Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014)

Damon Mayaffre

CNRS-Université de Nice Sophia Antipolis (UMR 7320, Bases, Corpus, Langage)

Abstract

This paper recalls the double constraint of text mining: tokenizing and co(n)textualizing. It also treats co-occurrences as (complex) text *units*. For this reason - as they are textual molecules and moreover also semantic molecules - text mining tools can be applied to them: such as for example the historic calculation of Specifics, especially in a contrastive approach to corpora. Because a co-occurring pair, unlike a single word, is rarely semantically ambiguous, constituting as it does the minimal form of co(n)text, it allows a better characterization of texts and their contents.

Résumé

Cette contribution rappelle la double contrainte de l'ADT : segmenter et co(n)textualiser. Elle pose par ailleurs les cooccurrences comme des *unités* (complexes) du texte. A ce titre – comme molécules textuelles et plus loin comme molécules sémantiques – les outils de l'ADT peuvent leur être appliqués comme par exemple l'historique calcul des *spécificités*, notamment dans une approche contrastive des corpus. Parce que la paire cooccurrentielle est sémantiquement rarement ambiguë et qu'elle constitue la forme minimale du co(n)texte, elle permet une meilleure caractérisation des textes et de leurs contenus.

Mots-clés : ADT, occurrences, cooccurrences, spécificités, statistiques textuelles, discours politique

1. Introduction

L'Analyse des Données Textuelles tient, à nos yeux, dans la difficulté d'un paradoxe ou la richesse d'une complémentarité : elle précède par atomisation et procède par co(n)textualisation. C'est là sa vertu et sa raison d'être qui la distinguent, dans l'inconfort heuristique d'un entre-deux fertile, à sa gauche des arts du texte de tradition « rhétorico-herméneutique » et à sa droite du TAL de tradition « logico-grammaticale »[1]. En effet, le pari épistémologique des chercheurs en ADT a toujours été, nous semble-t-il, de concilier l'exigence logique, technique ou computationnelle de *segmentation* qui précède bien au quotidien tout traitement et toute analyse (*tokenisation*, indexation, dictionnaire fréquentiel, matrice de données, etc.), et relève d'une non-linguistique nucléaire, et l'exigence herméneutique de *co(n)textualisation* sans laquelle on ne saurait, de notre point de vue, aborder ni la langue ni le texte, ni le sens ni l'interprétation.

Cette contribution entend au fond discuter l'idée même de « données textuelles » qui, si l'on n'y prend garde, peut être ressentie comme une réduction : un texte n'est pas seulement fait de données mais aussi de parcours de lecture, et une somme de données ne saurait à elle seule constituer un texte. Partant de *données* donc, de *jetons*, de *boules*, de *segments* ou *d'items*,

[1] Nous reprenons ici le distinguo fondamental de (Rastier, 2001).

l'enjeu de l'ADT est ainsi, aussi, de pouvoir les textualiser ou co(n)textualiser par des traitements statistiques ou parcours de lecture outillés qui rendent aussi bien compte des parties que du tout.

Loin d'abandonner l'acte interprétatif par co(n)textualisation à la seule subjectivité de l'analyste, qui finit toujours en ADT par *retourner au texte* (*i.e.* produire une lecture naturelle), nous pointerons quelques outils statistiques de co(n)textualisation, historiques dans la discipline, tels les segments répétés ou les motifs, pour espérer un programme de recherche, largement commencé par les pionniers de l'ADT mais jamais totalement abouti : une statistique pas seulement textuelle mais co(n)textuelle, c'est-à-dire une approche statistique du texte qui ne serait pas seulement atomique, élémentaire, discrétisante ou occurrentielle (le mot seul) mais moléculaire, compositionnelle, relationnelle ou cooccurrentielle (l'association de mots ; minimalement la paire de mots). D'un point de vue épistémologique, il ne s'agit là finalement que de bien comprendre une approche analytique du texte (segmentation ou *tokenisation*) qui ne perdrait pas de vue sa finalité herméneutique (interprétation par (re)-co(n)textualisation). L'exigence cooccurrentielle au cœur de cette contribution – définir la cooccurrence (la paire) comme *unité* essentielle du texte – sera appliquée au corpus présidentiel français (1958-2014) que nous avons par ailleurs largement décrit d'un point de vue occurrentiel ces dernières années (Mayaffre, 2012-a).

2. Des données au texte. Statut de la cooccurrence

S'il apparait difficile de dater le présupposé « co(n)textualiste » de l'ADT, affirmons que celui-ci épouse la discipline à partir du moment où elle a bien voulu identifier et respecter son objet : le texte. Non pas seulement la Langue en ses échantillons de laboratoire, mais le texte comme forme empirique d'un discours effectivement émis par un locuteur d'os et de chair, comme forme aboutie et complexe d'un langage actualisé par un énonciateur. Du contextualisme de l'école anglo-saxonne dès les années 1930-40-50 (Palmer au départ, Firth, Halliday, Sinclair) au positionnement rhétorico-herméneutique de Rastier dans *La Mesure et le grain* en 2011 ou celui d'Adam, ici même, aux JADTs en 2006, les théoriciens, les praticiens comme les compagnons de route de l'ADT posent que le sens naît du co(n)texte, que le tout subsume les parties ou que le global (le texte dans son entier voire le corpus) détermine le local (ses unités). Au plus profond de ce consensus disciplinaire fondamental se trouve la posture des chercheurs en ADT et plus généralement en Linguistique de corpus : vouloir travailler sur des données attestées (et non des exemples forgés), c'est pressentir que l'usage – c'est-à-dire le co(n)texte d'utilisation – prime sur le codage, que le sens n'est pas fait de référence (en dictionnaire) mais de différences (dans les corpus textuels) ; que l'item, au fond, n'est jamais seul.

En 1980, Maurice Tournier dans l'éditorial fondateur de la revue majeure en France pour la discipline, *MOTS*, pointait quelques principes élémentaires qui étaient devenus intangibles depuis déjà plusieurs années en ADT et au laboratoire de Saint-Cloud. Le premier principe, précisément, consistait à ne pas accorder précipitamment un contenu à un signe sans observer de manière systématique ses co(n)textes réels d'utilisation (Tournier, 1980: 5-6). Le deuxième principe était quant à lui plus précis pour devenir définitif :

> *La seconde règle de prudence consiste à n'en pas rester au vocable isolé. Le mot : cet acteur de sens que seuls d'autres mots peuvent actionner dans un sens. Tout, dans l'énonciation, est séquence, réseau, co-occurrences,*

> *équilibre ou fuites entre formes, scansions d'emplois affrontées* (Tournier, 1980:6)

Naissante, la lexicométrie occurrencielle (le mot isolé) s'affichait ainsi explicitement, déjà, en lexicométrie cooccurrentielle : c'est ce pressentiment, souvent travaillé, parfois oublié, qui a constitué un des agendas mi-séculaires de la discipline, et l'addenda que cette contribution propose. Les fameuses affirmations, théoriques mais non implémentées, de Firth au niveau du mot ("*You shall know a word by the company it keeps"* (Firth, 1957:11)) puis de Halliday et Hasan au niveau du texte ("…*cohesion that is achieved through the association of lexical items that regularly co-occur*…" (Halliday et Hasan, 1976:284)) furent et restent un programme de recherche pour les méthodologistes en ADT. Au-delà de la linguistique du texte – c'est-à-dire une linguistique de l'usage, une linguistique du co(n)texte –, ces prises de position se revendiquent aujourd'hui de Saussure même :

> *Avant tout on ne doit pas se départir de ce principe que la valeur d'une forme est tout entière dans le texte où on la puise, c'est-à-dire dans l'ensemble des circonstances morphologiques, phonétiques, orthographiques qui l'entourent et l'éclairent.*[2]

L'ensemble des circonstances qui entourent et éclairent le mot, dont parle Saussure, pourrait se définir, strictement, comme l'ensemble de ses *cooccurrences*.

3. Définition de la cooccurrence

Les travaux sur la cooccurrence sont depuis 50 ans nombreux et importants dans notre communauté. Et il est regrettable que la vivacité de cette production scientifique se trouve opacifiée par des différends terminologiques (qui cachent des différends plus fondamentaux), notamment entre le monde anglo-saxon et la France.

Par exemple, le terme concurrent de *collocation*, que l'on trouve majoritairement employé dans les travaux en langue anglaise depuis (Palmer, 1933) jusqu'à (Sinclair, 2003) semble désigner, le plus souvent, l'association syntagmatique de plusieurs mots contigus. La collocation est ainsi susceptible de pressentir des habitudes phraséologiques ou des raideurs idiomatiques (« the idiom principle » (Sinclair, 1991:115)) peut-être annonciatrices de figements en langue (pour le français : *pluie battante, passion dévorante, voix suave*, *battre en retraite*, etc). L'outil d'extraction de la collocation est le plus souvent le concordancier très répandu au-delà de l'hexagone (*Key Word in Context* - KWIC), lorsque le traitement statistique n'est utilisé qu'à grands traits (par exemple le *T-score* chez Sinclair 1991). L'intérêt des collocations est évident par exemple dans le cadre de la traduction automatique[3].

Le terme de *cooccurrence*, que l'on trouve majoritairement utilisé en France dans le domaine de l'ADT, désigne quant à lui le plus souvent des associations statistiques (critère de fréquence) dont on constate les relations textuelles (plus que phrastiques), sans distinction de

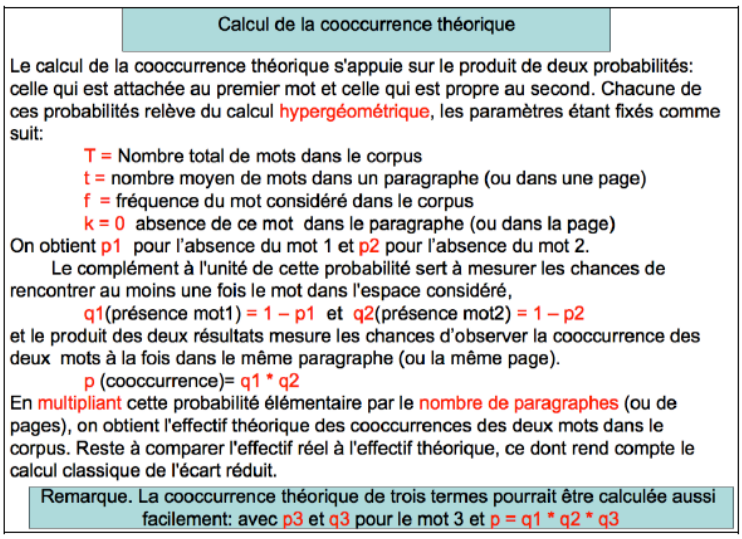
[2] Saussure [1894] 1922, « Sur le nominatif pluriel et le génitif singulier de la déclinaison consonantique en lituanien » dans Charles Bally et Albert Séchehaye (éd), Paris Payot, p. 514 (cité par J. P. Bronckart, « Les cadres organisateurs de la « vraie vie » des signes », in M. Monte et G. Philippe (eds), Genres et Textes, Lyon PUL, 2014, p. 40).

[3] Sur la collocation, notamment dans le monde anglo-saxon, *cf.* par exemple la thèse de (Williams, 1999).

nature phraséologique, sémantique, thématique, rhétorique, stylistique, etc. [4] C'est en tout cas ainsi que nous l'entendrons dans cette contribution en précisant deux aspects symétriques et essentiels ; essentiels car touchant à l'identité de l'ADT que nous entendons célébrer ici : une cooccurrence est une unité statistique pour le texte / une cooccurrence est une unité textuelle pour la statistique.

(i) Une cooccurrence est une unité statistique pour le texte

Pour l'ADT, ce n'est pas seulement l'existence ou la présence d'une association qui compte mais sa fréquence. Et celle-ci devra être significative. La cooccurrence est l'association statistiquement significative de deux items (en général deux mots) dans une fenêtre déterminée du texte (en général le paragraphe). De nombreux indices statistiques ont été donnés pour mesurer l'attraction entre deux mots, dont les plus répandus sont, dans le monde, l'Information mutuelle notamment reprise par (Church et Hanks, 1990) et, en France, les Méthodes des cooccurrences de (Lafon, 1984). Selon les canons de l'ADT, cette dimension statistique, nécessaire donc, est le plus souvent calculée par comparaison entre la fréquence attendue de la rencontre entre les deux items considérés, et la fréquence effectivement constatée dans le corpus. Nous reprenons ainsi dans cette communication l'indice proposé récemment par (Brunet, 2012:222) et implémenté dans le logiciel HYPERBASE car sa philosophie et sa mise en application sont simples, élégantes et orthodoxes en ADT, en mobilisant le modèle hypergéométrique :



Calcul de la cooccurrence théorique

Le calcul de la cooccurrence théorique s'appuie sur le produit de deux probabilités: celle qui est attachée au premier mot et celle qui est propre au second. Chacune de ces probabilités relève du calcul hypergéométrique, les paramètres étant fixés comme suit:

T = Nombre total de mots dans le corpus
t = nombre moyen de mots dans un paragraphe (ou dans une page)
f = fréquence du mot considéré dans le corpus
k = 0 absence de ce mot dans le paragraphe (ou dans la page)

On obtient p1 pour l'absence du mot 1 et p2 pour l'absence du mot 2.
Le complément à l'unité de cette probabilité sert à mesurer les chances de rencontrer au moins une fois le mot dans l'espace considéré,
q1(présence mot1) = 1 – p1 et q2(présence mot2) = 1 – p2
et le produit des deux résultats mesure les chances d'observer la cooccurrence des deux mots à la fois dans le même paragraphe (ou la même page).
p (cooccurrence)= q1 * q2
En multipliant cette probabilité élémentaire par le nombre de paragraphes (ou de pages), on obtient l'effectif théorique des cooccurrences des deux mots dans le corpus. Reste à comparer l'effectif réel à l'effectif théorique, ce dont rend compte le calcul classique de l'écart réduit.

Remarque. La cooccurrence théorique de trois termes pourrait être calculée aussi facilement: avec p3 et q3 pour le mot 3 et p = q1 * q2 * q3

Figure 1. Calcul de la cooccurrence théorique (Brunet, 2012:222)

[4] Pour d'autres précisions terminologiques sur la *cooccurrence*, la *collocation*, la *colligation*, le *segment répété*, le *motif*, *cf.* récemment par exemple (Legallois, 2012). Pour illustrer le débat, disons que la *collocation* est souvent rattachée à la phrase voire au syntagme, et est très sensible à l'ordre voire à la contiguïté des éléments. La *cooccurrence* opère quant à elle souvent au niveau du paragraphe et tolère voire recherche la discontiguïté. Néanmoins, Serge Fleury dans LE TRAMEUR propose sous la fonction *Cooc* un calcul statistique de cooccurrences fortement contraintes par leurs positions et/ou fonctions grammaticales (http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur.pdf). Le moteur de recherche de TXM (http://textometrie.ens-lyon.fr/) permet aussi, via des expressions régulières, de rechercher des cooccurrences/collocations dans le corpus. Les notions de *segment répété* ou de *motif*, développées infra, sont des mixtes à fort potentiel heuristique.

Et la formule hypergéométrique traditionnelle peut être largement simplifiée du fait de la valeur nulle de *k*.

$$prob : \frac{f!\ (T-f)!\ t!\ (T-t)!}{k!\ (f-k)!\ (t-k)!\ (T-f-t+k)!\ T!}$$

$$Soit\ pour\ k=0 : (T-f)!\ (T-t)!\ /\ (T-f-t)!\ T!$$

(ii) Une cooccurrence est une unité textuelle pour la statistique ADT

C'est ici l'affirmation convaincue de cette contribution, tacitement partagée mais insuffisamment implémentée dans nos outils et nos parcours d'analyse.

Certes historiquement présente dans les pratiques ADT et parfois superbement raffinée (*cf.* dès le départ (Demonet et al., 1975), puis (Lafon et Tournier, 1978)), la cooccurrence reste négligée en tant qu'unité textuelle, particulièrement dans le traitement contrastif des corpus, lorsqu'il s'agit de caractériser un texte ou un auteur au sein du corpus. La paire cooccurrentielle est rarement considérée comme une unité textuelle *sui generis* – nous pourrions dire : la cooccurrence perçue comme une occurrence en tant que telle – susceptible de discriminer un texte.

Pourtant, en passant de l'occurrence (le mot seul) à la cooccurrence (la paire de mots), nous effectuons un saut qualitatif décisif et passons de la forme au sens.

Nous avons défini ailleurs de manière ambitieuse la cooccurrence comme *la forme minimale du co(n)texte* ; ce co(n)texte condition de la maïeutique du sens (Mayaffre, 2008-b). Dans une relation essentielle et réciproque – mais non nécessairement symétrique – lorsque A et B occurrent ensemble, A et B se co(n)textualisent minimalement mais essentiellement l'un l'autre. Autrement dit, ils se sémantisent : cette co-présence matérielle de A et B au sein du paragraphe *fait sens*. La paire cooccurrentielle accède ainsi au statut de molécule sémantique, là où le mot-atome n'a pas de sens et à peine une signification. Qu'il nous soit permis d'exemplifier de manière caricaturale le propos : la fréquence de l'atome « classe » dans un texte ne nous dit rien encore, la fréquence de la molécule cooccurrencielle « classe_prolétariat » ou « classe_ouvrière » ou « classe_lutte » (ou au contraire « classe_élève ») nous apprend déjà beaucoup. Les mots comme encore les lemmes sont, comme on le sait, presque toujours ambigus ; la paire cooccurrentielle presque jamais. Qu'il nous soit permis maintenant d'illustrer plus subtilement le propos : la fréquence de l'atome « travail » aussi bien chez Nicolas Sarkozy que chez Arlette Laguiller durant la campagne de 2007 ne nous renseigne que faiblement sur les programmes respectifs du candidat de droite et de la candidate d'extrême gauche. En revanche, la fréquence chez le premier de la molécule cooccurrentielle « travail_mérite » et chez la seconde de « travail_capital » en dit long à l'analyste du discours politique : vision hégélienne du travail chez le premier (la libération par le travail), vision marxiste chez la seconde (l'aliénation par le travail). Autour du même mot (« travail »), grâce aux cooccurrences, se déclinent deux conceptions du monde (Mayaffre, 2012-b:60-69).

(Guiraud, 1954, 1960) et (Muller, 1968) s'en sont tenus à une approche occurrentielle des textes proposant un schéma d'urne nécessaire et fondateur mais dans lequel chaque boule (*i.e.* chaque mot) est indépendante. D'une autre manière, (Lebart et Salem, 1994) discutent, dans le chapitre 2 de leur livre de référence, des « unités du texte », et passent en revue la forme graphique, le lemme ou le segment, mais ne s'arrêtent sur la paire cooccurrentielle en tant

qu'unité que succinctement [2.7. Cooccurrences et quasi-segments : 39-42] ; comme Bolasco dans sa dernière livraison sur le *Text mining* (Bolasco, 2013). (Baayen, 2001), quant à lui, dans un ouvrage pourtant ambitieux, qui illustre au passage le regrettable écart entre les littératures scientifiques anglo-saxonnes et françaises[5], ne considère jamais les collocations ou les cooccurrences.

De fait, nos pratiques descriptives et exploratoires des textes demeurent encore essentiellement occurrentielles, et l'outillage statistique implémenté dans les logiciels aujourd'hui reste souvent au seuil de la cooccurrence, si l'on excepte la philosophie de la méthode Alceste (Analyse des Lexèmes Cooccurrents dans un Enonce Simple d'un Texte) de (Reinert, 1993 ; Ratinaud et Marchand, 2012) aujourd'hui développé dans IRAMUTEQ (http://www.iramuteq.org/) ou des outils spécifiques comme COOCS de William Martinez (http://williammartinez.fr/coocs/page.php) que l'on retrouve dans LE TRAMEUR de Serge Fleury (http://www.tal.univ-paris3.fr/trameur/). Et la plupart des travaux – dont les nôtres – se contente ainsi le plus souvent, dans un premier mouvement, d'un traitement statistique de l'occurrence (le mot-atome) ; charge ensuite aux fonctions documentaires d'une remise en co(n)texte par un concordancier par exemple ; ou charge à la statistique de produire dans un second temps – mais c'est déjà trop tard – un calcul de cooccurrence sur l'unité-mot d'abord considérée isolément.

4. Contrastes textuels

4.1. Spécificités cooccurrentielles

Ainsi l'outil de la statistique descriptive le plus utilisé de l'ADT française depuis 30 ans est le calcul des spécificités que (Lafon, 1984) et le laboratoire de Saint-Cloud (Demonet et al., 1975) ont proposé dans les années 1980. Sa performance n'a plus à être démontrée et des centaines d'articles, de thèses ou d'ouvrages reposent sur sa pertinence. Seulement, le calcul est presque toujours appliqué à l'item seul, qu'il prenne la forme du mot graphique, du lemme ou d'une étiquette morpho-syntaxique. Ainsi pouvons-nous caractériser le discours François Hollande (2012-2014) à l'intérieur du corpus présidentiel constitué à partir de 1958 des discours de De Gaulle, Pompidou, Giscard, Mitterrand, Chirac et Sarkozy[6] (tableau 1).

Forme graphiques	Lemmes	Codes
aujourd'hui (+17,8)	Est-ce que (+21,1)	Ponctuation forte (+18,5)
Entreprises (+16,6)	Emploi (+17,4)	Nom commun Mascul Sing (+14,5)
Mali (+14,3)	Entreprise (subst.) (+16,4)	Déterminant Article Pluriel (+14,5)
Compétitivité (+14,2)	Euros (+14,5)	Verbe principal Infinitif (+14,3)
Pour (+13,9)	Devoir (verbe) (+12,2)	Pronom démonstratif (+13,2)
Etc.	Etc.	Etc.

Tableau 1. Spécificités de Hollande ; calcul hypergéométrique (approximé ici en écarts réduits)

[5] L'auteur qui reprend pourtant souvent l'histoire de la discipline rend imparfaitement hommage à Guiraud (jamais cité) ou à Benzecri (jamais cité), à Lebart (une fois cité seulement) et Salem (une fois cité). Muller et Brunet sont néanmoins mentionnés 6 et 3 fois.

[6] Le corpus présidentiel (1958-2014) est constitué à ce jour de 573 discours de De Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy et Hollande, équivalents à 2 824 973 occurrences.

Conscient de la puissance évidemment, mais des limites de ce type d'approche monogrammique (un mot seul, un lemme seul, un code seul), André Salem a précocement appliqué le calcul aux *segments répétés* (Salem, 1986, 1987 ; Lebart et Salem, 1994:29-39), et la puissance du logiciel LEXICO réside, entre autres, dans sa faculté d'extraction statistique de tous les segments ou n-grams de longueur 2 à longueur *n*, qui caractérisent statistiquement un texte donné dans son corpus. Par exemple, dans le corpus présidentiel, les segments répétés de François Hollande sont (tableau 2) :

Segments répétés de longueur 9 et 8	**Fréq corpus**	**Fréq Hollande**	**Ecarts**
… c'est le rôle du président de la république…	4	3	+7,5
… il n'y a pas de temps à perdre…	4	3	+7,5
… le contrat de génération qui va donner une…	8	8	+14,3
…l'accord sur la sécurisation de l'emploi…	7	7	+13,4
… le barême de l'impôt sur le revenu…	3	3	+8
Etc.	Etc.	Etc.	Etc.

Tableau 2. Segments répétés spécifiques de Hollande de longueur 9 et 8, calcul hypergéométrique (approximé ici en écarts réduits)

Dès lors, pourquoi ne pas appliquer le calcul des spécificités aux paires de mots (cooccurrence), comme on l'applique aux mots (occurrence) ou à une suite de mots (segment) ? C'est ce que permet aujourd'hui HYPERBASE 9.0 (2014) avec des performances remarquables et des limites que l'on mentionne plus bas. Etienne Brunet a en effet implémenté en 2012 (Brunet, 2012) et amélioré ce printemps 2014, le calcul des spécificités cooccurrentielles[7] ainsi que de nombreux autres indices statistiques portant sur les paires (et non plus seulement les mots). Par ordre hiérarchique, les paires spécifiques (positives et négatives) de Hollande apparaissent dans la figure 2 (ci-dessous).

Les molécules sémantiques que constituent les paires de cette liste offrent un potentiel interprétatif évident en analyse de discours. François Hollande – en comparaison avec ses prédécesseurs – est moins le gardien des institutions (Article 5 de la Constitution) ou le chef des armées (article 15) que le PDG d'une entreprise-France en pleine crise économique : les 5 premières paires positives relèvent du domaine économique ; les 10 premières paires négatives relèvent du géo-stratégique ou de l'institutionnel. Ce déplacement des prérogatives présidentielles, telles qu'elles transpirent du corpus élyséen depuis 1958, est déjà marqué sous le quinquennat Nicolas Sarkozy mais s'accentue avec le présent Président. Sous le double effet de la construction européenne, qui dilue de fait la souveraineté nationale, et de la mondialisation qui exacerbe la compétition économique, les Présidents français se ministérialisent au fil du temps – jusqu'à envisager la suppression du 1er ministre en ce qui concerne Sarkozy. Le contexte économique et la crise post-industrielle les cantonnent le plus souvent au rôle de directeur des ressources humaines condamné à licencier ou de chef d'entreprise au bord du dépôt de bilan, assez loin des fonctions régaliennes traditionnelles. Plus fondamentalement encore le triomphe historique de l'idéologie libérale, commencé peut-être dès les années 1980, semble aujourd'hui achevé en faisant de l'homme –ici le Président – avant tout un *homo economicus*.

[7] Que l'on ne confondra pas avec le traditionnel calcul des « cooccurrences spécifiques » d'un mot-pôle (par exemple (Martinez, 2012)).

écart	corpus	texte	mot
13.7	58	35	avenir_emploi
12.8	22	22	contrat_génération
9.1	130	31	emploi_jeune
8.9	83	25	entreprise_salarié
8.5	25	15	crédit_impôt
8.4	62	21	euro_milliard
8.1	16	12	génération_jeune
8.1	13	11	avenir_génération
7.6	12	10	emploi_génération
7.4	152	27	année_fin
7.1	24	12	euro_million
7.0	37	14	marché_travail
7.0	31	13	avenir_jeune
7.0	25	12	euro_zone
6.9	9	8	avenir_contrat
6.8	75	18	emploi_entreprise
6.7	42	14	entreprise_jeune
6.1	86	17	emploi_formation
6.1	13	8	chômage_fin
6.0	32	11	contrat_emploi
6.0	19	9	contrat_jeune
5.8	15	8	accord_emploi
5.7	16	8	emploi_fin
5.7	11	7	avenir_fin
5.6	12	7	entreprise_euro
5.5	59	13	année_emploi
5.3	9	6	parlement_partenaire
5.3	9	6	formation_génération
5.3	9	6	économie_milliard
5.3	36	10	chômage_jeune
5.3	35	10	activité_entreprise

écart	corpus	texte	mot
-2.4	294	3	français_france
-2.4	239	2	france_problème
-2.1	767	15	président_républ
-2.0	337	5	monde_pays
-2.0	197	2	gaulle_général
-2.0	192	2	france_situation
-1.8	305	5	france_politique
-1.8	217	3	droit_homme
-1.8	166	2	action_france
-1.7	163	2	chef_gouvernemen
-1.6	463	10	france_monde
-1.6	145	2	pays_problème
-1.6	141	2	france_vie
-1.5	137	2	droit_france
-1.5	131	2	france_moyen
-1.4	159	3	fois_france
-1.4	127	2	action_gouvernem
-1.4	124	2	europe_union
-1.4	120	2	france_question
-1.3	218	5	allemagne_france
-1.3	152	3	pays_politique
-1.3	116	2	gouvernement_pro
-1.3	116	2	exemple_france
-1.3	116	2	décision_france
-1.3	112	2	français_vie
-1.3	112	2	effort_france
-1.3	105	2	année_pays
-1.2	98	2	affaire_france
-1.2	97	2	france_indépenda
-1.2	95	2	français_préside
-1.2	93	2	monde_paix

Figure 2. Spécificités cooccurrentielles ou paires spécifiques positives (à gauche) et négatives (à droite) de Hollande ; calcul hypergéométrique (approximé ici en écarts réduits)
Sortie HYPERBASE 9.0 (2014)

De fait, à l'autre extrémité chronologique du corpus, sous les présidences de De Gaulle ou de Pompidou, dans les années 1960, les paires cooccurrentielles positivement spécifiques sont toutes autres :

écart	corpus	texte	mot
13.8	118	77	algérie_france
8.9	36	28	parti_régime
8.1	68	37	etat_nation
7.7	34	24	algérie_pays
7.5	76	37	monde_peuple
7.5	23	19	indépendance_progrès
7.5	15	15	algérie_communauté
7.4	21	18	peuple_progrès
7.2	35	23	coopération_peuple
7.2	32	22	jour_peuple
7.2	20	17	algérie_etat
7.2	123	48	france_peuple
7.0	91	39	coopération_france
7.0	175	59	etat_france
6.9	19	16	milieu_monde
6.9	19	16	algérie_paix
6.6	14	13	algérie_avenir
6.6	12	12	algérie_oeuvre
6.5	50	26	paix_progrès
6.4	19	15	pays_référendum
6.4	19	15	allemagne_coopératio

écart	corpus	texte	mot
10.7	197	84	gaulle_général
7.0	13	13	civilisation_concept
6.0	69	29	homme_vie
5.7	9	9	général_peuple
5.7	9	9	gaulle_peuple
5.7	23	15	conception_vie
5.5	28	16	civilisation_homme
5.3	20	13	paris_région
5.1	38	18	france_gaulle
5.1	34	17	etats_unis_président
5.1	31	16	coopération_politiqu
5.0	32	16	besoin_homme
5.0	25	14	civilisation_monde
4.9	44	19	action_domaine
4.7	18	11	conception_société
4.5	13	9	effort_nécessité
4.4	30	14	communauté_domaine
4.4	20	11	communauté_négociati
4.4	11	8	paris_sommet
4.3	66	22	coopération_pays
4.3	28	13	conférence_sommet

Figure 3. Spécificités cooccurrentielles (ou paires spécifiques) positives de De Gaulle (gauche) et Pompidou (droite) ; calcul hypergéométrique (approximé ici en écarts réduits)
Sortie HYPERBASE 9.0 (2014)

Ce qui nous intéresse ici est la plus-value interprétative que l'unité cooccurrentielle apporte.

Par exemple, sur 60 ans, l'atome « France » et ses occurrences sont relativement bien distribués dans le corpus entre les Présidents, sans que l'on repère une tendance chronologique. Mais là où De Gaulle et Pompidou, dans les années 1960, s'adressaient plutôt aux citoyens (« France_Peuple »), Chirac, Sarkozy ou Hollande, dans les années 2000, s'adressent désormais aux agents économiques (« France_croissance ») (figure 4).

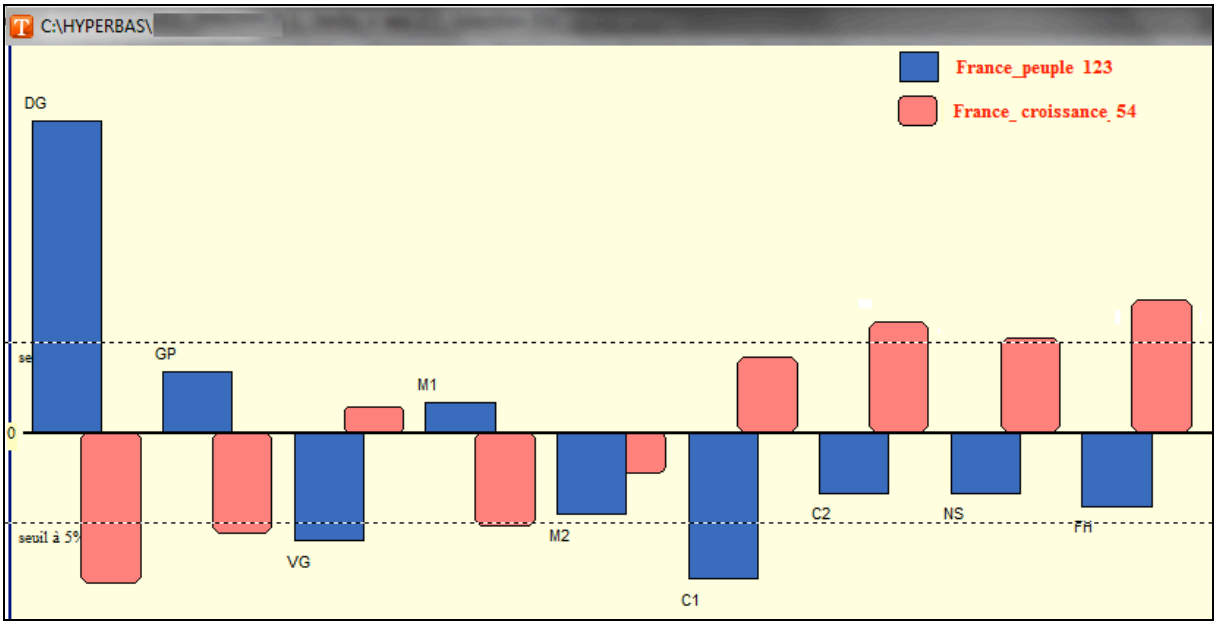


Figure 4. Distribution des paires « France_peuple » et « France-croissance » dans le discours présidentiel (1958-2014)

4.2. AFC cooccurrentielle

Si l'outil traditionnel de la statistique descriptive est la *spécificité*, l'outil de la « statistique exploratoire multidimensionnelle » (Lebart et al., 2006) le plus utilisé de l'ADT française reste, depuis la révolution benzécrienne, l'AFC (Analyse Factorielle des Correspondances).

Mais là encore, nos pratiques consistent le plus souvent à croiser, dans un tableau rectangulaire (tableau de contingence) des textes d'un côté et des unités textuelles élémentaires que sont les mots (lemmatisés ou non) de l'autre. Sur les nuages ainsi obtenus (par exemple et historiquement (Prost, 1974)), dont il ne s'agit pas de remettre en cause la puissance exploratoire et heuristique, le danger existe que l'on imagine que les mots voisins cooccurrent. Puisqu'ils sont à proximité dans un même quadrant, les mots entretiendraient des relations cooccurrentielles, appartiendraient peut-être aux mêmes thèmes, relèveraient d'une même isotopie. Conclure de la sorte ne constitue pas une erreur à coup sûr, mais est un raccourci méthodologique et une surinterprétation de la méthode.

L'idée devient dès lors de croiser les textes non plus seulement avec les mots-atomes, mais avec des paires-molécules dont on a souligné la dimension intrinsèquement sémantique (*i.e.* co(n)textuelle) (figure 5).

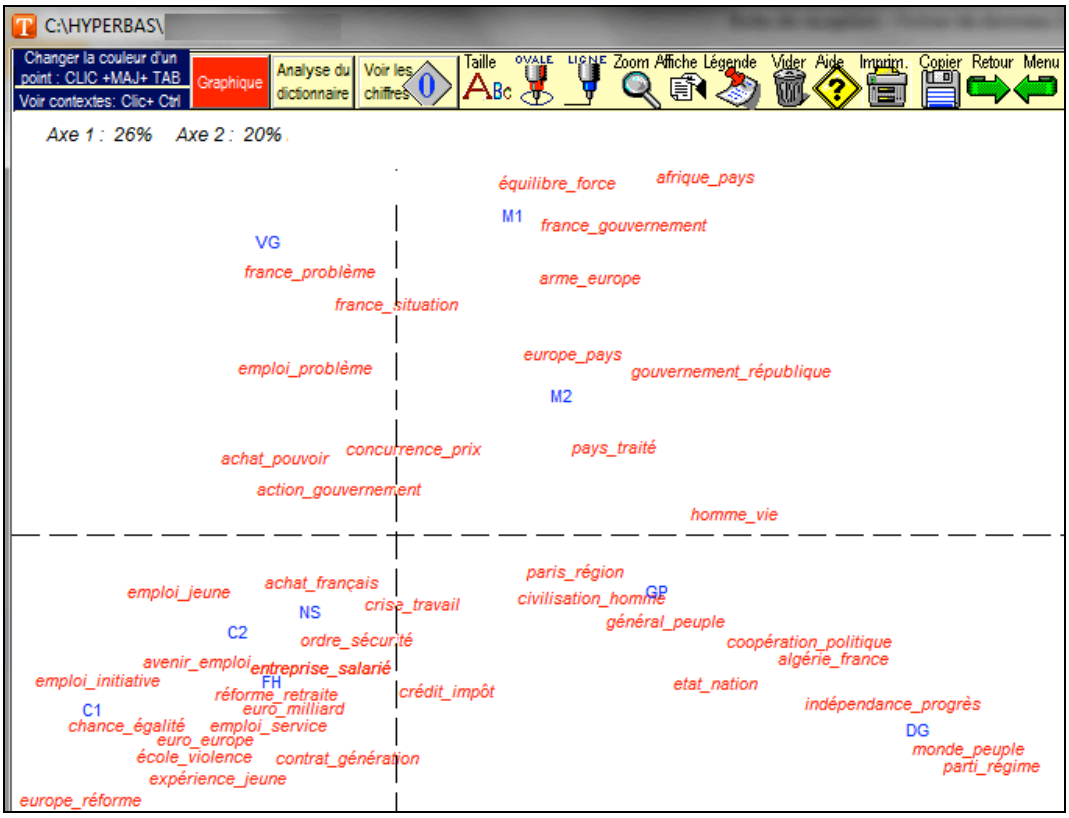


Figure 5. AFC de 50 paires dans le discours présidentiel (1958-2014)

4.3. Distance intertextuelle, accroissement chronologique, autres…et limites

Ce passage de l'occurrence à la cooccurrence, illustré sur les spécificités et l'AFC, peut être ainsi généralisé. (Brunet, 2012) montre par exemple que l'historique indice de corrélation chronologique, pertinent sur des corpus ordonnés et particulièrement des séries textuelles chronologiques, fonctionne ainsi aussi bien sur l'unité-cooccurrence que l'unité-occurrence. (Sur notre corpus 1958-2014, les paires qui augmentent le plus régulièrement sont ainsi très instructives du triomphe progressif de la problématique économique : «chomage_croissance » (+0,949), ou « emploi_Europe (+0.922), « formation_travail » (+0,921).) Nous montrons encore dans ces JADTs 2014 avec (Vanni et al., 2014) que le calcul de la distance intertextuelle (anciennement connexion lexicale) (*CORPUS* 2, 2003) se calcule, puis se représente aisément sur la paire comme il se calculait sur le gram. L'idée fondamentale étant ici que les Présidents sont condamnés es-qualité à utiliser un stock lexical prescrit (« France », « Europe », « politique », « gouvernement » etc.) et que leur originalité respective se situe, dés lors, dans les combinaisons (paires) préférentielles, c'est-à-dire dans la manière minimale de co(n)textualiser ces mots que leur fonction impose.

Ainsi, tous les indices que la communauté ADT a appris à utiliser gagnent aujourd'hui à être appliqués aux paires comme ils l'étaient aux formes. La limite d'HYPERBASE est de fonctionner certes de manière systématique, mais sur 300 mots déterminés seulement (soit 45 000 paires), là où nous pourrions généraliser encore l'analyse ; notamment, tous les exemples développés *supra* l'ont été sur la sélection grammaticale et fréquentielle des 300 substantifs

les plus utilisés dans le corpus, là où d'autres associations – et pourquoi pas toutes ? – auraient pu être traitées[8].

5. Cooccurrences et réseaux textuels

Dans ses livres baptismaux pour la discipline, Pierre Guiraud posait que le sens d'un mot est constitué par la somme de ses emplois, c'est-à-dire la somme de ses co(n)textes d'utilisation (Guiraud, 1954, 1960). Aussi, en définissant ici la cooccurrence comme la forme minimale et calculable du co(n)texte, nous pouvons préciser l'affirmation du côté de la statistique : *le sens d'un mot est la somme de ses cooccurrences*. Pour l'ADT en effet, la définition d'un mot peut être approchée ou calculée, en corpus, comme l'ensemble de ses fréquentations dans le texte. Non sans lien avec le distributionalisme harrisien (Harris, 1957), mais en ne mobilisant que la statistique, les mots pour nous prennent sens dans leur manière de distribuer leurs occurrences à l'intérieur du co(n)texte (*i.e.* vis-à-vis des autres mots) ; ou inversement, par la manière de recevoir des déterminations sémantiques de tous des mots partenaires qu'ils attirent. Plus loin, (Viprey, 1997, 2006), précisant les travaux de (Halliday et Hasan, 1976), attribue au texte une « texture » (à côté de la structure) ou une réticularité (à côté de la linéarité). Le texte est une *suite* bien sûr, mais aussi un entrelacs, un réseau, le croisement de mots qui entretiennent des relations (quantifiables), s'agencent, se structurent, se font écho ou se repoussent : *cooccurrent*. D'un point de vue étymologique, le *texte* est un *tissu*, un *tissage*… une tresse, un treillis, un maillage où le sens se tricote au fil des rencontrent mots-mots. De plain-pied avec la Linguistique textuelle (Adam 2006 ; Monte et Philippe, 2014), ces cooccurrences généralisées participent à la *cohésion* et à la *cohérence* du texte ; de plain-pied avec la sémantique de corpus (Condamines, 2005 ; Rastier, 2011), ces cooccurrences généralisées deviennent des corrélats sémantiques à l'occasion de parcours interprétatifs et rendent compte du sens des discours. Présentes à titre expérimental déjà chez Benzécri, les matrices mots x mots ou matrices cooccurrentielles semblent ainsi devenues depuis la fin du XX^{ème} siècle un objet statistique fondamental pour l'ADT (Tableau 3).

	Mot A	Mot B	Mot C	Mot D	Etc.
Mot A	***	*x* (cooc A_B)	*y* (cooc A_C)	*z* (cooc A_D)	…
Mot B	*x* (cooc A_B)	***	*v* (cooc B_C)	*w* (cooc B_D)	…
Mot C	*y* (cooc A_C)	*v* (cooc B_C)	***	*u* (cooc C_D)	…
Mot D	*z* (cooc A_D)	*w* (cooc B_D)	*u* (cooc C_D)	***	…
Etc.	…	…	…	…	***

Tableau 3. Matrice mots x mots ou matrice cooccurrentielle[9]

(i) L'ensemble de la matrice donne une représentation mathématique globale du texte-tissu.

[8] Pour l'heure, l'exploration systématique de la distribution de toutes les cooccurrences d'un gros corpus comme le nôtre reste techniquement difficile. Riche de 30 000 vocables différents (hors hapax), le nombre potentiel de paires s'élève à (30 000 x 30 000) / 2 dont on devrait ensuite regarder la distribution dans les 9 parties du corpus. C'est pourquoi HYPERBASE se contente d'une sélection de 300 lemmes choisis par critère de fréquence (les 300 plus nombreux) et sur critère grammatical (nom et/ou adjectif et/ou verbe). A noter que la fenêtre est par défaut le paragraphe et que l'ordre d'apparition des cooccurrents n'est pas contraint.

[9] Dans les cellules, *x, y, z*, etc. chiffrent la cooccurrence de A_B, de A_C, de A_D, etc. Ce chiffrage est d'abord une valeur absolue (le nombre de rencontres), mais peut devenir un indice (la probabilité de ce nombre de rencontres).

(ii) Chaque ligne (ou chaque colonne) donne le profil cooccurrentiel complet des mots : pour nous, une ligne représente le sens du mot en vertu de l'affirmation que *le sens d'un mot est la somme de ses cooccurrences*.

(iii) Les cellules enfin (croisements d'une ligne-mot avec une colonne-mot) sont les mailles fondamentales du texte : des cooccurrences, des co(n)textes minimaux mais fondamentaux, des noyaux de sens ou molécules sémantiques.

Fidèle à la sus-dite révolution benzécrienne (Viprey, 1997, 2006) à la suite de (Massonie, 1986) a proposé le traitement par l'AFC de ce type de matrices de contingence. Et notre discours présidentiel (1958-2014) peut être représenté comme un nuage de points (sur lequel nous avons colorié, seulement à titre illustratif, les mots spécifiques de De Gaulle).

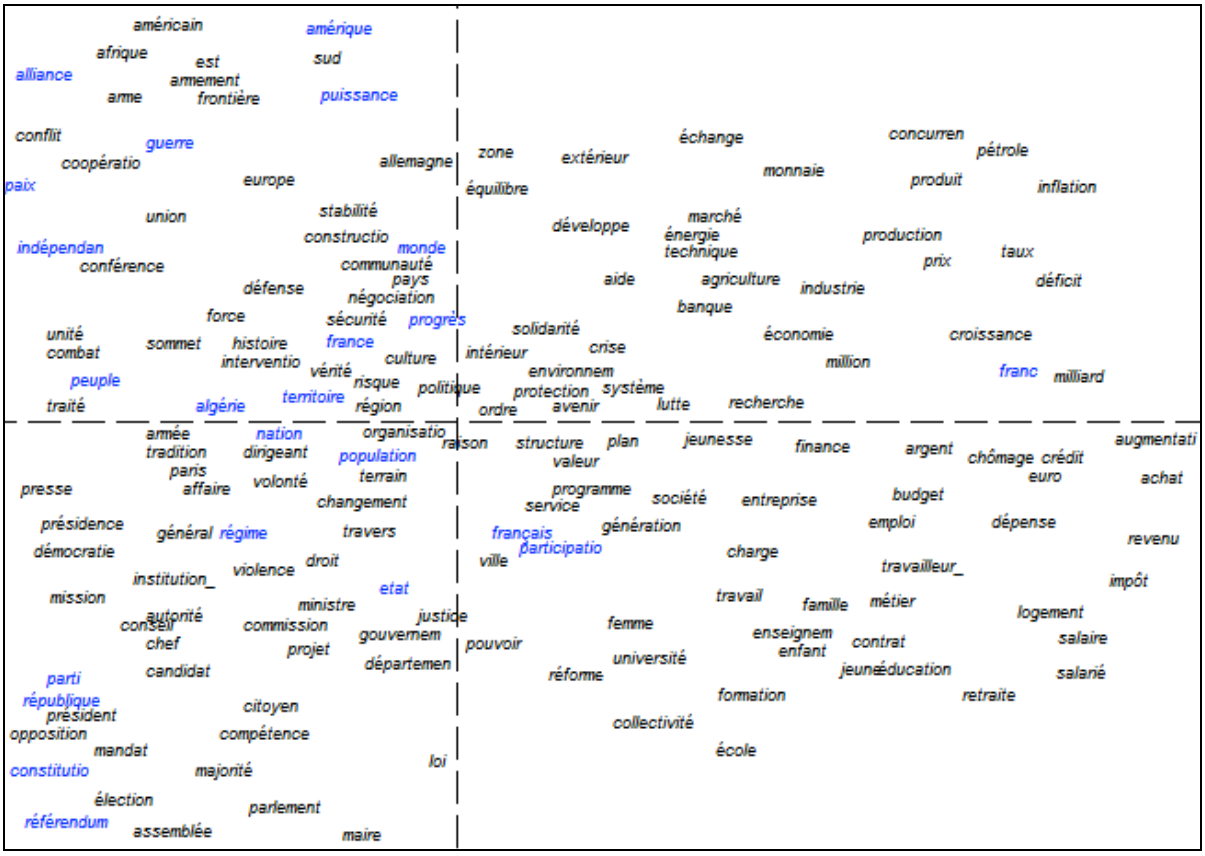


Figure 6. AFC de la matrice cooccurrentielle (200 substantifs les plus fréquents) ; en bleu les mots spécifiques de De Gaulle. Sortie HYPERBASE 9.0 (2014)

De manière très spectaculaire, en s'en tenant à une lecture par quadrant, 4 pôles structurent le lexique (ici des substantifs) de la communication présidentielle : la politique extérieure (haut-gauche), la politique institutionnelle (bas-gauche), la politique économique (haut-doit) et la politique sociale (bas-droit). Sans surprise pour ceux qui connaissent le gaullisme, les mots spécifiques de De Gaulle (en bleu) sont concentrés quasi-unanimement dans la partie gauche de la carte (vocabulaire institutionnel et de politique étrangère).

Enfin, de manière moins classique en ADT, les matrices mots x mots peuvent aujourd'hui donner lieu à de nombreux traitements puis représentations, grâce à des librairies d'outils statistiques en ligne, dont il appartient à la communauté ADT de juger de l'efficacité pour l'interprétation textuelle.

Nous proposons ici en guise de conclusion et grâce à l'articulation d'HYPERBASE et de GEPHI [http://gephi.org/], une représentation par réseau (*network*), spécification de la théorie des graphes (Bastian et al., 2009). La sortie machine, produite strictement sur la même matrice cooccurrentielle que précédemment, complète l'AFC ci-dessus.

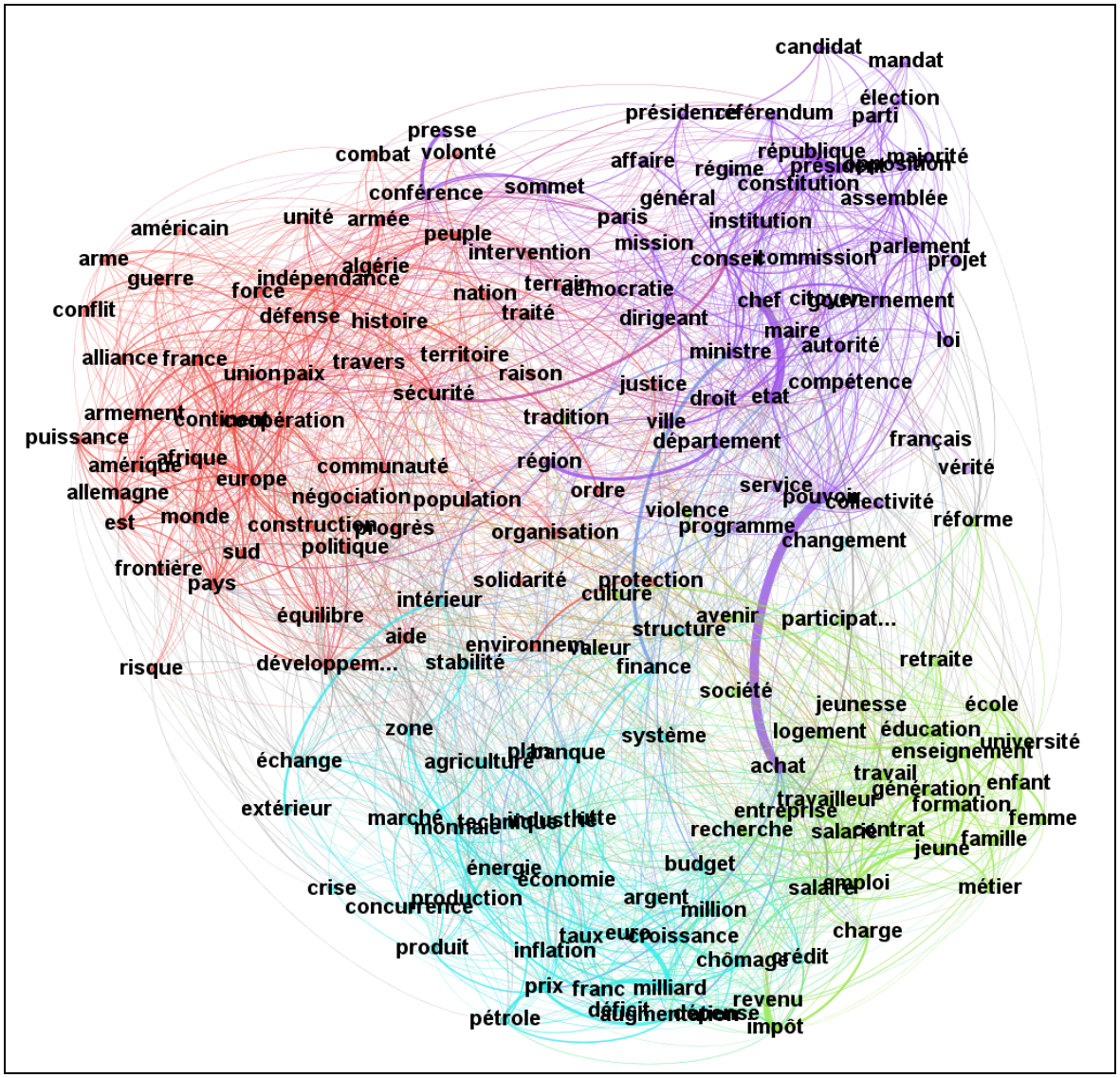


Figure 7. Matrice cooccurrentielle (200 substantifs) représentée en réseau
Sortie GEPHI 8.2-2013

Si le traitement par l'AFC relève d'une approche vectorielle portant sur le traitement des profils cooccurrentiels (chaque profil-vecteur concentrant la distribution d'un mot vis-à-vis de tous les autres), le réseau insiste sur la dimension directement relationnelle et l'interconnexion des mots-noeuds (relations directes mot-mot). C'est moins « d'isotropie » (Viprey, 1997, 2006) dont il s'agit que de « connexité ». L'image du texte comme un espace réticulaire où chaque mot est relié plus ou moins aux autres (*i.e.* cooccurre) est ainsi précisée. Les perfectionnements de ce type de traitement, fortement utilisés en sociologie (Hanneman et Riddle, 2005 ; Vergès et Bouriche, 2001) ou dans le TALN, laissent à l'utilisateur ensuite le loisir de calculer dans le réseau les « petits mondes lexicaux » (Watts et Strogatz 1998 ; Missen et al., 2008), les classes de modularité ou « communautés » (c'est-à-dire les pôles semi-connexes les mieux structurés dans le réseaux ; dans notre exemple représentés par

différentes couleurs) (Newman, 2006), ou, localement, depuis Dijkstra, « les plus courts chemins » entre deux mots. (Pour une utilisation en ADT des *networks* voir par exemple (Tauveron, 2012) ou aux dernières JADTs (Iezzi et al., 2012 ; Keller et Schultz, 2012 ; Gigante et Pellicia, 2010 ; Iezzi, 2010).)

Conclusion

Le propos de cette contribution n'était pas de lister les traitements existants en matière de cooccurrences dans le *Text mining* mondial, ni même dans l'ADT française.

Sans quoi, il nous aurait fallu accorder, par exemple, une attention particulière aux cooccurrences linguistiquement complexes comme les *motifs* (Mellet et Longrée, 2009 ; Longrée et Mellet, 2013). Trop souvent en effet – parfois pour de simples raisons didactiques – la cooccurrence est réduite à l'association de deux mots lorsque les associations d'éléments linguistiques de nature différente (lexique, grammaire, phonologie…) sont riches d'enseignement. Les *motifs* exposés plusieurs fois aux JADTS (outre (Mellet et Longrée, 2012), (Magri et Purnelle, 2012)) présentent l'avantage de mixer la cooccurrence lexicale et la colligation grammaticale et témoignent, de la langue au discours, de structures constitutives de la trame textuelle. De plus, ils accordent une place importante à l'ordre ou linéarité (ce qui n'est pas toujours le cas du traitement des cooccurrences) sans s'enfermer dans la stricte contiguïté. Ils constituent un de ces « nouveaux observables » (Rastier, 2011:13) linguistiques que l'ADT d'aujourd'hui peut inventer dans la continuité des occurrences, des segments répétés ou des cooccurrences *stricto sensu.*

Sans quoi, il nous aurait fallu aussi mieux souligner le prolongement de la cooccurrence le plus souvent binaire (la paire), en triplets, en quadruplets puis en Q-cooccurrences (Massonie, 1986). De nombreux outils et approches existent depuis le lexicogramme récursif de (Heiden, 2004) ou les poly-cooccurrences de (Martinez, 2003), en passant nous l'avons vu par la cooccurrence généralisée de (Viprey, 1997) ou l'étude aujourd'hui des réseaux. Et ces développements méritent d'être aujourd'hui poursuivis autour des cooccurrences de « deuxième ordre » (Bertels et Speelman, 2012 ; Lauf et al., 2012) ou cooccurrences indirectes (Mayaffre, 2008-b) c'est-à-dire, de manière générale, des *cooccurrences conditionnelles* (A → (?) → C), quitte à explorer une statistique bayésienne moins traditionnelle.

Enfin, sans prétendre être exhaustif, nous aurions pu revenir encore sur l'asymétrie cooccurrentielle que les précédentes JADTs ont accueillie (Luong et al., 2010 ; Ben Hamed et Mayaffre, 2012 ; Bonneau, 2012 ; Boulard et Poudat, 2012). Réciproques, disions-nous, les relations cooccurrentielles entre A et B ne sont pas nécessairement symétriques, comme le montre intuitivement le rapport entre un nom commun fréquent et un adjectif rare et particulier (« voie_lactée », « matière_fécale », « voiture_rutilante », etc.). Et dans cette asymétrie se joue les forces linguistiques qu'il reste à modéliser du côté statistique pour faire de la paire non seulement un objet, mais un objet orienté.

Mais l'objectif de cette contribution était plus modeste en espérant être plus fondamental. Il s'agissait seulement d'établir la cooccurrence comme un élément, comme un objet, comme un observable du texte ; la paire comme une unité ; la cooccurrence comme une occurrence ; le binôme comme un item.

La cooccurrence est une unité fondamentale du texte peut-être plus pertinente que l'occurrence seule car elle n'est pas seulement item mais co(n)texte, c'est-à-dire pas seulement forme mais sens. Seule, nous avons montré qu'elle avait un pouvoir de

discrimination important dans des corpus contrastifs ; généralisée, elle modélise le texte comme réticularité et rend compte aussi bien du sens que de l'organisation ou structure textuelle (*i.e.* la textualité).

Aussi, considérer la cooccurrence, c'est allier l'urne et le réseau, le paradigmatique et le syntagmatique. La cooccurrence porte l'exigence statistique : elle est elle-même calculée et devient une unité de calcul pour l'ADT. Mais la cooccurrence porte aussi en elle l'exigence co(n)textuelle : elle est la forme (certes minimale mais calculable) du co(n)texte. Son traitement rallie ainsi les contraires (segmenter/co(n)textualiser) et transforme le paradoxe de la discipline en complémentarité ; pour que vive l'ADT.

Références

Adam J.-M. (2006). « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse des données textuelles », conférence d'ouverture des JADT 2006 [http://lexicometrica.univ-paris3.fr/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf]

Baayen R.H. (2001). *Word frequency distributions*. Dordrecht : Kluwer Academic Publishers.

Bastian M., Heymann S., Jacomy M. (2009). « Gephi: an open source software for exploring and manipulating networks », International AAAI Conference on Weblogs and Social Media [http://gephi.org/publications/gephi-bastian-feb09.pdf].

Ben Hamed M. et Mayaffre D. (2012). « Saisir le sens dans les deux sens : exploration de la portée interprétative de l'énergie et de la disponibilité », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 121-134.

Benzécri J.-P. (1980). *Pratique de l'analyse des données*. Paris : Dunod.

Bertels A. and Speelman D. (2012) « La contribution des cooccurrences de deuxième ordre à l'analyse sémantique », *Corpus* 11, pp. 147-166.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma : Carocci.

Bonneau J. (2012). « La cooccurrence asymétrique : propriétés quantitatives en disponibilité et énergie », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Universitaires Saint-Louis, pp. 189-201.

Boulard A. Poudat C et Gauthier J.-M. (2012). « Des mots pour se dire : développement de la fonction narrative chez l'enfant », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Universitaires Saint-Louis, pp. 203-214.

Brunet E. (2012). « Nouveau traitement des cooccurrences dans Hyperbase », *Corpus*, 11, pp. 219-248.

Church K. W. & Hanks P. (1990). « Word Association Norms, Mutual Information, And Lexicography », *Computational Linguistics*, vol. 16(1), pp. 177-210.

Condamines A. (éd) (2005). *Sémantique et corpus*. Londres : Hermes

CORPUS 2 (2003). « La distance intertextuelle » (sous la direction de X Luong).

Demonet M. *et al*. (1975). *Des tracts en mai 68*. Paris : Colin.

Firth J. (1957). « A Synopsis of Linguistic Theory 1930-1955 », *Studies in Linguistic Analysis*, pp. 1-32.

Fleury S. (2013). *Annotations Rhapsodie pour le Trameur* [(http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur.pdf)]

Guiraud P. (1954). *Les Caractères statistiques du vocabulaire*. Paris : PUF.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : Larousse.

Halliday M. A. K. and Hasan R. (1976). *Cohesion in English*. London : Longman.

Hanneman R. A. and Riddle M. (2005). *Introduction to social network methods*. Riverside : University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/).

Harris Z. S. (1957). « Co-occurrence and transformation in linguistic structure », *Language*, 33, pp. 283-340.

Heiden S. (2004). « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex », *JADT 2004,* édité par G. Purnelle, C. Fairon et A. Dister, Louvain : Presses universitaires de Louvain, pp. 577-588.

Heiden S. et Lafon P. (1998). « Cooccurrences. La CFDT de 1973 à 1992 », in *Des mots en liberté, Mélanges Maurice Tournier*, Paris, ENS Éditions, tome 1, pp. 65-83.

Iezzi D. F. (2010). « Topic connections and clustering in text mining: an analysis of the JADT network », *JADT 2010*, édité par S. Bolasco, I. Chiari, L. Giuliano. Milan : Edizioni Universitarie di Lettere Economia Diritto, pp. 720-729.

Iezzi D. F., Mastrangelo M. e Sarlo S. (2012). « Text clustering based on centrality measures: an application on job advertisements », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 515-524.

Gigante A. e Pelliccia E. (2010). « L'immagine delle parole in rete. Applicazione di *Network Text Analysis* sui gruppi di *Fecebook* dedicati alla politica », *JADT 2010*, édité par S. Bolasco, I. Chiari, L. Giuliano. Milan : Edizioni Universitarie di Lettere Economia Diritto, pp. 631-642.

Keller D. B. and Schultz. (2012). « Morpheme networks reveal language dynamics », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 525-535.

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

Lafon P. et Tournier M. (1978). « Une Nouvelle approche lexicométrique des cooccurrences dans un texte », *Travaux de lexicométrie et de lexicologie politique*, 3, pp.135-148.

Lauf A., Valette M. et Khouas L. (2012). « Analyse du graphe des cooccurrents de deuxième ordre pour la classification non-supervisée de documents », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 577-589.

Lebart L., Piron M. et Morineau A. (2006 – 4ème éd.). *Statistique exploratoire multidimentionnelle. Visualisation et inférence en fouille de données*. Paris : Dunod.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris : Dunod.

Legallois D. (2012). « La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? », *Corpus*, 11, pp. 31-54.

Longrée D. et Mellet S. (2013). « Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours », *Langages*, 189, pp. 65-79.

Longrée D., Luong X. et Mellet S. (2008), « Les motifs : un outil pour la caractérisation topologique des textes », *JADT 2008*, édité par S. Heiden et B. Pincemin, Lyon : PUL, pp.733-744.

Luong *et al.* (2010). « La cooccurrence, une relation asymétrique ? », *JADT 2010*, édité par S. Bolasco, I. Chiari, L. Giuliano, Milan : Edizioni Universitarie di Lettere Economia Diritto, pp. 321-331.

Magri V. et Purnelle G. (2012). « Mot à mot, brin par brin : les suites [Nom préposition Nom] comme motifs », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 659-673.

Martinez W. (2012). « Au-delà de la cooccurrence binaire… Poly-cooccurrences et trames de cooccurrence », *Corpus*, 11, pp. 191-218.

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat, Université de la Sorbonne nouvelle-Paris 3, sous la direction d'A. Salem.

Massonie J.-P. (1986). *Pratique de l'analyse des correspondances*. Besançon : Annales Littéraires de l'Université de Franche-Comté.

Mayaffre D. (2008-a). « Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence », *JADT 2008*, édité par *S.* Heiden et B. Pincemin, Lyon : PUL, vol. 2, pp. 811-822.

Mayaffre D. (2008-b). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Sémantique & Syntaxe*, 9, pp. 53-72. [Hal : http://hal.archives-ouvertes.fr/hal-00551114].

Mayaffre D. (2012-a). *Le discours présidentiel sous la Vème République*. Paris : Presses de Sciences Po.

Mayaffre D. (2012-b). *Mesure et démesure du discours. Nicolas Sarkozy (2007-2012)*. Paris : Presses de Sciences Po.

Mellet S. et Longrée D. (2009). « Syntactical 'Motifs' and Textual Structures », *Belgian Journal of Linguistics*, 23, pp. 161-173.

Mellet S. et Longrée D. (2012). « Légitimité d'une unité textométrique : le motif », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 715-728.

Missen M., Boughanem M. et Gaume B. (2008). « The Small World of Web Network Graphs », *International Multitopic Conference* (IMTIC 2008), Vol. CCIS, Abdul Qadeer *et al*. (Eds.), Springer, CCIS, pp. 133-145.

Monte M. et Philippe G. (eds.) (2014). *Genres et Textes. Déterminations, évolutions, confrontations*. Lyon : PUL.

Muller Ch. (1968). *Initiation à la statistique linguistique*. Paris : Larousse.

Newman M. E. J. (2006). « Modularity and community structure in networks », Proc. Natl. Acad. Sci. USA, vol. 103, no 23, pp. 8577–8582.

Palmer (1933). *Second Interim Report on English Collocations*. Tokyo : Kaitakusha.

Prost A. (1974). *Vocabulaire des proclamations électorales de 1881, 1885, 1889*. Paris : PUF.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris : Champion.

Ratinaud P. et Marchand P. (2012). « Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRAMUTEQ », *JADT 2012*, édité par A. Dister, D. Longrée, G. Purnelle. Bruxelles : Université de Liège / Facultés Saint-Louis, pp. 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, pp. 5-39.

Salem A. (1986). « Segments répétés et analyse statistique des données textuelles », *Histoire et Mesure*, vol.1-n°2, pp. 5-28.

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

Sinclair J. M. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sinclair J. M. (2003). *Reading concordances*. Londres : Pearson Longman.

Tauveron M. (2011). « De la cooccurrence généralisée à la variation du sens lexical », *Corpus*, 12, pp. 219-248.

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

Vanni L, Luong X. et Mayaffre D. (2014). « Arbre et co-occurrences. Nouvel outil logométrique sur le net. Application au discours de François Hollande », in *JADT 2014*.

Vergès P. et Bouriche B. (2003). « L'analyse des données par les graphes de similitude », *Sciences Humaines* [http://www.scienceshumaines.com/textesInedits/Bouriche.pdf].

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

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

Watts D. J. et Strogatz S.H. (1998). « Collective dynamics of 'small-world' networks », *Nature*, 393 (6684), pp. 440-442.

Williams G. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. Thèse de doctorat, Université de Nantes.