Quand « travail », « famille », « patrie » co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence

Damon Mayaffre

CNRS – UMR, Bases, Corpus et Langage (Nice)

Abstract

Meaning is born in context. Starting from this assumption, we shall first offer a theoretical definition of co-occurrence as *the minimal form of linguistic context* required for the apprehension of meaning.
In the second part, we conduct a practical demonstration showing that the minimal contextualisation of the word "travail" by means of its co-occurrents ("famille", "patrie", etc.) in Nicolas Sarkozy's rhetoric in 2007 leads to political semantic implications heavy with historical sense.

Résumé

Le sens naît en/du contexte. Dans ce cadre, nous définirons, de manière théorique dans une première partie, la co-occurrence comme *la forme minimale du contexte linguistique* nécessaire à l'appréhension du sens.
Dans une seconde partie, nous montrerons, de manière pratique, que la contextualisation minimale du mot « travail » par ses co-occurrents (« famille », « patrie », etc.), dans le discours électoral de Nicolas Sarkozy en 2007, permet d'entrer dans une sémantique politique lourde de sens historique.

Mots-clés : co-occurrence(s), collocation, contexte, contextualisation, herméneutique numérique, Sarkozy.

1. Introduction

Tout semble avoir déjà été dit sur le traitement des co-occurrences. De la bibliographie française à la bibliographie anglo-saxonne, des études de la communauté ADT soucieuses de rendre compte de la textualité aux études de la communauté TALN portées sur l'extraction d'information et le web sémantique, des travaux pionniers de (Firth, 1957) ou de Saint-Cloud à ceux actuels de ATST, BCL, DELIC, ICAR, ILPGA, UQAM, etc., de la thèse de (Lafon, 1984), de (Salem, 1993) ou de (Viprey, 1997) à celle de (Martinez, 2003) exclusivement dédiée à la question…, la littérature foisonne de considérations méthodologiques, de modèles mathématiques et informatiques, d'indices statistiques complémentaires ou concurrents, de représentations graphiques variées des co-occurrences. A ceci, ajoutons que la bibliographie se démultiplie encore dès lors que l'on veut bien associer à la notion stricte de co-occurrences celle proche, notamment dans le milieu anglo-saxon, de collocation[1].

Ce foisonnement méthodologique contraste avec le peu d'études effectives d'historiens du texte, de politologues du discours, de linguistes de la parole utilisant la co-occurrence dans leur démarche interprétative. Il en est sans doute la raison, et l'indice qu'aucune pratique co-

[1] Ainsi nous avons très rapidement rassemblé plus de 100 références françaises ou internationales. Dans ces conditions, la bibliographie présentée en fin d'article est arbitraire et ne sera considérée qu'à titre indicatif.

occurrentielle stable des textes ne s'est réellement imposée aujourd'hui dans la boite à outils des interprètes de corpus textuels.

Concrètement, les grands logiciels d'ADT directement accessibles sur le marché offrent finalement peu de fonctions opératoires pour traiter de la co-occurrence. Et, un des résultats du projet « Textométrie » (2007-2009) financé par l'ANR et mené par Serge Heiden (ENS-ICAR) doit être de rendre accessible à tous des pratiques réservées jusqu'ici à certains utilisant des outils performants mais rarement interfacés pour le grand public scientifique.

Cette contribution a deux objectifs.

Elle propose, d'une part, une étude d'un cas concret et suggestif : l'approche co-occurrentielle du discours électoral de Nicolas Sarkozy autour de la valeur « travail ». Cette étude s'appuie sur le corpus exhaustif des discours de meeting du candidat victorieux à la présidentielle entre le 1er janvier 2007 et la veille de son élection le 6 mai (34 discours, 283 109 mots, 11 689 vocables ; 621 occurrences du lemme-pôle « travail »)[2]. Les fonctions variées du traitement de la co-occurrence mises en œuvre sont celles désormais implémentées sur le logiciel d'Etienne Brunet, HYPERBASE qui s'est appliqué ces derniers mois à rassembler plusieurs outils complémentaires (calcul probabiliste classique des co-occurrences hérité de Saint-Cloud (Lafon, 1984), graphes des co-occurences à l'image des lexicogrammes de (Heiden, 2004), représentation AFC des co-occurrences proche de celle proposée par (Viprey, 1997), analyses arborées et représentations topologiques de conception niçoise (ici, Brunet, 2008).

Cette contribution propose, d'autre part, en préalable, quelques considérations théoriques sur des enjeux de la co-occurrrence non pas dans le TALN mais dans l'ADT actuelle tournée résolument vers une linguistique textuelle à vocation rhétorico-herméneutique.

2. La co-occurrence : enjeu actuel pour l'ADT

Une des préoccupations de la linguistique de corpus, de l'ADT et de la textométrie est de passer d'une approche occurrentielle des données du corpus à une approche co-occurrentielle. Cette préoccupation a été identifiée comme cruciale très rapidement (par exemple Tournier, 1980) et reste en 2008 toujours d'actualité.

Le passage d'une statistique occurrentielle – étude de la distribution fréquentielle d'un terme dans un corpus partitionné –, à une statistique co-occurrentielle – étude du rapport fréquentiel entre deux termes co-présents dans le corpus au sein de fenêtres co-textuelles délimitées (le paragraphe par exemple) – représente un saut. On passe en effet d'une approche formelle, nucléaire ou positiviste du corpus à une approche contextualisante c'est-à-dire déjà sémantique. Avec la co-occurrence, la statistique textuelle met un pied dans une sémantique de corpus qui lui était jusqu'ici interdite et réaffirme par là sa vocation herméneutique.

2.1. Les deux traditions linguistiques de François Rastier

En termes rastiriens, la recherche d'occurrences s'inscrit dans une linguistique de tradition logico-grammaticale, pour laquelle il existe des entités indépendantes – des mots, ici, pour faire simple – qui renvoient, *in abstracto* ou hors contexte, à des ontologies – des références–. Cette linguistique, selon Rastier, est une linguistique du signe et non du sens. La remise du

[2] A ce corpus central, ajoutons d'autres corpus en contrepoint : les discours de meeting de Laguiller, Buffet, Royal, Bayrou et Le Pen (voir *infra*). L'ensemble de ces discours sont disponibles sur le site de Jean Véronis : http://sites.univ-provence.fr/veronis/Discours2007/.

mot dans un certain contexte linguistique, *via* par exemple des concordanciers, prolongera certes l'analyse, mais en renonçant déjà à la statistique et au traitement contrôlé pour revenir à une lecture d'essence traditionnelle.

La recherche de co-occurrences (leur mise en évidence par des calculs plus ou moins complexes, puis leur mise en forme sous l'apparence de tableaux, d'arbres ou de graphes cf. *infra*) renvoie, elle, à des notions de parcours interprétatifs, de mises en réseaux ou de mise en résonance[3], de textualité ou de texture. Elle s'inscrit en tout cas, comme nous allons essayer de le montrer, dans *une logique de contextualisation* qui est la condition de l'élaboration du sens et de l'interprétation. Nous entrons par là même dans la deuxième linguistique définie par Rastier, de tradition rhétorico-herméneutique. En d'autres termes encore, en privilégiant les co-occurrences sur les occurrences, on réintroduit les contextes (c'est-à-dire les *unités textuelles* qui font sens tels les *passages* récemment théorisés par [Rastier, 2007]) dans une pratique ADT, sans cela, seulement lexicographique.

2.2. La co-occurrence comme forme minimale du contexte

Comme on sait, le sens naît du/en contexte ; la contextualisation (et sa formalisation) est ainsi la condition d'un traitement sémantique/interprétatif.

A une extrémité, la *forme maximale* du contexte serait *tout le texte* (sauf à être, au-delà du texte, le corpus dans son ensemble[4]). A l'autre extrémité, nous voulons poser ici que la *forme minimale* du contexte est la *co-occurrence*.

Nous définirons en effet la co-occurrence comme le phénomène de contextualisation minimale d'un mot *par* un autre mot. (Charge à la statistique et au traitement informatique de repérer systématiquement toutes les attirances/répulsions lexicales co-occurrentielles parlantes ou significatives ; et leur degré de significativité).

Au sein du corpus, le contexte minimal d'un mot-pôle n'est pas le syntagme ou la phrase. Ceux-ci sont trop nombreux et trop variables pour être synthétisables : il y aurait alors, en effet, autant de contextes (c'est-à-dire de sens) du mot que de syntagmes ou de phrases le contenant, et il ne nous resterait plus qu'à les éplucher un par un[5]. La démarche serait contreproductive puisque cette multiplication de contextes plutôt que nous mener au sens nous compliquerait son accès.

Loin du syntagme ou de la phrase donc, le contexte minimal (mais formalisable) d'un mot est son co-occurrent attesté par la statistique, ou plutôt ses co-occurrents lexicaux attestés, systématiquement repérés.

---

[3] Suggérons cette image : le traitement co-occurrenciel espère saisir l'instant où le vocabulaire d'un texte entre en résonance avec lui-même.

[4] Selon l'expression souvent répétée par Rastier : « Le contexte, c'est tout le texte ». Nous savons néanmoins qu'un texte reçoit aussi, à un niveau supérieur, des déterminations du corpus dans lequel il se trouve plongé. Techniquement, la « loi endogène » des traitements textométriques –ici du calcul de la co-occurrence– prend en considération l'ensemble du corpus comme la norme par rapport à laquelle s'individualisent des événements linguistiques (*voir infra*).

[5] De fait, dans la tradition saussurienne, (Rastier, 2002) rappelle que « le principe différentiel de la linguistique saussurienne, appliqué aux contextes et aux textes, permet de conclure que chaque occurrence est un hapax. » Mais nous entrerions ici dans une réalité inaccessible à la statistique.

Par là, l'enjeu du traitement co-occurrentiel devient majeur pour l'ADT : il s'agit, *en corpus*, d'objectiver par la statistique, le contexte minimal (mais essentiel) des mots nécessaire à leur compréhension/interprétation.

On notera ici au passage tout l'avantage de l'ADT dont le traitement s'applique à articuler macro-contexte et micro-contexte, contextualisation maximale et contextualisation minimale, *approche globale* et *approche locale*. En effet, le repérage des co-occurrences opère toujours, dans son principe, en deux temps, de la même manière. (i) Toutes les occurrences des mots a, b, c (soit leur fréquence n_a, n_b, n_c) sont d'abord considérées *dans la totalité du corpus* (*N*). Le corpus dans son ensemble constitue bien le cadre (statistique et linguistique) qui fait sens : le macro-contexte des mots est le corpus. (ii) Puis les mots sont replacés dans de micro-contextes ou contextes locaux – le paragraphe, une fenêtre paramétrable, la phrase, etc. – pour y repérer les attirances (ou répulsions) lexicales saillantes. Ajoutons encore que selon les modèles statistiques proposés, l'articulation entre appréhension globale et appréhension locale du corpus va plus loin, au cœur même des calculs, puisque (iii) les attirances lexicales saillantes (entre a et b par exemple) s'apprécient, précisément, par la comparaison des fréquentations observées *localement* ($n_{(a+b)}$ *au sein du paragraphe*) d'une part et celles espérées *globalement* d'autre part, au regard de la fréquence totale des mots dans le corpus entier (n_a et n_b *au sein du corpus*)[6].

2.3. La co-occurrence comme nœud du tissu textuel

L'étymologie de « texte » est souvent rappelée. Tisser, tissus, tissage : on convient généralement qu'un texte est un entrelacement de deux fils ou deux axes, vertical et horizontal.

Le fil de chaîne, horizontal, représenterait la linéarité du texte ou l'axe syntagmatique ; il rendrait compte du déroulement et des *combinaisons* mises en place par le locuteur pour produire un texte ; fondamentalement, pour notre propos, cet axe renvoie à une logique de contextualisation puisque le mot est considéré *in praesentia* de son environnement linguistique naturel immédiat (le syntagme, la séquence, la phrase, la suite, la période, l'extrait, le passage, etc. : souvent appelés génériquement le *cotexte immédiat*).

Le fil de trame, lui, serait l'axe vertical ou axe paradigmatique qui représenterait la dimension non-linéaire du texte. Il mettrait à jour, sous forme de nomenclature, la *sélection* (sélection lexicale par exemple) opérée par le locuteur. Classiquement, nous avons alors à faire à *l'inventaire* des formes utilisées, au *dictionnaire* alphabétique du corpus, à *l'index* hiérarchique, à telle *liste* ou à tel *tableau* des spécificités, d'hapax, etc. Ici la décontextualisation aura présidé à l'analyse pour rendre les informations de corpus synthétisables en paradigme ; les mots sont considérés *in absentia* de leur environnement linguistique, la plupart du temps, en ADT, selon leur attribut quantitatif (fréquence, sous-fréquence, probabilité d'utilisation, etc.).

C'est dans ce cadre que les travaux d'ADT les plus récents cherchent à embrasser le texte dans ses deux dimensions. L'ADT, historiquement pertinente pour traiter l'axe paradigmatique, cherche aujourd'hui en effet à réintroduire la progression, le rythme et la structure linéaire du texte. Le concept de « topologie textuelle » développé par [(Mellet et Barthélemy, 2007) ou ici même (Longrée *et al*., 2008)] est une piste désormais balisée.

[6] Notre propos n'est pas d'entrer dans le détail du traitement statistique, mais la plupart des modèles reprennent dans son esprit cette mise en comparaison : (Lafon, 1984), (Salem, 1993), (Viprey, 1997), (Véronis, 2003), etc.

C'est dans ce cadre surtout que la co-occurrence peut être considérée comme un noeud essentiel du tissu-texte ; l'endroit même où se noue (et se dénoue pour l'analyste) le fil de trame et le fil de chaîne, l'axe paradigmatique et l'axe syntagmatique d'une production textuelle, la tabularité et la linéarité du texte.

L'approche ADT articule en effet dans le traitement co-occurrentiel un va et vient entre sélection et combinaison, entre décontextualisation et (re)contextualisation. La statistique met à jour une *sélection* lexicale : elle repère systématiquement par exemple les substantifs qu'un locuteur aura sur-utilisés (sur-sélectionnés) pour les considérer comme les mots-pôles à traiter. Mais les mots de la sélection seront considérés dans le cadre d'une *combinaison* ou d'une fenêtre syntagmatique données. Le traitement statistique des co-occurrences procède à la fois d'une décontextualisation lexicale (le mot *extrait du* corpus) et d'une (re)contextualisation (le mot *replacé dans* son paragraphe). Mieux : le traitement co-occurrentiel *produit* lui-même un effet paradigmatique (classiquement il aboutit à une liste des termes co-occurrents du mot-pôle, classés par ordre hiérarchique par exemple) et un effet syntagmatique : comme indiqué précédemment, il aboutit à la mise à jour ou à la définition du *contexte minimal* (ou combinaison minimale) du mot-pôle en question.

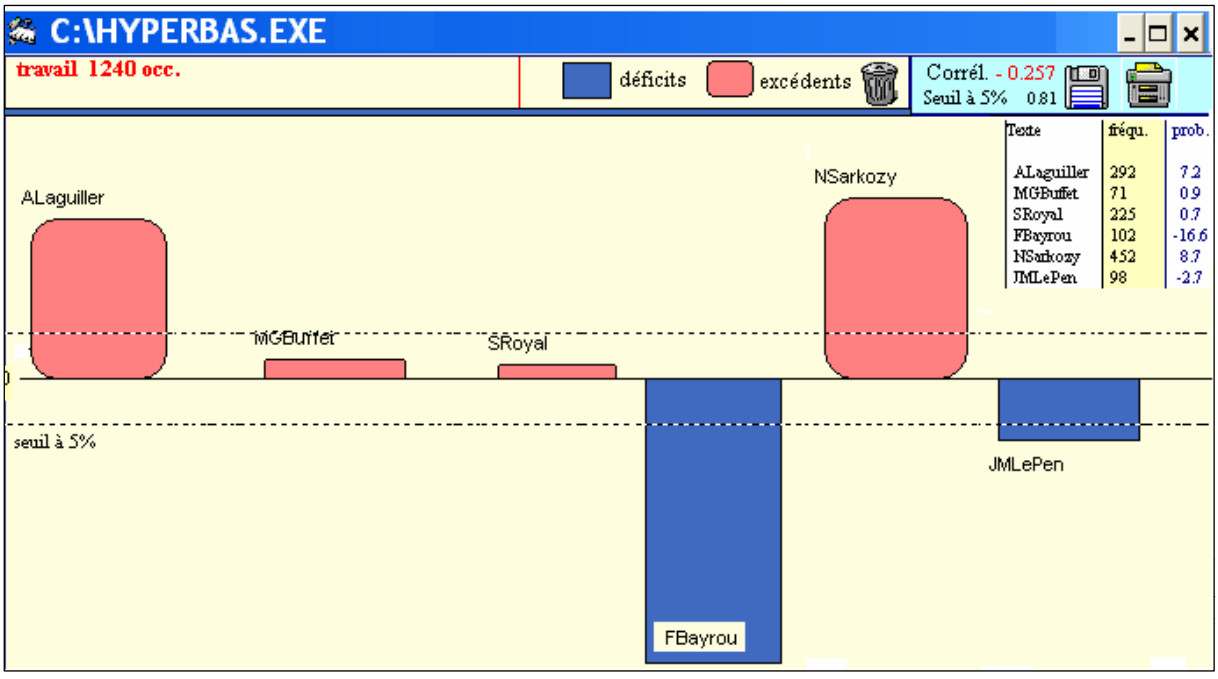
3. Co-occurrences de « travail » chez Sarkozy

Un traitement textométrique classique – que nous appelons occurrentiel –, sur le corpus partitionné de la campagne électorale 2007 (Laguiller / Buffet / Royal / Bayrou / Sarkozy / Le Pen), s'avère le plus souvent précieux. L'étude contrastive de la fréquence des mots, des lemmes, des catégories grammaticales ou autres unités linguistiques pertinentes dans le corpus, nous a permis de décrire par exemple, les grandes particularités du parler de Sarkozy (voir ailleurs : Mayaffre, 2007).

3.1. Aux limites du fréquentiel

Mais ce traitement touche parfois, jusqu'au discrédit, ses limites. La distribution du terme « travail » chez les différents candidats se présente en effet selon la figure 1.

Figure 1 : Distribution de « travail » dans la campagne électorale 2007[7]



Sur le graphique, Nicolas Sarkozy et Arlette Laguiller se rapprochent voire se confondent par la sur-utilisation massive et égale du mot « travail » dans leurs discours. Pour l'un comme pour l'autre, le traitement statistique indique que « travail » est une *spécificité* positive dans des proportions voisines ici mesurées en écart réduit (respectif de +8 et +7).

Dès lors, on comprend que la comparaison entre le discours de la candidate d'extrême gauche et le candidat qui séduira l'électorat d'extrême droite doit s'opérer non pas sur la fréquence d'utilisation du mot mais sur son type d'emploi c'est-à-dire ses contextes d'utilisation.

3.2. Du contexte minimal de « travail »

La contextualisation minimale de « travail » que réalise le traitement des co-occurrences dans le discours de Sarkozy et de Laguiller est instructive voire suffisante pour engager un processus interprétatif.

Le premier outil, fort connu, que propose le logiciel HYPERBASE est inspiré de l'approche saint-clousienne (Lafon, 1984 ; Lebart et Salem, 1994). Il s'agit d'extraire du corpus les passages contenant le mot-pôle choisi pour les constituer en sous-corpus, puis de comparer ce sous-corpus d'étude au corpus entier, pour en repérer les *spécificités* lexicales. Cette opération successivement réalisée pour les contextes-paragraphes de « travail » chez Sarkozy puis chez Laguiller apparaît spectaculaire car dans le début des deux listes (expurgées des mots outils) AUCUN co-occurrent n'est commun aux deux candidats ! (Figure 2)

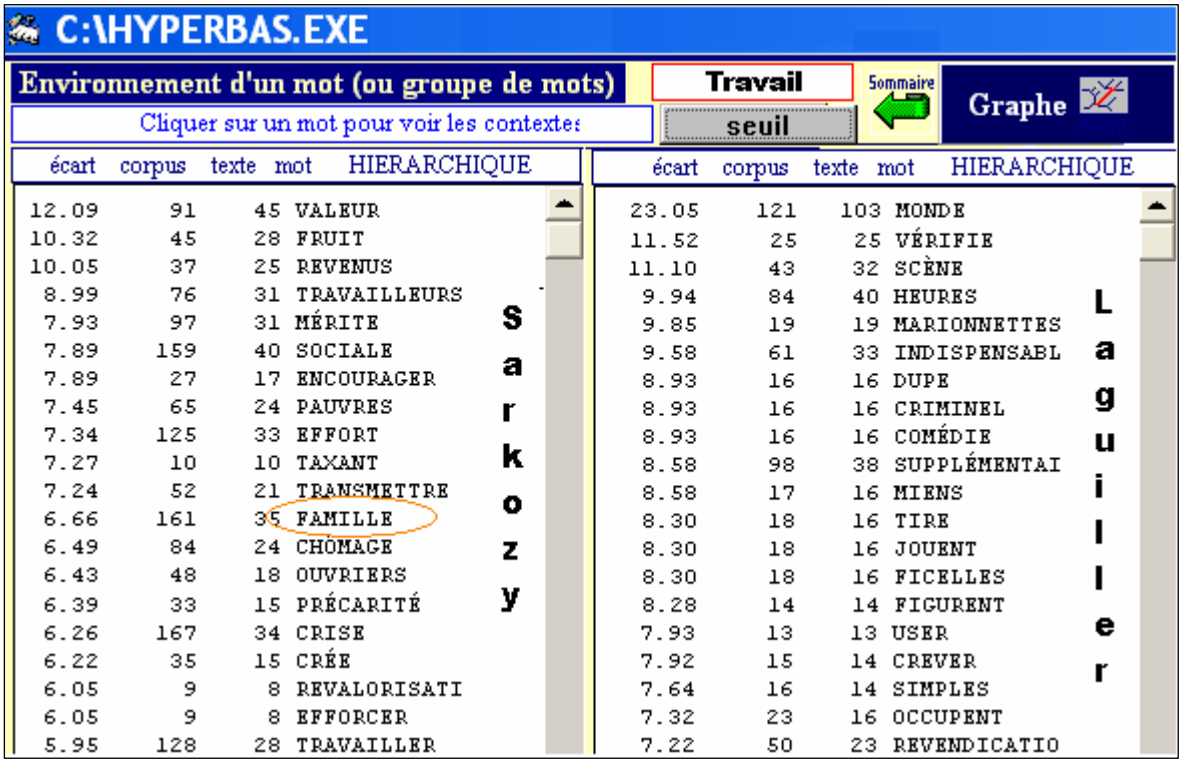


Figure 2 : co-occurrents de « travail » chez Sarkozy et Laguiller

Dans la *trame du texte*, le mot est donc utilisé dans des proportions voisines par Sarkozy et Laguiller, mais dans la *chaîne du texte*, l'environnement lexical du mot n'a rien de commun pour les deux candidats. Si Laguiller et Sarkozy *sélectionnent* « travail » à l'identique (i.e : dans les mêmes proportions) pour produire leur discours, ils le *combinent* ou l'articulent à d'autres termes différemment.

A la vue des co-occurrents privilégiés, une analyse rapide permet d'interpréter grossièrement que Sarkozy s'applique à *mythifier* le « travail » en l'associant notamment à « valeur », « mérite », « effort », « fruit », lorsque Laguiller s'applique à *démystifier* le discours dominant sur le travail en associant le terme à un vocabulaire théâtral (« scène », « ficelles », « jouent », « dupe ») et en prétendant incarner la réalité d'un « monde » qu'elle serait seule à connaître[8].

3.3. Co-occurrences généralisées (Viprey, 1997) autour de « travail »

Le discours de Sarkozy sur le travail – discours sarkozien que nous considèrerons seul désormais, et dans le corpus exhaustif des meetings de campagne, deuxième tour compris – nous parait plein, au sens où Sazkozy traite la question à plusieurs niveaux.

Après avoir repéré les co-occurrents principaux de « travail » (figure 2), il est possible de mesurer l'organisation des co-occurrents entre eux, pour entrer toujours plus finement dans l'entrelacement lexical et déterminer des contextes minimaux qui font sens (figure 3).

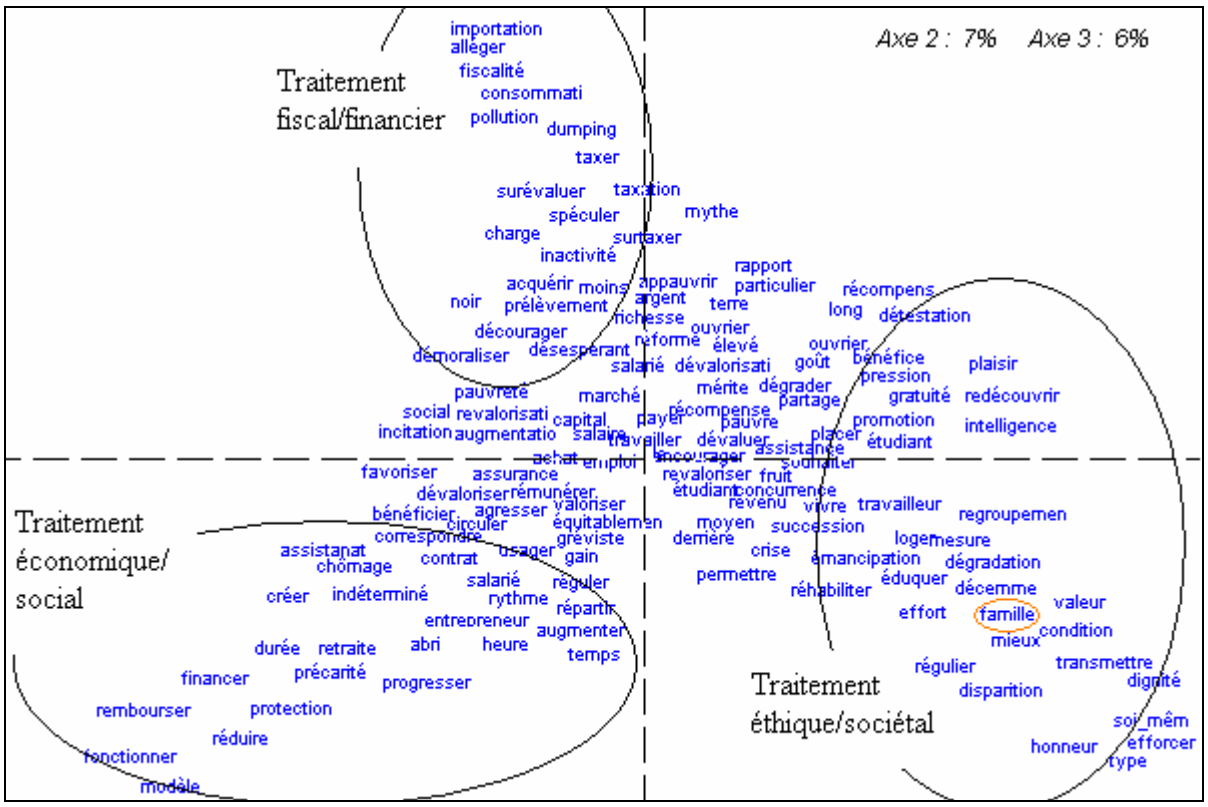


Figure 6 : co-occurrences généralisées autour de « travail »

Sur le modèle présenté par (Viprey, 1997) et repris dans plusieurs articles (Viprey, 2005 et 2006), HYPERBASE établit une matrice carrée qui croise tous les co-occurrents de « travail »

[8] Nous laissons là l'analyse de « travail » chez Laguiller. On remarquera par exemple qu'elle n'associe pas (contrairement à Sarkozy) particulièrement « travail » à son mot-signature « travailleurs ». Nous tenons ici une piste essentielle de son discours : son propos entier est adressé aux « travailleurs ». Le « travail », le « monde » du « travail », les « travailleuses » et les « travailleurs » ne sont pas un thème parmi d'autres mais le périmètre global de son discours partout présents mais/donc jamais traités spécifiquement. Lorsque le thème du travail est finalement abordé de front, c'est uniquement par le biais de la dénonciation du discours de l'autre pour révéler le jeu de « dupe » et, dans une tonalité prolétarienne, les difficultés de la tâche (« crever », « user », « criminel »).

entre eux pour mesurer leur attraction généralisée[9]. La représentation AFC d'un telle matrice (figure 3) permet ainsi de voir que Sarkozy organise les co-occurrents de « travail » en trois pôles lexico-thématiques distincts : il effectue un traitement économique et social de la question (travail => « rembourser », « protection », « entrepreneur », « heure », « salarié », « chômage », etc.), un traitement fiscal et financier (travail => « fiscalité », « alléger », « taxer », « surévaluer », etc.), enfin un traitement beaucoup plus ambitieux, idéologique pourrait-on dire, éthique ou sociétal (travail => « valeur », « effort », « famille », « émancipation », « éduquer », etc.).

3.4. « Travail » et « famille »

Dans la liste présentée en figure 2, comme au cœur de l'AFC en figure 3, l'analyste du discours politique ne pourra pas ne pas remarquer, dans une intertextualité historique chargée de sens, l'articulation du mot « travail » avec le mot « famille ». Que « travail » et « famille » occurrent (séparément) dans le discours de Sarkozy, cela n'étonnera guère tant ces termes sont présents dans tout programme politique, mais que les deux mots co-occurrent au sein des paragraphes et l'interprétation devient différente. Répétons par là même que si l'approche des co-occurrences est pour le TALN le plus souvent dominée par une volonté de désambiguiser les homographes ou de rechercher l'information (Véronis, 2003), elle renvoie, pour nous, en ADT, à la volonté d'offrir à l'analyste des parcours de lecture aux vertus herméneutiques afin de mieux interpréter les textes.

La co-occurrence constitue donc une contextualisation minimale nécessaire – elle balise le parcours interprétatif – mais, en dernier recours, seul le retour au texte permet l'interprétation. La convocation des passages, par simple clic dans HYPERBASE, où « travail » et « famille » co-occurrent est très parlante[10].

Certains passages apparaîtront assez neutres, s'il n'y avait cette association immédiate de deux valeurs pourtant de registres différents :

> Je crois au TRAVAIL et je crois à la FAMILLE.[11]

Mais d'autres nous renvoient directement à des discours entendus ailleurs et autrefois :

> J'ai voulu parler de l'identité nationale […] parce qu'il était interdit d'en parler sous peine d'être excommunié au nom de la pensée unique et du politiquement correct, comme il était interdit de parler de l'autorité, de la morale, de la FAMILLE ou de la valeur TRAVAIL. J'ai voulu parler de la France parce que depuis trop longtemps elle était dénigrée et parce qu'à force de l'abîmer, à force de l'abaisser, à force de renier son histoire, sa culture, ses valeurs, à force de tout détester, de détester la FAMILLE, la patrie*, la religion, la société, le TRAVAIL, la politesse, l'ordre, la morale, à force on finit par se détester soi-même.[12]

---

[9] Ici le mot-pôle *A* (« travail ») sert d'une part à constituer un corpus d'étude (C_A : les paragraphes dans lesquels *A* se trouve) et d'autre part à identifier une liste de mots (les co-occurrents de *A : B, C, D*, etc.). Mais désormais, c'est l'organisation « interne » des co-occurrents, entre eux, qui nous intéresse c'est-à-dire les (sous)-co-occurrences de *B* et *C*, de *C* et *D*, etc. à l'intérieur du corpus C_A (Viprey, 1997 et Brunet, 2008).

[10] Précisons, en trois temps, la valeur de ces citations. (i) le calcul des spécificités montre un sur-emploi de « travail » chez Sarkozy (fig. 1). (ii) Un traitement des co-occurrences montre que « famille » est un des co-occurrents majeurs de « travail » (fig. 2). (iii) Nous convoquons alors tous les passages qui contiennent les mots « travail » et « famille », persuadé qu'il s'agit de passages non anecdotiques du texte sarkozien.

[11] N. Sarkozy, meeting d'Issy-les-Moulineaux, 18 avril 2007 (répété au meeting de Rouen le 24 avril).

[12] N. Sarkozy, meeting de Tours, 10 avril 2007 (répété au meeting de Marseille, 19 avril 2007).

Ou encore :

> Oui, à force de tout détester, la FAMILLE, la patrie*, la religion, la société, le TRAVAIL, la politesse, la courtoisie, l'ordre, la morale. A force de tout détester, on finit par se détester soi-même. Beau résultat ![13]

Sarkozy sait tenir un discours moderne ou effectuer, parfois, des ouvertures à gauche (Mayaffre, 2007), mais nous voyons ici qu'il sait dialoguer avec le discours de l'extrême droite maurassienne ou vichyste. Poser, ensemble, dans des énumérations, la « famille » et le « travail » comme « valeurs cardinales »[14] (auxquelles sont adjoints comme dans ces citations et comme nous le verrons, la « patrie » ou encore la « religion », « l'ordre », la « morale ») produit un sens politique au-delà même des revendications idéologiques explicites. C'est la mise en résonance de « travail » par « famille » (et vice-versa) qui ici fait sens dans le corpus et oriente l'interprétation.

3.5. « Travail », « famille »… et « patrie »

Dans une démarche itérative enfin, après avoir constaté que « travail » était contextualisé par « famille », il est possible de rechercher, selon la même méthode, les co-occurrences qui contextualisent « famille ». Il s'agit là d'une logique en cascade dans laquelle la plupart des auteurs se sont laissés entraîner jusqu'à (Martinez, 2003) qui théorise la recherche des *poly-cooccurrences*[15].

écart	corpus	texte	mot	HIERARCHIQUE
11.21	711	80	enfant_2	
9.73	41	20	loger_1	
8.27	51	18	allocation_2	
7.43	60	17	familial_3	
6.86	38	13	obligation_2	
6.46	75	16	revenu_2	
6.28	32	11	immigré_2	
6.27	25	10	occuper_1	
5.54	14	7	entasser_1	
5.34	62	12	père_2	
5.23	11	6	scolariser_1	
5.18	621	41	travail_2	
5.11	7	5	traîner_1	
4.94	29	8	mère_2	
4.89	51	10	excellence_2	
4.77	9	5	éduquer_1	
4.62	46	9	mesure_2	
3.92	154	14	devoir_2	
3.66	61	8	situation_2	
3.40	7	3	politesse_2	
3.20	273	17	société_2	
3.19	62	7	difficulté_2	
3.00	52	6	religion_2	
2.73	44	5	patrie_2	
2.70	296	16	travailler_1	

Figure 4 : Co-occurrences de « famille » dans le discours de Sarkozy

[13] N. Sarkozy, meeting de Lyon, 5 avril 2007.

[14] N. Sarkozy, meeting en Guadeloupe, 22 mars 2007.

[15] Une solution plus contrainte est aussi possible. Prélever les paragraphes contenant et « travail » et « famille » pour étudier les mots sur-représentés dans ce sous-corpus.

De proche en proche (un mot-pôle => son co-occurrent => le co-occurrent du co-occurrent => etc.), le cheminement pourrait prétendre épuiser tout le texte (Tournier et Heiden, 1998). Plus modestement, l'objectif est d'établir des réseaux lexicaux de plusieurs degrés pour établir des faisceaux isotopiques et rendre compte de la textualité dans une certaine épaisseur. Ici, le parcours de lecture en trois temps ou trois degrés qui part de « travail », transit par « famille » et aboutit à « patrie » nous semble un parcours interprétatif, guidé par la statistique, particulièrement suggestif du sens à donner aux propos du candidat.

L'organisation visuelle de ces parcours co-occurrentiels passe par l'établissement de graphes de co-occurrences. (Heiden, 2004) ou (Véronis, 2003) ont présenté des modèles très construits. Le graphe que propose HYPERBASE est plus simple puisqu'il met en scène seulement le mot-pôle, ses principaux co-occurrents, ainsi que les co-occurrents des co-occurrents. Nous l'illustrons – à rebours du chemin emprunté jusqu'ici – avec le terme « patrie » (figure 5).

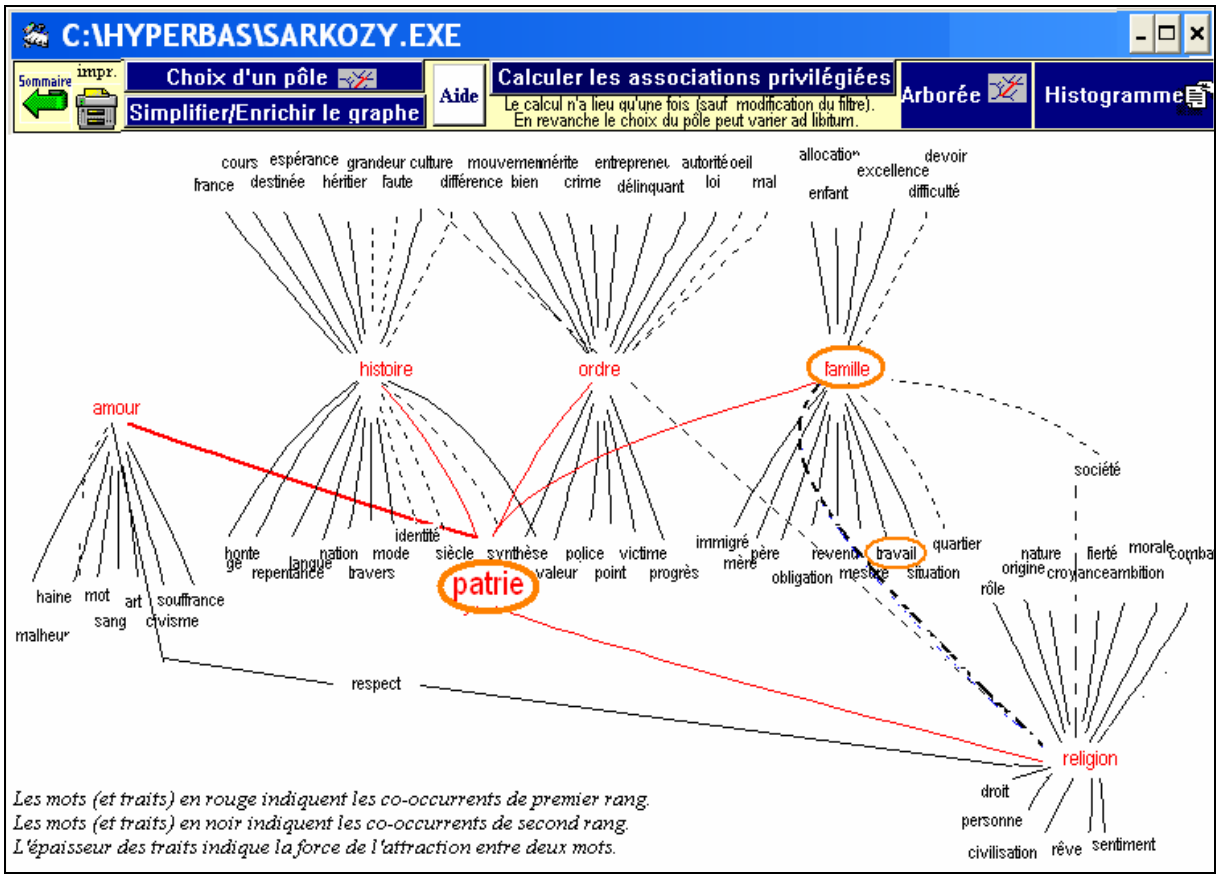


Figure 5 : graphe de co-occurrents à partir du mot-pôle « patrie » chez Sarkozy

De manière générale, « patrie » a 5 grands co-occurrents[16] qui marquent, si l'on veut bien considérer leurs co-occurrents respectifs, 5 dimensions du discours sarkozien : une dimension pathétique (« patrie » => « amour » => « sang », « haine », « souffrance », etc.) ; une dimension historique/patriotique (« patrie » => « histoire » => « France », « grandeur »,

[16] En réalité, les co-occurrents dépassant le seuil statistique sont bien plus nombreux. Nous les avons réduits aux cinq substantifs majeurs pour ne pas encombrer le graphique. De la même manière, les substantifs ont été privilégiés pour les co-occurrents de second rang. Ne cachons pas que la mise en graphe, pour des raisons techniques, réclame toujours des sélections problématiques ; sans rien dire des choix sémiotiques mis en oeuvre.

« destinée » etc.) ; une dimension politique/autoritaire (« patrie => « ordre » => « délinquant », « crime », « police », etc.) ; une dimension familiale (« patrie » => « famille » => « père », « mère », etc.) ; et une dimension religieuse/spirituelle (« patrie » => « religion » => « croyance », « rêve », « sentiment », etc.).

Pour le débat historico-discursif qui nous intéresse, le graphe propose un parcours de lecture en trois temps : « patrie » => « famille » => « travail » ou en sens inverse « travail » => « famille » => « patrie ».

4. Conclusion

La co-occurrence est un sujet complexe. Nombre de questions n'ont pu être abordées. La première d'entre-elle est, comme toujours en ADT, la question des unités linguistiques traitées : les lemmes sont souvent utilisés dans la recherche des co-occurrences afin de limiter les entrées, mais l'on objectera ici comme ailleurs que la lemmatisation consiste à projeter le sens dans le texte là où la recherche des co-occurrences se proposait de le chercher ; il y aurait là un vis de forme dans le procès de la démonstration.

La deuxième question porte sur la taille du co(n)texte, c'est-à-dire la fenêtre naturelle ou artificielle d'étude. La phrase et le paragraphe sont en général privilégiés. Pourtant l'avantage du traitement statistique est qu'il peut s'affranchir d'unités qui n'ont de naturelles que le mot. La question rebondit sur l'orientation de la recherche des co-occurrences à l'intérieur de la fenêtre (co-texte droit *versus* co-texte gauche), ou encore sur la prise en considération de l'empan existant entre deux co-occurrents (empan étroit voire contigu qui nous renvoie à des unités phraséologiques *versus* empan large qui renvoie à des corrélats). Ici rappelons le principe : la recherche est statistique : elle ne gagne pas, dans un mélange des genres, à s'encombrer de considérations grammaticales, phrastiques, syntaxiques ou distributionnelles.

Enfin, sans prétendre être exhaustif, la dernière question interroge la pertinence des calculs proposés : nous n'avons pas cherché à arbitrer les formules disponibles sur le marché scientifique (Rapport de Vraissemblance de Dunning, Information Mutuelle de Church, indice de Lafon, etc.) et avons utilisé le plus souvent le calcul hypergéométrique des cooccurrences désormais implémenté dans HYPERBASE (Brunet, 2006 et 2008).

L'objectif de cette contribution était ailleurs. Nous avons abordé la question par le biais de l'Analyse des données textuelles (et non du TALN) qui doit offrir, dans le cadre d'une linguistique des textes, des parcours de lecture susceptibles de nourrir l'interprétation.

Dans cet horizon, l'enjeu de l'ADT et de la textométrie est de ne pas se laisser enfermer dans une démarche purement lexicographique pour proposer des perspectives lexicologiques. Si la lexicographie est décontextualisante, la science lexicologique passe par une démarche contextualisante. Plus loin, si l'objectif de la textométrie est de rendre compte du texte dans sa complexité et non pas seulement de lexies nucléaires, cela passe par la prise en compte d'*unités textuelles* qui ne peuvent s'arrêter à la frontière du mot ni à celle, déjà reculée, d'enchaînement de mots comme les segments répétés.

Même si cela n'est pas intuitif, la co-occurrence est cette unité textuelle élémentaire, ce contexte minimal, producteur de sens et matrice d'interprétation.

Références

Brunet E. (2006). Navigation dans les rafales. In Viprey J.-M. (éd.), *JADT'06*. Presses Universitaires de Franche-Comté, pp. 15-29.

Brunet E. (2008). Les séquences (suite). In *Actes des JADT 2008*.

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

Ferret O. (2004). Discovering word senses from a network of lexical cooccurrences. In *Actes TALN 2004* (Fès).

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

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. In Purnelle G. (éd.), *JADT 2004, Le poids des mots*. Presses universitaires de Louvain, pp. 577-588.

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

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

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

Longrée D., Luong X. et Mellet S. (2008). Les motifs : un outil pour la caractérisation topologique des textes. In *Actes des JADT 2008.*

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, sous la direction d'André Salem, Paris.

Mayaffre D. (2007). Vocabulaire et discours électoral de Sarkozy : entre modernité et pétainisme. *La Pensée*, 352.

Mellet S. et Barthélemy J.-P. (2007). La topologie textuelle : légitimation d'une notion émergente. *Lexicométrica*, numéro thématique (http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf.)

Rastier F. (2001). *Art et science du texte*. Puf.

Rastier F. (2002). Saussure, la pensée indienne et la critique de l'ontologie. *Revue de sémantique et de pragmatique*, 11 : 123-146.

Rastier F. (2007). Passages. *Corpus*, 6 : 25-54.

Salem A. (1993). *Méthodes de la statistique textuelle*. Thèse d'Etat - Paris 3.

Tournier M. (1980). D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles. *Mots*, 1 : 189-209.

Tournier M. et Heiden S. (1998). Lexicométrie textuelle, sens et stratégie discursive. In Actes *I Simposio Internacional de Análisis del Discurso*.

Véronis J. (2003). Cartographie lexicale pour la recherche d'information, *Actes de TALN 2003*, pp. 265-274.

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

Viprey J.-M. (2005). Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus. In A. Condamines (dir.), *Sémantique et corpus*. Lavoisier, pp. 245-276.

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