte
télévision
heure
photo
at
présumer
dur
rapporter
version
enquêteur
wsj
avance
porte
subir
chambre
hôtel
salle
entrer
parole
femme
pari
portable
tribunal
reconnaître
respect
new
mai
york
sofitel
explication
midi
euro
bain
cour
élection
présidentiel
audience
ds
compter
fait
dollar
vigoureusement
hamon
lundi
dominique
victime
benoît
informel
user
utiliser
filmer
tonnerre
strauss
dette
long
petit
récupérer
prononcer
coup
nyc
kahn
manière
canada
autorité
crise
disposer
candidat
fmi
zone
ombre
fonction
preuve
détention
réagir
conseil
prochain
message
royal
primaire
fonds
montant
ségolène
administration
unir
client
maintien
prudent
fin
socialiste
paul
marier
expert
caution
juge
ouvrir
laurence
réunion
plaignant
accusation
délai
aubry
lipsky
jean
etats
américain
réunir
direct
toucher
cours
façon
fred
prison
martine
expliquer
partir
effet
prendre
décision
accuser
video
donner
dsk
journal
assumer
institution
position
élément
aéroport
côté
numéro
évoquer
premier
moment
acte
temps
prévenir
rassembler
porter
arrestation
vol
officiel
tenter
permettre
juger
relation
john
mot
démontrer
plaider
nommer
service
soir
partie
jackson
conséquence
imf
seulement
remettre
coupable
hollande
proposer
français
procureur
condamnation
ps
affaire
risque
contact
réaction
complot
sujet
encourir
grand
chercher
important
hostile
ensemble
temporairement
défense
mineur
faveur
publier
romain
exclure
agression
jouer
cas
exister
degré
égalité
recours
monde
id
tête
décider
charger
mois
suffisamment
plaindre
détail
liberté
fonctionner
garder
judiciaire
pénal
procès
los
peine
instruction
composer
jfk
françois
jour
tenue
juré
semaine
verbal
quai
http
angeles
obtenir
jury
code
leslie
finance
décharge
retrouver
question
inculpation
charge
système
tweet
roman
forcément
journée
écrivain
immunité
base
passible
note
classe
actualité
cumulatives
processus
blog
avenir
agir
suisse
contrairement
qualification
banon
occasion
purger
php
article
travailler
maximal
prélude
bouleverser
www
doute
impact
autrement
écouler
obligation
tristane
html
formel
politique
big
apres
lire
potentiellement
fressoz
abonné
seme
chat
fr
bigbrowser
ens_id
lemonde
browser
actu
françoise
synthese
confier
piroska
lien
extraire
semer
nagy
abonnes
romancier
lefigaro
incarceration
presomption
ordonner
dsk_1522962_1522571

3.4. Cooccurrences et analyse de similitude

- L'ADS est une technique, reposant sur la théorie des graphes, classiquement utilisée pour l'étude des représentations sociales. Son objectif est d'étudier la proximité et les relations entre les éléments d'un ensemble, sous forme d'arbres maximum (Marchand, & Ratinaud, 2012)



3.4. Cooccurrences et analyse de similitude



Quelques références

- **Lebart, L. & Salem, A.** (1994). *Statistique textuelle*. Paris : Dunod.
- **Marchand, P.** (1998). *L'Analyse du Discours Assistée par Ordinateur*. Paris : Armand Colin.
- **Reinert, M.** (1990). ALCESTE - Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de Méthodologie Sociologique*, 26, p. 24-54.
- **Ratinaud, P., & Dejean, S.** (2009). IRaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009), Toulouse, France.