Agnès Tutin

Sémantique lexicale et corpus : l'étude du lexique transdisciplinaire des écrits scientifiques

Lublin Studies in Modern Languages and Literature 32, 242-260

2008

LUBLIN STUDIES IN MODERN LANGUAGES AND
LITERATURE 32, 2008, HTTP://WWW.LSMLL.UMCS.LUBLIN.PL

Agnès Tutin
University of Grenoble,
Grenoble, France

Sémantique lexicale et corpus : l'étude du lexique transdisciplinaire des écrits scientifiques[1]

1. Introduction

La linguistique de corpus est une discipline qui a le vent en poupe, mais la façon dont les corpus sont exploités n'est pas toujours rapportée de façon explicite dans les écrits de linguistique. Nous souhaitons relater ici une expérimentation – encore en cours – basée sur corpus autour du lexique transdisciplinaire des écrits scientifiques, ce lexique en partie invariant qui renvoie aux procédures décrites par les chercheurs, au raisonnement et au métatexte. Après avoir défini notre objet, nous montrerons comment l'on peut exploiter un corpus pour circonscrire plus finement l'objet qui nous intéresse et en extraire un lexique de base de mots simples. Nous montrerons ensuite comment nous avons commencé à caractériser ce lexique sur le plan sémantique et comment les techniques issues du traitement automatique du langage peuvent faciliter la tâche du linguiste.

[1] Cette étude a été effectuée dans le cadre du projet ANR Scientext « Etude des marques de positionnement et de raisonnement dans les écrits scientifiques » piloté par le LIDILEM.

2. Notre objet : le lexique transdisciplinaire des écrits scientifiques

Le lexique qui nous intéresse est un lexique de genre plutôt qu'une terminologie : il comprend les mots qui sont spécifiques aux écrits scientifiques et communs, jusqu'à un certain point, à un large éventail de disciplines. Il intègre le lexique métascientifique, qui parle de la science (objets et procédures scientifiques) : par exemple, *collecter des données, analyser des résultats*, mais il renvoie aussi au métadiscours scientifique, au sens de Hyland (2005), c'est-à-dire aux marques linguistiques qui ne renvoient pas au sens propositionnel, mais aux interactions au sens large entre un auteur et son destinataire dans une même communauté, et renvoient à des relations internes au discours : par exemple, *méthode prometteuse, chapitre suivant, voir supra* (Tutin 2007c). Ce lexique a fait l'objet de plusieurs études, tout particulièrement en langue anglaise (Coxhead 2000 ; Cohead & Hirsh 2007 ; Pecman 2004 par exemple), mais pour le français, les travaux de référence restent ceux de Phal (1972) qui méritent d'être actualisés.

Notre objectif à plus long terme est l'étude des spécificités du discours scientifiques, en particulier des marques linguistiques qui indiquent le positionnement et le raisonnement de l'auteur, dans le cadre du projet ANR Scientext. Nous souhaitons également proposer un lexique de base qui permettrait d'élaborer des activités didactiques utilisables en Français Langue Etrangère et Seconde.

Définir clairement la notion de lexique transdisciplinaire des écrits scientifiques n'est cependant pas chose aisée. Ce vocabulaire n'est en effet pour une large part pas entièrement spécifique des écrits scientifiques, mais est simplement surreprésenté dans ce genre. Peu de termes, contrairement à la terminologie, y prendront une acception complètement spécifique. Certains lexèmes, en outre, seront seulement propres à un sous-ensemble de disciplines : par exemple, le lexique des évaluations quantitatives (résultats, tests statistiques) sera particulièrement représenté dans bon nombre de sciences expérimentales, mais probablement assez rare dans certaines sciences humaines comme la linguistique.

Nous proposons de schématiser le lexique à l'œuvre dans les écrits scientifiques comme suit.

1. **Le lexique transversal propre aux écrits scientifiques** renvoie aux procédures ou aux concepts génériques de l'activité scientifique, probablement présents dans nombre d'écrits du même type et dans des disciplines variées : *évaluation, théorique, réaliser des comparaisons, observation* (*directe*), *estimation, utiliser une procédure,*(*obtenir*) *résultats.*
2. **Le lexique abstrait non spécialisé**. Ce lexique n'est pas exclusif des écrits scientifiques et apparaît également dans d'autres types d'écrits argumentatifs ou informatifs : *poser un problème, hétérogénéité, la difficulté joue, influence, élément, dimension, conduire à …*
3. **Le lexique méthodologique disciplinaire**. Certains éléments lexicaux « métascientifiques » peuvent être considérés comme disciplinaires ou relevant d'une famille de disciplines : c'est ainsi le cas d'expressions comme *comparaison longitudinale, panel,* qui renvoient aux procédures scientifiques d'un ensemble de sciences sociales et humaines comme l'économie, la démographie ou la psychologie mais aussi à d'autres sciences appliquées comme la médecine.
4. **Le lexique terminologique** renvoie aux objets examinés dans la discipline et aux procédures spécifiques : *lexème, analyse syntaxique, collocation, phraséologie* en linguistique.
5. **Le lexique de la langue « générale » ou « commune »** intègre les mots qui ont essentiellement une fonction grammaticale (*l', de, ou, entre, être, …*) ou, peu spécialisés, ont une probabilité d'occurrence qui ne paraît ni liée à la discipline, ni au genre d'écrits (*enfant, préférence, arrivée, départ, …*).

Nous caractériserons comme **transdisciplinaire** le **lexique transversal propre aux écrits scientifiques**, ainsi que **le lexique abstrait non spécialisé** qui est particulièrement représenté dans les écrits scientifiques.

3. Circonscrire le lexique transdisciplinaire des écrits scientifiques à l'aide de corpus

Constituer un inventaire du lexique transdisciplinaire des écrits scientifiques nous paraît utile à plusieurs titres. Outre l'intérêt évident pour les activités didactiques en langue étrangère (qui a motivé des travaux comme ceux de Phal 1971, Pecman 2004 ou Coxhead 2000 et Coxhead & Hirsh 2007), inventorier ce lexique de base nous permet d'établir des comparaisons dans le fonctionnement des discours scientifiques, et en comprendre les spécificités. Comme dans tout inventaire, les limites en seront évidemment un peu incertaines, mais tous les linguistes s'accorderont à reconnaître dans les termes *hypothèse*, *notion* ou *valide* des lexèmes centraux des écrits scientifiques.

La plupart des inventaires lexicaux réalisés à ce jour recourent à des corpus diversifiés et à des mesures statistiques. De notre point de vue, le corpus idéal utilisé pour cette tâche devrait comporter un très grand nombre de mots, de l'ordre d'au moins 10 millions de mots, relevant de sous-genres variés – articles de recherche, mais aussi actes de colloques, thèses et mémoires, et de disciplines diverses des sciences humaines, sociales, expérimentales et appliquées également représentées. A ce jour, à notre connaissance, aucun inventaire lexical n'utilise ce type de grand corpus équilibré (Cf. tableau 1 ci-dessous). Pour l'anglais, Coxhead (2000) s'approche de cet objectif en exploitant un corpus volumineux et largement diversifié d'écrits universitaires. Pour notre part, nous utilisons, à ce stade de notre recherche[2], un corpus de 2 millions qui complète le corpus KIAP élaboré par l'équipe de Kjersti Fløttum[3], intégrant trois disciplines assez différentes, la médecine, l'économie et la linguistique. Le corpus contient des articles de recherche (corpus KIAP), mais aussi des thèses, des rapports et quelques cours.

---

[2] Dans le cadre du projet ANR Scientext, nous souhaitons baser nos travaux de lexicologie sur un corpus plus large et plus diversifié.

[3] http://www.uib.no/kiap/. Pour une synthèse des études réalisées sur ce corpus, voir Fløttum *et al.* (2006).

Pour déterminer le lexique propre à ces écrits de recherche, des techniques lexicométriques, parfois complexes, sont souvent employées. Elles mettent en jeu un ensemble de paramètres comme :

- La **fréquence**. Par exemple, nous sélectionnons les éléments du lexique apparaissant plus de 15 fois (dans chaque discipline).
- La **répartition**. On pourra ainsi sélectionner le lexique qui apparaît dans la moitié ou les trois-quarts des disciplines, ou qui apparaissent dans la moitié des tranches de textes.
- La **spécificité**. La méthode des spécificités permet de dégager un ensemble lexical propre à un type de texte. Par exemple, Drouin (2007) recourt à la méthode des spécificités établie par Lafon (1980) qui extrait le lexique spécifique par comparaison avec un corpus de référence.

Ces techniques permettent d'extraire le lexique qui est à la fois fréquent, bien réparti dans les différentes disciplines, et spécifique du genre examiné.

	Phal 1971 (Vocabulaire générale d'orientation scientifique)	**Coxhead 2000 (Academic Word List)**	**Drouin 2007 (Lexique Scientifique Transdisciplinaire)**	**Tutin, cet article**
Type de corpus exploité	Manuels de 4ème, 3ème, 2de, 1ère et terminales.	Corpus d'articles scientifiques, chapitres de livres, manuels universitaires, manuels de laboratoire, notes de cours.	Corpus de thèses	Corpus d'articles scientifiques, de thèses, de rapports de recherche, de cours. (Inclusion du corpus KIAP du français)
Disciplines traitées	Physique, mathématiques, chimie, sciences naturelles	28 disciplines dans les sections des facultés de *arts* (lettres et sciences humaines), commerce, droit et sciences		Médecine, linguistique, économie.

		« dures ».		
Taille du corpus (en mots)	1,8 million	3,5 millions	2,3 millions	2 millions
Technique employée pour circonscrire le lexique	Critères complexes faisant intervenir la fréquence mais aussi la répartition dans les différents ouvrages ainsi que la dispersion dans les disciplines. D'autres critères plus qualitatifs sont également employés (par exemple, inclusion des antonymes).	Mots qui apparaissent plus de 100 fois, dans au moins la moitié des 28 disciplines, et apparaissent au moins 10 fois dans chaque famille de 4 disciplines. Exclusion des mots courants.	Mots se répartissant dans 50% des tranches de textes et ayant une spécificité > 3,09 par rapport au corpus de référence du *Monde*. (Corpus prétraité avec étiqueteur morpho-syntaxique).	Mots apparaissant plus de 15 fois dans les trois disciplines. (corpus prétraité avec étiqueteur morpho-syntaxique). Filtrage manuel.

Tableau 1 : Listes lexicales du lexique des écrits scientifiques.

Dans notre expérimentation, nous avons utilisé des techniques statistiques simples, puisque nous avons extrait, après étiquetage[4] morpho-syntaxique des corpus avec Cordial, les mots à la fois fréquents (apparaissant plus de 15 fois) et transversaux (communs aux trois disciplines : linguistique, économie et médecine). Un filtrage manuel a été opéré pour les noms et les adjectifs lorsque des erreurs manifestes d'étiquetage avaient eu lieu ou lorsque les mots n'apparaissent que dans certaines expressions complètement figées. Nous avons également ôté les lexèmes polysémiques qui ne partageaient dans les trois disciplines de toute évidence que la forme et non une acception commune. Pour nous, les techniques

[4] L'étiquetage morpho-syntaxique permet d'associer aux mots en contexte leur partie du discours. Par exemple, le mot *fait* peut être un verbe ou un nom selon le contexte. Dans *ce fait doit être signalé*, c'est un nom. L'étiquetage morpho-syntaxique des corpus permet d'établir des statistiques plus fiables sur les mots du corpus.

lexicométriques ne peuvent être qu'une base dans la sélection des lexèmes, et un traitement manuel, qui examine le sens et l'emploi des lexèmes en corpus, reste indispensable.

Pour illustrer notre démarche, nous présentons ci-dessous (dans les tableaux 2, 3 et 4) les listes des adjectifs, noms et verbes transdisciplinaires les plus fréquents extraits du corpus[5]. A l'aide de la procédure décrite plus haut, nous obtenons 203 adjectifs, 363 noms et 300 verbes, soit un lexique de 866 éléments. Les adverbes n'ont pas été intégrés, mais devraient également l'être à terme[6].

Tableau 2 : Adjectifs transdisciplinaires des écrits scientifiques les plus fréquents.

Adjectifs	Economie (fréquence)	Linguistique (fréquence)	Médecine (fréquence)	Fréquence totale
1. Différent	635	536	506	1677
2. Important	500	197	672	1369
3. Grand	385	393	402	1180
4. Spécifique	248	369	323	940
5. Possible	306	406	226	938
6. Général	284	268	379	931
7. Certain	289	405	150	844
8. Faible	400	73	361	834
9. Elevé	418	32	382	832
10. Relatif	351	215	188	754
11. Supérieur	329	91	301	721
12. nécessaire	274	136	291	701

Tableau 3 : Noms transdisciplinaires des écrits scientifiques les plus fréquents.

[5] Les listes complètes sont disponibles sur : http://w3.u-grenoble3.fr/tutin/lexique/lexique.html.

[6] Drouin (2007) intègre les adverbes dans ses listes du lexique scientifique transdisciplinaire.

Noms	Economie (fréquence)	Linguistique (fréquence)	Médecine (fréquence)	Fréquence totale
1. effet	1621	576	1134	3331
2. cas	849	1036	1375	3260
3. étude	446	343	1727	2516
4. valeur	766	1052	621	2439
5. modèle	1413	174	805	2392
6. type	608	1083	648	2339
7. exemple	438	1603	207	2248
8. résultat	1058	241	924	2223
9. terme	977	966	202	2145
10. taux	1515	92	500	2107
11. forme	365	1004	425	1794
12. analyse	635	724	386	1745

Les premières listes extraites présentent des contrastes intéressants. La liste des adjectifs fréquents intègre des mots peu spécialisés, où la dimension quantitative et la comparaison sont cependant assez présentes (*important, grand, élevé, faible ; différent, supérieur*). Ces lexèmes sont bien entendu fortement polysémiques (comparons par exemple *concept important* et *nombre important*) et comme pour les prédicats verbaux, les adjectifs doivent surtout selon nous être considérés en association avec les arguments sur lesquels ils portent, ce qui motive notre intérêt pour les collocations dans ce lexique. En outre, on observe des différences remarquables entre disciplines. Par exemple, les adjectifs *faible, élevé, supérieur* sont très nettement sous-représentés en linguistique, par rapport à l'économie et à la médecine, ce qui semble indiquer la faible importance du paramètre quantitatif dans cette discipline (ce que la faible fréquence du terme *taux* dans la liste des noms semble confirmer). Contrairement aux adjectifs, les noms apparaissent beaucoup plus riches sémantiquement, nombre d'entre eux relevant du champ lexical de l'étude (*étude, analyse, modèle*) et de l'évaluation quantitative (*résultats, valeur, résultats, taux*). Des différences disciplinaires importantes se font également

jour pour cette catégorie : le terme *étude* par exemple est très souvent employé en médecine alors que les économistes se montrent particulièrement friands du concept de *modèle.* Enfin, en ce qui concerne les verbes fréquents, sans surprise ce sont les auxiliaires, les verbes supports et les modaux qui dominent (*être, avoir, pouvoir, faire, devoir, mettre*). Les verbes « pleins », assez polysémiques, relèvent de divers champs comme l'observation (*voir*) ou la démonstration (*montrer*). Comme les adjectifs, les verbes doivent être considérés en relation avec les arguments nominaux.

Tableau 4 : Verbes transdisciplinaires les plus fréquents.

Verbes	Economie (fréquence)	Linguistique (fréquence)	Médecine (fréquence)	Fréquence totale
1. être	14709	15518	15971	46198
2. avoir	4494	4703	7166	16363
3. pouvoir	2557	3573	2183	8313
4. permettre	1100	864	1114	3078
5. faire	783	1400	690	2873
6. devoir	732	638	737	2107
7. mettre	608	680	689	1977
8. présenter	392	503	786	1681
9. montrer	637	379	606	1622
10. considérer	656	618	275	1549
11. utiliser	541	414	579	1534
12. voir	519	734	125	1378

Nous n'avons ici commenté que les occurrences les plus fréquentes du lexique transdisciplinaire dégagé, qui apparaît bien entendu bien plus spécifique dans les fréquences moyennes.

4. Le traitement sémantique du lexique transdisciplinaire des écrits scientifiques

La liste de lexèmes dégagée n'est véritablement utile que si elle a été caractérisée au plan sémantique. Notre objectif, dans le projet Scientext, est d'étudier à travers les marques lexicales et syntaxiques le positionnement et le raisonnement de l'auteur dans les écrits scientifiques. Nous souhaitons dans ce cadre extraire des classes de mots permettant de constituer des grammaires locales entrant dans des patrons courants des marques de positionnement et de raisonnement. Ces grammaires seront ensuite intégrées dans une interface permettant d'interroger les textes de façon ciblée. Par exemple, la filiation scientifique dans les écrits scientifiques s'exprime souvent à l'aide d'expressions stéréotypées comme *nous (reprendrons/utiliserons /recourrons à) (la notion/le modèle/le concept) ... développé par ...* (Garcia 2008). Le recours aux classes de mots pourrait permettre de généraliser en quelque sorte ces expressions :

Nous UTILISER Det ARTEFACT_SCIENT ...

où la classe UTILISER intégrerait les verbes *recourir, utiliser, reprendre* et la classe ARTEFACT_SCIENT des noms comme *modèle, théorie, concept, idée ...* En outre, un tel traitement permet également de modéliser les associations lexicales ou collocations de façon utile.

Dans le cadre de notre travail, nous avons principalement souhaité proposer des classes sémantiques simples, fondées sur des propriétés linguistiques aisément reproductibles. Nous avons ainsi privilégié, en particulier pour le traitement des noms, des classes distributionnelles plutôt que des classes notionnelles comme celles qui ont été proposées par Pecman (2004). Ce travail n'est pas encore achevé pour l'ensemble du lexique transdisciplinaire (qui doit encore être affiné sur un corpus plus conséquent en cours de développement).

Un premier ensemble de **classes de noms** a été dégagé, un peu à la façon de Flaux et van de Velde (2000), à l'aide la combinatoire lexicale et syntaxique observée en corpus (voir Tutin 2007b pour une présentation plus détaillée), démarche que nous avons également mise en œuvre dans d'autres travaux sur le lexique des émotions (Tutin *et al.* 2006). Sur cette base, un ensemble de 60 noms transdisciplinaires

fréquents ont été répartis dans 7 classes, dont nous donnons quelques exemples dans le tableau 5 ci-dessous.

Tableau 5 : Quelques classes de noms du lexique transdisciplinaire des écrits scientifiques.

Classe de nom	Exemples	propriétés linguistiques
Objets construits par l'activité scientifique (artefacts scientifiques)	*analyse, application, approche, démarche, définition, idée, hypothèse, méthode, modèle, solution, système, technique, technologie, théorie, test*	- ne sont pas extensifs. - ont un agent humain. (le N_obj_const de Nhum). - se combinent avec des verbes comme *élaborer, construire.*
Observables de l'activité scientifique	*cas, données, exemple, facteur, paramètre, point, question, problème, résultat*	- ne sont pas extensifs. - se combinent avec le support *être.* - se combinent avec les verbes *analyser, examiner, étudier.*
Supports de la rédaction scientifique	*article, chapitre, document, figure, ouvrage, schéma, section, texte*	- sont à la fois concrets et abstraits non extensifs. - se combinent avec la préposition *dans.* - se combinent avec le verbe *présenter.* Ex : *ce chapitre présente.*

Par exemple, la classe des « artefacts scientifiques » présente un certain nombre de points communs : ils ne sont pas extensifs (au sens de Flaux et van de Velde (2000), ils ont un complément humain (le concepteur de l'artefact) et se combinent facilement avec des verbes comme *élaborer* ou *construire*. Ces classes peuvent être exploitées dans la modélisation des collocations, les patrons de collocations apparaissant davantage comme des associations de classes sémantiques, plutôt que des idiosyncrasies lexicales.

Des **classes de verbes** plus fines ont été proposées, un peu à la façon de Wordnet (Felbaum 1998), en prenant en compte l'association

avec les arguments nominaux (Voir tableau 6). Nous souhaitons en affiner la description en détaillant les structures argumentales. Le cadre théorique de Framenet proposé par Fillmore (Fillmore *et al.* 2003), nous paraît tout à fait adapté à cette tâche, en particulier dans la perspective de notre étude des marques du positionnement, en ce qu'il permet à la fois un traitement abstrait du lexique, mais fondé sur des propriétés observables en corpus.

Tableau 6 : Quelques exemples de classes de quasi-synonymes pour les verbes.

Etiquette	classe de quasi-synonymes
DÉCRIRE	décrire, détailler, exposer, présenter, retracer
ETUDIER	aborder, analyser, considérer, étudier, examiner, explorer, regarder
OPINION_FAVORABLE	avancer, défendre, postuler, préconiser, promouvoir, prôner, recommander, réhabiliter, soutenir

Enfin, une première classification simple a été proposée pour un sous-ensemble du **lexique adjectival évaluatif** fréquent, c'est-à-dire les adjectifs qui mettent en jeu une forme de jugement, par opposition à des adjectifs dit « objectifs » (Tutin à paraître). Dire d'une approche qu'elle est *nouvelle* ou *prometteuse* engage ainsi davantage l'auteur que la qualifier d'*exploratoire* ou de *théorique.* Suivant la typologie classique proposée par Kerbrat-Orecchioni (1980), nous avons ainsi réparti les adjectifs évaluatifs qui portent sur les noms transdisciplinaires en axiologiques et non axiologiques, en affinant cette dernière classe. Le tableau 7 présente quelques exemples de ce lexique évaluatif.

Tableau 7 : Typologie des adjectifs évaluatifs.

Axiologiques	Non axiologiques
résultats intéressants, analyse pertinente, mauvais résultats …	- temps : ***travaux récents, concept ancien***, ... - importance : ***rôle crucial, principal problème*** … - nouveauté : ***nouvelle méthode, problème classique*, caractère *novateur*** ... - degré et quantité : ***grande quantité, nombreux problèmes*** ... - comparaison : ***résultats comparables, méthode différente*** … - complexité : ***problème facile, analyse complexe*** … - autres : ***conclusion paradoxale***

Une première étude sur les associations entre noms transdisciplinaires et adjectifs évaluatifs a été effectuée sur notre corpus dans les domaines de la linguistique et de l'économie (Tutin, à paraître). Elle montre que l'emploi des axiologiques est sans surprise peu fréquent, l'écriture scientifique préférant des modalités d'évaluation plus subtiles, mettant en jeu des évaluatifs moins subjectifs comme les adjectifs évoquant la nouveauté ou l'importance.

5. Le recours aux outils de Traitement Automatique du Langage pour faciliter le traitement sémantique

Les premiers traitements sémantiques proposés ci-dessus restent à compléter et à affiner. Nous avons cherché à déterminer dans quelle mesure les outils de traitement automatique du langage permettaient de faciliter – en partie – certains de ces traitements sémantiques (Tutin 2007a). L'idée était d'extraire automatiquement les environnements lexicaux et syntaxiques, afin de constituer automatiquement des classes distributionnelles sémantiques homogènes. On reprend ainsi l'hypothèse triviale que les mots qui partagent les mêmes environnements seront sémantiquement proches. Nous ne cherchons pas à obtenir des classes très fines, mais plutôt des classes de co-

hyponymes du même type que celles qui ont été établies pour les noms transdisciplinaires.

Utilisant le corpus KIAP étendu déjà présenté ici, nous avons exploité les sorties syntaxique du logiciel SYNTEX développé par Didier Bourigault (2007 ; Bourigault & Lame, 2002). Cet analyseur syntaxique produit une analyse en dépendance et le système UPERY qui en est dérivé permet de calculer les fréquences des différents types de relations. Les relations syntaxiques des 50 noms transdisciplinaires les plus fréquents ont été extraites du corpus, et consignées dans un tableau. Par exemple, dans le tableau 8, on peut observer les relations syntaxiques les plus productives avec *hypothèse.* On voit ainsi que le mot *hypothèse* apparaît le plus souvent comme attribut du verbe *être* (1255 occurrences). Suivent ensuite les épithètes *autre, différent, même ...* Le second cooccurrent verbal le plus productif est *faire* (*une hypothèse*).

Tableau 8 : La combinatoire lexicale et syntaxique la plus fréquente du mot *hypothèse* dans le corpus KIAP étendu.

Relation	**Mot en relation**	**Catégorie du cooccurrent**	**Fréquence**
Attribut	être	V	1255
Epithète	autre	Adj	195
Epithète	différent	Adj	163
Epithète	même	Adj	138
Epithète	premier	Adj	132
Epithète	général	Adj	78
Epithète	nouveau	Nom	78
De	travail	Nom	59
De	capital	Adj	57
OBJ_DIR	faire	V	48

Nous avons ensuite calculé la distance sémantique entre les mots, en prenant en compte les environnements syntaxiques partagés, à l'aide de mesures classiques (ici, le jaccard). Par exemple, dans le tableau 9 ci-dessous, les mots *article* et *chapitre*, qui apparaissent à la première ligne du tableau, partagent 6 contextes identiques (= a). *article* apparaît lui-même dans 18 contextes différents (= n1), alors que *chapitre* apparaît lui-même dans 12 contextes différents (= n2). Le coefficient jaccard utilisé ici calcule la proximité sémantique entre les mots avec la formule suivante : $a/(n_1+n_2-a)$. Seuls sont sélectionnés les voisins pour lesquels le coefficient de jaccard dépasse 0,10 et qui ont au moins quatre types de contextes communs.

Tableau 9 : calcul des « voisins ».

Mot 1	**Mot 2**	**a (nombre de contextes partagés)**	**n1 (nbre de contextes du mot 1)**	**n2 (nbre de contextes syntaxiques du mot 2)**	**jaccard**
Article	chapitre	6	18	12	0.25
Article	section	6	11	19	0.25
Aableau	chapitre	21	84	21	0.25

Dans l'exemple présenté, la parenté sémantique entre les termes apparaît évidente : tous renvoient à des objets textuels, et le rapprochement entre eux a été possible du fait d'un nombre significatif de contexte partagés. Cependant, le calcul des « voisins », s'il permet de rapprocher des couples de termes, ne permet pas de regrouper les mots en classes, comme nous le souhaitions. Pour ce faire, il faut utiliser des techniques de « clustering » souvent utilisées en informatique. Pour cette expérimentation, nous avons choisi d'utiliser une classification par voisinage (neighbour joining cluster), effectuée à partir d'une matrice contenant tous les coefficients de proximité (jaccard) – sans seuil – calculés à partir de toutes les relations syntaxiques (Cf. un exemple Tableau 9). La figure 1 présente les résultats de cette classification.

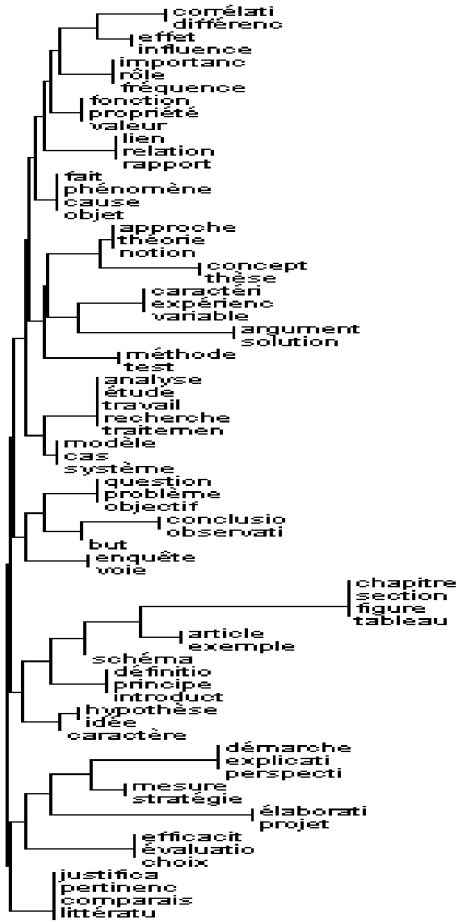

Figure 1 : Classification par voisinage à partir des coefficients de proximité (jaccard).

Nous avons comparé ces résultats avec la classification manuelle proposée sur des critères distributionnels (voir section 4). Sur les 27 classes finales obtenues avec la méthode automatique, 20 constituent des sous-ensembles des 7 classes définies manuellement (2 sous-ensembles ont des éléments uniques), ce qui apparaît un assez bon

résultat. Les sous-classes établies sont beaucoup plus fines que celles qui ont été élaborées manuellement, et beaucoup d'entre elles apparaissent pertinentes par rapport à nos objectifs. La méthode se révèle donc assez prometteuse pour étendre le traitement sémantique à l'ensemble du lexique. Deux problèmes apparaissent néanmoins. Tout d'abord, les classes proposées sont disjointes, ce qui ne permet pas le traitement de la polysémie. Par exemple, le mot *conclusion* est à la fois une partie textuelle et l'aboutissement d'un raisonnement, alors qu'il apparaît ici uniquement regroupé avec le mot *observation*, ce qui n'apparaît pas vraiment satisfaisant. Un traitement à l'aide de cliques serait plus pertinent, puisqu'il permettrait l'affection d'un élément à plusieurs classes. Le deuxième problème pour nous est la définition de l'environnement lexical. Tous les contextes n'ont en effet pas la même valeur informative pour le traitement sémantique, et nous aimerions limiter les cooccurrences lexicales aux mots du lexique transdisciplinaire, de façon à limiter les associations terminologiques qui faussent probablement les résultats. Pour une utilisation de cette méthode par des linguistes, il serait en outre nécessaire de connaitre les environnements partagés, afin de comprendre et d'évaluer la façon dont les regroupements sont effectués.

6. Pour conclure

Pour définir en extension un lexique de genre comme le lexique transdisciplinaire des écrits scientifiques et en étudier les caractéristiques sémantiques, le recours aux corpus apparaît indispensable. On peut tout d'abord appliquer des techniques lexicométriques simples pour en définir les contours, tout en filtrant au cas par cas les résultats obtenus. Les propriétés sémantiques de ce lexique peuvent ensuite être mises au jour par l'examen systématique de la combinatoire lexicale et syntaxique, et ce processus peut être facilité par l'utilisation d'outils de traitement automatique du langage (analyseurs syntaxiques, techniques de « clustering », etc.). Le linguiste devra néanmoins paramétrer finement ces outils afin de gérer adéquatement la polysémie et les expressions polylexicales.

Bibliographie

Bourigault D. (2007) : *Un analyseur syntaxique opérationnel : Syntex.* Habilitation à Diriger des Recherches. Juin 2007, Université Toulouse Le Mirail.

Bourigault, D., Lame, G. (2002) : Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit. In : *Traitement automatique du langage*, 43 (1), 129-150.

Coxhead, A. (2000) : A New Academic Word List. In : *TESOL Quarterly*, 34 (2), 213-238.

Coxhead, *A.,* Hirsh, *D.* (2007) : A pilot science word list for EAP. In : *Revue française de linguistique appliquée*, 12 (2), 65-78.

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

Fillmore, Ch.J., Johnson, C.R., Petruck, M. (2003) : Background to Framenet. In : *International Journal of Lexicography*, 16 (3), 235-250.

Fløttum,, K., Dahl, T., Kinn, T. (2006) : *Academic Voices.* Amsterdam/Philadelphia: John Benjamins.

Garcia, P. P. (2008): *Etude des marques de la filiation dans les écrits scientifiques.* Mémoire de Master 1, ss. dir. Francis Grossmann et Agnès Tutin, , Université Stendhal-Grenoble3 : Grenoble.

Hyland, K. (2005): *Metadiscourse.* London, New York: Continuum.

Kerbrat-Oreccioni, C. (1980): *L'énonciation : de la subjectivité dans le langage.* Paris : Armand Colin.

Lafon, P. (1980) : Sur la variabilité de la fréquence des formes dans un corpus. In *MOTS*, 1, 128-165.

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

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

Tutin, A. (2007a) : Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. In *Actes de TALN 2007. Communications affichées*, 283-292.

Tutin, A. (2007b) : Modélisation linguistique et annotation des collocations : application au lexique transdisciplinaire des écrits scientifiques. In S. Koeva, D. Maurel, M. Silberztein (Eds). *Formaliser les langues avec l'ordinateur.* Besançon : Presses universitaires de Franche-Comté, 189-216.

Tutin, A. (coord.) (2007c) : Lexique et écrits scientifiques. In : *Revue française de linguistique appliquée*, 12 (2), 5-14.

Tutin A. (à paraître) : Evaluative adjectives in academic writing in the humanities and social sciences. Communication acceptée au colloque Interlae. *Interpersonality in Written Academic Language*. Zaragoza, 11-13 décembre 2008.

Tutin, A., Novakova, I., Grossmann, F., Cavalla, C. (2006) : Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires. In : *Langue Française,* 150, 32-49.