La cooccurrence, une relation asymétrique ?

Xuan Luong [1], Etienne Brunet [1], Dominique Longrée [2], Damon Mayaffre [1], Sylvie Mellet [1], Céline Poudat [1]

[1] BCL, Université Nice – Sophia Antipolis, CNRS, MSH de Nice – Faculté des Lettres – 98, Bd Edouard Herriot – BP 3209 – 06204 Nice Cedex 3 – France

[2] LASLA – Université de Liège – Bâtiment A4 Quai Roosevelt 1B – B 4000 Liège – Belgique

Résumé

Les attractions cooccurrentielles sont au cœur de bien des pratiques en ADT : qu'il s'agisse de suivre l'évolution des emplois et du sens d'un mot au fil d'un corpus chronologiquement structuré, de dessiner des réseaux thématiques à travers une œuvre, de caractériser un discours, etc., la question se pose régulièrement de contextualiser les emplois des différents vocables. La présente contribution propose une nouvelle piste de calcul dont la nouveauté consiste à mesurer et comparer les associations lexicales au sein d'un corpus en recourant aux matrices de dissimilarités et aux calculs de distances, puis à la représentation graphique par l'analyse arborée.

Abstract

The computation of cooccurrence attraction is central to most tasks in text data analysis: the contextualization of term use is commonly put into question whether attempting to capture word evolution and meaning throughout a chronologically-structured corpus, or drawing a work's thematic network or when characterizing a specific discourse, etc. The present paper follows a new track of computation: the novelty of the method we present lies in the measure and comparison of lexical associations within a corpus using dissimilarity matrices, distance measures and tree representations.

Keywords : cooccurrence, dissimilarity matrix, Latin, political corpus, scientific corpus

1. Introduction

Les attractions cooccurrentielles sont au cœur de bien des pratiques en ADT [1]. Qu'il s'agisse de suivre l'évolution des emplois et du sens d'un mot au fil d'un corpus chronologiquement structuré, de dessiner des réseaux thématiques à travers une œuvre, de caractériser les isotopies d'un discours politique, etc., la question se pose régulièrement de contextualiser les emplois des différents vocables. Dès lors, si la cooccurrence peut être définie, selon la formule de Mayaffre (2008a), comme « la forme minimale du contexte », son traitement devient essentiel dans toutes les pratiques contextualisantes de l'ADT ou dans le cadre de ce qui fut un temps appelé la lexicologie quantitative.

[1] La perspective de cette contribution est ADT c'est-à-dire textuelle, et nous laissons de côté le traitement des cooccurrences dans le cadre du TALN ; notamment le traitement historique des collocations ou des unités phraséologiques, en vue du repérage des expressions idiomatiques, pour la traduction automatique.

Plus généralement encore, on soupçonne les cooccurrences, lorsqu'elles font système, et que la prise en compte de l'ensemble du vocabulaire peut être envisagée, de constituer un facteur primordial de la textualité. A la suite de Viprey (1997 ; 2005 ; 2006), on peut voir en effet le texte comme une entité réticulaire et la *cooccurrence généralisée* comme l'essence de cet entrelacs de mots corrélés, de ce tissu d'associations privilégiées.

A prendre au sérieux l'importance de tels constats, on est conduit à se demander si les outils habituellement utilisés pour rendre compte des attirances réciproques entre un certain nombre de termes lexicaux au sein d'un texte ou d'un ensemble de textes sont suffisamment affinés.

Après un bref rappel sur les méthodes existantes de calcul et de représentation graphique des cooccurrences dont on pointera les atouts et les faiblesses, on explorera une nouvelle piste dont les grands axes méthodologiques sont les suivants : on mesurera et comparera les associations lexicales au sein d'un corpus en recourant à des matrices de cooccurrences représentant un espace vectoriel, au calcul de distances approprié, et à la représentation graphique de celles-ci par l'analyse arborée ; on s'appliquera surtout à tenir compte de la potentielle asymétrie de la relation cooccurrentielle : là réside la principale originalité de notre proposition.

Si ce type d'approche est expérimenté, c'est pour répondre à la double préoccupation qui circonscrit la présente réflexion. D'abord il s'agira moins de considérer l'attirance des mots pris deux à deux, comme le calcul traditionnel de la cooccurrence le fait, que de représenter des systèmes de relations cooccurrentielles entre X termes (une trentaine de lemmes selon les exemples choisis pour cet exposé) : nous raisonnerons donc plus en termes de distributions cooccurrentielles, de réseaux ou de systèmes de cooccurrences que de binômes lexicaux ou de couples de mots. Ensuite, il s'agira moins de prendre en considération l'attirance mutuelle, globale ou moyenne, entre les cooccurrents que d'essayer de mesurer l'asymétrie de leur relation cooccurrentielle, étant entendu que l'attrait du mot A pour le mot B n'est pas nécessairement symétrique à l'attrait de B pour A : à cet effet, nous proposerons de distinguer *l'énergie cooccurrentielle* des mots (la part de ses propres occurrences qu'un mot consacre aux autres – ce que A donne à B, C, D, etc.) et la *disponibilité cooccurrentielle* des mots (la part d'occurrences que chaque mot reçoit des autres – ce que A reçoit de B, C, D, etc.).

2. Calculs de cooccurrence et cooccurrence généralisée

Depuis 60 ans maintenant et les travaux fondateurs de Firth (1951), puis Harris (1957), la communauté ADT comme la communauté TAL a mis au point de multiples algorithmes susceptibles de repérer et chiffrer les cooccurrences d'un texte. Ces algorithmes relèvent de trois types d'approche dont on peut résumer la philosophie ainsi :

- *recherche polarisée* : on étudie les mots qui sont statistiquement associés à un mot-pôle donné. La recherche est clairement orientée du mot-pôle vers ses cooccurrents. Le calcul des spécificités, selon le modèle hypergéométrique, reste dans le domaine français l'indice le plus usité pour ce type d'approche et on le trouvera implémenté dans les grands logiciels comme HYPERBASE (fonction *Thème*) ou LEXICO (fonction *Carte de section*). Martinez (2006), après d'autres, en rappelle le principe dans sa thèse consacrée à la cooccurrence Martinez (2003).
- *recherche systématisée* : dans un texte donné, toutes les paires de mots sont étudiées pour repérer celles statistiquement constituées. La recherche n'est plus orientée. Elle est systématique (toutes les paires) mais binaire (seulement les paires). Quant à l'asymétrie de l'attraction de A pour B (A => B) *versus* de B pour A (A <= B), elle se trouve fondue pour aboutir à un indice d'attirance unique (A <=> B) que l'on pourrait qualifier de moyen ou *mutuel*. Le monde anglo-saxon a alors le plus souvent recours au calcul d'Information Mutuelle présenté par Church et Hanks (1989). La textométrie française utilise le plus souvent l'indice présenté dans la thèse de Lafon (1984) comme le rappelle Heiden (2004).

- *recherche généralisée* : dans le tissu du texte, on étudie le comportement de X mots (idéalement tous les mots du corpus, en pratique une liste, par exemple les 400 premiers substantifs) dont on veut voir les relations et interrelations. A la suite de la thèse de Viprey (1997), on parlera de cooccurrences généralisées ; l'objectif étant de voir l'organisation cooccurrentielle générale du texte (*via* la liste des mots sélectionnés). L'approche n'est ni orientée (pas de mot-pôle) ni binaire (l'objectif n'est pas de constituer des paires) Elle est multidimensionnelle en croisant chaque mot avec tous les autres dans une matrice carrée ; Viprey (1997) propose alors le traitement de cette matrice par l'AFC traditionnelle puis par une AFC géodésique (Viprey 2006).

C'est ici que commence notre analyse en réfléchissant, d'une part, sur la matrice cooccurrentielle de départ et, d'autre part, sur le traitement et la représentation que l'on peut donner de cette matrice.

3. Construire et exploiter une matrice: principes méthodologiques

Soit un ensemble de vocables cooccurrents au sein d'un texte (nous reviendrons plus loin sur la façon de sélectionner cette liste de termes). La première étape, incontournable, consiste à construire la matrice carrée (et parfaitement symétrique) qui donne pour chaque vocable le nombre de fois où il est en cooccurrence avec chacun des autres.

Notre premier principe méthodologique sera de considérer que chaque ligne (et chaque colonne) de cette matrice est un vecteur. L'ensemble des mots constitue un espace vectoriel et chaque mot est une composante relativement homogène de la base. Pour traiter cette base, nous avons opté pour l'utilisation de la distance vectorielle normée. Celle-ci apparaît préférable à la distance du chi² qui se limite à ne comparer que les profils des vecteurs en écrasant les effets de fréquence. Les résultats de ce calcul sont ensuite représentés par un arbre tel que décrit aux JADT par Barthelémy et Luong (1998) sur la base de Luong (1988).

La validité de ce choix théorique qui est ici énoncé *a priori* a, dans notre démarche empirique, été testée sur plusieurs corpus et sur plusieurs matrices : ce n'est qu'au terme de la comparaison des résultats obtenus par le calcul de diverses distances que nous avons définitivement opté pour la distance vectorielle normée.

Mais le choix de la distance, pour délicat qu'il soit, ne constitue pas l'entier du problème. Il faut également déterminer à quelles données on l'applique. Or plusieurs options s'offrent à l'analyste.

- Soit celui-ci se contente de traiter les valeurs absolues qui correspondent à la fréquence observée des cooccurrences de chaque terme avec tous les autres. Cette méthode, quoique donnant des résultats intéressants, a l'inconvénient de travailler sur de simples effectifs et de ne pas tenir compte suffisamment de la fréquence respective de chacun des mots en présence : elle n'intègre donc au calcul ni la probabilité de leur rencontre au sein du texte, ni le poids relatif, dans la rencontre entre deux mots, des contributions respectives de chacun au contexte de l'autre.
- Soit à partir de la matrice des occurrences brutes, on divise le nombre de cooccurrences des mots A et B par la fréquence de A et par la fréquence de B simultanément (opération répétée pour chaque couple co-occurrent). On obtient ainsi une sorte de fréquence relative de la cooccurrence. La probabilité de la rencontre de deux mots au sein du texte est cette fois-ci prise en compte; en revanche, on confond toujours dans une même moyenne l'apport du mot A au contexte de B et l'apport du mot B au contexte de A. Or c'est précisément le point que nous souhaitons remettre en cause et auquel nous voudrions apporter un traitement innovant. D'où la troisième option ici présentée.
- Dans cette dernière méthode, il s'agit donc d'une part de rapporter le nombre de cooccurrences du terme A avec les autres termes B, C, D, E etc. au nombre d'occurrences de A (on mesure alors la part de ses effectifs que le terme A donne aux autres, soit l'« *énergie cooccurrentielle* » qu'il est prêt à consacrer à chacun d'eux); il s'agit d'autre part de rapporter ce même nombre de cooccurrences du terme A avec les termes B, C, D, E etc. au nombre d'occurrences de B, C, D, E respectivement (on mesure alors la part qui, dans l'effectif de B (resp. C, D ou E), est accueillie, peut-être sollicitée par une cooccurrence avec A : on mesure alors la « *disponibilité cooccurrentielle* » de A vis-à-vis de B (resp. C, D ou E).

Une telle conception des rapports cooccurrentiels est synthétisée par la matrice suivante (Tab. 1), dans laquelle N(AB) désigne le nombre de cooccurrences de A avec B, et où A désigne le nombre d'occurrences de A, B le nombre d'occurrences de B, etc.

	A	*B*	*C*	*D*	*E*
A	N(AA/A)	N(AB)/A	N(AC)/A	N(AD)/A	N(AE)/A
B	N(BA/B)	N(BB/B)	N(BC)/B	N(BD)/B	N(BE)/B
C	N(CA/C)	N(CB)/C	N(CC/C)	N(CD)/C	N(CE)/C
D	N(DA)/D	N(DB)/D	N(DC)/D	N(DD/D)	N(DE)/D
E	N(EA)/E	N(EB)/E	N(EC)/E	N(ED)/E	N(EE/E)

Tableau 1 : Matrice cooccurrentielle

On voit alors qu'il suffit d'appliquer le calcul de distances sur les lignes, puis sur les colonnes, pour évaluer la proximité ou l'éloignement de chacun des vocables en fonction soit de leur énergie cooccurrentielle, soit de leur disponibilité cooccurrentielle à l'égard de tous les autres. Si des différences apparaissent dans les arbres représentant ces distances, il conviendra alors d'une part de prendre acte de l'asymétrie de la relation, d'autre part d'interpréter celle-ci en fonction des résultats obtenus sur les différents corpus, éventuellement par un retour aux textes.

4. Applications

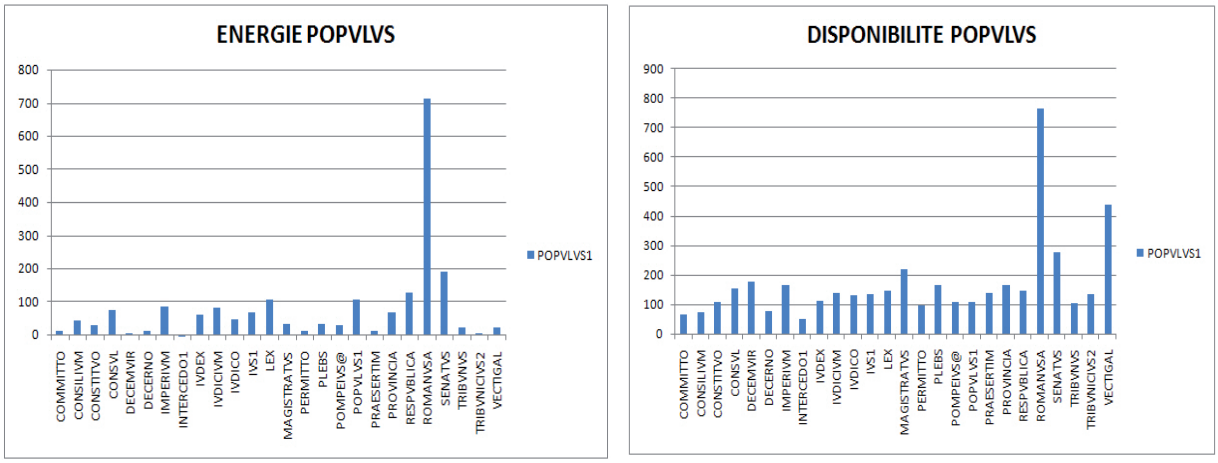
Nous avons choisi d'appliquer cette méthode à trois corpus différents pour en tester l'efficience et la stabilité. Ces trois corpus sont : l'ensemble de la littérature latine classique lemmatisée par le LASLA ; ASLF (224 Articles Scientifiques de Linguistique Française publiés autour de 2000) ; et Sarkozy_2007 (les 37 discours du candidat Sarkozy lors de campagne présidentielle 2007, équivalents à 250.000 mots).

La liste des vocables retenus (lemmes ou formes) est, dans chaque cas, assez restreinte (26 lemmes pour le corpus latin, 35 lemmes pour le corpus Sarkozy, 35 formes pour ASLF) ; elle a été constituée en fonction de la connaissance que chacun d'entre nous avait de son corpus et d'une hypothèse de travail précise. Elle reflète donc un arbitraire orienté par des présupposés de la recherche. Dans deux cas sur trois, la liste rassemble des lemmes qui ont été donnés par le logiciel HYPERBASE comme des cooccurrents spécifiques d'un mot-pôle qui nous semblait intéressant au sein du corpus (les concepts *sens* et *langue* dans ASLF et le mot *potestas* – « puissance, pouvoir » – dans le corpus latin) ; pour ne pas déséquilibrer les rapports cooccurrentiels, le mot-pôle lui-même a ensuite été retiré de la liste. Dans le troisième cas (Sarkozy-2007), ce sont les 35 substantifs les plus *spécifiques* de Sarkozy (*versus* les autres candidats à l'élection, étudiés ailleurs) qui ont été sélectionnés. On envisage à terme de poursuivre l'étude sur des listes plus importantes et plus aléatoires. Mais la place nous manque ici pour multiplier les tests.

4.1. Energie vs disponibilité

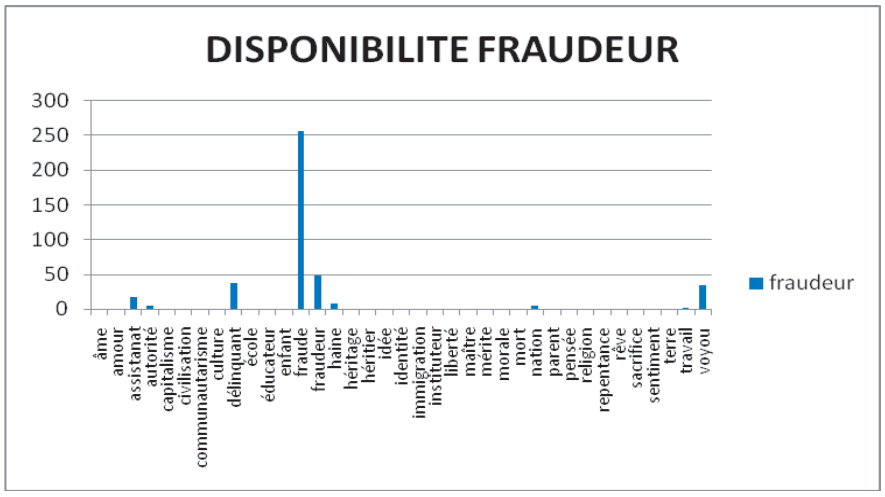
Commençons par illustrer la différence entre ce que nous avons appelé l'énergie cooccurrentielle et la disponibilité cooccurrentielle. Il suffit pour cela de représenter par une courbe respectivement la ligne et la colonne affectées à un mot donné. L'exemple du mot *populus* « peuple » est à cet égard parlant (graphiques 1a et 1b) : ses deux courbes sont assez semblables à l'exception

frappante de son rapport au mot *uectigal* « impôt »: cet effet est dû au déséquilibre entre le nombre d'occurrences de l'un et l'autre mots ; *populus* est très fréquent, *uectigal* beaucoup moins. Pour un nombre donné de cooccurrences, la proportion de son effectif consacrée par *populus* à cette attirance réciproque est sensiblement moindre que celle consacrée par *uectigal*. On peut donc penser que si *populus* donne peu à *uectigal*, on a faiblement besoin de ce dernier pour contextualiser et préciser le sens de *populus* en latin. En revanche, puisque *populus* accueille une très large part de l'effectif de *uectigal* dans son environnement proche, on a fortement besoin de *populus* pour cerner le sens contextualisé de *uectigal* dans la société romaine.



Graphiques 1a et 1b : Énergie et disponibilité de POPULUS

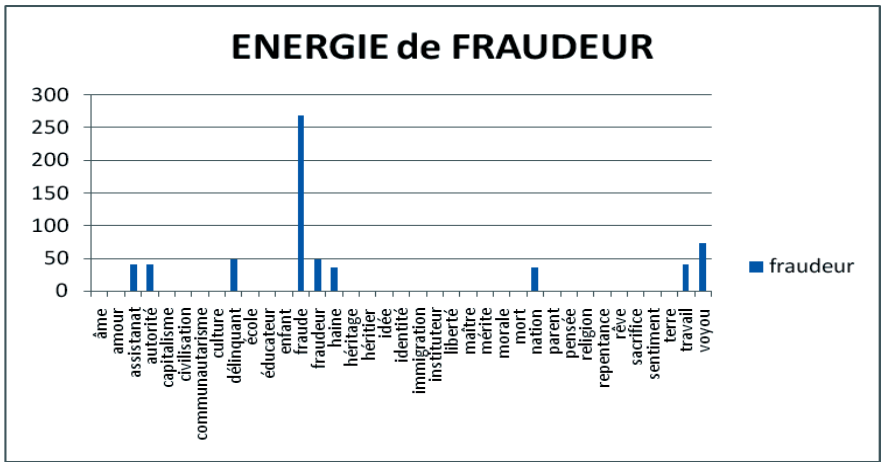
Prenons un autre exemple en français qui permettra de mieux appréhender cette double contextualisation. Le terme *fraudeur* avait fait polémique durant la campagne 2007 ; bien qu'utilisé que 41 fois, il est très spécifique de Sarkozy. Sa disponibilité cooccurrentielle est faible et ne renvoie qu'à des quasi-synonymes (*délinquant*, *voyou*, *fraude*, *assistanat*) dont on comprend, en langue, la parenté (graphique 2).



Graphique 2 : Disponibilité de FRAUDEUR

Fraudeur – sans doute du fait de sa faible fréquence – ne contextualise donc que peu de mots du discours, c'est-à-dire ne contribue que très faiblement à leur sens et, lorsqu'il le fait, c'est pour souligner une proximité lexicale évidente.

En revanche, l'énergie cooccurrentielle de *fraudeur* est plus variée et parfois inattendue (graphique 3) :



Graphique 3 : Énergie de FRAUDEUR

A côté des termes précédemment cités que l'on retrouve intégralement et sur lesquels il est inutile de revenir, nous trouvons ainsi *travail*, *autorité* ou *nation*. Dans ces cas, l'analyse nous renvoie, en discours, à la doxa sarkozienne. Par exemple, le terme *fraudeur* se laisse contextualiser de manière originale par *travail* (c'est-à-dire consacre une part importante de ses occurrences à la cooccurrence avec *travail*) comme dans la citation suivante :

Réhabiliter le TRAVAIL, c'est en finir avec les politiques d'assistanat généralisé, l'impunité des FRAUDEURS et le gaspillage des fonds publics. Quand l'assistanat paie plus que le TRAVAIL, quand la fraude reste impunie, quand l'argent public est détourné ou gaspillé, on démoralise la France qui travaille (Sarkozy, meeting de Meaux, 13 avril 2007)

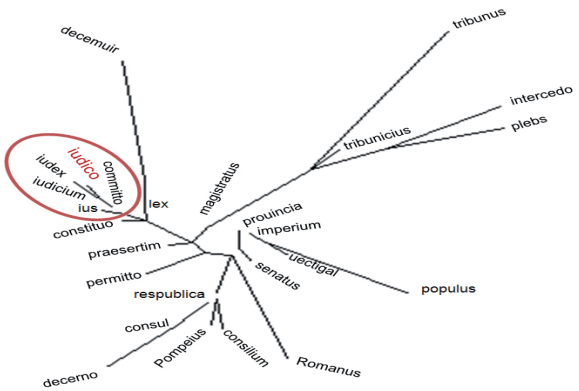
Au final, dans cet exemple, comme pour le couple *populus* et *vertigal*, la fréquence des deux termes *fraudeur* et *travail* joue à plein dans l'asymétrie constatée. Peu fréquent, le *fraudeur* ne saurait contextualiser le *travail*, thème omniprésent et polysémique dans le discours. Très fréquent, le *travail* se laisse le loisir de contextualiser pertinemment le *fraudeur* dans une relation textuelle ici riche de sens.

4.2. Calculs de distances, représentations arborées et variations dans les regroupements

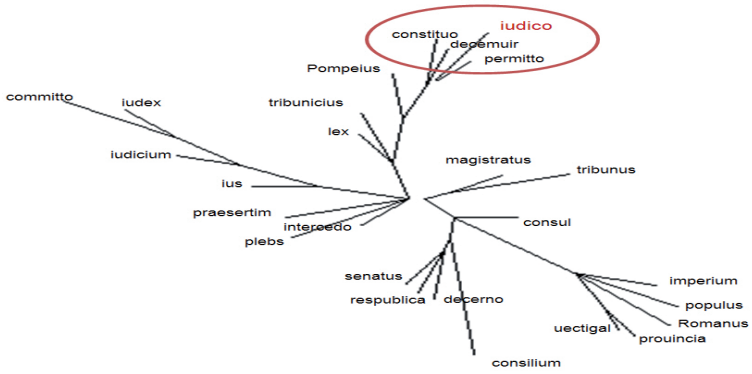
Ayant ainsi trouvé confirmation que la relation cooccurrentielle était asymétrique, il nous reste maintenant à traiter globalement les distances entre les mots en fonction de l'un ou l'autre points de vue (énergie *vs* disponibilité / lignes *vs* colonnes). Cela nous permettra (i) de confirmer l'asymétrie observée au niveau local à celui global du corpus et (ii) de mettre au jour les mots qui se rapprochent sur les deux niveaux de leur énergie et de leur disponibilité cooccurrentielles, ou au contraire, qui manifestent des comportements divergents. On présente donc ci-dessous les arbres issus des calculs de distances en colonnes puis en lignes de chacune des matrices étudiées (graphiques 4a et 4b).

Concernant les classifications des mots issus du corpus latin, on observe deux phénomènes importants :

1) une majorité de mots conserve à peu près les mêmes proximités et éloignements d'un arbre à l'autre ; mais certains se déplacent sensiblement. C'est le cas, par exemple, du verbe *iudico* (« juger, dire le droit ») qui, dans la distance des colonnes, se trouve logiquement rattaché au même noeud que *ius* (« le droit »), *iudex* (« le juge ») et *iudicium* (« le jugement »), alors que dans l'arbre des lignes il quitte cette branche pour aller rejoindre d'autres verbes tels que *permitto* (« permettre ») ou *constituo* (« décider ») ; *de la disponibilité à l'énergie, on passe donc aussi d'un regroupement par champ lexical stabilisé à un regroupement thématique contextualisé.*



Graphique 4a : Arbre des disponibilités cooccurrentielles



Graphique 4b : Arbre des énergies cooccurrentielles

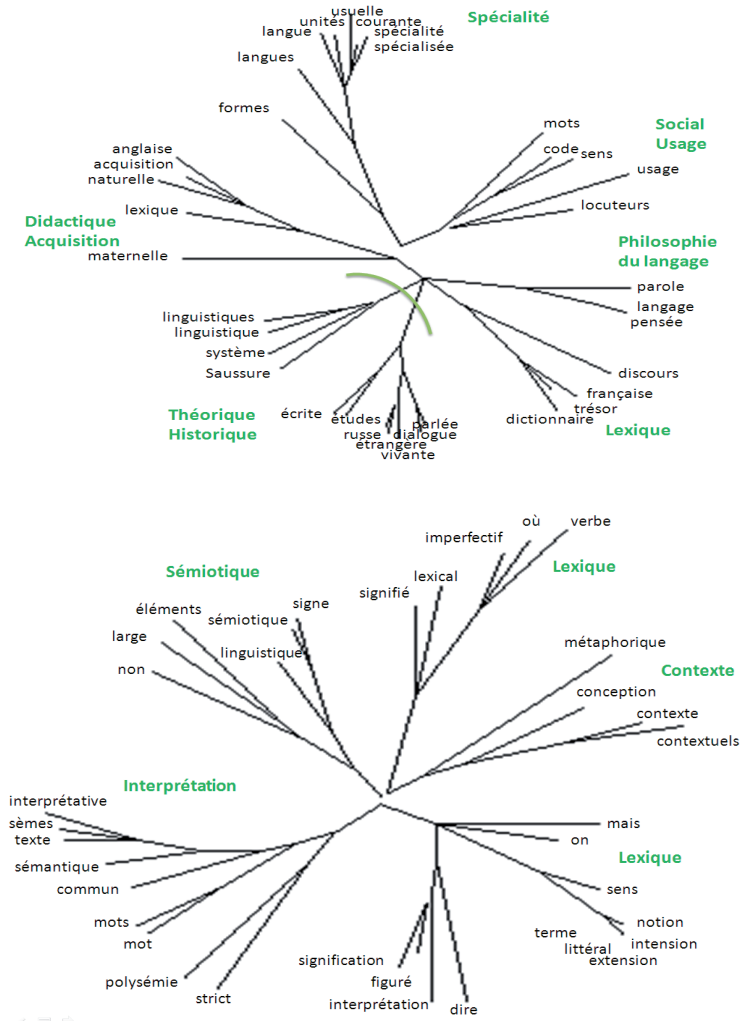
2) De fait, la structure des deux arbres n'est pas la même, non plus que l'interprétation qu'on peut donner aux principaux regroupements : ainsi une bipartition possible de l'arbre des lignes (énergie) oppose deux domaines conceptuels et référentiels, d'un côté celui de la gouvernance de la république puis de l'empire romain, notamment dans ses aspects exécutifs, et de l'autre celui du droit et de la loi. Dans l'arbre des colonnes (disponibilité), la structure est moins nette et l'on observe principalement deux champs lexicaux, celui qui est construit sur le thème *ius*, et celui qui constitue toute la phraséologie associée au tribun de la plèbe dont l'une des fonctions majeures est l'*intercessio* : pour un latiniste, ces rapprochements lexicaux vont de soi.

Or des phénomènes comparables s'observent aussi dans les corpus français ; l'asymétrie cooccurrentielle a ainsi été exploitée dans le corpus ASLF en suivant l'hypothèse [2] que *sens* et

[2] Émise originellement par Rastier (2005).

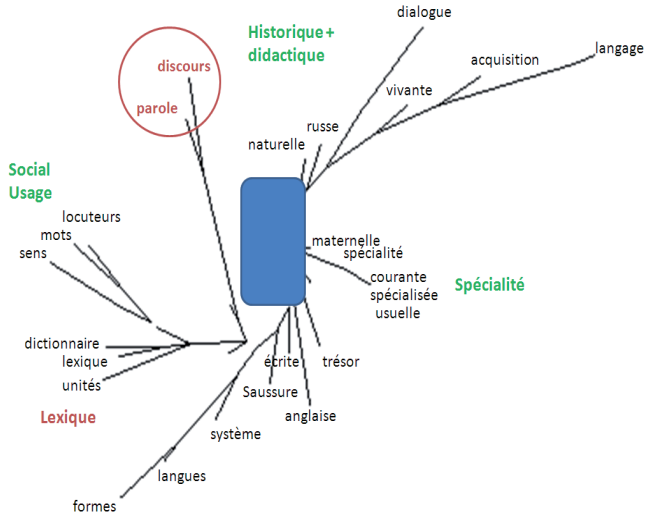
langue étaient respectivement des concepts de forme et de fond : nous avons déjà pu démontrer que *sens* était un thème scientifique débattu, participant à l'évolution des formes sémantiques le long du texte, tandis que *langue* était un concept non saillant, appartenant au fond sémantique général de la linguistique (Poudat, 2009).

Comme ce qui a été remarqué dans le corpus latin, les structures des arbres des lignes sont bien plus nettes que celles des colonnes : les cooccurrents du concept de forme *sens* sont organisés en autant de thèmes débattus autour du concept, tandis que les cooccurrents de *langue* laissent entrevoir les pôles disciplinaires principaux que le corpus contient (graphiques 5a et 5b) :



Graphique 5a et 5b : Arbres des énergies de langue *(en haut) et* sens *(en bas)*

Les mots qui ont les mêmes attirances sont donc regroupés thématiquement, selon les thèmes en usage dans la discipline ; il en va différemment des disponibilités, qui confèrent à chaque mot une distribution passive moins intuitive, car moins contextualisée. Nettement plus déstructurés, les arbres des disponibilités cooccurrentielles de *sens* et *langue* montrent ainsi des oppositions moins marquées et des branches plus rares et moins fournies – les cooccurrents sont significativement plus proches, à tel point qu'un grand nombre d'entre eux n'ont pu être distingués sur le graphe (cf. le conglomérat centré en bleu, graphique 6) :



Graphique 6 : Arbres des disponibilités de langue

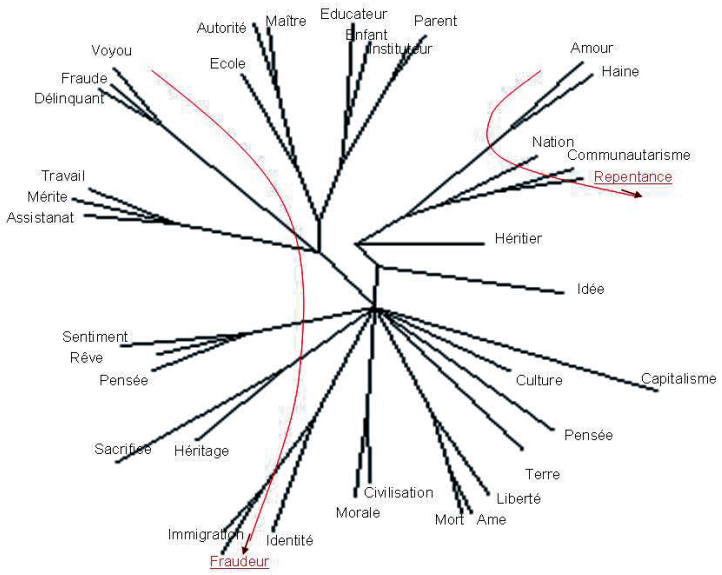
Si l'on ne retrouve pas tous les regroupements que nous avions mis au jour dans les arbres des énergies, on observe certaines persistances dont une particulièrement remarquable : la stabilisation des *langues de spécialité* en énergie et en disponibilité ; le champ thématique et la terminologie associée se superposent ainsi étroitement, ce qui est tout à fait remarquable. On peut enfin souligner la présence de certains regroupements inédits dans l'arbre (7) comparé à (6a) : ainsi *discours* et *parole* se trouvent réunis par l'usage en disponibilité, alors que leur proximité était moins marquée en énergie. Ce phénomène nous semble correspondre à ce qui a été remarqué pour le latin : dans l'arbre des disponibilités, le *discours* et la *parole* sont regroupés dans un même champ thématique stabilisé dans le lexique (hors contexte, *parole* et *discours* sont bien synonymes), tandis qu'ils sont contextualisés de manière différente par l'énergie qu'ils accordent à leurs cooccurrents.

Dans le corpus Sarkozy_2007, les deux arbres donnent les résultats proches pour l'essentiel mais différents dans le détail (graphiques 8 et 9). L'arbre de la disponibilité cooccurrentielle montre des associations lexicales obvies (et cette évidence souligne la pertinence du traitement). Ainsi, par exemple, *école*, *éducateur*, *instituteur*, *parent*, *enfant,* etc. sont-ils rassemblés sur une branche indépendante ; comme, l'on trouvera ensemble *repentance*, *haine et amour* ; ou *âme*, *mort*, *religion*, etc. Sur l'arbre de l'énergie cooccurrentielle, les regroupements sont proches. Pourtant d'intéressants déplacements peuvent être notés. Deux exemples seulement ont été soulignés sur le graphe. *Repentance* jusqu'ici rattaché au champ lexical du sentiment avec *haine* et *amour* est désormais rattaché à celui de la nation (*nation*, *communautarisme*) : il y a là une précision importante qui permet de bien caractériser le néo-patriotisme de Sarkozy durant la campagne, qui confine parfois au néo-colonialisme :

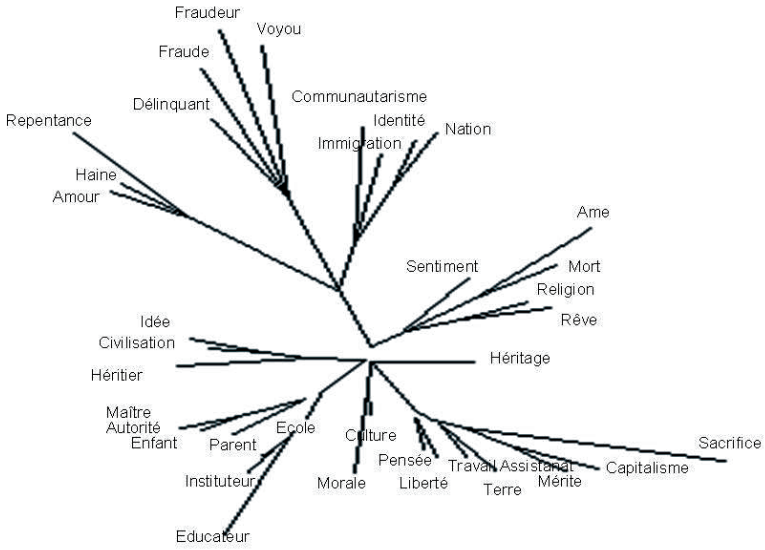
> Je n'accepte pas cette obsession de la repentance qui nourrit la détestation de la France et la détestation de soi. Je ne veux plus de ce dénigrement systématique de l'histoire de France, de ce révisionnisme historique qui n'a d'autre but que la destruction de notre pays en tant que nation (Sarkozy, meeting de Perpignan, 23 février 2007)

De la même manière, *fraudeur*, rattaché sur l'arbre de la disponibilité cooccurrentielle (graphique 7a) à ses para-synonymes (*fraude*, *délinquant*, *voyou*) se trouve désormais rattaché à *immigration* et *identité.* Nous l'avons vu, l'énergie cooccurrentielle de *fraudeur* (*versus* sa disponibilité) est complexe mobilisant des termes importants du discours sarkozien comme

travail, *nation* ou *autorité* et, par là, sa distribution cooccurrentielle le rapproche plus, dans des réticularités qu'il reste à étudier, d'*immigration* ou d'*identité* que de *délinquant* ou *voyou* (graphique 7b).



Graphique 7a : Arbres des disponibilités cooccurrentielles



Graphique 7b : Arbre des énergies cooccurrentielles

La similitude de ces constats et de l'interprétation qu'on est tenté d'en donner nous obligera à poursuivre l'expérimentation : si le phénomène était véritablement récurrent, il indiquerait que non seulement l'asymétrie que nous avons montrée existe bel et bien, mais qu'elle aurait une pertinence allant bien au delà de ce qu'on pouvait penser de prime abord : les différences de comportement en termes d'énergie et de disponibilité cooccurrentielles pourraient alors

coïncider avec les deux facettes du lexique, sa structuration en langue d'une part, fondée sur des apparentements morphologiques et sur l'organisation des sèmes inhérents, son emploi en discours d'autre part, fondée sur des constellations conceptuelles et/ou référentielles et sur l'attraction des sèmes afférents. Hypothèse hardie que nous ne défendrons qu'au terme d'études ultérieures plus approfondies.

Références

Barthélemy J.-P. and Luong X. (1998). Représenter les données textuelles par des arbres. In Brunet, É. and Mellet S., editors, *Actes des 4èmes JADT*, Université de Nice : UMR 6039, pp. 49-70.

Brunet É. (2007). Séquences et fréquences. Mises en œuvre dans Hyperbase. *Lexicometrica* 7.

Church K. and Hanks P. (1989). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, vol. 16, 1 : 22-29.

Firth J.R. (1951). Modes of Meaning. *Papers in Linguistics 1934-51*, *1957*. Oxford: Oxford University Press, pp. 190-215.

Harris Z.S. (1957). Cooccurrence and transformation in linguistic structure. *Language*, vol. 33 : 283-340.

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. In Purnelle, G., Fairon, C. and Dister, A., editors, *Actes des 7èmes JADT*, Presses Universitaires de Louvain, vol. 1, pp. 577-588.

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

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

Luong X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse d'Etat, Paris V.

Luong X., Juillard M., Mellet S. and Longrée D. (2007). Trees and after: The Concept of Text Topology. Some applications to Verb-Form Distribution in Language Corpora. *Literary and Linguistic Computing*, 22, 2 : 167-186.

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans des corpus textuels*. Thèse de doctorat, Université Sorbonne Nouvelle – Paris 3.

Martinez W. (2006). Coocs. Outils lexicométriques pour l'analyse des cooccurrences, en ligne sur http://www.cavi.univ-paris3.fr/ilpga/individus/martinez/download/Manuel%20COOCS.pdf.

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

Mayaffre D. (2008b). Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la cooccurrence. In Heiden, S. and Pincemin B., editors, *JADT2008*, PUL, vol. 2, pp. 811-822.

Poudat C. (in press). Concepts de forme et de fond en linguistique: une exploration en corpus. In Rastier, F. and Valette, M., editors, *Concepts en contexte. Analyses sémantiques de corpus théoriques*.

Rastier F. (2005). Pour une sémantique des textes théoriques. *Revue de sémantique et de pragmatique*, vol. 17 : 151-180.

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

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

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