Au-delà de la cooccurrence binaire... Poly-cooccurrences et trames de cooccurrence

William Martinez

Édition électronique

URL : <http://journals.openedition.org/corpus/2262>

DOI : 10.4000/corpus.2262

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 1 janvier 2012

ISSN : 1638-9808

Référence électronique

William Martinez, « Au-delà de la cooccurrence binaire... Poly-cooccurrences et trames de cooccurrence », *Corpus* [En ligne], 11 | 2012, mis en ligne le 18 juin 2013, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/2262> ; DOI : <https://doi.org/10.4000/corpus.2262>

Au-delà de la cooccurrence binaire... Poly-cooccurrences et trames de cooccurrence

William MARTINEZ

Instituto de Linguística Teórica e Computacional (ILTEC)

Rua Elias García, n°147, 5°D

1150-199 Lisbonne PORTUGAL

Introduction

Alors qu'elle se propose de dévoiler la structuration lexicale des textes, l'approche lexicométrique trahit d'emblée la réalité textuelle dans ce qu'elle a de fondamental : sa 'séquentialité' syntagmatique. En effet, au terme d'une *segmentation* – opération initiale qui vise à identifier les composantes lexicales du corpus étudié – la statistique textuelle produit un inventaire hiérarchisé qui classe les formes lexicales suivant leur fréquence d'apparition. Or, même si cette liste permet par des calculs contrastifs de mesurer la distribution du vocabulaire dans le discours et d'y repérer les ventilations irrégulières qui traduisent des préférences d'emploi, la décontextualisation des lexies crée une rupture que la statistique seule ne peut compenser.

A partir de cet état désarticulé du texte qu'est l'inventaire lexical, les méthodes lexicométriques se mobilisent avec un objectif double : hiérarchiser les unités repérées suivant leur originalité fréquentielle et les ordonner sous une forme articulée qui se rapproche du texte naturel. En effet, depuis les méthodes ouvertement cooccurentielles comme les *concordances* qui extraient les lignes de contexte d'apparition d'un pôle, les *segments répétés* et les *inventaires distributionnels* (Salem [1987]) qui hiérarchisent les associations lexicales contiguës, jusqu'aux approches multidimensionnelles (Lebart & Salem [1994]) comme l'Analyse Factorielle ou la Classification Hiérarchique qui, opérant comme des analyses exhaustives des similarités distributionnelles, esquissent au final une typologie des conco-

Corpus n°11, *La cooccurrence : du fait statistique au fait textuel*
(2012), 191-216

mitances quasi-contextuelles, on peut concevoir que toutes les approches lexicométriques visent *in fine* à restituer les structures cooccurentielles qui caractérisent le corpus étudié.

Pourtant, bien qu’omniprésente, cette *statistique des rencontres* ne parvient que rarement à combiner efficacement un classement probabiliste avec une désambiguïsation par association syntagmatique. Dans cet article, nous proposerons une méthodologie originale pour la mise en évidence des cooccurrences lexicales complexes¹, et, à partir d’un corpus textométrique qui réunit les professions de foi des candidats aux élections présidentielles françaises de 2007², nous montrerons comment l’analyse multi-cooccurentielle permet de dévoiler des réseaux de mots qui restituent, par degrés de virtualité, les systèmes lexicaux structurant le texte.

1. Cooccurrences spécifiques

La segmentation initiale du texte produit un dictionnaire des fréquences, ensemble clos et adynamique qui réunit un lexique isolé de son contexte et abstrait de ses conditions particulières d’emploi. Pourtant, on constate, en excluant les mots-outils³, que les formes les plus employées – *pays*, *travail*, *entreprises*, *sociale*, etc. – esquissent la dimension programmatique des déclarations, tandis que l’importance numérique des pronoms – *je*, *nous* et *vous* – révèle des textes essentiellement dialogiques. Et, même si cet inventaire ne fournit aucune information concrète sur la cooccurrence des formes, on peut y deviner des concomitances de deux, trois ou plus de ces formes dans les 1 114 phrases du corpus (ex. *politique* + *sociale*, *services* + *publics* et *je* + *veux* + *vous*).

1 Toutes les analyses cooccurentielles présentées ici ont été réalisées avec *CooCS* (www.williammartinez.fr/coocs) outil informatique pour formaliser et visualiser les réseaux de mots associés. Ce logiciel opère sur tout corpus constitué en base textuelle avec le programme *Lexico3* www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW.

2 Le corpus *Foi2007* compile les programmes politiques des douze candidats à l’élection présidentielle du 22 avril 2007. Segmenté en 1 114 phrases, le texte compte 20 078 occurrences pour 3 770 formes et 2 078 hapax.

3 Cette catégorie inclut les signes dont la fonction de dénotation est nulle ou dérivée d’un autre signe (articles, conjonctions, prépositions, etc.).

En automatisant cette exploration en contexte et en objectivant les associations inférées par un indice probabiliste, les méthodes de cooccurrence⁴ produisent typiquement des listes de formes associées deux à deux et hiérarchisées par un indice signalant le degré de probabilité de la *cooccurrence lexicale*⁵. On calculera ainsi que le pôle *je*, dans ses 142 phrases d'occurrence, apparaît très souvent avec *veux* (spécificité⁴ +24), mais aussi avec *que* (sp. +9), ainsi qu'avec *vous* (sp. +8).

A nouveau, on peut inférer des attractions lexicales au-delà de ce qui est garanti par l'inventaire lexicométrique et entrevoir une cooccurrence complexe telle que *je+veux+que+vous*. Toutefois, si le calcul confirme nos inférences initiales concernant les associations binaires, la méthode des cooccurrences ne dévoile qu'une infime partie des ensembles complexes opérant en contexte. En effet, notre assemblage intuitif de binômes qui alludent à des systèmes complexes reste une opération spéculative et laisse une part trop importante à l'inférence. Il faut donc étendre la mesure des attractions lexicales au-delà de la cooccurrence binaire...

2. Poly-cooccurrences

Une première approche de la cooccurrence complexe qui permet de dévoiler les systèmes lexicaux à l'œuvre dans le texte vise à déterminer, au-delà des associations binaires, des attractions multiples et simultanées en contexte : les *poly-cooccurrences*.

4 Parmi les méthodes les plus connues on citera le *Test de Significativité des Cooccurrences* de Beauchemin et Cucumel [1995], l'*Information Mutuelle* de Church et Hanks [1990] et les *Méthodes des Cooccurrences* de Lafon [1984]. Inspirée des travaux de Tournier [1988] et leur développement chez Heiden et Lafon [1998], notre méthode exploite le Modèle Hypergéométrique afin d'opérer une comparaison entre la totalité du corpus et l'ensemble des contextes porteurs du pôle étudié afin de dégager les originalités de cet échantillon sous la forme de mots qui y sont sur-employés ou à l'inverse sous-employés. Le calcul produit un indice signalant l'écart entre fréquence réelle et fréquence attendue : +3 signale une probabilité de 1/1000, +4 une probabilité de 1/10000, etc.

5 On adopte ici une acception ouverte du phénomène en tant que 'groupe de mots apparaissant fréquemment ensemble dans une fenêtre contextuelle donnée' – conception purement empirique qui écarte toute contrainte linguistique (contiguïté, orientation, etc.) et permet de saisir la réalité cooccurrence de manière exhaustive.

2.1 *Elaboration*

Les poly-cooccurrences résultent d'un calcul étendu des cooccurrences spécifiques. Succinctement dit, il s'agit, à partir d'un premier constat de coïncidence contextuelle, de réitérer l'analyse sur la base du système cooccurentiel déjà détecté et ce jusqu'à épuisement des originalités lexicales dans les contextes explorés⁶. Pour comprendre le processus calculatoire dont il est question, considérons l'extrait présenté au tableau 1 qui montre comment, au fur et à mesure de l'exploration, l'algorithme de prospection est amené à statuer sur l'originalité lexicale d'un échantillon contextuel de plus en plus réduit (les phrases où apparaît *je*, puis celles où apparaissent concurremment *je* et *veux*, etc.).

Comme l'illustrent ces exemples, les poly-cooccurrences, bien que filtrées à l'extrême⁷, reproduisent l'essentiel du message véhiculé en sanctionnant la récurrence en contexte d'une chaîne de cooccurrence. L'analyse de cette logique d'association complexe produit une approximation reconstructive qui résout les indéterminations des mots isolés. En effet toute association d'une forme avec une poly-cooccurrence déjà établie est une construction sémantique dont bénéficient à la fois l'élément ajouté et le système initialement repéré. En proposant une version du texte qui est à la fois ordonnée statistiquement et explicitée quasi-contextuellement, les poly-cooccurrences dévoilent des réseaux de cooccurrence authentiques autour d'un pôle et matérialisent

6 Schématiquement, là où la méthode des cooccurrences détecte une attraction significative entre les formes A et B puis entre A et C, la méthode poly-cooccurentielle met en évidence une association simultanée entre A, B et C. (Cf. Martinez [2003]).

7 Le résumé poly-cooccurentiel implique une perte d'information : l'ordre syntagmatique des mots est abstrait de l'analyse (puisque c'est l'ordre statistique qui prime dans la mise au jour des chaînes) et les systèmes extraits du corpus se trouvent dépouillés de plusieurs catégories de formes. En effet, les mots-outils (Cf. note 4), considérés comme vides de sens, sont écartés de l'analyse, et, les seuils en vigueur pour le calcul élaguent les cooccurents jugés trop faibles par co-fréquence et par spécificité. Par ailleurs, durant la phase de prospection contextuelle, les poly-cooccurrences déjà détectées sont écartées afin d'éviter l'exploration répétée des mêmes sous-ensembles contextuels : schématiquement, une fois le chemin de cooccurrence *ABCD* comptabilisé, on ignorera la détection des poly-cooccurents *ABDC*, *ACDB*, etc.

les contraintes sémantiques et syntaxiques de la langue qui conditionnent l'emploi des formes impliquées.

Tableau 1. Poly-cooccurrences du pôle *je* (*Sp.* ≥ 5) - extrait

<i>je</i>				<i>je + veux</i>				<i>je + veux + président</i>			
Forme	CF	Sp.	Ctx	Forme	CF	Sp.	Ctx	Forme	CF	Sp.	Ctx
<i>veux</i>	28	+24	26	<i>que</i>	16	+6	13	<i>être</i>	7	+7	7
<i>que</i>	49	+9	39	<i>président</i>	7	+6	7				
<i>suis</i>	10	+9	10	<i>pays</i>	8	+5	8				
<i>vous</i>	30	+8	26	<i>être</i>	8	+5	8				
<i>pense</i>	7	+6	7								
<i>ferai</i>	6	+6	6								
<i>serai</i>	6	+6	6								
<i>notre</i>	23	+5	20								
<i>me</i>	6	+5	6								

Guide de lecture du tableau 1 : Les trois parties du tableau exemplifient l'exploration progressive des contextes porteurs du pôle initial *je*. A partir de ce dernier on mesure, entre autres, une co-fréquence élevée (*CF*) avec la forme *veux*. En explorant les 26 contextes (*Ctx*) où ont lieu ces 28 associations, un second calcul révèle une spécificité (*Sp.*) de la forme *président*, co-présente à 7 reprises dans l'échantillon textuel. Enfin, une troisième exploration dans les contextes qui sont porteurs simultanément des formes *je*, *veux* et *président* rapporte une attraction spécifique avec *être*, co-présent à 7 reprises. Le diagnostic de la dernière phase du calcul garantit la réalisation de la poly-cooccurrence *je+veux+président+être* dans 7 phrases du corpus.

Sur la base des 'squelettes de phrase' relevées par la méthode, on extrait du corpus sept réalisations de la poly-cooccurrence *je+veux+président+être*, une dans le texte de F. Bayrou et six chez N. Sarkozy :

François Bayrou – UDF

- (1) **je veux être le président** de la république qui rendra son équilibre à notre pays.

Nicolas Sarkozy – UMP

- (2) **je veux être le président** du pouvoir d'achat...
- (3) **je veux être le président** qui garantit de nouvelles protections...
- (4) **je veux être le président** d'un état fort et impartial...

- (5) **je veux être le président** qui prépare l'avenir des générations futures...
- (6) **je veux être le président** d'une france fière de ses valeurs et de son identité...
- (7) **je veux d'abord être le président** de la valeur travail ...

A travers ces extraits on saisit la portée de l'exploration poly-cooccurentielle qui, au-delà de la cooccurrence classique embrasse une plus grande variété de phénomènes. Affranchie des contraintes de contiguïté et d'orientation, la poly-cooccurrence recouvre le segment répété *je veux être le président* et fusionne celui-ci avec le quasi-segment *je veux d'abord être le président*, fournissant une statistique plus juste sur les réalisations de ce système lexical en contexte.

2.2 Représentation graphique

Même si elle restitue fidèlement le mécanisme exploratoire à la base de la méthode, la transcription tabulaire des poly-cooccurrences appelle un autre mode de représentation. En effet, dans le texte d'origine, le sens émerge et se donne en situation, il est toujours local et transitoire. De manière analogue, le calcul des poly-cooccurrences invoque la solidarité contextuelle des cooccurrences multiples en interaction. Une visualisation a donc été imaginée sous forme d'arbre (figure 1) qui permet d'interroger le graphe en n'importe quel point et d'obtenir un résultat statistique qui, du fait des modalités de calcul, est inévitablement avéré en contexte.

Figure 1. Poly-cooccurrences à partir du pôle je ($Sp. \geq 4$) - extrait

Guide de lecture de la figure 1 : La présentation par tabulation des poly-cooccurrences chaînées distingue chaque niveau de l'exploration réitérée des contextes et la triple notation sous la forme de cartouches *co-fréquence [spécificité] (contextes de cooccurrence)* permet une lecture rapide des systèmes repérés (seuils : co-fréq. 4, spéc. 4 et au moins 4 contextes de réalisation). Ainsi, à partir du pôle *je*, on signale une cooccurrence avec la forme *suis* par une flèche notée 10[9](10) qui indique que le cooccurrent apparaît 10 fois aux côtés du pôle avec une spécificité de +9 dans un total de 10 phrases. En poursuivant la lecture à droite du graphe on découvre une cooccurrence de la paire *je + suis* avec la forme *candidat* à 7 reprises dans 6 des 10 phrases circonscrites à l'étape précédente. Enfin c'est la forme *contre* qui est détectée comme cooccurrente de la branche *je+suis+candidat* à 7 reprises dans 5 de ses 6 contextes.

Les poly-cooccurrences détectées autour de *je* permettent d'identifier des usages particuliers chez les candidats qui pour certains réduisent leur programme politique à son message essentiel. Chez les candidats de droite, outre le *je+veux+être+président* de N. Sarkozy, on trouve un *je+suis+candidat+contre* récurrent dans le discours de J.-M. Le Pen tandis que F. Bayrou privilégie un *je+serai+président*.

Jean-Marie LePen – FN

- (8) **je suis le candidat** de la vérité **contre** le mensonge.
- (9) **je suis le candidat** de la lucidité **contre** l'aveuglement, le **candidat** du bon sens **contre** l'incompétence.
- (10) **je suis le candidat** des vraies réformes **contre** les candidats des fausses promesses.
- (11) **je suis le candidat** de la france, de la vie, de l'avenir et de la paix **contre** les candidats du mondialisme, des communautarismes, de l'argent-roi et des guerres.

François Bayrou – UDF

- (12) **je serai le président** qui réunira les français pour les faire vivre ensemble.
- (13) **je serai le président** de ce grand changement.
- (14) **je serai le président** qui mobilisera les ressources de notre grand pays.
- (15) **je serai le président** qui réunira les français pour les faire vivre ensemble (...)

Etant données la richesse et la précision des résultats obtenus suivant cette méthode, on considérera que les poly-cooccurrences présentent l'état le plus concret de l'extraction cooccurentielle et offrent le plus juste compromis entre le texte d'origine et l'index lexicométrique. Pourtant, la subjectivité dans le choix du pôle initial tel qu'il est sélectionné dans les méthodes présentées jusqu'ici conditionne excessivement l'exploration contextuelle. Alors, afin de garantir une approche plus objective des structures de cooccurrence, une autre procédure a été imaginée qui explore le texte sans contrainte initiale.

3. Trames de cooccurrence

En empruntant à la métaphore de Maurice Tournier qui oppose 'une écriture de chaîne à une écriture de trame' (dans Demonet *et al.* [1975]), nous présenterons ici une méthode qui permet, en mobilisant une statistique exhaustive des coïncidences en contexte, de saisir l'organisation sous-jacente d'un texte : les *trames de cooccurrence*.

3.1 Elaboration

L'algorithme à la base du calcul des trames de cooccurrence se distingue de celui des poly-cooccurrences⁸ en ce qu'il ne restreint jamais le sous-corpus exploré aux contextes d'apparition d'un pôle choisi subjectivement mais au contraire multiplie les analyses cooccurentielles autour d'un ensemble de pôles à l'intérieur d'une *plage de fréquence*⁹. L'objectif est de dresser une liste d'associations dans un tableau qui croise toutes les formes étudiées et restitue l'intégralité de leurs cooccurrences réciproques¹⁰.

8 Chaque méthode répond à une interrogation différente. Pour les poly-cooccurrences : quels sont les cooccurents spécifiques des contextes du pôle composé ABC ? Pour les trames de cooccurrence : quels sont les cooccurents spécifiques de A, de B, de C ?

9 Soit un ensemble de formes comprises entre une fréquence maximale et une fréq. minimale.

10 De par ses principales propriétés de calcul la méthode des trames se rapproche de celle des *lexicogrammes récursifs* de Heiden et Lafon [1998]. Mais, à la différence de celle-ci et dans une visée d'exhaustivité, le calcul des trames explore l'intégralité des cooccurrences réciproques opérant dans le texte et restitue l'ensemble des systèmes cooccurentiels

Etant donnée cette visée d'exhaustivité, le calcul réitéré des cooccurrences suppose une combinatoire qui génère une masse de résultats pouvant rapidement devenir ingérable¹¹. Pour optimiser ce calcul, il est donc nécessaire de réduire les phénomènes rapportés aux plus importantes attractions dans le corpus. Pour ce faire, les seuils de calcul sont draconiens eu égard au volume du corpus d'étude : ne sont comptabilisées que les cooccurrences *récioproques*¹²⁻¹³ (de co-fréq. 5 et de spéc. 5). Une fois définis les paramètres de calcul, l'analyse des 184 formes de fréquence ≥ 5 livre une matrice d'adjacence qui croise tous les pôles étudiés et contient aux intersections l'indice de spécificité de leur cooccurrence. Avec cette matrice (tableau 2), l'objectif d'exhaustivité est atteint puisque l'ensemble des

qui s'y élaborent. Autrement dit, aucun système cooccurentiel présent dans le corpus n'est occulté par le choix initial d'un pôle et la méthode produit au final une liste exhaustive des réseaux de cooccurrences autonomes qui cohabitent dans le texte.

- 11 Etudier l'ensemble des cooccurrences récioproques reliant n formes revient à statuer sur n^2-n attractions. Dans notre corpus, on étudie les cooccurrences des 184 formes de fréquence ≥ 5 soient $184^2-184 = 33\,672$ associations.
- 12 Du fait des modalités de comptage des rencontres en contexte, la spécificité calculée pour une cooccurrence n'est pas forcément réciproque. Considérons l'exemple suivant, extrait du discours du candidat de la LCR : *contre l'europe de bruxelles, c'est une autre europe qu'il faut construire avec l'ensemble des peuples*. Tout contexte porteur ne pouvant être comptabilisé qu'une seule fois, on dénombre durant l'exploration de cette phase deux cooccurrences de la forme *europe* autour du pôle *construire* mais seulement une cooccurrence de *construire* autour du pôle *europe*. Afin de pallier cette asymétrie du comptage, il est possible de définir pour ce calcul un écart toléré entre deux spécificités. Cette marge permet d'inclure dans les résultats de l'analyse certaines cooccurrences qui n'ont pas une spécificité strictement réciproque, c'est-à-dire une intensité de spécificité identique quelle que soit l'orientation de la mesure. Dans notre exemple, *construire* attire *europe* avec une spécificité de +10 mais inversement *europe* attire *construire* avec une spécificité de +5. Pour valider cette cooccurrence il faudra tolérer un écart de spécificité d'au moins ± 5 .
- 13 On définit une liste de mots outils à écarter de l'exploration car on estime que leur profil cooccurentiel est peu pertinent et qu'ils entraînent l'algorithme d'exploration vers des espaces contextuels trop grands et sans originalité par rapport au corpus total (la comparaison local / global est à la base du modèle probabiliste exploité – Cf. Note n° 4).

cooccurrences binaires entretenues entre ces formes sont comptabilisées et tous les systèmes complexes que ceux-ci élaborent sont détectés¹⁴. Du fait de sa structure lacunaire, ce tableau est ce que l'on appelle une *matrice creuse* où le peu d'information pertinente est dispersé au gré des rares phénomènes d'attraction lexicale : 33 547 cases sur 33 856 (soit 99 % du total) contiennent un zéro. Paradoxalement, comme l'on verra plus loin, ce faible nombre d'attractions produit un nombre considérable de réseaux de cooccurrence dont certains forment même des systèmes de grande complexité.

Si la matrice d'adjacence contient toute l'information nécessaire pour décrire les systèmes d'attraction lexicale qui organisent le texte, elle ne fournit toutefois pas un résultat directement interprétable. Pour en saisir la structure et la dynamique, il faut y mesurer la capacité d'attraction de chaque forme et reconstituer les réseaux que les mots tissent entre eux.

Tableau 2. Matrice d'adjacence (extrait)

	st	e	l	ous	vec	e	ous	ous	...)
st									
e									
l									
ous									
vec				0					
e									
ous									
ous									
ous									
...)									

14 Aux seuils en vigueur, chacun des 184 pôles compte au moins une cooccurrence spécifique. Mais, si l'on étend la plage des formes étudiées, de nombreux 'orphelins' sont écartés de la liste pour absence de cooccurrence réciproque avec une autre forme de la plage.

Guide de lecture du tableau 2 : Le calcul des cooccurrences autour de tous les pôles dans une plage de fréquence donnée (ici $F \geq 5$) produit un tableau qui croise toutes les formes étudiées et présente, à l'intersection d'une ligne L avec une colonne C , l'indice de spécificité pour la cooccurrence du pôle L avec le cooccurent C . Si une attraction n'est pas réciproque, la spécificité 'à sens unique' est nullifiée, et, à l'issue du calcul, on obtient un tableau où à peine 1 % des cases présentent un indice de spécificité.

3.2 Valence lexicale

L'analyse de la matrice d'adjacence révèle que certains pôles entretiennent plus de cooccurrences spécifiques que d'autres. Ainsi, avec ses 7 cooccurents (*avons*, *voulons*, *nos*, *sommes*, *pouvons*, *ensemble* et *besoin*), *nous* apparaît comme une forme plus magnétique que *président* par exemple qui n'attire que 3 cooccurents (*veux*, *être* et *république*). En automatisant ce comptage, on dégage une liste hiérarchisée qui distingue les formes-pôles à forte capacité d'attraction et classe celles-ci suivant le critère de *valence lexicale* (tableau 3). On identifie alors des *ancres cooccurentielles*, c'est-à-dire des formes très magnétiques dont la présence dans le système lexical détermine souvent la présence d'autres formes qui dépendent de cette cooccurrence de manière exclusive (ex. : dans le réseau de la figure 2, *doit* dépend entièrement de *être*).

Lorsque l'on compare cet ordre de classement avec l'ordre fréquentiel du dictionnaire, on constate que la capacité d'attraction n'est pas corrélée à la fréquence. On trouve en effet des formes très fréquentes qui comptent peu de cooccurents et, à l'inverse, des formes à faible fréquence qui entretiennent beaucoup d'attractions. Ainsi la forme *nous* (fréq. 101) avec un ratio de 0.07 attire proportionnellement autant de cooccurents que *salaires* (fréq. 26), alors que le pronom est presque 4 fois plus fréquent que le nom. La forme *nous* serait donc moins structurante dans le texte que *salaires* puisqu'elle ne perturbe pas son environnement contextuel par l'introduction de formes statistiquement improbables¹⁵.

¹⁵ Le phénomène s'accroît lorsque l'on baisse les seuils de calcul pour évaluer la totalité des cooccurrences réalisées dans le corpus (voir 3.4.3). Le pôle *nous* compte alors 45 cooccurrences contre 23 pour *salaires* ce qui fixe leurs valences respectives à 0.44 et 0.88.

Tableau 3. Valence lexicale

Forme	CS	C/F	Forme	CS	C/F	Forme	CS	C/F
<i>nous</i>	7	0.07	<i>rupture</i>	4	0.36	<i>est</i>	2	0.10
<i>je</i>	6	0.04	<i>aux</i>	3	0.03	<i>il</i>	2	0.02
<i>violences</i>	6	0.75	<i>nos</i>	3	0.05	<i>sur</i>	2	0.02
<i>contre</i>	5	0.10	<i>entreprises</i>	3	0.08	<i>avec</i>	2	0.02
<i>être</i>	5	0.13	<i>européenne</i>	3	0.09	<i>ne</i>	2	0.02
<i>femmes</i>	5	0.21	<i>veux</i>	3	0.11	<i>pas</i>	2	0.02
<i>faites</i>	5	0.71	<i>depuis</i>	3	0.12	<i>notre</i>	2	0.03
<i>vous</i>	4	0.05	<i>loi</i>	3	0.12	<i>a</i>	2	0.03
<i>ans</i>	4	0.08	<i>22</i>	3	0.13	<i>s</i>	2	0.03
<i>éducation</i>	4	0.18	<i>président</i>	3	0.13	<i>n</i>	2	0.03
<i>euros</i>	4	0.21	<i>avril</i>	3	0.14	<i>pays</i>	2	0.04
<i>présidentielle</i>	4	0.21	<i>2007</i>	3	0.15	<i>vie</i>	2	0.04
<i>cadre</i>	4	0.36	<i>suis</i>	3	0.30	<i>(...)</i>		

Guide de lecture du tableau 3 : La valence lexicale mesure le magnétisme de chaque forme par le nombre de cooccurrences spécifiques (CS) qu'elle entretient dans le texte. On pondère cette mesure en rapportant le nombre de cooccurrences d'un pôle à sa fréquence (C/F), un ratio qui distingue les formes les plus actives sur le plan cooccurentiel.

Mesure simple mais juste de la cooccurrence spécifique, la valence lexicale révèle le magnétisme des formes qui est à la base de leur comportement cooccurentiel et signale les mots dont l'emploi dans le discours entraîne la présence rapprochée d'autres mots qui forment ensemble des cohortes statistiquement originales et définissent conjointement des noyaux sémantiques.

3.3 Trames de cooccurrence

La seconde étape de l'exploitation de la matrice d'adjacence vise à extraire les systèmes de cooccurrences multiples qui s'y développent. En effet, par le jeu des attractions binaires et réciproques, différents systèmes s'élaborent dont les composantes

tissent une toile de co-dépendances de deux en deux¹⁶. La totalité des formes listées dans la matrice sont en cooccurrence réciproque avec au moins une autre forme (qui elle-même compte éventuellement d'autres attractions) et l'on extrait du tableau plusieurs réseaux cooccurentiels de complexités différentes¹⁷.

Au final, l'exploration détermine 48 trames distinctes qui associent les 184 formes présentes dans la matrice d'adjacence. A la lecture de la liste (tableau 4) on constate que tous les réseaux détectés, du plus simple au plus complexe, présentent une homogénéité de sens. Ainsi, les ensembles que le calcul statistique a repérés en tant que structures de coïncidence contextuelle constituent des objets sémantiquement cohérents. Dans la vaste majorité des cas, les systèmes repérés par le dispositif se réduisent à des cooccurrences binaires, contiguës et orientées, c'est-à-dire des segments répétés (*petites retraites, au premier tour*) ou quasi-segments répétés (*mettre place, parti travailleurs*). Au-delà de ces associations simples, on trouve aussi des systèmes plus complexes, réels pour les uns, virtuels pour les autres.

Certaines des trames complexes mises au jour se réalisent en contexte dans leur intégralité. Par exemple, on trouvera dans le corpus deux réalisations de la trame *nicolas-sarkozy-françois-bayrou* :

Arlette Laguiller – Lutte Ouvrière

- (16) ni **nicolas sarkozy**, ni ségolène royal, ni **françois bayrou** ne proposent une politique susceptible de répondre aux besoins des masses populaires sur ces trois problèmes vitaux que sont le chômage, le coût de la vie et le logement.

Marie-George Buffet – Parti Communiste

- (17) 30 % des salarié-es gagnent moins de 1 300 euros : **nicolas sarkozy** et **francois bayrou** refusent toute augmentation des salaires.

16 Par construction de l'algorithme on peut partir de n'importe lequel des noeuds d'une trame donnée, c'est toujours le même système cooccurentiel qui surgit de l'exploration. Cette caractéristique du réseau montre que toute structure de dépendances mutuelles est sous-jacente au texte et qu'elle se dévoile à travers l'analyse cooccurentielle quel que soit le pôle de départ de l'exploration ou 'le fil sur lequel on tire'.

17 Voir note 14 ci-dessus.

D'autres systèmes parmi ceux détectés constituent des réseaux de *cooccurrence virtuelle* dans la mesure où, jamais dans le corpus, on ne trouve leurs composantes employées simultanément dans les mêmes contextes. C'est le cas de la trame *grandes-entreprises-profits-bénéfice* repérée progressivement par la statistique mais inexistante en l'état dans le texte. En effet, comme l'illustrent les extraits suivants, ce dont rend compte cette trame est la superposition de deux poly-cooccurrences avérées qui, ensemble, élaborent un réseau virtuel : *bénéfices+grandes+entreprises* et *bénéfices+profits+entreprises*.

Arlette Laguiller – Lutte Ouvrière

- (18) il faut augmenter les impôts sur les **bénéfices** des **grandes entreprises** et sur les revenus des plus riches.

Olivier Besancenot – Ligue Communiste Révolutionnaire

- (19) il faut s'en prendre aux **profits** des capitalistes en commençant par rétablir le taux de 50 % sur les **bénéfices** des **entreprises**.

Outre un grand nombre de systèmes cooccurentiels simples, on détecte dans le corpus des réseaux de cooccurrence qui associent un nombre élevé de composantes lexicales. Le calcul dégage notamment quatre grands réseaux de cohérence sémantique qui portent les principales thématiques évoquées dans les programmes :

- Union Européenne et Traité de Maastricht (8 formes)
- bilan négatif et avenir possible (9 formes)
- union nationale autour du pronom *nous* (13 formes)
- programme et dialogue autour du pronom *je* (37 formes)

A défaut de fréquence à proprement parler, on peut mesurer l'importance de ces trames suivant leur *surface de cooccurrence*¹⁸. D'après cet indice (donnée fournie entre parenthèses au tableau 4) on vérifie que ces systèmes constituent les principales structures lexicales à l'œuvre dans le discours. Aussi, en retournant au contexte on s'aperçoit que ces trames

18 Cette aire correspond au nombre d'occurrences cumulées des composantes d'une trame en relation de cooccurrence binaire avérée dans les contextes : chaque paire (ou plus) de cooccurents réalisée en contexte est comptabilisée car elle contribue à la présence structurante de la trame dans le texte.

Poly-cooccurrences et trames de cooccurrence

forment des pans de texte qui recouvrent plusieurs locuteurs. On pourra ainsi associer au discours global de la gauche un bilan critique du passé et des engagements européens tandis que la droite présente un discours programmatique autour du *je* et que l'union nationale reste un thème commun aux candidats de tous bords.

Soulignons que l'ensemble des 48 trames détectées cohabitent dans le corpus sans jamais fusionner¹⁹. En effet, leur fonctionnement moléculaire fait que les trames se juxtaposent au fil du texte pour construire des blocs textuels (syntagmes, phrases et même paragraphes) et forment ensemble le *treillis* qui structure le corpus dans son intégralité. Pour appréhender cette superstructure dans toute sa complexité, il faut recourir à une méthode de visualisation plus à même de promouvoir les mots suivant leur pouvoir structurant dans le discours.

Tableau 4. Trames de cooccurrence

<u>28 réseaux à 2 éléments :</u>	
- <i>toujours, plus</i> (surface de cooc. = 19 occ.)	- <i>mettre, place</i> (12 occ.)
- <i>une, politique</i> (78 occ.)	- <i>philippe, villiers</i> (18 occ.)
- <i>ceux, qui</i> (38 occ.)	- <i>aujourd, hui</i> (32 occ.)
- <i>emploi, formation</i> (15 occ.)	- <i>ma, candidature</i> (14 occ.)
- <i>pouvoir, achat</i> (23 occ.)	- <i>parti, travailleurs</i> (12 occ.)
- <i>j, ai</i> (25 occ.)	- <i>personnes, âgées</i> (10 occ.)
- <i>on, ferme</i> (11 occ.)	- <i>petites, retraites</i> (11 occ.)
- <i>services, publics</i> (54 occ.)	- <i>abrogation, lois</i> (10 occ.)
- <i>tous, toutes</i> (28 occ.)	- <i>dominique, voynet</i> (20 occ.)
- <i>tout, peut</i> (12 occ.)	- <i>gérard, schivardi</i> (18 occ.)
- <i>mais, aussi</i> (22 occ.)	- <i>jean, marie</i> (10 occ.)
- <i>son, ses</i> (18 occ.)	- <i>énergies, renouvelables</i> (10 occ.)
- <i>service, public</i> (24 occ.)	- <i>langues, régionales</i> (10 occ.)
- <i>1, million</i> (10 occ.)	- <i>ségolène, royal</i> (12 occ.)
<u>8 réseaux à 3 éléments :</u>	
- <i>il, y, a</i> (53 occ.)	- <i>augmentation, bas, salaires</i> (26 occ.)
- <i>au, premier, tour</i> (37 occ.)	- <i>impôts, taxe, sur</i> (14 occ.)
- <i>sécurité, sociale, démocratique</i> (32 occ.)	- <i>25, mai, 2005</i> (10 occ.)
- <i>engagé, e, s</i> (28 occ.)	- <i>changer, conditions, vie</i> (28 occ.)

¹⁹ La fusion des trames est fonction des seuils de calcul : si l'on baisse les seuils de co-fréquence et de spécificité alors les réseaux s'agrègent. Dans le cas extrême où les seuils sont fixés au minimum théorique de 1, c'est la totalité du corpus qui émerge dans une trame unique.

3 réseaux à 4 éléments :

- *sarkozy, bayrou, nicolas, françois* (47 occ.)
- *dans, classes, populaires, quartiers* (59 occ.)
- *grandes, entreprises, profits, bénéfices* (61 occ.)

5 réseaux à 5 éléments :

- *ans, depuis, 5, années, vingt* (64 occ.)
- *europe, sociaux, logements, construire, minima* (57 occ.)
- *22, avril, 2007, présidentielle, élection* (82 occ.)
- *éducation, nationale, enfants, culture, accès* (46 occ.)
- *moins, euros, milliards, 300, net* (37 occ.)

1 réseau à 8 éléments :

- *avec, européenne, union, rupture, finir, traité, constitution, maastricht* (110 occ.)

1 réseau à 9 éléments :

- *ne, pas, n, qu, ont, ils, possible, autre, peuvent* (271 occ.)

1 réseau à 13 éléments :

- *nous, notre, nos, pays, ensemble, communes, défense, avons, pauvres, voulons, besoin, sommes, pouvons* (281 occ.)

1 réseau à 37 éléments :

- *est, je, que, aux, vous, c, contre, être, doit, veux, loi, république, droite, femmes, président, votre, égalité, vos, soit, bien, candidat, voter, hommes, cadre, face, parce, pourquoi, suis, lutte, soient, pense, violences, faites, extrême, ferai, serai, avez* (938 occ.)

3.4 Représentation graphique

3.4.1 Elaboration

Dans le processus de calcul des trames de cooccurrence il n'y a pas d'immobilité : l'exploration contextuelle est sans cesse répétée jusqu'à révéler des systèmes lexicaux complexes à l'œuvre dans le texte. Par conséquent, il ne doit pas y avoir de figement dans le mode de présentation des données. Or, la visualisation de chaque trame sous forme d'une liste de mots brise la continuité de sens qu'élabore le chemin de cooccurrence et contraint la lecture tout comme le font les listes pétrifiées du dictionnaire de fréquence et celles issues du calcul des cooccurrences.

Nous avons donc cherché à combiner l'information du tableau de valence lexicale avec celle de la liste des trames afin

d'élaborer une représentation graphique qui traduit l'ensemble des interconnexions reliant les cooccurents dans un système lexical donné. Dans la figure 2 qui présente la plus importante trame détectée dans le corpus (37 formes et 46 connexions), les nœuds représentent les formes et les arcs figurent les attractions entre ces dernières.

Figure 2. Trame de cooccurrence (Spéc. ≥ 5)

Guide de lecture de la figure 2 : En associant l'information sur la valence lexicale des formes avec celle fournie par l'analyse des trames, le graphique de la trame de cooccurrence restitue une image fidèle des systèmes associatifs à l'œuvre dans le texte. Ici, dans le réseau impliquant le pôle *je*, nous avons enrichi l'information du schéma par un cadrage plus épais des formes à plus forte valence lexicale (≥ 5 cooccurents) et par une épaisseur des arcs proportionnelle à l'intensité de la spécificité (la plage de spécificité de 2 à 50 est divisée en trois intensités que l'on représente par trois épaisseurs de trait). On constate que les formes les plus magnétiques sont par ordre *je*, *que*, *violences*, *vous*, *contre*, *faites* et *femmes*, et que la cooccurrence binaire la plus forte associe les formes *je* et *veux*. Précisons que l'interprétation de ce graphe est soumise à la non-transitivité des cooccurrences : l'attraction des formes AB et celle des formes BC ne garantit pas une cooccurrence ABC.

3.4.2 Interprétation

Là où le texte présente des repères multiples et des points d'entrée prédéterminés pour appréhender l'information qu'il véhicule, la trame de cooccurrence offre une dynamique de lecture très différente. Si le texte est placé sous le signe de la durée, de la continuité, le réseau, lui, est discontinu, fragmentaire et impose constamment des choix de lecture. En lieu et place des blocs d'information du texte qui, destinés à être lus linéairement, sont disposés selon un ordre séquentiel (mots, syntagmes, phrases, etc.), le réseau offre des îlots de mots, distincts et de tailles variables, dont l'agencement se prête davantage à une approche *topologique* qui privilégie les propriétés combinatoires des noeuds et leurs positions relatives²⁰.

D'emblée on saisit l'efficacité synoptique de ce type de carte qui offre une vue claire de la distribution de l'information en signalant les relations pertinentes de dépendance entre variables et groupes de variables. En suggérant une vaste combinatoire lexicale, la dimension réticulaire du graphique favorise une attitude exploratoire par rapport à un matériau organisé non pas en liste mais quasi syntagmatiquement. L'information ainsi spatialisée permet au lecteur d'aller directement au mot qui l'intéresse et de choisir un point d'entrée dans le réseau. Dans la figure 2 on repère rapidement les branches principales de l'arbre *je-que-vous* et *contre-violences-faites-femmes* ainsi que des sous-branches pertinentes *voter-extrême-droite*, *c'est-pourquoi* ou encore *égalité-hommes-femmes*²¹.

20 Rappelant cette dimension spatiale du texte, Vandendorpe [1999] distingue différents modes d'accès aux réseaux – sélection, association, contiguïté et stratification – tous susceptibles de reconstituer la textualité d'origine : *'Les réseaux traduisent la textualité au sens de caractéristique d'un objet sensible appréhendé de façon spatiale et qui s'adresse à la compréhension d'un lecteur en jouant à la fois sur la mise en rapport systématique de propositions élémentaires placées en continuité et sur des rappels plus ou moins lointains, continus et réglés d'éléments présentés en amont.'*

21 La position centrale du nœud *contre* correspond à un rôle clef dans ce graphe puisque cette forme garantit la liaison entre deux univers distincts, l'un formé autour du binôme *je+veux* et l'autre autour de *violences+faites+femmes* (trio avéré dans 7 poly-cooccurrences). L'exclusion de cette forme-relais en tant que mot-outil aurait pour effet de disjoindre les deux univers.

Cependant, on saisit très vite les limites de l'interprétation de ces données et l'on comprend que l'exégèse de ce type de graphique sera difficile. Effectivement, si le réseau dessine une organisation lexicale qui est autrement imperceptible, il range les mots sans tenir compte de l'ordre dans lequel ils s'organisent et s'étend dans toutes les directions sans limites précises, ni début, ni fin. Certes, le schéma dynamique restitue fidèlement le texte en mouvement et invite le regard du lecteur dans ce flux en l'orientant peut-être vers quelque chose de plus important que le pôle initial choisi comme point d'entrée. Ceci étant, la prégnance des réseaux d'associations est telle qu'il s'avère difficile, voire impossible, de s'arrêter à quelque chose de définitif car il faut sans cesse relier la sélection faite au système global : la *factorisation de l'information* à la fois accélère et entrave la lecture²² !

Pour illustration, considérons sur la figure 2 la paire *je+que* (49 cooccurrences avérées). On pourra sans risque d'erreur associer ce couple à la forme *veux* ou encore à la forme *vous* (respectivement 28 et 29 cooc. avérées)²³. Mais, au-delà de ces systèmes simples, chaque ajout d'un cooccurrent poussera la spéculation un peu plus loin et l'on devra inexorablement vérifier la réalisation en contexte des séquences prélevées dans

22 Les raisons de cette difficulté d'interprétation sont à chercher dans le mode de calcul des réseaux dont il faut souligner deux conséquences directes sur le schéma produit :

- a connexité des sommets : il existe un chemin reliant toute paire de noeuds.

- la non-transitivité des cooccurrences : aucune attraction n'est garantie au-delà de la cooccurrence binaire.

Ces conditions d'interprétation opèrent en sens inverse : l'une invite à construire une chaîne de cooccurrence théoriquement illimitée, l'autre incite à la méfiance à l'égard de chaque association lexicale comptant plus de deux éléments. Alors, devant ces réseaux qui ne sont pas censés être lus dans un ordre précis, le fil de la lecture se redéfinit perpétuellement car tout nouveau développement modifie le sens : chaque ajout d'un nœud est construction d'un sens nouveau et/ou une destruction d'un sens ancien.

23 Notre inférence pourra s'appuyer sur la théorie des graphes qui identifie certains groupes de formes comme étant des *cliques*, c'est-à-dire des sous-ensembles dont chaque composante est reliée à toutes les autres. On vérifiera en contexte les fréquences d'apparition de *je+que+vous* (14 occ.), *je+que+veux* (16 occ.) et *je+que+veux+vous* (5 occ.).

le schéma. Cette réalité – totale mais virtuelle – restitue la coexistence de réseaux lexicaux dans le texte et s’oppose aux systèmes – partiels mais réels – révélés dans les poly-cooccurrences.

3.4.3 Vers des cartes conceptuelles

En dégageant des structures statistiques qui, parce qu’elles coïncident avec des structures sémantiques, font immédiatement sens à la lecture, l’approche multi-cooccurentielle du texte participe de ce que l’on appellera un *cognitivismes statistique*²⁴. De fait, l’aspect circulaire de certaines grappes dans le réseau clarifie la hiérarchisation des nœuds en dégageant les formes à valence lexicale élevée qui sont plus importantes sur le plan de la structuration sémantique du texte. Peu nombreuses, ces formes occupent une position centrale sur le graphe tandis que d’autres, telles de nombreux satellites, viennent les compléter pour former des constellations lexicales sémantiquement homogènes.

On peut tirer parti de cette topologie particulière afin d’améliorer la détection, la visualisation et la compréhension des réseaux de cooccurrence. Pour ce faire, on baisse d’une part les seuils de calcul et on élève d’autre part le seuil de filtrage des résultats. Cette manœuvre contradictoire permet de fonder la statistique cooccurentielle sur un maximum de phénomènes de cooccurrence réalisés dans le corpus tout en limitant les résultats visualisés aux pôles les plus magnétiques²⁵.

24 Nous empruntons le terme à Blayo et Verleysen [1996] qui expliquent que l’appréhension du monde par le cerveau humain procède souvent par un traitement statistique et, citant Delacour [1995], précisent que ‘*certaines formes de l’activité cognitive sont équivalentes à une réduction de la dimension des données*’, ‘*à une compression de celles-ci, comme le font de manière explicite des techniques statistiques d’analyse multivariée : elles donnent de façon en grande partie intuitive, imagée... une représentation simplifiée de données complexes, multidimensionnelles, en conservant le maximum d’information*’.

25 En appliquant les seuils les plus bas possibles (co-fréq. 2 et spéc. 2) on pousse le calcul des cooccurrences et la représentativité des résultats à l’extrême en choisissant d’ignorer le degré d’intensité des attractions au profit d’une capacité d’attraction au sens large qui se mesure aussi par les attirances les plus modestes. Cette tolérance propulse de nombreuses formes dans le classement du fait de leur valence lexicale élevée (fût-elle fondée sur des attractions faibles). La matrice d’adjacence intègre

La principale conséquence de ce nouveau seuillage est une diminution du nombre de trames détectées à seulement 18 réseaux (au lieu de 48 précédemment). En maximisant l'information on a donc aggloméré de nombreuses trames mineures dans un réseau principal qui compte désormais 1 256 nœuds et 4 003 liens. Toutefois, si la systématique totale du calcul à la base de cette *hypertrame* permet de penser le corpus dans son intégralité et de l'appréhender à une profondeur inédite, ce seuillage extrême a pour effet de produire des données illisibles si tant est qu'elles soient représentables graphiquement.

Afin de réduire la complexité visuelle du réseau au nécessaire, on applique un surfiltrage des résultats par indice de valence lexicale (≥ 20) ce qui permet d'obtenir une représentation compacte de l'information (fig. 3). Tel le 'pic de l'iceberg' cette superstructure de cooccurrence laisse apparaître 47 nœuds en surface mais repose sur une masse de 1 256 formes en profondeur. Les aires de sens que l'on esquissait dans la trame de cooccurrence ont laissé la place à une poignée d'*hyperpôles*²⁶ qui subsument la diversité initiale du réseau. En identifiant une base de faits primitifs dont les autres dépendent, l'*hypertrame* cartographie une constellation sémantique où les déterminations concrètes des cooccurrences binaires sont abstraites pour mettre en avant l'idéalité conceptuelle du réseau. Cet ensemble de pôles à très forte valence lexicale forme la maille principale du texte et permet de dessiner une carte conceptuelle très fidèle au corpus dont on mesure la représentativité notamment par son maillage serré²⁷ et son importante surface de cooccurrence²⁸.

alors 1 680 formes dont 1 256 s'associent en un réseau principal, 38 créent des réseaux secondaires et 386 restent orphelines.

26 Un parallèle s'établit ici avec l'*hypertexte* qui supprime la page comme unité élémentaire du texte et offre une interface nouvelle, un construit informatique de liens où tous les pliage sont possibles. Comme l'explique Lévy [1990], texte et hypertexte coexistent depuis toujours, de manière consubstantielle : '*La métaphore de l'hypertexte rend compte de la structure indéfiniment réursive du sens, car puisqu'il connecte des mots et des phrases dont les significations se répondent et se font écho par delà la linéarité du discours, un texte est toujours déjà un hypertexte, un réseau d'associations.*'

27 Par rapport au graphique de la trame précédente (fig. 2), on calcule un ratio liens / nœuds plus élevé : $76 / 47 = 1.61$ contre $46 / 37 = 1.24$.

Figure 3. Hypertrame de cooccurrence

Guide de lecture de la figure 3 : Le filtrage des formes à valence lexicale faible produit un macro-réseau qui ne présente que les ancres les plus magnétiques du corpus (val. lex. ≥ 20). Si la base du calcul garantit de par les seuils appliqués (co-fq. 2 et spéc. 2) une exploration totale du corpus, la présentation filtrée ne restitue que 3.7 % du réseau détecté (47 pôles sur 1 256). Malgré cet élagage massif, les nœuds restants tissent un réseau resserré et esquissent des aires sémantiques homogènes. L'importance de ces *hyper-pôles* comme éléments structurants du texte est suggérée par la taille des nœuds qui est proportionnelle à leur valence lexicale. On pourra, en repoussant ce seuil, extraire progressivement la colonne vertébrale du texte : *entreprises-europe-politique-présidentielle* (val. lex. ≥ 35) puis *nous-vie-communes* (vl. ≥ 40), et enfin *je* (val. lex. ≥ 45) qui s'avère être le noyau fondamental du texte.

28 Cette hypertrame recouvre 9 338 occurrences soit 46.5 % du corpus. Si l'on écarte les 28 % de mots-outils déclarés et les 10 % de hapax, on constate qu'à peine 15 % du texte échappe au résumé offert par ce réseau de cooccurrence.

Conclusion

En lexicométrie la désintégration du texte à l'état d'entités simplifiées est une heuristique nécessaire pour déterminer certaines propriétés lexicales et structurelles du corpus. À l'opposé, les méthodes de cooccurrence, par intégrations successives des éléments lexicaux, reconstituent progressivement le texte dans sa dimension syntagmatique afin d'identifier les préférences combinatoires de ses composantes²⁹. De fait, de par la place qu'elles occupent dans la chaîne de traitement, on désignera par le terme 'synthèse cooccurrentielle' l'ensemble des méthodes exposées ici et leurs résultats qui agrègent les mots de l'inventaire lexical en cooccurrences, poly-cooccurrences et trames de cooccurrence.

Nos expériences montrent qu'en fonction du format évolutif des données – texte, tableau et réseau – et étant donné leur mode d'accès respectif – linéarité, tabularité et connexité –, on aborde différemment ces résultats au fur et à mesure du processus d'agglomération lexicale. Ainsi, les focales employées pour le repérage et la hiérarchisation des systèmes de mots associés en contexte dévoilent différentes facettes de cette

29 Du complexe au simple puis du simple au complexe, toute entreprise lexicométrique peut se concevoir comme la combinaison de deux approches fondamentalement contradictoires :

- *analyse* : opération qui décompose un objet en ses éléments constitutifs.
- *synthèse* : opération qui procède du simple au complexe, des éléments au tout.

Par ailleurs, l'étymologie de ces termes livre la nature profonde des opérations qu'elles dénomment : le grec ancien *analysis*, dérivé de *analuein*, signifie 'défaire une trame' et *synthesis* signifie 'poser ensemble'. Aussi, comme la rappelle Vandendorpe [1999], la métaphore textile est au cœur même du matériau textuel : '*Sur le plan du contenu thématique et symbolique, les textes sont souvent loin d'être linéaires. En fait la notion même de texte, qui vient du latin textus, renvoie originellement à l'action de 'tisser, entrelacer, tresser', ce qui suppose le jeu de plusieurs fils sur une trame donnée et, par leur retour périodique, la possibilité de créer des motifs. Ainsi la métaphore visuelle est-elle présente dès les temps les plus anciens. Cet aspect paradigmatique du texte relève de l'ordre spatial. Le processus d'engendrement de la signification au cours de la lecture n'est donc pas forcément linéaire, et des sémioticiens comme Greimas et Courtès ont bien montré que l'existence de textes pluri-isotopiques contredit, au niveau du contenu, la linéarité de la signification.*'

architectonique : les *cooccurrences spécifiques* révèlent des attractions binaires, les *poly-cooccurrences* extraient des systèmes cooccurentiels avérés en contexte, enfin les *trames* dévoilent la totalité cooccurentielle du texte.

A chaque étape de cette méthodologie on constate que les informations lexicales, une fois synthétisées, prennent du relief. En effet, passées au crible de l'analyse cooccurentielle, les données essentielles se libèrent du volume et de la densité textuelles qui les occultent pour reformer un tissu de circonstances explicitant le sens de chaque composante lexicale. En particulier les *trames de cooccurrence*, à mi-chemin entre le texte d'origine et la liste fréquentielle, livrent une version épurée du corpus dont l'exploration, plus déliée que la lecture du texte, permet de cerner d'emblée les zones d'intérêt sémantique. Avec les *hypertrames* on réalise un saut qualitatif de l'information vers la connaissance, substituant à un recensement terminologique une conceptualisation des données qui rapproche le graphe final d'une représentation ontologique du corpus.

Virtualisation, actualisation, réalisation. La méthodologie présentée ici permet de mettre en évidence différents moments de la cooccurrence lexicale en tant qu'objet polymorphe et intrinsèque au texte. En effet, la transformation du texte en espaces de lecture possibles puis le choix d'un parcours singulier et enfin l'interprétation de ce chemin de cooccurrence par une lecture particulière, constituent autant de vues complémentaires d'une architectonique sous-jacente et spécifique au corpus étudié³⁰. En cartographiant l'ensemble des événements relationnels qui caractérisent le lexique employé, il devient possible de matérialiser autrement l'état naturel du texte, et ce, sans pour autant sacrifier ses traits organisationnels. Et, même s'il restitue

30 En dernière instance, le réel du texte est l'interprétation qu'en fait le lecteur, aidé ou non par le réseau cooccurentiel. Comme l'explique Lévy [1995] : 'On peut dire qu'un acte de lecture est une actualisation des significations d'un texte, actualisation et non réalisation puisque l'interprétation comporte une part inéliminable de création. L'hyper-textualisation est le mouvement inverse de la lecture, au sens où elle produit, à partir d'un texte initial, une réserve textuelle et des instruments de composition grâce auxquels un navigateur pourra projeter une multitude d'autres textes. Le texte est transformé en problématique textuelle.'

la structure non-linéaire des idées par opposition au format séquentiel du document d'origine, le réseau de cooccurrence reste un *mode d'être* du texte qui permet d'articuler celui-ci par-delà sa linéarité.

Bibliographie

- Blayo F. & Verleysen M. (1996). *Les réseaux de neurones artificiels*. Paris : Presses Universitaires de France, Collection Que sais-je ?
- Beauchemin J. & Cucumel G. (1995). « Stratégies discursives et test de significativité des cooccurrences lexicales », *Actes des 3^{èmes} Journées Internationales d'Analyse Statistiques des Données Textuelles*, vol. 2, Rome.
- Church K. & Hanks P. (1990). « Word Association Norms, Mutual Information and Lexicography », *Computational Linguistics* 16.
- Delacour J. (1995). *Le cerveau et l'esprit*. Paris : Presses Universitaires de France, Collection Que sais-je ?
- Demonet M., Geffroy A., Gouaze J., Lafon P., Mouillaud M. & Tournier M. (1975). *Des tracts en Mai 68. Mesures de vocabulaire et de contenu*. Paris : Armand Colin et Presses de la Fondation Nationale des Sciences Politiques.
- Harris Z. S. (1968). *Mathematical Structures of Language*. New York : Editions John Wiley and Sons.
- Heiden S. & Lafon P. (1998). *Cooccurrences, La CFDT de 1973 à 1992, Des mots en liberté, Mélanges Maurice Tournier*, tome 1. Fontenay-aux-Roses : ENS Editions.
- Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève : Slatkine-Champion.
- Lebart L. & Salem A. (1994). *Statistique textuelle*. Paris : Dunod.
- Lévy P. (1990). *Les technologies de l'intelligence - L'avenir de la pensée à l'ère informatique*. Paris : Editions La Découverte, Collection Sciences et Société.
- Lévy P. (1995). *Qu'est-ce que le virtuel ?* Paris : Editions La Découverte, Collection Sciences et Société.

W. MARTINEZ

- 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.
- Salem A. (1987). *Pratique des segments répétés*. Paris : Publications de l'InaLF, Collection Saint Cloud, Klincksieck.
- Tesnière L. (1959). *Eléments de syntaxe structurale*. Paris : Klincksieck.
- Tournier M. (1988). *Lexicometria* - Séminaire de lexicométrie, Lisbonne, Universidade Aberta.
- Vandendorpe C. (1999). *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*. Montréal : Editions La Découverte.