Textométrie

Analyse statistique du lexique politique

Jean-Claude Deroubaix

GRAID-ULB
SYLED Paris 3

Méthode quantitative et méthode qualitative

Faut-il vraiment distinguer ?

Les méthodes « quantitatives » d'analyse lexicale impliquent comme pour toutes méthode statistique un indispensable « retour aux données ».

Cela signifie :

1 que la « production » des données initiales doit être de qualité, contrôlée et critiquée ;

2 qu'on doit retrouver dans les données d'origine les conclusions de l'analyse, d'une manière ou d'une autre.

S'il est impossible ou difficile d'interpréter un résultat quantifié, il faut s'interroger et sur les données et sur la méthode.

En réalité, les méthodes quantitatives dans la recherche en sciences humaines doivent avoir pour but de permettre une « autre lecture » des données que la lecture naïve. Des méthodes qualitatives ont d'ailleurs le même objectif.

Méthode quantitative et méthode qualitative dans l'analyse de discours
Histoire

Dans l'analyse lexicométriqueil y a une grande intégration des approches statistiques et de remaniement lexical. Ce qui permet de vérifier pas à pas la qualité du travail statistique et de mettre l'accent sur son aspect heuristique.

La méthode est celle qui a été développé au Laboratoire de lexicologie de l'école normale supérieure de Saint-Cloud à partir de 1970 sous la direction de Maurice Tournier. De nombreux travaux ont été réalisés dans ce laboratoire sur les discours politiques et syndicaux dont une œuvre inaugurale, encore tâtonnante sur le choix des méthodes, « Des tracts en Mai 68 ».
Les méthodes (statistiques et informatiques) ont été mises au point essentiellement par Pierre Lafon et André Salem. Une grande partie des travaux ont nourris la revue MOTS, créée par le laboratoire de Saint-Cloud et actuellement dirigée par le politologue Paul Bacot (Lyon2)aux éditions des ENS (: http://mots.revues.org/)

Important : il s'agit de méthode dans lesquelles le **corpus** est la **référence**. Il ne s'agit donc pas d'une lexicologie qui supposerait la possibilité d'avoir des usages de mots fixés quantitativement dans la langue, mais seulement des usages dans un corpus particulier

D'où l'attention portée à la nécessité de constituer des corpus significatifs, c'est-à-dire des corpus qui représentent un univers cohérent et raisonnablement justifiable.

Le corpus

Le corpus qui servira d'exemple est celui que j'avais réuni pour ma thèse de doctorat à la Sorbonne : les déclarations gouvernementales faites au Parlement belge par les 37 gouvernements qui se sont succédé de 1944 à 1992.

Tout commence par la constitution d'un corpus, donc par un problème de recherche.

La Belgique se caractérise par un système politique qui favorise les **gouvernements de coalition.**

La **césure gauche/droite** très visible dans les pays comme la France de la Ve république est également présente dans la vie politique belge. Des partis se réclament de la gauche, d'autre de la droite, certains même du centre.

Mais se marquent ces clivages partisans dans le programme du gouvernement de coalition ?

A priori, l'hypothèse était que cette césure se marquerait encore mais tempérée par la présence presque systématique du parti (puis des partis) chrétien au gouvernement, fonctionnant comme pivot de la coalition.

Le choix du corpus était donc simple : **l'ensemble des discours** prononcé au parlement par le **gouvernement** en vue de demander **un vote de confiance** inaugural.

La constitution d'un corpus ce sont aussi des choix

Certains gouvernements rendaient publics parallèlement à leur déclaration, un programme écrit plus détaillé, fallait-il en tenir compte ?

Certaines déclarations sont extrêmement courtes car le gouvernement en question est explicitement un gouvernement de transition, etc.

L'utilisation des techniques lexicométriques implique des contraintes sur les données : les différentes **parties du corpus** doivent être **pertinentes** pour la question de recherche.
Mais elles doivent aussi être de poids pas trop dissemblables, et donc **de longueurs comparables**.

Elles doivent aussi représenter correctement l'univers exploré par la recherche :

soit le **corpus est exhaustif**,

soit il est issu d'un **échantillon** dont on peut penser qu'il est **représentatif**
statistiquement (échantillon aléatoire)
ou
raisonnablement mais avec le degré d'incertitude que cela peut entraîner sur la validité de la généralisation des résultats.

Notre corpus est

exhaustif

pertinent avec la question de recherche

les textes sont de **tailles comparable**s

(trois déclarations de transition dont le seul programme est la transition et une déclaration dont l'unique objet est la réforme de l'Etat seront toutefois soumis à un traitement particulier pour certaines procédures à cause de ces particularités).

N°	*Premier ministre*	*Période d'exercice*	*Composition de la coalition*
dg00	Hubert Pierlot	26.09.44 07.02.45	Sociaux-chrétiens, Libéraux, Socialistes, Communistes
dg01	Achille Van Acker I	12.02.45 15.06.45	Sociaux-chrétiens, Libéraux, Socialistes, Communistes
dg02	Achille Van Acker II	02.08.45 12.02.46	Socialistes, Libéraux, Communistes, Union démocratique belg
dg03	Paul-Henri Spaak I	13.03.46 20.03.46	Socialistes
dg04	Achille Van Acker III	31.03.46 09.07.46	Socialistes, Libéraux, Communistes
dg05	Camille Huysmans	03.08.46 13.03.47	Socialistes, Libéraux, Communistes
dg06	Paul-Henri Spaak II	20.03.47 19.11.48	Socialistes, Sociaux-chrétiens
dg07	Paul-Henri Spaak III	27.11.48 27.06.49	Socialistes, Sociaux-chrétiens
dg08	Gaston Eyskens I	11.08.49 06.06.50	Sociaux-chrétiens, Libéraux
dg09	Jean Duvieusart	08.06.50 11.08.50	Sociaux-chrétiens
dg10	Joseph Pholien	16.08.50 09.01.52	Sociaux-chrétiens
dg11	Jean Van Houtte	15.01.52 12.04.54	Sociaux-chrétiens
dg12	Achille Van Acker IV	23.04.54 02.06.58	Socialistes, Libéraux
dg13	Gaston Eyskens II	26.06.58 04.11.58	Sociaux-chrétiens
dg14	Gaston Eyskens III	06.11.58 03.09.60	Sociaux-chrétiens, Libéraux
dg15	Gaston Eyskens III bis	03.09.60 27.03.61	Sociaux-chrétiens, Libéraux
dg16	Theo Lefèvre	25.04.61 24.05.65	Socialistes, Sociaux-chrétiens
dg17	Pierre Harmel	28.07.65 11.02.66	Sociaux-chrétiens, Socialistes
dg18	Paul Vanden Boeynants I	19.03.66 01.04.68	Sociaux-chrétiens, Libéraux
dg19	Gaston Eyskens IV	17.06.68 08.11.71	Sociaux-chrétiens (PSC-CVP), Socialistes
dg20	Gaston Eyskens V	21.01.72 23.11.72	Sociaux-chrétiens (PSC-CVP), Socialistes
dg21	Edmond Leburton I	26.01.73 23.10.73	Socialistes, Sociaux-chrétiens (PSC-CVP), Libéraux (PLP-PV
dg22	Edmond Leburton II	23.10.73 19.01.74	Socialistes, Sociaux-chrétiens (PSC-CVP), Libéraux (PLP-PV
dg23	Léo Tindemans I	25.04.74 11.06.74	Sociaux-chrétiens, Libéraux
dg24	Léo Tindemans II	11.06.74 04.03.77	Sociaux-chrétiens, Libéraux, Rassemblement wallon

N°	Premier ministre	Période d'exercice	Composition de la coalition
dg21	Edmond Leburton I	26.01.73 23.10.73	Socialistes, Sociaux-chrétiens (PSC-CVP), Libéraux (PLP-PVV)
dg22	Edmond Leburton II	23.10.73 19.01.74	Socialistes, Sociaux-chrétiens (PSC-CVP), Libéraux (PLP-PVV)
dg23	Léo Tindemans I	25.04.74 11.06.74	Sociaux-chrétiens, Libéraux
dg24	Léo Tindemans II	11.06.74 04.03.77	Sociaux-chrétiens, Libéraux, Rassemblement wallon
dg25	Léo Tindemans IV	03.06.77 11.10.78	Sociaux-chrétiens, Socialistes, FDF, Volksunie
dg26	Paul Vanden Boeynants II	20.10.78 18.12.78	Sociaux-chrétiens, Socialistes (PS-BSP), FDF, Volksunie
dg27	Wilfried Martens I	03.04.79 16.01.80	Sociaux-chrétiens (PSC-CVP), Socialistes (PS-BSP), FDF,
dg28	Wilfried Martens II	23.01.80 09.04.80	Sociaux-chrétiens, Socialistes
dg29	Wilfried Martens III	18.05.80 07.10.80	Sociaux-chrétiens, Socialistes, Libéraux
dg30	Wilfried Martens IV	22.10.80 02.04.81	Sociaux-chrétiens, Socialistes
dg31	Marc Eyskens	06.04.81 21.09.81	Sociaux-chrétiens, Socialistes
dg32	Wilfried Martens V	17.12.81 14.10.85	Sociaux-chrétiens, Libéraux
dg33	Wilfried Martens VI	28.11.85 19.10.87	Sociaux-chrétiens, Libéraux
dg34	Wilfried Martens VII	21.10.87 13.12.87	Sociaux-chrétiens, Libéraux
dg35	Wilfried Martens VIII	09.05.88 29.09.91	Sociaux-chrétiens, Socialistes Volksunie
dg36	Wilfried Martens IX	29.09.91 07.03.92	Sociaux-chrétiens, Socialistes
dg37	Jean-Luc Dehaene	07.03.92 27.06.95	Sociaux-chrétiens, Socialistes

Quelques définitions

L'unité de comptage et l'unité d'observation doivent être définies.

A priori, le **mot** est un bon candidat.
Encore faut-il s'entendre sur ce qu'est un mot : était et être sont-ils le même mot ? travailleur et travailleuse ? etc.

Il y a un choix à faire.

Soit répondre oui et opérer ce qu'on appelle une **lemmatisation** : réduire chaque forme à une forme canonique : le masculin singulier pour les noms et adjectifs, l'infinitif pour les verbes (comme dans les entrées de dictionnaire).

Soit considérer que la marque du pluriel, du temps, de la personne, etc. sont des éléments signifiants et donc conserver à chaque **forme** distincte sa qualité d'unité observable.

La stratégie de la lexicométrie adoptée par le logiciel LEXICO 3 est
la **forme lexicale**.

Définition de la forme lexicale

On sépare l'ensemble des signes typographiques en **deux sous-ensembles** : les **signes séparateurs** et les **signes non séparateurs**.

Les premiers sont par exemple : l'espace, les signes de ponctuation, le retour à la ligne,etc.

Les seconds sont tous les autres : lettres et chiffres, etc.

Une forme lexicale est toute suite de signes non séparateurs comprise entre deux signes séparateurs.

Nous observerons et compterons donc **les occurrences de formes lexicales**.

La première opération statistique va être la **segmentation** du corpus en formes lexicales et le comptage de ces formes.

Notre corpus des déclarations peut être décrit ainsi :

Nombre d’occurrences : 115956 (longueur en mots du corpus)

Nombre de formes distinctes : 8496

Nombre de hapax : 3537 (formes qui n’apparaissent qu’une fois)

Fréquence maximale : 6645 (de)

Examinons d’abord les mots les plus fréquents. Voici la liste des plus fréquents.

N°	Forme attestée	Fréquence	Fréquence relative	N°	Forme attestée	Fréquence	Fréquence relative
1	de	6645	5,731%	26	ce	499	0,430%
2	la	4515	3,894%	27	qu	480	0,414%
3	et	3265	2,816%	28	aux	477	0,411%
4	l’	3092	2,667%	29	nous	450	0,388%
5	le	3082	2,658%	30	s’	448	0,386%
6	des	2891	2,493%	31	a	439	0,379%
7	les	2693	2,322%	32	être	434	0,374%
8	à	2566	2,213%	33	sera	433	0,373%
9	d	1948	1,680%	34	pays	402	0,347%
10	en	1530	1,319%	35	notre	382	0,329%
11	une	1394	1,202%	36	ne	381	0,329%
12	du	1347	1,162%	37	se	375	0,323%
13	dans	1264	1,090%	38	cette	373	0,322%
14	Gouvernement	1252	1,080%	39	économique	337	0,291%
15	un	1110	0,957%	40	pas	333	0,287%
16	il	1088	0,938%	41	avec	326	0,281%
17	qui	1008	0,869%	42	doit	284	0,245%
18	que	999	0,862%	43	n’	264	0,228%
19	est	855	0,737%	44	seront	263	0,227%

N°	Forme attestée	Fréquence	Fréquence relative	N°	Forme attestée	Fréquence	Fréquence relative
51	nos	230	0,198%	76	peut	140	0,121%
52	leur	225	0,194%	77	réforme	137	0,118%
53	ses	220	0,190%	78	dont	136	0,117%
54	mais	218	0,188%	79	faire	136	0,117%
55	sa	214	0,185%	80	vie	134	0,116%
56	entre	208	0,179%	81	programme	133	0,115%
57	État	205	0,177%	82	si	133	0,115%
58	elle	204	0,176%	83	y	131	0,113%
59	mesures	202	0,174%	84	entend	130	0,112%
60	Belgique	200	0,172%	85	notamment	129	0,111%
61	sans	200	0,172%	86	œuvre	128	0,110%
62	problèmes	191	0,165%	87	assurer	127	0,110%
63	tout	191	0,165%	88	enseignement	127	0,110%
64	Parlement	184	0,159%	89	projet	126	0,109%
65	aussi	184	0,159%	90	développement	124	0,107%
66	c’	181	0,156%	91	social	124	0,107%
67	été	177	0,153%	92	cadre	123	0,106%
68	même	174	0,150%	93	cet	123	0,106%
69	sociale	171	0,147%	94	travail	123	0,106%
70	matière	156	0,135%	95	économie	122	0,105%

Deux remarques d'emblée : la fréquence importante de mots outils comme *de le la les et,...* *De* à lui seul occupe près de 6% de la surface du texte mesurée en formes lexicales. C'est un marqueur du français. Un texte en français suffisamment long a une fréquence de *de* proche de six pour cent.

Mais outre cette importance des mots outils nous voyons se dessiner immédiatement l'univers socio-politique des textes : *Gouvernement, politique, pays, économique, loi, État, mesures Belgique, problèmes*, *Parlement*, etc.

Et même les rôles sont rapidement désignés : Gouvernement, Parlement, Belgique mais aussi *il, nous, s' , notre, se, tous, son, nos, leur, ses, sa, elle*. Le *il* comprend soit des formes dites impersonnelles soit renvoie au mot « gouvernement » mais le *nous* est particulièrement significatif de ce type de discours puisqu'il s'agit d'une auto-présentation d'un collectif. Une étude plus attentive du *nous* dans les concordances peut cependant amener à voir que se confond sous ce mot toutes sortes de nous : le gouvernement, les hommes politiques réunis en séances, nous les Belges, etc…

L'intensité mise sur *tous*, *notre*, *nos* et *pays* montre aussi qu'il ne s'agit pas d'une note confidentielle mais d'une adresse au pays par-delà l'assemblée parlementaire.

Le tableau de fréquence le premier outil que fournit le logiciel LEXICO 3

Deuxième outil : la recherche des segments répétés

Définition du **segment répété** : Les segments répétés sont des suites de formes comprises entre deux séparateurs de séquence et dont la présence est attestée dans le corpus avec une fréquence égale ou supérieure à deux.

L'intérêt principal de l'exploration des segments répétés est évidemment de rechercher des associations de formes qui complètent les analyses basées sur des formes simples.

L	Fréq.	Segment répété	L	Fréq.	Segment répété
2	1480	de la	2	118	sur le
2	1017	de l’	2	118	sur les
2	992	le Gouvernement	2	117	et d’
2	505	à la	2	114	de loi
2	420	à l’	2	113	et le
2	377	et de	2	111	politique de
2	343	dans le	2	109	dans l’
2	306	d’une	2	109	que les
2	227	la politique	2	106	toutes les
2	207	d’un	2	103	du pays
2	194	et la	2	102	il est
2	183	une politique	2	102	l’enseignement
2	180	et des	2	102	pour les
2	177	tous les	2	100	économique et
2	176	c’est	2	99	doit être
2	174	qu’il	2	98	en ce
2	169	de notre	2	97	que la
2	168	la Belgique	3	94	et de la
2	166	l’État	2	94	la loi
2	161	et les	2	94	sur la
2	159	dans la	2	92	le cadre
2	153	que le	2	92	projet de
2	151	dans les	2	91	la réforme
2	151	du Gouvernement	2	87	la vie
2	146	et à	2	86	entre les
2	134	et l’	2	86	il faut
2	132	de nos	2	86	notre pays
2	128	par la	3	85	dans le cadre
2	127	en matière	2	84	a été
3	124	de l’État	2	84	qui concerne
2	123	par le	2	83	à une
2	121	en vue	2	83	dans un
2	118	ce qui	3	82	ce qui concerne

Exemple : « en ce qui concerne » apparaît donc 82 fois dans le corpus ; cela dénote un discours assez « administratif »

Un autre exemple tiré de cette liste : « *le Gouvernement qui se présente devant vous* » apparaît 13 fois.

On peut cependant y ajouter quelques variantes comme :

le Gouvernement qui se présente aujourd'hui devant vous (5 fois)
le Gouvernement qui se présente à vous (1 fois)
le Gouvernement qui se présente devant les Chambres (1 fois)
le Gouvernement se présente donc aujourd'hui devant vous (1 fois)
le Gouvernement se présente à vous (1 fois)
le Gouvernement se présente avec un programme (1 fois)
le Gouvernement se présente devant les Chambres (2 fois)
le Gouvernement se présente devant vous (2 fois)
le Gouvernement qui se présente à vos suffrages (1 fois)

Cette liste a été trouvée en utilisant le concordancier de Lexico3

Concordancier

Pour les découvrir, nous utilisons le plus ancien des outils d'analyse lexicale : la concordance. Cette technique de réorganisation du texte est utilisée depuis l'antiquité. Mais ce qui prenait des années à la main pour les moines se fait en un clic de souris au temps de l'informatique.

Concordances : recherche des contextes d'occurrence d'une forme lexicale présenté de manière à aligner les contextes sur la forme en question.

ȝ08	tâche. mesdames, messieurs, le Gouvernement qui a l' honneur de se pré
ȝ11	leur gravité. au contraire. le Gouvernement qui a l' honneur de se pré
ȝ20	es et nationales nouvelles. le Gouvernement qui se présente à vos suff
ȝ00	mesdames, messieurs, le Gouvernement qui se présente devant vou
ȝ01	de fidélité et d' honneur. le Gouvernement qui se présente devant vou
ȝ03	sible aux intérêts du pays, le Gouvernement qui se présente devant vou
ȝ04	gagne. mesdames, messieurs, le Gouvernement qui se présente devant vou
ȝ06	posent à l' heure actuelle. le Gouvernement qui se présente devant vou
ȝ07	ligée. mesdames, messieurs, le Gouvernement qui se présente devant vou
ȝ09	tablie par la Constitution. le Gouvernement qui se présente aujourd_hu
ȝ11	és par l' intérêt national. le Gouvernement qui se présente à vous veu
ȝ12	avec des idées de revanche. le Gouvernement qui se présente devant vou
ȝ13	avail! mesdames, messieurs, le Gouvernement qui se présente devant vou
ȝ16	sera réalisée en son sein. le Gouvernement qui se présente devant vou
ȝ17	tard. mesdames, messieurs, le Gouvernement qui se présente aujourd_hu
ȝ18	ire aux situations claires. le Gouvernement qui se présente aujourd_hu
ȝ18	i de langue, ni de fortune. le Gouvernement qui se présente devant vou
ȝ19	venir. mesdames, messieurs, le Gouvernement qui se présente aujourd_hu
ȝ21	quart du vingtième siècle. le Gouvernement qui se présente devant vou
ȝ23	ue qui sait ce qu' il veut. le Gouvernement qui se présente devant les
ȝ23	intentions et le programme du Gouvernement qui se présente aujourd_hu
ȝ25	ont permis la constitution du Gouvernement qui se présente devant vou
ȝ26	es, messieurs, le programme du Gouvernement qui se présente devant vou
ȝ27	crise que traverse le pays. le Gouvernement qui se présente devant vou
ȝ33	ue dans sa vie quotidienne. le Gouvernement qui se présente devant vou
ȝ06	r la première fois, un nouveau Gouvernement se présente devant les Cha
ȝ15	me au mois d' août dernier, le Gouvernement se présente devant vous, r
ȝ21	ctive. mesdames, messieurs, ce Gouvernement se présente devant les Cha
ȝ29	histoire de notre pays que le Gouvernement se présente devant vous. n
ȝ30	sera dangereusement menacé. le Gouvernement se présente donc aujourd_h
ȝ31	ièCle. mesdames, messieurs, le Gouvernement se présente à vous avec un

Les tableaux lexicaux entier et tronqué

La ventilation de toutes les occurrences entre les parties du corpus (les 37 déclarations gouvernementales) aboutit à constituer un tableau croisé où les parties sont reprises en colonnes, les formes lexicales en ligne. Chaque case correspond à la fréquence de la forme dans la partie.

Exemple du tableau lexical entier du corpus des déclarations gouvernementales

	Formes	dg00	dg01	…	dg36	dg37	∑
0	de	203	87	…	48	153	6 645
1	la	127	71	…	23	87	4 515
…	…	…	…	…	…	…	…
1354	voisins	1	1	…	0	0	10
…	…	…	…	…	…	…	…
	∑	439	1 772	…	715	2 616	115 596

Généralement on tronque ce tableau en supprimant les formes qui apparaissent au total moins de 10 fois.

Il reste dans notre corpus 1355 lignes correspondant aux 1355 formes apparaissant au moins 10 fois.

Comment utiliser ce tableau ?

Deux stratégies sont possibles selon la recherche que l'on veut mener :

1. Quelles sont les formes caractéristiques de l'une ou l'autre déclaration ? Quelles sont les formes caractéristiques de plusieurs déclarations ?
2. Comment globalement se disposent les déclarations en fonction de l'utilisation plus ou moins fréquentes des mêmes formes lexicales ? Peut-on faire des regroupements de formes ou de déclarations en tenant compte de la répartition des fréquences ?

1. La première option relève de l'analyse des spécificités.
2. La seconde relève de l'analyse statistique multidimensionnelle et particulièrement de l'analyse des correspondances binaires et de la classification automatique.

Les spécificités

Étant donné une forme lexicale et sa fréquence dans une partie du corpus, la question de savoir si cette fréquence est significativement élevée ou basse par rapport à la fréquence de cette même forme dans le corpus est une question importante. On dira si la forme est particulièrement abondante qu'elle est « spécifique positive » dans cette partie et symétriquement qu'elle est « spécifique négative » dans l'hypothèse où elle serait sous-utilisée.

Pour calculer les spécificités d'une forme, LEXICO 3 tient compte de la fréquence de la forme dans la partie (f), de la taille de la partie (t), de la taille du corpus (T) et de la fréquence de la forme dans le corpus (F).

La loi statistique utilisée est la loi hypergéométrique.
Si dans une urne contenant T boules dont F sont marquées l'on tire t boules dont au moins f boules marquées, à condition d'effectuer le tirage sans remplacement, la probabilité de tirer un échantillon comprenant exactement f boules marquées est :

$$P\left(X = k\right) = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

On peut donc comparer la valeur observée avec la probabilité [qu'elle soit dépassée pour un certain seuil fixé.]

La spécificité positive est d'interprétation aisée : elle nous indique si la forme est plus utilisée dans une partie que dans le corpus.

La spécificité négative est plus difficile à interpréter puisqu'il nous dit que cette forme est sous-utilisée dans une partie par rapport au corpus.

Autant la spécificité positive est le résultat d'une action du locuteur, autant la spécificité négative est le résultat de l'examen des fréquences par le prisme de ce qu'ont dit les autres locuteurs.

Une forme qui n'est spécifique d'aucune partie est une **forme banale**. Les formes banales ne manquent pas d'intérêt puisque si on enlève les formes outils les autres formes banales renseignent sur le champ commun à tous les locuteurs.

Corpus : DG0037	Date : jeudi 21 février 2008 - 00:08	Partition = S

Spécifs - Part : S
Parties sélectionnées : 37,

Forme	Frq. Tot.	Fréquence	Coeff.
citoyen	17	10	13
contrat	8	6	9
autorités	18	6	7
équipe	9	5	7
Union	8	4	6
collègues	6	3	5
comportement	6	3	5
assainissement	54	7	5
Europe	71	8	5
gouvernement	6	3	5
plus	568	28	5
impact	6	3	5
réponse	7	3	4
monsieur	8	3	4
citoyens	63	7	4
achèvement	7	3	4
finances	71	7	4
président	10	3	4
société	81	8	4
1992	8	3	4
domaines	49	6	4
européens	13	3	4
demande	41	5	4
critères	9	3	4
publiques	83	8	4
pourquoi	36	5	4
environnement	23	4	4
contribution	9	2	3
intégrer	12	2	3
prévues	12	2	3
disposer	12	2	3
adoption	12	2	3
appartient	12	2	3
maîtrise	10	2	3
novembre	10	2	3
suffisamment	11	2	3
retard	12	2	3
fixer	11	2	3
biais	6	2	3
collectif	6	2	3
uniquement	7	2	3
24	6	2	3
satisfaire	5	2	3
souhaitent	5	2	3
brefs	5	2	3
liste	5	2	3
solidaire	7	2	3

Corpus : DG0037	Date : jeudi 21 février 2008 - 00:08	Partition = S

Spécifs - Part : S

Parties sélectionnées : 01,

Forme	Frq. Tot.	Fréquence	Coeff.
ceux	102	10	7
guerre	80	9	7
alliés	12	4	6
elle	204	13	6
titre	12	4	6
contre	84	8	5
continue	7	3	5
auront	20	4	5
écarter	5	2	4
ennemi	5	2	4
ma	5	2	4
rendus	6	2	4
sinistrés	6	2	4
veut	57	5	4
reprise	13	3	4
pain	5	2	4
tête	5	2	4
longtemps	18	2	3
tient	17	2	3
amitié	7	2	3
concorde	10	2	3
gravité	10	2	3
immense	9	2	3
Nation	8	2	3
devoirs	18	2	3
actes	16	2	3
il	1088	28	3
garantie	14	2	3
restera	12	2	3
oeuvre	12	2	3
répression	16	2	3
charbon	16	2	3
rendement	15	2	3
patrie	15	2	3
qu	480	13	3
leurs	109	5	3
sein	62	4	3
entreprises	76	5	3
dont	136	7	3
nos	230	8	3
n	264	9	3
pas	333	12	3
ses	220	8	3
mon	7	2	3
activité	35	3	3
est	855	23	3
besoin	21	3	3

La statistique multidimensionnelle. L'AFC

On peut se donner comme objectif d'avoir une vue synthétique des rapports qu'entretiennent les différentes parties dans leur utilisation du vocabulaire du corpus.

L'Analyse factorielle des correspondances peut être vue comme une méthode permettant de « cartographier » dans un espace à faible dimension (un plan, un espace à trois dimension, etc.) les distances entre les lignes ou entre les colonnes d'un tableau croisé.

Le tableau lexical tronqué est un tableau croisé. On peut donc le soumettre à l'analyse factorielle.

L'AFC va permettre de visualiser sur un plan les **distances**

entre profils des colonnes, c'est-à-dire des parties du corpus,

entre les profils des lignes, c'est-à-dire des formes lexicales attestées au moins dix fois,

et même de mettre en relation ces deux ensembles de points en les représentants simultanément.

La table des valeurs propres

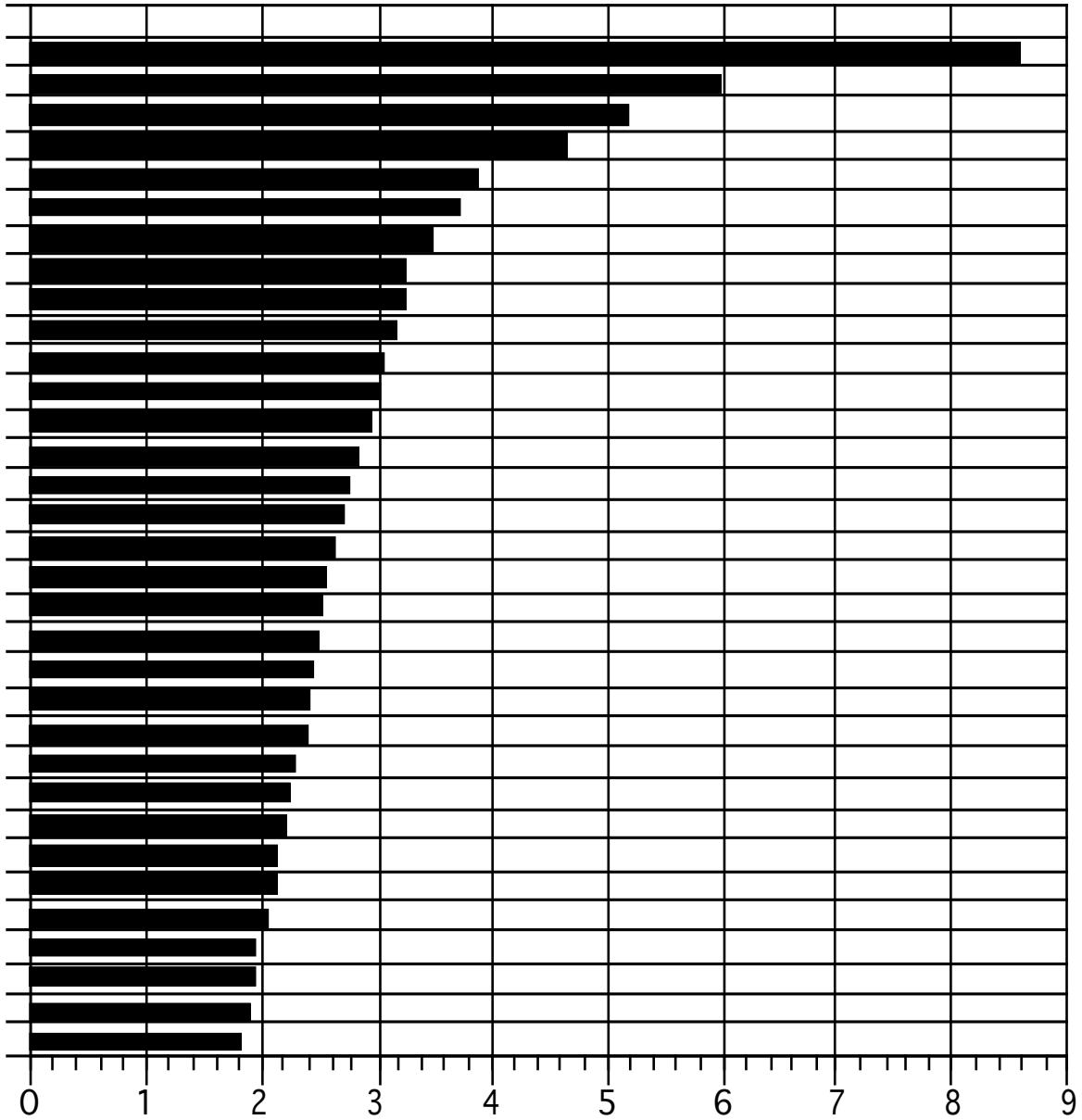
donne une mesure de la fidélité des la représentation du nuage des points sur les axes.

Le premier plan factoriel formé par les axes 1 et 2 donne une représentation de 14,5% de la dispersion des points autour du centre.

Cette valeur serait très faible s'il s'agissait d'un tableau croisant les réponses à deux questions fermées.

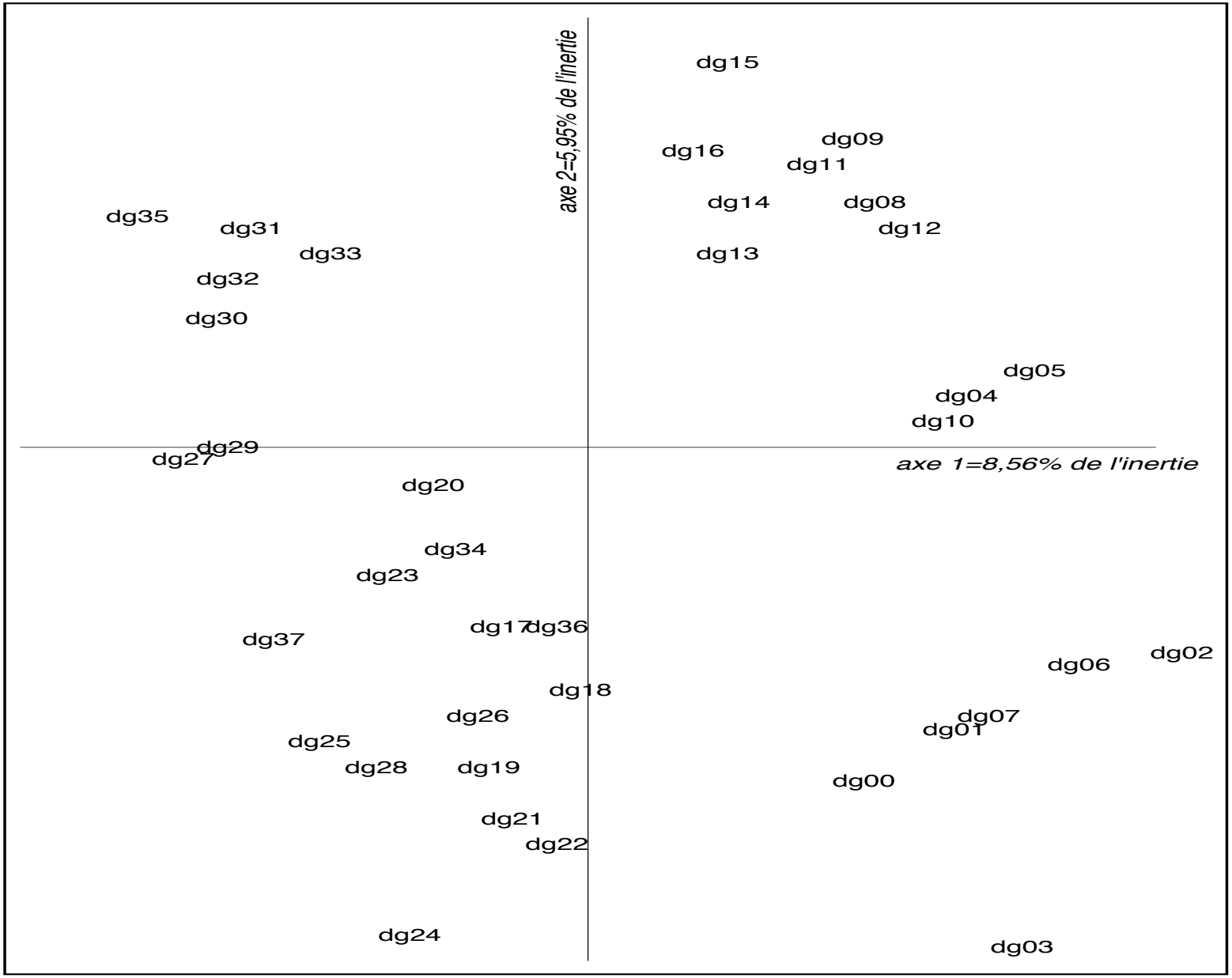
Mais dans le cas des tableaux lexicaux ces valeurs sont de bonne qualité car ces tableaux comportent des formes communes peu discriminantes mais très nombreuses (les formes outils entre autres qui accroissent la dispersion totale mais se répartissent de manière relativement uniforme sur tous les axes.

N°	Valeur	%	%Cum
1	.0615	8.56	8.56
2	.0428	5.96	14.52
3	.0369	5.14	19.66
4	.0331	4.61	24.26
5	.0278	3.87	28.13
6	.0266	3.70	31.83
7	.0251	3.49	35.32
8	.0232	3.24	38.56
9	.0230	3.21	41.77
10	.0228	3.17	44.94
11	.0219	3.05	47.99
12	.0216	3.00	50.99
13	.0211	2.94	53.93
14	.0202	2.82	56.75
15	.0198	2.75	59.50
16	.0194	2.70	62.20
17	.0187	2.60	64.80
18	.0184	2.56	67.36
19	.0181	2.52	69.88
20	.0177	2.47	72.35
21	.0176	2.45	74.80
22	.0174	2.42	77.23
23	.0169	2.36	79.58
24	.0163	2.27	81.85
25	.0158	2.21	84.06
26	.0156	2.17	86.22
27	.0153	2.12	88.35
28	.0151	2.10	90.45
29	.0148	2.05	92.50
30	.0138	1.92	94.42
31	.0138	1.92	96.34
32	.0134	1.86	98.20
33	.0129	1.80	100.00



premier plan factoriel de l'AFC du tableau lexical tronqué

représentation des parties.

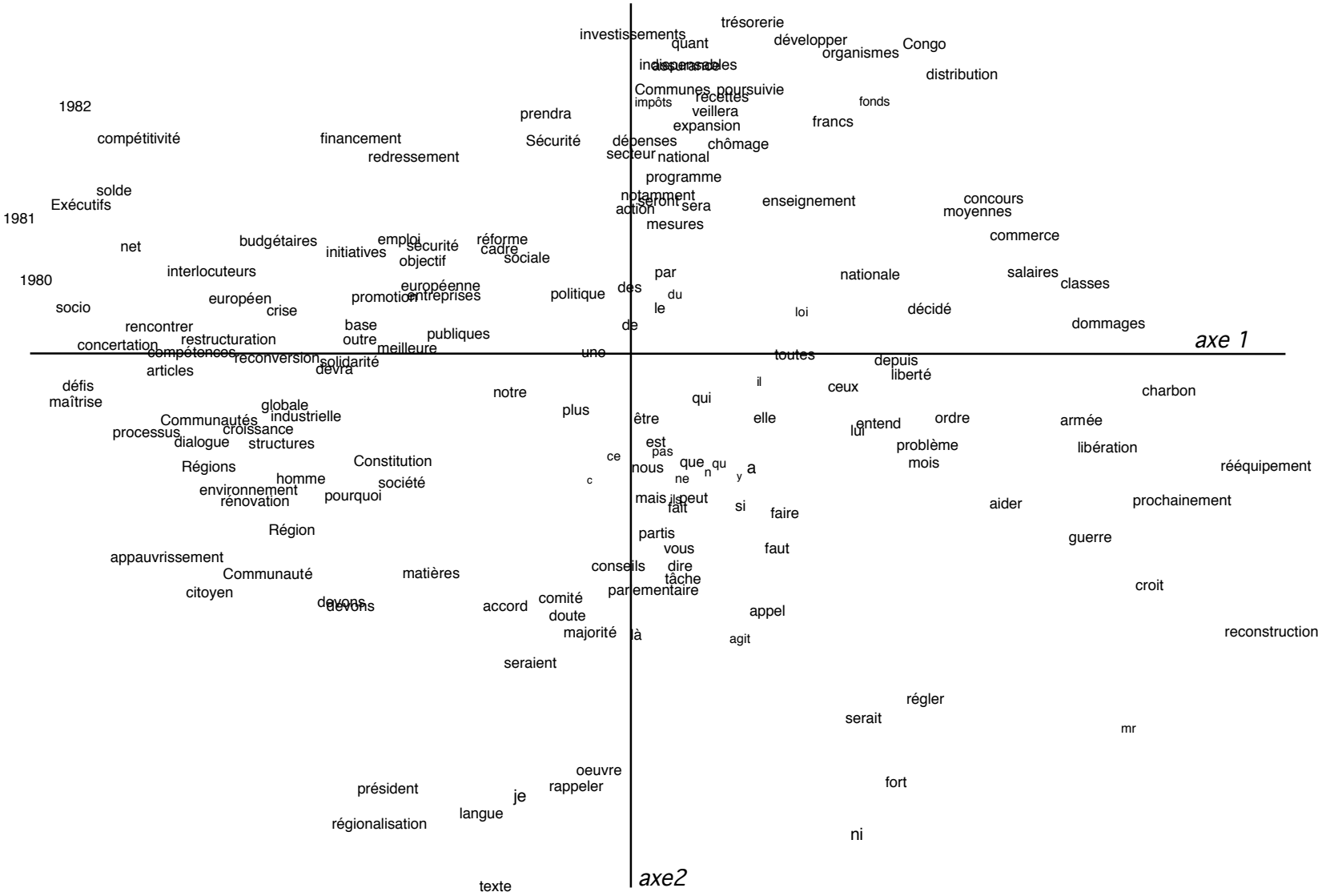


Ce plan appelle de nombreuses remarques

En première approche : on voit que les différentes parties du corpus s'organise plus ou moins de manière chronologique de manière horizontale (le long de l'axe 1). C'est la conséquence du fait que le vocabulaire se renouvelle tout au long de la période et qu'un gouvernement a tendance à puiser dans un vocabulaire proche de celui de son prédecesseur.

Ensuite il y a une rupture lorsqu'on passe de la partie gauche du graphique à sa partie droite, elle s'opère entre 1961 et 1965, alors que pour tous les politologues et historien contemporains de la Belgique, la rupture politique se trouve dans la crise suscitée par la grève de l'hiver 1960-1961. Mais le discours du gouvernement (DG17) qui se forme au début de 1961 continue à puiser son vocabulaire dans la période passée : le discours de de formation gouvernement est un texte censément programmatique, or il se réfère plus au passé qu'à l'avenir.

Pour le reste de l'analyse : article de J.C Deroubaix dans Mots, Histoire et Mesure, dans les actes des JADT.

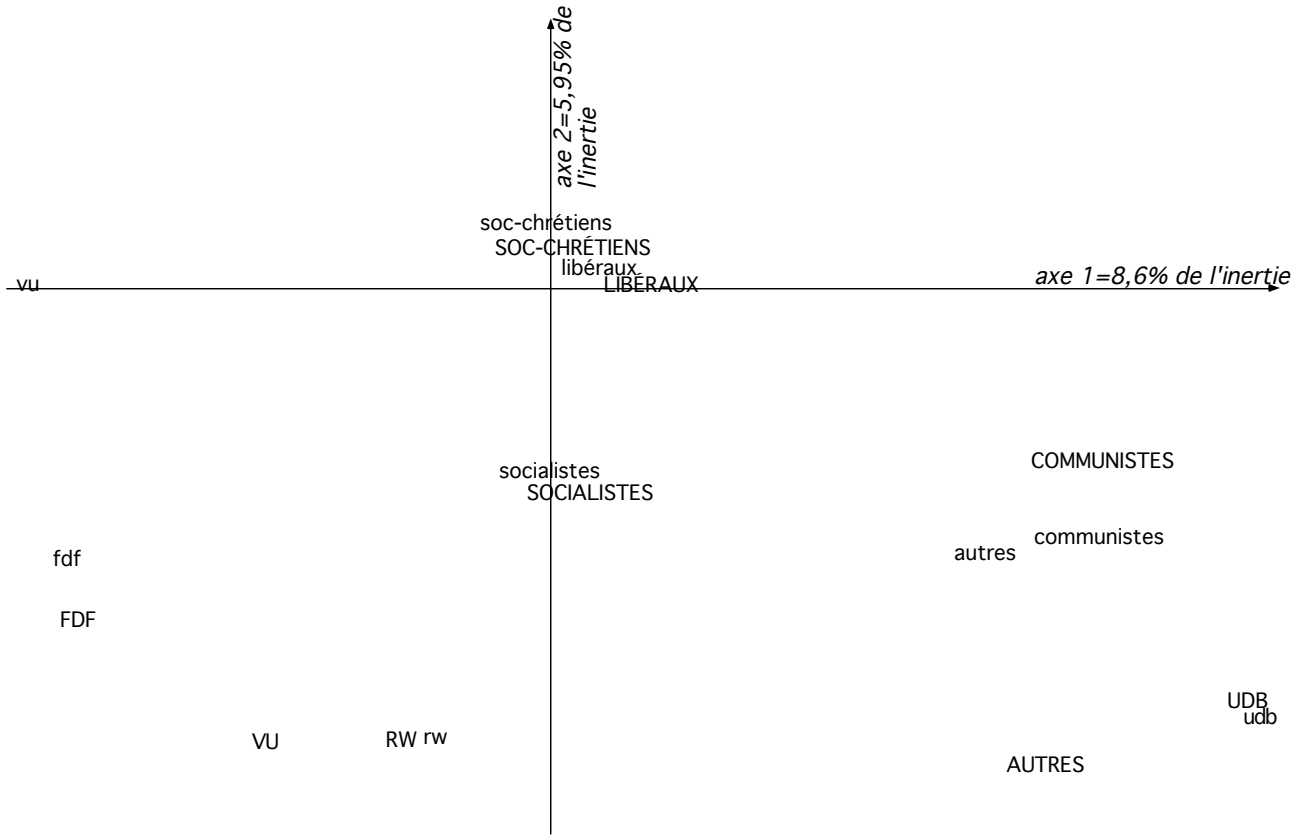

trésorerie
investissements
quant
développer
Congo
organismes
distribution
Communes
poursuivie
impôts
recettes
fonds
1982
prendra
veillera
expansion
francs
compétitivité
financement
Sécurité
dépenses
chômage
secteur
national
redressement
programme
solde
notamment
Exécutifs
seront
sera
enseignement
concours
moyennes
1981
action
mesures
commerce
net
budgétaires
emploi
sécurité
réforme
initiatives
cadre
objectif
sociale
interlocuteurs
par
nationale
salaires
1980
européenne
classes
européen
promotion
entreprises
politique
des
du
socio
crise
le
loi
décidé
rencontrer
base
de
dommages
restructuration
outre
publiques
concertation
meilleure
axe 1
compétences
reconversion
solidarité
une
toutes
depuis
articles
devra
liberté
il
défis
notre
ceux
charbon
maîtrise
qui
globale
plus
Communautés
industrielle
être
elle
entend
ordre
armée
processus
croissance
lui
dialogue
structures
est
problème
libération
pas
Régions
Constitution
ce
nous
que
qu
a
mois
réequipement
homme
n
y
ne
environnement
société
c
mais
ils
peut
aider
prochainement
pourquoi
rénovation
fait
si
faire
Région
partis
guerre
appauvrissement
vous
faut
conseils
dire
Communauté
matières
tâche
croit
citoyen
parlementaire
devons
accord
comité
appel
doute
majorité
à
agit
reconstruction
seraient
régler
serait
mr
oeuvre
président
rappeler
fort
je
langue
régionalisation
ni
texte
axe2

Le plan des formes permet de voir quelles formes sont sur-utilisées ou sous-utilisées par les gouvernements (par rapport à l'usage moyen du corpus).

En première analyse, on peut superposer les axes du plan des déclarations et celui des « mots » et associer les mots et les déclarations qui se situent dans le même quadrant : ces déclarations sur-utilisent ces « mots » et sous-utilisent les mots qui se trouvent dans le quadrant opposé en diagonale.

Mais : on ne peut s'arrêter à une analyse de ce type. Il faut étudier les règles d'interprétations. Voir Salem et Lebart, analyse textuelle ou Philippe Cibois, l'analyse factorielle, PUF.

De toute manière quelle que soit l'interprétation que l'on avance elle doit être confirmée par l'analyse des spécificités, vérifiée dans les concordances ou mieux dans le texte du corpus.



SOCIALISTES... : points représentants les partis en fonction du nombre de postes ministériels
socialistes... : points représentants les partis en fonction du nombre de partis membres de la coalition.
VU, RW et rw trop éloignés ont été ramenés dans le cadre.

On peut également faire figurer sur le plan les modalités d'une variable qui n'a pas été à la base de l'AFC; Cela s'appelle projeter la variable en élémément illustratif.

Ici il s'agit de la variable de la composition partisane des coalitions gouvernementales.

On distingue bien les partis « communautaires » présents dans le quadrant inférieur gauche (celui du discours de la réforme de l' État) et la présence par exemple du parti communiste dans le quadrant inférieur droit caractéristique du discours social de la reconstruction de l'immédiat après-guerre.

Peroraison

Je ferai miens les mots d'encouragement de celle d'Achille Van Acker (1954) peut-être pas pour la grandeur de la Belgique mais pour l'écriture d'une thèse

« Tous au travail ! »

À l'Université libre de Bruxelles

un groupe de recherche : le GRAID
(Groupe de recherche sur les acteurs internationaux et leurs discours)
dir. C. GOBIN

un cours : Analyse des discours politiques et médiatiques,
5 ECTS (Philosophie et lettres)
titulaires : C. Gobin et F. Heinderyckx

un laboratoire interfacultaire en formation : le LADISCO
dir. L. Rosier.

Quelques ressources

La revue MOTS. Les langages du politique : la collection commence en 1980 ; elle est disponible à la bibliothèque de l'ULB ; du numéro 1 à 64 : elle est téléchargeable en ligne sur le site public de Persée : ; les n°s 77 et 78 sont téléchargeables sur le site officiel de Mots ; quant aux derniers numéros : 79-85 sont accessibles leurs sommaires et les résumés des articles.

Histoire et Mesure : http://www.persee.fr/listIssues.do?key=hism

La revue électronique Lexicometrica (avec les actes des JADT)
http://www.cavi.univ-paris3.fr/lexicometrica/

Le livre de Salem et Lebart http://egsh.enst.fr/lebart/ST.html

Lexico3, le logiciel d'André Salem, laboratoire Syled (Paris3)

DTM, le logiciel de L.Lebart + articles http://egsh.enst.fr/lebart/

Le site de Jean Veronis : http://www.up.univ-mrs.fr/cgi-veronis/blog-cat
(Aller voir à "Outils" : il a mis en ligne des textes constitutionnels dont le projet de constitution européenne)