Thésaurus et corpus de spécialité sciences du langage : approches lexicométriques appliquées à l'analyse de termes en corpus

Evelyne Jacquey, Laurence Kister, Mick Grzesitchak, Bertrand Gaiffe, Coralie Reutenauer, Sandrine Ollinger, Mathieu Valette[1]

UMR ATILF-CNRS-Nancy Université, 44 avenue de la libération 54000 NANCY
Evelyne.Jacquey@atilf.fr, Laurence.Kister@atilf.fr

Résumé Cet article s'inscrit dans les recherches sur l'exploitation de ressources terminologiques pour l'analyse de textes de spécialité, leur annotation et leur indexation. Les ressources en présence sont, d'une part, un thesaurus des Sciences du Langage, le Thesaulangue et, d'autre part, un corpus d'échantillons issus de cinq ouvrages relevant du même domaine. L'article a deux objectifs. Le premier est de déterminer dans quelle mesure les termes de Thesaulangue sont représentés dans les textes. Le second est d'évaluer si les occurrences des unités lexicales correspondant aux termes de Thesaulangue relèvent majoritairement d'emplois terminologiques ou de langue courante. A cette fin, les travaux présentés utilisent une mesure de richesse lexicale telle qu'elle a été définie par Brunet (rapporté dans Muller, 1992) dans le domaine de la lexicométrie, l'indice W. Cette mesure est adaptée afin de mesurer la richesse terminologie (co-occurrents lexicaux et sémantiques qui apparaissent dans Thesaulangue).

Abstract This article aims to contribute to the field of the exploitation of terminological resources for the analysis of technical and scientific texts, their annotation and their indexation. The available resources are on one hand a thesaurus, Thesaulangue, which deals with Linguistics, and on the other hand, a corpus made of samples extracted from five books about Linguistics. More precisely, the article has two goals: first, studying how to determine which terms of Thesaulangue occur in texts. Second, attempting to measure if the lexical units which correspond to terms of Thesaulangue are used in texts in a terminological way or not. In this perspective, the presented work uses and adapts the Brunet's W-index designed in the area of lexicometry.

Mots-clés : sémantique lexicale, terminologie, corpus, richesse lexicale, lexicométrie
Keywords: lexical semantics, terminology, corpora, lexical richness, lexicometry

[1] Nous tenons à remercier chaleureusement les trois évaluateurs de cet article qui, par leurs remarques précises et argumentées, ont su nous aiguiller pour améliorer, nous l'espérons, la précision du contenu et de la structure de celui-ci. Tout élément faux ou incompréhensible résiduel est, bien sûr, de notre entière responsabilité.

1 Problématique et démarche de travail

L'augmentation de la quantité de données textuelles scientifiques, notamment grâce à l'accroissement des publications en ligne et via le recensement des publications dans la base HAL, ouvrent de nombreuses perspectives dont la classification et l'indexation d'articles et/ou d'ouvrages scientifiques, notamment à destination des centres de documentation. De plus, chaque domaine scientifique est en constante évolution et de plus en plus d'interpénétrations entre domaines apparaissent car de nombreuses questions sont abordées de manière multi- ou interdisciplinaire. La difficulté bien connue associée à ces évolutions réside dans la faible précision de l'information obtenue, lors d'une recherche bibliographique, par exemple, sa non exhaustivité et sa résistance à une représentation synthétique. Dans la lignée des travaux de (Aussenac-Gilles et al. 2000), qui ont montré la possibilité d'indexer et de classifier des documents à partir de ressources terminologiques, nous nous proposons d'exploiter un thésaurus en Sciences du Langage, le Thesaulange[2], sur un corpus de linguistique. L'objectif est de définir une méthodologie d'utilisation de ce thésaurus pour contribuer à l'indexation et à la classification de documents du même domaine. Cependant, pour atteindre cet objectif, il est nécessaire de réaliser une désambiguïsation des occurrences des unités lexicales dans les textes lorsque celles-ci correspondent à des termes de Thesaulangue. En effet, le thesaurus utilisé est composé de formes lexicales de termes, non définies, organisées hiérarchiquement dans 19 micro-domaines. Lorsque dans un texte, même de spécialité, apparaît une unité lexicale dont la forme correspond à celle d'un terme de Thesaulangue, on peut se demander à la suite de (L'Homme, 2004) si cette occurrence relève du langage courant ou s'il est de nature terminologique. Lorsque l'emploi étudié est terminologique, on peut considérer que l'on est face à une occurrence de terme. Dans les deux exemples ci-dessous à l'inverse, *transformation* et *moyen* ne sont pas des occurrences de termes bien que leur forme puisse correspondre à deux termes de Thesaulangue dans le micro-domaine de la syntaxe. Ils sont employés ici comme des lexèmes non spécialisés.

1. L'apparition de la grammaire théorique procède d'un besoin de régularité ; mais la **transformation** qu'elle opère est plus illusoire que réelle. [Bally, *Le langage et la vie*, 1952]

2. D'une part, l'emploi constant du même mot au milieu d'un contexte identique risque d'égarer l'esprit, qui, n'ayant pas le **moyen** de préciser par comparaison la valeur du mot, est exposé à la modifier. [Vendryes, *Langage*, 1921]

Pour déterminer la nature terminologique des occurrences d'unités lexicales correspondant à des termes de Thesaulangue, la démarche générale consiste à les repérer automatiquement dans les textes, puis à examiner leur environnement lexical à l'échelle du paragraphe. Nous y mesurons la quantité d'unités lexicales de formes différentes correspondant à d'autres termes de Thesaulangue. Nous faisons alors l'hypothèse que plus cette quantité est importante (nous reviendrons par la suite sur la manière dont cette quantité est mesurée), plus les unités lexicales étudiées se rapprochent d'un emploi terminologique, donc peuvent être considérées comme des occurrences de termes. Enfin, pour un terme donné, plus les occurrences de sa forme lexicale sont considérées comme terminologiques, plus ce terme sera considéré comme un indice fiable pour l'indexation du texte, et plus précisément pour l'indexation des paragraphes dans lesquels ce terme apparaît.

Enfin, puisque le thesaurus utilisé comporte uniquement des formes lexicales de termes non définies, nous utilisons une seconde ressource terminologique. Cette seconde ressource (Gaiffe et al. 2009) est le résultat de l'extraction des définitions du dictionnaire de langue générale, le TLFi, lorsque celles-ci dépendent directement ou par héritage de l'un des domaines suivants : grammaire, lexicographie, lexicologie, linguistique, philologie, phonétique, phonologie, rhétorique, sémiologie, sémiotique, stylistique, toponymie[3]. La comparaison entre les deux ressources terminologiques, le Thesaulangue qui est structuré et l'extraction du TLFi qui comporte des définitions, permet d'associer une définition à 307 termes de

[2] Ce thésaurus a été intégré au portail Termsciences accessible sur le site de l'INIST : http://www.termsciences.fr/. Il comporte 872 termes organisés dans 19 micro-domaines dont par exemple "syntaxe", "sémantique", "sémiotique", etc. Dans la suite, nous avons choisi, par commodité, d'appeler "micro-thesaurus" chaque ensemble de termes relevant d'un des micro-domaines de Thesaulangue.

[3] Les domaines dont nous parlons sont des éléments balisés dans la micro-structure des entrées du TLFi donc aisément repérables.

Thesaulangue, tous micro-domaines confondus. Dans la lignée de travaux précédents comme (Bourigault et Slodzian 1999) ou (Poibeau 2005), nous situons notre problématique à l'intersection de la terminologie et de la linguistique textuelle. Pour l'annotation sémantique de notre corpus, nous utilisons les définitions comme une ressource de traits sémantiques qualifiant les 307 termes de Thesaulangue. Cette ressource est désormais appelée TermTLF. Pour pouvoir à nouveau examiner l'environnement, non plus lexical, mais infra-lexical (traits sémantiques) des unités lexicales dont la forme correspond à un terme de Thesaulangue, les textes du corpus sont annotés sémantiquement à l'aide des traits sémantiques issus de l'ensemble du TLFi. Pour déterminer la nature terminologique ou non des occurrences des formes correspondant à des termes de Thesaulangue, la même hypothèse est appliquée mais cette fois en mesurant la quantité de traits sémantiques de TermTLF qui figurent dans les paragraphes environnants : la mesure de richesse est calculée avec N, le nombre d'occurrences de traits sémantiques issus du TLFi et spécifiques dans l'environnement et V, parmi ces traits sémantiques, ceux qui appartiennent à la ressource terminologique, TermTLF.

2 Données et outils de la démarche

Le corpus utilisé comporte 150.000 occurrences environ et correspond au regroupement de cinq œuvres : *Cours de Linguistique Générale* (F. de Saussure, 1916 [1965], Payot), *Le langage et la vie* (Ch. Bally, 1913 [1952], Droz), *Le Langage* (A. Martinet, éd., 1968, Gallimard), *La linguistique* (J. Perrot, 1963, PUF) et *Langage* (Vendryes, 1921, La Renaissance du Livre). Ces œuvres ont été choisies car elles sont disponibles et balisées dans la base FRANTEXT. Bien que ce corpus soit peu étendu et relativement ancien (1913-1968), il est suffisant pour mettre au point la méthodologie générale de notre approche[4].

Le thesaurus utilisé est un extrait de Thesaulangue. Nous avons choisi de considérer cinq micro-thesaurus parmi les 19 que compte Thesaulangue : sémiotique (201 termes) et syntaxe (185 termes) parce qu'ils comportent un très grand nombre de termes - lexicologie, phonétique et sémantique (29, 46 et 43 termes) parce qu'ils comportent relativement peu de termes. De plus, ces cinq micro-thesaurus ont des structures différentes dans Thesaulangue. Phonétique et sémiotique sont faiblement structurés (4 niveaux de profondeur maximum et une concentration de termes respectivement aux niveaux 3 et 4). A l'inverse pour syntaxe, lexicologie et sémantique, on observe une courbe en cloche atteignant son maximum au(x) niveau(x) intermédiaire(s). Enfin, les termes de ces micro-thesaurus ont des pourcentages de présence dans les textes assez différents : de 48 à 41% des termes des trois plus petits micro-thesaurus sont présents (lexicologie, phonétique, sémantique) alors que les termes des deux plus grands sont présents à environ 26%.

L'annotation sémantique est effectuée par la plateforme SEMY (Grzesitchak et al., 2008) qui utilise les traits sémantiques issus des définitions du TLFi par sélection des formes lemmatisées des lexèmes sémantiquement pleins, à savoir les adjectifs, les adverbes, les noms et les verbes. Pour toute forme lemmatisée d'un texte et de même catégorie que les traits sémantiques du lexique, l'annotation consiste à adjoindre à cette forme l'ensemble des traits sémantiques de l'entrée du dictionnaire correspondante. Ainsi, la forme *transformation* reçoit comme étiquettes sémantiques l'ensemble des traits sémantiques des définitions de la vedette TRANSFORMATION dans le TLFi : /passage/, /phrase/, /forme/, /grammatical/, /règle/, etc.

Le degré de présence de toute forme lexicale, celle d'un terme ou celle de n'importe quel lexème de la langue, est obtenue par le biais d'un calcul de spécificité, calcul lui aussi réalisé par SEMY dans la ligne des travaux de (Lafon, 1980) et qui fournit des résultats comparables à ceux obtenus par le logiciel de lexicométrie Lexico3 (Lebart & Salem, 1994). Le calcul de spécificité étant établi au niveau lexical (formes) et infra-lexical (traits sémantiques), les résultats obtenus sont les co-occurrents spécifiques du "voisinage" d'un terme, au niveau lexical et infra-lexical. Le voisinage d'un terme correspond l'ensemble des paragraphes qui contiennent un terme donné, par texte et sur l'ensemble du corpus. L'intérêt de

[4] Dans un futur proche, celle-ci sera appliquée sur un corpus scientifique et contemporain des sciences du langage (en particulier à l'aide des résultats du projet ANR Scientext qui a récemment mis à disposition le corpus scientifique constitué avec notamment 143 documents relevant de la linguistique, HDR, thèses, articles scientifiques et communication) ainsi que sur un corpus contemporain de vulgarisation en Sciences du Langage (revue des Sciences Humaines, convention en cours de discussion).

disposer de listes de co-occurrents spécifiques est de pouvoir se focaliser sur l'analyse de l'information lexicale et infra-lexicale (les traits sémantiques) qui est la plus significative[5].

La mesure de la quantité de formes ou de traits sémantiques de Thesaulangue au voisinage d'un terme est établie, à la suite de (Ollinger et Valette, 2010), sur la base de l'indice W de Brunet rapporté dans (Muller, 1992) et initialement destiné à comparer la richesse lexicale de plusieurs textes. A la lumière des travaux de (Mouelhi, 2007), l'indice de Brunet fournit des résultats significativement proches de la loi binomiale de Muller, mais est d'une réalisation nettement plus simple. De plus, comme la méthode de Muller, celle de Brunet permet de réduire l'influence des disparités de taille entre les textes à comparer. Or, dans le corpus que nous utilisons, les textes de Martinet et de Saussure représentent environ 80% de la totalité, le texte de Bally représente 12% et ceux de Perrot et Vendryes, respectivement 3 et 4%.

Le calcul de la richesse lexicale selon la méthode de Brunet passe par le calcul de l'indice W avec $W = N^{V^{-}}$: W est égal au nombre d'occurrences d'un texte (N) élevé à la puissance V, le nombre de formes différentes de ce texte, lui-même élevé à la puissance $(-\alpha)$, un coefficient estimé par Brunet à (- 0,172) pour les résultats les plus fiables. La richesse lexicale Rlex est alors égale à l'inverse de l'indice W, résultat que nous multiplions par 100 pour faciliter la lecture. On trouvera dans le tableau suivant, le relevé des données observées[6].

	Bally	Martinet	Perrot	Saussure	Vendryes	Totaux
Occurrences	8358	17437	2110	37508	2803	68216
Formes	3393	3593	1081	9654	1466	19187
Lexicologie	7	8	5	10	5	35
Phonétique	10	10	6	16	1	43
Sémantique	7	10	4	9	4	34
Sémiotique	21	31	5	24	11	92
Syntaxe	38	34	15	45	17	149
Nb_termes	83	93	35	104	38	353

Fig.1 - Fréquences et types observés de termes présents par micro-thesaurus et par texte

En adaptant la méthode de Brunet, nous avons ainsi calculé une mesure de richesse terminologique Rterm, avec N correspondant au nombre de formes différentes du texte et V correspondant au nombre d'occurrences de formes de termes dans Thesaulangue ou de traits sémantiques appartenant à la seconde ressource terminologique TermTLF[7].

3 Résultats

Richesse terminologique des textes du corpus et contributions respectives des micro-thesaurus : la comparaison de la richesse lexicale et de la richesse terminologique (Fig. 2) fait apparaître un contraste marqué entre deux extrêmes : Saussure est le plus riche lexicalement, mais le plus pauvre terminologiquement, tandis qu'on observe l'inverse pour Martinet.

	Bally	Martinet	Perrot	Saussure	Vendryes
Rlex	10,8	9,1	10,0	11,3	10,4
Rterm	2,2	2,3	2,2	1,6	2,0

[5] Afin d'éclairer la nécessité de choisir l'information analysée, nous dirons simplement que le corpus annoté en traits sémantiques est environ 23 fois plus étendu que le corpus brut. Un simple jeu de fréquences ne permet pas d'analyser le contenu lexical et sémantique sur l'unique sélection des co-occurrents fréquents.

[6] Le total est occurrences du corpus est inférieur à la taille initialement mentionnée de 150.000 car nous avons indiqué uniquement le nombre d'occurrences étiquetées en traits sémantiques par SEMY. Nous rappelons que SEMY ignore toutes les occurrences de mots grammaticaux et n'étiquette que les adjectifs, les adverbes, les noms et les verbes.

[7] Pour l'œuvre de Bally par exemple, Rterm est calculée avec N=3393 et V=83.

Fig.2 - Richesse lexicale et terminologique des cinq textes du corpus

Pour expliquer ce premier résultat, on peut tout d'abord noter que 50 ans environ séparent ces deux textes. Le *Cours* de Saussure a été en réalité écrit par ses éditeurs à partir des notes du troisième cours (1910-1911) prises par un de ses étudiants, Albert Riedlinger. D'autre part, Thesaulangue a débuté en même temps que la réalisation du TLFi et la constitution du fonds documentaire de l'InalF. Il en résulte une mise au point de Thesaulangue allant de 1960 à nos jours, puisque le thesaurus est régulièrement maintenu et complété. Par conséquent, la disparité de la richesse terminologique observée peut s'expliquer par un écart tant en terme de temps qu'en terme de maturité scientifique, entre les œuvres. A l'appui de cette explication, on peut ajouter que le second texte le moins riche terminologiquement du point de vue du calcul présenté est celui de Vendryes, très ancien lui aussi puisque nous travaillons avec une édition de 1921.

De manière transversale, les richesses terminologiques par texte et par micro-thesaurus (Fig. 3) montrent que le micro-thesaurus de la syntaxe est le plus présent globalement et en particulier pour Bally, Martinet et Perrot. A l'autre extrême, le micro-thesaurus de la lexicologie est le moins présent, en particulier chez Saussure.

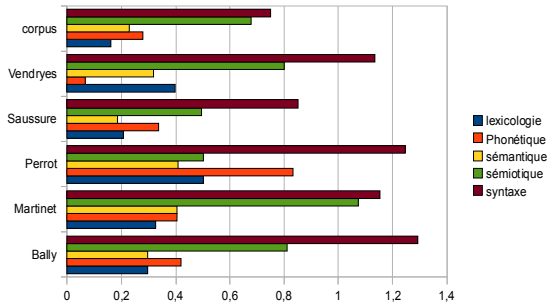

Fig. 3 - Richesse terminologique des cinq micro-thesaurus dans les cinq textes du corpus

Cependant, comme précisé précédemment, les mesures présentées sont établies sur la base de la présence ou non des formes des termes des cinq micro-thesaurus. Or, les occurrences de ces formes peuvent ou non relever d'un emploi terminologique. La section suivante a pour objectif de contribuer à discriminer entre emplois terminologiques et emplois de langue courante.

Caractère terminologique des occurrences de formes de termes dans les textes : afin de déterminer dans quelle mesure les occurrences de formes de termes de Thesaulangue relèvent d'un emploi terminologique ou non, nous avons fait l'hypothèse suivante : *plus les occurrences de la forme d'un terme sont environnées d'autres formes de termes dans le même paragraphe, plus ces occurrences seront considérées comme terminologiques, de même sur le plan des traits sémantiques.* Nous adaptons maintenant cette hypothèse en disant que *le caractère terminologique des occurrences d'une forme de terme peut être approché par la richesse terminologique des paragraphes qui contiennent cette forme*. Afin d'évaluer cette hypothèse, nous procédons comme précédemment à une mesure de la richesse terminologique des paragraphes contenant chaque occurrence de forme de terme pour les deux micro-thesaurus de la lexicologie et de la syntaxe, respectivement le moins présent et le plus présent dans les textes (cf. Fig. 3).

La mesure de richesse est obtenue à partir de l'indice W de Brunet avec N, le nombre total de formes co-occurrentes spécifiques autour des formes de termes de chaque micro-thesaurus présentes dans les textes, et V, le nombre de ces formes qui sont des termes dans Thesaulangue. De même, pour mesurer la richesse terminologique au niveau infra-lexical (traits sémantiques), le même calcul est appliqué mais en confrontant les traits sémantiques spécifiques co-occurrents (N) et parmi eux, ceux qui appartiennent à TermTLF, la ressource terminologique en traits sémantiques.

Pour la lexicologie (Fig. 4), on observe ainsi que des termes potentiellement peu ambigus comme *dictionnaire* ont une richesse terminologique forte par rapport à la moyenne. De manière plus étonnante, le

terme *notions* figure en bonne place. Comme le montre l'exemple « *Et d'abord on se souviendra que les notions sur lesquelles opère la linguistique* », provenant de Bally, cette occurrence ou forme peut être employée comme terme. En revanche, la position du terme *mots* est peu étonnante étant donné le nombre de contextes non terminologiques dans lesquels il peut apparaître. Pour la syntaxe[8], les résultats sont comparables : des termes comme *transformation* (R_formes = 4,55) ou *moyen* (R_formes = 4,2), dont la forme est très ambiguë entre occurrences de terme et de langue générale, ont une richesse terminologique fortement négative par rapport à la moyenne (Moy_R_formes = 9,44). A l'inverse, un terme comme *préposition* (R_formes = 16,67), moins ambigu potentiellement et de manière avérée dans les textes, se trouve dans une position symétrique, c'est-à-dire caractérisé par une richesse terminologique fortement positive par rapport à la moyenne.

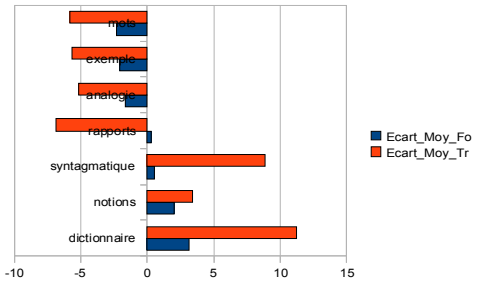

Fig 4 - Richesse terminologique des termes de la lexicologie

4 Conclusion

A l'aide des mesures de richesse terminologique, adaptées de la mesure de richesse lexicale de Brunet, nous pouvons établir une échelle de fiabilité de la présence d'une forme de terme selon l'heuristique suivante : *plus les richesses terminologiques, au niveau lexical et infra-lexical, sont au-dessus de leur moyenne respective pour les paragraphes contenant une forme de terme T, plus les occurrences de ce terme dans les textes pourront être considérées comme un bon indicateur du micro-domaine dans lequel ce terme est positionné dans Thesaulangue*. Cependant, ces travaux pourraient être poursuivis dans deux directions. D'une part, il serait utile de pouvoir mesurer le caractère terminologique de chaque occurrence de forme de terme : pour cela, la mesure de richesse présentée pourrait être calculée pour chaque paragraphe et comparée avec la méthode de désambiguïsation lexicale adaptée de (Véronis 2003). D'autre part, la ressource terminologique lexicale, Thesaulangue, n'est pas satisfaisante en l'état pour l'indexation et la classification manuelle par les documentalistes, notamment du fait de l'absence étonnante de certains termes. Ainsi, le terme *moyen* ne figure pas dans les valeurs possibles de *voix* dans le micro-thesaurus de la syntaxe, mais il apparaît uniquement comme un type possible du terme référant aux *subordonnées circonstancielles*. Or, comme le montre l'extrait de Saussure ci-dessous, cette lecture est possible et elle est de nature terminologique.

3. Dans le parfait grec, à côté de l'actif *Pépheuga, *Pépheugas, *Pephéugamen, etc., tout le **moyen** se fléchit sans *A : * * , et la langue d'*Homère nous montre que cet *A manquait anciennement au **pluriel** et au duel de l'actif / * * /.

Par conséquent, il pourrait être intéressant d'adapter la méthodologie présentée afin de proposer de nouveaux termes pour intégration dans Thesaulangue. Mais pour pouvoir proposer une intégration accompagnée d'une proposition de position, il serait nécessaire de calculer la richesse terminologique relativement à un même micro-thesaurus et non plus seulement relativement à Thesaulangue ou à TermTLF dans son ensemble.

[8] Ce micro-thesaurus comporte 185 termes dont 59 sont présents de manière spécifique dans le corpus. Faute de place, les données ne sont pas reproduites ici.

Références

AUSSENAC-GILLES N., BOURIGAULT D. (2000). The Th(IC)2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents. Proceedings *EKAW'2000 workshop : Ontologies and texts,* Juan-Les-Pins, octobre, 71-78.

BOURIGAULT D., SLODZIAN M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, 19, 29-32.

GAIFFE B., JACQUEY E. & KISTER L. (2009). Approche lexico-sémantique de l'extraction termino- logique : utilisation de ressources lexicographiques et validation sur corpus. *In Actes de la conférence Toth'09,* Annecy.

GRZESITCHAK M., JACQUEY E., BAIDER, F. (2008). Annotation sémantique : profilage textuel et lexical. Actes de la conférence Lexicographie et Informatique : Bilan et Perspectives. Nancy, France.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS,* 1, 128–165.

Presses de la Fondation Nationale des Sciences Politiques.

LEBART L. & SALEM, A. (1994). *Statistique Textuelle*, Dunod, 344 p.

L'HOMME M.C., (2004), *La terminologie : Principes et Techniques*. Presses Universitaires de Montréal.

MOUELHI, Z. (2007). La richesse lexicale dans une perspective de lexicométrie arabe. Etude de cinq méthodes de mesure. *Texte et Corpus, Actes de la journée Corpus 2007*, 271-284.

MULLER CH. (1992). *Principes et méthodes de statistique lexicale*, Larousse, 1977, réimpression Champion-Slatkine, 1992, 211p.

OLLINGER S., VALETTE M. (2010), « La créativité lexicale : des pratiques sociales aux textes », *Actes del I Congrés Internacional de Neologia de les llengües romaniques (CINEO'08) (Barcelona, 07-10 maig 2008)*, M. Teresa Cabré i Castellví, Ona Domènech Bagaria, R. Estopà Bagot, Judit Freixa Aymerich, Mercè Lorente Casafont (Ed.), Publicacions de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), pp. 965-876.

POIBEAU T. (2005). Parcours interprétatifs et terminologie. *Actes TIA 2005*. Rouen.

VÉRONIS J. (2003). Hyperlex : Cartographie lexicale pour la recherche d'information. http://www.up.univ-mrs.fr/veronis/pdf/2003-hyperlex-rapport.pdf.