Répulsions lexicales : expériences autour de la cooccurrence négative

William Martinez

SYLED, Université de la Sorbonne nouvelle - Paris 3 – will_martinez@hotmail.com

Abstract

Traditionally the study of lexical co-occurrences has focused on detecting phenomena of attraction between words in order to reveal significant pairs, or beyond binary co-occurrences, to unveil more complex systems of word association at work in context. This paper aims to draw attention towards the opposite type of phenomenon that is lexical repulsion or put in the form of a question: does a given pole-word avoid other words? Lexicometric experiments carried out on different types of corpora have enabled the detection of significant cases of words eluding each other. An initial study of these corpora by means of multidimensional statistics reveals different examples of lexical repulsion which are later analyzed as anti-co-occurrences by way of adapted measuring tools.

Keywords: lexicometrics, co-occurrences, negative co-occurrences, anti-co-occurrences, lexical valence, anaphora.

Résumé

La recherche autour des cooccurrences lexicales s'est souvent attachée à mettre en évidence les phénomènes d'attraction entre formes lexicales, ou au-delà des cooccurrences binaires, d'associations plus complexes entre systèmes de formes. Cet article s'intéresse au phénomène inverse, celui de la répulsion lexicale ou posé en d'autres termes : un pôle donné évite-t-il certaines formes ? Des expériences lexicométriques menées sur des corpus différents permettent de relever des cas significatifs d'évitements lexicaux. Une première approche typologique du phénomène grâce à la statistique multidimensionnelle facilite le repérage de répulsions entre des formes que l'on pourra ensuite analyser en tant qu'anti-cooccurrences par le biais d'outils de mesure adaptés.

Mots-clés : lexicométrie, cooccurrences, cooccurrences négatives, anti-cooccurrences, valence lexicale, anaphore.

1. Introduction

Les différentes méthodes consacrées à l'étude des cooccurrences permettent de repérer des couples de formes qui se rencontrent beaucoup plus souvent dans les mêmes phrases que ne le laissent prévoir des calculs fondés sur des modèles probabilistes. En effet, dans son approche des corpus textuels la statistique syntagmatique privilégie ce qui est présent, et surtout, ce qui est présent en masse. Ainsi, même si elle s'attache à des critères contextuels tels que l'orientation et la distance des collocats, l'analyse des cooccurrences vise en priorité à détecter la surreprésentation de certains mots dans le voisinage contextuel d'une forme pôle.

Mise en oeuvre dans de nombreuses méthodes de cooccurrence telles que l'*Information Mutuelle* (Church et Hanks [1990]), les *Cooccurrences Significatives* (Beauchemin et Cucumel [1995]), les *Cooccurrences Spécifiques* (Lafon [1984]) ou encore les *Segments Répétés* (Salem [1987]) et l'*Inventaire Distributionnel* (Salem [1987]), cette statistique des rencontres fréquentes se révèle très efficace en ce qu'elle identifie les attractions lexicales en contexte autour d'un pôle. La priorité accordée dans l'investigation lexicométrique aux

suremplois lexicaux s'accorde tout à fait avec la logique de production à l'origine de nombreux textes. Les *répétitions*, *redites* et *rafales* de mots repérés par l'appareil lexicométrique correspondent dans le discours à des préférences d'emploi sémantiques ou trahissent la récurrence de patrons syntaxiques. Ainsi, dans le cas des textes littéraires, les répétitions lexicales correspondent au développement d'effets stylistiques et de champs sémantiques. Dans les textes politiques, elles instaurent des rituels et on les retrouve à la base des slogans et des formules figées de la langue politique.

En dressant un inventaire hiérarchisé des associations répétées autour d'un pôle, la méthode cooccurrentielle produit une véritable cartographie de son univers lexical. Ce profil distributionnel caractéristique du pôle définit ce que l'on peut appeler sa *valence lexicale*[1] c'est-à-dire sa capacité à attirer d'autres formes de manière récurrente en contexte. Nous proposons d'enrichir cette notion en considérant cette fois les sous-emplois lexicaux autour de certains pôles afin de montrer comment ces derniers se caractérisent également par les formes qu'ils repoussent : leurs *anti-cooccurrents*[2].

Dans les expériences dont nous rendons compte ici, la cooccurrence négative se révèle difficile à étudier car elle élude les méthodes typologiques, résiste aux mesures cooccurrentielles et ne se dévoile qu'au prix d'explorations contextuelles récursives et paramétrages évolutifs de l'appareil de mesure.

2. Approche typologique de la répulsion lexicale

2.1. Présentation

L'analyse typologique d'un corpus fournit une vue synthétique et structurée du texte par la mise en relief des relations entre les formes lexicales et les parties identifiées dans le corpus. Nous exploiterons ici une méthode de la statistique multidimensionnelle[3]- l'Analyse Factorielle des Correspondances - pour l'appliquer à un corpus chronologique, *Affaires Etrangères*, qui réunit les allocutions parlementaires prononcées entre 1986 et 1996 par le Ministre français des Affaires Etrangères[4].

[1] En chimie le terme *valence* désigne le potentiel de liaison d'un atome avec d'autres atomes.

[2] Pour conduire les expériences dont nous rendons compte ici, nous avons exploité deux logiciels. Les modules lexicométriques du logiciel *Lexico3* (Cf., Fleury (S.), Lamalle (C.), Martinez (W.), Salem (A.) *et al.* [2004]) permettent dans un premier temps de transformer les corpus étudiés en bases textuelles auxquelles on peut appliquer une statistique comparative et une analyse typologique. Les modules d'analyse cooccurrentielle du logiciel *Coocs* (Martinez [2003]) permettent une étude complète et détaillée des phénomènes d'attraction et de répulsion qui s'opèrent en contexte autour d'un pôle donné. *Lexico3* est disponible en téléchargement sur le site du Centre Audio-Visuel & Informatique de la Sorbonne nouvelle Paris 3. Le mode d'emploi de ce logiciel ainsi que les modalités de son utilisation (usage privé, usage universitaire, etc.) sont disponibles sur la même page : *www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW*. Le programme *Coocs* est quant à lui téléchargeable sur le site *www.cavi.univ-paris3.fr/ilpga/individus/martinez/accueil.htm*.

[3] Pour plus d'information sur la statistique multidimensionnelle nous renvoyons à Lebart et Salem [1994] et Lebart *et al.* [1995].

[4] Le corpus *Affaires Etrangères* (Martinez [2003]) réunit les allocutions prononcées à l'Assemblée Nationale et au Sénat par les ministres J.-B. Raimond (avril 1986 à déc. 1987), R. Dumas (juillet 1988 à déc. 1992), A. Juppé (avril 1993 à déc. 1995) et H. De Charette (juin 1995 à déc. 1996). Le texte compte 307 discours et 441 803 occurrences pour 16 571 formes et 6 220 hapax.

L'analyse du corpus s'effectue suivant une partition chronologique par année qui divise le corpus en onze parties[5]. L'inventaire des formes suivant cette division du corpus produit un Tableau des Formes Graphiques (TFG) qui contient la fréquence de chaque forme dans chaque partie du corpus. Les données de ce tableau rassemblent 39 050 nombres (11 parties sur 3 550 formes de fréquence ≥10) et constituent le point de départ de l'analyse typologique qui suit.

2.2. Analyse chronologique du corpus Affaires Etrangères

L'Analyse Factorielle des Correspondances (AFC) résume l'information contenue dans le TFG par un jeu de facteurs qui permet de synthétiser la structure de ses lignes et colonnes[6]. Cette méthode décompose l'information, souligne les faits les plus saillants du corpus, et permet d'atteindre les régularités et les ruptures dans la structuration du discours. Une fois les facteurs calculés, ceux-ci sont représentables deux à deux sur un plan graphique sous forme d'axes croisés. Dans l'espace factoriel ainsi créé, l'algorithme permet de situer les variables et les individus statistiques - formes et parties - les uns par rapport aux autres.

Sur la figure 1 qui présente le plan factoriel calculé à partir du corpus *Affaires Etrangères* divisé en onze parties suivant la clef *année*[7], on observe un agencement des points en parabole qui est caractéristique des corpus chronologiques : l'effet *Guttman*. Cette disposition des points est la signature factorielle des *Séries Textuelles Chronologiques* (STC). En effet, dans les compilations de textes écrits sur une longue durée de temps on observe une périodisation des discours qui voient leur vocabulaire se renouveler progressivement suivant un facteur que Salem [1991] nomme le 'temps lexical'.

La courbe d'évolution de la figure 1 est loin d'être parfaite ce qui indique que si le renouvellement du vocabulaire est progressif, il connaît plusieurs accidents. Dans le continuum temporel certaines années se suivent de près et d'autres se singularisent, et malgré le découpage chronologique, des regroupements se produisent qui font apparaître la marque des quatre locuteurs que rassemble le corpus. Autour de chaque ministre un vocabulaire caractéristique traduit des préférences lexicales liées notamment à l'évolution du concept européen ainsi qu'à l'essor des organisations supranationales[8] : *français* et *cee* chez Raimond, *président* de la *république* et les *douze* chez Dumas, *omc*, *onu*, *gatt* et *européen* chez Juppé. Le ministre De Charette quant à lui se distingue par un vocabulaire non spécifique qu'il puise chez son prédécesseur et dans le fonds lexical commun. En effet, au centre du plan on trouve le vocabulaire neutre, employé par tous les locuteurs et uniformément durant les 11 années[9].

---

[5] Dans *Lexico3* l'analyse typologique d'un corpus implique au préalable son découpage en sous-parties que l'on identifie en contexte par un codage méta-textuel sous la forme de balises du type <*clef* = *contenu*>.

[6] Cibois [1994] résume l'objectif de l'opération : '[…] l'analyse factorielle traite des tableaux de nombres et elle remplace un tableau difficile à lire par un tableau plus simple à lire qui soit une bonne approximation de celui-ci'.

[7] Bien que l'analyse porte sur 3 550 formes de fréquence ≥10, par souci de clarté la figure ne présente qu'une sélection de points représentant les années et certaines formes-actants du discours.

[8] Pour éviter la dilution des formes on modifie la casse de toutes les formes du texte si bien que tous les mots y compris les noms propres sont écrits en lettres minuscules.

[9] A ce stade visuel de l'interprétation rappelons que par construction de la méthode factorielle le centre d'un axe représente la moyenne de la population statistique et signale les individus peu spécifiques tandis qu'aux deux extrémités d'un axe on trouvera des éléments qui s'opposent entre eux.

On note que les noms propres *europe* et *france*, l'adjectif et nom *européen*, et enfin les pronoms *je*, *nous* et *elle* appartiennent à cet ensemble de vocabulaire commun.

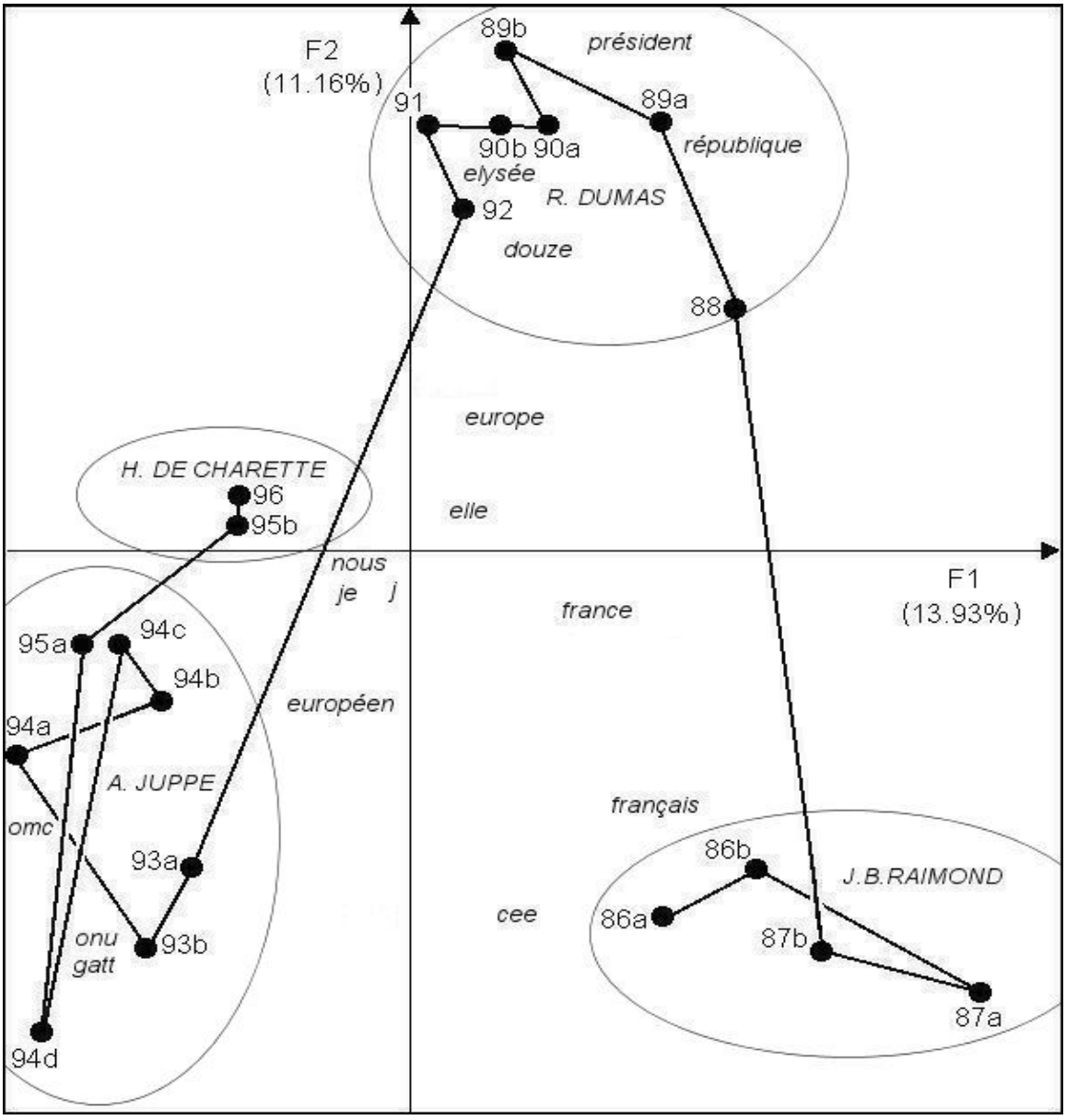

Figure 1 : Plan des facteurs 1 et 2 issus de l'AFC du tableau lexical
[3 550 formes (F ≥ 10) x 11 années]

Guide de lecture de la figure 1 : L'Analyse Factorielle des Correspondances résume les liens d'attraction et de répulsion entre les formes et les parties du corpus sur les deux premiers facteurs. On observe de droite à gauche une évolution chronologique des vingt parties qui restent cependant regroupées en quatre ensembles correspondant aux ministres-locuteurs.

Considérons le quatuor *europe*, *france*, *nous* et *elle* dont la proximité sur le plan factoriel laisse inférer une coïncidence de ces formes dans le texte justifiée *a priori* par la fonction anaphorique pouvant relier indifféremment les deux noms propres aux deux pronoms. Au-delà de cette spéculation en amont, le plan factoriel n'autorise qu'une interprétation limitée sur la cooccurrence de ces formes. En effet, l'AFC révèle la ventilation des formes dans les parties et non pas dans les unités contextuelles. L'information fournie sur la concomitance de

ces formes est donc ambiguë : elles appartiennent au fonds commun du corpus mais apparaissent-elles dans les mêmes contextes ou pas ?

En retournant à leurs contextes de rencontre, on vérifie qu'il s'agit là effectivement de formes ventilées uniformément dans le texte : elles sont présentes dans les allocutions de chacun des quatre ministres-locuteurs et employées également au fil des 11 années de discours. Pourtant, comme le montrent les extraits du tableau 1, ces formes font parfois fi de leur distribution similaire et s'évitent en contexte. On s'aperçoit que la stabilité syntaxique du rapport anaphorique (ou cataphorique) liant les formes *elle* et *france* (exemples 1 à 4) et *elle* et *europe* (ex. 5 à 8), n'est pas du tout reflétée par la configuration contextuelle liant chaque couple de formes. En effet, cet agencement est à chaque fois différent : dans le cas de *elle* on trouve tantôt le pronom dans la même phrase que la forme *france*, tantôt dans la même phrase que la forme *europe*, tantôt avec les deux formes, tantôt avec aucune des deux.

La variabilité de la configuration contextuelle réunissant les formes nous, elle, europe et france est telle que, combinée au volume du corpus et aux fréquences des formes impliquées (respectivement 4 396, 1 309, 1 306 et 2 118 occ.), il est impossible de tirer des conclusions générales sur ces associations lexicales. Et, même si une lecture cursive permet de déterminer le référent du pronom dans chaque cas particulier avec plus ou moins de certitude, on ne peut pas identifier un type d'agencement inter-contextuel ou intra-contextuel particulier qui corresponde systématiquement à une relation anaphorique entre les pronoms elle et nous et les formes europe et france[10].

Guide de lecture du tableau 1 : L'extraction des contextes d'apparition des formes *europe*, *france* et *elle* fournit une série d'exemples montrant la variété de la disposition contextuelle des trois actants qui s'associent de différentes manières dans le texte. Les exemples 1 à 4 montrent qu'autour du pôle *elle* (phrases grisées) - lorsqu'il est anaphoriquement lié à la forme *france* - on observe une orientation, une distance et un nombre des cooccurrents qui varie dans des contextes phrastiques contigus créant ainsi une combinatoire inter- et intra-contextuelle difficilement analysable par les méthodes typologiques. La même variabilité est constatée dans les contextes liant *elle* à *europe* (ex. 5 à 8)

Tableau 1 : Contextes de cooccurrence des formes france, europe et elle

Ex.1 - J.- B., Raimond, 1986

*c'est compte tenu de cette évolution que la **france** avait retiré la plainte qu'**elle** avait déposée avec quatre autres pays devant la commission européenne des droits de l'homme. par la suite, un consensus assez large s'était dégagé au sein du conseil de l'**europe** sur l'amélioration de la situation en turquie et ce pays a été élu à la vice-présidence du conseil de l'**europe** avant de reprendre son tour de présidence en novembre prochain.*

Ex.2 - R. Dumas, 1991

*monsieur le président, mesdames, messieurs les députés, oui, la **france** est présente sur tous les fronts de l'avenir, pour la paix, pour la prospérité, le développement. **elle** a entraîné sur ces chemins faits tour à tour d'embûches et d'embellies l'**europe** et les européens.*

Ex.3 - H. De Charette, 1995

*quand **elle** parle d'**europe**, la **france** ne met pas d'eau dans son vin. la grande affaire de la **france** reste l'**europe**.*

Ex.4 - A. Juppé, 1994

[10] Pour compléter cette approche typologique, une expérience a été menée par Classification Ascendante Hiérarchique (CAH) qui comme l'Analyse Factorielle s'applique au TFG et permet de rapprocher les vocables qui apparaissent souvent ensemble dans les mêmes sous-parties du corpus. Comparant les ventilations de chaque forme, l'algorithme de CAH effectue par étapes successives des regroupements jusqu'à ce que tous ces éléments lexicaux soient unis en un seul ensemble. A l'issue de la classification, la méthode livre, sous la forme d'un *dendrogramme*, un arbre de la hiérarchie des partitions qui reflète certaines thématiques locales de notre corpus mais ignore comme l'AFC les subtilités d'agencement contextuel.

*la **france** a participé à l'élaboration de cette initiative communautaire. c'est **elle** qui, en particulier, a émis l'idée de critères permettant de juger le moment où ces pays seront effectivement en mesure de rejoindre la communauté. il faut bien s'entendre : il ne s'agit pas de faire une manoeuvre dilatoire, mais d'aider les pays d'**europe** centrale et orientale, de les guider sur le chemin de l'adhésion,(...)*

Ex.5 - H. De Charette, 1995

*je pense que ce n'est pas le lieu d'évoquer les questions de financement communautaire, qui correspond d'ailleurs, je vous l'ai dit, à une étape précise dans le calendrier des travaux européens. telle est l'**europe** d'aujourd'hui. **elle** est donc toujours chargée d'autant de projets et, de la part de la **france**, d'autant de volonté.*

Ex.6 - R. Dumas, 1989

*l'**europe** sans monnaie commune, sans banque centrale, sans cohésion fiscale n'était qu'une europe adolescente. mais **elle** serait de la même façon une **europe** étiolée si **elle** n'acquérait pas en même temps sa dimension sociale, voulue par la **france** dès 1982.*

Ex.7 - A. Juppé, 1993

*dans tout cela, l'**europe** a un rôle déterminant à jouer. **elle** est le premier donateur dans la région. 500 millions d'écus annoncés à washington, l'année dernière - et j'ai moi-même proposé qu'au-delà de l'aide aux territoires palestiniens, l'**europe** puisse ajouter 500 millions d'écus pour les autres pays de la région au fur et à mesure que les accords de paix seront signés. j'ai parlé de la jordanie, de la syrie, je voudrais aussi parler du liban, parce que la **france** tient à ce qu'on ne l'oublie pas.*

Ex.8 - A. Juppé, 1993

*c'est pourquoi il faut donner à l'**europe** un nouvel élan, qui lui permette de trouver enfin des solutions à la crise économique qu'**elle** traverse et d'affirmer son existence et son identité politiques dans un monde à la recherche de nouvelles valeurs. comme elles l'ont fait si souvent dans le passé, la **france** et l'allemagne doivent pour cela constituer ensemble une force d'impulsion et de proposition.*

3. Analyse cooccurrentielle de la répulsion lexicale

Qu'il s'agisse d'éléments éloignés sur le plan graphique de l'Analyse Factorielle ou d'agrégats reliés très tard sur le dendrogramme de la Classification Hiérarchique, les résultats obtenus par les méthodes typologiques suggèrent des phénomènes de répulsion entre différentes formes au sein du corpus sans toutefois mesurer leur degré de cooccurrence négative. Afin d'étudier en détail ces cas particuliers, une méthode fondée sur le module hypergéométrique[11] a été développée pour le repérage de cooccurrents spécifiques à l'intérieur de fenêtres d'exploration contextuelle délimitées. Le programme - *Coocs*[2] - détecte tant les formes sur-employées dans l'environnement contextuel d'un pôle que celles qui y sont sous-employées[12].

[11] Fondé sur la distribution en probabilité du nombre de rencontres de toutes les permutations possibles des formes étudiées dans l'hypothèse d'équiprobabilité, le modèle hypergéométrique détermine la valeur la plus probable d'après les paramètres suivants :

T : le nombre d'occurrences dans le corpus
t : le nombre d'occurrences dans les contextes du pôle
F : la fréquence du cooccurrent dans le corpus
f : la fréquence du cooccurrent dans les contextes du pôle

$$P\,[X = f\,] = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

A partir de cette valeur probable on calcule un indice de spécificité (Cf. note 12).

[12] Pour chaque cooccurrent on détermine un diagnostic de spécificité signalant l'écart par rapport à la valeur attendue - un écart qui peut être positif, négatif ou nul. Si la fréquence réelle est supérieure à la fréquence attendue, alors la forme est spécifique positive et nous l'indiquons par le code *+Exx*. Si la fréquence réelle est inférieure à la fréquence attendue, la forme est spécifique négative et nous l'indiquons par le code *-Exx*. Enfin, si la fréquence réelle est égale à la fréquence attendue, alors la forme est banale. La valeur numérique indique le

3.1. Cooccurrences spécifiques

Pour tenter de saisir le lien référentiel existant entre les formes *europe*, *france*, *nous* et *elle*, on peut dans un premier temps appliquer l'analyse des cooccurrences au pôle *france* afin de déterminer ses tendances attractives et répulsives. Dans l'échantillon de phrases où apparaît cette forme (70 386 occurrences soit 16% du volume du corpus), le module lexicométrique relève à la fois les 13 formes qui y sont sur-employées et les 6 formes en sous-emploi (tableau 2). D'emblée on constate une cooccurrence positive avec la forme *elle* (+E13) et une répulsion intense avec la forme *nous* (-E50) qui tend à suggérer une anaphore locale (dans la phrase) avec le premier pronom. En répétant le calcul aux mêmes seuils (spéc. ≥E10 et co-fréq. ≥10) dans la fenêtre d'exploration du paragraphe (tableau 3), on observe que l'attraction *france-elle* se confirme à une plus grande distance du pôle : sa co-fréquence double et sa spécificité passe de +E13 à +E42. En revanche, la répulsion *france-nous* diminue fortement en passant de -E50 à -E13 tandis qu'un cooccurrent négatif émerge aux cotés du pronom collectif : *europe* (-E11).

Tableau 2 : Cooccurrents spécifiques du pôle france (Sp ≥E10) - fenêtre phrase

Cooccurrents positifs				**Cooccurrents négatifs**			
Forme	**F**	**CF**	**Sp**	**Forme**	**F**	**CF**	**Sp**
la	15617	4058	+E50	*nous*	4396	351	-E50
sa	702	229	+E28	*il*	3425	378	-E17
a	3977	847	+E19	*avons*	1024	86	-E13
ses	673	198	+E19	*des*	7507	976	-E13
elle	1309	307	+E13	*les*	7937	1057	-E12
position	198	71	+E12	*notre*	1413	141	-E11
prête	43	27	+E12				
son	865	213	+E11				
entend	76	37	+E11				
rôle	226	74	+E10				
présente	69	33	+E10				
voix	61	31	+E10				
attachée	14	13	+E10				

Guide de lecture du tableau 2 **:** L'analyse des cooccurrences révèle les principales attractions et répulsions autour d'un pôle en comparant, entre autres données, la fréquence globale de chaque cooccurrent (*F*) avec sa co-fréquence (*CF*) dans les phrases où apparaît le pôle et fournit un indice de spécificité (*Sp*) signalant son suremploi (*+Ex*) ou son sous-emploi (*-Ex*). (Cf. notes 11 et 12).

Tableau 3 : Cooccurrents spécifiques du pôle france (Sp ≥E10) - fenêtre paragraphe

Cooccurrents positifs				**Cooccurrents négatifs**			
Forme	**F**	**CF**	**Sp**	**Forme**	**F**	**CF**	**Sp**
la	15617	6362	+E50	*les*	7937	2279	-E14
elle	1309	662	+E42	*nous*	4396	1215	-E13
sa	702	349	+E21	*europe*	1306	318	-E11
a	3977	1520	+E15				
ses	673	311	+E14				
son	865	380	+E12				
entend	76	55	+E12				
liban	210	114	+E11				
présente	69	48	+E10				

degré de probabilité de l'évènement : un indice de E03 signalera une probabilité de 1 sur 1000, E04 une probabilité de 1 sur 10 000, etc.

En reproduisant ce calcul pour chaque pôle, on détermine progressivement un système cooccurrentiel multiple : dans le contexte phrastique le pôle *elle* attire les formes *france* (+E11) et *europe* (+E6) mais repousse *nous* (-E20) alors que le pôle *nous* refoule à la fois *elle* (-E18) et *france* (-E37) sans attirer la forme *europe* qui, elle, n'attire - aux seuils en vigueur - aucun des deux pronoms[13]. De ces comparaisons, on comprend que le nombre de pôles impliqués dans ce système engendre une complexité des liens de cooccurrence qui exige un dispositif de mesure adapté permettant une perception statistique correcte des phénomènes cooccurrentiels impliqués.

3.2. Ecarts de spécificité

En relevant les caractéristiques quantitatives générales de l'activité cooccurrentielle du pôle *france*, on constate qu'elle est de même ordre dans les deux types de contexte - phrase et paragraphe[14]. Ces deux volumes étant proches, nous avons pu les comparer et mettre en évidence les particularités qui distinguent l'univers cooccurrentiel de *france* dans un type de contexte par rapport à l'autre. On observe alors que la très large majorité des formes spécifiques sont réparties parallèlement dans les deux unités contextuelles. Autrement dit, si une forme est en cooccurrence spécifique (positive ou négative) avec le pôle dans ses phrases d'apparition, elle l'est également dans le contexte plus large du paragraphe[15].

Pour trouver des exceptions à cette règle de répartition et dégager des écarts de spécificité notables, il faut baisser le seuil de spécificité des cooccurrences à E05 et rechercher les écarts de spécificité supérieurs ou égaux à 5. On dresse alors deux listes de dissemblances - 6 positives et 14 négatives - observées entre deux types de fenêtre (tableau 4). L'information statistique hiérarchise les cooccurrences suggérées par les méthodes typologiques et donne une mesure exacte des attractions et des répulsions dans chaque type de contexte.

Tableau 4 : Cooccurrents spécifiques du pôle france (F ≥10, Sp ≥E05) classés par écart de spécificité (≥5) entre fenêtre-phrase et fenêtre-paragraphe

Ecarts positifs			
Forme	**Phrase**	**Paragr.**	**Ecart**
europe	-4	-11	7
sa	28	21	7
position	12	6	6
unique	-3	-9	6
ses	19	14	5
acte	0	-5	5

Ecarts négatifs			
Forme	**Phrase**	**Paragr.**	**Ecart**
nous	-50	-13	-37
elle	13	42	-29
il	-17	-5	-12
notre	-11	0	-11
avons	-13	-3	-10
sommes	-7	0	-7
nos	-9	-3	-6
liban	5	11	-6
des	-13	-8	-5
c	-5	0	-5
ont	-5	0	-5
sont	-9	-4	-5
tchad	0	5	-5
absence	-3	2	-5

[13] Dans l'unité contextuelle du paragraphe on observe des phénomènes du même ordre.

[14] On y dénombre respectivement 3 267 et 3514 cooccurrents dont 546 et 701 cooccurrents spécifiques (positifs et négatifs) dont 61 et 59 de spécificité ≥E10.

[15] De fait, on dénombre 301 cooccurrents (soit 92%) de spécificité semblable dans les deux fenêtres (c'est-à-dire avec une différence de spécificité supérieure ou égale à 4).

Guide de lecture du tableau 4 **:** Le calcul des cooccurrences spécifiques autour du pôle *france* livre un ensemble de cooccurrents positifs et négatifs à indices très élevés tant dans la phrase que dans le paragraphe. En comparant les indices dans chaque unité contextuelle on repère des formes avec des écarts importants, positifs pour les uns, négatifs pour les autres. La partie gauche du tableau présente les écarts de spécificité positifs qui correspondent à des cooccurrents qui sont plus fréquents dans les phrases que dans les paragraphes où apparaît le pôle. A l'inverse, la partie droite montre les formes cooccurrentes qui privilégient le paragraphe au détriment de la phrase. On notera qu'un écart positif ou négatif peut résulter de la comparaison de deux indices de spécificité positive, négative ou nulle. Par exemple, une forme à spécificité négative dans les deux types de contexte sera considérée comme privilégiée dans le contexte où elle est moins sous-représentée. C'est le cas de *europe* qui est moins sous-employée dans les phrases (-4) que dans les paragraphes (-11) où apparaît le pôle *france*, et qui se voit attribuer un écart de +7. A l'inverse, on associe la forme *nous* au paragraphe (-13) car elle y est moins absente que dans la phrase (-50), ce qui lui vaut un écart de -37.

Dans la partie gauche du tableau 4 où apparaissent les cooccurrents de *france* privilégiés dans la phrase, on trouve des cooccurrents de nature syntaxique (*sa*, *ses*) et de type sémantique (*position*, *unique*, *acte*) dont la forme *europe*. Dans la partie droite du tableau, on observe que les écarts de spécificité sont bien plus élevés pour les formes privilégiant le paragraphe. Parmi celles-ci on trouve les pronoms *elle* et *nous* qui répondent comme en écho au duo *france-europe* des phrases, et parallèlement, le paradigme *nous-notre-nos* répond aux pronoms *sa-ses*. Cette nouvelle statistique, en reproduisant la configuration contextuelle très variable que nous avons aperçue dans les extraits du tableau 1, dessine deux systèmes cooccurrentiels : le premier sur une orbite intérieure et resserrée autour du pôle *france*, le second sur une orbite extérieure et éloignée. Sans toutefois nous renseigner sur le lien anaphorique entre noms et pronoms, cette représentation reflète la réalité de deux systèmes entremêlés.

C'est en poursuivant la comparaison cooccurrentielle autour du pôle *europe* que l'on précise la fonction référentielle du pronom *nous*. Le tableau 5 montre que la comparaison de l'activité cooccurrentielle autour du pôle *europe* produit, par rapport à l'analyse de *france*, davantage d'écarts (9 positifs et 20 négatifs), mais d'intensité moindre. Concernant le rapport anaphorique entre le pôle et les pronoms, on remarque que *elle* a disparu des cooccurrents privilégiant les paragraphes du pôle. La cooccurrence avec le pronom *nous* avec *europe* quant à elle est jugée plus forte si l'on s'en tient à la spécificité nulle qui signale une attraction 'normale' dans la mesure où elle est prévue par la probabilité (par rapport à -13 dans le cas de *france*, tableau 4).

Tableau 5 : Cooccurrents spécifiques du pôle europe (F ≥10, Sp ≥E05) classés par écart de spécificité (≥5) entre fenêtre-phrase et fenêtre-paragraphe

Ecarts positifs				**Ecarts négatifs**			
Forme	**Phrase**	**Paragr.**	**Ecart**	**Forme**	**Phrase**	**Paragr.**	**Ecart**
occidentale	51	42	9	*européenne*	4	16	-12
et	12	5	7	*nous*	-8	0	-8
de	8	3	5	*union*	11	19	-8
en	8	3	5	*maastricht*	3	11	-8
sécurité	20	15	5	*continent*	7	14	-7
ministère	-9	-14	5	*identité*	6	13	-7
autorités	-6	-11	5	*pas*	-6	0	-6
étranger	-6	-11	5	*traité*	0	6	-6
peuple	0	-5	5	*européens*	2	8	-6
				alliance	6	12	-6
				devons	0	6	-6
				idée	0	6	-6
				il	-8	-3	-5
				cette	-5	0	-5
				n	-8	-3	-5

unique	0	5	-5
perspective	0	5	-5
ouest	5	10	-5
armements	4	9	-5
partenariat	0	5	-5

Les écarts négatifs du tableau 5 montrent qu'autour du pôle *europe* s'élabore un univers cooccurrentiel qui définit un champ sémantique homogène et ce tout particulièrement dans le contexte du paragraphe où l'on retrouve la thématique européenne portée par les formes *union européenne*, *européens*, *alliance européenne* ou encore *continent* et *partenariat*. Du fait de leur nombre ces (quasi-) synonymes du pôle *europe* contribuent à diluer sa référence dans les paragraphes, et ce au détriment du pronom *nous* qui n'est plus la seule anaphore possible du nom propre.

Le système décrit par ces écarts est plus cohérent que celui observé autour de *france*. En effet, en évacuant les formes *france* et *elle*, le réseau cooccurrentiel de *nous* en devient plus interprétable car il n'y a plus d'ambivalence référentielle. Là où le tableau 4 montre le télescopage de deux univers lexicaux, le tableau 5 circonscrit une activité cooccurrentielle spécifique du pole *europe* qui s'associe entre autres formes au pronom *nous*.

4. Conclusion

Dans de nombreuses méthodes de cooccurrence l'information rapportée est souvent constituée du vocabulaire sur-employé dans le voisinage du pôle étudié. Nos expériences autour de la cooccurrence négative montrent que le lexique qui est négligé ou absent est lui aussi révélateur des tactiques discursives en jeu. En soumettant nos données expérimentales à plusieurs dimensions d'analyse par l'application combinée de méthodes typologiques et cooccurrentielles, nous avons observé que l'anti-cooccurrence se manifeste en contexte dans des configurations diverses et sert des stratégies lexicales variées : champs sémantiques mutuellement exclusifs, phénomènes anaphoriques, effets dialogiques... (voir exemples en annexe).

Sur le plan statistique les anti-cooccurrents se définissent comme des formes qui ne sont pas simplement absentes du voisinage contextuel d'un pôle, mais qui faisant fi des lois de la probabilité qui prévoient leur association sont expressément maintenues séparées l'une de l'autre à des distances variables. Pour détecter ces exclusions mutuelles la méthode des cooccurrences négatives se fonde sur les résultats d'explorations réalisées dans deux unités contextuelles différentes - la phrase et le paragraphe. La comparaison de ces résultats montre que la répulsion lexicale locale correspond le plus souvent à une attraction globale et afin d'interpréter ce phénomène d'écho cooccurrentiel nous reprendrons la terminologie de Kintsch et Van Dijk [1978] pour qui la cohérence d'un texte se trouve dans les liens qui se forment entre la *microstructure* (ce qui est dit au niveau de la phrase) et la *macrostructure* (le thème développé d'une phrase à l'autre). Avec l'analyse des anti-cooccurrents nous ciblons précisément cette relation entre deux types de contexte qui correspondent à deux unités de pensée. En observant les écarts de spécificité pour un cooccurrent entre ses apparitions dans le contexte de la phrase et dans celui du paragraphe, on constate des intervalles plus ou moins importants qui révèlent des figures d'évitement lexical et esquissent des structures qui s'opposent dans le discours. Les exemples observés dans le corpus illustrant cet article montrent que la cohérence du discours repose tout aussi bien sur des présences marquées que des absences notables.

Les systèmes lexicaux à l'œuvre dans le texte se dévoilent différemment suivant la méthode que l'on emploie pour les appréhender et c'est le choix de la fenêtre d'exploration qui conditionne la lecture lexicométrique. Entre parallélismes et antagonismes lexicaux, les contiguïtés du texte livrent tantôt des analogies (telle forme implique telle autre), tantôt des oppositions (telle forme exclut telle autre). Les résultats obtenus suivant nos expériences plaident pour une vision dualiste de la valence lexicale qui doit se définir tant par la capacité d'attraction d'un pôle à l'égard de certaines formes que par sa tendance à en repousser d'autres. Aussi, cet enrichissement de la notion de valence s'inscrit dans une vision du texte en tant que système différentiel : là où le cooccurrent contribue à construire le sens du pôle par complétion, l'anti-cooccurrent participe par opposition.

Références

Beauchemin J., Cucumel G. (1995). Stratégies discursives et test de significativité des cooccurrences lexicales. *3es Journées Internationales d'Analyse Statistiques des Données Textuelles*, 11-13 décembre 1995, Rome.

Church K., Hanks P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, n°16.

Cibois P. (1994). *L'Analyse Factorielle*. Collection Que sais-je, Presses Universitaires de France, Paris.

Fleury S., Lamalle C., Martinez W., Salem A., *et al.* (2004). *Lexico3 Textometric toolbox User's manual*. Travaux du *SYLED-CLA2T*, Université de la Sorbonne nouvelle - Paris 3, Paris.

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

Kintsch W. & Van Dijk T. A. (1978), Toward a model of text comprehension and production. *Psychological Review*, n°85 Vol. 5. American Psychological Association, Washington.

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

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

Lebart L., Piron M., Morineau A. (1995). *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

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

Leblanc J.-M., Martinez W. (2006). L'analyse contrastive des réseaux de cooccurrence**.** Le 'monde' dans les discours des présidents de la cinquième République. *8es Journées d'Analyse Statistique des Données Textuelles*, Besançon.

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.

Martinez W. (2005). *COOCS - Outils lexicométriques pour l'analyse des cooccurrences - Manuel d'utilisation.* SYLED-CLA2T (Centre d'analyse de lexicométrie et d'analyse automatique des textes), Université de la Sorbonne nouvelle - Paris 3.

Salem A. (1986). Segments répétés et analyse statistique des données textuelles. Étude quantitative à propos du Père Duchesne de Hébert. *Histoire & Mesure*, Vol. 1, no 2, Paris.

Salem A. (1987). *Pratique des segments répétés*. Publications de l'InaLF, collection Saint Cloud. Klincksieck, Paris.

Salem A. (1991). Les séries textuelles chronologiques. *Histoire & Mesure*, Vol. 7, nos 1/2, Paris.

Annexe - Expériences et résultats complémentaires

A.1. Un cas d'agencement dialogique – corpus Père Duchesne

Un dépouillement initial du corpus *Père Duchesne* - journal publié durant la Révolution Française[16] - montre que les actants sont des composantes importantes de ce texte qui s'inscrit dans le genre du discours-appel et s'organise sous la forme d'un échange entre locuteur et interlocuteurs. De fait, l'analyse des anti-cooccurrents du pôle *je* (979 occ.) révèle des configurations lexicales avec *tu* et *nous* qui structurent le texte de mobilisation :

> « *si **je** m'étais cru, j'aurais mis cette tigresse en chair à pâté, que t'avait fait *marat, lui dis **je** ? **tu** as menti quand **tu** as avancé que **tu** le regardais comme un ennemi de ton pays. toi-même l'as reconnu pour un bon citoyen et un brave bougre, puisque pour le voir, **tu** as cherché à exciter sa pitié. »*

> « *quand j'entendais ces propos de jean-foutres, **je** commençais par examiner ces viédases de la tête aux pieds, et **je** remarquais toujours qu'ils avaient les mains blanches et délicates. ces bougres là, disais-**je**, ne sont que des manoeuvres de contrebande. **nous** autres, gens de fatigue, **nous** ne nous servons pas de pâte d'amande pour avoir de jolis doigts, et le travail est écrit sur nos mains couvertes de poireaux et de durillons. défions-**nous** de ces endormeurs qui viennent moucharder au milieu de **nous**. »*

A.2. Un cas d'agencement sémantique – corpus Voeux

Dans une expérience rapportée par Leblanc et Martinez [2006], une analyse du pôle *monde* dans une compilation d'allocutions présidentielles[17] produit une ségrégation statistique de deux systèmes lexicaux qui expose une structuration particulière du discours : un texte informatif et évènementiel qui est borné, en début et en fin de message, par un vocabulaire de protocole, stable et régulier. Les extraits suivants de discours de F. Mitterrand montrent comment les deux classes de vocabulaire - cooccurrents et anti-cooccurrents du pôle *monde* - coexistent dans l'unité contextuelle du paragraphe tout en s'excluant au niveau plus précis de la phrase :

> « ***mes** chers compatriotes, ce soir **mes voeux** tiendront en quelques mots très simples, ceux que vous emploierez vous-mêmes quand vous vous direz **"bonne année"**. que 1991 vous soit aussi heureuse que la vie le permet, que vous soient épargnées les grandes peines, la souffrance et la solitude, que vous vous sentiez solidaires, là où vous êtes, de ceux qui vous entourent et, d'une façon plus large, que vous ayez l'envie, l'ambition de contribuer au succès de la france qui reste, grâce à vous, l'un des premiers pays du **monde**. **vive** la république ! **vive** la france ! »*

> *« **mes** chers compatriotes, **je** vous adresse **mes voeux** de **bonne** et heureuse **année**. vous penserez ce soir avec moi à ceux des nôtres, qui, partout dans le **monde**, en somalie, au cambodge, en bosnie, portent le message de la france, vous penserez à ceux qui souffrent et qui ont besoin d'amitié. **vive** la république ! **vive** la france ! »*

[16] Publié entre 1793 et 1794, *Le Père Duchesne* est l'organe de presse des hébertistes qui défend l'homme du peuple et dénonce les injustices. Le texte compte 142 177 occurrences pour 10 988 formes. Cf. Salem [1986].

[17] Le corpus *Vœux* rassemblé par Leblanc [2005] réunit les 43 allocutions de Noël du Président de la République française depuis 1959 jusqu'à 2001. Il compte 41 125 occurrences pour 5 201formes.