Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation

Sylvain Hatier
LIDILEM, BP 25, 38040 Grenoble Cedex 09
sylvain.hatier@u-grenoble3.fr

RÉSUMÉ

Nous présentons dans cet article les premiers résultats de nos travaux sur l'extraction de mots simples appartenant au lexique scientifique transdisciplinaire sur un corpus analysé morpho-syntaxiquement composé d'articles de recherche en sciences humaines et sociales. La ressource générée sera utilisée lors de l'indexation automatique de textes comme filtre d'exclusion afin d'isoler ce lexique de la terminologie. Nous comparons plusieurs méthodes d'extraction et montrons qu'un premier lexique de mots simples peut être dégagé et que la prise en compte des unités polylexicales ainsi que de la distribution seront nécessaires par la suite afin d'extraire l'ensemble de la phraséologie transdisciplinaire.

ABSTRACT

EXTRACTION OF ACADEMIC LEXICON'S SIMPLE WORDS IN HUMANITIES WRITINGS

This paper presents a first extraction of academic lexicon's simple words in french academic writings in the fields of humanities and social sciences through a corpus study of research articles using morpho-syntactic analysis. This academic lexicon resource will be used for automatic indexing as a stoplist in order to exclude this lexicon from the terminology. We try various extraction methods and show that a first simple words lexicon can be generated but that multiwords expressions and words distribution should be taken into consideration to extract academic phraseology.

MOTS-CLÉS : corpus – écrits scientifiques – lexique - phraséologie

KEYWORDS : corpus – scientific writings – lexicon - phraseology

1 Introduction

Notre but est la constitution d'une ressource lexicale du lexique scientifique transdisciplinaire en procédant à son extraction automatique. Ce travail s'inscrit dans le cadre du projet ANR Contint Termith (Terminologie et Indexation de Textes en sciences Humaines), dont le but principal est l'indexation automatique d'écrits scientifiques en sciences humaines. Cette indexation requiert l'identification de la terminologie afin de lister les termes les plus significatifs pour un document donné, terminologie dont un des principaux critères de reconnaissance est la spécificité (propriété des mots statistiquement sur-représenté dans un corpus en comparaison d'une référence). Or, les écrits scientifiques font une large part à un autre lexique spécifique, le lexique scientifique transdisciplinaire, que l'on peut définir comme le lexique servant à la description et la présentation de l'activité scientifique (Tutin, 2007c). Cela nous amène à identifier ce lexique, afin de diminuer le bruit qu'il provoque lors de l'indexation automatique en l'utilisant comme filtre d'exclusion.

Nous nous limiterons ici à l'extraction des mots simples, en nous restreignant aux catégories syntaxiques des noms et des verbes. Nous testons, à l'instar de (Paquot, 2010), différentes méthodes statistiques et procédons à une évaluation humaine des extractions résultantes pour identifier la plus performante. À terme, ces lexiques constitueront une ressource couvrant l'ensemble de la phraséologie de l'écrit scientifique telles les expressions polylexicales (collocations, expressions figées, etc.), et devront être structurés selon une typologie restant à déterminer (syntaxique, notionnelle, fonctionnelle, rhétorique) afin de permettre des applications telles que la caractérisation automatique de documents, l'aide à la rédaction scientifique (pour natifs ou apprenants), l'identification des segments introducteurs de définition ou dénomination.

Notre travail est ciblé sur les écrits dans les sciences humaines et sociales, ceci pour plusieurs raisons. Le projet Termith, auquel nous sommes associé, a comme objet d'étude les écrits en sciences humaines. La distinction entre lexique scientifique transdisciplinaire et terminologie est plus complexe pour les sciences humaines que pour les sciences exactes, dans la mesure où la frontière inter-lexiques y est davantage indéterminée. Enfin, dans une optique de développement de ressource à but pédagogique, un tel travail s'avère particulièrement utile en sciences humaines et sociales où l'écriture académique se révèle plus complexe.

Après avoir présenté dans un premier temps les caractéristiques de l'écrit scientifique et des lexiques le composant, nous reviendrons sur les travaux traitant de notre objet de recherche. Nous détaillerons par la suite la méthodologie d'extraction avant d'analyser les résultats puis nous conclurons sur les apports et limites de notre procédure.

1.1 Écrits scientifiques et lexiques associés:

1.1.1 L'écrit scientifique

Le genre de l'écrit scientifique est particulièrement normé, homogénéisé. Il est fonction de la communauté de discours dans laquelle s'inscrit le scripteur et à laquelle est adressé le discours (Swales, 1990). Dans l'écrit scientifique sont combinés plusieurs types de

lexique : lexique scientifique transdisciplinaire, lexique « abstrait » général (non spécifique aux écrits scientifiques mais très fréquent en rapport de la langue générale), lexique terminologique (lié à la discipline, non traversant), lexique de la langue générale (défini par exclusion des lexiques précédents). Le lexique scientifique transdisciplinaire fait donc partie d'un continuum de lexiques aux frontières floues, lexiques de langue spécialisée dont l'univocité et la monosémie ne sont qu'apparentes (Bertels, 2007). L'extraction d'un lexique commun aux écrits en sciences humaines et sociales, dont la langue, comme le note (Blumenthal, 2007), est différente de celle des sciences exactes, nous permet de mieux caractériser la production des savoirs. De plus, cette extraction participe à concrétiser l'existence d'une communauté de discours qui donne sens à la notion de « transdisciplinarité ».

1.1.2 Définition du lexique scientifique transdisciplinaire

A la suite de (Tutin, 2007a), nous définissons le lexique scientifique transdisciplinaire (désormais LST) comme le lexique renvoyant au discours sur les objets et les procédures scientifiques. Il est par nature non terminologique, et a pour fonction la désignation des procédures et outils de l'activité scientifique. (Da Sylva, 2010) le décrit comme abstrait et largement transdisciplinaire. Pour (Drouin, 2007), le LST se situe au cœur de l'argumentation et de la structuration du discours et de la pensée scientifique. C'est donc un lexique méta-scientifique et méta-discursif (c'est-à-dire qui prend pour objet le discours lui-même).

Le LST a pour principales propriétés d'être :

- transversal aux différentes disciplines, donc réparti dans différents corpus disciplinaires. Ce critère exclut la terminologie, intra-disciplinaire et thématique.
- spécifique à l'écrit scientifique étudié ici, donc absent ou moins fréquent dans la « langue générale » qui sera représentée dans cette étude par un lexique général du français.

Nous présentons ci-dessous un extrait d'article de recherche en psychologie, pour illustrer les différents lexiques présents dans ce genre d'écrit.

Les segments en **gras** appartiennent au LST ou au lexique abstrait général, les segments soulignés à la terminologie.

> […] l'organisation matricielle a **renforcé** et **multiplié** les situations de coopération directe tout au long du **processus**. Zarifian (1996) **estime** qu'on a assisté à un **changement** de **paradigme** et **identifie** une « version faible » de la coopération qui **prévaut** dans les organisations traditionnelles de la conception. **L'objectif** est d'assurer une bonne coordination du travail.[1]

[1]Françoise Darses « Résolution collective des problèmes de conception », *Le travail humain* 1/2009 (Vol. 72), p. 43-59.

2 Travaux sur le lexique de l'écrit scientifique

Plusieurs travaux ont porté sur un lexique spécifique aux écrits scientifiques, majoritairement en anglais. (Coxhead, 2002), par exemple, dans un but didactique, a extrait, en se basant sur les fréquences, une liste de mots anglais. Pour le français, (Phal, 1971) a analysé ce vocabulaire général d'orientation scientifique sur un corpus de manuels scolaires et d'ouvrages non universitaires concernant les sciences dures. Peu d'études se sont donc intéressées à l'écrit scientifique en français dans le domaine des sciences humaines et sociales.

Les différences de procédure, de méthodologie, entre sciences expérimentales et sciences humaines se retrouvent dans les lexiques. Or, la majorité des travaux ont porté sur les sciences expérimentales, ou sur un mélange de sciences exactes et sciences humaines (Paquot, 2010).

Nous nous situons dans la continuité des travaux de (Tutin, 2007c), (Drouin, 2007) et (Da Sylva, 2010), mais en nous appuyant sur un corpus analysé morpho-syntaxiquement uniquement composé d'articles de recherche et ce, seulement en sciences humaines et sociales.

Plusieurs types de statistiques sont utilisés dans les travaux cités. Nous reprenons en partie la méthodologie de (Drouin, 2007) qui combine au critère de fréquence (fréquence relative dans le corpus d'analyse en comparaison avec un corpus de référence) le critère de répartition par tranche. Ce critère permet de s'assurer de la répartition d'un mot dans l'ensemble du corpus et évite ainsi d'extraire des mots certes fréquents mais limités à une sous-partie du corpus. Contrairement à ces travaux ciblés sur les sciences dures, notre corpus d'analyse se restreint aux sciences humaines et sociales. De plus, nous avons pour but la constitution d'un corpus de référence intégrant des textes littéraires, journalistiques ainsi que des transcriptions de l'oral afin de disposer d'un corpus de référence le plus large possible.

En plus de ces critères statistiques de fréquence et de répartition, nous ajoutons un filtrage des segments répétés (afin de ne pas traiter isolément les mots les constituant), et nous nous basons sur un corpus analysé morpho-syntaxiquement. Nous utilisons les méthodes lexicométriques basées sur les spécificités du lexique examiné par comparaison de fréquences, ce qui mène à l'identification de particularités lexicales, subdivisées en spécificités positives (sur-représentation), négatives (sous-représentation) et banales (scores comparables) à partir desquelles nous pouvons décomposer, contrastivement, les différents lexiques constituant notre corpus.

Ces travaux préliminaires devront être poursuivis en se basant sur un corpus de référence du français à large échelle. Le traitement des éléments polylexicaux du LST sera intégré et une typologie des éléments de notre ressource lexicale devra être effectuée afin de la structurer et de permettre des applications didactiques d'aide à la rédaction d'écrits scientifiques.

3 Méthodologie

3.1 Corpus

Pour garantir une homogénéité maximum, le corpus d'analyse est composé d'articles de revues préalablement sélectionnées, dont la qualité est vérifiée par la notation ERIH et/ou AERES et dont le(s) auteur(s) sont francophones natifs. Nous utilisons une sous-partie (ultérieurement augmentée) du corpus du projet Scientext[2].

Notre corpus d'analyse comporte plus de 3,5 millions de mots et sera étendu pour atteindre les 5 millions. Les textes ont été formatés au format TEI Lite (Burnard, 1995) et analysé avec le logiciel Syntex (Bourigault, 2000). Nous utilisons, comme lexique de comparaison de fréquences, la base de données lexicales *lexique3* (New, 2006) qui intègre des informations de fréquence.

Notre corpus d'analyse est précisément composé de 339 articles et de 3 511 716 mots provenant de dix disciplines des sciences humaines et sociales : anthropologie, économie, géographie, histoire, linguistique, sciences de l'éducation, sciences politiques, sciences de l'information, sociologie, psychologie.

3.2 Extraction automatique

Nous avons précédemment défini le LST comme un lexique fréquent, traversant (donc réparti dans les diverses disciplines) et spécifique aux écrits scientifiques (donc plus fréquent dans ces écrits que dans un corpus de référence). Ces hypothèses sur les propriétés linguistiques du LST peuvent se traduire sous forme de critères statistiques que nous appliquons à notre corpus afin d'extraire les éléments qui nous intéressent. Nous combinons les critères suivants :

1. Répartition : le corpus est découpé en 100 tranches de tailles égales. Les mots extraits doivent apparaître dans un minimum de 50 tranches, et un minimum de 5 disciplines sur les 10 composant le corpus d'analyse.
2. Fréquence et Spécifité : les éléments doivent être sur-représenté par rapport au corpus de référence (spécificité positive + seuil du nombre d'occurrences minimal dans l'ensemble du corpus fixé à 100)
3. non-présence systématique dans un segment répété : nous ôtons à la fréquence d'un mot simple le nombre d'occurrences des segments répétés dans lesquels il intervient pour ne pas intégrer des composantes d'unités lexicales qui n'ont pas d'appartenance autonome au LST (par exemple, *point* apparaît une fois sur deux au sein du polylexical *point de vue* et a donc sa fréquence divisée par 2).
4. non-appartenance à une stop-liste ad hoc permettant d'amoindrir le bruit généré par exemple par les segments en langue étrangère (par exemple l'article anglais *the* lemmatisé en nom français *thé*)

Nous calculons le critère de spécificité selon trois formules (ratio de fréquence, chi-carré, rapport de vraisemblance) afin d'identifier, à l'instar de (Paquot, 2009), la plus adaptée,

[2] http://scientext.msh-alpes.fr

en confrontant les différentes listes de mots extraits. Plusieurs calculs peuvent être envisageables, certains offrent de meilleurs résultats sur les événements rares lorsque d'autres fonctionnent mieux sur les événements fréquents : (Labbé, 2001) pointe par exemple la faible efficacité du calcul de spécificité sur les fréquences basses.

Pour les trois calculs statistiques, nous reprenons les formules décrites par Drouin sur le site de son logiciel *TermoStat*[3].

Le fait de travailler sur un corpus analysé morpho-syntaxiquement nous permet d'une part d'effectuer un regroupement flexionnel et ainsi d'amoindrir la dispersion de fréquence, et d'autre part d'utiliser les relations de dépendances récurrentes afin d'ajouter aux lemmes extraits des informations d'ordre lexico-syntaxique. Ces relations, utilisées à ce jour seulement pour contextualiser les mots lors de l'évaluation, devront être intégrées à terme dans le processus d'extraction pour la désambiguïsation.

La fréquence est utilisée de manière absolue, par le biais de seuils, pour valider la présence minimale d'un candidat-LST à l'intérieur d'une discipline et d'une tranche de corpus dans le corpus d'analyse, et de manière relative lors de la comparaison à la base de données qui fait office de corpus contrastif.

Comme il n'existe pas à ce jour de lexique de référence pour vérifier la validité des mots extraits comme éléments du LST, cette appartenance est jugée dans le cadre d'une évaluation effectuée par trois juges experts (chercheurs en linguistique travaillant sur les écrits scientifiques), dont la tâche est présentée dans la section suivante.

3.3 Évaluation

Nous avons créé deux listes, une de 100 verbes et une de 100 noms, compilant les résultats d'extraction des trois différents calculs, en prenant soin de représenter les tranches hautes, moyennes et basses pour chacun d'eux.

Les juges avaient pour consigne de classer ces 200 candidats-LST monolexicaux dans 3 lexiques :

- LST : lexique scientifique transdisciplinaire et lexique abstrait général
- LT : lexique terminologique
- LG : lexique de la langue générale

Ils devaient classer chaque mot dans un lexique au minimum et dans tous au maximum : ceci pour tenir compte de la difficulté à circonscrire ces lexiques comme le note (Tutin, 2007b).

[3]Université de Montréal. Patrick Drouin
http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html
[consulté le 22/03/2013]

Candidat	LST	LT	LG	Rel()	Contexte
illustrer_V	+	-	-	exemple_N_SUJ cas_N_SUJ	Cet exemple paradigmatique illustre le choix que nous faisons [...]
synthèse_N	+	-	-	proposer_V_OBJ réaliser_V_OBJ	[...] la seconde propose une synthèse des principales interventions [...]

TABLEAU 1 – Exemple de grille d'évaluation

Dans cet extrait de grille d'évaluation, la première colonne présente le candidat-LST sous sa forme lemmatisée suivie de son étiquette de catégorie syntaxique (*N* pour nom et *V* pour verbe).

Les deux dernières colonnes, *rel()* et *contexte,* ont pour fonction d'aider à la décision en apportant une recontextualisation des candidats-LST.

La colonne *rel()* indique les deux associations lexico-syntactiques les plus fréquentes, c'est à dire les couples mot-relation syntaxique les plus fréquemment reliés au candidat-LST (après filtrage des relations peu informatives sur le contexte, telles celles impliquant un verbe auxiliaire ou un pronom). Y sont détaillés le cooccurrent syntaxique, par son lemme et sa catégorie, suivi du type de la relation *(objet* ou *sujet* par exemple).

La dernière colonne *contexte* permet la visualisation d'une phrase contenant l'association lexico-syntaxique la plus fréquente impliquant le candidat-LST.

4 Résultats

4.1 Analyse des évaluations

Pour les verbes, les 3 juges sont en accord sur 68 des 100 candidats (55 sont validés, 13 sont invalidés comme élément du LST). De plus, 27 verbes sont validés par 2 des 3 juges.

Pour les noms, les 3 juges sont en accord sur 79 des 100 candidats-LST (55 sont validés, 24 sont invalidés). 21 noms sont validés par 2 des 3 juges. Dans le tableau ci-dessous, un + représente la validation d'un juge sur l'appartenance au LST, un - représente la non-validation en tant qu'élément du LST.

Appartenance au LST	+ + +	+ + -	- - -	- - +	Accord à 3
Verbes	55	27	13	5	68
Noms	55	21	24	0	79

TABLEAU 2 – Résultats d'évaluation

Ces faibles pourcentages d'accord s'expliquent par deux principaux facteurs :

- L'objet d'étude, le LST, est un lexique à frontière floue, inscrit dans un continuum de lexiques. Il existe ainsi une grande variabilité quant à sa perception selon le juge.
- Les mots s'étudient en contexte. L'apport des associations lexico-syntaxiques récurrentes et la mise en contexte phrastique ne permettent pas une désambiguïsation systématique des différentes acceptions des candidats-LST.

Cette première observation sur l'évaluation met en lumière la difficulté de circonscrire notre objet d'étude, partie d'un continuum qu'il nous faut discrétiser en vue de l'extraction. L'évaluation doit être complétée par une tâche d''annotation en contexte, ce qui permettrait une évaluation précise du silence occasionné par notre méthode d'extraction.

De plus, les cas problématiques d'évaluation sont le plus souvent liés à des mots au sens vague entrant dans des collocations tel *formuler* dans *formuler une hypothèse*. L'ajout futur d'une phase de traitement des expressions polylexicales permettra d'éviter cet écueil.

La difficulté majeure se situe au niveau de la frontière entre LST et lexique de la langue générale : dans tous les cas où les juges ont validé l'appartenance d'un mot à plusieurs lexiques, les lexiques concernés étaient le LST et le lexique de la langue générale.

4.2 Analyse par méthodes statistiques

Nous comparons ci-dessous les 3 formules statistiques utilisées étudiant plus particulièrement les cas où les trois juges sont en accord sur l'appartenance ou non d'un élément au LST.

Nous observons les rangs d'extraction, parmi les trois calculs utilisés, des différents mots de nos listes (validés ou non), le mot de rang 1 étant considéré comme l'élément le plus caractéristique du LST selon le critère de spécificité positive. Les trois calculs que nous utilisons ne donnent pas d'indice certain sur l'appartenance ou non au LST pour un candidat-LST extrait d'après les critères détaillés dans la partie précédente (fréquence haute et répartition large). Par exemple *revenu* est un candidat-LST invalidé par les 3 juges, et classé 82ème selon le ratio, 87ème selon le rapport de vraisemblance et 79ème selon le chi-carré. Nous ne pouvons cependant pas conclure sur l'appartenance au LST des candidats-LST ayant un rang inférieur ou supérieur à celui de *revenu* : *coût*, candidat-LST invalidé, a un rang entre 7 et 13 selon les calculs, alors que *argument*, candidat-LST validé, a un rang entre 214 et 235.

Des mots à spécificité positive, traversant les disciplines, peuvent donc être éléments du LST (*argument, méthode, expliquer*) ou du lexique de la langue générale ou du lexique terminologique *(coût, mobilisation, multiplier)*. Même si 158 des 200 mots sont validés par au moins 2 juges sur 3, le critère de spécificité ne suffit pas à confirmer ou infirmer l'hypothèse d'appartenance au LST. La prise en compte de la fréquence des segments répétés permet d'éliminer certains composants de ces segments mais ne peut gérer les cas d'expressions polylexicales acceptant des modifieurs et donc sujettes à des variations.

En examinant la répartition des occurrences par discipline, nous pouvons dégager un problème récurrent lié aux acceptions variées que peut recouvrir un mot. Par exemple, le nom *coût* a une fréquence totale (dans les 10 disciplines combinées de notre corpus d'analyse) de 976 occurrences, dont 688 dans le seul sous-corpus d'économie, et entre 10 et 80 dans les 9 autres disciplines. En calculant l'écart-type, ce phénomène de sur-présence dans une discipline serait identifié et nous pourrions alors différencier la fréquence du mot dans son acception générale de sa fréquence dans son acception disciplinaire. D'autres exemples plus complexes peuvent être rencontrés, par exemple le cas de *sujet :* il peut avoir plusieurs sens disciplinaires (*Le sujet de la phrase* en linguistique, *Le sujet de l'expérience* en psychologie) ou un sens transdisciplinaire (*Le sujet de l'article*).

Nous pouvons donc observer qu'aucun calcul n'est suffisant pour valider ou invalider de façon certaine un candidat-LST. La fixation d'un seuil ou d'un rang engendre un silence et un bruit trop importants pour ne tenir compte que de ce critère. Nous privilégierons donc la méthode la plus simple, à savoir le ratio de fréquence, et ajouterons le calcul de l'écart-type afin de diminuer le bruit occasionné par les mots à multiples acceptions dont l'une est sur-représentée dans une discipline.

Malgré les apports certains des techniques lexicométriques, celles-ci ne sont pas suffisantes pour extraire automatiquement les mots simples du LST. Une prise en compte de la distribution est nécessaire, que ce soit en vue de la gestion des éléments polylexicaux ou pour la discrimination des diverses acceptions que peut recouvrir un candidat-LST.

Enfin, l'absence d'un lexique de référence du LST, à grande échelle, nous impose une évaluation humaine et ne permet pas de confronter les résultats d'extractions dans leur totalité (nous avons sélectionné 200 mots parmi plus de 1000 composant nos différentes listes) pour quantifier précisément le bruit et le silence générés.

5 Conclusion et perspectives

Nous avons pu, en combinant méthodes statistiques et informations morpho-syntaxiques, procéder à une première extraction des mots simples du lexique scientifique transdisciplinaire dans les articles de recherche en sciences humaines et sociales.

Ces premiers résultats, intéressants du strict point de vue des mots simples, soulignent l'importance de la prise en compte de leur contexte d'apparition, afin de gérer le cas des expressions polylexicales (pour filtrer les mots simples non autonomes), et pour identifier les cas de polysémie et ainsi discriminer les différentes acceptions, transdisciplinaires ou non. La prise en compte de l'écart-type au niveau de la répartition des fréquences intra-disciplinaires est une piste d'identification de ces phénomènes, la distribution en étant une autre, par exemple à l'aide des cooccurrents de deuxième ordre (Bertels 2012).

Le traitement automatique des expressions polylexicales devra être la prochaine étape, étant donné leur part importante dans le LST (Pecman, 2007).

Les travaux de (Kister, 2012) sur le lien syntaxique récurrent entre lexème scientifique transdisciplinaire et terme sont également une piste pour identifier l'acception transdisciplinaire d'un élément ambigu.

L'étape d'évaluation (que nous effectuerons avec plus de juges) devra être remaniée sur deux principaux points :

- Une plus grande recontextualisation des candidats-LST via un plus grand nombre d'exemples phrastiques et d'associations lexico-syntaxiques récurrentes.
- Un travail d'annotation manuelle sur corpus pour évaluer précisément le silence.

Dans l'optique d'affiner les calculs de spécificité, la constitution d'un corpus contrastif « hybride » de grande échelle, intégrant une partie orale, journalistique, et littéraire, sera nécessaire. Le travail ici présenté utilise pour l'étude contrastive une base de donnée ne permettant pas les comparaisons distributionnelles ou la recherche de segments répétés.

Notre corpus d'analyse suivant le format TEI Lite, les informations de structure textuelle pourront être croisées avec les fréquences pour caractériser les parties textuelles d'un article : quel est le lexique le plus présent dans les résumés, conclusions, en regard par exemple des observations sur la prédominance de la terminologie dans les introductions (Rinck, 2010).

Au niveau théorique, l'extraction aura pour but une analyse fine de la phraséologie (au sens large des combinaisons récurrentes et stabilisées) des écrits en sciences humaines et sociales menant à une description de ce type d'écrit. Par ailleurs, compte tenu de la difficile acquisition du métadiscours propre à la production scientifique, une application didactique d'aide à la rédaction pourrait profiter d'une ressource lexicale structurée du LST.

Remerciement

Nous remercions la région Rhône-Alpes qui finance nos travaux de recherche, le projet ANR-Contint Termith, Olivier Kraif pour son aide précieuse sur les aspects lexicométriques ainsi que les relecteurs pour leurs nombreux conseils.

6 Références

Bertels, A. et Geeraerts, D. (2012). L'importance du recoupement des cooccurrents de deuxième ordre pour étudier la corrélation entre la spécificité et la monosémie. *Actes de JADT 2012*, 135-147.

Blumenthal, P. (2007). Sciences de l'Homme vs sciences exactes: combinatoire des mots dans la vulgarisation scientifique. *Revue française de linguistique appliquée*, *12*(2), 15-28.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M. P., et Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*.

Burnard, L., et Sperberg-McQueen, C. M. (1995). TEI lite: An introduction to text encoding

for interchange (*pp. 23-152*). SURFnet.

Coxhead, A. (2002), The academic word list : a corpus-based Word List for Academic Purposes *in Teaching and Language Corpora (TALC) 2000 Conference Proceedings. Atlanta : Rodopi.*

Da Sylva, L. (2010), Extraction semi-automatique d'un vocabulaire savant de base pour l'indexation automatique , *Actes de TALN 2010, 2010.*

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. Revue française de linguistique appliquée, 12(2), 45-64.

Evert, S. (2008). Corpora and collocations. Corpus Linguistics. An International Handbook, 2.

Kister, L., et Jacquey, E. (2012). Relation syntaxiques entre lexique terminologique et transdisciplinaire: analyse en texte intégral. *CMLF 2012*.

Labbé C. et Labbé D. (2001). Que mesure la spécificité du vocabulaire?, *Lexicometria, no 3, 23 p.*

New, B., Pallier C., Ferrand L. et Matos R. (2001) Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique, 101, 447-462. http://www.lexique.org* [consulté le 22/03/2013]

Paquot, M. et Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers*, *68*(1), 247-269.

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. *Continuum*.

Pecman M. (2004), Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique. *Thèse de doctorat. 9 déc. 2004. Dir. Henri Zinglé. Université de Nice-Sophia Antipolis. 467 p.*

Pecman, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée, 12(2), 79-96.*

Phal, A. (1971), Vocabulaire général d'orientation scientifique (V.G.O.S) – Part du lexique commun dans l'expression scientifique ». *Paris : Didier, Crédif*

Rinck, F. (2010), L'analyse linguistique des enjeux de connaissance dans le discours scientifique, *Revue d'anthropologie des connaissances* 3/2010 (Vol 4, n° 3), p. 427-450.

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings: *Cambridge Applied Linguistics. Cambridge University Press.*

Tutin, A. (2007a). Modélisation linguistique et annotation des collocations: une application au lexique transdisciplinaire des écrits scientifiques. *Formaliser les langues avec l'ordinateur: actes des sixièmes, Sofia 2003, et septièmes, Tours 2004, journées Intex-Nooj*, *3*, 189.

Tutin, A. (2007b). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée, 12(2),* pages 5-14.

Tutin, A. (2007c). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. *Actes de Traitement Automatique des Langues Naturelles (TALN),* pages 283-292.

Tutin, A., Grossmann, F., Falaise, A. et Kraif, O. (2009). Autour du projet Scientext: étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Actes des 6es journées de linguistique de corpus*.