L'affaire du Mediator au prisme de la textométrie

Philippe GAMBETTE (1) - William MARTINEZ (2)
(1) Univ. Paris-Est Marne la Vallée
(2) ILTEC, Lisbonne

Résumé. Sur la base d'un corpus de plus de 2000 articles de la presse française relatant l'affaire du Médiator, nous appliquons les méthodes de la statistique textuelle afin d'étudier les tendances d'emploi du vocabulaire, les thèmes privilégiés et les stratégies discursives mises en place au fil de la couverture journalistique de l'affaire. Objective et exhaustive, la lexicométrie étudie les fréquences d'emploi des mots pour déterminer la variabilité du discours en distinguant les articles de commentaire et les textes factuels, en opposant les avis scientifiques aux opinions politiques et interprétations journalistiques. En particulier, l'analyse des cooccurrences identifie les ancres conceptuelles du corpus et une représentation des textes dans un nuage arboré permet de visualiser les réseaux de mots qui structurent la narration
Mots-clés. Textométrie, presse écrite, cooccurrences, nuage arboré.

Introduction

Le Médiator - médicament pour diabétiques dévoyé comme coupe faim - et son effet secondaire mortel ont largement défrayé la chronique durant ces dernières années. Volumineuse et dense, l'information diffusée dans la presse sur l'affaire du Médiator depuis fin 2010 est une superposition éditoriale de discours médicaux et législatifs où la mise en scène d'acteurs multiples - victimes, coupables, experts et critiques - forme une trame complexe à interpréter.

La statistique textuelle apporte un éclairage nouveau sur l'exposition médiatique de ce problème de santé publique devenu au fil des pages une affaire d'état. Les méthodes lexicométriques permettent à partir d'une compilation d'articles de presse qui se veut représentative de la couverture journalistique de l'affaire[1] d'aborder le texte de manière objective et exhaustive pour y étudier la distribution des lexies afin d'estimer, entre autres caractéristiques, les fréquences d'emploi des

[1] Afin d'évaluer la représentativité des articles accessibles gratuitement en ligne par rapport, d'une part, aux articles de la presse écrite traditionnelle - gratuite ou payante - et, d'autre part, par rapport aux articles réservés aux abonnés payants des sites des grands quotidiens, nous avons comparé les chiffres de diffusion des uns et de consultation des autres. On s'aperçoit alors, sur la base des résultats produits par ces quotidiens eux-mêmes ainsi que par ceux de l'OJD (Office de Justification de la Diffusion - www.ojd.com), que la presse en ligne - et gratuite - constitue actuellement la principale source de diffusion des textes journalistiques.
A titre d'exemple pour l'année 2011 et concernant l'affaire du Mediator :
- Le Dauphiné Libéré (263041 exemplaires papier tirés par jour versus 766666 pages consultées en ligne par jour) propose 70 articles accessibles gratuitement contre 17 réservés aux abonnés - soit 76% versus 24% du total.
- Le Monde (386177 ex. vs 7433333 consultations) compte 149 articles gratuits contre 61 payants (60% vs 40%).

mots et expressions au fil du développement de l'affaire. On détermine alors la variabilité du discours en distinguant les articles de commentaire et les textes factuels, en opposant les avis scientifiques aux opinions politiques et interprétations journalistiques.

Cet article s'intéresse aux clivages internes du corpus qui permettent d'opposer notamment presse de droite et presse de gauche (I), presse nationale et presse régionale (II) ou encore presse de journaliste et presse d'agence (III).

I - Le corpus de presse et ses clivages

La presse met à disposition de ses lecteurs internautes un grand nombre d'articles en ligne qui couvrent les actualités au quotidien. Ces documents sont pour la plupart accessibles gratuitement et téléchargeables très longtemps après leur parution, ce qui facilite aujourd'hui la compilation de corpus journalistiques et chronologiques. Ainsi, nous avons pu télécharger depuis 22 portails de la presse française[2] un total de 2091 articles parus dans la presse quotidienne nationale et régionale française pour l'année 2011[3]. On obtient alors un corpus de 808 642 occurrences de 18 567 formes lexicales distinctes, auquel on peut appliquer les méthodes de la textométrie afin d'étudier les tendances d'emploi du vocabulaire, les thèmes privilégiés et les stratégies discursives mises en place au fil de la couverture journalistique de l'affaire.

Dans un premier temps on abordera le corpus par le biais de l'Analyse Factorielle des Correspondances (AFC), méthode de statistique multidimensionnelle qui permet de croiser l'ensemble des formes lexicales employées dans le texte avec le titre de chaque publication en vue de produire une carte qui visualise les similitudes lexicales entre les organes de presse. Ci-après, le plan factoriel montre comment les 22 publications se rapprochent ou se séparent du fait du vocabulaire qu'elles emploient dans leurs articles couvrant l'affaire du Mediator en 2011.

---

2 Le corpus a été construit à partir de textes glanés sur les portails suivants (listés par ordre de volume lexical dans le corpus de travail) : Le Progrès de Lyon (93675 occurrences soit 11.6% du corpus total), Le Parisien (93526 - 11.6%), Le Figaro (91200 - 11.3%), Le Nouvel Observateur (88425 - 11%), Libération (71437 - 9%), Le Point (68295 - 8.5%), 20 Minutes (45643 - 5.6%), Le Monde (43746 - 5.5%), France Soir (43654 - 5.5%), La Dépêche du Midi (34230 - 4.2%), Le Républicain Lorrain (265573.3%), Le Dauphiné Libéré (19058 - 2.3%), La Charente Libre (17577 - 2.2%), L'Est Républicain (17086 - 2.1%), La Croix (15919 - 2%), La Provence (9876 - 1.2%), Marianne (8737 - 1.1%), Le Courrier Picard (7529 - 0.9%), Ouest France (7370 - 0.9%), L'Humanité (3202 - 0.4%), Le Journal du Dimanche (1260 - 0.2%) et Valeurs Actuelles (640 - 0.8%). NB : Le Point, Le Nouvel Observateur et Marianne ne sont pas des quotidiens mais leur portail actualisé tous les jours nous permet de les assimiler à des titres quotidiens.

3 Bien que le scandale du Mediator ait éclaté dans la presse dès le mois de mai 2010 et que le volet judiciaire se poursuive tout au long de l'année 2012 (année en cours au moment de la rédaction de cet article), nous avons restreint notre étude à la couverture journalistique de l'année 2011. Nous nous concentrerons donc sur cette période chronologique close considérée comme l'épisode politique de l'affaire, un volet calé entre la dénonciation (2010) et la judiciarisation (2012) avec pour objectif ultérieur d'étendre le corpus et son étude.

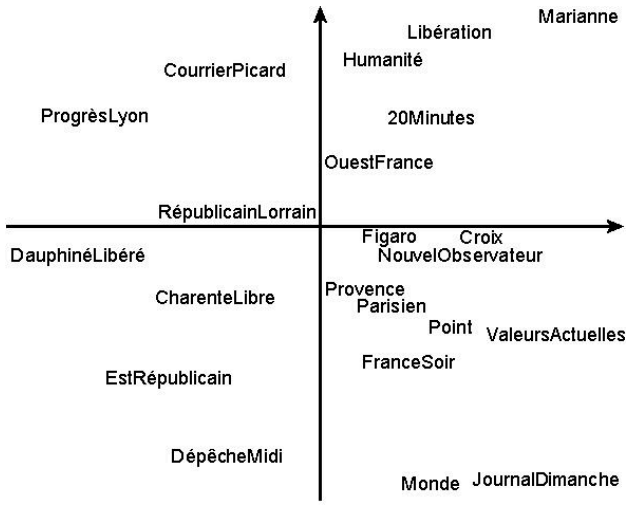


Figure 1 : Analyse Factorielle des Correspondances
22 titres de presse

Suivant une première interprétation toute visuelle du graphe (schématiquement, la proximité graphique indique une proximité lexicale), on constate que l'axe horizontal semble - à quelques exceptions près - distinguer les organes de presse suivant un critère géographique tandis que l'axe vertical les agence suivant un critère plutôt politique. En effet, l'axe horizontal sépare les quotidiens nationaux (à droite du graphe) des régionaux (à gauche) tandis que l'axe vertical sépare les titres de gauche (partie haute du graphe) de ceux de droite (partie basse).
Pour confirmer cette lecture de l'AFC, il faut revenir à la statistique comparative de la *Méthode des Spécificités*[4] qui rend compte des formes qui sont sur-employées dans chaque pan du corpus (presse nationale versus régionale puis presse de droite versus presse de gauche). Ci-après les principales formes spécifiques de chaque catégorie de presse suffisent pour saisir les thématiques et les acteurs[5] mis en avant par chaque type de quotidien.

Presse nationale	Presse régionale
étude, *cnam*, *assemblée*, *autorités*, *intérêts*, *assurance*, *labo*, *mardi*,	*hier*, *mediator*, *victimes*, *j*, *association*, *partage*, *justice*, *ai*, *besançon*, *crci*, *fnath*,

4 La Méthode des Spécificités est fondée sur la distribution en probabilité des mots dans le corpus (cf. Lafon, 1984). Cette statistique prend en compte le nombre total d'occurrences dans le corpus et le nombre total d'occurrences dans une partie du corpus (ex. : les articles de la presse régionale ou encore ceux du mois de juin 2011) pour établir des fréquences théoriques d'apparition des mots. En comparant ces résultats avec les fréquences réelles observées, le modèle statue sur la surreprésentation de certains mots à certains endroits du corpus.

5 Les acronymes des associations et organismes cités sont explicités ci-après :
CRCI - Commission Régionale de Conciliation et d'Indemnisation.
FNATH - Fédération Nationale des Accidentés du Travail et des Handicapés.
JURAVEM – Institut National d'Aide aux Victimes et de Médiation du Jura.
EMA – European Medical Agency.
CNAM – Caisse Nationale d'Assurance Maladie.

commissions, affirme, *document*, *défend*, *laboratoire*, *scientifique*, *maladie* et *entreprises*.

aide, *plaintes*, *médiation*, *dossiers*, *expertise*, *indemnisation* et *conciliation*.

Presse "de gauche"

juravem, *on*, *partage*, *hier*, *association*, *j*, *ai*, *labo*, *prescrire*, *jurassiens*, *je*, *ne*, *peur*, *berra*, *travail*, *rien*, *consommateurs*, *difficile*, *généralistes*, *médecins*, *hirsch*, *houssin* et *médiation*.

Presse "de droite"

desprez-prévost, *escroquerie*, *servier*, *juge*, *tromperie*, *antidiabétique*, *ema*, *nanterre*, *bettencourt*, *hayat*, *témime*, *audience, réforme, douste-blazy, juges*, *chambre*, *examen*, *laboratoire* et *déssaisissement*.

Ces spécificités lexicales majeures montrent que la presse régionale privilégie la question de l'aide et de l'indemnisation des victimes par l'action des associations tandis que la presse nationale s'intéresse aux aspects scientifico-commerciaux de l'affaire et leur traitement par les hautes instances législatives et administratives. Suivant une opposition politique, la presse de droite explore presqu'exclusivement le volet juridique de l'affaire et ses rebondissements au Tribunal de Grande Instance de Nanterre (juges, magistrats et avocats sont souvent cités) tandis que la presse de gauche privilégie les citations directes (*je* et *j*) et l'action de hauts responsables (Directeur et Secrétaires d'Etat à la Santé) tout en couvrant l'action associative régionale.

L'Analyse Factorielle permet une première prise de contact avec le corpus mais pour dépasser ce niveau de synthèse, il faut appréhender le texte directement là où le sens se construit, c'est-à-dire dans le syntagme où la combinatoire des mots élabore les thématiques des articles. Nous emploierons ci-après deux méthodes statistiques dédiées à l'analyse des affinités lexicales pour approfondir deux pistes de recherche : l'opposition entre la presse nationale et la presse de région puis celle entre les articles signés par des journalistes et ceux inspirés par ou recopiés de dépêches d'agence.

II - Presse nationale versus presse régionale

Une fois identifiée l'opposition entre presses nationale et régionale, il faut mettre en évidence les structures lexicales récurrentes qui construisent cette distinction au fil des textes. On exploite pour cela une méthode dite *de cooccurrence*. Cette statistique syntagmatique permet, à partir de leur nombre de rencontres en contexte, de calculer le degré d'attraction entre deux formes lexicales[6]. En réitérant ce calcul pour chaque mot dans le corpus on obtient une matrice volumineuse qui contient l'ensemble des attractions lexicales opérées en contexte. On peut alors

[6] Pour cette analyse nous avons eu recours au logiciel CooCS. Un modèle statistique (cf. note 3) comptabilise chaque coïncidence de deux mots dans une même phrase et estime au final la probabilité d'un tel nombre de rencontres par rapport au volume du corpus.

substituer à ce tableau complexe un graphe qui représente les mots par des nœuds et les attractions par des arêtes.

La figure 2 présente le graphe obtenu pour les formes ayant au moins 400 coïncidences en contexte. Le calcul statistique sanctionnant une attraction exceptionnelle entre différentes paires de mots, des agrégats de plusieurs lexies se forment et quatre sous-ensembles émergent clairement. Une lecture inférentielle permet alors de reconstruire les principales thématiques :

- un médicament pour diabétiques dévoyé comme coupe-faim
- un problème de santé publique
- indemnisation des victimes
- témoignages et discours rapporté

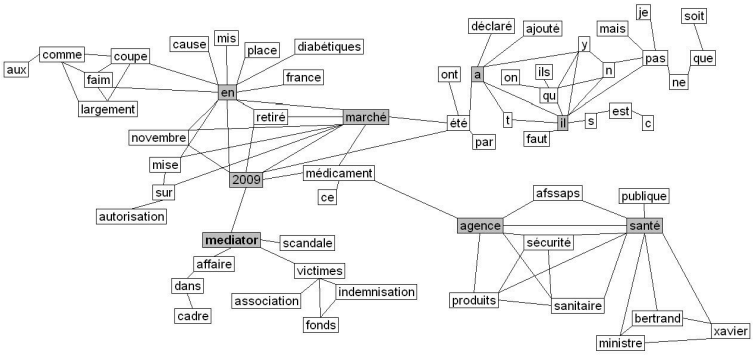


Figure 2 : Trame de cooccurrences

Notre interprétation de la trame se confirme par l'extraction de *segments répétés* (Lebart & Salem, 1994) c'est-à-dire des séquences de mots récurrentes dans le texte qui sont plus explicites que les formes simples. On discerne alors le traitement institutionnel d'un problème de *santé publique* tel qu'il est présenté dans la presse nationale des oppositions binaires *laboratoire-victimes* et *tromperie-indemnisation* opérés dans la presse régionale. Aussi une troisième catégorie de vocabulaire, qui se caractérise par sa neutralité, émerge du fait de sa distribution égale entre presses nationale et régionale.

Presse nationale	Voc. commun	Presse régionale
scandale du mediator	*sécurité sanitaire*	*laboratoires servier*
industrie pharmaceutique	*produits de santé*	*victimes du mediator*
santé publique	*mise sur le marché*	*tromperie aggravée*
inspection générale		*indemnisation des victimes*
assurance maladie		*accidents médicaux*
projet de loi		*effets secondaires*

III - Articles de journalistes versus articles d'agence

On peut aborder le corpus des articles suivant un autre paramètre de rédaction - leur auteur. En effet si l'on assimile les nombreux articles anonymes à des articles écrits par la rédaction du quotidien pour les opposer aux articles explicitement signalés comme des reprises de dépêches, on estime la représentation des agences comme suit : l'AFP (Agence France Presse), l'AP (Associated Press) et Reuters respectivement à hauteur de 529, 5 et 61 articles contre 1496 articles rédigés sans agence (soit 25.3%, 0.2% et 2.9% contre 71,5%).

La visualisation en nuage arboré[7] des deux sous-corpus ainsi créés fait émerger une structuration intéressante du vocabulaire utilisé dans le corpus des articles de journalistes.

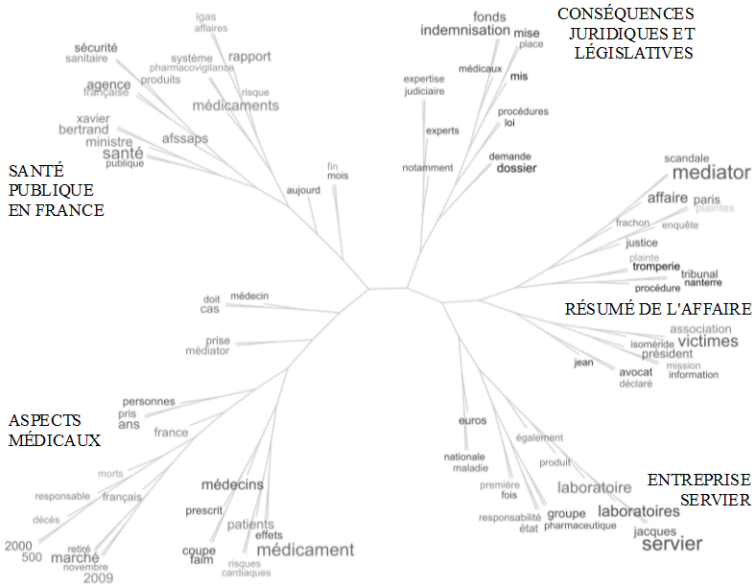


Figure 3 : Nuage arboré des articles de journalistes

L'arbre de la figure 3 peut en effet être découpé en plusieurs sous-arbres thématiques :

- celui du résumé de l'affaire, contenant le mot le plus fréquent, *mediator* et mêlant des mots de couleurs variées ;
- celui des aspects médicaux, aux couleurs plutôt claires ;
- celui concernant le système français de santé, et notamment des organismes comme l'afssaps, l'igas, le ministère, également en couleurs plutôt claires ;
- celui sur l'entreprise Servier, généralement dans un gris médian ;

---

[7] *Un nuage arboré (Gambette & Véronis, 2009) présente les mots les plus fréquents du texte, regroupés au sein d'un même sous-arbre quand ils sont fréquemment cooccurrents dans le texte. Celui de la figure 3 a été construit par le logiciel TreeCloud avec la formule de cooccurrence Liddell (Evert, 2005) sur des fenêtres glissantes de 20 mots de largeur, pour les 100 mots les plus fréquents (en excluant les mots outils). La coloration des mots reflète leur localisation moyenne, en gris clair pour ceux du début du corpus (donc plutôt employés début 2011) et en noir pour ceux de la fin (utilisés fin 2011).*

- celui sur les conséquences juridiques, aux couleurs plutôt foncées, qui dévoile beaucoup moins les aspects législatifs que l'arbre équivalent pour les articles avec agence.

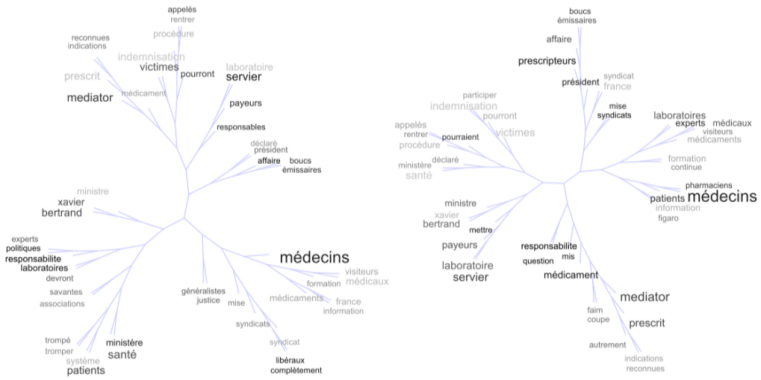
Cette structuration apparaît également, mais de manière plus éclatée, dans le nuage arboré équivalent pour les articles d'agence. Ainsi, les articles d'agence font mieux apparaître les regroupements thématiques du vocabulaire. On peut émettre l'hypothèse que cela provient de sujets d'articles plus ciblés, par rapport à des articles d'agence qui, en proposant des paragraphes récapitulatifs, ou en citant des déclarations générales, brouillent l'organisation thématique en rapprochant des termes ayant une relation sémantique de faible intensité.

Afin de tester cette hypothèse, nous pouvons employer de nouveau la Méthode des Spécificités afin d'identifier le vocabulaire spécifique de chaque sous-corpus.

Articles de journaliste	Articles d'agence
hier, *médiator*, *est*, *c*, *j*, *ai*, *comment*, *effet*, *cas*, *autres*, *médicaments*, *effets*, *troubles*, *service*, *explique*, *pharmaco-vigilance*, *loire*, *car*, *quand*, *cliniques*, *saint*, *jamais*, *mal*, *médecins*, *clair*.	*a*, *t*, *afp*, *assemblée*, *déclaré*, *dit*, *bertrand*, *ajouté*, *avait*, *servier*, *précisé*, *cour*, *ministre*, *ps*, *enquête*, *groupe*, *2000*, *mercredi*, *sante*, *députés*, *gauche*, *selon*, *500*, *bapt*, *xavier*, *indiqué*, *député*, *cassation*, *accoyer*, *millions*.

Outre un lexique signalant le discours rapporté, dans des citations courtes (*déclaré*, *dit*, *ajouté*, *précisé*, etc.), par des acteurs au plan national (*bertrand*, *servier*, *ministre*, *ps*, *groupe*, *députés*, *gauche*, etc.) les articles d'agence privilégient l'aspect législatif (ce qui confirme l'impression donnée par la comparaison des nuages arborés), ainsi que l'aspect judiciaire. Un retour au texte, par l'analyse des contextes de *déclaré* montre une tendance des articles d'agence à enchaîner les citations très courtes, ce qui conduit à des juxtapositions thématiques qui confirment l'hypothèse expliquant la moins bonne structuration du nuage arboré des articles d'agences.

Parmi le vocabulaire sur-représenté dans les articles de journalistes, on constate d'une part la présence de termes indiquant des lieux et interlocuteurs de proximité (*loire*, *cliniques*, *saint*), mais aussi un champs lexical relatif au traitement des patients (*médicaments*, *service*, *pharmacovigilance*, *médecins*, etc.) plus développé.
Pour étudier les emplois de *médecins*, nous avons extrait dans chacun des deux sous-corpus les contextes (10 mots à gauche et 10 mots à droite) de chaque occurrence de *médecins* afin d'en visualiser le nuage arboré, dont les mots sont colorés en noir en fonction de leur degré de cooccurrence avec *responsabilité*, en figure 4. Ainsi, on remarque qu'outre le mot *médecins*, les cooccurrents de *responsabilité* dans les contextes du mot *médecins* sont *servier*, *politiques* et *laboratoires* dans les articles avec agence, alors que dans les articles de journaliste, il s'agit de *question*, *syndicats* et *prescripteurs*, ce qui montre des interrogations plus fortes sur la responsabilité des médecins prescripteurs. Cette démarche d'investigation et de recherche d'explication est cohérente avec la surreprésentation de *expliquer* et *comment* dans le corpus des articles avec agence.



ARTICLES AVEC AGENCE ARTICLES DE JOURNALISTE

Figure 4 : Nuage arboré des contextes de *médecins*

Conclusion

L'approche textométrique employée dans cet article illustre la manière d'exploiter le contenu de la presse en ligne afin d'effectuer des analyses mêlant la démarche exploratoire de recherche d'hypothèses et la confirmation des signaux par l'objectivité des calculs statistiques. La focalisation sur des sous-corpus d'intérêt permet en particulier de faire émerger des contrastes utiles soit pour faire émerger des thématiques intéressantes du sujet étudié, soit des caractéristiques pertinentes sur les sources et auteurs de l'information.

Sur le plan méthodologique, les deux techniques employées ici - réseaux de cooccurrence et nuages arborés - exploitent une même statistique syntagmatique qui met en évidence les attractions entre des mots qui s'écartent sensiblement de la distribution lexicale attendue. Ces méthodes ont permis de relier entre elles différentes variables de production du texte (auteur, géographie, etc.) afin de définir des classes de mots à forte cohérence sémantique puis de rechercher ces catégories dans des segments de message particulièrement caractéristiques du traitement journalistique de l'affaire du Mediator.

Références bibliographiques

Evert S. (2005). *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Thèse de doctorat, Université de Stuttgart.

Gambette, P. & Véronis J. (2009). *Visualising a Text with a Tree Cloud*. In H. Locarek-Junge & C. Weihs (Ed.), *Clasification as a Tool for Research,* Studies in Classification, Data Analysis and Knowledge Organization 40 (pp. 561-570). Berlin Heidelberg : Springer-Verlag.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie,* Paris : Slatkine-Champion.

Lebart, L. & Salem, A. (1994). *Statistique textuelle.* Paris : Seuil.

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels,* Thèse pour le Doctorat en Sciences du langage, Université de la Sorbonne nouvelle – Paris 3.

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

Logiciels

Lexico3 - www.tal.univ-paris3.fr/lexico/lexico3.htm

Coocs - www.williammartinez.fr/coocs

TreeCloud - www.treecloud.org

Une version étendue de cet article avec des graphiques supplémentaires est téléchargeable sur le site *mediator.treecloud.org*.