Journée R
MUSEUM NATIONAL D'HISTOIRE NATURELLE,
vendredi 24 mai 2013

Analyse de textes avec des outils R de statistique textuelle

Bénédicte Garnier (garnier@ined.fr)
Service méthodes statistiques
Institut National d'Etudes Démographiques

Introduction

Projet
(7e PCRD)

EuroBroadMap
VISIONS OF EUROPE IN THE WORLD

www.eurobroadmap.eu

WP 2

Produire une vision de l'Europe à partir d'un point de vue non-centré sur l'Europe

Volonté des partenaires du projet de partager des savoir-faire, d'acquérir et de diffuser des méthodologies notamment dans le domaine de l'analyse spatiale et de la statistiques textuelle (15 équipes internationales)

- Questionnaire destiné à 9340 étudiants dans leur langue native (18 pays, 43 villes)
- Question (D2) : Quel sont les mots que vous associés le plus à « Europe ». Choisissez 5 mots au maximum
- Corpus simple mais original par son contexte de production
- Saisie en ligne, en anglais par chaque pays partenaire

des réponses données

Statistiques textuelles - Projet EuroBroadMap

- Connaissance en statistique, analyste textuelle, environnement informatique et logiciels très divers entre les partenaires du projet
- Moyens financiers limités
- Corpus en anglais
- Solution proposée

un langage libre, gratuit et multiplateforme : R

Les outils de statistique textuelle

Evolutions

Analyse de données textuelles (Wikipédia)

L'analyse de données textuelles (ou ADT) est une approche méthodologique des sciences humaines qui envisage les textes comme des données qui peuvent être analysées par un ensemble de manipulations informatiques.
Ces analyses, inspirées par la linguistique structurelle et l'analyse de discours, utilisent des approches qualitatives et quantitatives.
C'est-à-dire qu'elles cherchent le plus souvent à qualifier les éléments du texte à l'aide de catégories et à les quantifier en analysant la répartition statistique des éléments du texte.

Dans les années 1960 : Traitement des variables qualitatives (Benzecri, Escoffier) - Analyse factorielle (Etudes sur Phèdre)
Depuis les années 1980 (Lebart, Salem) : Interaction entre linguistique, analyse du discours, statistique, informatique et traitement des questions ouvertes dans les enquêtes

Deux familles d'application :

Lexicométrie/Analyses factorielles et classifications

- Comparaison de textes sur la base d'une étude quantitative du vocabulaire = s'intéresser à la forme des textes en faisant abstraction de leur contenu
  - Attributions d'écrits historiques ou littéraires, comparaison du style de différents auteurs, etc. (Corneille/Molière)

- Analyse du contenu des textes pour en extraire le sens et mettre en évidence une structure dans les données
  - Traitement des réponses aux questions ouvertes, analyse d'entretiens, discours, trajectoires, etc.

Les méthodes de statistique textuelle indispensables aux outils

- Identifier les mots à analyser
  - Corpus et Lexique
  - Réduire le vocabulaire
- Créer les tableaux lexicaux
- Analyses factorielles et Classifications
- Calculer le vocabulaire spécifique

Comparatif en 2010

Garnier B., Guérin-Pace F. 2010. (Les Clefs pour)Appliquer les méthodes de la statistique textuelle Paris, CEPED, 86 p.

	Spad 7	Dimvic 4.3	Trideux 5	Lexico 3	Alceste 4.8	Tm de R
Initiateurs	L. Lebart, Morineau	L. Lebart	Ph Cibois	A. Salem	M. Reinert	I. Feinerer, K. Hornik, D. Meyer
Année (mise à disposition)	1993	2005	1986	1990	1986	2008
Type de texte	réponses à des questions ouvertes	réponses à des questions ouvertes	réponses à des questions ouvertes	réponses à des questions ouvertes et textes courts	Tout type	Tout type
Mise en forme (préparation du corpus)	simple	assez fastidieuse	assez fastidieuse	fastidieuse	assez fastidieuse	simple
Lemmatisation	manuelle	sans	sans	assistée	automatique	automatique
Découpage des textes	manuel (si nécessaire)	manuel (si nécessaire)	manuel (si nécessaire)	manuel (si nécessaire)	automatique	manuel (si nécessaire)
Site fournisseur	www.spad.eu	http://www.dtmvic.com	http://pagesperso-orange.fr/cibois/SitePhCibois.htm	www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW	http://www.image-zafar.com/	http://www.r-project.org/
Point fort	Plans factoriels dynamiques	Complet	Simple	Cartographie des données chronologiques	Lemmatisation	*Open source*
Prix	Environ 700€ (Prix recherche)	gratuit	gratuit	Env 600€ (licence simple) 3000€ (multipostes) V1 et 2 gratuites	1000 à 3000€ (standard ou entreprise selon la taille des corpus à traiter)	Gratuit
Interface graphique	excellente	bonne	moyenne	moyenne	bonne	excellente
Présentation des résultats	excellente	bonne	bonne	bonne	bonne	bonne

http://textometrie.ens-lyon.fr/

Dans la même logique :
TXM (Heiden, Magué & Pincemin, 2010)

Fonctionnalités du package tm

- **Data preparation** : importing, cleaning and general preprocessing

Intègre des options permettant de rapporter des mots à leurs radicaux ou d'enlever des mots communs comme les articles (lemmatisation)

- **Stemming** : Synonyms (wordnet package)
- **Dictionary** (> (d <- Dictionary(c("prices", "crude", "oil"))

Comptage de mots, calcul d'associations et création des tableaux lexicaux

- **Analysis** : finding associations for a given term based on counting
  - Co-occurrence frequencies, Filters
  - Creating Term-Document Matrices (TLE)
  - > findFreqTerms(),> findAssocs()
  - Clustering

Le package tm : essai de vulgarisation

Dans le cadre du projet EuroBroadMap

(Meeting de février 2010)

- → Package tm et Rweka, Java, RODBC, Snowball, Rcmdr, FactoMineR, dynGraph
- → Mise à diposition de tutoriels
  - Textual Statistics Methods for exploring D2 question
  - Hows to install R (SetupR.exe)
  - R functions for text mining

R functions for text mining (Morand, Taché, Garnier)

Données sous forme de tableau Excel

Mise à disposition d'un *programme-type* (.r) contenant les fonctions crées

- fvocabulary() ; fvocspec()

Exemple : **fvocabulary(folder="***D:\\myfolder***\\", data_workbook_Excel = "***mydatawb***.xls", data_sheet_Excel = "***sheet1***", vector_variables_words = c(***"ID_D2_1","ID_D2_2","ID_D2_3"***), byFreqDecr =T)**

- fReturnTLE() ; freturnTLA()

analyses factorielles/classifications sous FactomineR

R.TeMiS

Présentation de R.TeMiS [R Text Mining Solution]

30 juin 2012

Par Gilles Bastin

http://mediacorpus.hypotheses.org/104

R.TeMiS [R Text Mining Solution] est un environnement graphique de travail sous R permettant de créer, manipuler et analyser des corpus de textes. Il a été conçu pour limiter les effets de « boîte noire », souvent inhérents aux logiciels de statistique lexicale, et favoriser la réflexivité dans l'usage sociologique des données textuelles.

L'architecture statistique de l'environnement **R.TeMiS** est fournie par le paquet tm développé par Ingo Feinerer (Feinerer, 2008; 2011; Feinerer, Hornik & Meyer, 2008). Celui-ci a été complété par d'autres paquets classiques de R comme ca pour la représentation des analyses factorielles des correspondances (Nenadic & Greenacre, 2007). Enfin des paquets spécifiques ont été développés pour faciliter l'usage de **R.TeMiS** dans le domaine des études sur les médias, par exemple pour la gestion des corpus constitués depuis la base de données d'articles de presse Factiva.

Afin de faciliter l'usage de **R.TeMiS** aux néo-utilisateurs de R, le développement d'un environnement graphique a été privilégié. Celui-ci se présente donc comme un menu de R Commander (Fox, 2005).

R.TeMiS a été développé par Milan Bouchet-Valat et Gilles Bastin. Il est encore en cours d'amélioration mais depuis cette page vous pouvez télécharger la présentation faite à Quanti-Lille le 3 juillet 2012 [PDF] ainsi que le corpus de démonstration "Assange(FR)" [dossier de fichiers HTML].

- Menu intégré dans la fenêtre R Commander
- Commandes affichées dans la fenêtre de script (possibilité d'y intervenir et de les sauvegarder)
- Résultats dans la fenêtre de sortie et dans un fichier html

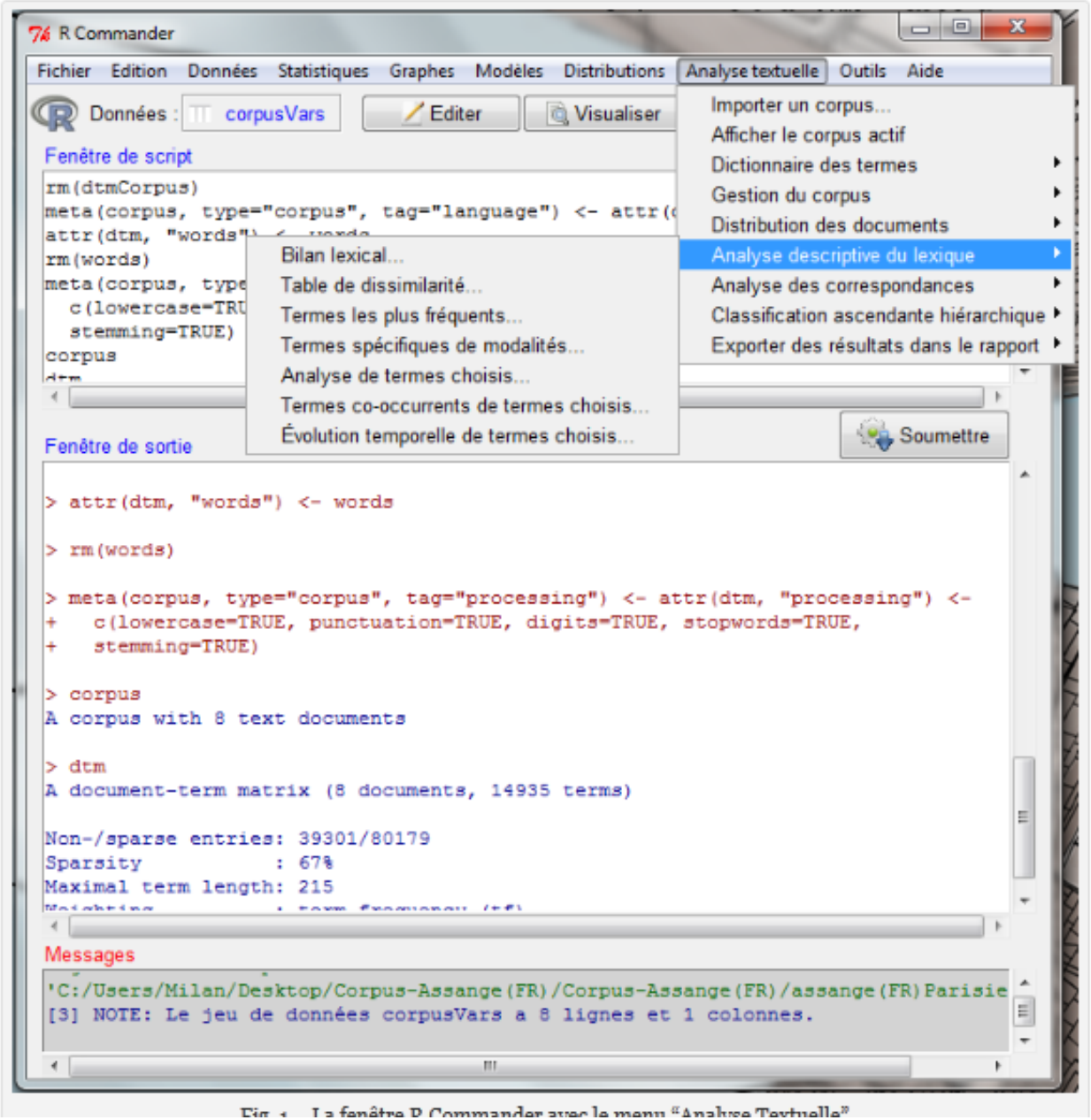


Fig. 1 – La fenêtre R Commander avec le menu "Analyse Textuelle"

Voir sur la page : http://mediacorpus.hypotheses.org/104

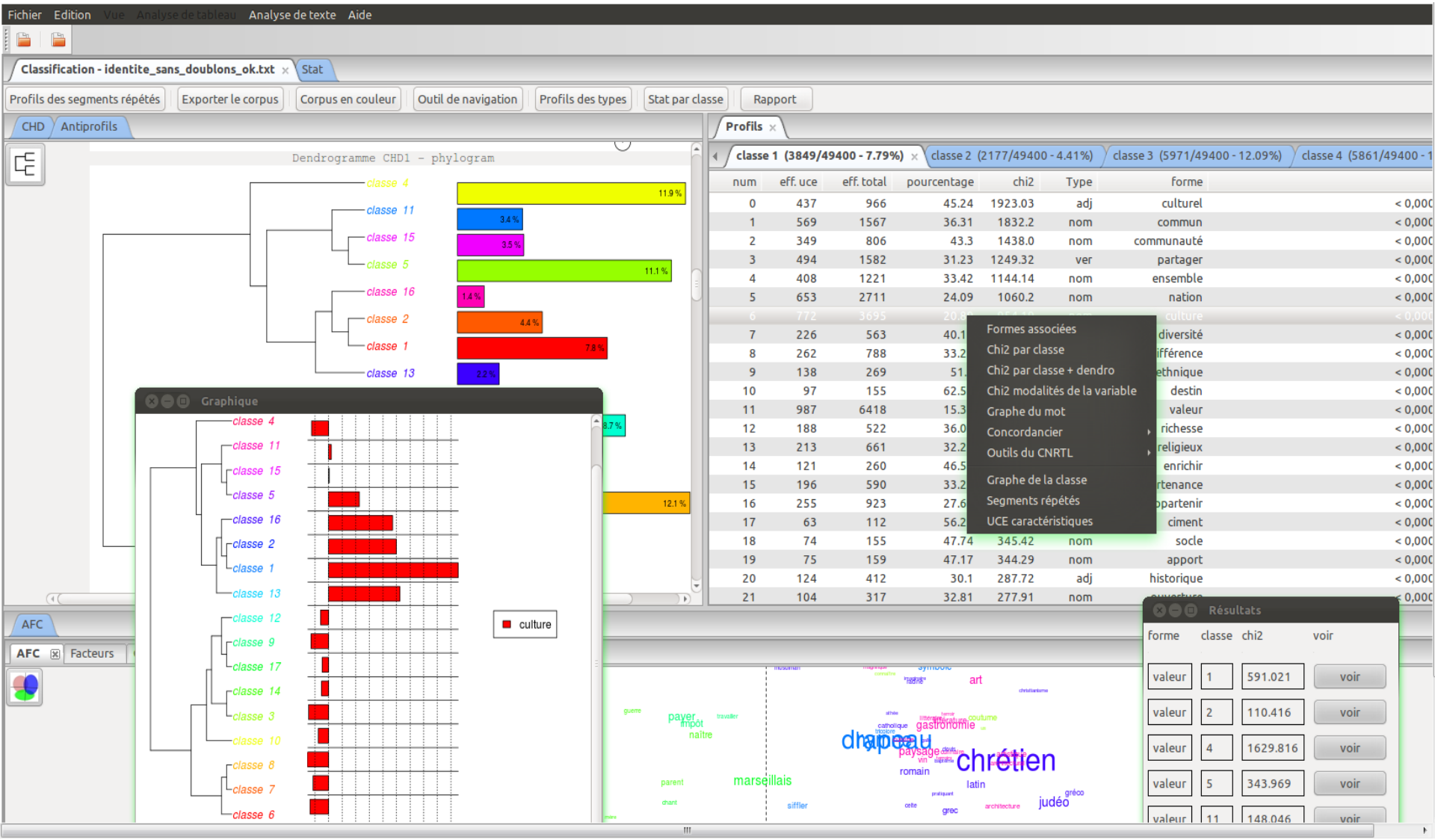
R.TeMiS

Facile à utiliser pour un débutant en R (R Commander)

Les fonctions d'analyse textuelle sont basées sur le package tm

Mais il existe aussi d'autres logiciels comme IraMuTeQ (plus spécifique; utilisant R mais non intégré)

IraMuTeQ Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires



http://www.iramuteq.org/

Les données textuelles

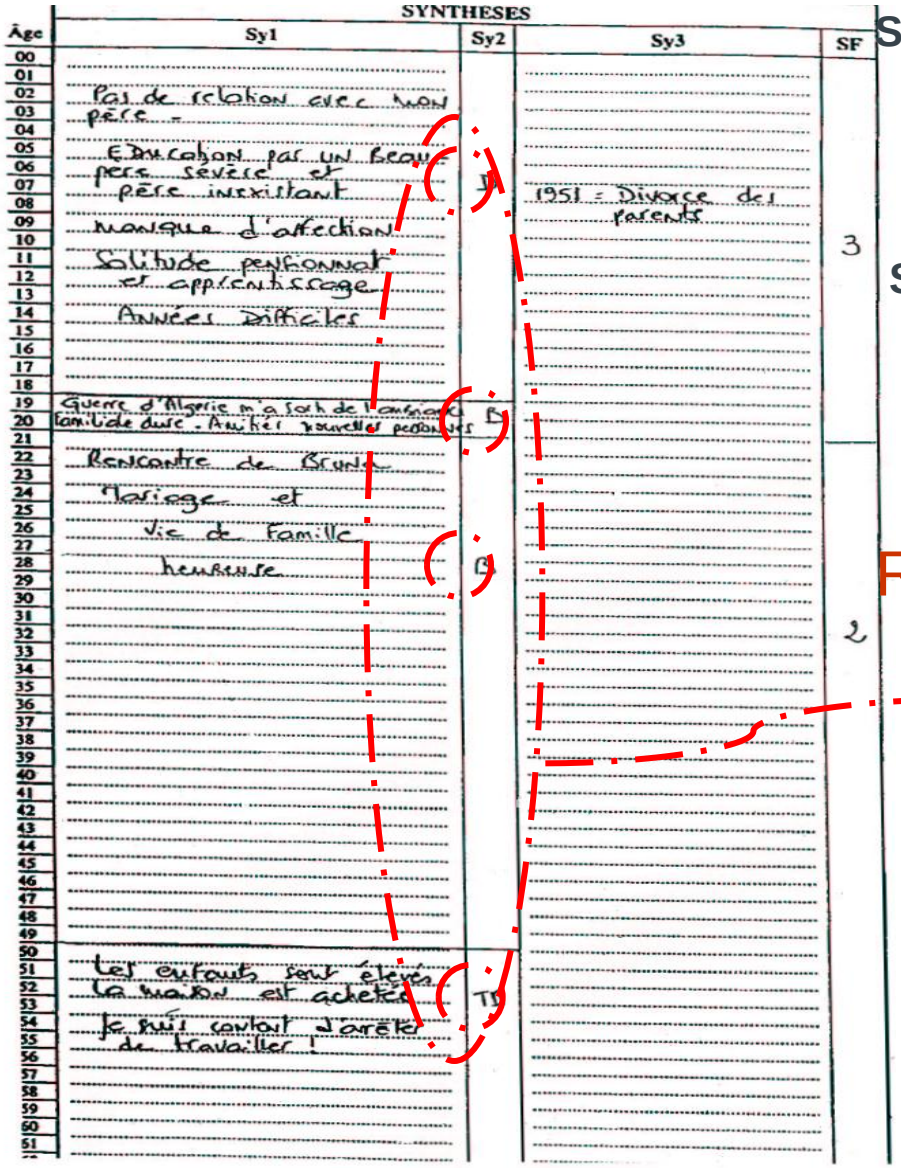
Corpus

- Pièce(s) de théâtre, ouvrages littéraires
- Discours politique(s), Pages Web

- Entretien (un ou plusieurs)
- Réponses à une (ou plusieurs) questions ouvertes
- Mots associés

Enquête Ined : Biographies et entourage

Découpage de moments de la vie en périodes et en tonalités



Sy1-Pouvez-vous découper votre vie en époques différentes ?

Caractériser ces périodes, en particulier, en identifiant ce qui les différencie les unes des autres et ce qu'elles représentent dans votre vie.

Sy2-Pour chacune de ces époques, était-ce :

- **TB** de très bonnes années
- **B** de bonnes années
- **SP** des années sans problème
- **D** des années difficiles
- **TD** des années très difficiles

Rythmes factuels et rythmes perçus

TRAJECTOIRE DE VIE :

Difficile-Bien-Bien-Très Bien

4 périodes

3 tonalités

Biographies et entourage, 2001

Mise en forme des textes courts (pour R.TeMiS) (texte à analyser en première colonne)

Exemple de commentaires sur l'Enfance

discoursEnfance	ROW_LABEL	sexegof	nationf
enfance heureuse malgré 5 enfants, la pauvreté et la jalousie de mes sœurs	101	Femme	français
enfance heureuse enfance difficile due a l exode puis la vie en hôtel a paris	102	Femme	français
enfance pauvre malheureuse, mon père était malade, je devais nourrir mes frères et sœurs	103	Homme	autre
enfance heureuse	104	Homme	français
enfance heureuse, on était pauvre, on marchait pieds nus mais on s'entendait très bien avec les arabes	105	Homme	français
jeunesse heureuse pas de soucis	106	Femme	français
enfance sans problèmes	107	Homme	français
petite enfance je ne m'entendais pas trop avec mes parents j'étais confinée dans deux pièces seule. c'était étouffant mes parents étaient âgés je dialoguais peu bouffée d'oxygène avec les parents de mon amie chantal ils ne sortaient m'emmenaient en vacances parents de substitution	108	Femme	français
enfance merveilleuse, que de bon souvenir, beaucoup de chaleur familiale.	109	Femme	français
vie difficile, problèmes matériels, il faut toujours se serrer la ceinture	110	Homme	français
enfance sans problème, bonne entente familiale	111	Homme	français
enfance heureuse toujours très entoure	112	Femme	français

Méthodes et mise en œuvre

Mots à analyser

Exemple de description de corpus

Le vocabulaire varie selon les logiciels

- Nombre de Documents/unités textuelles : 2678
- Nombre de termes/mots : 35241
- Nombre de termes/mots distincts : 2544
- Pourcentage de termes distincts : 7,2%
- Nombre d'hapax : 1158

Corpus *Synthèse de la trajectoire de vie pendant l'enfance*
(Biographies et entourage, 2001) (R.TeMiS)

Le lexique

	Occ. globales	% global
de	2198	6.17
enfanc	1352	3.79
la	970	2.72
pas	837	2.35
Le	819	2.30
vi	719	2.02
et	718	2.01
période	710	1.99
heureux	600	1.68
tres	576	1.62
parent	454	1.27
je	414	1.16
avec	401	1.12
on	397	1.11
en	377	1.06
me	363	1.02
bon	339	0.95
adolescent	332	0.93
famill	311	0.87
était	309	0.87
ma	306	0.86
mon	290	0.81
per	279	0.78
difficil	269	0.75
san	269	0.75
guerr	264	0.74
une	261	0.73
dan	251	0.70
mer	250	0.70
mais	249	0.70
est	245	0.69
bien	244	0.68
anné	238	0.67
un	232	0.65

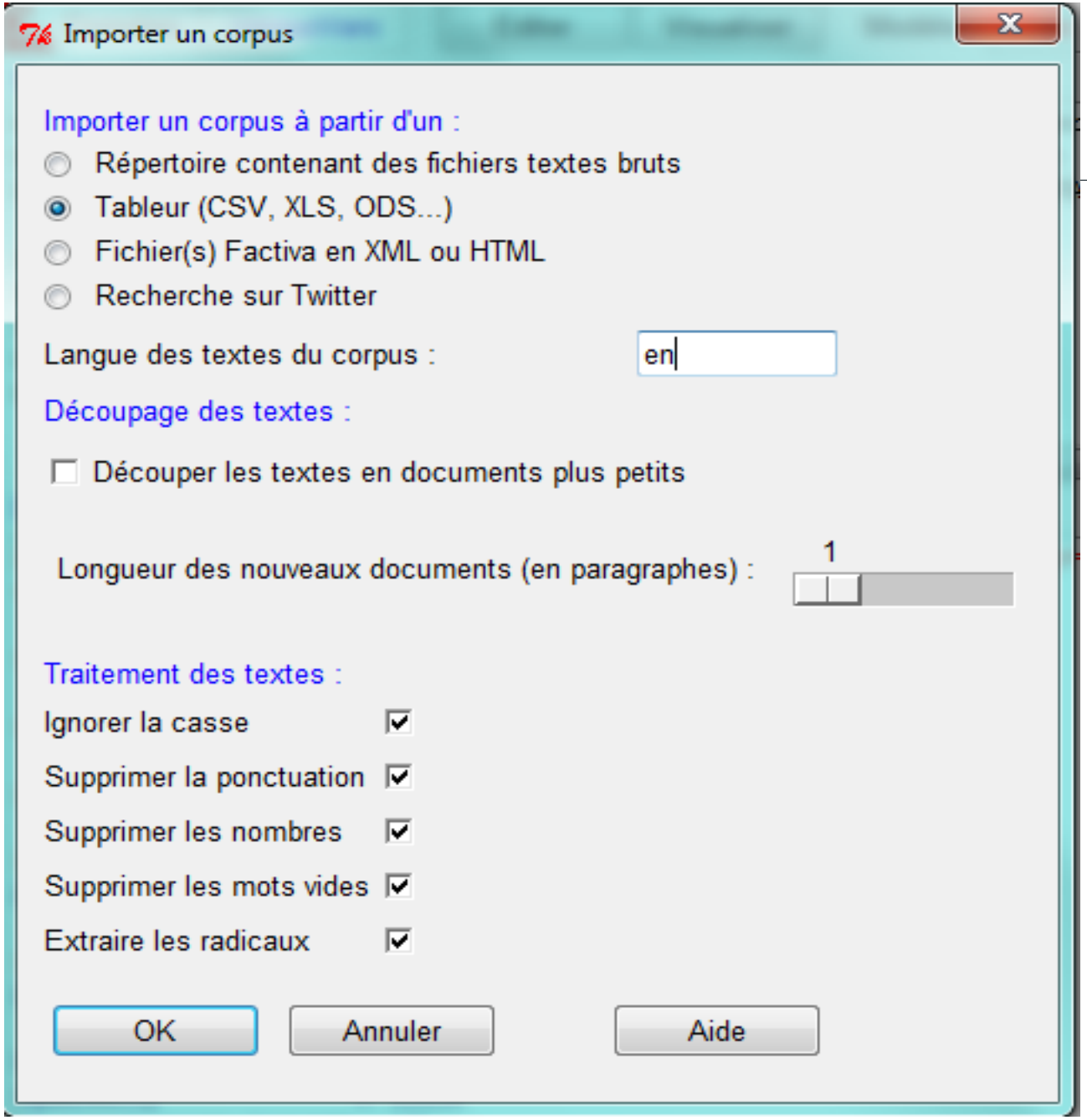
Extrait du lexique associé au corpus *Synthèse de la trajectoire de vie pendant l'enfance* (Biographies et entourage, 2001) (R.TeMiS avec racinisation)

Réduire le vocabulaire : la lemmatisation

- *Lemmatisation* = rattacher un ou plusieurs mots à une *forme* dite *racine* (Lebart, Salem, 1994)
- Des dictionnaires permettent de différencier les mots du lexique selon leur catégorie grammaticale (articles, prépositions, mots-outils, noms propres, verbes,…)
- Son utilisation divise les spécialistes
  - → Pour des corpus de taille importante (vocabulaire riche et varié)
  - → Pour augmenter la fréquence des mots (petits corpus)

	Occurrences	Racine	Occ. racine	Mot vide ?	Supprimé ?
de	1821	de	2178	Mot vide	
des	357	de	2178	Mot vide	
enfanc	1	enfanc	1345		
enfance	1343	enfanc	1345		
enfances	1	enfanc	1345		
la	958	la	958	Mot vide	
pas	831	pas	831	Mot vide	
le	375	le	812	Mot vide	
les	437	le	812	Mot vide	
vie	715	vi	715		
et	714	et	714	Mot vide	
période	684	période	706		
périodes	22	période	706		
heureuse	524	heureux	599		
heureusement	2	heureux	599		
heureuses	28	heureux	599		
heureux	45	heureux	599		
très	571	tres	571	Mot vide	
parent	7	parent	452		
parents	445	parent	452		
je	412	je	412	Mot vide	

Extrait de la « lemmatisation » Synthèse de la trajectoire de vie pendant l'enfance (Biographies et entourage, 2001) (R.TeMiS)


Importer un corpus
Importer un corpus à partir d'un :
Répertoire contenant des fichiers textes bruts
Tableur (CSV, XLS, ODS...)
Fichier(s) Factiva en XML ou HTML
Recherche sur Twitter
Langue des textes du corpus :
en
Découpage des textes :
Découper les textes en documents plus petits
Longueur des nouveaux documents (en paragraphes) :
1
Traitement des textes :
Ignorer la casse
Supprimer la ponctuation
Supprimer les nombres
Supprimer les mots vides
Extraire les radicaux
OK
Annuler
Aide

R.TeMiS

Les tableaux lexicaux

- Les logiciels de statistique textuelle transforment le corpus de textes initial en tableaux dits *lexicaux*
- Selon la structure du corpus et le volume des textes, il est parfois nécessaire de procéder à un découpage des textes en séquences de taille réduite appelées *unités textuelles*
- Selon les logiciels, le découpage des textes est automatisé ou à à effectuer manuellement

Extrait du tableau lexical entier (TLE/document-term matrix)

associé au corpus *Synthèse de la trajectoire de vie pendant l'enfance*

ROW_LABEL	adolescence	agréable	argent	bien	bons	campagne	copains	difficile	dur	décès	enfance	enfant	enfants	entourée	facile	famille	fille	france	frère	grands-parents	guerre	heureux	insouciance	je	jeunesse	liberté
101	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
102	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
103	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
104	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
105	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
106	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
107	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
108	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
109	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
110	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
111	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
112	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
113	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
114	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
115	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
116	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0

Tableau *hyper creux* (environ 95% de 0) Les lignes correspondent aux unités textuelles et les colonnes aux mots du lexique : Permet de repérer les cooccurrences des mots dans les réponses

Objet crée sous tm (R) : **document-term matrix** (2678 documents, 2544 terms)
Non-/sparse entries: 30900/6781932
Sparsity : 100%
Maximal term length: 13
Weighting : term frequency (tf)

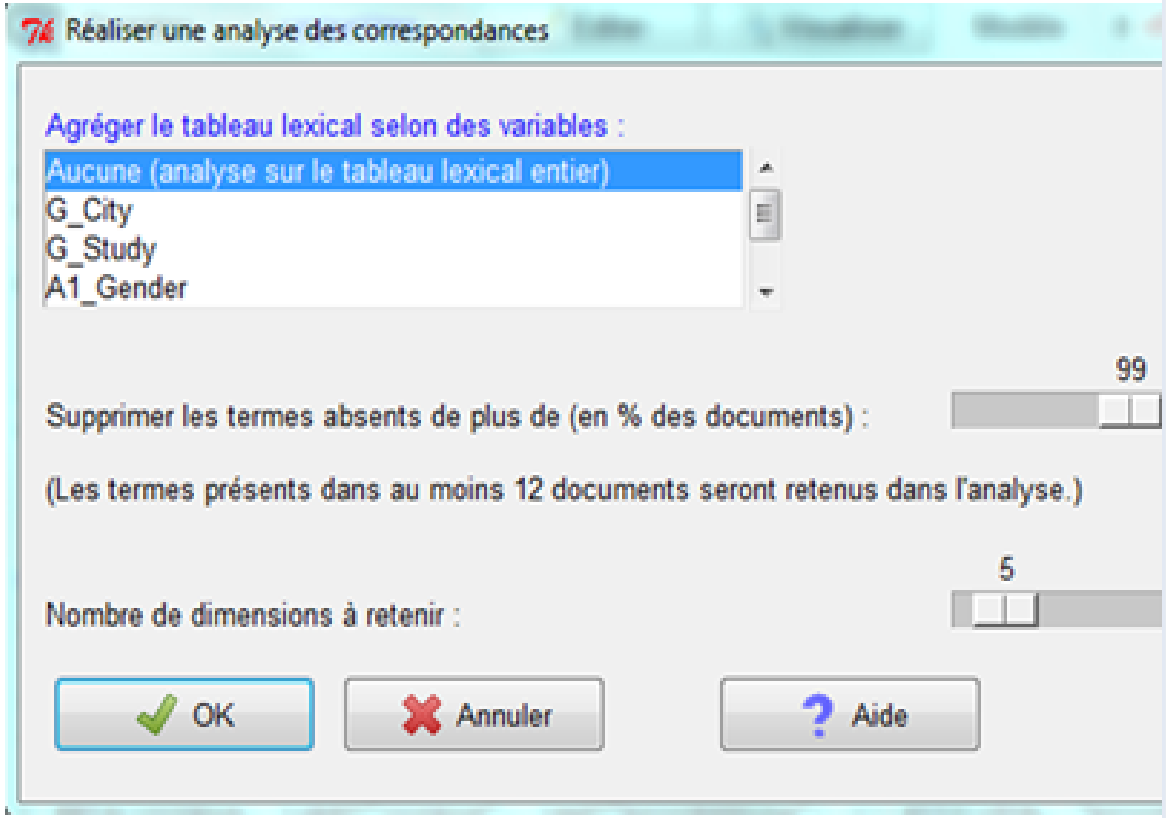
Extrait d'un Tableau Lexical Agrégé (TLA)

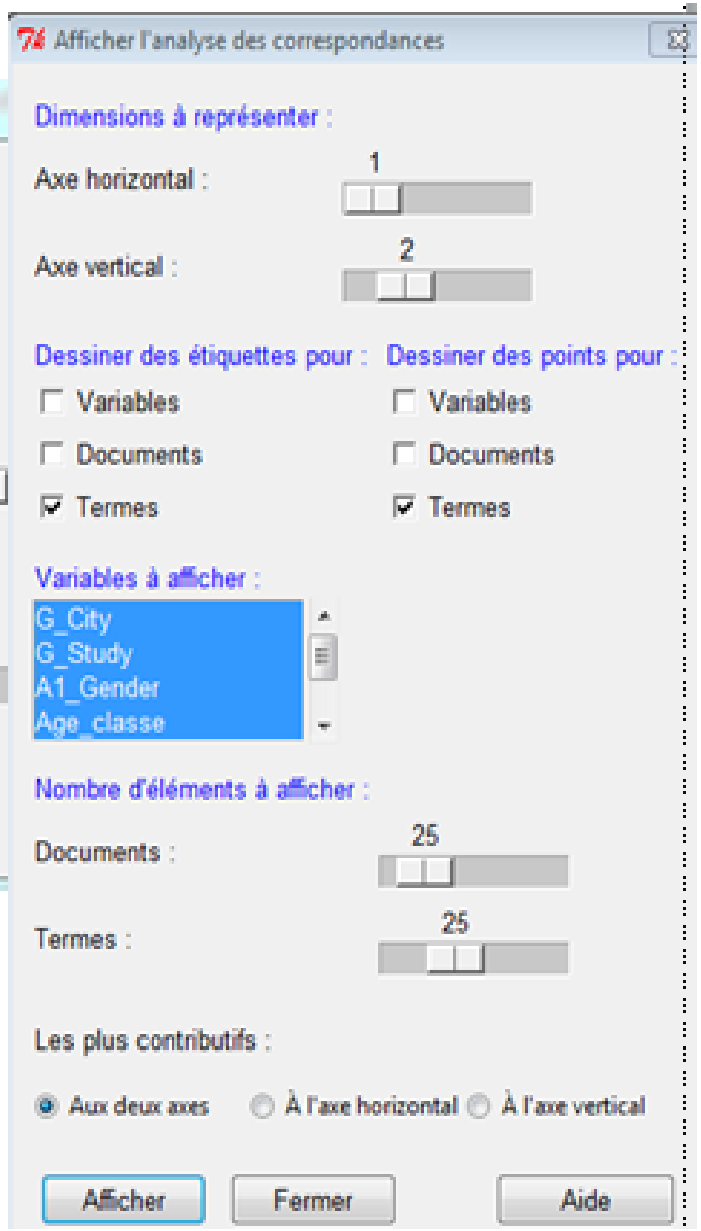
associé au corpus *Synthèse de la trajectoire de vie pendant l'enfance*

	trajE=EAutres	trajE=EB	trajE=EBD	trajE=EDB	trajE=EDP	trajE=ESP	Sex_H	Sex_F	….
adolescence	29	196	17	44	1	36			
agréable	3	48	6	2	3	7			
algérie	2	15	6	3	0	8			
argent	0	20	4	24	0	4			
bien	14	141	22	25	7	35			
bons	30	226	27	23	10	23			
campagne	8	32	5	11	2	7			
copains	5	24	1	2	0	4			
difficile	26	46	33	122	12	28			
dur	10	23	12	50	2	6			
décès	4	25	11	18	0	4			
enfance	102	774	101	191	16	163			
enfant	3	25	7	10	1	5			
enfants	5	25	3	10	0	12			
entourée	0	26	2	0	0	2			
facile	2	19	1	12	0	6			
famille	18	280	25	70	4	45			
fille	1	18	3	3	0	5			
france	5	11	4	9	0	4			
frère	7	43	18	24	2	9			
grands-parents	13	45	24	8	6	10			
guerre	39	83	34	69	13	24			
heureux	35	390	57	33	6	49			
insouciance	10	55	3	4	1	14			
je	97	372	69	167	25	87			
jeunesse	7	89	5	18	2	18			
liberté	6	19	4	6	2	3			
lycée	10	18	1	3	1	1			

Biographies et entourage, 2001

R.TeMiS : Menu Analyse des correspondances





AFC sur le TLE ou TLA (et choix des variables acives/illustratives)

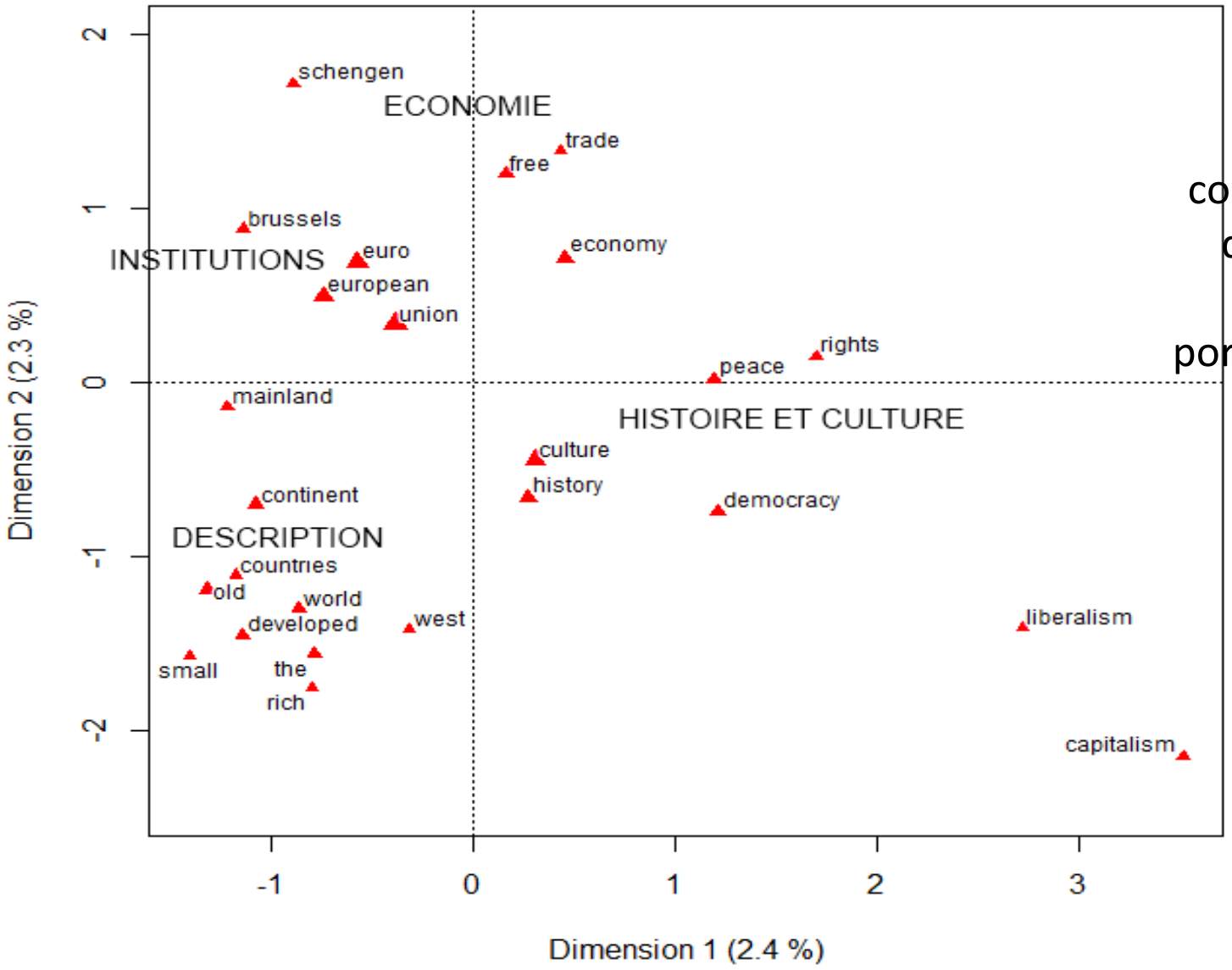
Analyse Factorielle des Correspondances sur le *Tableau Lexical Entier*

- Permet de structurer l'ensemble des *mots* en fonction de leur répartition dans les unités textuelles
- La représentation des résultats sous forme de *plans factoriels*, permet de visualiser les proximités de mots, les oppositions, les tendances, …

Deux mots seront d'autant plus proches que leurs contextes d'utilisation se ressemblent et d'autant plus éloignés qu'ils seront rarement utilisés ensemble

Les *cooccurrences* de mots ainsi mises en évidence permettront de *repérer des thèmes et de visualiser des oppositions entre ces thèmes*

AFC sur le Tableau Lexical Entier



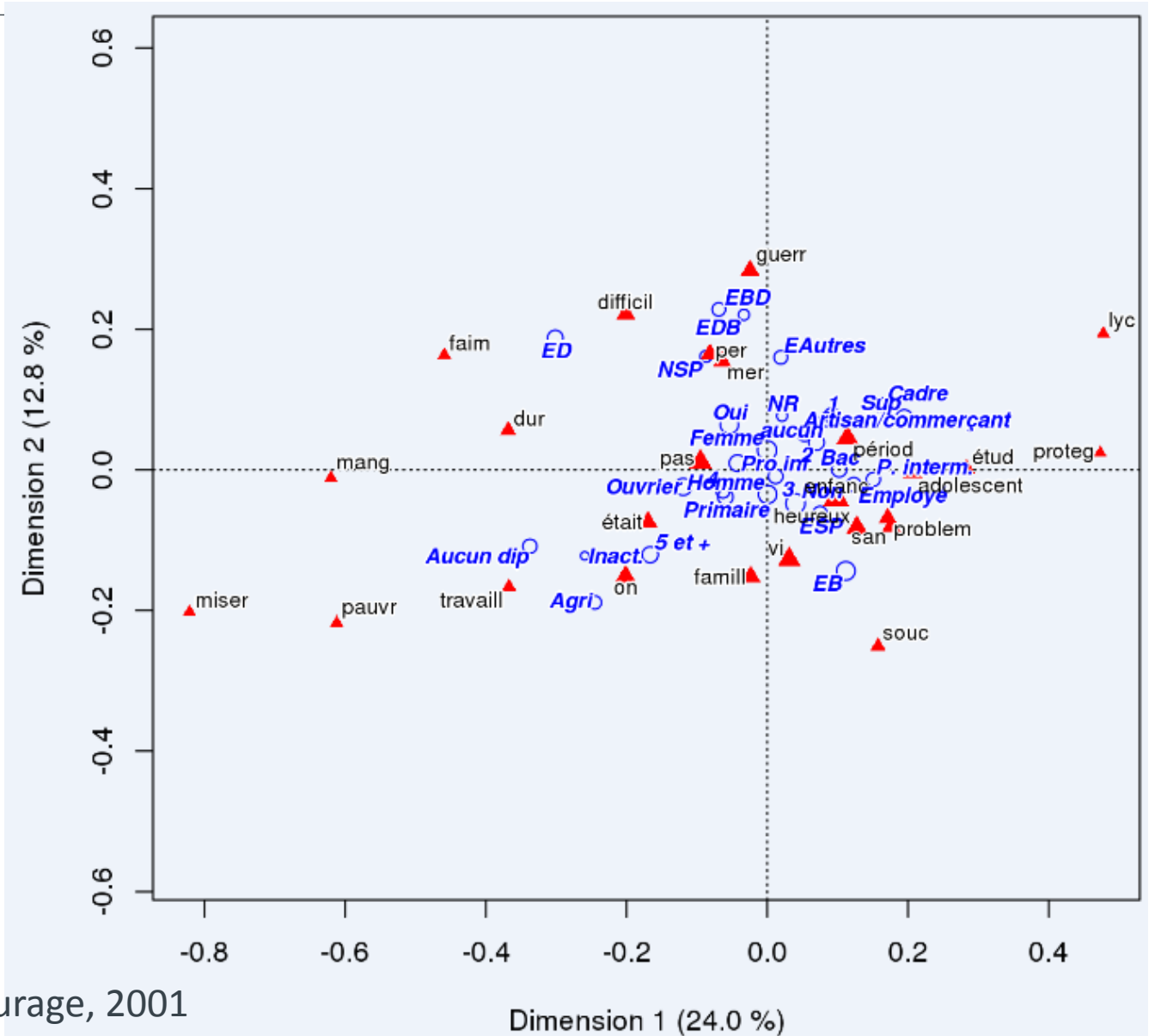
Exemple sur le corpus des réponses de 1 676 étudiants français, belges, portugais et hongrois (EuroBraodMap) Plan (1-2)

Analyse Factorielle des Correspondances sur le Tableau Lexical Agrégé

- Permet de structurer l'ensemble des « mots » en fonction des caractéristiques des locuteurs (par exemple)
- Les plans factoriels issus de l'analyse permet d'observer la position réciproque des mots et des variables sociodémographiques et d'interpréter les proximités en répondant à la question *Qui dit quoi ?*

AFC sur le TLA (trajE, PCS du père, taille de la fratrie, diplôme, avoir parlé de guerre)

Corpus "*Synthèse de la trajectoire de vie pendant l'enfance*"
Plan 1-2
R.TeMiS avec racinisation



Biographies et entourage, 2001

Résultats d'une l'AFC (Contributions et coordonnées) avec R.TeMiS

Termes les plus contributifs du côté négatif de l'axe 1 :

	Position	Contribution (%)	Qualité (%)
miser	-0.821	5.420	75.66
on	-0.201	4.294	47.26
dur	-0.368	4.175	69.81
pauvr	-0.612	3.711	73.00
mang	-0.620	3.298	86.92
difficil	-0.201	2.869	35.30
était	-0.170	2.363	59.44
travaill	-0.367	2.275	67.32
pas	-0.095	2.008	42.77
faim	-0.459	1.581	53.06
per	-0.082	0.484	14.63
mer	-0.064	0.274	8.14
famill	-0.023	0.045	1.01
guerr	-0.024	0.041	0.31

Voir également le contexte d'utilisation des mots (ici par exemple mang) pour interpréter/intituler les plans factoriels

Corpus "*Synthèse de la trajectoire de vie pendant l'enfance*"

Modalités actives du côté négatif de l'axe 1 :

	Position	Contribution (%)	Qualité (%)
ED	-0.30131	20.966850	57.6920
Aucun dip	-0.33742	17.143176	70.3647
5 et +	-0.16646	7.903675	40.4134
Agri	-0.24522	6.487044	31.6916
Ouvrier	-0.11947	5.613924	45.6362
Oui	-0.05388	1.494646	9.7375
Primaire	-0.05956	0.810966	7.2188
Pro inf	-0.04233	0.588195	8.1344
EBD	-0.06904	0.499548	2.8638
4	-0.06302	0.444591	5.2984
NSP	-0.08738	0.350161	2.8131
Inact.	-0.25949	0.108245	1.1966
EDB	-0.03321	0.031619	0.3659
Femme	-0.00038	0.000098	0.0011

Biographies et entourage, 2001

Aides à l'interprétation

Documents les plus éloignés côté négatif de l'axe 1 :

	Position	Qualité (%)
51	-5.3	3.3
249	-3.8	7.9
2662	-3.6	5.7
1358	-2.2	8.5
2041	-2.2	1.2
383	-2.1	1.8
1239	-2.0	13.8
2206	-2.0	3.5
40	-1.8	2.5
2430	-1.8	9.7
1266	-1.7	3.0

Corpus "*Synthèse de la trajectoire de vie pendant l'enfance*"

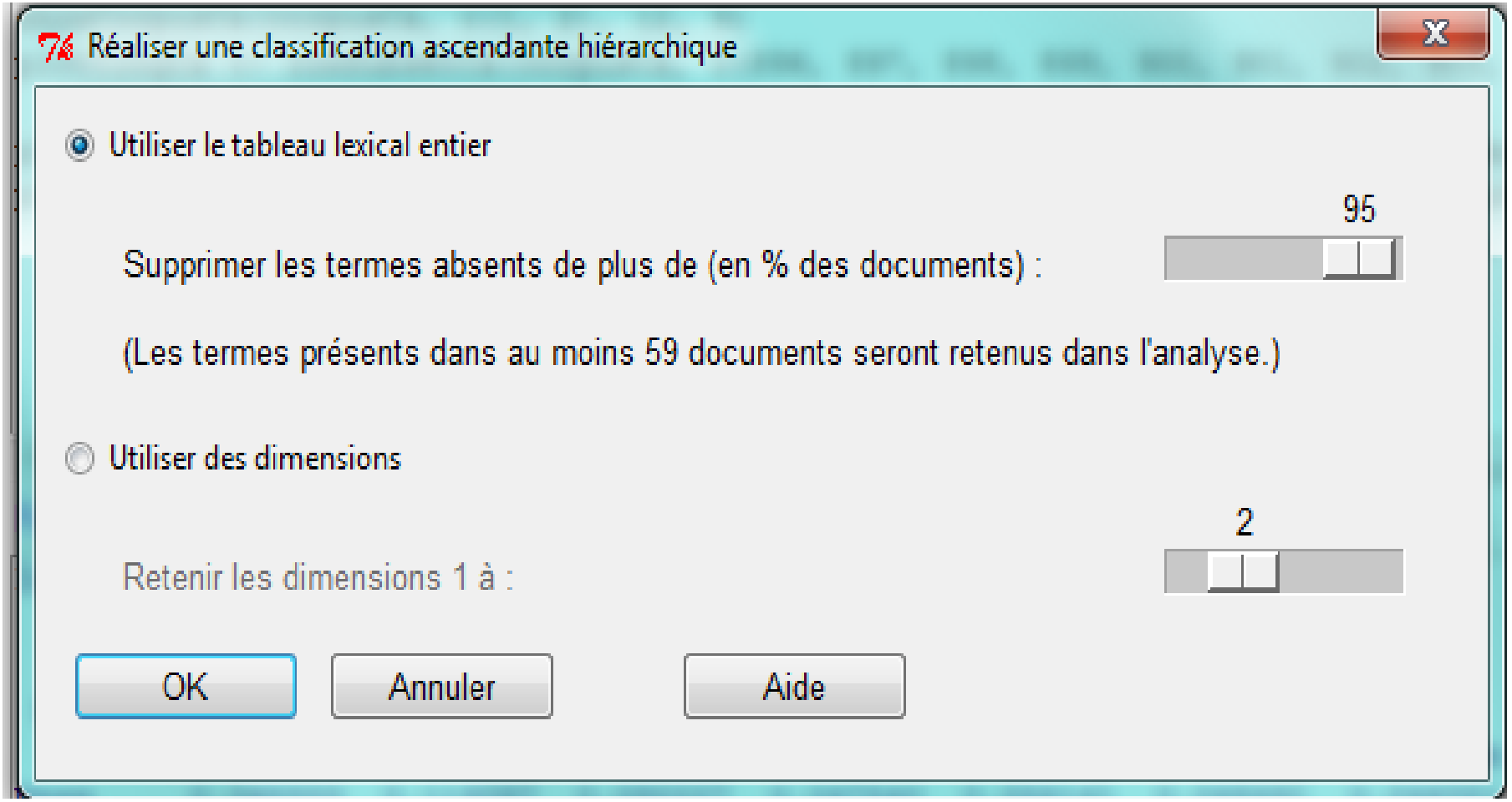
Documents spécifiques de l'axe 1

51
à manger à volonté
249
enfance pauvre, solitaire, misérable, famine
2662
la misère
1358
enfance misérable misère noire pas de père j ai mange de l herbe
2041
travail
383
travailler toujours

Les classifications automatique

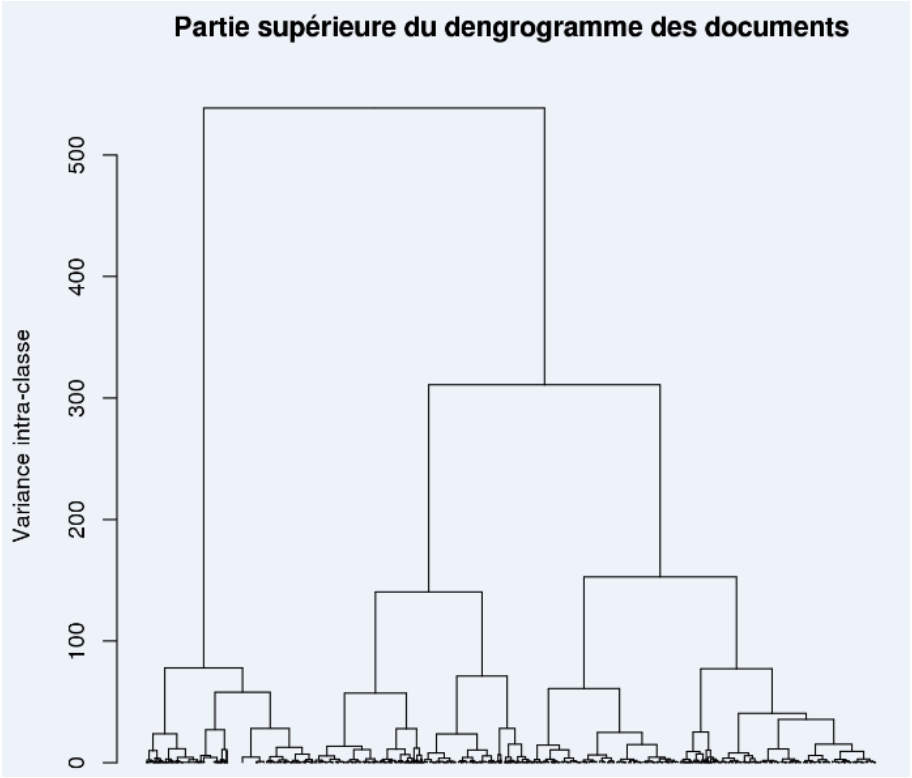
- Destinées à « mettre en évidence une dimension d'organisation du corpus de textes qui « mémorise » ses conditions de production» selon une partition des unités textuelles et à révéler des mondes lexicaux (Reinert, 1983)
- Directement sur le TLE (Document Term Matrix)
- Sur axes factoriels

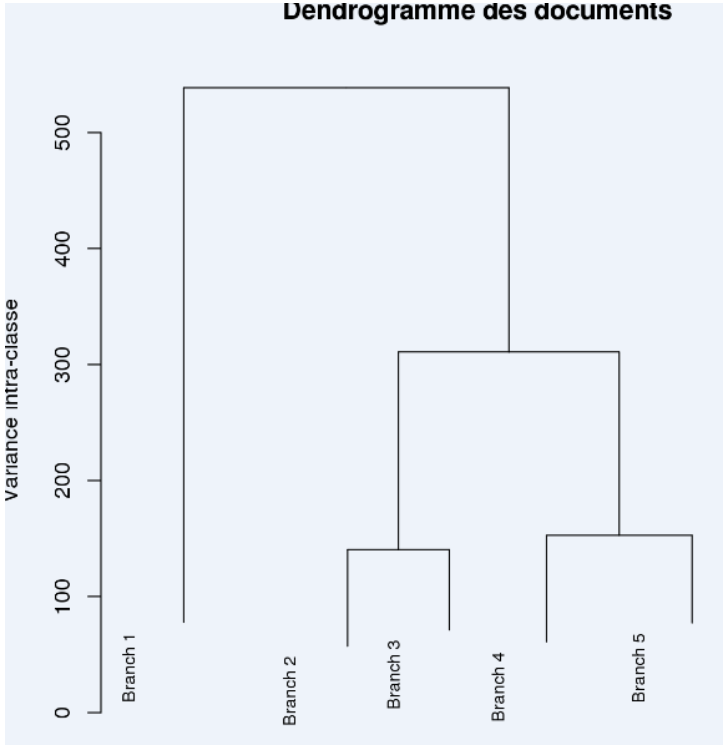
R.TeMiS :
Menu classification ascendante hiérarchique



Classification à partir des facteurs de l'AFC sur le TLA

On retient 5 classes



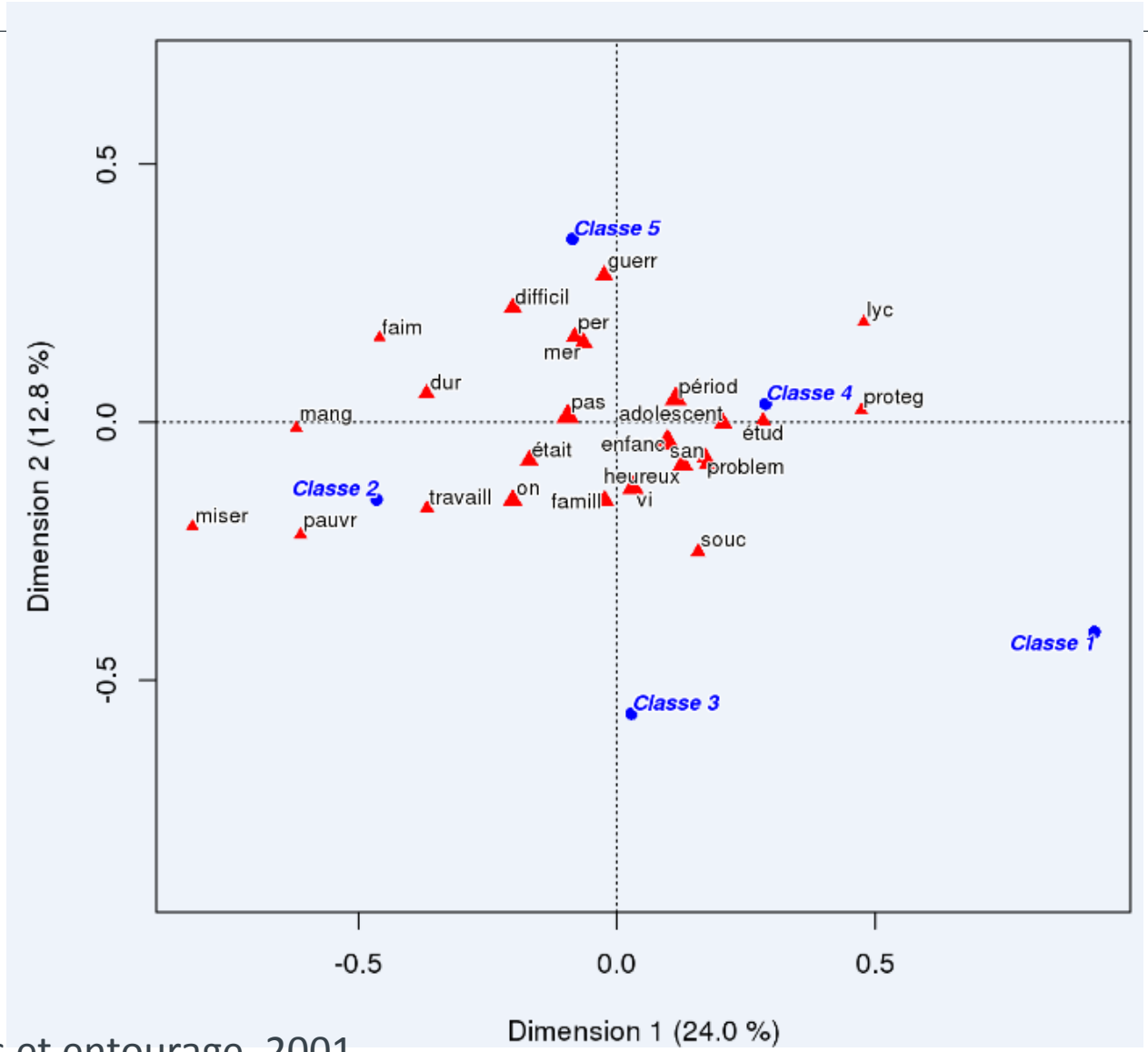


Résumé des classes :

	1	2	3	4	5
Nombre de documents	609.0	395.0	398.0	533.0	714.0
% des documents	23.0	14.9	15.0	20.1	27.0
Variance intra-classe	77.9	57.1	71.2	60.9	77.3

Biographies et entourage, 2001

Projection de la variable de classes sur le plan factoriel



Biographies et entourage, 2001

Aides à l’interprétation

- Vocabulaire spécifique

Le corpus est découpé selon les modalités d’une variable et le vocabulaire de chacun des sous-corpus (:modalité) est comparé au vocabulaire d’ensemble (global) afin de repérer les mots sur ou sous représentés (Test statistique)

- Contexte d’utilisation d’un mot

Aides à l'interprétation : Les spécificités lexicales

Vocabulaire spécifique de la modalité « B » (enfance bien) (R.TeMis)

On pourrait utiliser la variable de classe

	% terme/mod.	% mod./terme	% global	Modalité	Global	Valeur t	Proba.
heureux	2,47	68,11	1,70	408	599	0,00	0,00
enfanc	4,70	57,77	3,82	777	1345	8,04	0,00
bon	1,37	66,86	0,96	226	338	7,35	0,00
vi	2,44	56,50	2,03	404	715	5,12	0,00
adolescent	1,22	61,03	0,94	202	331	5,10	0,00
famill	1,14	60,84	0,88	188	309	4,86	0,00
difficil	0,28	17,36	0,75	46	265	-10,09	0,00
pas	1,90	37,91	2,36	315	831	-5,29	0,00
pension	0,04	12,73	0,16	7	55	-5,26	0,00
dur	0,17	24,35	0,33	28	115	-4,90	0,00

Termes spécifiques de la classe 1 corpus *Synthèse de la trajectoire de vie pendant l'enfance*

	% terme/mod.	% mod./terme	% global	Modalité	Global	Valeur t	Proba.
adolescent	5.032	40.18	0.94	133	331	Inf	0
enfanc	16.534	32.49	3.82	437	1345	Inf	0
étud	1.854	32.89	0.42	49	149	Inf	0
familial	2.043	32.73	0.47	54	165	Inf	0
heureux	6.735	29.72	1.70	178	599	Inf	0
insouci	1.476	36.11	0.31	39	108	Inf	0
problem	2.573	30.77	0.63	68	221	Inf	0
san	3.973	39.18	0.76	105	268	Inf	0
scolair	0.795	41.18	0.14	21	51	6.6	0
souc	1.059	30.11	0.26	28	93	6.4	0
période	3.859	14.45	2.01	102	706	6.3	0
et	3.784	14.01	2.03	100	714	6.0	0
de	2.270	2.75	6.19	60	2178	-9.8	0
pas	0.303	0.96	2.36	8	831	-8.9	0
je	0.038	0.24	1.17	1	412	-7.2	0
la	0.832	2.30	2.72	22	958	-7.1	0
on	0.076	0.51	1.12	2	395	-6.6	0
était	0.000	0.00	0.87	0	307	-6.5	0
per	0.000	0.00	0.77	0	272	-6.1	0
difficil	0.000	0.00	0.75	0	265	-6.0	0

Biographies et entourage, 2001

Réponses spécifiques de la classe 1 corpus *Synthèse de la trajectoire de vie pendant l'enfance*

206

période d'enfance sans problème et très heureuse

769

enfance et adolescence très heureuse

1018

enfance et adolescence très heureuse

1281

enfance et adolescence très heureuse

1394

enfance et adolescence très heureuse

Documents spécifiques de la classe 1 :

	Dist. du Khi2 au centroïde
206	7.8
769	8.7
1018	8.7
1281	8.7
1394	8.7

Biographies et entourage, 2001

Conclusion sur la mise en œuvre de la statistique textuelle sous R (R.TeMiS)

1. Description du corpus (bilan lexical)
2. Mots du lexique les plus fréquents,
3. Croiser les mots et des caractéristiques sur ces textes (variables qualitatives) -> Tableau lexical agrégé
4. Mots spécifiques par sous catégories, par classe etc.
5. Classification pour rechercher des « mondes lexicaux » ->Tableau lexical entier

Mise en œuvre facile des méthodes de la statistique textuelle

Permet de s'approprier peu à peu la syntaxe du package tm pour aller ensuite plus loin

R. TeMis

- Bon outil pédagogique mais outil statistique
- Le lexique des mots et les vocabulaire spécifique sont simples à analyser, interpréter et présenter
- Les analyses factorielles/classification requièrent des connaissances de base en analyse des données (choix des variables, rôle actif ou illustratif, interprétation et présentation des résultats)

Quelques éléments de bibliographie (articles)

- Brennetot A., Emsellem K., Guérin-Pace F., Garnier B. 2013. *Dire l'Europe à travers le monde.Les mots des étudiants dans l'enquête EuroBroadMap*, Cybergéo
- Collomb Ph., Guérin-Pace F. 1998. Les contours du mot « environnement » : enseignements de la statistique textuelle Espace Géographique, L'espace géographique, 41 (1), p. 41-52 (1)
- Guérin-Pace F., Saint-Julien T., 2012 - *Les mots de L'Espace Géographique. Une analyse lexicale des titres et mots-clés de 1972 à 2010.* L'espace géographique*, 41 (1), p. 4-30*
- Golaz V., Lelièvre E., 2012 - *L'entourage familial pendant l'enfance et l'adolescence, entre faits et perceptions. Une analyse rétrospective des parcours de vie des Franciliens des générations 1930-1950* – In Population, (3), Ined. Paris.
- Sites
  - → www.cavi.univ-paris3.fr/lexicometrica (actes des JADT)
  - → www.eurobroadmap.eu (projet)

Quelques éléments de bibliographie (méthodes)

- Lebart L., Salem A. 1994. *Statistique textuelle*. Paris, Dunod, 342 p. (épuisé mais sur le site de Ludovic Lebart http://www.dtmvic.com/

Voir *Publications* et choisir *Télécharger* (format pdf) ou Lecture sur écran

- Reinert M. 1983, Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. Cahiers de l'Analyse des Données, 3, pp. 187-198
- Benzecri J.-P., 1973 – *L'analyse des Données* (tome 1 et 2). Dunod, Paris
- Guérin-Pace F. 1997. La statistique textuelle : un outil exploratoire en sciences sociales. In: Population, (4), Ined. Paris. pp. 865-887
- Garnier B., Guérin-Pace F., 2010 - Appliquer les méthodes de la statistique textuelle, Ceped, les clefs pour, Paris

Approches complémentaires

Term frequency (TF) : nombre de citations du terme ti dans le document dj (wij) ramené au nombre de terme dans le document dj

$$TF_{ij} = \frac{w_{ij}}{|d_j|}$$

Inverse document frequency (IDF) : N nombre de documents du corpus, dfi nombre de documents contenant le terme t_i

Terme dans peu de documents plus discriminant qu'un terme présent dans tous les documents

$$IDF_i = \log \frac{N}{df_i}$$