Extraction des isotopies d'un corpus textuel : analyse systématique des structures sémantiques et des cooccurrences, à travers différents logiciels textométriques

Margareta Kastberg Sjöblom, ELLIADD, Université de Franche-Comté

Jean-Marc Leblanc, Ceditec, Université de Paris Est UPEC

Introduction

Cet article s'intéresse à l'analyse des données textuelles et plus précisément à la sémantique lexicale appliquée ici au rituel politique. Les outils fournis par la statistique lexicale et par la textométrie (ou lexicométrie) ouvrent aujourd'hui la voie à de nombreuses pistes de recherche dans le domaine de la linguistique textuelle et à l'analyse du discours politique, permettant de reconstruire les thématiques majeures d'un corpus de façon systématique.

On se propose ici de prendre pour terrain d'expérimentation un corpus de discours politique rituel, constitué d'allocutions de vœux faites à la presse par quelques Premiers ministres français de la Cinquième République de 1976 à 2007 (Barre, Mauroy, Rocard, Cresson, Balladur, Juppé, Jospin, Raffarin, Villepin).

Les allocutions de vœux politiques ont déjà fait l'objet d'études statistiques et d'analyses, mais on a plus souvent analysé les vœux présidentiels et non les vœux ministériels émanant de Matignon[1]. Il nous semble cependant intéressant de confronter les deux discours et de prendre comme point de référence les messages de vœux aux Français des présidents de la Cinquième République sur fond méthodologique.

Il s'agit de tester la validité des méthodes textométriques sur des données sinon réduites, du moins quantitativement limitées, en comparaison à de très gros ou de gros corpus habituellement soumis à ces analyses. Au-delà de la dimension politique et rituelle, ce corpus constituera avant tout un terrain d'expérimentation et d'application méthodologique. Nous nous procéderons ici à une comparaison des méthodologies proposées aujourd'hui par les différents logiciels textométriques.

Une problématique essentielle portera sur les différentes formes de rapprochement d'items lexicaux dans un corpus textuel : quels sont les différents regroupements et constellations sémantiques d'un texte ? Nous tenterons ici d'éprouver un certain nombre de méthodes d'analyse des données textuelles pour mettre à jour les isotopies (Rastier 1987-1996) ou isotropies du corpus (Viprey 2005).

L'étude automatique de la micro-distribution des termes (Viprey 2005) nous incite à aborder l'extraction automatique des univers sémantiques par des biais différents : d'un côté par l'extraction d'un univers thématique, gravitant autour d'un mot-pôle, de l'autre par le recensement des cooccurrents et des séquences d'items.

1. Les méthodes cooccurrentielles

[1] Cf Leblanc 2005

La cooccurrence se définit comme la présence simultanée de deux ou plusieurs items lexicaux dans un même énoncé ou dans un même espace donné. On cherche ainsi la corrélation entre deux items, autrement dit la proximité de ces items. Ce traitement des cooccurrences et l'extraction automatique des réseaux cooccurenciels est depuis longtemps l'un des enjeux majeurs de la lexicométrie.
La collocation est l'association habituelle d'un mot à un autre au sein d'une phrase, un rapprochement de termes qui, sans être fixe, n'est pas pour autant fortuit, comme : « prendre peur », « petite voix », « grièvement blessé ».

Des expressions, en français, comme « passer outre », « mauvaise foi » ou « flagrant délit » sont des locutions, des lexies dont la signification ne se déduit pas de celles de leurs constituants et de la structure syntaxique de leur association. En ce sens, selon J.-M. Viprey (Viprey 2012), elles participent de ce que l'on entend par lexique du français et elles sont décrites, par les dictionnaires, au même titre que les « mots simples » (ces unités lexicales atomiques sont des « mots composés » lorsque ce terme est admis).

La notion de *segment répété*, élaborée par André Salem (1987), est une spécification statistique de la collocation. Conçu pour fonctionner sur la surface graphique (formes fléchies), il est constitué d'occurrences contiguës de formes simples fléchies, et repéré par une fréquence remarquable, à l'origine en nombre d'occurrences, dans des versions ultérieures en probabilité composée. Le segment répété peut être interprété (Viprey 2012) comme une voie particulière pour identifier des lexies composées en discours, des répétitions insistantes et des semi-figements discursifs. En principe on exclut du relevé des segments répétés les occurrences de locutions déjà bien identifiées « en langue » (dictionnaires et grammaires). Dans toutes ces dimensions, le segment répété se présente bien comme une variété de collocation au sens classique du terme.

Il y a cooccurrence lorsque deux items, deux « types », deux classes présentent chacun conjointement au moins une occurrence dans un espace considéré. Il n'est jamais simple d'établir un rapport bi-univoque entre un type et ses occurrences, sauf à en rester à la surface graphique.

Cet espace est délimité dans la linéarité de l'énoncé. Il peut être délimité par des instances structurelles identifiées préalablement (entre deux pauses d'un même type, à l'oral, entre deux ponctuations – fortes ou faibles – ou dans un paragraphe, ou dans un vers, etc. à l'écrit), ou déterminé par des critères quantitatifs (nombre de mots, de caractères à l'écrit, unités de passage du temps à l'oral), ou par une quelconque combinaison de ces deux ordres de paramètres.

Dans les recherches littéraires l'étude de la cooccurence s'avère un outil précieux, aussi bien que dans les études sur le discours politique. Le traitement statistique cooccurrentiel d'items lexicaux nous fait franchir un pallier important dans la statistique lexicale pour basculer du côté de la sémantique et du côté de la phraséologie.

Toutefois, aujourd'hui l'on ne peut plus se limiter à une vision simpliste et binaire de la cooccurrence, il s'agit désormais de prendre en compte les différentes formations et constellations sémantiques d'un texte qui sont complexes et multiples.

Nous nous intéressons ici à deux méthodes complémentaires : les cooccurrents associés à un pôle et les cooccurrences dites généralisées[2].

Du point de vue des outils logiciels la cooccurrence est présentée de manière parfois assez différente. On parle ainsi de cooccurrents spécifiques pour le logiciel *Lexico 3* ou bien d'environnement thématique pour *Hyperbase*, pour s'en tenir ici à la recherche de cooccurrents associés à un pôle[3]. On parle aussi de lexicogrammes simples ou récursifs, associés ou non à un pôle, pour *Weblex*[4], d'associations ou de corrélats sémantiques avec *Hyperbase*, de classes sémantiques établies par *Alceste* sur la base cette fois de cooccurrences généralisées au sein d'énoncés[5], de micro et de macro-distribution pour *Astartex*, pour ne citer que quelques exemples.

Les différents logiciels procèdent de manières parfois différentes non seulement dans le calcul, mais aussi dans la manière de visualiser les résultats, préoccupations majeure de l'utilisateur du logiciel, non nécessairement spécialiste en calcul statistique. Nous présenterons ici quelques-unes de ces méthodes[6].

2. Présentation du corpus : vœux à la presse des Premiers ministres français

Nous disposons ici d'un recueil de 59741 occurrences (selon *Lexico 3*) pour 7011 formes, qui s'étend de janvier 1979 à janvier 2008, avec quelques écueils[7].

Ceci constitue un total de 21 messages prononcés par 10 locuteurs différents (Barre, Rocard, Cresson, Balladur, Juppé, Jospin, Raffarin, Villepin, Fillon). Il conviendra de souligner que nous disposons pour Jospin de cinq messages devant la presse, trois pour Raffarin et pour Rocard, deux pour Balladur et Mauroy, un seul pour Fillon, Juppé, Cresson et Barre.

Nous produisons ici les principales caractéristiques quantitatives de notre corpus. A côté du nombre d'occurrences de chaque intervention, nous notons la longueur moyenne des messages disponibles pour chaque locuteur, ainsi que, dans la dernière colonne, le nombre

[2] Nous reprenons ici la terminologie de JM Viprey et développons plus loin cette notion.

[3] Ces deux approches relèvent de principes tout à fait similaires : on recherche dans une fenêtre contextuelle donnée, une forme, un segment, un motif, posés comme pôle puis un calcul probabilisé permet d'apprécier quels sont les mots plus particulièrement représentés autour de ce pôle ; dans cette fenêtre contextuelle, par rapport à ce qui se passe à l'extérieur de la fenêtre contextuelle. Le calcul revient à créer deux groupes, l'un étant la référence de l'autre.

[4] *Weblex* développé par Serge Heiden, n'est plus maintenu, mais proposait une approche intéressante concernant le calcul des cooccurrents. A ce jour ces fonctionnalités ne sont pas implémentées dans la plate-forme TXM.

[5] Précisons qu'un tri croisé sur une forme avec Alceste revient à une recherche cooccurrentielle autour d'un pôle.

[6] Dans une autre contribution, nous faisons référence aux travaux de William Martinez sur les cooccurrences, en particulier à une approche très intéressante qui s'intéresse aux polycooccurrences et aux squelettes de phrases. Cette conception de la cooccurrence apporte un complément intéressant aux méthodes présentées ici (Leblanc Martinez W. (2005).

[7] Manquent en effet plusieurs textes (1980, 1981, 1982), (1985, 1986, 1987, 1988), (1993) (1996). Certains sont disponibles aux archives nationales, [Voeux de la presse à Alain Juppé (16 janvier 1996) ; Voeux à la presse de Jacques Chirac, Premier ministre : discours (6 janvier 1988) ; Fabius Vœux à la presse (9 janvier 1986) ; Fabius Vœux à la presse (11 janvier 1985) ; Mauroy Vœux à la presse (14 janvier 1982) ; Barre Vœux à la presse (13 janvier 1981) ; Chirac Vœux à la presse (13 janvier 1976)], d'autres sont difficilement trouvables. Les données textuelles datant de quelques années n'ont pas encore la disponibilité des textes récents.

d'occurrences total pour chacun, c'est-à-dire le matériau discursif dont nous disposons pour chacun des Premiers ministres.

Partie	Nb occurrences	Longueur Moyenne	Total Locuteur
1979 barre	2786	2786	2786
1983 mauroy	2140		
1984 mauroy	2178	2159	4318
1989 rocard	2059		
1990 rocard	1222		
1991 rocard	1405	1562	4686
1992 cresson	2429	2429	2429
1994 balladur	2166		
1995 balladur	2506	2336	4672
1997 juppé	2098	2098	2098
1998 jospin	3174		
1999 jospin	3774		
2000 jospin	3409		
2001 jospin	4476		
2002 jospin	2826	3532	17659
2003 raffarin	3840		
2004 raffarin	3288		
2005 raffarin	3698	3608	10826
2006 villepin	3597		
2007 villepin	2790	3193	6387
2008 fillon	3880	3880	3880
Données générales	**59741 occ**	**Moyenne: 2844**	**21 discours**

Figure 1 : Principales caractéristiques quantitatives du corpus *Premiers minis*tres

Ce tableau mérite qu'on s'y arrête. Les occurrences disponibles ne sont pas identiques pour chacun des locuteurs : 17659 pour Jospin, 2098 occurrences pour Juppé, pour ne retenir que les cas extrêmes. Ce déséquilibre nous autorise à poursuivre notre analyse, la taille des parties restant dans un rapport acceptable (1 à 5 environ). Il convient bien sûr de tenir compte des circonstances et du matériau disponible. Jospin, on l'a dit, a présenté ses vœux à la presse cinq fois. Barre ou Juppé, soit parce qu'ils n'ont été Premiers Ministres qu'une année, soit que les données n'eussent été disponibles, n'apparaissent qu'une fois dans notre recueil.

On retiendra aussi la longueur moyenne des messages, exprimée en occurrences pour chacun des locuteurs. La moyenne la plus haute est celle de Fillon avec 3880 occurrences (le message de Sarkozy comptait aussi parmi les plus longs du corpus des voeux présidentiels), la moyenne la plus faible est représentée par Rocard, avec 1562 occurrences en moyenne et un seuil au plus bas de 1222 occurrences en 1990. La moyenne globale de longueur des messages étant de 2844 occurrences sur 21 messages.

Peut-on noter une évolution à ce sujet et traduire ce phénomène en tendance ? Nous pouvons souligner que les moyennes les plus fortes concernent les 4 derniers locuteurs. Ces quatre

dernières moyennes s'opposent ainsi à toute la première moitié du tableau. Les Premiers Ministres seraient ils devenus de plus en plus loquaces avec le temps[8] ?

L'histogramme qui suit matérialise le nombre d'occurrences de chaque discours (en effectif et non plus en moyenne).

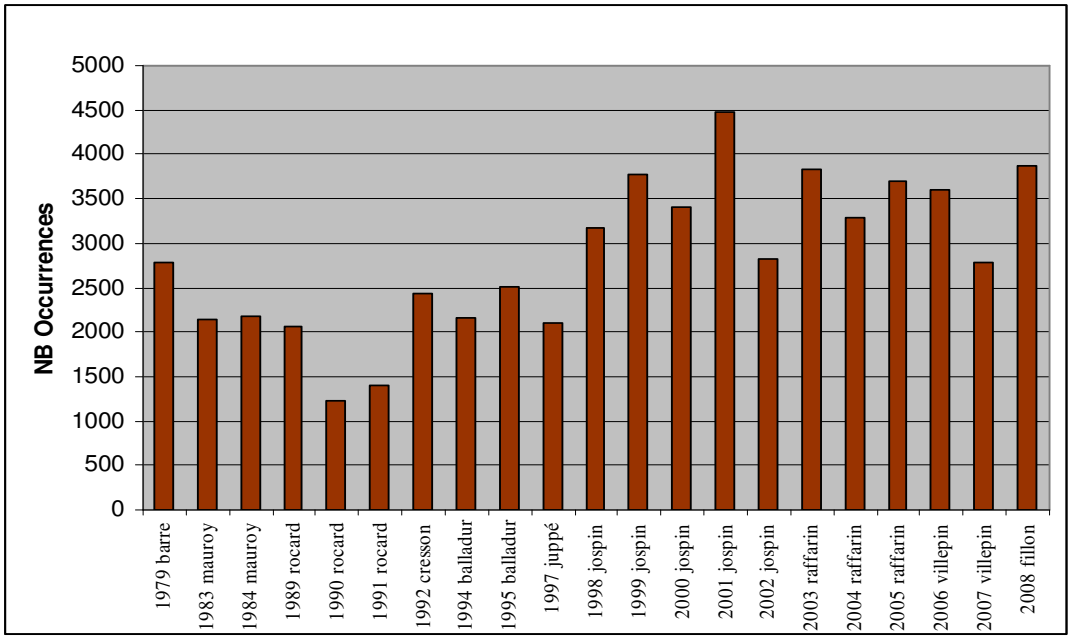


Figure 2 : Distribution des fréquences (absolues) par discours[9]

On y observe que tous les messages disponibles de la période 1979 à 1997 sont situés sous le seuil des 2844 occurrences correspondant à la moyenne et que les autres messages dépassent (ou atteignent pratiquement, pour Jospin 2002) cette moyenne. Il y a bien, semble-t-il, une tendance réelle à l'allongement des messages de vœux à la presse, ce que nous devrons confirmer en complétant notre corpus. Doit on y voir le signe d'une importance de plus en plus grande donnée aux médias ? Qu'en est il des vœux des Présidents de la République aux Français mais aussi à la presse ?

3. Premiers éléments descriptifs du tableau lexical

L'analyse factorielle réalisée sur la partition en textes nous offre une première configuration globale du corpus. Précisions qu'il s'agit ici d'une analyse portant sur le tableau lexical comprenant toutes les formes graphiques jusqu'à la fréquence 5, analyse établie par *Lexico 3*, et que ce tableau croise les 21 textes de notre corpus et la ventilation en fréquence absolue des formes graphiques comprises dans ce corpus.

[8] Dans une analyse des vœux des présidents de la République, nous avions noté une diminution sensible de la longueur des phrases sur l'ensemble de la période.

[9] *Hyperbase* propose une approche un peu différente pour juger de la taille des différentes parties d'un corpus en produisant un histogramme de l'étendue probabilisée de ce corpus.

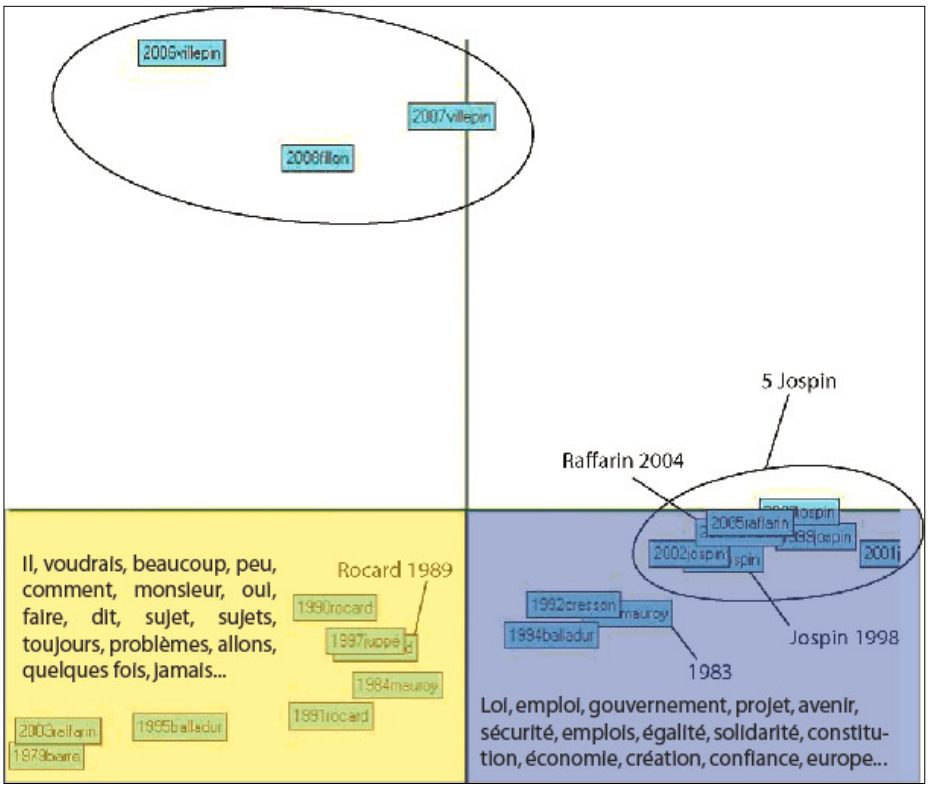


Figure 3 : Analyse factorielle de la partition textes (lexico 3) [12%, 10%] Fréquence 5.

Ce tableau factoriel oppose sur le deuxième axe les trois derniers messages, 2006, 2007, 2008.

Les trois messages de Rocard, 1989, 1990, 1991, sont relativement proches les uns des autres, bien que voisinant avec Juppé 1997, et Mauroy 1984.

Exception faite de Jospin (et de Rocard) pour qui on note une certaine cohérence, cette AFC ne privilégie ni la chronologie ni la personnalité de locuteurs.

Raffarin se trouve pour deux messages à proximité de Jospin, ce qui pourrait relever d'une logique diachronique puisque les deux locuteurs se succèdent à Matignon. Cependant, le premier message de Raffarin (2003) s'oppose aux deux suivants sur le premier axe. Peut-on noter une rupture entre Jospin et Raffarin, et un retour de Raffarin au style Jospin ? D'autres mesures pourraient nous apporter des éléments de réponse.

Les messages de Balladur, 1994 et 1995, sont en opposition de part et d'autre de l'axe vertical, de même que Mauroy 1983 et 1984.

Comment interpréter sur l'axe 2 la « singularité » des trois derniers messages ?

Ce ne sont pas nécessairement les spécificités positives des trois parties 2006, 2007, 2008 qui livrent l'information la plus pertinente. On note cependant le suremploi du « civilisation » exclusif de cette dernière partie qu'il est intéressant de rapporter au message de Sarkozy en décembre 2007. C'est bien sûr le message de Fillon qui évoque la notion de civilisation.

Mais il est plus intéressant ici d'interpréter l'opposition qui construit le premier axe.

Les formes les plus contributives de cet axe, qui correspondent aux spécificités de la partie droite nous livrent des informations intéressantes :

Loi, emploi, gouvernement, projet, avenir, sécurité, emplois, égalité, solidarité, constitution, économie, création, confiance, Europe, sont autant de termes du politique de l'action, du volontarisme. Ainsi, plus on avance sur la droite de l'axe, plus en se dirige vers un vocabulaire politique…

Les spécificités négatives sont plus parlantes encore :

Il, voudrais, beaucoup, peu, comment, monsieur, oui, faire dit, sujet, sujets, toujours problèmes, allons, quelque fois, jamais. Ces formes relevant de l'énonciation ou du métadiscours, des formes d'adresse également, sont donc celles qui sont sous-employées dans la partie droite de l'AFC par rapport à l'ensemble du corpus. Toutes les formes du non-politique, du non-thématique sont en sous-emploi dans cette partie. Notons cependant un peu plus bas les termes de *guerre*, de *pluralisme* et de *changement*, qui apparaissent parmi les sous-emploi du groupe. Ainsi pouvons nous noter sur cet axe une sorte de gradation allant de l'énonciatif au politique, de la gauche à la droite de l'axe.

Le poids des discours de Jospin est bien sûr déterminant, mais l'espace manque ici pour préciser cet aspect des données. Cependant les spécificités positives de Jospin confirment le caractère éminemment politique de ses interventions : *projets, loi, emploi, emplois, gouvernement, cumul…* Quant aux négatives elles sont fortement ancrées dans l'énonciation : *je vous dire, cela, bien peut, voudrais, souhaite, président, problème…*

Ainsi, une AFC, portant sur le tableau lexical, est apte à nous livrer des informations qui ne sont pas nécessairement liées aux seules caractéristiques individuelles mais peuvent avoir une résonance plus globale, ici l'opposition du politique/non politique.

Les autres approches mettent-elles cette opposition du politique et de l'énonciatif en lumière ?

4. Cooccurrences multiples entre énoncés : ALCESTE

Doit-on revenir sur le principe et sur la philosophie du logiciel *Alceste* ? Rappelons simplement que sur la base d'unités de contexte élémentaires le logiciel produit des classes d'énoncés à partir de la distribution des formes réduites qui composent ces UCE, ce qui revient à parler de cooccurrences multiples. Cette méthodologie, basée sur une approche Harrisienne du texte est désormais implémentée dans un outil logiciel libre, *Iramuteq* développé et conçu par Pascal Marchand et Pierre Ratinaud.

Alceste, après la reconnaissance des formes, l'élimination d'hapax et la réduction des désinences de conjugaison, effectue les calculs de données et l'analyse des unités de contexte élémentaires (U.C.E.).

Il est intéressant de mentionner un mode d'interprétation possible : il s'agit d'attribuer à chaque unité une appartenance à une classe topique qui divise le vocabulaire par rapport à la posture : le réel, l'imaginaire et le symbolique. Cette division en classes topiques est l'aboutissement d'une réflexion sur la matérialité de l'activité langagière et la notion de posture, une approche bien différente de celle de divisions sémantiques conventionnelles qui divisent le vocabulaire en classes catégorielles selon le contenu[10] ou bien d'une analyse onomasiologique traditionnelle. On ne trouvera pas dans cette démarche une grille de contenu ou une catégorisation formelle, mais la recherche d'un découpage permettant d'isoler des mondes lexicaux stabilisés à partir d'un calcul purement statistique des probabilités et des cooccurrences. Pour l'estimation statistique *Alceste* fait appel à un corpus d'étalonnage d'environ 26 millions de caractères réunissant des textes du 19ème et du 20ème siècles[11].

Nous avons soumis notre corpus à *Alceste* et obtenons la configuration suivante :

Cinq classes émergent de l'analyse, pour 68.8% des énoncés classés. On remarque assez vite une configuration qui oppose des classes du politique (classes 5, 3 2) et deux classes du rituel et de l'énonciation, ou plutôt du métadiscours (classes 4 et1).

Il s'agit en effet des discours très ritualisés ; des vœux pour la nouvelle année, exprimés devant un public de journalistes, encadrant, comme le témoignent les classes 2, 3 et 5, un message politique et gouvernemental.

On notera l'opposition forte qui naît de la classification descendante entre politique et énonciation.

- La classe 5, couvre un lexique purement économique gravitant autour des formes *chômage*, *emploi, croissance, augmentation* et *diminution achat, salaires*…

- La classe 3 est plus sociétale, même si l'homogénéité n'est pas aisément détectable. Les énoncés qui la composent portent à la fois sur l'Europe, sur l'immigration, sur la violence à l'école et sur la délinquance, le tout sur fond de modèle républicain.

- La classe 2, qui se situe toujours dans cette première moitié de l'arborescence, est fortement ancrée dans un vocabulaire législatif : *loi, réforme, parlement, texte, projet* (de loi), *législatif*… sont en effet les formes réduites les plus significatives de cette classe.

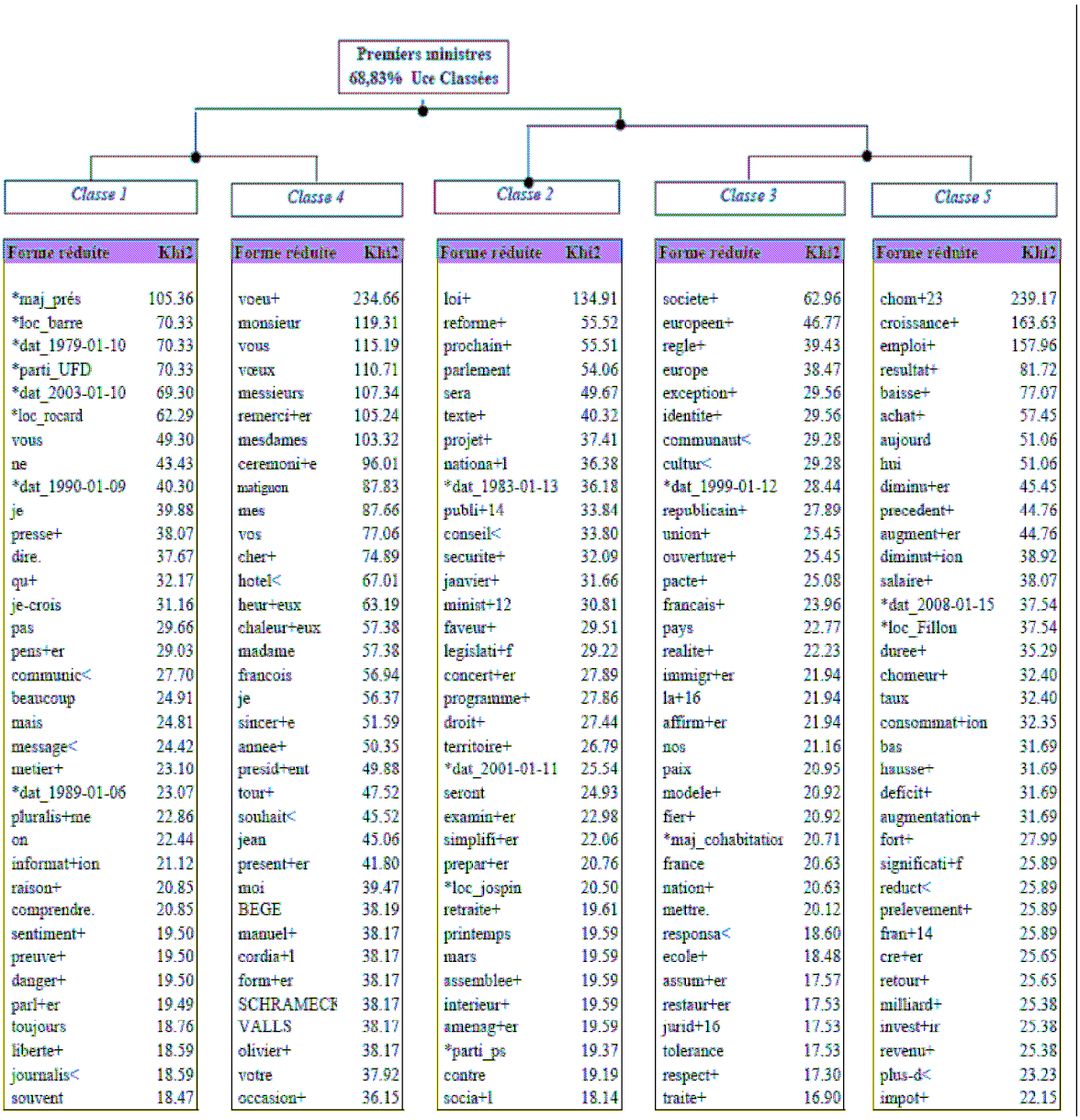
Les deux dernières classes s'opposent diamétralement aux trois premières, à la fois sur le plan de la structure de l'arborescence et sur le plan sémantique. C'est bien une opposition du politique et de l'énonciatif qui se construit ici.

- La classe 4 est essentiellement rituelle : *vœux, vœu, cérémonie, remercier, présenter heureux, chaleureux, sincère* et riche en formules d'adresses : *monsieur, messieurs, madame, chers*, …

[10] Voir M. Reinert (2008) pour plus d'explication sur cette méthode et cette approche philosophique du langage.

[11] Idem.

- La classe 1 est plus anecdotique et fortement liée à des variables individuelles. (Rocard et Barre en particulier.). Ces deux locuteurs présentent en effet un ethos atypique et se définissent souvent comme étant apolitiques (Barre surtout). Ils manifestent par ailleurs une certaine connivence avec les journalistes, qui se présente différemment des autres chefs de l'exécutif. Par ailleurs, au-delà de cette connivence et du caractère anecdotique des énoncés, la relation entre la presse et le pouvoir, mais aussi la condition des journalistes, leur mission, leur indépendance, sont des thématiques contributives de cette classe. (*Métier, pluralisme, information, liberte, journalisme*…)



Forme réduite	Khi2
*maj_prés	105.36
*loc_barre	70.33
*dat_1979-01-10	70.33
*parti_UFD	70.33
*dat_2003-01-10	69.30
*loc_rocard	62.29
vous	49.30
ne	43.43
*dat_1990-01-09	40.30
je	39.88
presse+	38.07
dire.	37.67
qu+	32.17
je-crois	31.16
pas	29.66
pens+er	29.03
communic<	27.70
beaucoup	24.91
mais	24.81
message<	24.42
metier+	23.10
*dat_1989-01-06	23.07
pluralis+me	22.86
on	22.44
informat+ion	21.12
raison+	20.85
comprendre.	20.85
sentiment+	19.50
preuve+	19.50
danger+	19.50
parl+er	19.49
toujours	18.76
liberte+	18.59
journalis<	18.59
souvent	18.47

Forme réduite	Khi2
voeu+	234.66
monsieur	119.31
vous	115.19
vœux	110.71
messieurs	107.34
remerci+er	105.24
mesdames	103.32
ceremoni+e	96.01
matignon	87.83
mes	87.66
vos	77.06
cher+	74.89
hotel<	67.01
heur+eux	63.19
chaleur+eux	57.38
madame	57.38
francois	56.94
je	56.37
sincer+e	51.59
annee+	50.35
presid+ent	49.88
tour+	47.52
souhait<	45.52
jean	45.06
present+er	41.80
moi	39.47
BEGE	38.19
manuel+	38.17
cordia+l	38.17
form+er	38.17
SCHRAMECK	38.17
VALLS	38.17
olivier+	38.17
votre	37.92
occasion+	36.15

Forme réduite	Khi2
loi+	134.91
reforme+	55.52
prochain+	55.51
parlement	54.06
sera	49.67
texte+	40.32
projet+	37.41
nationa+l	36.38
*dat_1983-01-13	36.18
publi+14	33.84
conseil<	33.80
securite+	32.09
janvier+	31.66
minist+12	30.81
faveur+	29.51
legislati+f	29.22
concert+er	27.89
programme+	27.86
droit+	27.44
territoire+	26.79
*dat_2001-01-11	25.54
seront	24.93
examin+er	22.98
simplifi+er	22.06
prepar+er	20.76
*loc_jospin	20.50
retraite+	19.61
printemps	19.59
mars	19.59
assemblee+	19.59
interieur+	19.59
amenag+er	19.59
*parti_ps	19.37
contre	19.19
socia+l	18.14

Forme réduite	Khi2
societe+	62.96
europeen+	46.77
regle+	39.43
europe	38.47
exception+	29.56
identite+	29.56
communaut<	29.28
cultur<	29.28
*dat_1999-01-12	28.44
republicain+	27.89
union+	25.45
ouverture+	25.45
pacte+	25.08
francais+	23.96
pays	22.77
realite+	22.23
immigr+er	21.94
la+16	21.94
affirm+er	21.94
nos	21.16
paix	20.95
modele+	20.92
fier+	20.92
*maj_cohabitation	20.71
france	20.63
nation+	20.63
mettre.	20.12
responsa<	18.60
ecole+	18.48
assum+er	17.57
restaur+er	17.53
jurid+16	17.53
tolerance	17.53
respect+	17.30
traite+	16.90

Forme réduite	Khi2
chom+23	239.17
croissance+	163.63
emploi+	157.96
resultat+	81.72
baisse+	77.07
achat+	57.45
aujourd	51.06
hui	51.06
diminu+er	45.45
precedent+	44.76
augment+er	44.76
diminut+ion	38.92
salaire+	38.07
*dat_2008-01-15	37.54
*loc_Fillon	37.54
duree+	35.29
chomeur+	32.40
taux	32.40
consommat+ion	32.35
bas	31.69
hausse+	31.69
deficit+	31.69
augmentation+	31.69
fort+	27.99
significati+f	25.89
reduct<	25.89
prelevement+	25.89
fran+14	25.89
cre+er	25.65
retour+	25.65
milliard+	25.38
invest+ir	25.38
revenu+	25.38
plus-d<	23.23
impot+	22.15

Figure 4 : Principales formes réduites et classes thématiques établies par Alceste

5. Traitement cooccurrentiel dans *Hyperbase*

Nous avons soumis le corpus à un traitement par le logiciel *Hyperbase*, dans sa dernière version, 9.0. Ce logiciel, désormais bien connu, n'a plus besoin d'être présenté[12] ; signalons cependant quelques innovations car le traitement statistique des données textuelles permet désormais des approches différentes des réseaux thématiques d'un corpus.

Après la fonctionnalité « thématique » ou « thème » *Hyperbase* propose en effet d'autres fonctions pour l'analyse cooccurrentielle : « topologie », « corrélats » et « associations » , qui relèvent de la même approche. On y considère les mots dans leur environnement immédiat en ignorant la partition en textes.

Il s'agit ici d'une approche fondée sur les corrélats. Au lieu de reposer sur une segmentation pré-établie, d'un thesaurus préalablement constitué, ici les mots reconnaissent d'eux-mêmes leur parenté, du seul fait de leur voisinage dans les mêmes contextes.

5.1 Tableau général des cooccurrences

Le tableau général des cooccurrences reprend quelque chose qui existe depuis bien longtemps et que l'on trouve sous *Weblex* sous le nom de *cooccurrences*, par opposition à *lexicogramme* simple, récursif, associés ou non à un pôle. Il s'agit ici de faire la recension de toutes les rencontres attestées dans notre corpus.

10.11 ministre premier	3.83 mise oeuvre
9.73 loi projet	3.82 formation jeune
6.75 année début	3.79 cours travail
6.17 action gouvernement	3.78 projet sens
5.67 information rôle	3.77 année voeu
5.62 liberté presse	3.74 compte équilibre
5.59 croissance emploi	3.74 justice réforme
5.54 majorité soutien	3.72 accord partenaire
5.52 jeune plan	3.70 europe france
5.26 emploi jeune	3.67 droit respect
4.93 gouvernement majorité	3.66 démocratie presse
4.67 mise place	3.62 cadre loi
4.64 mesdames voeu	3.56 croissance résultat
4.55 emploi plan	3.54 messieurs voeu
4.49 loi parlement	3.53 état solidarité
4.47 monde voix	3.48 domaine rapport
4.38 mesdames messieurs	3.45 conférence partenaire
4.36 année messieurs	3.38 ensemble profession
4.29 chômage emploi	3.38 an fois
4.17 fin janvier	3.37 avenir europe
4.06 partenaire travail	

Figure 5 : Tableau général des cooccurrences

[12] E. Brunet (2012) *Hyperbase, Manuel de référence, versions 8.0. et 9.0.*

5.2 Corrélats :

L'extraction des corrélats, pourrait s'apparenter à la démarche du logiciel *Alceste*[13], mais plus encore d'*Astartex* puisque c'est le principe de la *cooccurrence généralisée* qui est repris ici. :

Le programme commence par établir une liste de mots (les substantifs ou adjectifs qui ne sont ni trop rares, ni trop fréquents) et enregistre toutes leurs rencontres, occasionnelles ou insistantes, dans la même page[14]. Un lien est établi entre deux mots quand ils ont tendance à se donner rendez-vous. La "tendance" tient compte du nombre de cooccurrences, dont le registre est tenu dans un tableau carré où les mêmes éléments sont portés sur les lignes et les colonnes (Viprey, 2004).

Le choix de la page permet d'échapper en partie aux contraintes syntaxiques qu'imposerait le choix d'une unité linguistique plus courte (syntagme, phrase ou paragraphe). L'élimination des mots fréquents et des mots-outils contribue aussi à privilégier les relations sémantiques ou thématiques plutôt que les rapports de dépendance syntaxique. La division en textes est également ignorée.

La cohabitation à longue distance dans un même texte n'entre pas dans le calcul. Seule compte la proximité immédiate dans la même page, là où l'on a le plus de chances de relever les isotopies.

L'extraction des items est automatique. Compte tenu de l'étendue du corpus, le programme de sélection s'arrange pour retenir entre 200 et 400 items. Ensuite vient une phase, assez longue, d'exploration séquentielle et d'analyse d'associations du corpus. Dans chaque page on teste la présence ou l'absence des éléments de la liste, on calcule les distances et on trie le détail des associations deux à deux, en notant les cooccurrences. Nous aboutissons à une matrice carrée qui croise chaque mot avec tous les autres ou plutôt avec lui-même. Cette méthode reprend celle imaginée par JM Viprey dans *Astartex*, comme nous le verrons plus bas.

On regroupe donc les items lexicaux les plus fréquents (par défaut 400) et les mieux répartis dans le corpus et on établit une carte synthétique de leurs cooccurrences par une analyse factorielle des correspondances. On voit bien ici la différence primordiale qu'il y a entre cette factorielle et celle pratiquée sous Lexico 3 (cf. figure 3). JM Virpey propose d'autres visualisations de ces matrices. (cf Figure 12)

[13] Pour plus de détails voir Reinert M. *Notice d'Alceste*.

[14] Voir E. Brunet, *Manuel d'Hyperbase*.

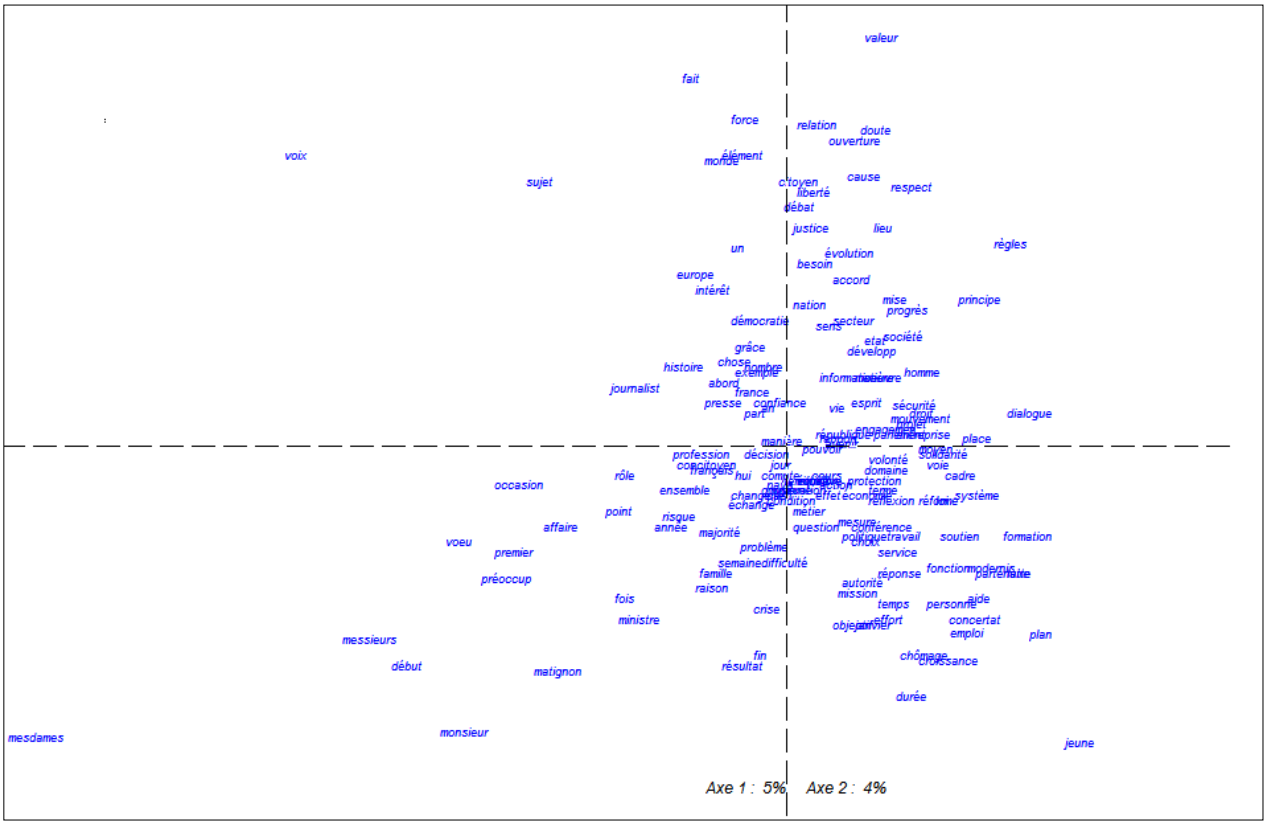


Figure 6 : analyse factorielle des corrélats

Le graphique obtenu est assez encombré mais l'interprétation en est pourtant claire : de la gauche à la droite on passe du moment de l'énonciation et du rituel des voeux *ministre vœux, messieurs, mesdames, Matignon* à des considérations éminemment politiques voire économiques (*Chômage, croissance, emploi, durée, travail, formation...*).

Quant au deuxième facteur il paraît séparer notamment sur la droite de l'axe les mesures politiques aux valeurs (*règles, principe, dialogue, respect, progrès, justice citoyen...*). Cette configuration rappelle celle obtenue au moyen d'Alceste, dans les grandes lignes.

5.3 Choix d'un pôle parmi les corrélats

On peut envisager d'aborder les isotopies spécifiques à partir d'un pôle, Cette analyse permet de considérer les associations de mots dans leur environnement immédiat, toujours en ignorant la partition des textes.

L'extraction des corrélats par le logiciel *Hyperbase* regroupe comme nous venons de le mentionner les items lexicaux qui ne sont ni rares ni trop fréquents dans le corpus. En voici l'illustration.

abord accord action affaire aide an année autorité avenir besoin cadre cas cause changement choix chômage chose citoyen compte concertation concitoyen condition conférence confiance cours crise croissance débat début décision démocratie développement dialogue difficulté domaine doute droit durée échange économie effet effort élément emploi engagement ensemble entreprise équilibre esprit etat europe évolution exemple fait famille fin fois fonction force formation français france gouvernement grâce histoire homme hui information initiative intérêt janvier jeune jour journaliste justice liberté lieu loi lutte majorité manière matière matignon mesdames messieurs mesure métier ministre mise mission modernisation mois monde monsieur mouvement moyen nation nombre objectif occasion oeuvre ouverture parlement part partenaire pays personne place plan point politique pouvoir premier préoccupation presse principe problème profession progrès projet protection question raison rapport réflexion réforme règles relation réponse république respect résultat risque rôle secteur sécurité semaine sens service situation société solidarité soutien sujet système temps terme territoire travail un valeur vie voeu voie voix volonté

Figure 7 : Formes les mieux réparties, telles qu'identifiées par *Hyperbase*

Plutôt que de calculer, comme précédemment, l'analyse factorielle des corrélats et donc, à partir des formes les mieux réparties, de considérer des cooccurrences généralisées, nous pouvons choisir un substantif dans cette liste et en étudier l'environnement.

Ceci revient à proposer une représentation des liens préférentiels qui tissent un réseau autour d'un mot choisi pour pôle, sous forme de graphe, comme ici le mot *action* :

Les cooccurrents les plus proches d'*action* sont, d'après les résultats fournis par *Hyperbase* (les plus proches du pôle) *action, gouvernement, durée, politique, semaine, cadre, lutte, loi, occasion, projet, emploi, concitoyen*.

Le calcul du graphe arborescent, des nœuds et des arcs, est assuré par le logiciel libre GRAPHVIZ[15]. Les données sont fournies à ce programme selon les spécifications du langage DOT et les résultats bruts sont repris par *Hyperbase* dans une représentation graphique qui tient compte non seulement des positions mais aussi des pondérations[16].

Les mots en rouge correspondent ici à des nœuds de forte fréquentation, et ceux en noir à des nœuds moins fréquentés, n'ayant pas de contact direct avec le mot-pôle. Les traits gras correspondent aux cooccurrences directes avec le pôle, et les traits fins aux cooccurrences indirectes, c'est-à-dire en cooccurrence avec les cooccurrents.

[15] Voir E. Brunet, *Manuel d'Hyperbase* et S. Heiden, *Notes de Weblex*.

[16] La mesure de la cooccurrence est empruntée au Rapport de Vraisemblance, proposé par Dunning en 1993. Cet indice s'appuie sur quatre paramètres

- a : nombre de cooccurrences des deux mots dans le champ exploré (ici le paragraphe)
- b : nombre d'occurrences du premier mot en l'absence du second
- c : nombre d'occurrences du second mot en l'absence du premier
- d : nombre d'occurrences des autres mots RV = -21ogL = 2(sl-s2)

pour s 1 = a log a + b log b + c log c + d log d + (a+b+c+d)log(a+b+c+d)

s2 = (a+c)log(a+c) + (b+d)log(b+d) + (a+b)log(a+b) + (c+d)log(c+d)

A partir de la valeur 4, l'indice de Dunning est considéré comme échappant au hasard, au seuil de 5%.

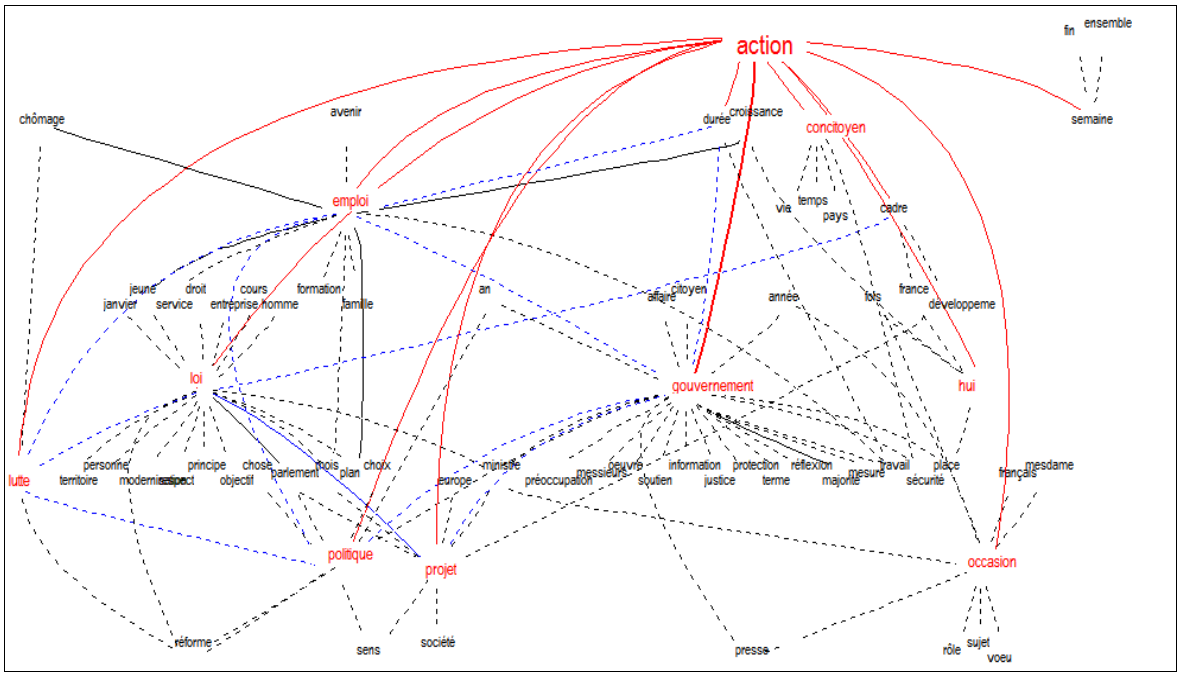


Figure 8 : Associations à partir du pôle *action*

Nous voyons donc ici, non seulement la proximité des items, mais aussi leur force d'attraction. Les liens les plus forts de l'item *action* sont *gouvernement, emploi, politique, projet* et *occasion*. Le graphique permet aussi de recenser les liens qui entretiennent les cooccurrents avec d'autres items lexicaux, *le gouvernement* étant lié à l'*information*, à la *réflexion, à la sécurité, à la préoccupation, à la justice* etc.

Il est possible par ailleurs de produire un graphe simplifié qui ne conserve que les associations privilégiées, nous ne le produirons pas ici par manque de place.

5.4 Choix d'un pôle dans l'environnement thématique:

On peut se rapprocher d'une association, d'un microcosme cooccurrentiel d'une autre manière, en recourant à une analyse plus classique, celle qu'Etienne Brunet appelle « environnement thématique ».

Prenons un des lemmes les plus fréquents « croissance »et cherchons le contexte immédiat de ce terme.

L'extraction automatique du contexte d'un item lexical permet la création d'un sous-corpus qui est soumis à un calcul de spécificité particulier, puisqu'on ne recherche plus une relation entre un mot et un texte, mais une relation privilégiée entre les mots eux-mêmes. Cette procédure ne se réduit pas ici à deux mots confrontés, mais à l'ensemble indéfini de tous les mots qui peuvent se trouver dans l'entourage d'un mot (ou d'un groupe de mots) qu'on définit comme étant le pôle. En confrontant le lemme *croissance* au sous-corpus constitué par les mots qui gravitent autour du pôle, ici le paragraphe, nous pouvons extraire l'environnement thématique suivant (ordre hiérarchique, début de liste) :

Le principe peut être comparé aux cooccurrents spécifiques de *lexico3*. Il n'est pas très éloigné non plus d'un tri croisé sur une forme avec Alceste.

Classiquement, l'environnement thématique du terme croissance (ici du lemme) se présente sous *Hyperbase* sous la forme de deux listes : l'une alphabétique et l'autre hiérarchique.

A partir de la liste hiérarchique les cooccurrents les plus proches du pôle permettent d'établir un graphe de connexion.

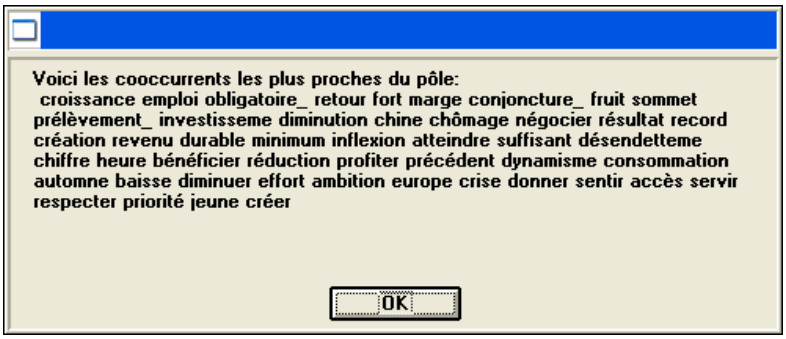


Figure 9 : cooccurrents les plus proches du pôles établis par Hyperbase

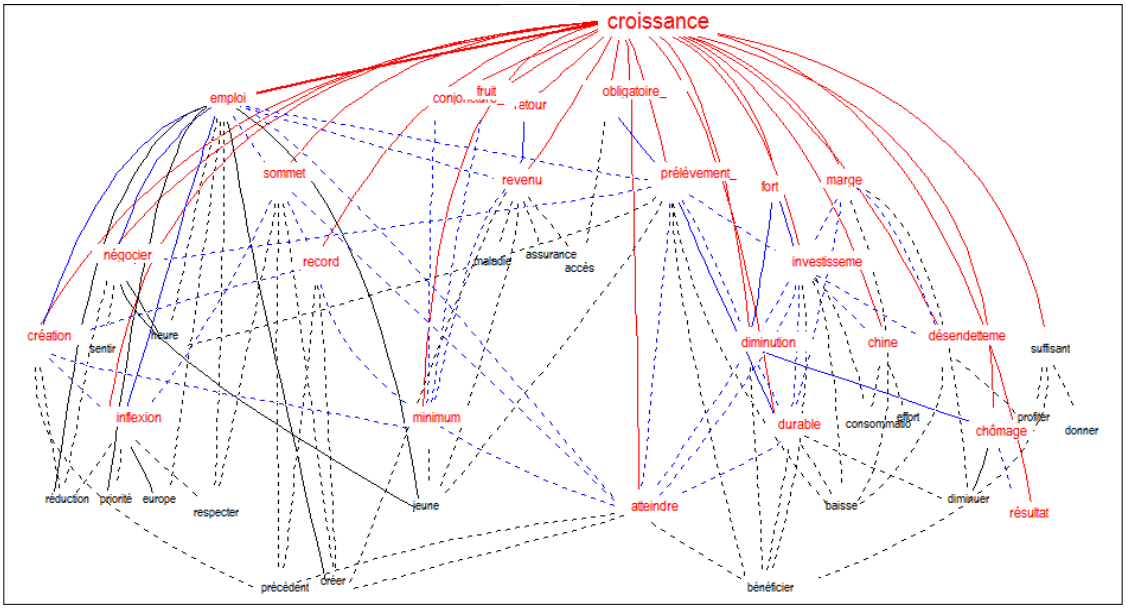


Figure 10 : relation des cooccurrents à l'intérieur d'un thème

Il s'agit ici d'un univers thématique qui est créé à partir d'un mot-pôle.

Ici le calcul et le tri sont effectués de la même manière que dans le calcul précédent, mais on ne prend en compte que l'univers thématique, c'est-à-dire le paragraphe entourant le pôle, extrait et traité comme un sous-corpus indépendant du corpus dans sa globalité.

Ces différentes manières d'envisager les associations et les environnements thématiques à partir d'un mot-pôle, à l'intérieur d'un corpus, non seulement permettent de comparer différents calculs statistiques, mais fournissent aussi des preuves quant à la solidité de ces analyses. Les résultats sont en effet souvent très semblables et la complétude des différentes études constitue une base solide pour l'étude des isotopies et des collocations dans un texte.

6. Cooccurrences généralisées : Astartex

Pour le logiciel *Astartex* il s'agit d'extraire le profil cooccurrentiel, l'isotropie, des unités non-séquentielles, à partir d'une matrice statistique.

Dans cette analyse, après que le corpus ait été lemmatisé avec l'outil *Diatag*, le logiciel extrait les 250 items les plus fréquents, en ôtant au-préalable les mots-outils. L'analyse factorielle de correspondances montre ici la relation, les réseaux associatifs entre les différents items extraits comme étant cooccurrentiels.

Le tableau n'est pas facile à interpréter, mais en réalité les listes de droite du graphe sont cliquables ce qui permet de repérer l'item concerné. L' interprétation de l'AFC est semblable à celle que fournit le logiciel *Hyperbase* ; en haut, à gauche nous observons le rituel de vœux pour la nouvelle année, en bas à gauche, le discours de la presse, devant un public de journalistes et à droite le discours porteur de message politique.

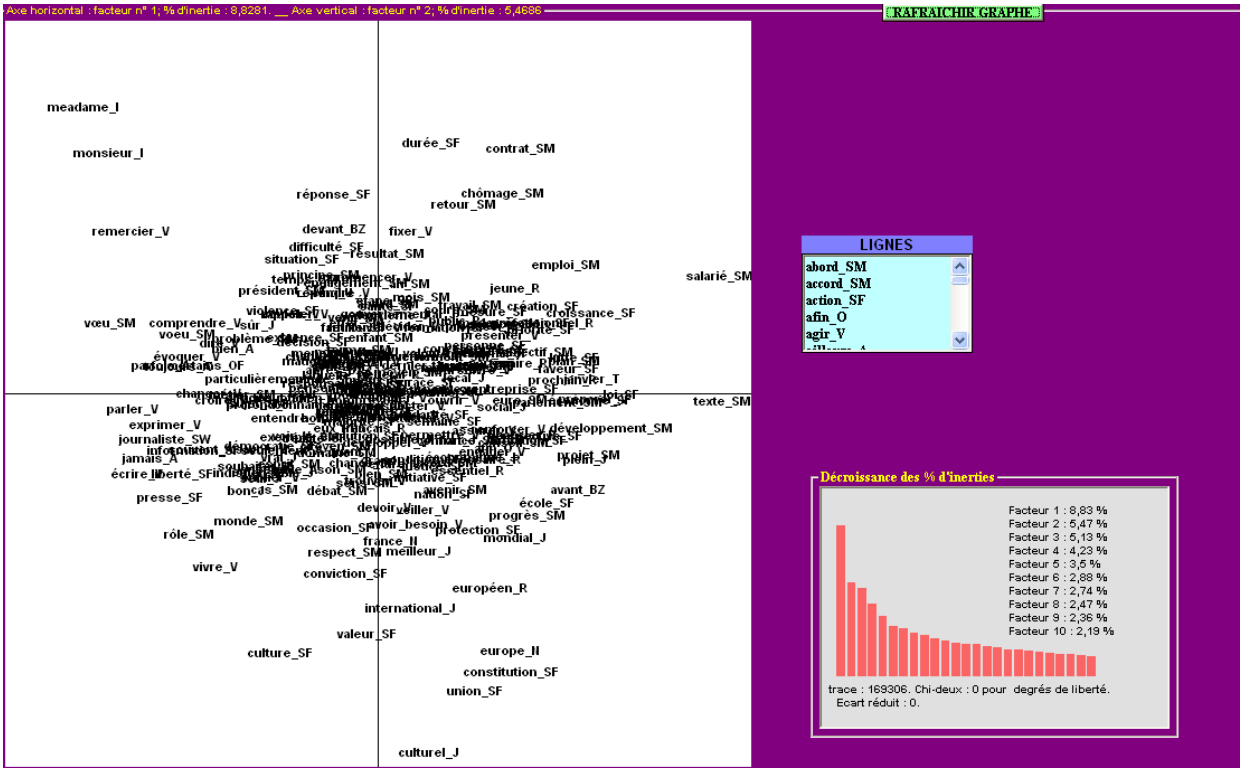


Figure 11 : Projection des deux premiers axes factoriels

Le logiciel *Astartex* permet non seulement visualiser les deux premiers facteurs de l'analyse factorielle de correspondances, mais de visualiser simultanément les trois premiers facteurs de l'AFC, visualisation qu'il est le seul logiciel à permettre dans l'état actuel. *Iramuteq*, XLSTAT3D ou encore TextObserver permettent la visualisation de l'analyse factorielle en trois dimensions mais *Astartex* va plus loin que la simple implémentation du troisième axe en proposant des visualisations originales.

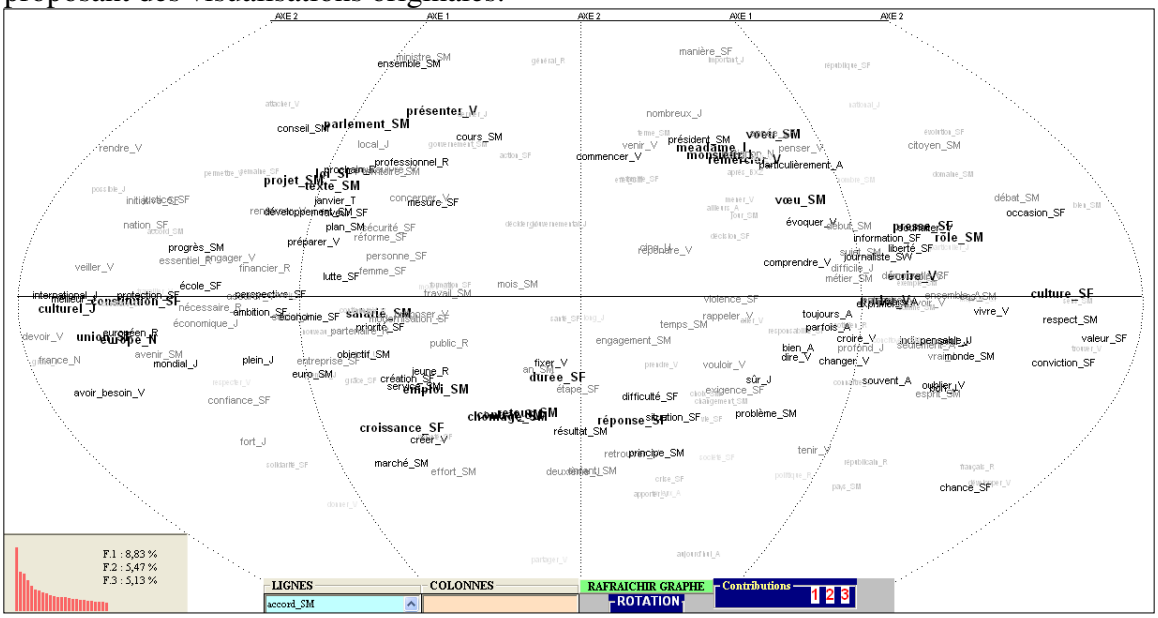


Figure 12 : Projection de l'AFC sur trois axes ou micro distribution des 250 items les plus fréquents du corpus sous le mode « géodésique »

Cette visualisation prend l'aspect d'un planisphère, d'une représentation géodésique, qui en réalité peut tourner et donne une dimension visuellement plus riche aux constellations d'associations d'items lexicaux et de réseaux sémantiques à l'intérieur d'un corpus. On notera l'opposition chez Viprey entre macro distribution et micro distribution qui oppose la répartition des différentes parties d'un corpus à la répartition des *items* de ce corpus et de leurs cooccurrents.

Un double travail est ici effectué sur la visualisation, puisqu'on introduit le troisième axe, mais plutôt que de s'en tenir à une représentation sur des axes X Y Z, un système de projection qui permet une visualisation plus fine, transforme cette AFC en une visualisation géodésique.

Nous observons ici les mêmes constellations et proximités lexicales que dans les analyses factorielles de correspondances "traditionnelle" à deux axes, les items lexicaux provenant du rituel des vœux regroupés, les items relatifs aux discours oratoire devant les journalistes également, divisé du contenu politique.

Des zooms permettent de se focaliser sur des pôles particuliers, comme ici, celui d'un volet du politique que nous pourrions opposer au rituel.

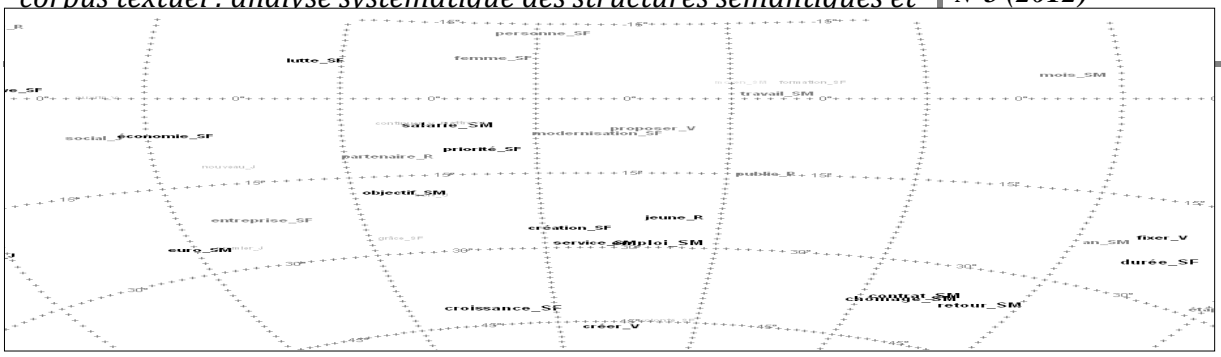


Figure 13 : Zoom sur une région micro distributionnelle du corpus

La visualisation à trois dimensions permet en effet une interprétation plus fine et plus imagée. Regroupées comme des îles sur notre planisphère nous trouvons par exemple la politique intérieure avec les items relevant du marché du travail comme, *salaire*, *travail*, *marché*, *emploi*, *chômage*, etc., et des éléments de la politique extérieure relevant notamment la construction de l'Europe et l'échange culturel international. Ces pôles rappellent globalement certaines des classes identifiées par *Alceste*.

Conclusion

La collaboration dont cette contribution est le fruit aura permis de mettre en relation deux points de vue très proches mais issus de traditions différentes. Ces traditions dictent les orientations des logiciels et des fonctionnalités, le recours à un élagage des mots outils ou à une lemmatisation par exemple, le choix de tel ou tel tableau à soumettre à une analyse factorielle. Ici nous aurons confronté une méthodologie proche d'*Hyperbase* et d'un protocole « niçois » de description plus littéraire à une vision plus proche de *Lexico,* plus socio-politique avec des traitements plus minimalistes, tenant en compte les différents paramètres de la lemmatisation.

Sans s'opposer mais en étant complémentaires au contraire, on constate à travers la terminologie que deux « écoles » se rencontrent : *item, macro et micro-distribution* (Viprey, Besançon), *forme graphique, occurrences* (Salem, Fiala, Heiden) d'autres encore parleraient de *vocables*…les acceptions et les terminologies décrivant ou désignant l'unité minimale sont tout à fait révélatrices des ces origines différentes.

Les initiatives actuelles en matière de mise en commun de méthodologies ont encore du chemin à parcourir mais la nécessité de recourir à la complémentarité est désormais bien reconnue. La réflexion sur la complémentarité des approches, sur les fonctionnalités des logiciels, sur les visualisations et l'ergonomie ont permis le développement d'outils tels que TXM, Iramuteq ou TextObserver qui proposent, d'expérimenter, de fédérer, de mettre en commun ces différentes approches.

Au-delà de la dimension technique et méthodologique et des algorithmes eux-mêmes, les différentes écoles de la lexicométrie sont désormais unies dans l'idée de considérer le texte comme une structure ordonnée ou un espace organisé. Le traitement quantitatif des données avec ses comparaisons et ses classifications automatiques tient aussi compte de la dimension syntagmatique des textes, de la structure linéaire, de la dynamique interne et de la progression du texte. Faute de place nous ne pouvons pas nous intéresser à tous ces aspects qui forment un ensemble cohérent de l'analyse d'un corpus numérisé.

Références bibliographiques

Adam J.-M. Heidemann U. (2006). *Sciences du texte et analyse de discours, Enjeux d'une interdisciplinarité*, Genève, Slatkine Erudition.

Brunet E. (2011). *Hyperbase, Manuel de référence, versions 8.0. et 9.0.*

Heiden S. (2004). *Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex* , 7èmes Journées internationales d'Analyse Statistique des Données Textuelles (JADT 2004), Louvain-la-Neuve.

Heiden S., Lafon P. (1998). *Cooccurrences, La CFDT de 1973 à 1992. Des mots en liberté, Mélanges Maurice Tournier*, ENS Éditions, tome 1, Fontenay-aux-Roses.

Kastberg Sjöblom M., (2006). « Peut-on refuser les genres littéraires ? Etude quantitative d'un corpus informatisé », in Olsen M. et Swiatek E. (éditeurs) *XVIe Congrès des Romanistes Scandinaves*, Roskilde Universitetcenter, http://www.ruc.dk/cuid/publikationer/publikationer/XVI-SRK-Pub/

Kastberg Sjöblom M, (2006). « La sémantique lexicale et les genres : analyse systématique d'un corpus québécois », in Williams G. (éditeur) *Les 4èmes journées de la Linguistique du corpus*, Rennes, Presses universitaires de Rennes.

Kastberg Sjöblom M. (2006). *L'écriture de J.M.G. Le Clézio – des mots aux thèmes*, Paris, Honoré Champion.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion. Paris.

Leblanc, JM (2005). *Les vœux des présidents de la cinquième République (1959-2001). Recherches et expérimentations lexicométriques à propos de l'ethos dans un genre discursif rituel*, Thèse de Doctorat en Sciences du Langage, Université de Paris 12 Val-de-Marne.

Leblanc J.-M., Martinez W. (2005). "Positionnements énonciatifs dans les vœux présidentiels sous la cinquième République", *Corpus*, Numéro 4, Les corpus politiques : objet, méthode et contenu - décembre 2005, [En ligne], mis en ligne le 1 septembre 2006. URL : http://corpus.revues.org/document347.html.

Mayaffre D. (2004). *Paroles de président*, Paris, Honoré Champion.

Martinez, W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3 Paris.

Rastier, F. (1987). *Sens et textualité*, Paris, Hachette.

Reinert Max (2008). "Mondes lexicaux stabilisés et analyse de discours", in Heiden S., Pincemin B., , Actes des Journées internationales d'Analyses des Données Textuelles, ENS, Lyon, p. 981-993.

Salem A. (1987) *Pratique des segments répétés, essai de statistique textuelle*, Paris, Klincksieck.

Viprey J.-M. (2004). Analyse séquenée de la micro-distribution lexicale, in Purnelle G., Fairon C., Dister A. (éds), *JADT, Le poids des mots, Actes des 7e journées internationnelles d'analyse statistiques des données textuelles*, p. 1166-1177.

Viprey J.-M. (2005)."Philologie numérique et herméneutique intégrative" in Adam J.-M., Heidemann U. (2006). *Sciences du texte et analyse de discours, Enjeux d'une interdisciplinarité*, Genève, Slatkine Erudition.

Viprey J.-M. 2012. *Colloque international "La cooccurrence : du fait statistique au fait textuel", Besançon, 8-9-10 février 2012*, actes à paraître.