Cooccurrences, contrastes et caractérisation textuels. Applications à un corpus de professions de foi électorales (1958 – 2007)

Magali Guaresi

Cooccurrences, contrastes et caractérisation textuels. Applications à un corpus de professions de foi électorales (1958 – 2007)

Magali Guaresi

BCL (UMR 7320), Université Nice Sophia Antipolis – France

Abstract

This contribution defines the cooccurrence as a textual unit for statistical text analysis. Through the study of a corpus of statements of principles of women and men candidates under the French Fifth republic (1958-2007), it aims to expose some of the main benefits of using cooccurrences for textual comparisons. As minimal forms of co(n)text and semantic units, cooccurrences are a qualitative leap in the characterization of texts.

Résumé

Cette contribution pose les cooccurrences comme des unités textuelles pour le traitement statistique des textes. Elle vise, à travers des analyses menées sur un corpus de professions de foi électorales rédigées sous la Cinquième République (1958-2007), à exposer quelques-uns des principaux apports de l'application des outils de contrastes textuels à l'unité cooccurrentielle. Forme minimale du co(n)texte et déjà porteuse de sens, la cooccurrence incarne pour les entreprises de caractérisation des textes une avancée qualitative notable.

Mots-clés : cooccurrence, statistique textuelle, statistique différentielle, contextualisation, discours électoral

1. Introduction

Enjeu des débuts de la lexicométrie (Tournier, 1980) réaffirmé avec force ces dernières années comme programme d'avenir de l'analyse de données textuelles (ADT), la statistique cooccurrentielle entend constituer une avancée dans l'appréhension et l'interprétation des textes comme des entités réticulaires organisées.

Le repérage de la cooccurrence, c'est-à-dire de la co-présence matérielle et statistiquement significative de deux items au sein d'une fenêtre textuelle définie, apparaît en effet comme une voie efficace de formalisation de co(n)textes porteurs de sens (Firth, 1957, Halliday et Hasan 1976). *A minima*, elle permet de dépasser l'inévitable ambiguïté du mot seul pour former des paires de mots déjà sémantiques (Mayaffre, 2008-a). Généralisée à l'ensemble des relations des unités textuelles au sein d'un corpus (Viprey, 1997, 2006, Martinez 2012), la statistique co-occurrentielle va jusqu'à révéler la structure sémantique, si ce n'est thématique, des textes qui le composent (Ben Hamed et Mayaffre 2015).

De nombreuses fois éprouvées dans les entreprises de description co(n)textuelle des textes (Heiden et Lafon 1998, Mayaffre 2008-b), la démarche cooccurrentielle demeure sous-employée dans les tentatives de discrimination et de caractérisation des textes entre eux. Dans nos protocoles méthodologiques les plus fréquents, la recherche cooccurrentielle ne vient

souvent que dans un second temps[1]. En dépit des avancées méthodologiques importantes de ces dernières années, dont nous nous attacherons à exposer quelques exemples dans cette contribution, les outils d'exploration, de classification et de caractérisation implémentés dans les logiciels d'ADT demeurent essentiellement appliqués à l'unité seule, atomisée et décontextualisée. Très concrètement, les analyses factorielles des correspondances, les analyses arborées, les calculs de spécificités, etc., utilisés pour rendre compte des textes qui constituent les corpus se fondent, encore majoritairement, sur l'occurrence considérée isolément, hors de son co(n)texte d'emploi. La fonction « Thème », par exemple, implémentée dans Hyperbase permet de calculer tous les mots associés de façon significative à un mot-pivot, repéré au préalable par la statistique occurrentielle. Précieuse pour baliser des parcours interprétatifs en renouant avec le contexte d'apparition d'un terme sans renoncer trop tôt à l'encadrement de la statistique, cette fonctionnalité n'intervient qu'après la première approche élémentaire du mot seul.

« Unité statistique » pour l'analyse du texte, la cooccurrence a également été redéfinie comme une « unité du texte pour la statistique » (Mayaffre, 2014 : 18). C'est dans cette seconde acception qu'elle nous paraît sous-utilisée. En dépit des apports de traitements tels celui des segments répétés (Salem, 1987) ou des motifs (Mellet et Longree, 2012), la cooccurrence ne constitue que rarement une unité de contraste fondamentale entre les textes d'un corpus. Pourtant, si l'on estime que le sens naît de la différence (Hébert, 1996) et de la contextualisation (Rastier, 2001)[2], l'usage de la cooccurrence comme une unité textuelle pertinente pour la statistique différentielle nous paraît doublement justifié.

Dans cette perspective, deux tentatives méthodologiques sont particulièrement abouties. La première se trouve dans les récents développements d'Hyperbase qui autorisent l'élaboration de bases d'analyse exclusivement cooccurrentielles par le programme HYPOCCUR (Brunet, 2012). La seconde est issue de la méthode Reinert (1993) implémentée dans le logiciel libre Iramuteq (Ratinaud et Marchand, 2012). A partir du traitement d'un corpus de professions de foi électorales de député-e-s élu-e-s sous la Cinquième République (1958 – 2007)[3], cette contribution en expose les principales plus-values dans le contraste et la caractérisation des textes d'un corpus.

[1] Certaines analyses, certes encore minoritaires, se distinguent toutefois par le primat accordé à l'approche cooccurrentielle. Par exemple, récemment, (Lauf, Valette et Khouas, 2012).

[2] La démarche logométrique pose en effet que le sens des textes naît du co(n)texte, formalisé de façon maximale par le corpus. C'est en fonction de la norme qu'il incarne que les variations constatées entres les différents textes qui le composent prennent sens.

[3] Le corpus d'étude rassemble la quasi-totalité des professions de foi d'élues députées entre 1958 et 2007 (à l'exception de celles du scrutin de 1986, exceptionnellement pour la Cinquième République tenues au scrutin de liste) et un corpus de comparaison de textes d'hommes rédigés dans des conditions politiques, géographiques et chronologiques comparables (700 textes). Pour cette contribution, le corpus de travail oscille entre une version restreinte (à chaque proclamation de femme répond une unique profession de foi d'homme ; 500 000 mots) et une version élargie à l'ensemble des professions de foi des député-e-s des départements ayant au moins une femme élue (1 200 000 occurrences).

2. Spécificités cooccurrentielles et caractérisation du discours électoral féminin (1958 – 2007)

2.1. Les trois temps de la parole féminine

Les outils de la textométrie tentent depuis ces dernières années de proposer des fonctionnalités susceptibles de dialoguer avec les propositions théoriques sur le texte et les corpus. Hyperbase, par exemple, permet depuis 2014 de traiter de corpus ramenés à leurs seules paires cooccurrentielles significatives, préalablement repérées. Tous les autres mots sont ignorés par le traitement ; seuls les binômes cooccurrentiels sont traités dans l'ordre de leur apparition dans le texte originel. Définie comme une occurrence du texte, comme une unité textuelle en tant que telle, la cooccurrence peut dès lors être soumise aux calculs historiques de la lexicométrie, et en particulier aux fonctionnalités de caractérisation.

Dans le cadre de notre thèse, nous avons eu recours au traditionnel calcul des spécificités, décomptées sur les occurrences (les mots seuls), pour décrire la parole électorale des femmes au fil d'un demi-siècle de scrutins législatifs (Guaresi, 2015). Il est aujourd'hui possible de prolonger ces analyses en appliquant le calcul des spécificités sur les paires de mots (plus précisément ici, les couples de substantifs). Sur le corpus de professions de foi restreint, divisé en six textes – un par sexe et par période historique (1958-1973, 1978-1988, 1993-2007)[4] –, on peut ainsi obtenir une liste de cooccurrences spécifiquement sur-utilisées par chaque groupe de locutrices et locuteurs.

Femmes (1958-1973)	Hommes (1958-1973)	Femmes (1978-1988)	Hommes (1978-1988)	Femmes (1993-2007)	Hommes (1993-2007)
Gaulle_paix (5,4)	république_cinquième (8,6)	français_parti (7)	gauche_programme (5,6)	répartition_retraite (3,7)	élan_France (4,4)
construction_ logement (5,1)	candidat_parti (5,2)	gouvernement_ ministre (6,9)	changement_union (5,3)	environnement_ qualité (3,6)	croissance_pouvoir (3,7)
expansion_vie (5)	candidat_programme (5,6)	changement_union (6,9)	liberté_société (5,6)	éducation_protection (3,3)	énergie_recherche (3,5)
femme_promotion (4,3)	France_Gaulle (4,8)	changement_parti (6,2)	franc_smic (4,9)	droit_enfant (3,1)	assemblée_terrain (3,4)
crédit_million (4,1)	pouvoir_régime (4,7)	justice_liberté (5,2)	changement_majorité (4,4)	engagement_femme (2,9)	immigration_ sécurité (3)

Figure 1 : Spécificités cooccurrentielles des femmes et des hommes selon les périodes (1958-1973, 1978-1988 et 1993-2007), exprimées en écarts-réduits.

A la simple lecture de ce tableau, il est possible de cerner l'originalité thématique des scrutins qui se succèdent. De façon déjà minimalement sémantique, la liste des duos lexicaux résume les principaux enjeux des engagements à la députation des élu-e-s. Là où l'occurrence seule est informative, la paire spécifique laisse déjà connaître les mutations chronologiques fines du

[4] Pour des raisons de lisibilité, nous utilisons ici une chronologie endogène du corpus issue des traitements exploratoires et classificatoires menés dans la thèse, qui permettent la caractérisation de trois grands ensembles (plutôt que des douze scrutins individuellement).

discours mais également les modalités de la construction genrée[5] des professions de foi à une époque donnée.

Du point de vue de l'évolution générale des professions de foi au fil de la République, la distribution des spécificités cooccurrentielles indique le passage d'un discours programmatique centré sur les débats politiques nationaux, l'affrontement partisan voire idéologique à une promesse de représentation organisée sur des thématiques sociétales ou économiques (orientées sur le local, le « terrain »).

Plus intéressant au regard de notre problématique sur le genre, le tableau suggère les reconfigurations des identités et engagements « féminins » légitimes, à l'œuvre dans les campagnes électorales. Si les premières élues font référence au père du régime quinto-républicain naissant, le général De Gaulle, c'est avant tout pour souligner son action en faveur de la paix en Algérie. Elles entrent moins que leurs homologues masculins dans les considérations institutionnelles que suscite l'établissement de la république (« pouvoir_régime »), pour s'en tenir à un discours plus axé sur des propositions de politiques publiques et d'utilisation des fruits de l'expansion (« construction_logement »). Par exemple, la députée de Seine-et-Oise s'engage à :

« Concevoir une politique du district plus rationnelle et plus humaine qui en finisse avec le paradoxe consistant à laisser s'implanter des centaines de milliers de nouveaux habitants sans avoir prévu l'infrastructure d'ordre matériel, culturel et social qui doit précéder ou au moins accompagner toute construction de logement (Thome-Patenôtre, 1967, FGDS, Seine-et-Oise).

Primo-élues de la République, les femmes des cinq premières législatures rejettent le jeu politique pour construire un engagement conforme aux représentations traditionnelles de l'activité publique féminine, pacifique, consensuelle et pragmatique. Toutefois, leur présentation de soi témoigne d'une implication en faveur d'une amélioration de la condition des femmes dont elles portent les intérêts de façon substantielle (« femme_promotion »). La seconde période, marquée par une forte politisation des débats et la victoire de la gauche, impose un vocabulaire politique qui neutralise le poids du genre dans les proclamations électorales. Durant cette décennie, les femmes accèdent à un discours de militantes plus assumé (« français_parti », « changement_union »). Le déclin de l'affrontement partisan à compter de la fin des années 1980 s'accompagne, en revanche, d'une re-sexuation des thèmes des candidatures. Aux femmes, l' « éducation_protection », l' « environnement_qualité » et les enfants ; aux hommes, les visions politiques (« élan_France »), les questions macro-économiques (« croissance_pouvoir ») ou les sujets régaliens (« immigration_sécurité »). De façon significative, les candidates sous-utilisent le discours sécuritaire qui domine les tribunes à la fin du XX° siècle et que ces quelques exemples, dans les professions de foi masculines, illustrent :

[5] Nous entendons le genre comme « un système de bi-catégorisation hiérarchisée des sexes et des valeurs et des représentations qui y sont associés » (Achin et Bereni, 2013). Nous postulons que le discours (politique) en est l'un de ses agents et de ses acteurs privilégiés.

« Nous devons choisir la voie du courage et de la raison, continuer de rénover l'Etat, […] renforcer la sécurité, mener une politique volontariste d'intégration sans complaisance vis-à-vis de l'immigration clandestine » (Barre, 1997, UDF, Rhône).

« Dès le 1er tour, voter Jacques Masdeu-Arus, c'est pour la France, pour notre circonscription, poursuivre une politique courageuse en matière d'immigration clandestine et de sécurité des biens et des personnes » (Masdeu-Arus, 1997, RPR, Yvelines).

La plus-value interprétative de telles listes, minimalement sémantiques, est nette. La distribution des paires cooccurrentielles révèle le monopole masculin sur le politique. Bien qu'ils changent de nature en cinquante ans, les lieux du pouvoir, signifiés par les paires cooccurrentielles, se concentrent dans les textes des hommes. Au début de la République, l'importante question institutionnelle est surreprésentée dans les textes masculins. Mais, encore ces dernières décennies, à l'heure où les alternatives politico-idéologiques se tarissent au profit d'un primat de l'économie, ce sont les hommes qui s'emparent et sur-utilisent les sujets relatifs à la croissance, à l'énergie, etc. pour bâtir leurs déclarations électorales. A l'heure où la méfiance à l'égard du personnel politique est de plus en plus criante, ce sont encore les hommes qui monopolisent le thème de l'immigration dans un sens sécuritaire pour ré-affirmer leur *ethos* et leur autorité. Les élues, quant à elles, embrassent des contenus, moins prestigieux dans le *cursus honorum* politique, qui flattent leurs supposées expertises « naturelles » développées dans la sphère familiale ou privée. En témoignent par exemple, ces extraits composés de la paire cooccurrentielle spécifique « engagement_femme » :

« Dans mon engagement au service de ces idées, je mettrai toute ma détermination d'Aveyronnaise, ma sensibilité de femme, mon attachement à la France, à Paris, à notre 14° arrondissement » (Catala, 1993, RPR, Paris).

« Mon engagement politique est celui d'une femme, d'une mère qui pense que nous devons à nos enfants un pays stable socialement, fort économiquement, dans un monde de paix » (Roig, 1993, RPR, Vaucluse).

2.2. *Thématisation partisane : les cooccurrences de « femme » et de « famille » dans le discours de gauche et de droite*

La plus-value de la statistique cooccurrentielle dans la caractérisation des textes est encore plus évidente lorsqu'il s'agit d'étudier la sémantisation d'un même mot sous la plume de groupes de locutrices ou de locuteurs différents, que l'on confronte termes à termes.

Par exemple, l'étude de la distribution des paires cooccurrentielles formées autour des mots-pôles « femme » ou « famille », très redondants dans les tracts des femmes, permet de rendre compte de la variation partisane et idéologique de discours articulés aux mêmes lemmes dans les textes des élues de gauche d'une part et de droite d'autre part.

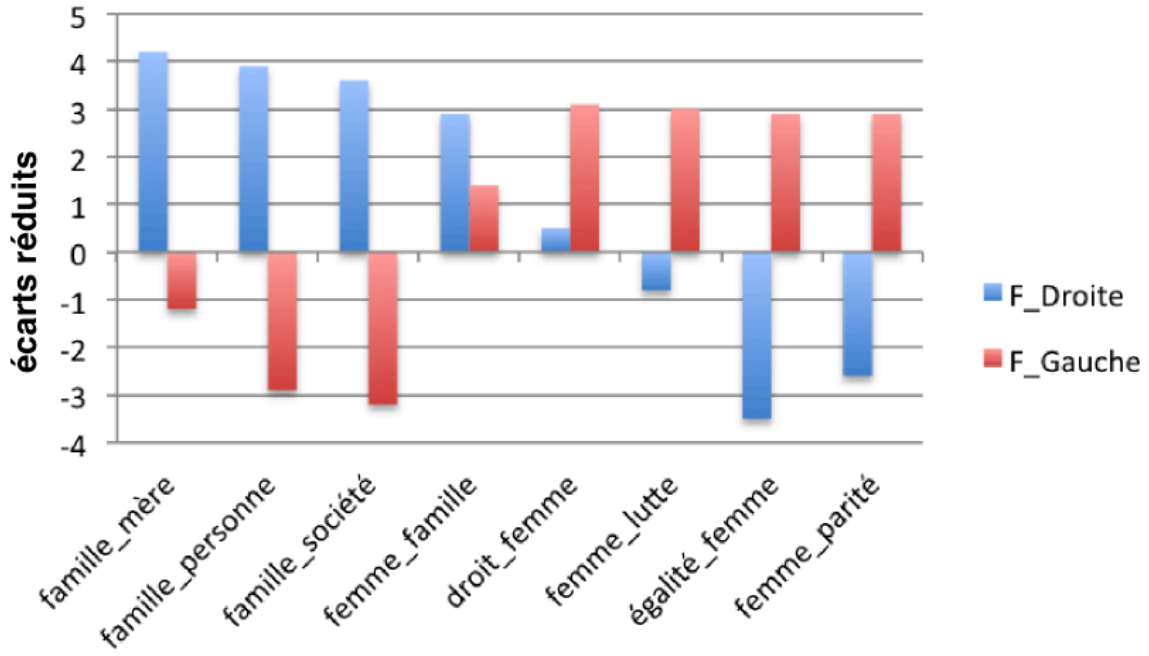

Figure 2 : Distribution des paires cooccurrentielles autour de « femme » et de « famille » chez les députées de gauche versus les députées de droite (1958 – 2007), exprimées en écarts-réduits.

Considérés de façon isolée, les vocables « femme » et « famille » ne permettent pas de rendre compte d'une ligne de clivage qui traverse pourtant le discours féminin aux législatives[6]. Appréhendés dans leurs contextes d'emploi minimaux, les mots révèlent, en revanche, deux types de traitement des questions de genre par les candidates. Les progressistes sur-emploient de façon caractéristique des unités cooccurrentielles militantes en faveur du combat pour l'égalité et de l'obtention de nouveaux droits pour les femmes. Les candidates des partis de l'ordre, pour leur part, assoient et tirent leur légitimité du développement de la thématique familiale dans une perspective relativement conservatrice. Ce sont moins les femmes qui sont au cœur de leurs prises de parole que la famille et c'est moins une critique de la famille héritée du modèle patriarcal qui domine leurs projets qu'une réaffirmation de la division genrée des rôles parentaux. En témoignent trois extraits composés des paires cooccurrentielles typiques du corpus des femmes de droite :

« Revaloriser les allocations pour permettre à la mère de famille de rester à son foyer » (Martinache, 1958, UNR, Nord).

« Protéger la famille en donnant aux mères le libre choix entre une vie professionnelle et l'éducation de leurs enfants » (Boutin, 1993, UDF, Yvelines).

« La famille est le lieu des apprentissages essentiels de la vie en société : le respect de l'autre, la prise de conscience des responsabilités, l'éducation à la citoyenneté ; autant de valeurs qu'il faut vite restaurer. Plus que jamais, notre société a besoin des familles » (Ramonet, 2002, UMP, Finistère).

[6] Nous considérons ici les lemmes, c'est-à-dire l'ensemble des emplois des mots au singulier et au pluriel.

A travers ces quelques exemples, c'est le genre comme objet politique, comme vision du monde, qui est aperçu. Les élues sont, certes, globalement soumises au paradoxe des minoritaires, c'est-à-dire à l'exigence de traiter des problèmes spécifiques des femmes pour prétendre à la représentation de la nation. Mais, leur originalité politique se manifeste dans les combinaisons préférentielles qu'elles effectuent autour des mots indépassables de la candidature féminine.

3. Cooccurrences généralisées, exploration non-supervisée et description du discours électoral

3.1. *Les mondes lexicaux des professions de foi*

Unité complexe utile pour discriminer des sous-parties de corpus, la cooccurrence peut également être utilisée à des fins exploratoires, susceptibles de faire émerger des contrastes significatifs depuis le corpus non préalablement partitionné. Dans sa dimension généralisée (Viprey, 1997, 2006), le traitement de la cooccurrence permet d'établir les structures sémantiques (sous-jacentes) des corpus. Si l'on veut bien admettre que le sens d'un mot s'approxime à partir de l'étude de ses contextes d'emploi et donc de ses associés privilégiés, alors on comprendra que la formalisation des profils cooccurrentiels de plusieurs centaines de mots laissent apparaître la trame sémantique des corpus.

Concrètement, le logiciel toulousain Iramuteq établit, par une classification hiérarchique descendante, des classes de vocabulaire, aussi stables et homogènes que possible, rendant compte des « mondes lexicaux » (Reinert, 1993) des corpus. L'analyse initiée par M. Reinert et implémentée dans Iramuteq (Ratinaud et Marchand, 2012) repose sur une série de bipartitions construites sur la base d'un tableau croisant des segments textuels – ici 40 items – et les mots sélectionnés pour le calcul – ici les substantifs, les verbes, les adjectifs et les pronoms personnels lemmatisés – dont on décompte la rencontre. Une classe de vocabulaire est formée et distinguée des autres lorsque l'inertie interclasse la plus importante est atteinte, c'est-à-dire lorsque la « meilleure partition » est effectuée. Ainsi, dans la combinaison d'une approche fréquentielle et séquentielle, paradigmatique et syntagmatique, la cooccurrence généralisée permet l'émergence inductive de polarisations lexicales informant l'architecture d'un corpus.

Appliquée au corpus de professions de foi électorales élargi, la classification hiérarchique descendante discrimine 6 classes sur la base de l'analyse cooccurrentielle de 3000 formes. La projection factorielle des principaux mots composant les différents mondes lexicaux est donnée ci-dessous (Figure 3).

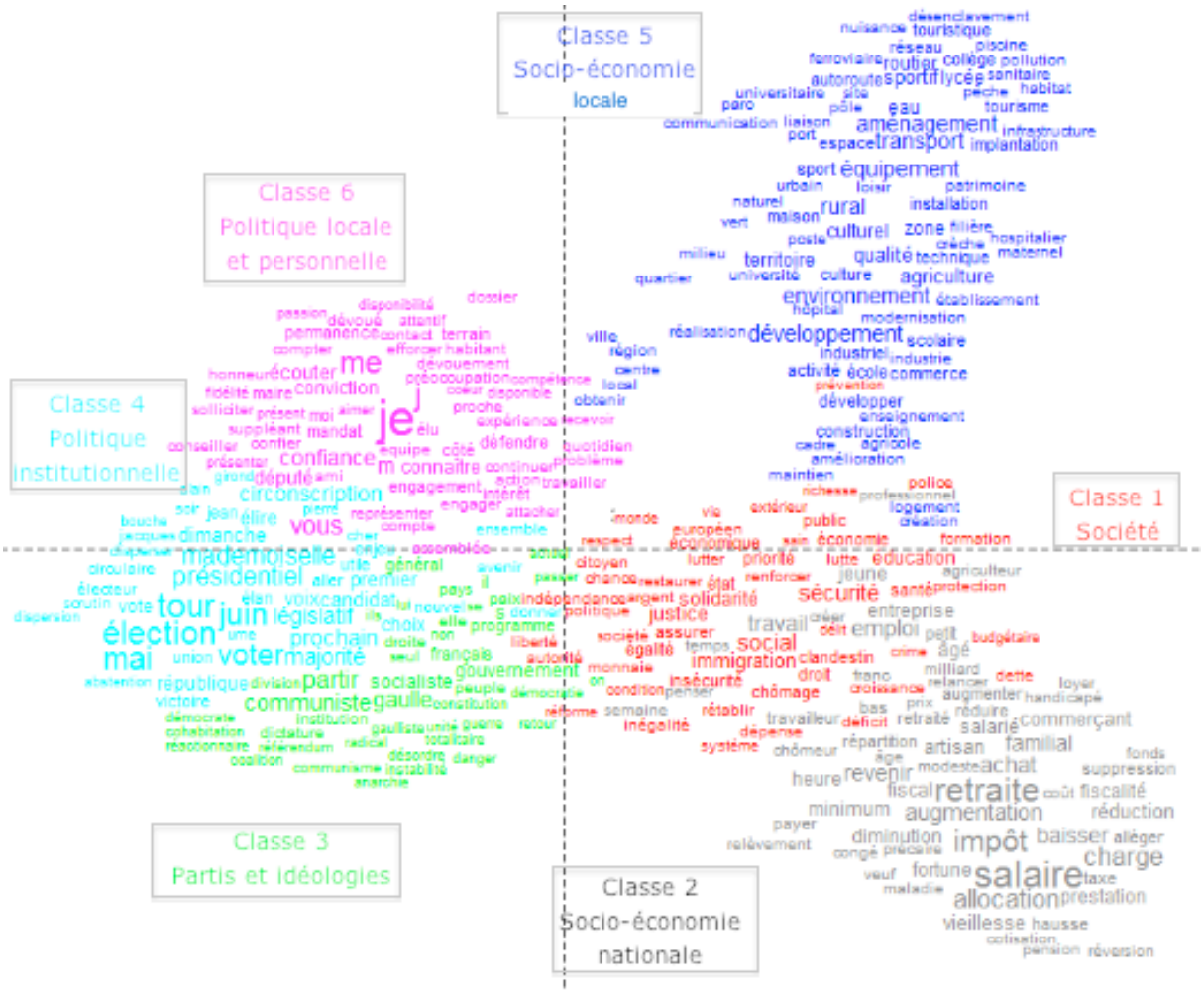

Figure 3 : Projection factorielle des cinq classes du corpus Professions de foi 1958 – 2007 (Facteur 1 : 40,1% - Facteur 2 : 20,71)

Le premier apport de ce type d'analyse est de rendre compte des pôles isotropiques structurants de la parole électorale quinto-républicaine. Celle-ci s'organise, en synchronie, en quatre pôles dominants (pour s'en tenir à une lecture de la figure par quadrant) qui opposent le long de l'axe horizontal un vocabulaire politico-institutionnel à gauche (« élection », « législatif », « république » à des mots de la socio-économie à droite (« entreprise », « formation », « budgétaire ») ; et le long de l'axe vertical le vocabulaire national (en bas) aux termes plus locaux ou territoriaux (en haut). Sans revenir en détail sur la richesse de ces classes de vocabulaire pour établir les thèmes fondamentaux du discours aux législatives (Ben Hamed et Mayaffre 2015, Guaresi 2015), insistons davantage ici sur le potentiel heuristique de cette figure dans la caractérisation textuelle.

3.2. *Caractérisation endogène : distribution des scrutins selon les classes de vocabulaire*

Ces groupements lexicaux formés par le repérage des fréquentations des mots entre eux, au sein de segments textuels de 40 items, dressent un panorama des saillances sémantiques du

corpus à partir duquel il est possible, de façon endogène ou inductive, de faire émerger des contrastes pour discriminer et décrire les textes qui le composent.

Soit, par exemple, le corpus des professions de foi étiqueté selon les douze années électorales[7]. Alors que la partition chronologique du corpus n'a pas présidé à la construction des classes, projetées sur le plan factoriel (Figure 3), il est toutefois possible de la réintroduire, dans un second temps, dans l'analyse. Il s'agit alors d'observer et de quantifier la participation de chacun des textes chronologiques à chacun des mondes lexicaux préalablement définis, de façon non supervisée.

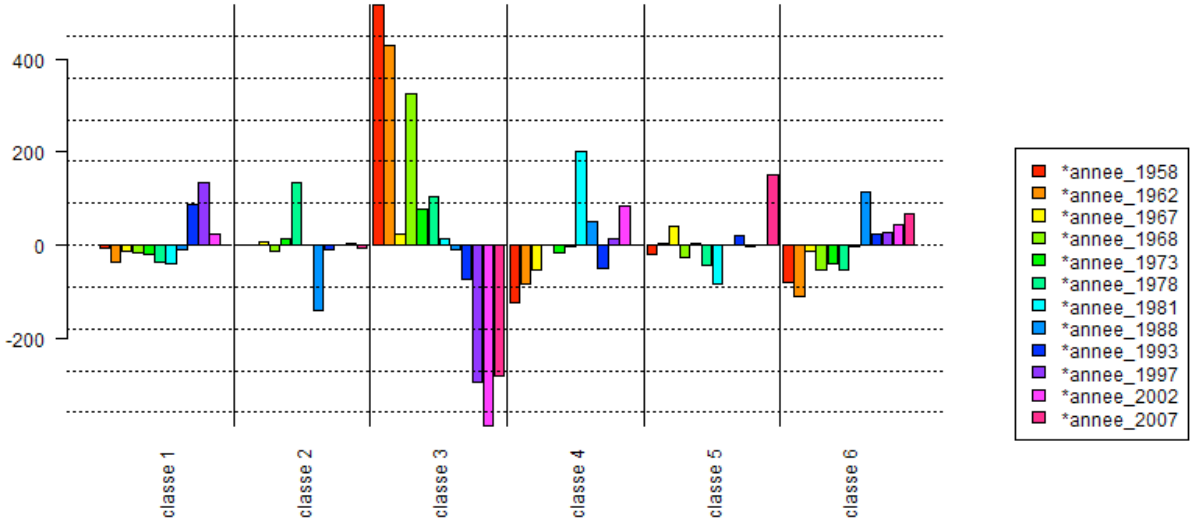

Figure 4 : Chi2 des modalités de la variable « année d'élection » selon les classes de vocabulaires du corpus Professions de foi 1958 – 2007 (sortie Iramuteq – scores donnés en Chi2)

Ce graphique représente la force (indiquée par un Chi2) des relations entre une modalité de la variable chronologique (i.e. une date d'élection) et chaque monde lexical. Apparaissent ainsi des clivages et des contrastes dans un mouvement *buttom-up*, sur la seule base de l'organisation cooccurrentielle du corpus, hors de l'intervention de l'analyste et de partitions pré-imposées pour la comparaison. Outre la plus-value heuristique, la démarche incarne une avancée qualitative. En effet, les particularités discursives des candidatures au fil du temps s'apprécient à l'aune de classes de vocabulaire significatives, du point de vue quantitatif, mais également déjà sémantiques.

Sans prétendre épuiser la richesse d'un tel graphique, nous nous contenterons de formuler deux interprétations principales et heuristiques pour l'historien-ne du discours politique. La distribution des textes en fonction des mondes lexicaux semble répondre à deux grandes logiques. La première est d'ordre chronologique. Les classes 1, 3 et 6 témoignent d'une évolution temporelle presque parfaitement linéaire dans le corpus. Ainsi, le plus fort contraste chronologique s'opère dans la classe 4, au niveau du vocabulaire idéologico-partisan (« parti », « communiste », « constitution », « démocratique »). Les scrutins précédant l'année 1988 se qualifient par un sur-investissement de la thématique alors que tous les suivants

[7] Nous considérons ici l'ensemble des textes de femmes et d'hommes produits à une même date. Le clivage sexué n'est pas analysé dans le développement qui suit.

délaissent toujours plus ces enjeux dans les débats électoraux. *A contrario*, les thématiques sociétales de la classe 1 (telles l' « immigration », la « santé », l' « éducation ») ne s'imposent qu'à partir des élections législatives de 1993, 1997 et 2002 alors qu'elle sous-composent les engagements des élu-e-s antérieur-e-s. De la même manière, les textes rédigés à partir de 1988 se spécifient dans la classe 6 par les mises en scène personnelles et locales, en quantités inédites, des actions des futur-e-s élu-e-s (« je », « permanence », « disponibilité »).

La seconde lecture du graphique tend moins à rendre compte d'une évolution linéaire qu'à décrire des types de scrutins ayant ponctuellement recours aux mêmes spécificités isotropiques. C'est le cas des élections de 1981, 1988 et 2002 par exemple, qui sont surreprésentés dans la classe 4, consacrée au vocabulaire politico-institutionnel national (« élection », « voter », « présidentiel »). Scrutins tenus après des élections présidentielles notables (les deux élections mitterrandiennes et le choc du 21 avril 2002 avec la présence du Front National au second tour), ils sont fortement imprégnés des enjeux et des argumentaires nationaux. La bataille électorale de 2007, en revanche, s'originalise en inscrivant particulièrement ses programmes dans une dimension micro-économique territorialisée et infrastructurelle (dans la classe 5 : « équipement », « aménagement », « territoire », « école ») et témoignent d'un scrutin axé sur des enjeux majoritairement localistes.

Par ce protocole de recherche, la description du discours électoral aux législatives est affinée et chevillée à des indices statistiques et sémantiques solides. En effet, les principales lignes de fracture (chronologiques ou historico-politiques) de la parole aux législatives s'animent sur une toile de fond constituant l'organisation forte et significative du corpus de professions de foi. La caractérisation des textes et de leurs contenus s'opère ainsi, dès le premier mouvement de recherche, sur des éléments déjà porteurs de sens.

4. Conclusion

En ADT, l'atomisation du texte « est une heuristique nécessaire pour déterminer certaines propriétés lexicales et structurelles du corpus » (Martinez, 2012 : 213). Pourtant l'exigence de contextualisation pour l'émergence du sens est affirmée depuis les débuts de la lexicométrie. Point de rencontre entre ces deux présupposés, *a priori* contradictoires, la cooccurrence prolonge le dialogue entre les approches paradigmatiques et syntagmatiques des textes.

Admettre la paire coocurrentielle comme une unité textuelle à part entière multiplie les perspectives d'exploration et d'interprétation des corpus. Dans la démarche contrastive en particulier, la cooccurrence approfondit la statistique différentielle ; celle-ci ne se contentant plus de traiter de mots atomisés mais de noyaux déjà minimalement sémantiques pour comparer les textes. La compréhension de la prose féminine aux législatives a ainsi pu être précisée par l'étude de la sémantisation différenciée des mots spécifiques des élues, en fonction de leur bord politique.

Dans sa dimension généralisée, la cooccurrence renouvèle les protocoles d'exploration des corpus en formalisant leurs saillances cooccurrentielles (isotropiques voire thématiques) structurantes. Sur cette base, elle autorise les contrastes entre les textes selon des parcours interprétatifs non supervisés et sur des critères déjà qualitatifs tout en ne renonçant pas à l'exigence de représentativité statistique. Les puissants parcours de lecture cooccurrentielle ont ainsi permis, dans cette contribution, de décrire les singularités du parler électoral au fil de douze scrutins sur la base de groupements lexicaux, non pas anecdotiques, mais déjà significatifs en quantité et en qualité.

Références

Achin C. et Bereni L. (2013). *Dictionnaire Genre et Science politique*. Paris : Presses de sciences po.

Ben Hamed M. et Mayaffre D. (2015). Thèmes et thématiques du discours politique. Du concept à la méthode. *Mots. Les langages du politique*, 108 : 5-13.

Brunet E. (2012). Nouveau traitement des cooccurrences dans Hyperbase. Corpus, 11 : 219-248.

Firth R. (1957). *Papers in Linguistics 1934-1951*. London : Oxford University Press.

Guaresi M (2015). *Parler au féminin. Les professions de foi des député-e-s sous la Cinquième République (1958-2007)*. Thèse de doctorat, Université Nice Sophia Antipolis.

Guaresi M. (2015). Les thèmes dans le discours électoral de candidature sous la Cinquième République : perspective de genre (1958-2007). *Mots. Les langages du politique*, 108 : 15-37.

Halliday M.A.K. et Hasan (R.). *Cohesion in English*. London : Longman.

Hébert L. (1996). Une sémantique différentielle unifiée. *RS/SI, Association canadienne de sémiotique*, 1-2 : 275-285.

Heiden S. et Lafon P. Cooccurrences. La CFDT de 1973 à 1992. In *Des mots en liberté, Mélanges Maurice Tournier*, Paris : ENS Lyon, T1, pages 65-83.

Lauf A., Valette M. et Khouas L. (2012). Analyse du graphe des cooccurrents de deuxième ordre pour la classification non supervisée de documents. In Dister A., Longrée D., Purnelle G., editors, *JADT 2012*, pages 525-535.

Martinez W. (2012). Au-delà de la cooccurrence binaire… Poly-cooccurrences et trames de coocurrence. *Corpus*, 11 : 191-216.

Mayaffre D. (2008-a). De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. *Sémantique et Syntaxe*, 9 : 53-72.

Mayaffre D. (2008-b). Quand travail, famille, patrie co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence. In Heiden S. et Pincemin B. editors, *JADT 2008*, pages 811-822.

Mayaffre D. (2014). Plaidoyer en faveur de l'analyse de donnée co(n)textuelles. Parcours coocurrentiels dans le discours présidentiel français (1958-2014). In Née E., Valette M., Daube J.-M. , Fleury S., editors, *JADT 2014*, pages 5-32.

Mellet S. et Longrée D. (2012). Légitimité d'une unité textométrique : le motif. In Dister A., Longrée D., Purnelle G., editors, *JADT 2012*, pages 715-728.

Rastier F. (2001). *Arts et sciences du texte*. Paris : Presses universitaires de France.

Ratinaud P. et Marchand P. (2012).Application de la méthode ALCESTE aux "gros" corpus et stabilité des mondes lexicaux : analyse du CableGate avec Iramuteq. In Dister A., Longrée D., Purnelle G. , editors, *JADT 2012*, pages 835-844.

Reinert M. (1993). Les mondes lexicaux et leur logique à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 66 : 5-39.

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

Tournier M. (1980). En souvenir de Lagado. *Mots*, 1 : 5-9.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*. Paris : Champion.

Viprey J.-M. (2006). Structure non-séquentielle des textes. *Langages*, 163 : 71-85.