Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS

Sylvain Hatier

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Sciences du langage Spécialité Informatique et Sciences du langage**

Arrêté ministériel : 25 mai 2016

Présentée par

Sylvain Hatier

Thèse dirigée par **Agnès Tutin**
codirigée par **Marie-Paule Jacques**

préparée au sein du **Laboratoire LIDILEM – EA 609**
dans l'**École Doctorale Langues, littérature et sciences humaines**

Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS

Thèse soutenue publiquement le **7 décembre 2016**,
devant le jury composé de :

Mme Marie-Claude L'Homme
Professeur à l'Université de Montréal, Présidente
Mme Cécile Fabre
Professeur à l'Université de Toulouse 2, Rapporteur
M. Alain Polguère
Professeur à l'Université de Lorraine, Rapporteur
Mme Agnès Tutin
Professeur à l'Université Grenoble Alpes, Directrice
Mme Marie-Paule Jacques
Maître de conférences à l'Université Grenoble Alpes, Co-encadrante

Remerciements

Je remercie en premier lieu mes directrices de thèse Agnès Tutin et Marie-Paule Jacques, pour leurs nombreuses remarques, relectures, suggestions (de références, de pistes de travail, de collaborations), en résumé de leur soutien et de leur patience qui m'ont permis de bénéficier d'un encadrement idéal pendant la thèse.

Je tiens également à remercier Cécile Fabre, Alain Polguère et Marie-Claude L'Homme, d'avoir accepté de faire partie de mon jury et d'avoir pris le temps d'examiner ce travail.

Je suis spécialement reconnaissant d'avoir pu effectuer ces travaux en bénéficiant d'une bourse de la région Rhône-Alpes et des possibilités offertes en participant à un projet financé par l'Agence Nationale de la Recherche. Je remercie d'ailleurs l'ensemble des partenaires du projet TermITH, et plus spécifiquement Évelyne Jacquey et Laurence Kister, pour leur expertise du LST, Sabine Barreaux pour nos échanges sur les méthodes d'extraction et Yannick Toussaint pour son accueil et son enseignement des treillis. J'exprime également ma reconnaissance à Patrick Drouin pour son accueil et nos échanges lors de mon séjour de recherche à Montréal et pour la collaboration renforcée entre nos laboratoires.

Je remercie aussi Magdalena Augustyn, Rui Yan, Hoai Tran, Agnès Tutin et Marie-Paule Jacques pour le travail collaboratif sur la ressource du LST. Remerciements également à Olivier Kraif notamment pour son aide dans les traitements de corpus et l'utilisation du Lexicoscope, ainsi que Claude Roux et Ágnes Sándor pour leurs conseils dans l'utilisation de XIP.

Je souhaite également remercier mes collègues du LIDILEM, Zohra et Isabelle pour leur incessante aide et disponibilité, ainsi que mes partenaires (ir)réguliers de café, de pause de bas du D et de patio. Un grand merci aussi pour vos relectures, Magda, Fred et Fanny. Je remercie enfin mes amis et ma famille — angloye, brachoise, saillansonne, toscadine, tlublublienne, d'ici et d'ailleurs — pour tous les moments « extrathétiques » partagés pendant ces quatre années.

Table des matières

Index des illustrations

Index des tableaux

Introduction

Notre travail a pour objet l'identification et l'analyse linguistique d'un lexique particulier, mobilisé dans les écrits scientifiques : le lexique scientifique transdisciplinaire (LST). Nous nous centrerons ici sur le LST présent dans le champ scientifique des sciences humaines et sociales (SHS).

À la suite de Tutin (2007a), nous définissons le LST comme le lexique renvoyant au discours sur les objets et procédures de l'activité scientifique, et assurant une fonction métadiscursive et métatextuelle dans l'argumentation et la structuration du discours scientifique.

L'exemple ci-dessous, extrait d'un article de recherche en économie, met en évidence les éléments de ce lexique partagé par un ensemble de disciplines :

- *Cet article permet de préciser certains éléments méthodologiques du test de la théorie des anticipations, dans le cadre d'une représentation explicite des anticipations des agents, à partir d'un modèle RVAR.*[1]

Le LST est constitué d'unités lexicales et d'expressions polylexicales référant aux textes scientifiques eux-mêmes (*article*), aux objets (*élément, théorie, modèle*) et procédures (*préciser, représentation*) relevant de l'activité scientifique. Ce lexique est fréquemment mobilisé dans les écrits scientifiques et constitue en cela un objet d'importance pour l'étude de ce genre ainsi que pour les applications connexes, notamment dans les domaines de la didactique et l'extraction terminologique.

Notre étude propose une approche outillée du LST propre aux écrits en français de dix disciplines des SHS, champ scientifique encore peu étudié au niveau de son lexique. Les travaux sur les écrits scientifiques des SHS ont jusqu'à présent porté principalement sur la dimension énonciative et la structuration textuelle. L'étude du lexique présente pourtant un intérêt particulier dans les SHS

[1] Jondeau, É. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. *Annales d'Économie et de Statistique*, 139-174.

où construction du discours et construction du savoir vont de pair (Grossmann, 2010). Le LST permet en cela d'analyser comment cette double construction s'effectue dans les différentes disciplines, notamment par l'étude des routines sémantico-rhétoriques, formulations récurrentes et servant des fonctions rhétoriques régulières, par exemple : *ces résultats/analyses montrent que…*

Étant donné ce rôle premier du LST dans les écrits scientifiques, il est essentiel pour tout scripteur, étudiant ou chercheur débutant, allophone ou francophone, de le maîtriser dans le contexte de la rédaction/compréhension de textes scientifiques. Nous proposons dans ce travail d'apporter une réponse à ce besoin, en élaborant une ressource lexicale adaptée, construite à partir des usages du genre en corpus.

Jusqu'à présent, les études à des fins didactiques sur les lexiques spécifiques aux écrits scientifiques ont principalement consisté[2] à dresser l'inventaire de ces éléments, en vue de proposer une liste de vocabulaire à enseigner aux apprenants. La ressource que nous élaborons intègre une analyse sémantique et une étude de la combinatoire des éléments du LST. En inventoriant et en organisant dans une typologie les acceptions mobilisées par les noms du LST, nous constituons une ressource offrant un accès sémasiologique et onomasiologique. De plus, le fait de procéder à cette classification sémantique des unités du LST nous permet d'effectuer une analyse des constructions syntactico-sémantiques récurrentes dans les articles de recherche en SHS et ainsi de proposer plusieurs expérimentations sur la dimension phraséologique du LST.

Outre leurs applications pour la description linguistique du genre des écrits scientifiques et en didactique, l'identification et la caractérisation sémantique du LST ont également un intérêt dans le domaine de l'indexation terminologique telle qu'elle a été envisagée au sein du projet TermITH[3]. Ce projet avait pour objet l'accès à l'information des articles de recherche, par l'indexation de termes présents dans ces écrits en SHS. Le LST, en tant que lexique de genre, est fortement mobilisé dans les articles de recherche au côté de la terminologie, et

---

[2] Ce constat ne concerne pas les travaux de Pecman (2004b) qui s'intéresse pour sa part à la dimension phraséologique de la *Langue Scientifique Générale,* non limitée aux SHS contrairement à nos travaux présentés ici.

[3] TermITH (Terminologie et Indexation de Textes en sciences Humaines) : ANR-12-CORD-0029 CONTINT. ATILF, INIST, LIDILEM, LINA, INRIA NGE and Saclay. http ://www.atilf.fr/ressources/termith/ [consulté le 04/04/2016]

permet à ce titre d'améliorer la reconnaissance des termes dans les textes. Il remplit alors une double fonction. D'une part, utilisé en tant que liste d'exclusion, le LST permet d'invalider un supposé statut terminologique des candidats-termes repérés automatiquement. D'autre part, certaines classes d'unités lexicales du LST constituent un indice de la présence de termes et permettent alors de valider le statut terminologique des candidats-termes.

Les études précédentes sur les lexiques spécifiques des écrits scientifiques ont permis de mettre en évidence la complexité de la tâche de délimitation des différents lexiques présents dans ce genre, tels que la terminologie, le LST et le lexique de la langue générale. Il nous apparaît alors nécessaire, pour mieux circonscrire notre objet d'étude, d'identifier les différentes propriétés de ces éléments. Nos traitements et analyses nous permettent d'aboutir à une définition enrichie du LST, tant au niveau extensionnel, en identifiant les éléments de ce lexique, qu'au niveau intensionnel, en proposant une analyse des propriétés lexicales, syntaxiques et sémantiques de ces éléments. L'interprétation manuelle des résultats de traitements automatiques nous permet alors de caractériser les éléments du LST à ces différents niveaux de l'analyse linguistique. Pour l'identification des éléments du LST, nous nous concentrons sur les mots simples des catégories dites de mots pleins : adjectifs, adverbes, noms et verbes. L'ensemble de ces traitements a alors pour but la constitution d'une ressource lexicale représentative de l'usage du LST dans les articles en SHS, dans la mesure où les applications envisagées se situent dans le cadre de ce champ scientifique. En conséquence, notre travail repose sur l'analyse des propriétés linguistiques des éléments du LST dans un corpus actualisant ce genre des articles de recherche en SHS. L'adéquation aux applications précitées de notre ressource sémantique du LST s'appuie donc sur son ancrage discursif, et sur la nécessaire étape d'interprétation des résultats des méthodes automatiques de classification (Habert & Nazarenko, 1996).

Nous nous situons pour le présent travail dans une approche de linguistique de corpus, outillée, où le traitement automatique du langage intervient, dans un premier temps, pour faire émerger des phénomènes linguistiques récurrents susceptibles de caractériser notre objet d'étude. L'analyse de ces phénomènes, dans un second temps, nous permet de valider manuellement les propriétés du LST que nous souhaitons inclure dans notre ressource lexicale. Les traitements

automatiques et l'analyse manuelle sont ainsi associés dans l'étude des unités lexicales du LST à trois niveaux : leur extraction, l'étude de leurs propriétés lexico-syntaxiques et leur caractérisation sémantique. Le recours à la validation manuelle a alors pour fonction de s'assurer de la pertinence des éléments automatiquement extraits et ce faisant d'enrichir la ressource du LST au fur et à mesure des traitements. L'analyse du LST ainsi que l'élaboration de la ressource lexicale sont ainsi effectuées de manière incrémentale, les résultats des différents traitements étant ré-exploités dans les phases suivantes.

Nous adoptons une méthode hybride, adaptée aux propriétés particulières du LST, objet d'étude se révélant complexe à circonscrire et à caractériser. Nous verrons que ce lexique spécifique des écrits scientifiques, se situant entre lexique terminologique et lexique de la langue générale, partage des points communs avec ces deux lexiques et de ce fait a des frontières peu marquées, ce qui complexifie son étude. De plus, bien que mobilisé dans l'écrit scientifique, le LST n'échappe pas à l'écueil de la polysémie et nécessite en conséquence un traitement sémantique adapté.

Le plan de notre étude est comme suit. Le chapitre 1 détaille en premier lieu les motivations à la source de notre travail et précise les applications, présentes et futures, de la ressource lexicale. Nous y examinons par la suite notre objet d'étude, en précisant sa définition et en revenant sur les différents travaux sur les lexiques associés aux écrits scientifiques. Nous verrons, au travers des dénominations multiples de ces lexiques, que la question de la délimitation du LST n'est pas entièrement résolue. L'état de l'art des travaux du domaine nous permettra ensuite de distinguer les différentes approches adoptées pour l'étude de ces lexiques. Les différentes méthodes d'extraction mettent ainsi en évidence deux propriétés principales pour le LST : la transdisciplinarité et la spécificité. Nous situerons alors notre objet d'étude et le genre dans lequel il est convoqué afin d'identifier les particularités des textes composant notre corpus et par la même les particularités du LST. Nous conclurons ce premier chapitre en précisant notre approche de linguistique outillée, associant traitement automatique du langage et analyse linguistique.

Le chapitre 2 présente la première phase de nos traitements, à savoir l'extraction semi-automatique des éléments du LST. Nous détaillerons dans un

premier temps les outils et corpus sur lesquels nous nous appuyons pour identifier les mots simples de ce lexique. Dans un second temps, nous présenterons notre méthode d'extraction, consistant en l'application de critères lexicométriques suivie d'une phase de validation manuelle. Enfin, nous proposerons une comparaison entre notre méthode et deux autres protocoles afin d'évaluer l'effet des mesures et du corpus sur les résultats de l'extraction.

Le chapitre 3 est consacré aux traitements sémantiques que nous avons effectués sur les éléments du LST identifiés lors du chapitre précédent. Nous reviendrons en premier lieu sur les travaux en sémantique et typologie lexicale afin de poser le cadre de constitution de notre classification. Nous présenterons ensuite les traitements sémantiques nous permettant d'aboutir, pour les noms du LST, à une ressource dont les entrées sont des acceptions structurées dans une typologie à deux niveaux, classes et sous-classes sémantiques.

Le chapitre 4 présente deux expérimentations de catégorisation sémantique automatique que nous avons effectuées pour les noms du LST. Nous poursuivons un double objectif à travers ces expérimentations. D'une part, leurs résultats peuvent faire émerger des classes de noms pertinentes à intégrer dans la classification effectuée manuellement. D'autre part, ces méthodes automatiques permettent de faciliter l'intégration de nouveaux éléments dans la perspective de la maintenance de notre ressource lexicale. La première expérimentation, reprenant la méthode des prototypes, nous permettra d'introduire la notion de degré d'appartenance à une classe pour les éléments du LST et d'identifier les propriétés partagées par les prototypes d'une même classe. Dans la seconde partie, nous expérimenterons une méthode d'analyse en concepts formels, basée sur la théorie des treillis de Galois, dont l'intérêt premier est de faire émerger des attributs définitoires pour les classes de notre typologie.

Le chapitre 5 est consacré aux expressions polylexicales impliquant les éléments du LST. Nous aborderons ainsi la dimension phraséologique du LST en étudiant plus particulièrement trois types de constructions définies par des contraintes syntaxiques et sémantiques. La première partie détaillera nos travaux en collaboration avec Rui Yan sur les patrons verbaux du LST, dans une perspective didactique d'aide à la rédaction scientifique. La deuxième partie présentera une première expérimentation sur les cooccurrences entre éléments du

LST et éléments terminologiques, avec pour objectif la détection de patrons lexico-syntaxiques délimiteurs de termes. Pour conclure, la troisième partie présentera une application de la classification sémantique du LST dans la détection de routines sémantico-rhétoriques, unités phraséologiques exprimant des fonctions typiques de l'écrit scientifique telles que le positionnement ou la définition de la problématique.

La conclusion nous permettra d'indiquer les perspectives ouvertes par l'ensemble de notre travail. Nous verrons que la classification sémantique du LST offre de multiples possibilités d'applications pour l'étude de la dimension phraséologique dans les écrits scientifiques. La détection des expressions polylexicales contenant un élément du LST nous permettra alors de proposer des pistes de travail au niveau de l'indexation terminologique, de l'enrichissement de la ressource du LST et plus globalement au niveau de l'analyse de la phraséologie scientifique transdisciplinaire.

Le mot doit faire naître l'idée ; l'idée doit peindre le fait : ce sont trois empreintes d'un même cachet ; et, comme ce sont les mots qui conservent les idées et qui les transmettent, il en résulte qu'on ne peut perfectionner le langage sans perfectionner la science, ni la science sans le langage, et que, quelque certains que fussent les faits, quelque justes que fussent les idées qu'ils auraient fait naître, ils ne transmettraient encore que des impressions fausses, si nous n'avions pas des expressions exactes pour les rendre.

Lavoisier, Traité élémentaire de chimie (1789).

Chapitre 1
Lexique scientifique transdisciplinaire : intérêts, propriétés et approches

Sommaire

Comme nous l'avons abordé précédemment, le présent travail a pour objectif l'analyse d'un lexique partagé par un ensemble de disciplines des sciences humaines et sociales (SHS). L'étude de ce lexique, qui renvoie aux objets et procédures de l'activité scientifique, présente à nos yeux plusieurs intérêts.

1.1 Motivations de l'étude du LST

En premier lieu, en tant que lexique associé au genre de l'écrit scientifique, le lexique scientifique transdisciplinaire (LST) tient un rôle essentiel dans la communication scientifique et de ce fait dans la science elle-même, tel que peut l'exprimer Lavoisier dans la citation liminaire. L'étude de ce lexique présente ainsi un double intérêt, linguistique et épistémologique.

Au niveau linguistique, le LST peut être envisagé comme une entrée préférentielle dans l'étude du genre de l'écrit scientifique. Rinck (2006, p. 240) rappelle d'ailleurs que « comme le lexique varie avec le genre, la caractérisation d'un genre peut prendre comme objet ses aspects lexicaux ». Cette étude permet ainsi d'interroger et de caractériser le genre de l'article de recherche par son lexique propre.

Au niveau épistémologique, le lexique commun à un ensemble de disciplines permet de mettre en évidence les procédures et concepts partagés dans le contexte de la communication scientifique, donc de la construction du savoir. Car, selon Grossmann (2010, p. 419), dans les SHS, « construction du discours et construction du savoir vont de pair ». Le LST est alors utilisé dans l'analyse des épistémologies présentes dans les disciplines, par exemple par le repérage de routines sémantico-rhétoriques (Tutin & Kraif, 2016), telles les routines introductrices de nouveaux termes (Jacques, 2011). Les particularités et points communs peuvent alors être identifiés par une analyse contrastive des fonctions rhétoriques et routines mobilisées dans les différentes disciplines.

En plus de cet apport du LST dans l'étude de la construction du savoir, plusieurs objectifs ont motivé notre travail d'extraction et d'analyse d'un lexique dont l'inventaire mène à de nombreuses applications (Rinck, 2010). Malgré l'importance de ce lexique et la profusion de travaux sur le sujet, peu d'études récentes se sont intéressées à un lexique scientifique du français dans le domaine

des SHS. La question de l'ensemble disciplinaire convoquant ce lexique est pourtant essentielle. Ainsi, Blumenthal (2007) compare deux corpus de sciences exactes et de SHS et observe un recoupement de seulement 47 % pour les 500 noms les plus fréquents dans les deux corpus. En analysant la combinatoire de ces noms, il fait alors apparaître les nombreuses différences au niveau lexical et sémantique entre ces deux ensembles disciplinaires. La définition des disciplines des textes sur lesquels portera notre analyse du LST est donc primordiale. Nous proposons dans le présent travail de pallier l'absence d'étude sur ce lexique spécifique aux SHS en élaborant une ressource du LST, représentative de cet ensemble disciplinaire et répondant aux objectifs détaillés ci-après. Cette ressource sera exploitable de façon directe dans deux champs principaux.

1.1.1 LST et extraction terminologique

D'une part, dans le champ de la recherche d'informations et plus précisément dans le cadre du projet TermiTH[1], le LST est utilisé dans l'indexation automatique en termes, à l'instar de Da Sylva (2010). Sa fonction dans le processus d'identification de termes est alors double :

- Le LST sert de filtre d'exclusion pour l'extraction des termes simples et complexes. Les candidats termes appartenant au LST sont donc dans un premier temps moins susceptibles d'être validés. Inversement, les candidats termes n'appartenant pas au LST, et qui par la cooccurrence présentent un environnement susceptible de favoriser l'acception terminologique, peuvent être validés. Ainsi, le nom *sujet* a une acception terminologique en cooccurrence avec les mots *phrase*, *syntaxique*, *verbe* et une acception transdisciplinaire en cooccurrence avec *étude, aborder, article,* comme l'illustrent les deux exemples ci-dessous :
    - *Il est manifeste que le nombre d'études sur le sujet est impressionnant.*[2]

[1] TermITH (Terminologie et Indexation de Textes en sciences Humaines) : ANR-12-CORD-0029 CONTINT. ATILF, INIST, LIDILEM, LINA, INRIA NGE and Saclay. http://www.atilf.fr/ressources/termith/ [consulté le 04/04/2016]

[2] Crahay, M. (2004). Peut-on conclure à propos des effets du redoublement?. *Revue française de pédagogie*, 11-23.

- *Ce test n'est quand même pas inutile, à preuve ces exemples de l'impossibilité d'une dislocation entre le sujet et le verbe.*[3]

- Outre cet emploi comme lexique d'exclusion, le LST peut servir d'indice dans la recherche de termes. Il constitue d'ailleurs « un critère de différenciation des emplois terminologiques des termes et des emplois en langue courante » selon Kister & Jacquey (2012, p. 909). Autrement dit, le LST est un indice de la présence de termes car majoritairement délimiteur de termes. Par exemple, le verbe *nommer* et le nom *étude* font partie du LST. Or, les suites *'nous nommons'* et *'étude de'* introduisent fréquemment des unités terminologiques. Plus précisément, certaines classes d'unités lexicales du LST sont préférentiellement introductrices de termes (Jacquey *et al.*, 2013).

1.1.2 LST et enjeux didactiques

L'autre champ d'application de la ressource du LST se situe au niveau didactique, plus particulièrement dans l'aide à la rédaction/compréhension d'écrits scientifiques. Les éléments du LST, ainsi que nous le verrons, permettent la structuration de l'argumentation et du discours scientifique. Or, les scripteurs étudiants allophones, face à un genre d'écrit codifié comme l'écrit scientifique, manifestent des lacunes dans la maîtrise des organisateurs textuels, comme le remarque Tran (2014), ce qui engendre une organisation des idées peu cohérente en production.

De la même façon, Pecman (2004b) constate la difficulté des apprenants à assimiler ce lexique, comme Granger & Paquot (2010) qui observent également une sous-utilisation d'un grand nombre de verbes du LST par les apprenants. Elles notent qu'en prenant en compte les mots de la *General Service List* (*GSL*)[4], de l'*Academic Word List* (*AWL*)[5] et les mots spécifiques du domaine (la terminologie), le seuil critique de 95 % de couverture lexicale, pourcentage des mots dont la connaissance est nécessaire pour la compréhension d'un texte, est atteint. Ce seuil de 95 % est un fort argument en faveur de l'utilisation du LST pour

[3] Dugas, A. (2001). Une analyse des constructions transitives indirectes en français. *Travaux de linguistique*, (1), 111-120.

[4] *General Service List* : mots fréquents de la langue générale en anglais extraits par West (1953).

[5] *Academic Word List* : lexique académique extrait par Coxhead (2002).

les applications didactiques. En ce qui concerne l'aide à la rédaction de l'écrit scientifique, une ressource du LST, en s'ajoutant aux connaissances sur le lexique de la langue générale (correspondant à la *GSL*) et celles sur le lexique du domaine particulier (correspondant à la discipline concernée), permet à l'apprenant d'avoir une maîtrise de l'ensemble des mots convoqués dans le genre.

Les applications didactiques des éléments du LST peuvent par ailleurs servir de mesure pour une évaluation de la ressource, comme le proposent Paquot & Bestgen (2009). Dans la visée d'applications pédagogiques, nous avons d'ailleurs spécifiquement enrichi la ressource au niveau des constructions concernant les verbes du LST, comme détaillé dans le chapitre 5. Nos travaux sur les patrons verbaux[6] permettent ainsi de dégager les constructions posant problème pour les apprenants, par comparaison aux usages dans les articles de recherche, et ainsi d'adapter au mieux la ressource aux besoins de ces apprenants.

En cohérence avec l'ensemble de ces objectifs et perspectives, notre étude du LST se compose d'analyses de plusieurs ordres :

- Une analyse lexicologique, en procédant à l'inventaire des éléments le constituant.
- Une analyse syntaxique, au travers des patrons verbaux et l'analyse de constructions syntaxiques intégrant un élément du LST.
- Une analyse sémantique, en procédant à l'identification des acceptions du LST mobilisées dans le corpus et en élaborant une classification sémantique des noms du LST.

Étant donné les applications envisagées, il est nécessaire de procéder à une définition double du LST :

- en intension, à travers ses propriétés linguistiques dans notre corpus d'analyse ;

[6] Nous avons collaboré avec Rui Yan, doctorante au LIDILEM, pour plusieurs études (Yan & Hatier, 2016; Hatier & Yan, à paraître) sur les constructions verbales fréquentes dans l'écrit scientifique, dont l'intérêt pédagogique a déjà été prouvé (Granger & Paquot, 2009b). Nous avons mis en place une méthode d'extraction des constructions verbales basée sur les dépendances. Suite à cette étape, Yan a modélisé les patrons verbaux mobilisés dans les articles de recherche, élaborant ainsi une ressource adaptée aux spécificités de l'écrit scientifique, intégrant des informations sur les constructions aux niveaux lexical (cooccurrents préférentiels), syntaxique (arguments, alternances) et sémantique (définition, étiquette sémantique).

- en extension, en répertoriant l'ensemble de ses éléments, acception par acception.

Ayant ainsi déterminé nos objectifs, nous détaillons dans la partie suivante les propriétés de notre objet d'étude, le lexique scientifique transdisciplinaire, en tirant parti des travaux antérieurs dans le domaine.

1.2 Lexique Scientifique Transdisciplinaire

Notre objet d'étude, le lexique scientifique transdisciplinaire (LST), est considéré pour le présent travail dans son usage pour un corpus particulier, composé d'article de recherches en SHS. Cette restriction à un ensemble défini de textes s'appuie sur la définition du LST comme un lexique de genre, genre alors représenté par un corpus dont la composition influe sur l'ensemble de nos traitements et analyses[7]. Les ressources élaborées, ainsi que les analyses effectuées, le sont dans le cadre du genre de l'écrit scientifique en SHS. Conséquemment, les applications didactiques d'une ressource du LST sont envisagées dans le cadre de la rédaction scientifique en SHS. De la même manière, l'emploi du LST pour l'identification de termes est considéré pour les articles en SHS (domaine de connaissances traité dans le projet TermITH), l'intérêt de notre ressource pour d'autres ensembles disciplinaires restant à évaluer.

1.2.1 Définition du LST

Nous reprenons la définition de Tutin (2007a) pour qui le LST renvoie au discours sur les objets et les procédures scientifiques. Le LST est un lexique métascientifique et métadiscursif, portant aussi sur les interactions au sens large entre un auteur et son destinataire dans une même communauté.

Bien que spécifique, le LST ne fait pas partie des langues de spécialité au sens où il n'est pas restreint à un domaine de connaissances précis. Il n'intègre donc pas la terminologie des domaines, liée aux thématiques des textes. La terminologie, disciplinaire, renvoie à des concepts dont la définition n'est pas forcément partagée par l'ensemble des disciplines de SHS représentées dans notre corpus.

[7] Ou tel que le formule Sinclair (1991, p. 13) : « The results are only as good as the corpus ».

Dans l'extrait d'article d'anthropologie suivant, les segments soulignés appartiennent au LST :

- *Du point de vue$_{LST}$ anthropologique ce déplacement$_{LST}$ de paradigme$_{LST}$ met en jeu$_{LST}$ la définition$_{LST}$ même de l'identité en renforçant$_{LST}$ le divorce entre l'identité civile et l'identité personnelle ou sociale*[8]

Le LST est composé de mots simples (*déplacement, paradigme, définition, renforcer*) et d'expressions polylexicales (*mettre en jeu, du point de vue*) appartenant aux quatre catégories syntaxiques ouvertes : nom, verbe, adverbe et adjectif. Ce lexique est non terminologique (*identité* renvoie ici à un terme de l'anthropologie), largement abstrait, et très fréquent dans l'écrit scientifique.

Le LST est le lexique mobilisé dans le discours sur les objets et les procédures scientifiques. Il est en cela primordial dans l'argumentation et la structuration du discours et de la pensée scientifique (Drouin, 2007). Cependant, un grand nombre de ses éléments est potentiellement présent dans d'autres genres que l'écrit scientifique (tel le genre journalistique). Tel est le cas par exemple pour : *du point de vue*, *définition, considérer, hétérogène*. Le LST permet d'organiser la communication du savoir, de remplir des fonctions rhétoriques essentielles dans la communication scientifique : introduire un concept, argumenter, souligner un fait, modaliser son discours, se positionner par rapport aux pairs.

Nous retenons, pour le LST, les trois propriétés principales suivantes :

- la transdisciplinarité : il n'est pas spécifique à une discipline mais à un genre ;
- la fréquence : il inclut des éléments lexicaux fréquents dans le discours scientifique ;
- la spécificité : il est spécifique au genre de l'écrit scientifique, par sa fonction de désignation des procédures et outils de l'activité scientifique.

Si la plupart des travaux reprennent ces propriétés dans leur définition du lexique de l'écrit scientifique, nous verrons dans la partie suivante que les approches du LST sont multiples, de même que les objectifs qui motivent son étude.

[8] Dubey Gérard, « Nouvelles techniques d'identification, nouveaux pouvoirs.. Le cas de la biométrie », *Cahiers internationaux de sociologie* 2/2008 (n° 125), p. 263-279

1.2.2 Études des lexiques spécifiques de l'écrit scientifique

L'étude des lexiques scientifiques pose le premier problème de leur identification. Il est alors nécessaire dans un premier temps de délimiter les différents lexiques présents dans l'écrit scientifique, ainsi que la section 1.2.2.1 le détaille. L'inventaire des lexiques permet alors de définir les propriétés particulières au LST, propriétés traduites en critères quantifiables abordés section 1.2.2.3. Ces critères sont ensuite intégrés dans différentes méthodes d'extraction que nous présentons dans la section 1.2.2.4.

1.2.2.1 Délimitation des lexiques mobilisés dans l'écrit scientifique

Différences de méthodes mises à part, l'ensemble des travaux définit le LST comme un lexique traversant, essentiel dans l'exposé de l'activité scientifique et dans l'organisation du discours scientifique. Les autres lexiques intervenant dans ce discours scientifique sont différemment détaillés, selon les études, reflétant la complexité inhérente à toute tâche de délimitation de lexiques.

Tutin (2007a) distingue cinq lexiques intervenant dans les écrits scientifiques : LST, lexique abstrait, lexique méthodologique disciplinaire, lexique terminologique, lexique de la langue générale. Selon elle, la difficulté première est alors de circonscrire ce lexique, de délimiter l'objet d'étude. Pecman, citant Depecker (2002, p. 63), rappelle également cette difficulté dans l'étude de ce qu'elle nomme la *Langue Scientifique Générale* :

> « Chacune des langues forme un ensemble qui n'est ni clos ni imperméable. La langue technique et la langue scientifique ne sont que des spécialisations de la langue commune ».

On peut d'ailleurs constater que l'étude et l'inventaire des lexiques à l'œuvre dans l'écrit scientifique ont été et restent au centre de nombreuses études. Nation (2001) classe, lui, les unités lexicales rencontrées dans l'écrit scientifique en quatre ensembles :

- les mots très fréquents : mots outils et mots de la langue générale ;
- les mots du vocabulaire académique : communs à un ensemble divers de textes académiques ;

- les mots techniques : liés aux sujets abordés dans les textes ;
- les mots peu fréquents : ensemble le plus important, ni très fréquents, ne relevant ni du LST, ni du vocabulaire technique.

Nous retrouvons ainsi dans les différentes études un découpage comparable en lexiques qui permet de mettre en lumière le fait que ce genre ne mobilise pas uniquement les lexiques de la « langue générale » et de la terminologie. Plusieurs sortes de lexiques interagissent dans l'écrit scientifique, dont un, le LST, a la particularité d'être à la fois propre au discours scientifique (et proche en cela de la terminologie) et partagé par des disciplines différentes (propriété également vraie pour la langue générale). Afin d'identifier ce lexique dans le contexte de l'écrit scientifique, les travaux le définissent comme indépendant des thématiques (ce qui le distingue de la terminologie) et fréquent dans le genre académique (pour le différencier de la langue générale).

Pour autant, les frontières sont complexes à tracer entre ces différents lexiques, particulièrement entre le LST et un ensemble lexical composé de mots abstraits également mobilisé dans la langue générale. Dans son travail sur le vocabulaire académique, Paquot (2010, p. 20) insiste sur l'importance de ces « General Service Words[9] », absents de l'*Academic Word List* de Coxhead et étudiés par Cowan (1974) sous le terme de *sub-technical vocabulary*. Le lexique abstrait général (LAG), ainsi que nous le nommons, est constitué d'éléments non spécifiques au genre mais sur-représentés dans l'écrit scientifique, tels : *manière, difficulté, enjeu, rôle, inhérent, abstrait, précédent, aborder, confronter, participer, effectivement, strictement.*

Ainsi que le note Paquot (2010, p. 212), ces mots servent également à l'organisation du discours scientifique et sont en cela aussi important à maîtriser que le LST dans l'optique de la rédaction d'écrits scientifiques. La différence majeure entre le LST et le LAG est liée au fait que ce dernier ne mobilise pas d'acceptions réellement spécifiques au genre scientifique, et intègre des éléments lexicaux fréquents dans la langue générale. À ce titre, le LAG présente des recoupements avec d'autres lexiques, dont l'intérêt en didactique a notamment motivé les travaux de West (1953) sur la *General Service List*. Nous verrons ainsi

[9] Elle donne pour exemple en anglais : *result, group, cause, develop, model, plan*

section 2.2.2.2.2 que nous intégrons dans la ressource du LST certains éléments du LAG qui répondent aux critères lexicométriques d'extraction.

Si les études se rejoignent sur la définition de ces lexiques présents dans les écrits scientifiques, la section suivante montre que les approches adoptées pour leur description sont diverses et en grande partie liées aux applications et objectifs envisagés.

1.2.2.2 Approches et analyses des lexiques scientifiques

Les travaux sur les lexiques scientifiques sont généralement motivés par, ou étroitement liés à, un objectif déterminé tel que l'élaboration d'une ressource à perspectives didactiques (Coxhead, 2002; Pecman, 2004b; Paquot, 2010), ou l'aide à l'indexation terminologique (Da Sylva, 2009, 2010). D'autres études abordent le LST sous l'angle de la description linguistique (Phal & Beis, 1972; Drouin, 2007; Tutin, 2007c). L'ensemble de ces travaux ont en commun de se baser sur des corpus pour faire émerger un lexique particulier, dont la nature varie avec la composition desdits corpus. Les études se rejoignent également pour mettre en avant le rôle essentiel du LST dans la communication scientifique.

Nous montrons dans cette partie que trois types d'approches des lexiques scientifiques peuvent être distingués dans la littérature. Un premier groupe de travaux, présenté section 1.2.2.2.1, se concentrent sur les propriétés lexicales, dans le but d'élaborer une liste généralement à visée didactique pour l'apprentissage de vocabulaire. Les travaux réunis dans le deuxième groupe, section 1.2.2.2.2, s'intéressent aux lexiques scientifiques dans une perspective énonciative et discursive du genre de l'écrit scientifique. Le dernier type d'approche intègre un niveau sémantique dans l'analyse du lexique, en proposant une classification ou un étiquetage de ces éléments, comme nous le détaillons section 1.2.2.2.3.

1.2.2.2.1 Dimension lexicale

La majorité des études sur les lexiques similaires au LST est guidée par des objectifs didactiques. En effet, nous avons vu que la maîtrise de ce lexique est un enjeu pour tout scripteur, tant dans un contexte de Français Langue Étrangère que

de Français sur Objectif Universitaire[10]. Conséquemment, nombre de travaux visent à l'élaboration d'une liste de vocabulaire à destination des scripteurs de ce genre.

Ainsi, Phal & Beiss ont procédé, dans un but didactique, à l'étude du *Vocabulaire Général d'Orientation Scientifique* (*VGOS*), lexique « commun à toutes les spécialités[11] [servant] à exprimer les notions élémentaires dont elles ont toutes également besoin […] et les opérations intellectuelles que suppose toute démarche méthodique de la pensée » (1972, p. 9). Ils poursuivent en cela les études qui dressent l'inventaire d'un *français fondamental* (Gougenheim, Michéa, Rivenc & Sauvageot, 1964), à partir d'un corpus de manuels scolaires en français, dans une optique didactique en direction des non-natifs. Leur large description du lexique prend également en compte les lexies complexes et certaines structures lexico-syntaxiques. Ils proposent une ressource riche du *VGOS*, dont les informations sur les contextes d'apparition permettent une meilleure appropriation par les apprenants de ce lexique complexe.

Coxhead (2002), s'intéresse également à la perspective de l'enseignement de la langue scientifique, mais pour scripteurs anglophones natifs ou apprenants[12]. Elle recense les éléments lexicaux qu'il est nécessaire de maîtriser pour la rédaction scientifique en anglais à partir d'un corpus de manuels et d'articles. L'*Academic Word List* (*AWL*)*,* est composé des mots les plus fréquents dans un corpus de textes académiques. Notant qu'il est plus facile d'apprendre les mots par famille morphologique (*evidence, evidential, evidently*), Coxhead procède au regroupement de ces dérivés afin de proposer une ressource la plus utile possible pour l'aide à la maîtrise de ce lexique particulier. Cependant, malgré ce type de traitement, cette ressource intègre peu d'informations linguistiques, mis à part la structuration du lexique en dix classes de fréquence (dont la première est constituée des éléments les plus fréquents et les plus répartis). Par exemple, l'*AWL* ne permet pas la distinction entre les catégories nominale et verbale pour les mots

---

[10] Le français sur objectif universitaire a pour but de préparer des étudiants allophones à suivre des études au sein d'universités francophones. Le français langue étrangère ou FLE ne définit pas de but précis dans l'apprentissage du français par les allophones.

[11] Les spécialités ici sont représentées par un corpus de manuels en physique, chimie et sciences naturelles.

[12] Son travail se situe dans la lignée de West, dont la *General Service List* est destinée aux natifs et non-natifs (« developped with the needs of ESL/EFL learners in mind » selon les mots de Coxhead (2000, p. 213)).

appartenant potentiellement aux deux catégories (*survey, function, approach*). Comme nous le verrons par la suite, la qualité de l'extraction et de la caractérisation du LST est soumise à la qualité des traitements d'enrichissement de corpus (simple segmentation, lemmatisation, annotation en parties du discours, en relations de dépendance). L'absence d'informations syntaxiques et sémantiques ne serait pas compatible avec les applications que nous avons détaillées précédemment.

Parallèlement à ces travaux d'inventaire du LST en vue d'objectifs didactiques, plusieurs études abordent ce lexique dans un but d'amélioration de l'indexation en termes (Drouin, 2007; Da Sylva, 2009, 2010). Da Sylva, dans son étude du *Vocabulaire Savant de Base* (VSB) constate ainsi que ce lexique permet une meilleure extraction terminologique et conclut que le VSB peut avoir une utilité dans la phase de description et d'indexation des documents.

Les travaux de Drouin (2007), dont la méthode est détaillée section 1.2.2.4, s'intéressent au LST dans une double perspective didactique et de description linguistique. Il procède ainsi à l'extraction automatique du LST, en anglais et en français, rendant ainsi possible une utilisation de ce lexique pour la traduction. Il propose également une première analyse des collocations dans ce lexique et note qu'une approche syntaxique plutôt que purement statistique de ces phénomènes est une perspective à explorer. Nous nous situons dans ce cadre d'analyse, mettant au centre des traitements les propriétés lexico-syntaxiques du LST (voir notamment section 1.3.5.2).

1.2.2.2.2 Dimension énonciative et discursive

Les lexiques scientifiques sont également étudiés dans leur dimension énonciative et discursive, avec pour objectif une caractérisation du genre de l'écrit scientifique. Ainsi, Rinck (2006, p. 242), dans ses travaux sur l'article de recherche, s'intéresse au lexique transdisciplinaire dont les aspects énonciatifs comportent « un intérêt [...] pour l'étude de l'argumentation et de la construction des textes ». Rinck interroge dans son étude les pratiques discursives à l'œuvre dans deux disciplines des SHS (lettres et sciences du langage) et montre ainsi que l'article scientifique est un genre à la dimension rhétorique très présente.

Tran (2014) s'intéresse à la dimension discursive du LST, et prend pour point de départ les marqueurs discursifs dans l'écrit scientifique. Son travail aborde spécifiquement les marqueurs polylexicaux dont elle dégage les propriétés syntaxiques et sémantiques. Elle les intègre dans une typologie des marqueurs discursifs, composée de deux grands ensembles correspondant aux deux fonctions les plus convoquées dans ce genre : la fonction métadiscursive et la fonction argumentative. Elle propose ainsi une description riche de la phraséologie adverbiale du LST, avec l'objectif d'améliorer l'enseignement/apprentissage de ces marqueurs discursifs.

Les dimensions discursive et énonciative du genre de l'écrit scientifique sont également au centre des travaux de Fløttum, Dahl, & Kinn (2006) qui proposent une caractérisation du genre, et de certaines disciplines, au niveau de phénomènes tels que la présence de l'auteur, les variations disciplinaires ou l'expression de l'évidence. Les phénomènes rhétoriques sont par ailleurs au centre de nombreux travaux issus du projet Scientext[13], dont certains sont présentés par Tutin & Grossmann (2014), et ont notamment pour objet : le lexique évaluatif, les verbes d'opinion et de constat, la rhétorique de la surprise, de la filiation ou l'auctorialité.

Nous notons ici que la ressource du LST que nous voulons élaborer, y compris la classification sémantique, pourrait avoir un apport certain en rendant automatiquement identifiables plusieurs de ces phénomènes linguistiques, énonciatif ou discursif, ainsi que le chapitre 5 en donne plusieurs exemples. Un des apports principaux de notre ressource du LST se situe alors au niveau de la description linguistique du genre de l'écrit scientifique aux niveaux énonciatif, discursif et rhétorique, à condition d'intégrer à notre ressource une description sémantique du LST.

[13] Projet dans le cadre de l'ANR « Corpus et outils de la recherche en sciences humaines et sociales » (2007-2010). Site du projet : http://scientext.msh-alpes.fr

1.2.2.2.3 Dimension sémantique

L'intégration du niveau sémantique dans l'analyse des lexiques scientifiques intervient à différents degrés. Le fait d'identifier la ou les acceptions mobilisées par les mots de ces lexiques constitue alors une première étape. Certains travaux proposent également une classification de ce lexique et attribuent des étiquettes sémantiques aux éléments du lexique étudié.

Par exemple, Paquot (2010, p. 81), dans son étude de l'*Academic Vocabulary*, intègre plusieurs types d'analyses sémantiques. Outre le recours à un étiqueteur sémantique automatique (voir section 3.2.2.2), elle étudie les fonctions rhétoriques associées à l'emploi de ce lexique, telles l'expression de la cause, la reformulation, la comparaison ou la mise en contraste. Ce lexique constitue alors une entrée intéressante dans l'étude des routines sémantico-rhétoriques, comme nous le détaillerons dans la partie 5.4. Paquot ne propose cependant pas de classification du lexique scientifique à proprement parler et ne désambiguïse pas les éléments de son lexique, se limitant alors au niveau du mot, et non de l'acception.

En termes d'analyse lexicale, syntaxique et sémantique d'un lexique scientifique, les travaux de Pecman (2004b, 2007) sur la *Langue Scientifique Générale* (*LSG*), nous paraissent les plus poussés. Pecman (2004b) adopte une approche phraséologique de la *LSG* dans un but didactique, et d'analyse contrastive, sur un corpus multilingue d'articles de recherche en sciences exactes (anglais, français et serbe). Elle propose une classification notionnelle de la phraséologie de la *LSG* en 125 concepts reposant sur quatre grandes sphères conceptuelles (2004b, p. 293) :

- La sphère de la « scientificité » intègre les concepts référant à un objet, une action, une qualité, de nature purement scientifique tels que {expérience}, {découverte}, {évaluation} ou {variation}.
- La sphère de « l'universalité » renvoie à un univers conceptuel commun ({temporalité}, {quantité}), plus large que la *LSG*, et complexe à distinguer de la sphère précédente selon Pecman. Cette sphère peut pour ces raisons s'apparenter à ce que nous nommons le LAG.

- La sphère de la « modalité » concerne les unités phraséologiques « dont le sémantisme met au premier plan l'auteur du discours » (Pecman, 2004b, p. 294), modalité épistémique, déontique, etc.
- La sphère de la « discursivité » : renvoie à des impératifs discursifs tels que l'indication d'exemples, d'une citation, la présentation de l'objet d'étude.

Les concepts sont eux-mêmes liés par des relations typées : genre-espèce, partie-tout, analogie, antonymie, etc. Elle associe ainsi unité phraséologique et étiquette notionnelle, par exemple le concept {hypothèse} est relié aux collocations suivantes : *adopter une hypothèse, confirmer une hypothèse, émettre|formuler|avancer une hypothèse*. Ces collocations sont ensuite regroupées dans des schémas collocationnels qui listent l'ensemble des unités phraséologiques correspondant à un même concept, à la manière des routines que nous proposons d'identifier dans la section 5.4. Pecman propose de cette manière un accès onomasiologique à la LSG en permettant de consulter pour une notion les différentes réalisations possibles. Par une approche empirique, sur corpus, elle identifie et caractérise au niveau sémantico-syntactique les unités phraséologiques de la LSG, « fond de formules préfabriquées permettant d'exprimer les notions fondamentales et les raisonnements communs aux différentes sciences exactes » (Pecman, 2004b, p. 128).

Le travail de Pecman se distingue ainsi par une analyse fine de la LSG (aux niveaux syntaxique, notionnel, fonctionnel) et par le type d'unités qu'elle choisit d'étudier, les unités phraséologiques, dont l'importance est également soulignée par Nation (2001), qui constate la difficulté à maîtriser les cooccurrences spécifiques à ce genre pour les scripteurs. Comme nous le verrons dans les chapitres suivants, nous n'aborderons l'aspect phraséologique du LST que pour en caractériser les mots simples. Nous ne négligeons cependant pas les expressions polylexicales (collocation, expression figée) dont l'analyse et le traitement ont été et/ou sont effectués par d'autres membres du laboratoire LIDILEM. Ainsi, en plus des locutions adverbiales déjà étudiées par Tran (2014), seront ajoutées dans la ressource du LST des expressions polylexicales (nominales, verbales, adjectivales, prépositionnelle[14]). De plus, la ressource du LST pourra également bénéficier des

[14] Ces expressions, collocations et phrasèmes figés, ont été identifiées et validées par Agnès Tutin et Marie-Paule Jacques.

précédents travaux de Yan (2012), qui propose une description et une modélisation des constructions verbales typiques du genre en intégrant une composante sémantique dans cette description par l'identification des acceptions mobilisées et l'attribution de paradigmes sémantiques aux arguments verbaux.

Enfin, les travaux de Tutin sur le LST portent sur les différentes dimensions entrevues dans la présente partie. Elle s'intéresse notamment à la dimension phraséologique du LST (2007b, 2014), à sa dimension sémantique (2007c, 2008) ainsi qu'énonciative (2010, 2011). Les angles d'approches du LST sont ainsi multiples, de même que le niveau de caractérisation effectué sur le lexique. L'ensemble des travaux que nous venons d'évoquer a cependant pour point commun de faire appel aux données quantitatives pour l'identification du lexique, première étape des traitements.

1.2.2.3 LST et lexicométrie

Nous abordons dans cette section la dimension quantitative inhérente à toute tâche d'extraction lexicale automatique. Les données lexicométriques (fréquence absolue, relative, mesures d'association, etc.) constituent en effet le matériau premier pour l'identification de phénomènes lexicaux récurrents.

La grande majorité des études sur le lexique/vocabulaire de l'écrit scientifique se base sur des données de fréquence afin d'identifier les éléments lexicaux saillants dans le corpus de travail. Par éléments lexicaux, nous entendons tout token (ou regroupements de tokens) issu de la segmentation de l'analyseur et n'appartenant pas à la ponctuation. La fréquence dans le corpus d'analyse intervient alors de plusieurs manières dans les paramètres employés pour l'extraction de lexique. Dans un premier temps, le calcul d'occurrences donne accès aux fréquences brutes et relatives des unités lexicales. Ces fréquences sont ensuite intégrées dans des calculs de comparaison de fréquence entre corpus pour mesurer la sur-représentation d'un élément. Pour chiffrer une telle sur-représentation, différentes mesures sont utilisées dans le champ de l'extraction terminologique et de lexiques de l'écrit scientifique, notamment la mesure de spécificité (Drouin, 2004).

Cependant, Drouin (2004, p. 351) rappelle, suite à Labbé & Labbé (1994), que « la fiabilité du calcul des spécificités diminue lorsque la fréquence des événements

considérés est basse ». Il est donc nécessaire d'avoir à l'esprit les limites de ces métriques et de ne pas valider les résultats des extractions sans évaluation a posteriori. Bien que le LST ne soit pas, selon notre définition, un lexique terminologique (du fait de sa distribution transdisciplinaire), les méthodes utilisées pour identifier lexique scientifique et terminologique se rejoignent sur plusieurs points. Ces méthodes se composent généralement de cinq phases principales (Scott & Tribble, 2006, p. 58-60) :

1. Extraction des mots fréquents.
2. Fixation d'un seuil minimum de fréquence.
3. Comparaison entre le corpus d'analyse et le corpus de contraste (pour mesurer la sur-représentation des éléments lexicaux dans le premier corpus par rapport à leurs occurrences moindres dans le second corpus).
4. Filtrage des mots sous-représentés.
5. Ordonnancement selon la métrique choisie (spécificité, rapport de vraisemblance, khi², ratio, etc.).

Plusieurs paramètres influent ainsi sur les résultats de l'extraction (mis à part les outils de segmentation et d'analyse de corpus) : le choix des métriques et des seuils ainsi que la composition des corpus.

Dans une précédente étude (Hatier, 2013), nous avons comparé trois mesures statistiques pour évaluer leur apport dans l'extraction du LST. Le chi-carré, le rapport de vraisemblance et le ratio de fréquence ont été appliqués sur un corpus d'articles de recherche en SHS et ont ainsi généré trois listes de candidats au statut LST. Après observation des éléments en tête de ces listes, nous avons constaté qu'aucune mesure ne permet de valider automatiquement un mot extrait comme élément du LST. Paquot & Bestgen (2009) testent également plusieurs mesures et notent que celles qui sont habituellement utilisées pour extraire les mots caractéristiques d'un corpus (LLR[15], khi²) sont fondées sur des fréquences absolues, sans prise en compte de la variabilité interne des corpus. Ils remarquent que nombre de mots-clés extraits à l'aide du LLR sont dépendants du sujet ou de la discipline. Bien que l'intégration de métriques représentant la répartition améliore sensiblement les résultats, une validation manuelle reste cependant

[15] Log-Likelihood Ratio ou rapport de vraisemblance

nécessaire (Hatier, 2013). Ceci s'explique notamment par le fait que chaque mesure présente certains avantages et inconvénients dans l'identification des spécificités. On peut noter par exemple la propension du calcul du khi² à fournir trop de résultats significatifs (2012). Ainsi, une utilisation efficace de ces critères statistiques exige de prendre en compte le silence et le bruit pouvant en résulter.

Il est à noter que la phase manuelle, nécessaire pour valider les listes extraites d'après les différentes mesures, permet de traiter directement le bruit présent dans les résultats d'extractions. Pour le silence, il est alors nécessaire de « rattraper » les mots n'ayant pas passé le filtre des critères statistiques. La difficulté est alors supérieure puisqu'il faut dans un premier temps identifier les mots manquants : en générant des listes avec des seuils différents, en comparant les résultats à des listes déjà existantes (pour la même langue et le même ensemble disciplinaire de préférence) ou par l'ajout de mots en relation sémantique ou morphologique avec des mots déjà repérés[16]. Le rattrapage se fait également au cas par cas, lors d'applications (didactique, extraction de routines), en ajoutant des éléments nous paraissant intéressants et cependant absents de la ressource du LST.[17]

Il existe ainsi autant de paramétrages statistiques différents que d'études. Nombre d'entre elles incluent une étape de validation manuelle, car comme le remarque Pecman (2004b, p. 169) : « le critère de fréquence est le plus souvent invoqué et est à considérer comme un instrument de localisation [...] plutôt que comme un critère définitoire ». Puisque les éléments les plus fréquents sont souvent les plus cohérents (Biber, Conrad, & Cortes, 2004; Simpson-Vlach & Ellis, 2010), la fréquence est le premier indicateur utilisé, parmi d'autres, pour l'identification de ces lexiques particuliers.

1.2.2.4 Méthodes d'extraction des lexiques scientifiques

Au niveau des méthodes employées dans les divers travaux, nous pouvons constater que les perspectives d'utilisation (notamment, pour les applications

---

[16] Paquot (2010, p. 53-55) consacre ainsi une section pour ces mots, qu'elle identifie en tirant parti de liens morphologiques et sémantiques avec des mots automatiquement extraits. Ainsi, le nom *analysis* est « rattrapé » de par son lien avec le verbe *to analyse* lui automatiquement extrait.

[17] Nous prévoyons ainsi de rattraper certains mots et de les inclure dans la ressource du LST. Tel est notamment le cas des verbes *dire* et *voir* très fréquents dans le corpus mais non-spécifiques.

didactiques, les groupes à qui sont destinés les ressources, par exemple apprenants natifs ou non-natifs) déterminent généralement la composition des corpus d'analyse ainsi que le type d'analyse effectués sur le lexique.

Paquot (2010), dans la lignée de Rayson (2008), opte pour un modèle *data-driven*. En combinant les critères de répartition et de spécificité, elle extrait, pour l'anglais, l'*Academic Vocabulary*. Son travail se base sur un corpus académique de SHS et de sciences exactes, mêlant ouvrages et articles de recherche. L'ensemble lexical qu'elle décrit, à des fins didactiques, est donc un lexique partagé par un ensemble disciplinaire plus hétérogène que le nôtre, qui se focalise sur les SHS et le genre des articles de recherche.

Parmi les travaux se rapprochant le plus de nos objectifs, ceux de Drouin sur le lexique scientifique transdisciplinaire (extrait pour l'anglais et le français) nous intéressent pour plusieurs raisons. Ce sont, d'une part, les seuls travaux d'extraction, à notre connaissance, d'un lexique correspondant à notre définition du LST, pour le français, et ce au niveau des mots simples (et non pas de la phraséologie), pour les catégories adjectivale, adverbiale, nominale et verbale. Son corpus d'analyse, de 2 millions de mots, est composé de thèses, couvrant neuf disciplines : anthropologie, chimie, droit, géographie, histoire, informatique, ingénierie, physique et psychologie. L'ensemble transdisciplinaire représenté est donc plus diversifié que le domaine des SHS. D'autre part, Drouin, à la différence de nombreux travaux, intègre pour son analyse du LST un corpus de « référence », à des fins contrastives, ce qui lui permet d'identifier les éléments surreprésentés dans le corpus d'analyse. Drouin (2007) utilise précisément un corpus journalistique (une année de parution du *Monde*) et applique le calcul de spécificité de Lafon (1980). Cette mise en opposition de corpus, dont les genres diffèrent, nous paraît essentielle dans la mesure où nous avons défini le LST comme un lexique associé à un genre, l'écrit scientifique, par opposition aux autres genres dans lesquels il est moins représenté. Il nous paraît important ici de préciser que la spécificité est employée pour identifier les mots simples du LST. Dans l'optique d'une étude de la phraséologie transdisciplinaire, ce critère de sur-représentation devrait alors être appliqué au niveau des unités phraséologiques et non au niveau des éléments constitutifs de ces unités. Nous considérons ainsi l'utilisation d'un tel corpus de contraste comme essentielle dans l'identification de notre objet d'étude. Nous détaillerons dans le chapitre suivant les points que nous

retenons dans la méthode présentée par Drouin et nous dédierons la section 2.2.3.1 à une comparaison des résultats de notre méthode avec la liste du LST de Drouin. Parmi les autres études intégrant un corpus de contraste afin de mesurer la spécificité du LST, nous pouvons également citer Biber (2006) qui opte pour un corpus hétérogène (conversation, fiction, presse) comparable au corpus de contraste que nous présentons dans le chapitre suivant.

À l'instar de Drouin, Paquot (2010, p. 46-47) applique le calcul du rapport de vraisemblance par comparaison à un corpus de fiction, partant de l'hypothèse que le LST est particulièrement sous-représenté dans ce genre. À l'inverse, la plupart des autres travaux, tels ceux de Coxhead (2002), n'utilisent pas de corpus de contraste pour mesurer la spécificité (ou surreprésentation) du LST dans le genre de l'écrit scientifique mais font appel à des listes d'exclusion représentant la langue générale. Coxhead exclut ainsi de l'*AWL* les 2000 mots les plus fréquents en anglais donnés dans la *General Service List* de West (1953). Ce type de méthode présente toutefois deux inconvénients principaux. D'une part, des mots sur-représentés dans l'écrit scientifique sont ignorés s'ils sont présents dans la liste d'exclusion. D'autre part, des mots non spécifiques à ce genre d'écrits peuvent être faussement identifiés, étant donné qu'aucune comparaison de fréquence n'est effectuée pour les mots absents de la liste d'exclusion. Le tableau 1.1 résume ci-dessous les particularités de ces méthodes afin de mettre en évidence les éléments que nous retenons et d'identifier les limites de certaines ressources que nous voulons dépasser.

	Corpus d'analyse	**Référence**	**Critères**	**Analyse sémantique**	**Entrée**
Drouin ***LST*** *français, anglais,*	thèses : informatique, SHS	journalistique *Le Monde*	spécificité, répartition	liens entre acceptions Fr ↔ En	mot
Paquot ***Academic Vocabulary*** *anglais*	livres, articles : SHS, sciences dures Essais d'étudiants : SHS	fiction *BNC*	keyness répartition	étiquetage sémantique automatique	mot
Pecman ***LSG*** *français, anglais, serbe*	articles rapports comptes rendus : physique chimie biologie		fréquence répartition ré-exploitabilité (didactique)	typologie notionnelle, liens entre les langues	phrasème, concept
Phal ***VGOS*** *français*	manuels : mathématiques physique chimie sciences naturelles		fréquence répartition	repérage des sens	mot, lexie complexe, structure lexico-syntaxique

Tableau 1.1: Méthodes et ressources des travaux sur les lexiques scientifiques

Nous relevons de cette compilation des travaux antérieurs plusieurs points essentiels. Premièrement, au niveau des critères du LST correspondant à notre définition, la répartition et la spécificité (ou sur-représentation) nous semblent les plus importants. Bien que peu convoquée dans les travaux, l'utilisation d'un corpus de contraste représentant des genres différents de l'écrit scientifique nous paraît alors essentielle pour identifier cette spécificité. De la même façon, Gardner & Davies (2013), dans la perspective d'élaboration d'une liste du LST, insistent sur l'importance des corpus et recommandent ainsi l'utilisation d'un corpus de contraste contemporain, hétérogène et de grande taille.

Deuxièmement, nous constatons qu'aucune étude ne nous paraît susceptible de répondre aux objectifs que nous avons définis, i.e. se concentrant sur les mots simples en français, du lexique mobilisé dans les articles de recherche, dans le domaine des SHS. De plus, aucune étude ne se base sur des corpus analysés

syntaxiquement. Or, les informations sur les relations de dépendance sont essentielles pour caractériser le LST au niveau sémantique et syntaxique, caractérisation nécessaire en vue des applications définies section 1.1. Un des apports importants de notre méthode dans l'étude du LST est alors la prise en compte de ses propriétés lexico-syntaxiques, subordonnées à une analyse en dépendance préalable.

1.2.3 Propriétés du LST

La revue des travaux du domaine nous montre que la définition du LST, sous toutes ses dénominations, les méthodes d'identification du LST ainsi que les résultats des processus d'extraction, varient principalement selon les corpus d'analyse : articles de recherche scientifiques de différents domaines, thèses (Drouin, 2007), manuels scolaires ou universitaires (Phal & Beis, 1972), combinaison de différents types de textes, etc. De la même manière, le choix du genre de textes analysés (articles en SHS) et les objectifs d'utilisation du LST ont influé sur la définition de notre méthodologie d'extraction et de caractérisation du LST.

Ainsi, nous avons montré l'importance de la prise en compte des contextes d'apparition du LST, en mettant en évidence l'apport du contexte pour les applications en indexation terminologique ou pour les perspectives didactiques. De plus, si nous voulons caractériser finement le LST, nous devons dépasser le simple niveau lexical et prendre en compte les propriétés syntaxiques et sémantiques des éléments de ce lexique. Ces propriétés, pour qu'elles soient représentatives de l'usage du LST dans le genre qui nous intéresse, doivent alors être dégagées d'un corpus d'articles de recherche, et ce pour une bonne adéquation de notre ressource aux objectifs annoncés.

Parmi ces propriétés, la spécificité et la transdisciplinarité sont les deux principales, tant au niveau de la définition du LST que de son extraction. Nous détaillons, dans les deux sections suivantes, ces deux propriétés et précisons leur rôle dans les différentes étapes de caractérisation du LST.

1.2.3.1 Spécificité

Le LST est constitué d'éléments dont l'emploi est spécifique dans le cadre de l'écrit scientifique. La spécificité se traduit tant au niveau lexicométrique (nombre d'occurrences des unités lexicales) qu'au niveau sémantique (acception particulière mobilisée) ou syntaxique (collocation spécifique au genre telle *confirmer une hypothèse*).

Afin d'identifier les éléments lexicaux spécifiques au corpus scientifique, nous utilisons un corpus de contraste (présenté section 2.2.1.2) faisant office de « référence ». Ainsi, nous rejoignons Lerat (1997, p. 6) qui précise que, dans l'optique de l'étude des langues spécialisées, « mérite d'être pris en considération tout ce qui n'est pas prédictible à partir soit d'une compétence linguistique générale soit des connaissances de sens commun », ce que nous traduisons ici par le critère de spécificité, le corpus de contraste représentant alors la compétence linguistique générale.

À l'instar de Drouin (2004, p. 345), nous utilisons alors deux corpus combinés pour « mettre en opposition le comportement des unités lexicales de corpus de niveaux de spécialisation différents ». Un mot appartenant potentiellement au LST doit donc être plus fréquent dans un corpus de genre scientifique que dans un corpus représentant d'autres genres. Les mots fréquents dans les autres genres ne sont pas automatiquement éliminés, à l'inverse de Gardner & Davies (2013) qui ne retiennent pas dans leur inventaire des mots pourtant spécifiques. La sur-représentation d'un mot dans le genre scientifique n'implique pourtant pas qu'il ne soit pas mobilisé fréquemment dans un autre genre, tel le genre journalistique ou la fiction. Le critère de spécificité nous permet donc ici de contraster les genres pour faire émerger un lexique associé à une pratique et à un type de texte particuliers.

1.2.3.2 Transdisciplinarité

Nous avons également défini le LST comme un lexique traversant, couvrant l'ensemble des disciplines des SHS présentes dans notre corpus d'analyse (dont la composition est détaillée section 2.2.1.1). Cette propriété de transdisciplinarité est ainsi au coeur de l'ensemble des travaux sur des lexiques similaires, tels que le

Vocabulaire Savant de Base défini par Da Sylva (2010) comme non spécifique à un domaine de connaissance. Ce critère de transdisciplinarité sous-tend ainsi l'existence d'éléments lexicaux communs à un groupe de disciplines, ne relevant pas de la terminologie.

Granger & Paquot (2009a), posant la question de l'existence d'un LST réellement partagé, repèrent plus de 100 verbes communs à trois sous-domaines académiques. Pour Tutin (2007a), les propriétés de ce lexique non terminologique sont davantage corrélées au sous-genre d'écrit (article, manuel) qu'à la discipline. Tutin remarque cependant des différences intra-disciplinaires, notamment sur le type de langage évaluatif (2010) ou les verbes de positionnement (2011). Blumenthal (2007), en comparant un corpus de sciences exactes et un corpus de SHS, observe cependant une intersection importante en ce qui concerne les éléments lexicaux les plus fréquents, spécifiquement pour ceux appartenant au lexique abstrait général. Il note néanmoins que les sens mobilisés dans ces deux ensembles disciplinaires ne se recouvrent pas entièrement pour une même unité lexicale. Ainsi le nom *théorie* mobilise deux sens différents en SHS ou en sciences exactes, comme peuvent le révéler ses cooccurrents : *quantique/nombre/relativité* en sciences exactes, *pratique/élaborer/développer* en SHS[18]). L'acception d'un élément peut donc varier d'un groupe de disciplines à un autre comme le confirme Hyland (2008b). Granger & Paquot (2009a) font la même analyse et remarquent pour l'anglais que le verbe *to analyse* réfère à la décomposition d'une substance en ingénierie alors qu'en SHS, il dénote une étude précise. Nous devons ainsi nous assurer que les éléments du LST que nous identifions ont bien une acception transdisciplinaire stable, en appliquant le critère de répartition tant pour les occurrences que pour les autres propriétés que nous retenons, notamment les relations lexico-syntaxiques.

En nous basant sur un corpus d'articles en SHS (pour être en adéquation avec les objectifs définis en introduction de ce chapitre), nous nous assurons d'une certaine transdisciplinarité dans le lexique employé bien que les disciplines aient une influence sur celui-ci. Ainsi, Hyland & Tse (2007) notent que certains mots de l'*AWL* sont spécifiques à un domaine. La distribution des occurrences dans les

[18] En observant la combinatoire de *théorie* via le *Lexicoscope* (voir section 1.4), nous relevons parmi les cooccurrents significatifs de *théorie* (dépassant le seuil de 10.83 du log-likelihood) : *anticipation, pragmatique, économique, unifier, développer, proposer*.

disciplines, critère vérifiant la transdisciplinarité, permet cependant d'exclure ces éléments lexicaux propres à une minorité de disciplines et/ou d'articles. Ainsi, le nom *chômage*, ayant 485 occurrences dans notre corpus, est principalement convoqué dans trois disciplines : économie (337), sociologie (67) et sciences politiques (37). La transdisciplinarité opère alors, selon Simpson-Vlach & Ellis (2010), un rôle de filtrage des phénomènes idiosyncrasiques. L'information sur la répartition des occurrences, comme toute mesure de fréquence, présente néanmoins comme limite majeure le fait d'être basée sur le lemme et nom l'acception. Ainsi, un mot peut renvoyer à plusieurs acceptions, tant dans le corpus d'analyse que dans le corpus de contraste. Les occurrences n'étant pas désambiguïsées, nous nous situons dans un premier temps au niveau lexical, avant d'aborder le niveau sémantique au chapitre 3.

Après avoir circonscrit notre objet d'étude et l'avoir caractérisé par ses fonctions et propriétés, nous nous intéressons maintenant au cadre dans lequel il est mobilisé, le genre de l'écrit scientifique.

1.3 LST et genre de l'écrit scientifique

Nous nous situons, pour l'étude du LST, dans la description d'un lexique inscrit dans un des genres de l'écrit scientifique, celui de l'article de recherche en SHS. Plusieurs travaux se sont également intéressés à un ensemble lexical associé au genre de l'écrit scientifique et ont tenté d'en donner une définition.

1.3.1 Inscription du LST au sein du genre

Paquot (2010, p. 4) définit ainsi l'*Academic Vocabulary* comme un ensemble qui réfère « aux activités caractérisant le travail académique, organisant le discours scientifique et permettant la construction rhétorique des écrits académiques[19] ». Cette fonction structurante, métadiscursive, du LST est également soulignée par Paquot & Bestgen (2009), pour qui ces mots spécifiques de l'écrit académique fournissent l'ossature sémantico-pragmatique du texte et sont utilisés pour référer aux idées abstraites, aux processus, à la recherche et l'évaluation. Un certain nombre d'études du LST adopte ainsi une approche fonctionnelle de ce lexique, en

[19] « a set of options to refer to those activities that characterize academic work, organize scientific discourse, and build the rhetoric of academic texts ». Notre traduction.

mettant en avant la fonction métadiscursive du LST, défini par sa présence dans un groupe de textes homogènes.

D'autres travaux se distinguent de cette approche en s'attachant à une définition du LST qui prend en compte le contexte de production des écrits. Ainsi, en définissant le LST comme un lexique de genre, Tutin (2007a) envisage ce lexique comme élément d'un genre codifié par une communauté de discours (Swales, 1990). Cette communauté scientifique mobilise alors un lexique particulier dans une pratique commune : la construction et la communication du savoir scientifique par les articles de recherche. La pratique est, de façon similaire, centrale chez Pecman (2004b, p. 128), qui définit la *Langue Scientifique Générale* (LSG) comme « une pratique langagière spécifique à une communauté de discours composée de chercheurs dont les objectifs communicatifs poursuivis émanent des préoccupations partagées par des scientifiques à travers le monde et indépendamment de leurs spécificités disciplinaires. »

Nous considérons le LST comme le lexique du genre de l'article de recherche en SHS, genre défini par une communauté linguistique, des pratiques sociales et des productions textuelles. Le LST doit notamment être envisagé au niveau de la pratique sociale de construction de la connaissance. Cette construction s'accompagne de pratiques et de discours particuliers qui se manifestent dans les articles notamment par les interactions intertextuelles, au travers des citations, des références, des positionnements, exprimés par le LST. Le LST n'est ainsi pas un lexique de spécialité : son caractère transdisciplinaire transcende les domaines restreints des langues de spécialité. Il est mobilisé dans un usage spécialisé de langue mais n'en constitue pas à lui seul le lexique. Le LST fait ainsi partie de la langue de la linguistique, au même titre que la terminologie de ce domaine. Il peut alors être considéré comme un lexique mobilisé dans un usage spécialisé de la langue, au sein d'un genre déterminé.

Afin de clarifier la notion de genre, nous reprenons la définition du genre donné par Poudat (2006, p. 380) dans ses travaux sur l'article de recherche en linguistique :

> « Niveau normatif de contraintes et de prescriptions positives ou négatives régulant la production et l'interprétation d'un texte. Puisque tout texte relève d'un genre, et que tout genre relève d'un discours, le genre permet de relier les textes aux discours. »

Concernant notre travail, la notion de genre se justifie en premier lieu par l'homogénéité des textes compilés dans notre corpus d'analyse (présenté section 2.2.1.1), représentant un sous-genre particulier du discours scientifique (qui comprend aussi d'autres genres tels les présentations de conférence, les ouvrages, etc.). L'homogénéité des textes s'explique alors par les objectifs communicatifs partagés (Swales, 1990) qui contraignent la structure et le contenu des textes. Swales a ainsi mis en évidence le caractère récurrent, associé au genre de l'article, en particulier le modèle *CARS* (« Create a Research Space ») associé aux introductions, qui font appel à des procédés rhétoriques récurrents.

Le concept de genre permet ici de s'interroger sur la caractérisation lexicale de l'écrit et du discours scientifique, et sur les contraintes exercées sur le LST, lexique remarquable par sa surreprésentation vis-à-vis d'autres genres et par sa distribution traversant les disciplines.

1.3.2 LST, genre et langue de spécialité

Le LST, en tant que lexique associé à un discours spécifique, est également rattaché, dans certains travaux, au concept de langue de spécialité, au sens de Kocourek (1991, p. 24-42), cité par Pouliot (2012)« sous-ensemble de la langue défini à partir de critères du domaine où cette variété est employée pour communiquer un contenu spécialisé à l'aide de ressources linguistiques comprenant notamment des unités lexicales au sens précis ». Bien que le LST soit convoqué dans un usage spécialisé de la langue, il nous apparaît paradoxal de l'associer à un domaine limité dans la mesure où il est défini comme transdisciplinaire.

Pecman (2007, p. 80), dans son étude de la LSG, utilise également le terme de langue de spécialité dans le sens de « sous-système linguistique correspondant à l'emploi de la langue dans un domaine de connaissance unique ». Pecman précise néanmoins que la LSG est une « forme particulière de langue de spécialité », la rattachant à la notion de communauté de discours dans la mesure où l'ensemble des scripteurs/lecteurs partagent un ou plusieurs buts notoires. La langue de spécialité est donc envisagée dans le cadre du discours de spécialité, au côté d'autres systèmes sémiotiques (tels les schémas, figures), comme le note Kocourek (1991). Cependant, la notion de langue de spécialité, définie par

opposition à celle de langue générale, pose un problème définitoire similaire au LST. Comme nous l'avons mentionné dans la section 1.2.1, la première difficulté dans l'analyse du LST est sa délimitation, ce lexique étant défini à l'intérieur d'un continuum de lexiques mobilisés dans l'écrit scientifique, comprenant notamment le lexique de la langue générale et la terminologie. Cette difficulté de délimitation se retrouve au niveau de la langue de spécialité, dont les frontières floues en font un « univers langagier [...] à cheval entre la langue générale et les langues de spécialité d'orientation scientifique », selon Pecman (2007, p. 91), au même titre que le LST se situant entre lexique de la langue générale et terminologie.

Cependant, il paraît complexe de segmenter la langue en autant de langues de spécialité que de domaines décrits. En ce qui concerne les travaux ici présentés, le genre de l'écrit scientifique prend forme à travers la définition même de notre corpus d'analyse. Ainsi, l'ajout ou la suppression de certaines disciplines (des SHS ou des sciences exactes) impliquerait des variations dans les résultats de l'extraction du lexique que nous étudions, comme le montre l'ensemble lexical renvoyant aux objets d'étude des SHS, que nous abordons dans la section 2.2.2.2. Les textes retenus sont alors, selon Beacco (2004, p. 111), considérés comme l'actualisation du genre étudié. La catégorie du genre tend ainsi à être « définie de façon extensionnelle plutôt qu'intensionnelle ». Autrement dit, il nous faut préciser le fait que les spécificités identifiées ne prouvent aucunement l'existence d'un lexique de genre, puisque tout regroupement de textes mène inévitablement à l'observation de points communs (et de différences). L'existence, parmi les genres de l'écrit scientifique, d'un genre de l'article de recherche en SHS, s'exprimant par des pratiques sociales déterminées, un lexique particulier, est donc posée comme axiome, préalablement à la définition de ses attributs et en particulier du LST.

Nous devons donc envisager l'étude du LST comme l'étude d'un sous-ensemble intégré à la langue générale (avec laquelle il partage de nombreuses propriétés en commun) mais également comme l'étude d'un système « distinct », dont certaines propriétés linguistiques sont spécifiques par rapport à la langue générale. Ces considérations rejoignent la notion de *sublanguage* (Kittredge & Lehrberger, 1982; Harris *et al.*, 1989), sous-langage correspondant à un usage spécifique de la langue, identifiable notamment par l'analyse des propriétés distributionnelles et syntaxiques. Cette notion de sous-ensemble est également présente dans la définition des langues de spécialité : « sous-système linguistique

tel qu'il rassemble les spécificités linguistiques d'un domaine particulier » (Dubois *et al.*, 2001, p. 440) .

Ainsi, nous définissons, et étudions, le genre de l'écrit scientifique à travers ces particularités linguistiques, des « gradients de typicalité [...] des faisceaux de régularités », pour reprendre les mots d'Adam (1999, p. 93-94). Le LST, en tant que lexique de genre, est étudié selon ce principe de propriétés spécifiques à un genre, tels certains marqueurs mis en évidence par Beacco (2004). De la même manière que l'expression *il était une fois* est un marqueur, un trait prototypique, du genre du conte, un patron du type — *Cet article/ouvrage/étude/travail s'intéresse/aborde le sujet/problème/problématique* — sera considéré comme marqueur du genre de l'écrit scientifique s'il est suffisamment fréquent et réparti dans notre corpus.

Ce sont ces propriétés, de fréquence et de répartition dans le corpus, qui permettent de faire émerger le LST, mobilisé par des scripteurs du genre de l'article de recherche, ou genre du « discours scientifique spécialisé » selon la typologie de Desmet (2006).

1.3.3 Communautés et pratiques discursives

Comme précisé ci-dessus, nous abordons ce lexique, à la suite de Tutin (2007a), comme un lexique de genre, codifié par une communauté de discours. Cette notion, définie sur des bases socio-rhétoriques renvoie à un groupe d'individus partageant un genre d'expression spécifique. Selon Swales (1990, p. 24), les communautés de discours possèdent six caractéristiques principales :

- les membres partagent un ensemble d'objectifs communs ;
- il existe des mécanismes d'intercommunication entre les membres ;
- le but de l'intercommunication est l'échange d'informations ;
- la communauté possède et utilise plusieurs genres ;
- la communauté partage un lexique spécifique ;
- la communauté a un seuil minimal de membres possédant un degré d'expertise approprié.

La communauté des chercheurs, inscrits dans l'activité de publication scientifique, correspond donc à la définition donnée par Swales puisqu'elle intègre effectivement l'ensemble de ces critères, à savoir l'objectif commun de communication du savoir au travers de plusieurs genres (dont l'un est l'article de recherche), et un lexique spécifique partagé, ou en d'autres termes, scientifique transdisciplinaire.

Le genre, en particulier le genre scientifique, peut ainsi être défini par la communauté dans laquelle il s'inscrit, ou le contexte dans lequel il apparaît. Ainsi, Biber (1993a), dans son étude des écrits académiques, définit le genre dans une perspective situationnelle. Le genre et/ou la communauté discursive sont alors essentiellement définis au travers des textes produits et de leur contexte de production. Desmet (2006) envisage également l'écrit scientifique sous l'angle situationnel, en tant que variation de la langue générale s'actualisant en discours. Elle identifie alors différents facteurs de variation : le récepteur (les chercheurs), les normes (via les relectures), la structure textuelle et le but (convaincre de l'intérêt de ses travaux). Bazerman (1988) opte pour sa part pour une approche socio-rhétorique des textes scientifiques[20], et s'intéresse aux stratégies persuasives dans ces écrits. Teufel (1998) insiste sur les objectifs communicatifs de l'article et étudie également les procédés rhétoriques à l'œuvre dans l'article de recherche, dont le but communicationnel est de présenter, raconter et de se référer aux résultats de recherches spécifiques. Elle constate ainsi que les raisonnements à propos des problèmes, des traitements et des solutions suivent des patrons prédictibles. Les dimensions inter-textuelle, discursive et rhétorique sont ainsi très présentes dans le genre de l'article de recherche. En conséquence, le LST est fréquemment mobilisé pour remplir ces fonctions particulières : citation par rapport à un texte, modalisation du discours, argumentation, etc.

Ces définitions diverses du genre de l'écrit scientifique nous mènent à deux constats. D'une part, le LST, par son rôle dans la construction du savoir, est un objet d'étude permettant d'interroger les aspects discursifs, rhétoriques, sociologiques, épistémologiques et lexicaux dans l'exposé de l'activité scientifique. D'autre part, ces différentes approches de l'écrit scientifique ont mis en évidence des universaux et des spécificités disciplinaires, tant au niveau de la structure

[20] « Scientific texts can be understood as part of the social process of organizing scientists and creating science » (Bazerman, 2011, p. 14)

textuelle des articles qu'au niveau des procédés rhétoriques et des éléments lexicaux mobilisés. La partie suivante aborde ainsi ces phénomènes de variabilité et d'invariabilité entre les disciplines.

1.3.4 Variabilité et invariabilité du lexique dans les sciences

Comme la revue des travaux antérieurs l'a mis en évidence, les ensembles disciplinaires au fondement des études sur le LST sont variés : sciences humaines et sociales, sciences exactes, combinées ou séparées. La constitution de corpus couvrant plusieurs disciplines part de l'hypothèse que certains phénomènes linguistiques traversent ces disciplines. Il n'en existe pas moins des variations disciplinaires, des spécificités selon les domaines.

1.3.4.1 Divergences disciplinaires

Cette diversité a, pour Fløttum & Rastier (2003), une base sociologique. Bazerman (2011) note quant à lui que les disciplines ont différentes épistémologies, stratégies, procédures, et qu'elles s'attachent à des objets divers. Or, les manières de dire dans les disciplines sont indissociables de la façon de faire la recherche, ainsi que l'observe Rinck (2010). Ainsi, même si le genre que nous étudions est notamment caractérisable par des éléments lexicaux transdisciplinaires, des spécificités disciplinaires subsistent au niveau des lexiques mobilisés. Ces spécificités sont par ailleurs davantage liées aux communautés disciplinaires qu'aux communautés linguistiques (Fløttum *et al.*, 2006, p. 15).

Ainsi, en analysant des corpus disciplinaires, Tutin (2011) relève plusieurs de ces variations : en psychologie, l'accent est davantage mis sur les hypothèses et les résultats ; en sciences de l'éducation, les intentions du chercheur, ses opinions et son questionnement sont particulièrement présents ; la linguistique se caractérise par la présence de nombreux verbes dénotant l'apport scientifique, les intentions et opinions. Ces différences disciplinaires se manifestent au niveau lexical, énonciatif, rhétorique et également textuel.

En ce qui concerne la structure textuelle, Grossmann (2010) insiste sur l'hétérogénéité présente dans les sciences, en prenant pour exemple les différences existant entre un article d'ethnologie sous la forme de récit et un article de

physique au format IMRAD[21]. Cependant, l'existence de telles structures textuelles traversant les disciplines est le signe, selon Teufel (1998), que certaines d'entre elles partagent des méthodologies et méthodes d'évaluation et ont ainsi plus ou moins trouvé une définition commune de comment faire de la recherche.

1.3.4.2 Convergences disciplinaires

D'un autre côté, on peut donc affirmer qu'il existe une certaine homogénéité au sein des SHS, renforcée par les collaborations et les influences interdisciplinaires. Il est d'ailleurs notable que les SHS constituent une famille de disciplines reconnue par diverses institutions (CNRS, Universités, éditeurs).

Grossmann (2010, p. 414) rappelle ainsi certaines caractéristiques, traversant les disciplines, du discours des chercheurs, « en particulier [la] forte dimension intertextuelle et [la] double contrainte d'effacement énonciatif[22] et de construction d'un point de vue de l'auteur. » Ces ressemblances entre disciplines confortent ainsi la pertinence d'une étude sur ce genre de l'écrit scientifique en SHS, qui n'a pas pour unique but de véhiculer la connaissance mais porte également, dans toutes les disciplines, l'image du chercheur, du public et de la recherche (Fløttum & Rastier, 2003).

Rinck (2010), en proposant une analyse textuelle des articles, identifie des propriétés transdisciplinaires, notamment l'importance de l'introduction pour asseoir la légitimité du scripteur ou celles des sections résultats, discussion et conclusion qui permettent d'interroger les épistémologies en jeu. Tutin (2011) met également en évidence des propriétés traversant les disciplines, en observant la présence importante du positionnement de l'auteur dans les conclusions. Les différents niveaux d'analyse (lexical, pragmatique, textuel) permettent ainsi de dégager un ensemble de propriétés partagées par les disciplines des SHS. Le fait de se circonscrire à ce champ scientifique nous garantit alors une certaine proximité entre les sous-corpus disciplinaires, proximité confirmée par la présence du lexique des objets d'étude des SHS dans l'ensemble de ces sous-corpus (voir section 2.2.2.1).

---

[21] IMRAD (*introduction, methods, results, and discussion*) : renvoie à la structure textuelle dans un article de recherche des sciences expérimentales.

[22] Rinck (2010) et Tutin (2011) ont également étudié et caractérisé cet effacement énonciatif caractéristique des écrits en SHS.

Nous cherchons ainsi des « stabilités » au niveau lexical, présentes dans l'ensemble des disciplines composant notre corpus, à la manière de Münchow qui rappelle les bases de l'analyse en linguistique de discours comparative :

> Il s'agit de comparer différentes cultures discursives par l'intermédiaire des productions verbales qui en relèvent […] en mettant en rapport les manifestations d'un même genre discursif dans au moins deux communautés ethnolinguistiques différentes, genre dont il s'agit alors de décrire les invariabilités (ou "stabilités") et les variabilités (ou "instabilités"). (Münchow, 2007, p. 109)

Les articles de recherche de SHS peuvent être considérés comme relevant d'un même genre discursif qui s'actualise dans des disciplines différentes et qui présente des « stabilités » lexicales. Notre analyse du LST dans les SHS part donc de l'hypothèse que les proximités dans les méthodologies employées, dans la manière de construire le savoir, se manifestent dans le langage, et plus particulièrement dans le lexique. L'homogénéité des articles de recherche des différentes disciplines peut également être renforcée par l'éditorialisation dans les revues à comité de lecture, au travers de la sélection et de la correction des soumissions (Rinck, 2006, p. 97). La spécification de sous-genre (tel l'article de recherche expérimental) ou la création de sous-ensembles disciplinaires peut toutefois s'avérer nécessaire pour garantir une certaine cohérence entre les textes.

Nous avons ici défini le LST en tant que lexique de genre et avons ensuite circonscrit le genre concerné. D'autres travaux abordent le discours scientifique en tant que langue de spécialité et proposent alors une analyse du LST en tant que lexique de spécialité. Nous présentons dans la section suivante ces travaux sur les langues spécialisées en nous intéressant aux propriétés qui leur sont attribuées, dans la mesure où ces propriétés s'appliquent au LST et peuvent nous guider dans la mise en place de notre méthodologie d'extraction et de caractérisation du LST.

1.3.5 Propriétés des lexiques scientifiques

Les lexiques de l'écrit scientifique se voient attribuer dans la littérature certaines propriétés qu'il nous paraît important de mettre en évidence avant de proposer une méthode adaptée et d'élaborer une ressource capable de représenter ces propriétés. Bien que distincts de la terminologie (qui a une portée seulement disciplinaire), les lexiques de l'écrit scientifique (dont le LST) partagent plusieurs propriétés avec celle-ci : haute fréquence, sur-représentativité, domaine

sémantique couvert relativement restreint (à une ou des disciplines) et majoritairement abstrait. La section 1.3.5.1 se concentre sur ces points communs et différences entre terminologie et LST, particulièrement au niveau sémantique. La section 1.3.5.2 aborde le niveau lexico-syntaxique du LST, partant de l'observation de Lerat (1997) qui note que la spécificité des langues de spécialité est essentiellement lexico-syntaxique, au niveau des cooccurrences et des patrons récurrents. En conclusion de cette partie, la section 1.3.5.3 détaille l'intérêt de ces patrons dans le cadre de l'analyse du LST.

1.3.5.1 Propriétés sémantiques du LST

Le LST partage, comme nous l'avons vu, certaines propriétés avec la terminologie, en tant que lexique mobilisé dans certains usages spécialisés de la langue. Cependant, à la différence de la terminologie dont la dimension principale est conceptuelle, le LST assure aussi un rôle argumentatif, métadiscursif et métatextuel dans les écrits scientifiques.

Le LST et la terminologie présentent des similitudes que l'on retrouve au niveau des méthodes employées pour l'extraction et la caractérisation du LST ou d'éléments terminologiques : utilisation des données de fréquence, comparaison de la fréquence dans un ensemble de textes par rapport à un autre ensemble de textes, prise en compte de la distribution de l'unité lexicale. De plus, les informations sur les contextes d'apparition des unités lexicales (les profils combinatoires que nous présentons section 1.3.5.2) sont également utilisées pour la caractérisation de la terminologie, en tant que « données terminologiques », comme le note L'Homme (2004, p. 38). Le parallèle effectué entre les méthodes est révélateur des points communs entre la lexicologie et la lexicographie d'une part, et la terminologie et la terminographie d'autre part.

Cependant, malgré les nombreux points communs entre LST et terminologie, ces deux lexiques spécifiques (d'un genre et d'un domaine) sont des ensembles aux propriétés sémantiques différentes. Ainsi que nous le détaillons par la suite, le LST est un lexique largement abstrait, potentiellement polysémique, associé à un large ensemble disciplinaire tandis que la terminologie a pour propriété de renvoyer à un « sens spécialisé » d'un « domaine de spécialité » (L'Homme, 2005, p. 1125). Biber (2006), en s'appuyant sur un corpus spécialisé, remarque effectivement que

le discours scientifique est caractérisé au niveau lexical par une présence importante de noms abstraits, dénotant notamment les processus cognitifs. De la même façon, Paquot (2010), en procédant à une analyse sémantique, note que 87 % des éléments de ce lexique appartiennent à la catégorie « termes génériques et abstraits ». Rappelons cependant que le LST n'est pas uniquement nominal et couvre également les catégories adjectivale, adverbiale et verbale alors que la terminologie « classique » est essentiellement nominale, bien que certains chercheurs en proposent une conception plus large (L'Homme, 2005).

Afin de mettre en évidence les deux principales divergences entre le LST et la terminologie, nous reprenons la définition du terme donnée par Dubois *et al.* (2001, p. 480) : « unité signifiante d'un mot (terme simple) ou de plusieurs mots (terme complexe), qui désigne une notion de façon univoque à l'intérieur d'un domaine. »

Le premier point se situe au niveau de l'univocité du terme qui s'oppose à la polysémie des éléments du LST, comme nous le verrons dans la section 3.3.1.2. Cette distinction reste opératoire dans le cadre de l'écrit scientifique, comme le rappellent Bertels & Geeraerts (2012) qui ajoutent que la terminologie présente dans l'écrit scientifique a pour but l'univocité, en privilégiant les éléments monosémiques. Cependant, les observations de Kayser (1995) nuancent cette remarque lorsqu'il note que les mots et les termes font face à des problèmes comparables d'ambiguïté. Par ailleurs, la polysémie est de toute façon présente pour les termes lorsque l'on se situe dans le contexte d'une analyse automatique. Ainsi, dans un même corpus, une unité lexicale peut renvoyer à une acception terminologique ou à une acception non technique, tel que l'a observé Drouin (2004) dans son étude sur la détection automatique de termes. Le LST, en tant que lexique spécifique, associé au genre de l'écrit scientifique, mais non terminologique, est utilisé pour communiquer entre spécialistes avec précision. Cependant, ce lexique n'évite pas les problèmes de la polysémie. Ainsi, les noms *conclusion, définition*, *figure* ou *indice* renvoient à plusieurs acceptions relevant du LST. La désambiguïsation[23] est alors nécessaire pour identifier les différentes acceptions convoquées, pour passer du niveau lexical au niveau sémantique.

[23] La désambiguïsation et les expressions polylexicales font ainsi partie des problèmes centraux en TAL (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002).

D'autre part, le LST n'est pas restreint à un domaine de connaissance aussi particulier que la terminologie, ainsi que le précise la définition du terme donnée par L'Homme (2004, p. 22) : « unités lexicales dont le sens est envisagé par rapport à un domaine de spécialité, c'est-à-dire un domaine de la connaissance humaine ». La dénomination même de lexique transdisciplinaire doit nous permettre de distinguer les éléments du LST des termes, en ne retenant que les éléments traversant le corpus. Cependant, comme les exemples donnés en fin de paragraphe suivant le montrent, nombreux sont les éléments pouvant avoir pour certaines occurrences un statut terminologique et pour d'autres, renvoyer à une notion relevant du LST.

Le LST se situe ainsi entre les lexiques de la langue générale et de la terminologie, tant en ce qui concerne la polysémie que le domaine de connaissance associé. Meyer & Mackintosh notent que « lorsqu'il est repris par la langue générale, un terme adopte un sens plus large que lorsqu'il est confiné à un domaine spécialisé » (2000, p. 199), et nomment ce processus déterminologisation. Ce processus est également à l'œuvre lorsqu'un terme est utilisé dans un discours spécialisé en tant qu'élément du LST. De même, des éléments du LST peuvent être mobilisés dans d'autres genres que l'écrit scientifique, et adoptent alors un sens plus large, non restreint au domaine scientifique. Ainsi en est-il des noms du LST suivants, qui ont tous une ou plusieurs entrées terminologiques dans le *Grand Dictionnaire Terminologique*[24] et sont également mobilisés dans d'autres genres (littéraire et journalistique) : *réseau, schéma, échantillon, commentaire, analyse, paradigme*.

La présence attestée de la polysémie, et le caractère abstrait du lexique, nous orientent alors vers une méthodologie combinant méthodes automatiques et analyse manuelle (pour valider les informations potentiellement ambiguës). Pour identifier ces différentes propriétés sémantiques du LST, nous nous basons sur le corpus en tirant parti de la combinatoire des éléments de ce lexique.

1.3.5.2 Profils combinatoires et propriétés lexico-syntaxiques

Notre intérêt pour les propriétés lexico-syntaxiques des éléments du LST a pour origine plusieurs observations. Premièrement, cette dimension est

[24] http://www.granddictionnaire.com/index.aspx [consulté le 17/09/2016]

pleinement adaptée à l'étude d'un lexique de genre tel que le LST. Ainsi, Lerat (1997) note que les dimensions lexicale et syntaxique sont essentielles dans l'étude des langues spécialisées. Il ajoute que « la meilleure conception d'ensemble est celle qui lexicalise au maximum la syntaxe et qui en matière conceptuelle, tire le plus parti du jeu des relations typiques entre les mots. » (1997, p. 2) L'ensemble de nos traitements part de ce constat sur l'importance de la combinatoire. D'une part, notre analyse des patrons verbaux du LST, présentée section 5.2, intègre des relations de dépendance et des cooccurrents préférentiels, conformément au principe de lexicalisation de la syntaxe. D'autre part, nous nous basons, pour l'élaboration de notre typologie sémantique du LST, sur les propriétés lexico-syntaxiques de ses éléments, reprenant ainsi le principe d'une étude des concepts fondée sur les relations lexicales.

Le fait de travailler sur des corpus analysés syntaxiquement nous permet ainsi d'avoir accès aux profils combinatoires[25] des éléments lexicaux que nous étudions. Les profils combinatoires sont, selon Blumenthal (2008, p. 38), « l'image que donne du comportement d'un mot de base l'ensemble de ses collocatifs ». Nous pouvons ainsi caractériser un élément du LST par sa combinatoire : cooccurrents significatifs, patrons lexico-syntaxiques dans lesquels il s'intègre préférentiellement, classes sémantiques avec lesquelles il entre en cooccurrence. D'une manière beaucoup plus formalisée et systématique, la caractérisation du lexique par sa combinatoire est également à la base de la *Lexicologie Explicative et Combinatoire* (Mel'čuk, Clas, & Polguère, 1995), qui propose d'associer à chaque entrée une définition analytique ainsi qu'une description de la combinatoire syntaxique et lexicale.

1.3.5.3 Des propriétés aux constructions et patrons lexico-syntaxiques

Comme nous le verrons dans le chapitre 5, le fait de structurer le LST en classes sémantiques permet l'identification de motifs incorporant des contraintes syntaxiques, lexicales et sémantiques. Ce type d'association privilégiée est d'ailleurs mis en exergue par Lerat (1997, p. 4) qui note qu'un « intérêt particulier de la syntaxe des langues spécialisées est qu'elle a une grande pertinence dès lors

[25] Les profils combinatoires sont, selon Blumenthal (2008, p. 38), « l'image que donne du comportement d'un mot de base l'ensemble de ses collocatifs ».

que la construction du verbe, du nom, ou de l'adjectif, impose un type de complément spécifique ». Un autre aspect syntaxique important des langues spécialisées se trouve également au niveau des alternances, des nominalisations et des constructions à la voix impersonnelle. Ce style abstrait est d'ailleurs également remarquable au niveau du LST. Biber (2006, p. 64) a souligné, en tant que spécificité de l'écrit scientifique, la surreprésentation de phrases à la voix passive (20 % contre 5 % dans les autres genres). Biber note, de plus, à l'instar de Kocourek (1991), une surreprésentation des nominalisations, prototypiques du genre de l'écrit scientifique. Il est alors possible de typer le genre en s'intéressant aux constructions présentes dans les corpus. Fifielska (2015) se concentre ainsi sur les constructions les plus typiques de l'article de recherche et relève notamment une grande proportion de constructions à l'impersonnel, au passif impersonnel et au passif réduit. Le genre peut alors être caractérisé par ces constructions particulières, notamment dues à une forte propension dans ce genre à modaliser son propos en évitant de nommer l'agent du procès.

Il nous apparaît alors primordial de prendre en compte ces alternances syntaxiques dans nos différents traitements. Notre travail d'analyse des constructions verbales du LST, présenté section 5.2, intègre d'ailleurs ces alternances afin de pouvoir proposer à des apprenants une ressource des structures verbales attestées en corpus. Cette étude des constructions verbales s'inspire des travaux de Hanks (2013) et de son approche sur corpus des patrons verbaux, définis par des contraintes aux niveaux syntaxique et sémantique. Cette approche des cooccurrences, collocations et patrons récurrents fait notamment suite aux travaux de Sinclair (1991) en linguistique de corpus. Les propriétés distributionnelles sont utilisées, dans le présent travail, pour caractériser le LST au niveau sémantico-syntaxique, et pour l'élaboration d'une ressource incorporant ces deux dimensions.

Notre travail de caractérisation du LST s'effectue ainsi de façon incrémentale. Les analyses syntaxiques nourrissent la classification sémantique qui autorise à son tour l'exploration de routines de niveau supérieur. En termes de description linguistique, les applications possibles de notre ressource du LST sont ainsi multiples. De plus, ceci nous permet de procéder à une auto-évaluation continue de la ressource, pour chaque enrichissement. Ainsi, la qualité de la classification

sémantique des noms du LST que nous opérons est notamment mesurée au niveau de son apport sur l'extraction de routines.

Nous avons dans un premier temps défini notre objet d'étude, le LST, en étudiant ses fonctions et le genre dans lequel il est mobilisé. Nous avons également tracé les principaux objectifs poursuivis dans l'élaboration d'une ressource du LST, afin de préciser quelles informations linguistiques nous voulons y intégrer. Enfin, notre revue des travaux sur le domaine nous a permis d'identifier les éléments dont nous nous inspirons, mais également de mettre en évidence les limites dans l'ensemble de ces travaux, afin de proposer une approche que nous présentons dans la partie suivante.

1.4 Approche outillée du LST

Notre travail s'inscrit dans le domaine de la linguistique outillée, au sens de Habert (2004), dans la mesure où notre objectif est en premier lieu une description linguistique fine du LST, menant à l'élaboration d'une ressource de ce lexique, et non pas à la mise en place d'un protocole automatiquement reproductible.

Les analyses et traitements que nous présentons, dans le présent chapitre et les suivants, s'inscrivent dans le cadre de cette approche, dont nous reprenons les points principaux, rappelés par Tanguy & Fabre (2014, p. 11), en y précisant entre parenthèses la contextualisation dans notre travail :

- Observation des données linguistiques et circonscription des objets d'étude (éléments du LST, propriétés syntaxiques et sémantiques).
- Traitement des données (annotation de corpus) et utilisation de ressources (dictionnaires).
- Interrogation des données par des méthodes automatiques (basées sur la fréquence et la combinatoire).
- Analyse et description linguistique (des propriétés du LST)
- Classification et élaboration/enrichissement d'une ressource (du LST).

L'analyse est, pour chaque étape, précédée par la mise en place de traitements spécifiques (annotation de corpus, extraction de cooccurrents, etc.). Dans le cadre de l'étude de ce lexique particulier, nous pouvons distinguer, parmi les principales spécificités de notre démarche par rapport à l'état de l'art, les points suivants :

- l'intégration d'une phase d'évaluation manuelle des traitements automatiques par des juges experts ;
- l'utilisation d'un corpus de contraste mixte (genre journalistique, oral et fiction) de grande taille (120 millions de mots) ;
- l'analyse des propriétés lexico-syntaxiques du LST à l'aide d'un analyseur syntaxique en dépendance ;
- l'utilisation de ressources lexicographiques à large couverture (*Les Verbes Français* (*LVF*) (Dubois & Dubois-Charlier, 1997) et le *Dictionnaire Électronique des Mots* (*DEM*) (Dubois & Dubois-Charlier, 2010), voir section 3.2.3.2) pour les traitements sémantiques ;
- l'utilisation d'un outil d'exploration de corpus et de calcul de profils combinatoires, le *Lexicoscope*[26] (Kraif & Diwersy, 2012), pour la prise en compte des contextes d'occurrences des éléments du LST.

L'ensemble des procédures d'extraction et de caractérisation du LST suppose dans un premier temps, ainsi que nous l'avons détaillé dans la section 1.2.2.3, une quantification de données linguistiques pour identifier les éléments saillants de l'écrit scientifique au centre de notre analyse. L'approche quantitative nous permettra d'identifier des formes, lemmes, unités lexicales répondant à nos critères de fréquence, de dispersion, de spécificité.

Dans un deuxième temps, par la prise en compte du profil combinatoire (voir section 1.3.5.2), et donc des propriétés lexico-syntaxiques des éléments du LST, nous effectuerons un enrichissement de la ressource du LST. Les profils combinatoires, et plus généralement toute information sur un cooccurrent syntaxique récurrent, interviennent alors pour la plupart de nos analyses : évaluation manuelle de l'extraction, classification sémantique manuelle,

[26] Disponible ici : http : ://dip01.u-grenoble3.fr/~kraifo/lexicoscope/index.php [consulté le 30/07/2015]

expérimentation sur les classifications automatiques, analyse de constructions verbales, étude de routines intégrant un élément du LST et un élément terminologique. L'analyse détaillée des profils combinatoires permet également de gérer la polysémie des unités du LST, dans la mesure où les multiples acceptions d'une même unité sont le plus souvent différenciables par la nature sémantique de leurs cooccurrents.

Nous proposons dans l'illustration ci-dessous une vue d'ensemble de ces différents traitements et analyses que nous effectuons sur le LST.

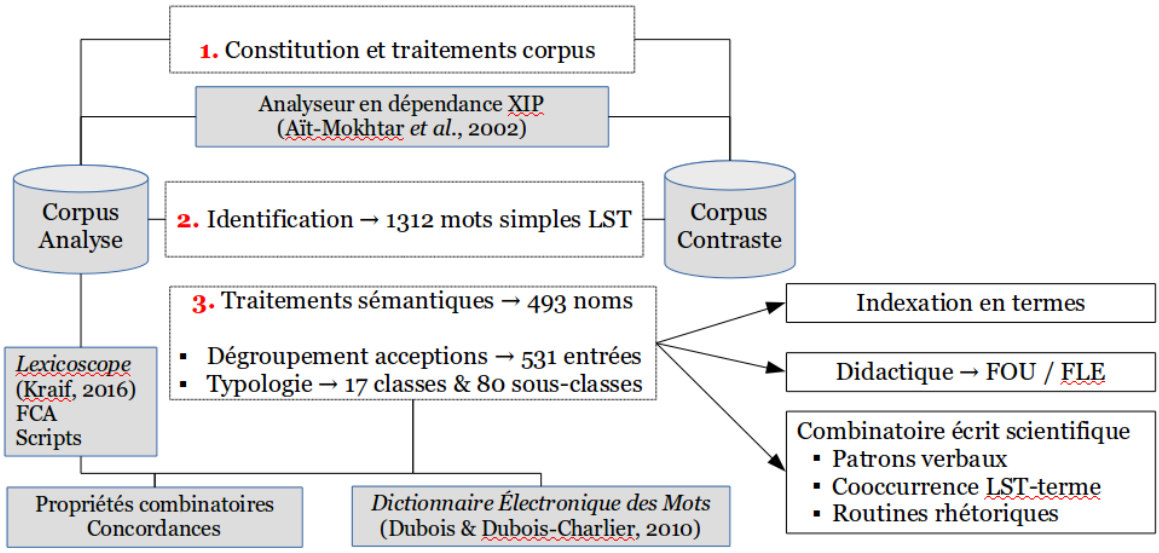


Illustration 1.1 : Traitements et utilisation du LST

La figure permet de mettre en évidence le caractère incrémental de notre étude du LST. Les outils d'enrichissement (*XIP*) et d'interrogation de corpus (*Lexicoscope*) autorisent l'exploitation des informations de combinatoire lexicale en vue de la caractérisation sémantique du LST. Ces traitements sémantiques du LST sont alors réinvestis dans les divers domaines d'applications précédemment évoqués : indexation, didactique, etc. De plus, en intégrant informations sémantiques et propriétés lexico-syntaxiques, nous pouvons alors proposer plusieurs expérimentations semi-automatiques visant à l'étude de phénomènes phraséologiques spécifiques au genre de l'écrit scientifique.

Comme le montre le schéma précédent, notre étude du LST se situe dans le cadre de la linguistique de corpus. À l'instar de Biber (2006), Paquot (2010) et Drouin (2007), les méthodes et outils du traitement automatique de la langue nous

permettent de constituer et d'analyser un ensemble de données linguistiques (occurrence, combinatoire) issues de l'usage en corpus, donc « fondée[s] sur l'observation de régularités à partir de textes » (Habert, 2004, p. 13).

1.5 Traitements automatiques et caractérisations manuelles du LST

Nous revenons dans cette partie conclusive du chapitre sur les différents choix méthodologiques adoptés suite à la définition de notre objet d'étude et de ses propriétés.

Nous venons de montrer que le LST est un ensemble lexical spécifique, entre terminologie et langue générale. Les traitements et les analyses mis en place doivent alors prendre en compte la particularité de ce lexique, notamment par la constitution d'un corpus actualisant le genre de l'écrit scientifique, par l'intégration de la dimension lexico-syntaxique et de la polysémie du LST. Ainsi, au-delà de la définition que nous donnons du LST, les outils et ressources que nous utilisons ont un effet direct sur les résultats de son extraction et sa caractérisation. Les traitements automatiques (segmentation, annotation syntaxique, extraction d'unités lexicales et de leur combinatoire) génèrent bruit et silence, explicables notamment par la complexité de certains phénomènes linguistiques tels que les lexiques aux frontières peu marquées dans l'écrit scientifique ou la polysémie des éléments de ces lexiques.

Nous verrons dans les chapitres suivants que les différents critères, mesures et seuils permettent un premier filtrage des phénomènes linguistiques susceptibles de correspondre à notre définition du LST et répondant aux besoins des applications considérées. Ces premiers résultats automatiquement générés sont, dans notre méthode, systématiquement suivis d'une analyse manuelle visant à les enrichir puis les valider. Les techniques automatiques sont ainsi au service de l'analyse linguistique dans les différentes tâches d'exploration et d'interrogation de corpus. Le présent travail s'inscrit alors dans un cadre heuristique, dans lequel le TAL permet dans un premier temps l'émergence de phénomènes linguistiques remarquables. Il revient alors, dans un second temps, au linguiste (lexicographe, lexicologue) d'analyser et de décrire ces phénomènes. L'alliance d'une approche quantitative et d'une analyse des propriétés lexico-syntaxiques et sémantiques du

LST nous permet d'adopter une méthodologie appropriée aux questions soulevées dans ce chapitre, à savoir la complexité de notre objet d'étude, ses spécificités, ainsi que les perspectives d'utilisation que nous envisageons pour la ressource du LST. Cette méthodologie est notamment à l'œuvre pour la phase d'extraction et d'identification des éléments du LST, présentée dans le chapitre suivant.

Chapitre 2
Identification des mots simples du LST

Sommaire

2.1 Cadre et objectifs de l'extraction du LST

Nous présentons dans ce chapitre la mise en place de notre méthode d'extraction du LST, dont le but est la constitution d'une liste de mots simples du LST, pour l'ensemble des catégories pleines. L'extraction du LST repose sur des traitements automatiques, pour fournir une première liste de candidats-LST, et est validée manuellement afin d'aboutir à une ressource de qualité, correspondant à nos besoins.

Nous avons présenté dans le chapitre précédent notre objet d'étude, le LST en nous situant dans une approche de linguistique de corpus outillée. Une telle approche nécessite l'utilisation et/ou l'élaboration de ressources et outils que nous détaillons dans le présent chapitre. La première ressource à constituer est notre corpus de travail. Ayant défini le LST comme un lexique de genre, ce corpus d'analyse se doit de représenter le genre scientifique objet de nos travaux, à savoir les articles en SHS. Le fait de définir le LST comme associé à ce genre implique sa sous-représentation dans d'autres genres, que nous représentons dans nos travaux par un corpus de contraste diversifié de grande taille. Les différents traitements que nous mettons en place pour l'extraction et la caractérisation du LST nécessitent de pouvoir interroger un corpus annoté en dépendances, comme nous le justifions section 2.2.1.3. En effet, nous tirons parti des informations de cooccurrences et de relations syntaxiques pour caractériser les éléments du LST.

Nous avons précédemment défini pour le LST des propriétés linguistiques : il est spécifique au genre scientifique et traverse les disciplines des SHS. À l'instar de Drouin et Paquot, nous partons de ces propriétés pour définir des critères lexicométriques afin de mettre en place une méthode d'extraction semi-automatique du LST, basée sur les critères de spécificité (ou sur-représentativité) et de répartition (pour juger de la transdisciplinarité).

Le processus se déroule en trois phases principales. Dans un premier temps, nous procédons à la constitution et à l'annotation des corpus nécessaires à nos traitements. Dans un deuxième temps, nous effectuons une extraction automatique de mots répondant aux critères du LST. Enfin, à la vue des résultats générés, nous mettons en place une phase de validation manuelle de ces candidats-

LST. Cette dernière étape est nécessaire pour filtrer des éléments n'appartenant pas au LST mais dont les propriétés ne permettent pas de l'en distinguer automatiquement.

Nous commençons ainsi par la constitution de deux corpus : un corpus d'analyse et un corpus de contraste (section 2.2.1). Le premier définit l'ensemble multidisciplinaire, donc le genre, duquel nous cherchons à extraire le lexique spécifique. Le second sert de base pour calculer la surreprésentation des éléments du LST dans le corpus d'analyse, pareillement aux travaux de Drouin (2007) ou Simpson-Vlach & Ellis (2010). Ce corpus de contraste représente ainsi la « langue générale » par opposition à la langue spécialisée de l'écrit scientifique. L'analyse syntaxique en dépendance que nous effectuons pour ces deux corpus autorise alors les traitements de la seconde étape.

Dans un second temps, l'extraction automatique d'éléments lexicaux candidats-LST est basée sur des critères statistiques (spécificité et transdisciplinarité) et sur les résultats de l'analyse des corpus (section 2.2.2). Ces résultats sont également utilisés pour la troisième phase. Les informations sur les cooccurrents syntaxiques permettent aux évaluateurs une recontextualisation des éléments lexicaux candidats-LST.

La méthode d'extraction se conclut par une phase de validation manuelle, qui fait appel à plusieurs juges experts dans le domaine de l'écrit scientifique (section 2.2.2.2). La mise en place de cette étape finale est motivée par l'objectif d'élaborer une ressource de qualité, adaptée aux applications du LST que nous avons présentées dans le chapitre précédent. Le concours des juges nous a permis d'éliminer le bruit produit par les traitements automatiques et de juger de l'acception effectivement transdisciplinaire des candidats-LST. Afin d'estimer la fiabilité des jugements et la pertinence de la ressource, nous avons procédé au calcul de l'accord inter-annotateurs, et avons ainsi mesuré la difficulté à circonscrire notre objet d'étude.

Ces différents traitements ont été guidés par les applications que nous projetons pour le LST. Ainsi, par son utilisation dans des traitements d'indexation automatique en termes[1], le LST doit être issu d'un corpus représentant les futurs

[1] Ces traitements s'inscrivent dans le cadre du projet TermITH, comme détaillé dans la section 1.1.1.

textes indexés. Le corpus d'analyse est donc constitué d'articles de recherche en SHS, champ scientifique où le lexique présente un fort degré d'ambiguïté entre la langue générale et la langue terminologique de spécialité. Il en est ainsi de *profit, marché, mot, santé, solidarité, loisir,* mots de la langue générale ayant une ou plusieurs acceptions terminologiques[2] en sociologie, psychologie, linguistique, économie. Le corpus de contraste répond au besoin d'identifier les éléments lexicaux absents, ou peu présents, dans d'autres genres que l'écrit scientifique. Cette sur-représentation du LST dans les articles en SHS est d'ailleurs la preuve de son utilité dans l'extraction terminologique. En isolant ainsi le lexique spécifiquement mobilisé dans ce contexte, nous nous assurons également de proposer une ressource lexicale adaptée pour l'aide à la rédaction scientifique. Les difficultés rencontrées par les apprenants et natifs dans la rédaction scientifique concernent ainsi ce lexique particulier et les constructions dans lesquels il s'inscrit (Hatier & Yan, à paraître).

La phase de validation manuelle participe des mêmes préoccupations. D'une part, les perspectives didactiques, présentées section 1.1.2, imposent une vérification manuelle de la ressource, qui permettra la réalisation d'exercices et/ou de cours ciblés sur ces éléments essentiels à l'organisation du discours scientifique. D'autre part, au niveau des processus d'indexation, le fait de disposer d'une ressource validée manuellement, enrichie ultérieurement aux niveaux sémantique et syntaxique, assure une meilleure détection des termes, comme le montrent Jacquey *et al.* (2013). Ces différents objectifs ont ainsi motivé plusieurs choix méthodologiques que nous abordons dans la partie suivante.

2.2 Extraction semi-automatique du LST : ressources et traitements

Nous présentons dans cette partie l'ensemble des traitements, automatiques et manuels, menant à l'extraction des mots simples du LST.

Dans la première section, nous détaillons les traitements liés au corpus, notamment la segmentation et l'annotation morpho-syntaxique permettant l'identification des mots simples candidats-LST. Par mots simples, nous entendons

[2] Tel que vérifié dans Le *Grand Dictionnaire Terminologique* : http://www.granddictionnaire.com/[consulté le 1/09/2016]

les mots identifiés comme « token » par l'analyseur, et définis par un lemme et une catégorie. Ainsi, dans les deux exemples suivants, la forme *alternative* correspond à deux lemmes et deux catégories distincts :

- le nom *alternative* : *Une alternative$_{LST}$ à l'utilisation d'un corpus lemmatisé consiste à dresser un inventaire exhaustif des formes verbales envisagées ;*[3]
- l'adjectif *alternatif* : Ce *développement de la presse alternative$_{LST}$ se retrouve nationalement et dans différentes régions, avec par exemple Fakir.*[4]

L'ambiguïté peut porter uniquement sur la catégorie, comme dans les exemples suivants où une même forme (*caractéristique*) correspond à un lemme identique mais à deux catégories différentes :

- nom : *On sait en effet qu'une caractéristique$_{LST}$ du langage écrit est qu'il comporte des régularités ;*[5]
- adjectif : *Un lexique employé de manière systématique, caractéristique$_{LST}$ des « aventures dont vous êtes le héros » interpelle l'internaute.*[6]

L'extraction du LST se base alors sur ces unités formelles composées d'un lemme et d'une catégorie (le nom *alternative*, l'adjectif *caractéristique*).

2.2.1 Constitution et annotation des corpus

Nous présentons dans cette partie les étapes de constitution et d'enrichissement de corpus. Notre approche se base en premier lieu sur les corpus et sur l'hypothèse que le LST, de par ses propriétés de spécificité et de répartition, a un comportement statistiquement repérable. Cette identification du LST par des critères statistiques présuppose la présence de deux corpus, permettant la comptabilisation et la comparaison des occurrences.

---

[3] Rebeyrolle, J., & Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, *25*, 153-174.

[4] Frisque C. (2010). Des espaces médiatiques et politiques locaux ? *Revue française de science politique* 5/2010 (Vol. 60), p. 951-973

[5] Doignon, N., & Zagar, D. (2006). Les enfants en cours d'apprentissage de la lecture perçoivent-ils la syllabe à l'écrit ?. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *60*(4), 258.

[6] Broudoux, E. (2011). Le documentaire élargi au web. *Les Enjeux de l'information et de la communication*, (2), 25-42.

Nous avons procédé à l'élaboration d'un corpus d'analyse, constitué d'articles de recherche en SHS, et d'un corpus de contraste afin de mettre en place une analyse contrastive des lexiques mobilisés et conséquemment l'émergence du LST.

Nous nous situons ainsi, pour l'ensemble de nos travaux, dans le cadre de la linguistique de corpus. Cette approche envisageant les corpus comme base de l'analyse, nous devons précautionneusement les définir et les constituer. Biber (1993b) note qu'avant toute constitution de corpus, il faut en définir l'intension et les spécifications, en fonction des buts poursuivis, des besoins.

Nous avons défini dans le premier chapitre le genre étudié, l'article de recherche en SHS, appartenant au genre de l'écrit scientifique. Nous définissons dans la section suivante les spécifications de ce corpus d'articles dont la taille de 5 millions de mots peut être considérée comme conséquente au regard de travaux précédents[7]. Nous détaillons par la suite l'élaboration du corpus diversifié utilisé comme corpus de contraste, nécessaire au calcul de spécificité, critère définitoire du LST.

2.2.1.1 Corpus d'analyse

Le corpus d'analyse est composé de 500 articles de recherche en français dans 10 disciplines des SHS : linguistique, psychologie, sciences de l'éducation, économie, sciences politiques, anthropologie, histoire, géographie, sciences de l'information et de la communication, sociologie. Issu en partie du projet Scientext[8], ce corpus est constitué d'articles de revues à comité de lecture. Pour la constitution du corpus d'analyse, nous avons tout d'abord repris le corpus constitué par Tran (2014), pour ses travaux sur les marqueurs discursifs dans l'écrit scientifique. Nous avons ensuite complété ce corpus, pour l'étendre de 300 à 500 articles.

La sélection des articles constituant notre corpus, dont la partie originale est détaillée par Tran (2014, p. 54-57), se base premièrement sur le classement des revues scientifiques de l'AERES[9]. Cette classification permet de s'assurer de la

---

[7] Granger & Paquot (2009b) effectuent des travaux similaires sur un corpus de 2 millions de mots, Drouin sur un corpus de 2,3 millions de mots.

[8] Projet ANR « Corpus et outils de la recherche en sciences humaines et sociales » (2007-2010). Site du projet : http://scientext.msh-alpes.fr. [Consulté le 10/01/2016]

[9] AERES : Agence d'évaluation de la recherche et de l'enseignement supérieur.

qualité des revues en SHS. Nous avons ainsi sélectionné ces articles en prenant en compte plusieurs critères :

- La revue doit avoir une note de A ou B par l'AERES, garantissant ainsi la qualité des articles.
- La date de publication doit se situer entre 2005 et 2013, afin de travailler en synchronie sur des textes actuels.
- Le thème de la revue doit être assez central, non marginal, pour ne pas sur-représenter un domaine particulier d'une discipline. Nous avons ainsi évité les revues de psycholinguistique ou de sociolinguistique pour constituer le sous-corpus de linguistique.
- Un auteur ne doit pas être représenté par plus de deux articles pour éviter de sur-représenter des phénomènes idiosyncratiques.
- Les auteurs sont publiants dans des revues francophones à comité de lecture, et possèdent ainsi une certaine maîtrise de la rédaction scientifique en français. Par ailleurs, des précédents travaux montrent que le français académique comporte des universaux transcendant les nationalités des scripteurs (Fløttum *et al.*, 2006). Ce constat nous conforte dans l'hypothèse d'un socle lexical commun dans un corpus composé de textes dont les auteurs ne sont pas nécessairement francophones natifs.

Ces critères ont pour but la constitution d'un corpus aussi représentatif que possible du genre de l'article scientifique en SHS. En effet, les recherches empiriques, sur des échantillons, ne peuvent être généralisées que si ces échantillons sont représentatifs (Fløttum *et al.*, 2006). Les contraintes sur la qualité des revues, la date de publication, la généricité des sujets, sont définies dans ce but.

Le lexique non-terminologique spécifique aux écrits de recherche est notamment fonction, selon Fløttum *et al.* (2006), de l'identité disciplinaire, de la nationalité et de la langue maternelle du scripteur. En constituant un corpus multidisciplinaire d'articles dont les scripteurs sont francophones, nous nous assurons d'extraire un lexique commun, correspondant aux propriétés du LST, en limitant au maximum les variations autres que disciplinaires. Nous constituons

ainsi un corpus d'analyse d'environ 5 millions de mots, dont la composition est détaillée dans le tableau ci-dessous.

Discipline	**Nombre d'articles**	**Nombre de mots**
Anthropologie	50	493 988
Économie	50	417 944
Géographie	50	400 533
Histoire	50	773 170
Linguistique	50	425 952
Psychologie	50	417 846
Sciences de l'éducation	50	417 069
Sciences de l'information	50	399 007
Sciences politiques	50	548 222
Sociologie	50	540 630
Total Corpus Analyse	**500**	**4 834 361**

Tableau 2.1: Composition du corpus d'analyse

Chaque sous-corpus disciplinaire est ainsi composé de 50 articles, soit 50 fichiers XML au format TEI-Lite. Les métadonnées de chaque article intègrent l'auteur, le titre, l'éditeur, la date et l'adresse URL à laquelle le fichier a été récupéré (au format XHTML) avant d'être adapté à notre DTD[10]. Chaque article a également été automatiquement balisé en parties textuelles afin de pouvoir identifier des sections telles que résumé, mots-clés, introduction, notes, etc. La structure des documents, conforme au schéma utilisé pour le corpus Scientext[11] présentée en détail par Tran (2014, p. 59-61), est comme suit :

- Un en-tête (<*teiHeader*>) sur les métadonnées du corpus et de l'article, informant sur :
  - la bibliographie du document (<*fileDesc*>) ;
  - la méthodologie d'encodage du texte (<*encodingDesc*>) ;
  - les aspects non bibliographiques du texte (<*profileDesc*>) : langue utilisée, mots-clés, genre, discipline.

[10] Nous remercions ici Thi Thu Hoai Tran et Marie-Paule Jacques pour leur aide dans la constitution du corpus.

[11] Pour une présentation de la constitution du corpus Scientext, voir Falaise (2014).

- Le contenu de l'article (<*text*>), intégrant :
  - ce qui précède le corps de l'article, i.e les informations métatextuelles, présentes dans l'article (<front>) : auteur, titre, résumé et mots-clés en plusieurs langues ;
  - le corps de l'article (<*body*>) ;
  - ce qui suit le corps du texte (<*back*>) : notes et références.

Un extrait du corpus au format TEI-Lite est présenté dans l'illustration ci-après :

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="file:teilite.rnc" type="compact"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
            </titleStmt>
            <publicationStmt>
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <author>Nicolas Penin  </author>
                    <title>Le sexe du risque</title>
                    <editor></editor>
                    <date>4/2006</date>
                    <extent>651-658</extent>
                    <title>Ethnologie française</title>
                    <pubPlace>http://www.cairn.info/revue-ethnologie-francaise-2006-4-page-651.htm</
                    pubPlace>
                    <publisher>P.U.F.</publisher>
                </bibl>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
        </encodingDesc>
        <profileDesc>
            <langUsage>
                <language ident="fr">français</language>
            </langUsage>
            <textClass>
                <keywords scheme="subjects"><list>
                    <item>sports, prises de risque, sexe, masculin, domination</item></list>
                </keywords>
                <keywords scheme="genre">
                    <list><item>article</item></list>
                </keywords>
                <keywords scheme="discipline"><list><item>anthropologie</item></list>
                </keywords>
                <keywords scheme="mothertongue">
                    <list><item>français</item></list>
                </keywords>
            </textClass>
        </profileDesc>
    </teiHeader>
<text>
    <front>
```

Illustration 2.1: Extrait du corpus XML – Article Anthropologie

Nous disposons ainsi, pour chaque article, de l'intégralité du texte, ainsi que des métadonnées correspondantes, présentes dans les balises correspondantes.

2.2.1.2 Corpus de contraste

Dans la perspective de comparer des fréquences de mots dans différents genres, nous devons disposer d'un corpus de contraste, que nous « opposerons » au corpus d'analyse, représentant le genre scientifique. Ce corpus de contraste doit représenter d'autres genres, dans le but de faire émerger des phénomènes spécifiques au genre scientifique. Il faut également nous assurer de la contemporanéité des écrits constituant ce corpus de contraste, pour que la comparaison entre les deux corpus ne dégage pas des phénomènes imputables à des variations diachroniques. Comme il n'existe pas, à notre connaissance, un corpus correspondant à ces critères (en français, hétérogène, contemporain, et de grande taille), nous avons élaboré un corpus de ce type.

Nous avions utilisé dans un premier temps une base de données de fréquences lexicales, *Lexique 3*[12]. Cette ressource, librement disponible, présente cependant quelques limites dans le cadre de nos travaux. Bien qu'elle intègre des romans contemporains et des sous-titres, elle ne comporte pas de sous-corpus journalistique, genre particulièrement intéressant dans l'étude contrastive du lexique scientifique (cf. le corpus de contraste *Le Monde* utilisé par Drouin (2007)). De plus, la base de données ne permet pas un accès aux usages en contexte des mots, n'autorisant donc pas l'extraction d'informations lexico-syntaxiques, sur les cooccurrences notamment. Or, ces informations de cooccurrences nous sont nécessaires pour les processus d'enrichissement sémantique et syntaxique développés dans les chapitres suivants. De plus, dans un souci d'homogénéité des traitements sur les différents corpus, la segmentation et les annotations se doivent d'être réalisées par le même outil pour les deux corpus. Or, la base de données mentionnée ci-dessus ne se base pas sur le même analyseur que celui que nous utilisons.

Nous avons donc opté pour l'élaboration d'un corpus de grande taille nous permettant d'opérer des comparaisons de fréquences, de cooccurrences, ceci afin d'isoler les éléments spécifiques à notre corpus d'analyse, constitué d'articles de recherche en SHS. Notre corpus de contraste ne prétend pas être parfaitement représentatif de la langue française, mais pour répondre à notre objectif contrastif,

[12] New B., Pallier C., Ferrand L., 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 10/01/2016]

il est constitué de genres divers. Le but est ici de pouvoir comparer le lexique mobilisé dans l'écrit scientifique aux lexiques présents dans des textes aux caractéristiques différentes, ne se limitant pas à la fiction ou au genre journalistique, mais incluant également des transcriptions d'oral et des sous-titres de films, afin de disposer d'un corpus de contraste diversifié.

Chaque genre ayant ses spécificités, nous souhaitons donc isoler celles de l'écrit scientifique en le comparant à cet ensemble hétérogène. Nous faisons l'hypothèse, à la suite de Paquot & Bestgen (2009) et Drouin (2007), que les mots décrivant l'activité scientifique, le raisonnement et les procédures seront sur-représentés dans le corpus d'analyse.

Les écrits scientifiques sont, comme le note Pecman (2004b), normés et peu favorables à la créativité. Les études comparatives sur les genres ont ainsi permis de mettre en évidence des propriétés de genre. Thoiron & Béjoint (1989, p. 662) observent par exemple que le genre littéraire s'affranchit « des groupements trop prévisibles » alors que la « langue techno-scientifique, au contraire, évite l'originalité », observation cependant remise en cause par Siepmann (2016) qui observe des collocations récurrentes dans des corpus littéraires.

Nous avons constitué, dans ce but contrastif, un corpus de contraste de grande échelle, intégrant trois sous-corpus :

- Le sous-corpus littéraire est composé de 331 romans en français original ou traduit, paru après 1950. Ce sous-corpus est issu du projet ANR Emolex[13].
- Le sous-corpus journalistique comprend des articles de diverses thématiques (politique, société, économie, sport, etc.) de journaux de presse quotidienne régionale et nationale. Ce sous-corpus est également issu du projet Emolex.
- Le sous-corpus « d'oral » provient de 4 sources différentes :
    - les sous-titres de films (français ou non) rassemblés dans le cadre du projet OPUS (Tiedemann & Nygaard, 2004)[14], représentant 38 millions de mots :

[13] Projet franco-allemand ANR/DFG Emolex (ANR-09-FASHS-017) sur l'étude du lexique des émotions dans cinq langues européennes. Site : www.emolex.eu [Consulté le 10/01/2016]

[14] Site du projet OPUS : http://opus.lingfil.uu.se/ [Consulté le 10/01/2016]

- les transcriptions du corpus de parole (radio) issu de la campagne d'évaluation ESTER[15], 1 200 000 million de mots ;
- les transcriptions du Corpus de Français Parlé Parisien (CFPP)[16], 535 000 mots ;
- les transcriptions issues du projet TCOF[17], 340 000 mots.

Le tableau ci-après détaille la composition de ce corpus de contraste.

	Nombre de mots
Fiction	**39 984 513**
Le Monde (2008)	6 666 637
Libération (2008)	6 666 594
Le Figaro (2008)	6 666 607
Presse Quotidienne Nationale	**19 999 838**
Ouest France (2008)	9 999 883
Sud Ouest (2008)	9 999 819
Presse Quotidienne Régionale	**19 999 702**
Sous-titres OPUS	**38 266 218**
Ester (Radio)	1 027 227
CFPP (Parler Parisien)	366 668
TCOF (Entretien)	339 888
Oral	**1 733 783**
Total Corpus Contraste	**119 984 054**

Tableau 2.2: Composition du corpus de contraste

Le corpus de contraste est constitué sur le modèle du corpus d'analyse, et inclut les mêmes informations. Les métadonnées, déjà présentes sur les fichiers sources récupérées (issus de différents projets) sont renseignées pour l'ensemble des sous-corpus composant le corpus de contraste.

---

[15] Ester : Évaluation des Systèmes de Transcription d'Émissions Radiophoniques. http://www.technolangue.net/article60.html [Consulté le 10/01/2016]

[16] CFPP : ensemble d'interviews sur les quartiers de Paris et de la proche banlieue. http://cfpp2000.univ-paris3.fr/ [Consulté le 10/01/2016]

[17] Traitement de Corpus Oraux en Français : projet mené par le laboratoire ATILF http://www.cnrtl.fr/corpus/tcof/ [Consulté le 10/01/2016]

En plus de ces métadonnées, nous enrichissons par la suite nos corpus d'une annotation en catégories syntaxiques, lemme, traits morpho-syntaxiques et relations de dépendance, résultat de l'analyse syntaxique que nous détaillons dans la partie suivante.

2.2.1.3 Analyse syntaxique des corpus

L'enrichissement des corpus d'analyse et de contraste, par l'application d'une analyse syntaxique en dépendance, se situe dans le cadre de notre approche de linguistique outillée. Nous voulons tirer parti, autant que possible, des informations linguistiques présentes dans les corpus. La caractérisation du LST se basant sur les corpus, l'adéquation de ces derniers à nos besoins est donc essentielle. Au-delà de la constitution des corpus (critères de qualité et de répartition des articles), l'annotation et plus précisément la qualité de l'annotation déterminent en partie la qualité de l'ensemble des traitements. Les traitements sémantiques et syntaxiques que nous envisageons pour le LST (chapitres 3, 4 et 5) nécessitent un corpus analysé syntaxiquement, rendant ainsi possible l'exploitation automatique de données lexico-syntaxiques.

Les corpus d'analyse et de contraste sont traités avec l'analyseur en dépendance *XIP* (Aït-Mokhtar, Chanod, & Roux, 2002). Le choix de cet analyseur performant est le résultat d'une collaboration prolongée avec le centre de recherche XRCE de Xerox. Ceci nous a notamment permis d'intégrer des grammaires adaptées, de nouvelles dépendances, afin d'ajuster l'outil à nos besoins spécifiques[18]. En traitant les deux corpus avec un outil et des paramètres identiques, nous nous assurons ainsi d'une homogénéité dans la segmentation, l'étiquetage morpho-syntaxique et l'analyse en dépendance. Avant de présenter le format de sortie intégrant les résultats de l'analyse syntaxique, nous détaillons dans la partie suivante les règles de grammaires et étiquettes de dépendance que nous avons ajoutées afin d'adapter les sorties à nos besoins spécifiques.

[18] Nous remercions Claude Roux et Ágnes Sándor pour leur aide précieuse lors de la mise en place de ces post-traitements.

2.2.1.3.1 Dépendances syntaxiques

XIP est un analyseur en dépendance robuste, qui permet d'obtenir, pour chaque phrase analysée, une liste de tokens suivie d'une liste de dépendances syntaxiques entre ces tokens. Le fonctionnement de *XIP* peut être résumé en trois phases (Hagège & Tannier, 2007, p. 492) :

- une phase de pré-traitement, gérant la segmentation, l'étiquetage morphologique et la catégorisation en parties du discours ;
- une phase d'analyse syntaxique de surface, résultant d'une première analyse en chunks (Abney, 1991), permettant l'annotation des relations syntaxiques et des entités nommées ;
- une phase d'analyse syntaxique profonde, basée sur la phase précédente, annotant des relations syntaxiques profondes, tels les sujets et objets profonds (notamment pour les phrases à voix passive et les constructions avec un verbe support).

Les résultats de ces analyses (basées sur un riche ensemble de règles) sont notamment récupérables au format XML. Nous reformatons à l'aide de routines *Perl* les sorties *XIP* pour obtenir un format comme illustré ci-dessous :

```xml
<s id="1">
    <tc>
        <t id="1" c="DET" l="le" f="MASC FEM SG">L'</t>
        <t id="2" c="NOUN" l="analyse" f="FEM SG">analyse</t>
        <t id="3" c="PREP" l="de" f="">de</t>
        <t id="4" c="DET" l="le" f="FEM SG">la</t>
        <t id="5" c="NOUN" l="phrase" f="FEM SG">phrase</t>
        <t id="6" c="VERB" l="comprendre" f="SG P3 IND">comprend</t>
        <t id="7" c="DET" l="une" f="FEM SG">une</t>
        <t id="8" c="NOUN" l="liste" f="FEM SG">liste</t>
        <t id="9" c="PREP" l="de" f="">de</t>
        <t id="10" c="NOUN" l="tokens" f="GUESSED MASC FEM PL">tokens</t>
        <t id="11" c="COORD" l="et" f="">et</t>
        <t id="12" c="DET" l="une" f="FEM SG">une</t>
        <t id="13" c="NOUN" l="liste" f="FEM SG">liste</t>
        <t id="14" c="PREP" l="de" f="">de</t>
        <t id="15" c="NOUN" l="relation" f="FEM PL">relations</t>
        <t id="16" c="SENT" l="." f="END">.</t>
    </tc>
    <dc>
        <d t="SUBJ" h="6" d="2"/>
        <d t="OBJ" h="6" d="8"/>
        <d t="OBJ" h="6" d="13"/>
        <d t="COORDITEMS" h="8" d="13"/>
        <d t="PREPOBJ" h="15" d="14"/>
        <d t="PREPOBJ" h="10" d="9"/>
        <d t="PREPOBJ" h="5" d="3"/>
        <d t="DETERM" h="13" d="12"/>
        <d t="DETERM" h="8" d="7"/>
        <d t="DETERM" h="5" d="4"/>
        <d t="DETERM" h="2" d="1"/>
        <d t="U3_DE_NMOD" h="8" d="10"/>
        <d t="U3_DE_NMOD" h="13" d="15"/>
        <d t="U3_DE_NMOD" h="2" d="5"/>
    </dc>
</s>
```

Illustration 2.2: Représentation de l'analyse d'une phrase au format XML

La phrase prise pour exemple ici est : *L'analyse de la phrase comprend une liste de tokens et une liste de relations*. Nous pouvons observer que l'analyse syntaxique d'une phrase est composée de deux ensembles :

- une liste de formes (tokens, balise <*t*>) correspondant à la segmentation de l'analyseur. Un token, outre sa forme fléchie, est défini par les attributs suivants : identifiant, catégorie syntaxique, lemme, traits morpho-syntaxiques ;
- une liste de relations syntaxiques (balise <*d*>) entre ces tokens. Une relation est définie par les attributs suivants : type de la relation, identifiants du gouverneur (*h*) et du dépendant (*d*) de la relation.

Nous pouvons également représenter le résultat de cette analyse syntaxique sous la forme d'un graphe dont les nœuds sont les tokens et les arêtes les relations de dépendance. La sortie XML présentée ci-dessus peut ainsi être visualisée comme dans l'illustration suivante :

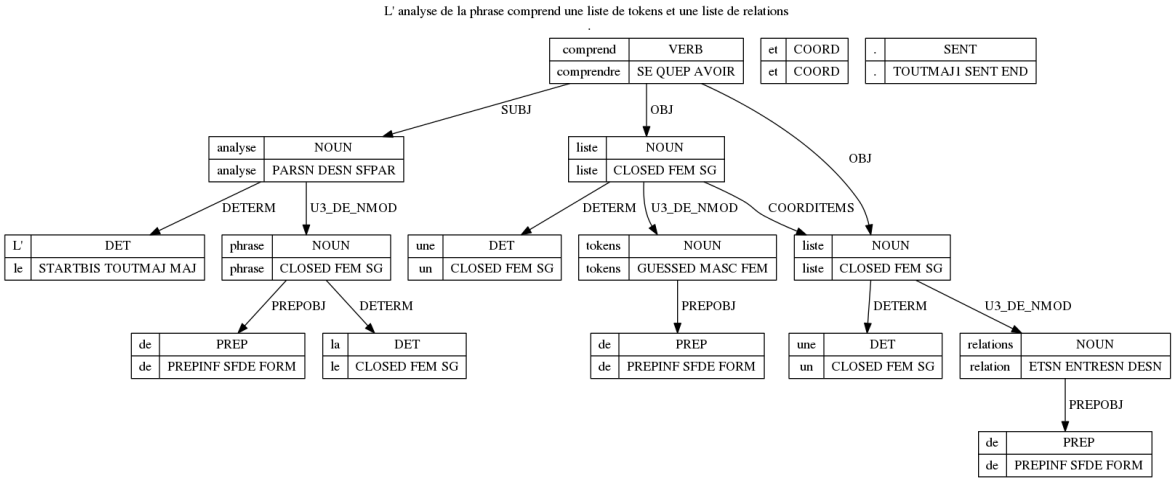


Illustration 2.3: Exemple d'analyse syntaxique sous forme de graphe

La dépendance de type sujet entre le token *analyse* et le token *comprend*, définie par ailleurs dans le document XML, est ici visible sous la forme d'une flèche orientée (pointe vers le dépendant) et typée (par l'étiquette *SUBJ*) entre les tokens précités, dont les lemmes, catégories et traits morpho-syntaxiques sont renseignés.

Nous utilisons ces représentations graphiques pour observer les sorties de l'analyseur sur un échantillon du corpus. Nous identifions ainsi des contextes pour lesquels nous voulons affiner les résultats de l'analyse de *XIP*. Nous présentons dans la partie suivante une partie de ces post-traitements.

2.2.1.3.2 Post-traitements syntaxiques

Les post-traitements constituent l'étape de raffinement de l'analyse syntaxique. L'objectif est alors d'adapter le mieux possible les sorties de l'analyse syntaxique aux traitements ultérieurs. Les dépendances déterminent la qualité des informations sur les cooccurrents syntaxiques. Or, plusieurs traitements sur le LST prennent appui sur ces informations de combinatoire (voir notamment les sections 3.3.2, 4.4 et 2.2.2.2.1 de ce chapitre).

Nous présentons dans cette section les modifications que nous avons effectuées au fur et à mesure de l'observation des résultats de l'analyse automatique.

L'analyseur retenu, *XIP*, permet l'ajout de règles pour construire de nouvelles dépendances basées sur les résultats de l'analyse (Hagège & Roux, 2003). Ces règles, rédigées selon le formalisme *ad hoc* de l'analyseur, permettent, en présence de configurations prédéfinies, de supprimer ou d'ajouter de nouvelles relations et/ou de nouveaux traits. La section A.I en annexe présente un aperçu du fonctionnement de ces règles.

Nous observons par exemple que l'étiquetage des cas de coordination n'est pas adapté aux traitements de calculs de cooccurrences que nous projetons de faire. Ainsi, si nous prenons pour exemple la phrase suivante :

- *Parallèlement, le paysage constitue un enjeu reconnu d'un point de vue économique et culturel.*[19]

Les résultats de l'analyse indiquent une relation de dépendance de type adjectival entre *point de vue* et *économique* ainsi qu'une relation de dépendance de type coordination entre *économique* et *culturel*. L'analyse n'indique pas de relation de dépendance directe entre *point de vue* et *culturel*. Cette relation peut être déduite automatiquement du contexte en propageant la relation déjà existante entre *point de* vue et l'adjectif *économique*. Ce type de normalisation, déjà éprouvé en analyse distributionnelle, permet alors d'accéder « à des relations syntaxiques plus profondes [et] d'expliciter certaines relations » (Fabre & Bourigault, 2006, p. 123).

Nous obtenons ainsi une analyse enrichie en nouvelles dépendances, tel qu'illustré ci-dessous :

[19] Fourault-Cauët V., (2010). Le paysage, outil de territorialisation et d'aménagement incomplet pour les forêts méditerranéennes ? *Annales de géographie* 3/2010 (n° 673), p. 268-292.

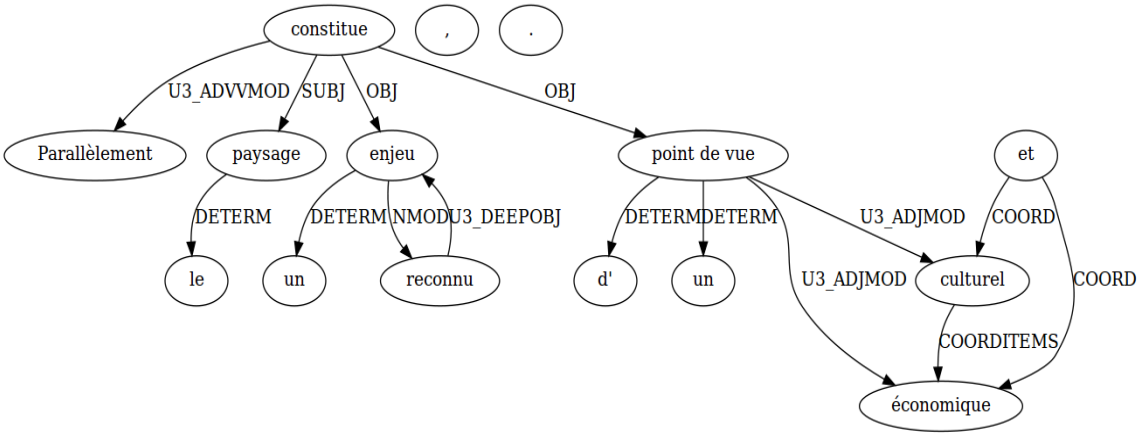


Illustration 2.4: Exemple d'analyse syntaxique après post-traitement

Dans cet exemple, nous pouvons identifier les relations de dépendance résultant de nos grammaires par le préfixe *u3_*. Ainsi, une relation directe entre *économique* et *point de vue* a été ajoutée, ainsi qu'une relation d'objet profond entre *reconnu* et *enjeu*.

Nous avons rédigé un ensemble de règles similaires concernant des phénomènes tels que la coréférence, l'attribut du sujet, le passif réduit, les complétives en *si* et *que,* certaines constructions avec verbes de contrôle et verbes modaux. Dans l'optique des calculs de cooccurrences, ceci nous permet de prendre en compte le maximum de relations de dépendance. Nous dégroupons également certaines relations pour ajuster la granularité à nos besoins. La richesse de certaines étiquettes, ou au contraire la limite d'autres, nous ont décidé à mettre en place des scripts de réécriture des relations, afin d'avoir à notre disposition un ensemble de relations définies. Ainsi, *XIP* définit une relation modifieur de nom (*nmod*) entre les noms *manifestation* et *réforme* pour les deux exemples suivants :

- *la manifestation pour la réforme aura lieu demain* ;
- *la manifestation contre la réforme aura lieu demain*.

Nous souhaitons pouvoir faire la distinction entre ces deux types de constructions sans devoir accéder à la forme de la préposition. Nous intégrons alors la préposition elle-même à l'étiquette de la relation, et créons de nouvelles relations telles que *pour_nmod*, *contre_nmod, de_nmod,* etc. Ces ajouts de relations nous garantissent une visualisation affinée des propriétés lexico-

syntaxiques et ainsi une analyse plus qualitative du lexique que nous étudions. Ces post-traitements ont pour but de s'assurer de la pertinence de ces propriétés, primordiales dans ce travail de caractérisation du LST dans la mesure où la suite de nos traitements s'appuie sur ces relations de dépendance.

Nous opérons, comme le présentent Baroni & Lenci (2010), des normalisations et regroupements de relations de dépendance : allocation de nouvelles étiquettes, suppression de certaines dépendances, calcul de relation indirecte (dans les constructions avec auxiliaires ou verbes supports).

Ainsi, si nous prenons pour exemple le verbe *intervenir*, *XIP* annote une relation de dépendance de type sujet entre ce verbe et le nom souligné pour les deux exemples suivants :

- *En d'autres termes, une <u>firme</u> **intervenant** sur le segment de la haute qualité est incitée à produire [...]*[20] ;
- *Les <u>méthodes</u> d'aide à la décision **interviennent** pour faciliter les choix entre différentes décisions ou évaluations [...]*[21]

XIP renseigne également le sujet profond d'un verbe à la voix passive et permet ainsi de distinguer les sujets syntaxique et sémantique pour ce type d'alternance.

Cette étiquette partagée par des constructions différentes nous permet de comptabiliser de la même manière les deux noms (*firme* et *méthode*) lors du calcul du profil combinatoire (voir section 1.3.5.2) du verbe *intervenir*. Ce dernier aura alors notamment pour propriété lexico-syntaxique le fait de prendre pour sujet les deux noms précités.

Considérons maintenant l'exemple suivant :

- *Il joue le rôle d'un pivot <u>qui</u> **intervient** dans les deux relations prédicatives*[22].

---

[20] Coestier, B. (1998). Asymétrie de l'information, réputation et certification. *Annales d'économie et de statistique*, 49-78.

[21] Renard, F., & Chapon, P. M. (2010). Une méthode d'évaluation de la vulnérabilité urbaine appliquée à l'agglomération lyonnaise. *L'Espace géographique*, *39*(1), 35-50.

[22] Pierrard, M. (1999). Grammaticalisation et contexte : l'extension des emplois de comme. *Revue de sémantique et de pragmatique*, *6*, 133-144.

D'un point de vue informationnel, il nous est apparu plus intéressant de prendre en compte la cooccurrence avec l'antécédent du pronom *qui* qu'avec le pronom lui-même. L'analyse en dépendance de cet exemple recense donc une relation sujet entre *qui* et *intervient* mais également une relation de coréférence (étiquetée *COREF* par *XIP*) entre le pronom *qui* et le nom *pivot*. Nous avons donc une relation sujet indirecte entre *pivot* et *intervient*, que nous traduisons en relation directe. Nous avons élaboré de cette manière plusieurs règles de grammaires locales pour recalculer certaines relations et en renommer d'autres.

Tanguy, Sajous & Hathout (2015) remarquent cependant que les normalisations des passifs, participes présents, ainsi que la recherche de l'antécédent des pronoms relatifs, n'apportent pas de gains significatifs dans la tâche déterminée de calcul de voisinage, et sont donc inopérantes dans la perspective d'une classification sémantique automatique des noms du LST, présentée chapitre 4. Nous avons cependant choisi d'intégrer ces modifications dans la mesure où plusieurs utilisations des dépendances sont prévues. Par exemple, les phases de validation du LST (voir section 2.2.2.2.1) et de classification sémantique manuelle (voir section 3.3.2), s'appuient sur ces dépendances. Le but est alors d'identifier des informations lexico-syntaxiques suffisamment informatives pour que les juges puissent recontextualiser les occurrences et cooccurrences au travers de ces propriétés, permettant ainsi une meilleure représentation des différentes acceptions mobilisées.

Les résultats des traitements automatiques sont dépendants du corpus et des outils employés pour l'annotation des textes. Ce premier traitement peut ainsi donner lieu à diverses erreurs : de segmentation (la locution adverbiale *etc.* segmentée avec ou sans le point), de lemmatisation (le déterminant anglais *the* analysé comme le nom *thé*), de catégorisation morphosyntaxique (le nom féminin *contrée* annoté comme adjectif), de relations syntaxiques.

En définitive, à la suite de l'ensemble de ces traitements, nous disposons de deux corpus, d'analyse et de contraste, analysés syntaxiquement et enrichis en métadonnées (éléments de bibliographie, parties textuelles, etc.). Nous nous basons sur ces corpus annotés pour identifier semi-automatiquement les éléments constituant le LST.

2.2.2 Procédure d'extraction du LST

Nous présentons dans cette section la procédure d'extraction des mots simples du LST que nous avons mise en place, en détaillant les différents apports de notre méthode au regard des travaux du domaine.

La première étape dans ce processus d'extraction est la délimitation des unités lexicales que nous voulons identifier. Nous nous basons pour ce faire sur les résultats de la phase de segmentation opérée par *XIP*, dont les lexiques comportent certaines expressions polylexicales. Ainsi, *point de vue* ou *dans le cadre de* ne sont pas segmentés et correspondent pour l'analyseur à un unique token. Les lexiques de *XIP* ne sont cependant pas exhaustifs et nous devons ainsi prendre en compte les conséquences de la segmentation sur l'ensemble de nos traitements. Notre étude se concentre sur les mots simples du LST, les unités phraséologiques d'ordre supérieur relevant de travaux distincts. D'autres membres du laboratoire LIDILEM, étudiant également l'écrit scientifique et le LST, travaillent ainsi spécifiquement sur les expressions polylexicales (expressions figées, collocations) et effectuent des traitements adaptés afin de les intégrer, à terme, dans la ressource lexicale (Tutin & Kraif, 2016).

Comme signalé précédemment, notre unité lexicale associe un lemme et une catégorie (adjectifs, adverbes, verbes et noms). Nous nous démarquons ainsi de Coxhead (2002) qui ne fait pas de distinction catégorielle systématique.

2.2.2.1 Application des critères statistiques

À l'instar de Drouin (2007) ou Paquot (2010), nous partons de notre définition du LST et traduisons ses propriétés linguistiques en critères statistiques afin d'extraire automatiquement les éléments le composant. Nous verrons qu'il est complexe, pour certaines des propriétés du LST, de les transcrire efficacement en termes statistiques. La nature méta-discursive et méta-scientifique du LST, inscrite dans notre définition, ne peut ainsi se traduire directement au moyen de seuils de fréquence. À l'inverse, plusieurs autres propriétés définitoires du LST sont des amorces pertinentes dans la mise en place d'une extraction automatique. Ayant défini le LST comme le lexique spécifique au genre de l'écrit scientifique et

commun à un ensemble de disciplines des SHS, nous devons définir les critères statistiques vérifiant ces propriétés que sont la spécificité et la transdisciplinarité.

2.2.2.1.1 Fréquence et spécificité

La spécificité se calcule en prenant appui sur le corpus de contraste. Le LST étant spécifique aux écrits scientifiques, sa fréquence relative[23] doit y être supérieure par rapport au corpus de contraste. Cette sur-représentation d'un élément peut être calculée à l'aide de différentes mesures : LLR[24], ratio, khi², log-odds, etc. Comme détaillé dans la section 1.2.2.3, plusieurs travaux ont mis en évidence les effets particuliers de certaines mesures. De façon générale, les mesures habituellement utilisées ne prennent pas en compte la distribution dans le corpus, comme le notent Paquot & Bestgen (2009). Pourtant, ils notent également, à l'instar de Simpson-Vlach & Ellis (2010), que l'extraction du lexique académique doit prendre en considération la répartition des occurrences dans le corpus, ne se limitant ainsi pas au seul critère de fréquence brute ou relative.

Nous avons par ailleurs montré les limites d'un traitement automatique reposant uniquement sur ces mesures de spécificité (Hatier, 2013). Après avoir comparé les éléments extraits par trois différentes mesures statistiques (LLR, ratio, khi²), nous avons conclu que le critère de spécificité n'est pas suffisant pour valider ou non l'appartenance d'un élément au LST.

En effet, certains mots ayant un haut score de spécificité renvoient à la langue générale (*flux, utilité, échelle, contrainte*) ou au lexique terminologique (*coût, management, usager*) alors que d'autres mots au score de spécificité inférieur appartiennent effectivement au LST (*argument, méthode*).

Bien que ce critère de spécificité ne soit pas un indice suffisant, il n'en demeure pas moins important et permet d'extraire des unités lexicales reflétant la nature des textes, selon les mots de Scott & Tribble (2006, p. 55-56). Ces derniers observent également que, parmi les mots-clés extraits à l'aide du critère de spécificité, deux groupes peuvent être distingués : les indicateurs thématiques et les indicateurs stylistiques[25]. Les premiers renvoient au sujet des textes du corpus d'analyse et, dans notre cas, correspondent à la terminologie disciplinaire :

[23] Par fréquence relative, nous entendons le nombre d'occurrences par millions de mots.
[24] Log-Likelihood Ratio ou rapport de vraisemblance.
[25] « aboutness indicators and stylistic indicators » (Scott & Tribble, 2006, p. 55).

commerce, politique, entreprise, revenu, département sont ainsi extraits de par leur spécificité par rapport au corpus de contraste. Le LST n'inclut pas les éléments de ce premier groupe mais correspond en revanche au second groupe, qui renvoie aux unités lexicales typiques d'un ensemble de textes, appartenant en l'occurrence à l'écrit scientifique : *fonctionnement, méthode, concept, analyse, contexte, catégorie* sont alors identifiés de par leur spécificité et leur répartition. S'il est abusif de réduire le LST à un indicateur stylistique, il reste vrai qu'il renvoie davantage à la structuration du discours et à l'argumentation qu'il ne réfère aux thématiques abordées dans les articles.

Le fait de disposer d'un corpus de contraste nous permet de ne pas nous baser uniquement sur la fréquence absolue des unités lexicales, à la différence de travaux précédents (Coxhead, 2002). Nous retenons, comme calcul de spécificité, le ratio de fréquences (fréquence relative dans le corpus d'analyse / fréquence relative dans le corpus de contraste). Nous avons vu dans de précédents travaux que les différentes mesures (tel le ratio ou le rapport de vraisemblance) donnent des résultats similaires dans l'extraction du LST (Hatier, 2013). Nous posons le seuil de 1 pour ce ratio ce qui revient à extraire les mots dont la fréquence relative dans le corpus d'analyse est au moins égale à la fréquence relative dans le corpus de contraste. Ainsi, le nom *variable* a une fréquence relative de 268 occurrences par million de mots dans le corpus d'analyse contre 2 occurrences par million dans le corpus de contraste. Son ratio de fréquence de 134 valide ainsi le critère de spécificité. Parmi les candidats-LST validant ce critère, se trouvent ainsi des éléments très spécifiques (au ratio supérieur à 100) – *catégorisation, organisationnel, typologie, normatif, corréler* – et des éléments moins spécifiques (au ratio inférieur à 2), bien que surreprésentés : *échec, constater, idéal, classement, opération*.

À l'inverse, les mots sous-représentés dans le corpus d'analyse ne sont pas retenus comme candidats-LST. Par exemple, le nom *journaliste,* bien que fréquent et réparti, est filtré en raison de son ratio de fréquence inférieur à 1. Sa fréquence relative dans le corpus de contraste, 78 occurrences par million de mots, est supérieure à celle du corpus d'analyse, 60.

De plus, afin de s'assurer que la spécificité positive d'un mot ne soit pas due à des occurrences particulières à une discipline, ou à un thème développé dans un

article ou un numéro de revue, nous posons un critère statistique supplémentaire, combinant spécificité et répartition. Nous calculons pour cela la spécificité par corpus disciplinaire[26] et imposons à un candidat LST d'avoir un score de spécificité positif dans un minimum de 3 disciplines.

2.2.2.1.2 Transdisciplinarité et répartition

À ce critère de spécificité, s'ajoute celui de transdisciplinarité. Ayant défini le LST comme un lexique partagé par plusieurs disciplines, nous transposons cette propriété par un critère de répartition dans le corpus d'analyse multidisciplinaire, afin de ne pas intégrer des mots spécifiques à un domaine, renvoyant à la terminologie disciplinaire ou plus largement à la thématique des articles. Ainsi, Hyland & Tse (2007) remarquent que l'*Academic Word List* (*AWL*) inclut des mots non équitablement répartis dans les sous-corpus académiques, et ce faisant ne correspondant pas à notre définition du LST. Afin de garantir le mieux possible le critère de transdisciplinarité, nous associons au critère de spécificité (ratio de fréquence positif) dans au moins 3 corpus disciplinaires le critère de répartition dans le corpus d'analyse.

Le LST, étant transdisciplinaire et traversant, doit être présent dans la majorité des disciplines constituant notre corpus. Nous segmentons notre corpus en 100 tranches égales d'environ 500 000 mots. Nous posons plusieurs seuils de répartition après observation des listes extraites selon les valeurs choisies. Une unité lexicale valide ainsi le critère de transdisciplinarité lorsque ses occurrences recouvrent un minimum de 40 tranches sur 100, correspondant à un minimum de 4 disciplines différentes. Le but est alors d'aboutir à une liste de candidats-LST réduisant autant que possible bruit et silence.

2.2.2.1.3 Limites des critères statistiques

La combinaison des critères retenus a notamment pour objectif de ne pas filtrer les mots ayant une acception terminologique (spécifique mais non traversante) et une acception appartenant à la langue générale (traversante mais non spécifique). Le genre de l'écrit scientifique étant caractérisé par un ensemble de lexiques aux frontières floues (voir section 1.2.2), nous isolons ainsi le LST en

[26] Par exemple, la spécificité pour le corpus de sociologie est le résultat du ratio de la fréquence relative dans le corpus sociologie sur la fréquence relative dans le corpus de contraste

éliminant terminologie et langue générale. Nous verrons cependant qu'un lexique particulier, le lexique abstrait général, répond positivement aux critères statistiques définis, et nous permet d'élargir ainsi notre définition du LST.

Ce lexique, non spécifique au genre de l'écrit scientifique, y est fréquemment mobilisé, au même titre que dans le genre journalistique. Cependant, à la différence du LST, il ne prend pas d'acception particulière dans le contexte de l'écrit scientifique. Nous pouvons donner, pour plusieurs catégories, quelques exemples de ce lexique, que nous pouvons rapprocher du *sub-technical vocabulary* de Cowan (voir section 1.2.2.1) :

- verbes : *varier, conserver, contenir, déboucher, ressortir, élargir, modifier* ;
- noms : *tendance, domaine, aspect, manière, durée, situation, problème* ;
- adjectifs : *important, favorable, futur, précédent, élevé, difficile*.

Nous reviendrons dans la suite de ce chapitre plus en détail sur ce lexique particulier, difficile à distinguer du LST au niveau statistique.

Puisque l'ensemble des critères d'extraction du LST repose sur les calculs de fréquence d'occurrences, nous avons souhaité affiner au maximum cette mesure. Ainsi, certains mots simples sont fréquemment intégrés à une expression polylexicale figée tel *point* dans *point de vue*. Afin de ne prendre en compte que les occurrences autonomes de ces mots simples, nous soustrayons à la fréquence totale d'un mot son nombre d'occurrences dans des segments répétés non ambigus. Pour ce faire, nous mettons en place une simple extraction de segments répétés, au seuil minimal de 10 occurrences, et évaluons manuellement l'aspect non ambigu du segment. Ce traitement ne prend cependant pas en compte certaines expressions polylexicales lorsqu'elles autorisent l'insertion d'un élément, comme *mettre en place* dans les exemples suivants :

- *Dès le XVIe siècle, l'Europe met progressivement en place une division du travail à l'échelle mondiale.*[27]

[27] Grasland, C., & Van Hamme, G. (2010). La relocalisation des activités industrielles : une approche centre-périphérie des dynamiques mondiale et européenne. *L'espace géographique*, *39*(1), 1-19.

- *Le professeur met ensuite en place les premiers éléments du jeu.*[28]

L'intérêt de ces unités supérieures, telles que les collocations, n'est pas remis en cause mais leur traitement n'entre pas dans le cadre de cette étude et est abordé par ailleurs dans d'autres travaux (Tutin, Tran, Kraif, & Hatier, 2015). De plus, il nous paraît complexe, dans la limite de notre sujet, d'identifier, pour l'ensemble des mots simples du LST, leurs occurrences correspondant à une collocation, ou plus généralement à une expression polylexicale. Ainsi, nombre de mots fréquents du LST (par exemple les noms *place, sens, compte, lieu*) sont mobilisés dans le corpus de façon « autonome » et dans des expressions verbales figées (*faire sens, mettre en place, prendre en compte, donner lieu à*). La prise en compte de ces phénomènes concernant chaque candidat-LST impliquerait des traitements manuels trop importants.

La combinaison de ces différents critères, à la base de l'extraction du LST, peut se résumer comme suit :

- spécificité positive globale : en comparaison du corpus de contraste diversifié ;
- spécificité positive dans au moins 4 disciplines ;
- dispersion : présence dans un minimum de 40 tranches sur 100 dans le corpus d'analyse ;
- occurrences en tant qu'unité monolexicale : nous retranchons à la fréquence absolue de chaque candidat-LST son nombre d'occurrences dans des expressions polylexicales contiguës, pour filtrer les éléments non autonomes, seulement intégrés dans ce type d'expressions.

Suite à l'application de ces critères, nous obtenons une liste de 1 976 candidats-LST, qui se répartit de la façon suivante :

- 513 adjectifs : *qualitatif, spécifique, méthodologique, judiciaire, urbain, marchand* ;
- 213 adverbes : *ibid., significativement, explicitement, autrement* ;

[28] Forest, D. (2008). Agencements didactiques : pour une analyse fonctionnelle du comportement non-verbal du professeur. *Revue française de pédagogie. Recherches en éducation*, (165), 77-89.

- 786 noms : *concept, auteur, méthode, nord, ministère, religion* ;
- 464 verbes : *définir, raisonner, questionner, adhérer, gouverner, recruter.*

Bien que la majorité des unités lexicales extraites correspondent effectivement à notre définition du LST, nous avons observé qu'un certain nombre d'éléments répondaient aux critères de spécificité et de répartition sans toutefois appartenir au LST. L'extraction est bruitée principalement par cet ensemble de mots renvoyant aux objets d'étude des SHS, que l'on retrouve dans plusieurs catégories :

- les verbes : *négocier, institutionnaliser, gouverner, financer, recruter, fréquenter* ;
- les noms : *école, ouvrier, ménage, politique, industrie, santé, parent, démocratie, religion* ;
- les adjectifs : *judiciaire, urbain, marchand, commercial, financier, artistique, gouvernemental, occidental*.

Ceci s'explique par le fait que ces mots renvoient à des thématiques potentiellement présentes dans la plupart des disciplines de notre corpus donc transdisciplinaires. Ce lexique relève plus de la terminologie que du LST, dont la nature est davantage métascientifique et métadiscursive.

De plus, en comparaison du corpus de contraste (composé d'oral, d'articles de journaux et de fiction), ces mots ont une fréquence élevée puisqu'ils sont fortement mobilisés pour la désignation des objets d'étude dans les différentes disciplines composant notre corpus d'articles en SHS. Le critère de spécificité permet cependant d'écarter un certain nombre de ces noms (*assemblée, syndicat, février, anglais, paix, combat, mariage, police*), et de minimiser le bruit dans la liste des candidats-LST.

Il apparaît donc impossible de circonscrire entièrement automatiquement le LST. C'est pourquoi nous avons choisi de recourir, comme dans d'autres études (Pecman, 2004b), à des experts humains, spécialistes du domaine, pour valider l'appartenance au LST des candidats. Paquot (2010, p. 63) note d'ailleurs, à la suite de l'élaboration de l'*Academic Keywords List,* que cette ressource, constituée de mots appartenant potentiellement à *l'Academic Vocabulary*, nécessiterait une

validation, à la vue du bruit et du silence provoqués par les traitements automatiques. Nous illustrons ci-dessous ces deux phases, automatique et manuelle, qui aboutissent à l'identification des mots simples du LST.

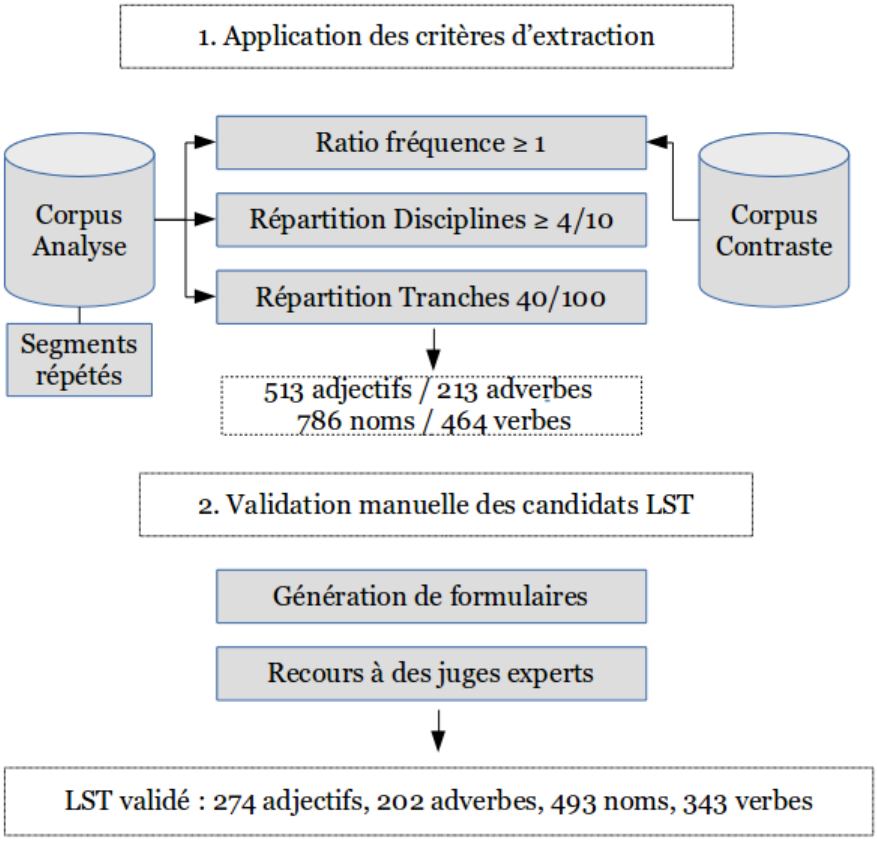


Illustration 2.5 :Extraction des mots simples du LST

La seconde étape manuelle permet également d'identifier les erreurs d'annotation automatique, liées par exemple à une mauvaise catégorisation ou lemmatisation. Ainsi, le verbe *contrer*, qui répond aux critères statistiques du LST, correspond pour un grand nombre d'occurrences à une erreur d'analyse automatique comme le montrent les exemples suivants, dans lesquels l'analyseur annote ces formes en tant que verbe *contrer*.

- *De la même façon, les familles nobles formaient des réseaux transfrontaliers, la noblesse de Flandre se mélangeant ainsi aux élites nobles des autres contrées.*[29]
- *« Voici venu votre tour d'échanger l'héritage contre un peu de vous-mêmes », annonce le journal de l'armée Bled.*[30]

[29] Buylaert, F. (2010). Les anciens Pays-Bas : nouvelles approches. *Revue historique*, (1), 3-25.
[30] Bantigny, L. (2007). Temps, âge et génération à l'épreuve de la guerre : La mémoire, l'histoire, l'oubli des appelés en Algérie. *Revue historique*, (1), 165-179.

Nous présentons dans la partie suivante cette étape de validation manuelle des résultats des traitements automatiques de la première phase.

2.2.2.2 Validation manuelle des candidats-LST

Les résultats de l'extraction automatique se révélant trop bruités, nous avons mis en place un protocole de validation manuelle des candidats LST. Un jugement humain est nécessaire pour confirmer l'appartenance au LST des candidats. Paquot & Bestgen (2009) proposent dans leurs travaux que les mots soient évalués par des formateurs amenés à enseigner le Français sur Objectif Universitaire, afin qu'ils jugent de leur utilité dans cet enseignement. L'appartenance d'un mot au LST peut donc être évaluée à la mesure de son apport dans diverses applications.

Comme nous l'avons détaillé dans la section 1.1, les perspectives d'utilisation de notre ressource du LST sont diverses : aide à l'indexation terminologique, aide à la rédaction scientifique, description linguistique. Étant donné que notre définition du LST s'adapte à l'ensemble de ces perspectives, la validation manuelle a pour but de s'assurer, pour chaque candidat-LST, qu'il correspond en effet à cette définition. Si nous revenons sur l'exemple des objets d'étude des SHS, nous pouvons constater que cet ensemble lexical ne correspond pas à notre définition du LST, et conséquemment, ne répond pas à nos besoins. Ainsi, le fait de considérer les objets d'études comme élément du LST aurait probablement pour conséquence, en extraction terminologique, de générer du silence en ne sélectionnant pas ces mots renvoyant aux thématiques. L'intérêt de ce lexique nous semble également limité dans le cadre de l'enseignement/apprentissage du français pour l'écrit scientifique. Ces mots renvoyant aux objets disciplinaires communs, bien qu'essentiels à maîtriser pour tout apprenant du domaine, relèvent plus de la terminologie que du LST et présentent un intérêt moindre lorsque l'on s'intéresse à ce lexique partagé par l'ensemble des disciplines.

Nous choisissons de confronter les résultats au jugement d'experts du domaine de la linguistique pour les trois catégories des noms, adjectifs et adverbes. Nous avons en effet pu observer que la catégorie adverbiale n'offre pas les mêmes difficultés. Ainsi, le bruit généré par les objets d'études des SHS ou certains éléments terminologiques ne concernent pas les adverbes, qui ne réfèrent pas aux

objets ou aux procédures de l'activité scientifique mais ont pour principale fonction de structurer le discours et de modaliser ce qui est dit.

Pour faciliter le processus d'évaluation pour les experts, nous avons mis en place l'élaboration automatique d'un formulaire intégrant les données issues du corpus pour chaque candidat : cooccurrents syntaxiques les plus fréquents, exemples phrastiques issus de différentes disciplines, etc. Les formulaires nous permettent dans un premier temps de proposer aux juges un profil des occurrences pour chaque mot, afin que l'évaluation de l'appartenance au LST soit fondée sur l'usage en contexte de chaque unité lexicale. Ces formulaires sont ensuite comparés afin d'identifier les éléments collectivement validés ou invalidés par l'ensemble des experts.

2.2.2.2.1 Jugement et recontextualisation

La majorité des travaux sur le lexique scientifique n'intègrent pas de phase de validation manuelle des résultats des méthodes d'extraction. Dans la mesure où nous prévoyons des applications, didactiques et linguistiques, utilisant cette ressource du LST, il nous apparaît pourtant essentiel de procéder à une telle étape, afin d'assurer la pertinence des unités lexicales retenues.

Paquot (2010) s'appuie dans ses travaux sur un logiciel d'étiquetage sémantique[31] pour valider les éléments extraits ; les mots retenus sont ceux répondant positivement aux différents seuils et appartenant à une des 6 classes sémantiques suivantes : *termes généraux et abstraits, nombres et mesures, action/état/processus psychologique, action/état/processus social et langage/communication*). Paquot identifie également les mots n'ayant pas répondu totalement aux critères statistiques mais appartenant à l'une de ces catégories et « rattrape » ainsi 330 éléments, soit plus de 30 % des unités finalement retenues. Ce type de traitement implique ainsi un étiquetage automatique en catégorie sémantique des candidats-LST du corpus, pour chacune de leur occurrence.

Ne disposant pas d'une ressource pour l'annotation sémantique automatique en français, nous choisissons, dans notre cas, de valider manuellement les éléments extraits. De plus, comme nous le verrons section 3.3.1.2, le fait d'avoir à

[31] Le système *UCREL Semantic Analysis System* (Rayson, Archer, Piao, & McEnery, 2004).

disposition une telle ressource ne réglerait pas les problèmes posés par l'annotation sémantique automatique d'éléments potentiellement polysémiques.

Nous optons donc pour une évaluation du statut LST des candidats basée sur leurs propriétés linguistiques, lexico-syntaxiques et sémantiques. Ces informations sur les propriétés distributionnelles des éléments du LST permettent alors un accès au sens mobilisé, en usage dans le corpus d'analyse, et donc de pouvoir avoir une représentation du sens en fonction du corpus spécialisé que nous avons constitué.

Les évaluateurs, chercheurs en linguistique, et sensibilisés à la problématique du LST à travers plusieurs de leurs travaux, doivent dans un premier temps confirmer ou infirmer l'appartenance des unités lexicales au LST. L'évaluation prend ici pour objet une unité formelle représentée par un lemme et une catégorie, tel le nom *terme*. Les juges doivent confirmer l'existence d'au moins une acception transdisciplinaire pour chaque mot. La potentielle polysémie transdisciplinaire sera gérée dans l'étape suivante des traitements du LST, présenté dans le chapitre 3. Ainsi, lors de l'évaluation, la question est de savoir si l'une des acceptions du nom *terme* mobilisée dans le corpus relève effectivement du LST, et non pas d'identifier l'ensemble de ses acceptions. Ce travail d'identification des acceptions, fait par la suite, nous révélera par exemple que *terme* renvoie à 2 acceptions du LST : sens 1 de 'mot, vocable', sens 2 de 'fin, limite'.

Nous proposons aux juges, en cas de non-appartenance au LST, d'évaluer l'appartenance au lexique abstrait général (LAG, présenté par ailleurs section 1.2.2.1). Ce lexique (*année, changer, fin*), présent dans l'écrit scientifique, peut être considéré comme moins spécifique à ce genre puisqu'il apparaît également dans les genres journalistique et littéraire notamment. Les résultats de l'extraction reposant sur l'analyse syntaxique automatique du corpus, nous proposons également aux évaluateurs de signaler toute erreur d'analyse (de catégorisation, de segmentation ou de lemmatisation).

L'évaluation du statut LST des candidats est aidée par une visualisation de leur emploi en contexte. Ainsi, pour permettre un accès aux acceptions mobilisées dans le corpus, le formulaire intègre des informations sur les associations lexico-syntaxiques les plus fréquentes et propose des exemples phrastiques représentant le maximum de disciplines différentes. Cette variété dans les exemples permet de

s’assurer d’un usage transdisciplinaire, en contexte, du mot candidat. Nous pouvons observer dans l’illustration suivante une page type du formulaire pour les noms.

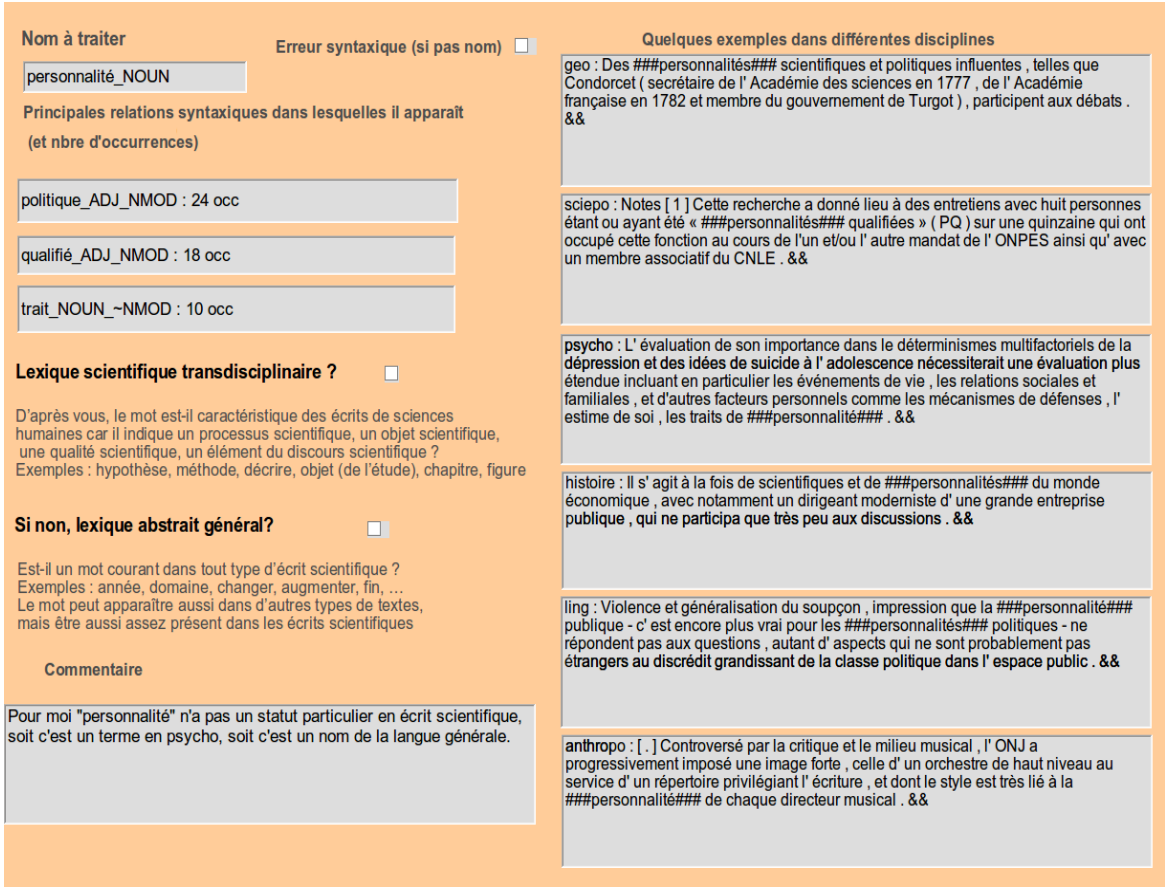


Illustration 2.6: Formulaire d'évaluation pour les noms

L'évaluateur a ainsi accès aux consignes détaillées de la tâche, ainsi qu'aux informations de recontextualisation. Les consignes ont pour but de guider les juges dans leur évaluation, ainsi que de clarifier le type d'informations qui leur sont présentées. Nous reproduisons ci-dessous, l'ensemble de ces consignes pour l'évaluation des différentes catégories.

3 cas sont possibles :

1. Le mot à traiter appartient au LST (LST coché) ;
2. Le mot à traiter appartient au LAG (LAG coché) ;
3. Le mot à traiter n'appartient ni au LST ni au LAG (aucune case cochée, catégorie générique).

Les informations concernant les relations syntaxiques et les exemples phrastiques ont pour but l'aide à la décision en permettant une recontextualisation du mot.

Les autres champs sont :

- -Erreur syntaxique : à cocher si le mot n'appartient pas à la catégorie à traiter (permet d'identifier les erreurs de catégorisation de l'analyseur) ;
- -Principales relations : sont listées n associations lexico-syntaxiques les plus productives dans notre corpus. Ces relations s'interprètent de la façon suivante :
  - Le formalisme est du type : *obtenir_verb_nmod* : 177 occ → il y a 177 occurrences dans notre corpus d'une relation de type *nmod* entre le nom à traiter et le lemme-catégorie obtenir_verb.
  - Si la relation (ex : ~*subj*) est précédée du caractère ~, le mot à traiter est régi par le cooccurrent dans cette relation. Si la relation est du type *subj*, le mot à traiter est recteur du cooccurrent dans cette relation.
  - Les relations sont : *nmod* (modifieur de nom), *vmod* (modifieur de verbe), *obj* (complément d'objet du verbe) et *subj* (sujet du verbe).
- Exemples disciplinaires : n exemples phrastiques sont proposés (précédés de la discipline dans laquelle ils apparaissent), issus de disciplines différentes, afin d'avoir des exemples en usage transdisciplinaire du mot à traiter.
- Commentaire : la case commentaire permet de noter vos remarques si besoin.

Cas complexes :

- Même si parmi les exemples, l'acception LST ou LAG est minoritaire (i.e. 4 exemples sur 6 mettent en évidence un usage non pertinent alors que les deux autres mettent en avant un usage de type LST), on valide le candidat comme LST sachant que nous devrons traiter ultérieurement la désambiguïsation afin de filtrer les acceptions qui nous intéressent.
- Si les exemples ne concernent que des expressions figées, on ne valide pas l'appartenance.
  - S'il existe au moins un exemple avec le candidat hors locution, on peut le traiter en tant que mot simple.
- Le lexique des objets d'étude des SHS, lié aux activités humaines, aux sociétés, (du type *ouvrier, société, pays, collectivité*) ne doit pas être inclus dans le LST ou le LAG

Le lexique qui nous intéresse porte (comme défini dans le formulaire) sur le discours scientifique (processus, qualité, observable de l'activité scientifique). Il intègre également le lexique méta-discursif et méta-textuel.

Illustration 2.7: Consignes pour l'évaluation du LST

Dans un premier temps, est donnée une rapide description de la tâche, qui est de discriminer trois ensembles lexicaux, le LST, le LAG, et un dernier ensemble défini par exclusion de deux précédents. Le formalisme permettant d'interpréter les propriétés lexico-syntaxiques est ensuite détaillé dans un deuxième temps. Pour conclure les consignes, plusieurs cas complexes sont abordés, afin de guider au mieux les juges dans la gestion de la polysémie et du bruit généré par les objets d'étude des SHS.

Les différents pages de formulaire[32] (une par candidat LST) sont générées automatiquement à partir des données lexicométriques (fréquence, relations, exemples) extraites à l'aide de scripts *Perl*. Chaque évaluateur dispose ainsi d'informations issues du corpus pour juger de l'appartenance des candidats au LST ou au LAG. Nous avions également prévu un champ *commentaire* libre pour garder une trace des cas problématiques ou ambigus.

Ces formulaires sont ainsi au centre de l'évaluation des candidats LST, que ce soit pour la catégorie nominale, traitée par cinq juges, ou pour les verbes et les adjectifs, évalués par trois juges.

Les noms ayant été traité au moment de l'évaluation des adjectifs, nous avons intégré dans les formulaires concernant la catégorie adjectivale une information supplémentaire sur la proportion de cooccurrents nominaux appartenant au LST, tel que le montre l'illustration 2.8.

[32] Nous remercions Marie-Paule Jacques pour son aide dans l'élaboration des formulaires à partir des bases de données lexicométriques.

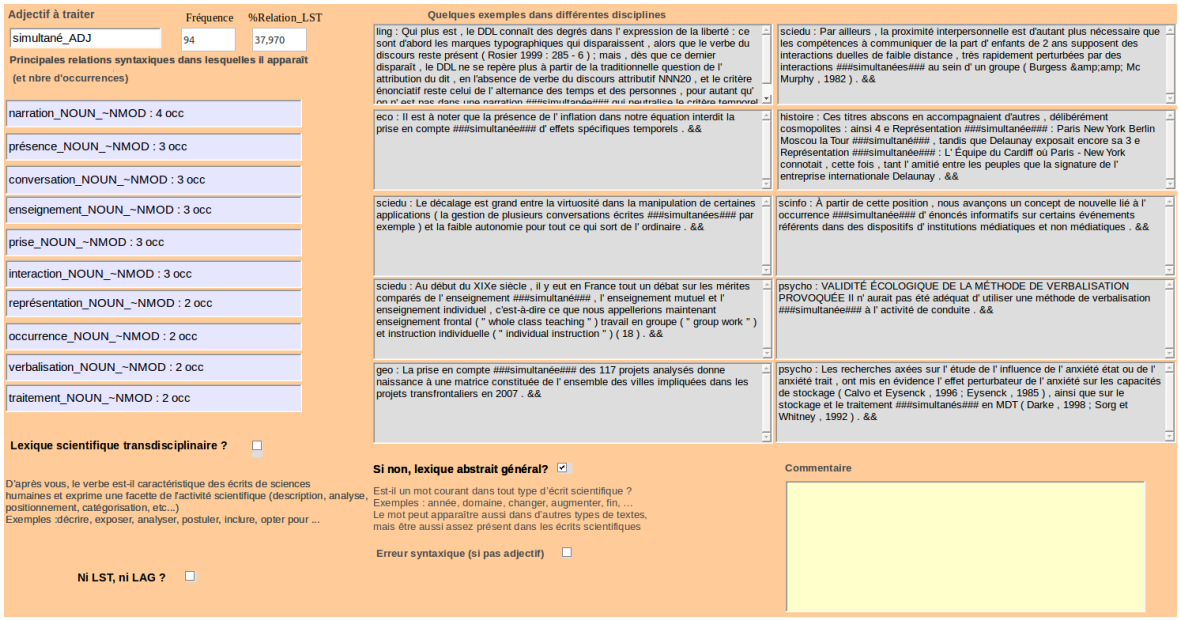


Illustration 2.8: Formulaire d'évaluation pour les adjectifs

Nous enrichissons ainsi les formulaires, au fur et à mesure de l'évaluation des différentes catégories, afin de proposer aux juges les informations les plus à même de représenter les usages dans le corpus d'analyse.

Dans l'exemple de formulaire illustré ci-dessus, diverses informations permettent une recontextualisation des emplois de l'adjectif *simultané*. Ce candidat-LST, 94 occurrences dans le corpus d'analyse, a dans 37 % de ses occurrences une relation avec un nom du LST, tel *traitement, interaction, apparition, évolution*.

Le formulaire révèle également que plus de 89 % des cooccurrents nominaux de l'adjectif *crucial* ont été validés comme LST (*rôle, question, élément, point, importance, étape,* etc.). Nous avons plus généralement observé que les éléments du LST entrent fréquemment en cooccurrence, ce type d'association étant alors un indice supplémentaire du statut LST d'un candidat.

Cette information sur les cooccurrents des candidats à évaluer a pour but de faciliter la tâche des juges qui est de valider ou non la présence d'une acception transdisciplinaire du mot en question (qui peut renvoyer à d'autres acceptions, ne relevant pas du LST). Dans le même objectif, nous augmentons le nombre

d'exemples de relations syntaxiques et de phrases après discussion avec les évaluateurs. Le travail de constitution du LST est donc ici envisagé de façon incrémentale, en tirant parti des résultats des traitements précédents.

En outre, nous ajoutons dans le formulaire des verbes des champs permettant de signaler l'emploi spécifique du candidat dans une forme pronominale ou passive afin d'intégrer ces informations dans notre ressource du LST. Ainsi, le verbe *avérer* ne se trouve que dans un emploi pronominal dans notre corpus, alors que le verbe *dédier* n'est réalisé qu'à la voix passive, comme nous pouvons l'observer dans les exemples suivants.

- *Cela a des implications économiques qui peuvent s'avérer considérables*.[33]
- *Ce dernier sera rapidement dédié à la création circassienne contemporaine*.[34]

En se basant sur les résultats de ces formulaires remplis par l'ensemble des juges[35], nous pouvons procéder à l'étape de validation manuelle du LST pour l'ensemble des candidats des trois catégories concernées : 786 noms, 513 adjectifs, 464 verbes.

Nous présentons dans la partie suivante les résultats de l'évaluation en procédant à une analyse quantitative et qualitative de la validation des formulaires par les juges. Nous nous intéressons notamment à l'accord inter-annotateurs et aux critères pouvant expliquer des jugements hétérogènes.

2.2.2.2.2 Accords inter-annotateurs

Nous intégrons, à la différence des autres travaux sur le sujet, une analyse de l'accord inter-annotateurs dans l'évaluation du LST.

Avant d'aborder les scores d'accord inter-annotateurs, nous revenons dans un premier temps sur les commentaires (champ proposé dans les formulaires d'évaluation) notés par les juges. Ces traces permettent d'identifier les unités

---

[33] Wieviorka, M. (2008). L'intégration : un concept en difficulté. *Cahiers internationaux de sociologie*, (2), 221-240.

[34] Sizorn, M. (2008). Une ethnologue en « Trapézie » : sport, art ou spectacle ?. *Ethnologie française*, *38*(1), 79-88.

[35] Nous remercions Evelyne Jacquey, Laurence Kister, Agnès Tutin et Marie-Paule Jacques pour leur aide dans ce travail fastidieux.

lexicales ayant posé problème, et incidemment de faire émerger les difficultés récurrentes pour la tâche d'évaluation. Nous observons ainsi plusieurs commentaires convergents :

- des commentaires sur les mots trop « techniques, spécialisés, terminologiques », ne relevant pas du LST : *récit, enseignement, croissance, religion, lecteur, sexe, pays, travailleur* ;
- des commentaires sur les mots dont les occurrences s'intègrent dans des « expressions polylexicales, locutions » : *base, cause, point, mise, lieu, prise* ;
- des commentaires sur des erreurs d'analyse (segmentation, catégorisation, langue étrangère). Par exemple, les « noms » *multiple, double, court* se révèlent souvent être, après observation des concordances, des adjectifs mal catégorisés ;
- des commentaires sur la difficulté à discriminer le LST et le LAG : *observateur, cohérence, apparition, effectif* ;
- des commentaires sur la difficulté à identifier le sens pour les unités polysémiques : *figure, lien, loi, sujet, revue*.

Émergent dans ces commentaires les principales difficultés inhérentes à l'identification du LST de façon automatique. L'étape de validation manuelle est ainsi justifiée par la présence de ces erreurs syntaxiques, et la non pertinence de certains candidats, due à leur statut d'éléments d'expressions polylexicales, ou en raison de leur polysémie.

Nous avions mesuré, dans de précédents travaux, l'accord inter-annotateurs pour une tâche d'évaluation d'un sous-ensemble nominal et verbal du LST (Hatier, 2013). Trois juges devaient classer un ensemble de 200 mots dans un des lexiques

suivants : LST, lexique terminologique (LT), lexique de la langue générale (LG). Pour chaque mot à classer, les juges disposaient d'informations sur les deux principaux cooccurrents ainsi que d'un exemple extrait du corpus, comme l'illustre le tableau suivant.

Candidat	**LST**	**LT**	**LG**	**Cooccurrences**	**Contexte**
illustrer_verb	+	-	-	*exemple_nom#subj cas_n#subj*	*Cet exemple paradigmatique illustre le choix que nous faisons […]*[36]
synthèse_nom	+	-	-	*proposer_verb#~obj réaliser_verb#~obj*	*[…] la seconde propose une synthèse des principales interventions […]*[37]

Tableau 2.3: Grille d'évaluation de l'expérimentation 2013

Dans ce tableau est indiqué le fait que le verbe *illustrer* entre en cooccurrence dans la relation sujet (en tant que gouverneur) avec les noms *exemple* et *cas*. Le nom *synthèse* est lui dépendant (caractère ~) dans les relations données dans le tableau.

Les juges étaient en accord, pour cette expérimentation préalable, pour 68 des 100 verbes et 79 des 100 noms. Ces pourcentages d'accord pourraient être encore plus élevés, si ce n'était la présence des deux facteurs suivants :

- le LST est un lexique à frontière floue, inscrit dans un genre intégrant un ensemble de lexiques dont la délimitation est complexe. Il existe ainsi une grande variabilité quant à sa perception selon le juge ;
- les informations de recontextualisation, pour cette expérimentation, ne permettent pas un accès systématique aux différentes acceptions des candidats LST. Autrement dit, l'accès au sens se complexifie avec le nombre d'acceptions mobilisées par un même verbe et par le type d'alternances dans lesquelles il s'inscrit (voir notamment l'exemple du verbe *situer* dans l'illustration 2.9).

[36] Lassègue, J., Rosenthal, V., & Visetti, Y-M. (2009). Économie symbolique et phylogenèse du langage. L'Homme, 192, 67-100.

[37] Laurière, C. (2008). L'anthropologie et le politique, les prémisses. L'Homme, 187-188, 69-92

Comme nous l'avons illustré dans la partie précédente, des formulaires intégrant plus d'informations (cooccurrents syntaxiques, exemples phrastiques) ont été créés afin de pallier la limite identifiée ci-dessus. Nous avons également élargi le nombre de juges, pour en impliquer 5, pour la catégorie nominale, 3 pour les verbes et les adjectifs, afin de minimiser les effets de la subjectivité individuelle.

Nous procédons alors, en ce qui concerne cette phase de validation manuelle, à l'analyse des accords inter-annotateurs, en nous concentrant sur deux mesures principales :

- le pourcentage d'accord brut : proportion des candidats ayant reçu le même jugement par l'ensemble des évaluateurs ;
- le kappa de Fleiss[38] (1971): coefficient d'accord inter-annotateurs permettant des calculs sur plus de 2 juges, sur des données nominales, et tenant compte de données manquantes (en cas de non-décision d'un juge).

Le tableau ci-dessous donne les scores pour ces deux mesures pour les trois catégories des noms, verbes et adjectifs.

	LST	**LAG**	**GEN**
Noms	0,391 (82 %)	0,139 (57 %)	0,283 (68 %)
Verbes	0,228 (66 %)	0,052 (54 %)	0,253 (72 %)
Adjectifs	0,393 (79 %)	0,252 (68 %)	0,486 (74 %)

Tableau 2.4: Accords inter-annotateurs : kappa et %

Les 3 colonnes correspondent aux choix possibles pour chaque candidat :

- LST : le mot appartient au lexique scientifique transdisciplinaire ;
- LAG : le mot appartient au lexique abstrait général ;
- GEN : le mot n'appartient ni au LST ni au LAG.

Le premier score correspond au kappa de Fleiss, le score entre parenthèses est celui du pourcentage d'accord brut. Selon Landis & Koch (1977), la valeur du kappa de Fleiss s'interprète comme suit :

- inférieur à 0 → désaccord ;

[38] Calculé à l'aide de l'interface proposé par : Geertzen, J. (2012). Inter-Rater Agreement with multiple raters and variables. https://nlp-ml.io/jg/software/ira/ [consulté le 01/06/2016]

- [0,01- – 0,20] → accord très faible ;
- [0,21- – 0,40] → accord faible ;
- [0,41- – 0,60] → accord modéré ;
- [0,61- – 0,80] → accord fort ;
- [0,81- – 1] → accord presque parfait.

Nous pouvons ainsi observer que l'accord inter-annotateurs se révèle faible ou modéré pour la plupart des dimensions (LST, LAG et GEN) et des catégories. Nous remarquons cependant certaines tendances. Ainsi, la catégorie verbale s'avère la plus complexe à évaluer, ce qui n'est pas surprenant compte tenu de l'importance du contexte (et des arguments) dans l'analyse lexicale des verbes.

Considérons certains de ces verbes pour lesquels les trois juges ont fait une évaluation différente (LAG, LST, ou aucun des deux) : *intégrer, intervenir, dépasser, disperser, affranchir, interroger*. La majorité de ces verbes recouvre plusieurs acceptions et autorise les alternances, aux voix pronominale et passive. Si l'on se reporte à la ressource du *LVF*[39], nous constatons cette forte polysémie : 7 entrées pour *intégrer* et *intervenir*, 12 pour *dépasser*, ou encore 8 pour *affranchir*. Il en résulte, pour les juges, un difficile accès au(x) acception(s) mobilisée(s) pour chaque verbe. Malgré les informations de recontextualisation, les évaluateurs peinent quelquefois à circonscrire les sens à juger, d'autant plus que le verbe en question est polysémique.

L'apport des informations lexico-syntaxiques est cependant réel. Nous observons ainsi que certains verbes, tel *situer*, bien que génériques et fortement polysémiques (5 entrées de *situer* dans le *LVF*), ont été évalués pareillement par l'ensemble des juges[40]. Le formulaire se révèle, notamment pour ces exemples, efficace dans l'aide à la recontextualisation, comme illustré ci-dessous.

---

[39] Le *LVF* (Dubois & Dubois-Charlier, 1997), ressource élaborée par les auteurs du *DEM,* a été utilisé comme ressource de référence pour le repérage des acceptions transdisciplinaires des verbes du LST par Yan (Hatier *et al.*, 2016).

[40] En l'occurrence, *situer* a été jugé comme appartenant au LAG par les 3 juges.

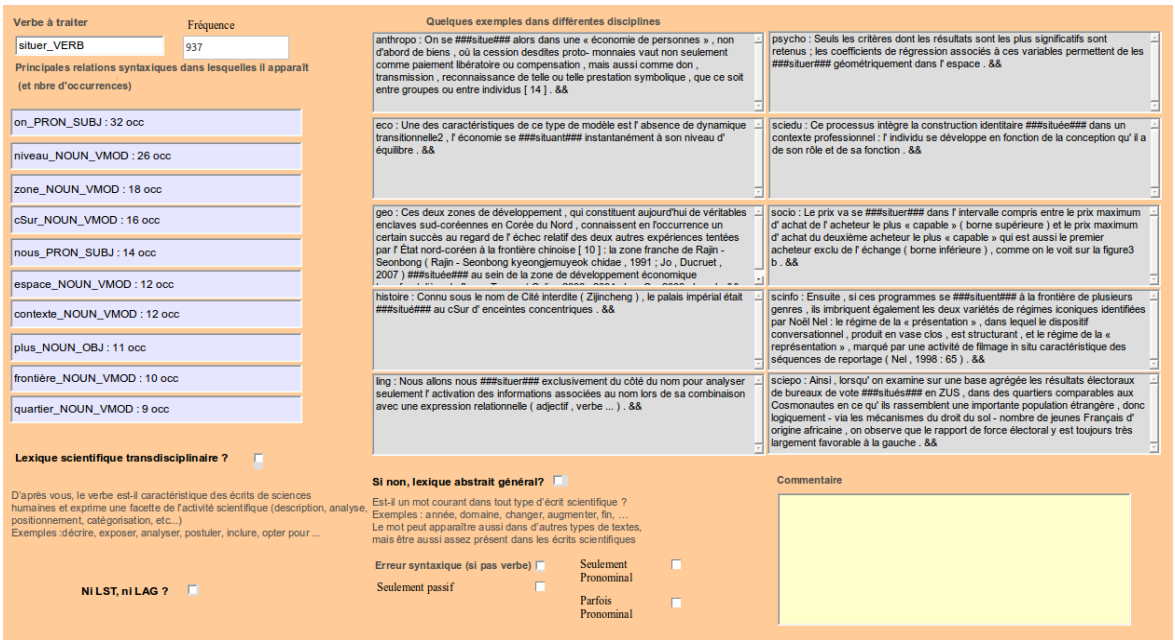


Illustration 2.9: Formulaire d'évaluation pour les verbes

Dans cet extrait de formulaire pour le verbe *situer,* nous constatons que malgré la polysémie à l'œuvre dans le corpus, les informations sur les relations fréquentes et les concordances permettent d'identifier les acceptions mobilisées. Les exemples autorisent notamment ici l'identification rapide du sens principal, employé à la forme pronominale au sens abstrait et concret de 'être localisé'. Les cooccurrences avec les pronoms sujets *on* et *nous* révèlent que le verbe *situer* est fréquemment employé dans un but de positionnement face aux pairs, à la littérature ou aux notions convoquées.

Une difficulté spécifique à cette catégorie est par ailleurs l'identification des verbes aux formes strictement pronominale, passive ou impersonnelle (voir les exemples avec les formes *s'avérer* ou *être dédié* section 2.2.2.2.1). Dans le cadre d'une utilisation de la ressource comme aide à la rédaction, il nous apparaît indispensable d'intégrer cette information sur les constructions verbales préférentielles, afin que l'appropriation de l'usage particulier des verbes soit la plus effective possible pour les apprenants.

De plus, comme nous le verrons dans la partie sur l'analyse des acceptions du LST (voir section 3.3.1.2), les verbes sont la catégorie pour laquelle la polysémie est la plus productive.

Nous observons également que le jugement d'appartenance au LAG est celui menant à l'accord le plus faible. La distinction entre LST et LAG se révèle complexe à juger. On observe en effet qu'un grand nombre de noms sont ainsi évalués comme LST par plus d'un évaluateur et LAG par le reste des juges. Les substantifs suivants ont par exemple tous été annotés LST ou LAG : *augmentation, étape, rang, biais, phase, classe, fonction, proposition, réponse, objectif, solution, outil, mesure, indice, cas, participant*.

Nous avons précédemment justifié l'intérêt de ce lexique abstrait général au niveau didactique : bien que ce lexique ne soit pas spécifique à l'écrit scientifique, il y est fréquemment mobilisé et conséquemment sa maîtrise se révèle indispensable dans les tâches de rédaction et de compréhension de textes scientifiques, comme le note Paquot (2010).

Dans l'optique de l'utilisation du LST dans les processus d'extraction terminologique, l'intégration du LAG au LST, alors employé comme filtre d'exclusion, se justifie tout autant. En effet, le LAG étant un lexique surreprésenté dans notre corpus, ses éléments peuvent être potentiellement identifiés comme mots-clés d'un texte scientifique. Le fait de combiner LAG et LST dans le cadre de tels processus ne peut ainsi qu'aider à la diminution du bruit dans l'extraction de mots-clés.

En plus de partager certaines propriétés lexicométriques dans notre corpus d'analyse (spécificité et répartition), ces deux lexiques ont en commun leur utilité dans la détection de mots-clés, en tant que filtre d'exclusion mais également en tant qu'indices de présence de terme. Considérons, pour illustrer cet intérêt du LAG, un échantillon de ces noms ayant été jugés LAG par la majorité des évaluateurs : *ensemble, défaut, présence, expression, mot, domaine, aspect, enjeu, rôle, situation, durée, texte, groupe, type*.

Plusieurs de ces noms ont été identifiés par ailleurs (Jacquey *et al.*, 2013) comme potentiels délimiteurs de termes – *groupe, ensemble, type, présence* – lorsqu'ils sont intégrés dans un patron du type Nom_{LAG} Prep $\text{Nom}_{\text{Terme}}$, comme illustré par les exemples suivants :

- *Vient ensuite un groupe*$_{LAG}$ *de*$_{PREP}$ *six provinces*$_{Terme}$ *de rang moyen.*[41]

[41] Quertamp F., (2010). La périurbanisation de Hanoi. Dynamiques de la transition urbaine vietnamienne et métropolisation, *Annales de géographie* 1/2010 (n° 671-672), p. 93-119.

- *La mobilisation du lien ethnique met en évidence trois types*$_{LAG}$ *de*$_{PREP}$ *solidarité*$_{Terme}$.[42]
- *[…] il pourrait provenir de la présence*$_{LAG}$ *d'*$_{PREP}$ *hétéroscédasticité*$_{Terme}$ *dans les données.*[43]

La complexité à différencier le LAG du LST et leurs intérêts similaires nous conduisent donc à une redéfinition du LST, dans lequel nous intégrons le LAG.

En analysant les résultats des formulaires, nous remarquons par ailleurs que pour chaque catégorie, la proportion de LST, LAG et GEN est stable pour l'ensemble des juges sauf un. L'illustration suivante met ainsi en évidence ces déséquilibres pour les noms.

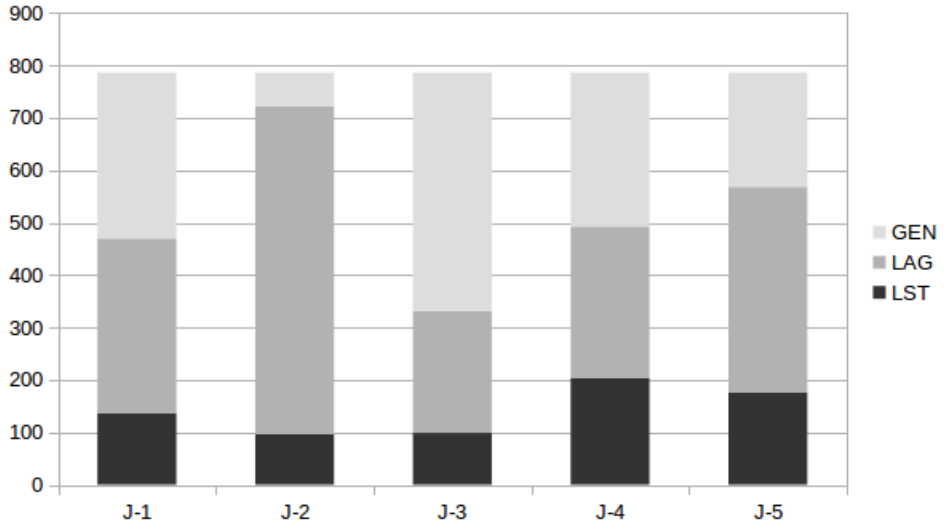


Illustration 2.10: Proportion des lexiques par évaluateurs

Nous constatons alors que l'évaluateur 2, de par la proportion de candidats validés comme LAG, se démarque de l'ensemble des autres évaluateurs.

Nous procédons alors à un nouveau calcul de l'accord inter-annotateurs en ignorant, pour chaque catégorie, les résultats de l'évaluateur le plus singulier. Nous aboutissons alors aux résultats résumés dans le tableau ci-dessous.

---

[42] Augé A., (2007) Les solidarités des élites politiques au Gabon : entre logique ethno-communautaire et réseaux sociaux, *Cahiers internationaux de sociologie* 2/2007 (n° 123), p. 245-268.

[43] Jondeau, É. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. *Annales d'Économie et de Statistique*, 139-174..

	LST	LAG	GEN
Noms	0,45 (83 %)	0,251 (64 %)	0,428 (72 %)
Verbes	0,234 (63 %)	0,081 (55 %)	0,337 (82 %)
Adjectifs	**0,755** (90 %)	0,615 (83 %)	**0,746** (88 %)

Tableau 2.5: Accords inter-annotateurs (n-1) : Kappa et %

Ces nouveaux calculs donnent des accords inter-annotateurs supérieurs, majoritairement modérés (soulignés) ou forts (en gras) selon l'échelle de Landis & Koch. Nous relevons des tendances similaires à celles observées dans le tableau 2.4. D'une part, les scores montrent que les dimensions LST et GEN sont moins complexes à évaluer que la dimension LAG. D'autre part, nous constatons que les catégories adjectivales puis nominales obtiennent des scores d'accord plus élevées que la catégorie verbale, confirmant la difficulté à juger du sens mobilisé par les verbes par le biais du formulaire, malgré l'apport des concordances et cooccurrences.

Ces résultats nous confortent dans la nécessité et l'utilité de l'étape d'évaluation manuelle des candidats LST. Nous retenons ainsi comme élément du LST :

- les noms validés LST ou LAG par au moins 3 des 5 juges experts ;
- les adjectifs validés LST ou LAG par au moins 2 des 3 juges experts ;
- les verbes validés LST ou LAG par au moins 2 des 3 juges experts ;
- 202 des 213 adverbes validés LST après filtrage des erreurs syntaxiques.

Nous aboutissons à une liste de mots simples du LST de 1311 mots pleins :

- 493 noms : *totalité, généralisation, fonction, démarche, composition*. Voir annexe A.II pour le détail.
- 274 adjectifs : *possible, singulier, ambigu, observable, homogène*. Voir annexe A.III pour le détail.
- 342 verbes : *importer, correspondre, juger, analyser, localiser*. Voir annexe A.IV pour le détail.

- 202 adverbes : *successivement, ibid., différemment, fondamentalement.* Voir annexe A.V pour le détail.

Nous n'avons pas mis en place un protocole défini pour gérer le silence produit par notre méthode. Le repêchage de certains mots n'ayant pas satisfait aux critères statistiques a été fait, pour les verbes et les adverbes, en croisant les résultats de l'extraction avec des listes déjà existantes, émanant de travaux précités (voir pour les verbes Yan (2012), et pour les adverbes Tran (2014)). Ainsi sont rattrapés les verbes – *amener, remplacer, rassembler, apporter, juger, progresser, succéder* – ou les adverbes *environ, alors, très*. Dans la perspective de la mise à jour de la ressource, nous pourrions automatiquement mettre en place une extraction d'une nouvelle liste afin d'identifier de potentiels éléments du LST passés sous silence, en modifiant notamment certains seuils, tels ceux de la spécificité ou de la répartition, afin de faire émerger de nouveaux candidats. Leur intégration finale dans la ressource nécessiterait toutefois d'effectuer à nouveau la phase de validation manuelle.

Nous pouvons comparer la composition catégorielle de notre LST en regard des listes de Paquot[44] (2010) et Drouin (2007). Le diagramme ci-dessous résume la proportion par catégorie pour les différents travaux.

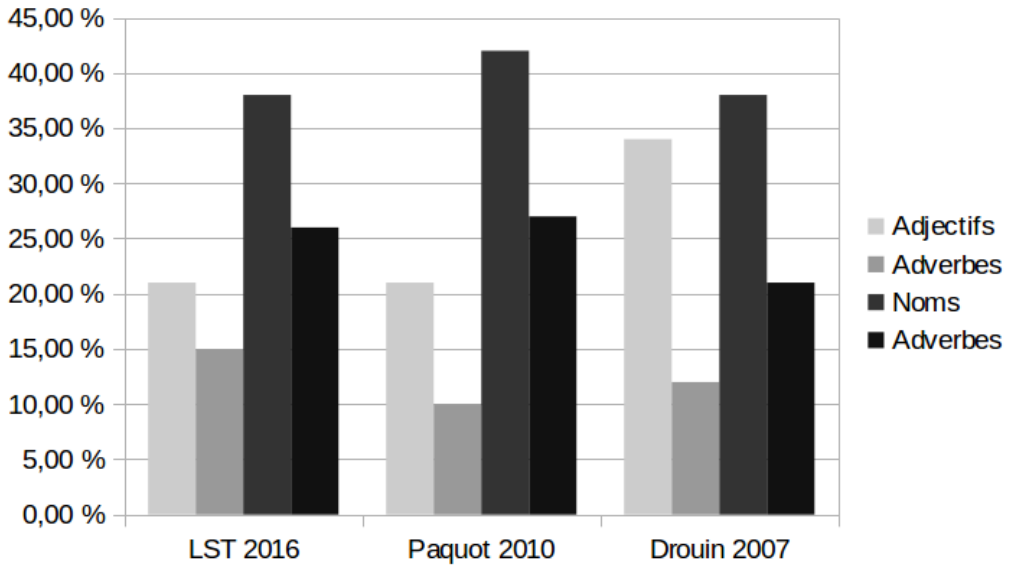


Illustration 2.11: Proportion des catégories du LST

[44] La proportion de noms pour Paquot est initialement de 38 %, mais ces travaux incluent d'autres catégories (préposition, déterminant, etc.) que nous ne prenons pas en compte.

Les noms représentent la catégorie la plus productive pour le LST pour l'ensemble des travaux, et ce dans une proportion autour de 40 %[45] du lexique extrait. Les noms sont donc la catégorie la plus fréquente dans les textes scientifiques, ce phénomène étant probablement accentué par la surreprésentation des nominalisations telle que relevée par Biber (2006) et Kocourek (1991).

Avant de présenter dans les chapitres suivants les traitements d'enrichissement sémantique et syntaxique, nous confrontons dans la suite de ce chapitre les résultats de notre extraction du LST à d'autres méthodes afin d'évaluer l'influence des corpus et des critères statistiques sur les éléments extraits.

2.2.3 Effets du corpus et des mesures dans l'extraction du LST

Nous présentons dans cette partie deux études de comparaison de notre liste du LST avec des listes résultant de méthodes différentes. Dans un premier temps, nous proposons une comparaison qualitative et quantitative entre notre liste du LST et celle obtenue par Drouin (2007), disponible en ligne[46]. Dans un deuxième temps, nous présentons les conclusions d'une expérimentation menée au sein du projet TermITH avec la collaboration de l'Institut de l'Information Scientifique et Technique (INIST[47]). Cette expérience avait pour but de confronter différents types de métriques et de corpus afin d'évaluer les méthodes pour l'identification du LST dans l'optique de l'extraction terminologique.

2.2.3.1 Méthode par répartition et spécificité

Comme nous l'avons précisé dans la section 1.2.2.4, Drouin a proposé une méthode d'extraction dont nous nous sommes inspiré. Il extrait une liste du LST, d'après un corpus de thèses dans différentes disciplines des SHS (psychologie, histoire, archéologie, droit, géographie) et d'autres domaines (informatique, chimie, physique, ingéniérie) de 2,3 millions de mots. Pour ce faire, il calcule la spécificité des mots (le corpus est analysé par *TreeTagger*) par comparaison avec un corpus de référence, constitué de l'ensemble des articles du journal *Le Monde* pour l'année 2002 (dont la taille est de 30 millions de mots). Il applique les

---

45 Il est à noter que Biber (2006) obtient également 38 % de noms dans son lexique académique.

46 À l'adresse suivante : http://olst.ling.umontreal.ca/lexitrans/nomenclature.php [Consulté le 10/01/2016]

47 Pour une présentation de l'INIST, voir: http://www.inist.fr/? lang=fr [Consulté le 10/01/2016]

critères de spécificité et de répartition dans le corpus pour identifier 1113 mots du LST des catégories pleines : adverbe, verbe, adjectif, nom.

Comme nous l'avons détaillé dans ce chapitre, nous reprenons plusieurs points de la méthode de Drouin et y ajoutons toutefois certains traitements spécifiques : la prise en compte des segments répétés, l'utilisation d'un analyseur syntaxique au lieu d'un étiqueteur morpho-syntaxique, et surtout une phase de validation manuelle avec prise en compte des profils combinatoires. Nous reprenons de sa méthode les critères de répartition et de spécificité ainsi que l'utilisation d'un corpus de contraste. Mais à la différence de Drouin qui utilise le corpus *Le Monde*, nous avons constitué un corpus de contraste non-uniforme de grande taille, estimant que la confrontation du genre scientifique à différents genres était nécessaire pour faire émerger le LST.

Plusieurs paramètres sont ainsi à prendre en compte dans l'analyse des différences entre le LST extrait par Drouin et les résultats de nos travaux :

- différences au niveau du corpus d'analyse : genre différent (articles *vs* thèses), disciplines et domaines différents (SHS pour notre expérimentation, corpus plus hétérogène pour Drouin qui inclut par exemple l'ingénierie et l'informatique) ;
- différences au niveau des métriques (ratio *vs* calcul de spécificité Lafon (1980)), de la référence pour la spécificité (nature du corpus de contraste) ;
- différences au niveau des traitements : type d'analyse de corpus (analyse en dépendance *vs* étiquetage morpho-syntaxique), prise en compte des segments répétés, analyse manuelle des candidats par observation de concordances et des propriétés lexico-syntaxiques, etc.

Nous comparons ainsi les 1312 éléments de notre LST avec les 1600 éléments du LST identifiés par Drouin[48]. Une première comparaison par catégorie des deux listes est résumée dans le tableau ci-dessous.

[48] Nous ne retenons pas les mots composés pour cette comparaison, pour que les différences de segmentation d'analyseur n'interfèrent pas dans notre comparaison des méthodes.

	LST-2016	Drouin-2007	Recoupement
Noms	493	610	348 (70 %)
Adjectifs	274	458	209 (76 %)
Verbes	343	334	233 (68 %)
Adverbes	202	198	104 (51 %)

Tableau 2.6: Comparaison LST-liste Drouin 2007

Le pourcentage entre parenthèses de la colonne *Recoupement* correspond à la proportion des éléments de notre liste du LST apparaissant également dans la liste de Drouin.

Nous pouvons constater que le recoupement entre listes est majoritaire pour l'ensemble des catégories. De plus, nous avons pu noter que parmi les 145 noms du LST présents dans notre ressource (LST-2016) mais absents de celle de Drouin (D-2007), 70 ont un dérivé morphologique dans cette liste (ainsi *implication* appartient à la liste LST-2016 et *impliquer* à la liste Drouin 2007). Si nous intégrions une phase dans laquelle nous « rattrapions » les dérivés morphologiques des mots du LST, à la manière de Coxhead (2002), il est probable que le recoupement entre les deux listes augmenterait considérablement.

L'analyse contrastive des résultats par catégorie nous apporte des informations supplémentaires. Ainsi, pour les noms, un grand nombre d'éléments de D-2007 et absents de LST-2016 renvoient aux objets d'études des articles de recherche, précédemment abordés dans la section 2.2.2.1.3 : *homme, eau, histoire, pays, famille, cour, ville.* Le traitement manuel que nous avons effectué a permis d'éliminer ces éléments, ne correspondant pas à notre définition du LST bien que répondant positivement à ses critères statistiques. Nous identifions également des éléments résultant d'expressions polylexicales segmentées (*main, œuvre, sein, tour*, *parti*) filtrés de notre ressource par le critère sur les segments répétés[49].

Au niveau des noms communs aux deux listes, nous constatons que l'intersection est composée de noms scientifiques généraux appartenant tant au LST (*traitement, modèle, section*) qu'au LAG (*seuil, série, rôle, manière, groupe*).

[49] Ce traitement a permis d'identifier notamment les expressions suivantes : *prendre en compte, à l'œuvre, mise au point, à leur tour, faire parti* (tel que lemmatisé dans notre corpus).

Si l'on s'intéresse aux noms uniquement présents dans notre liste, nous pouvons identifier certains noms spécifiques au domaine de notre corpus d'analyse (les SHS), plus restreint que le corpus de Drouin, qui intègre des disciplines des sciences exactes : *vocabulaire, paradigme, corpus, débat, observateur*. Un autre ensemble de noms, propre à notre liste, renvoie aux événements et acteurs de la recherche scientifique : *publication, conférence, colloque, université, chercheur*. Les types différents d'écrits des deux corpus d'analyse (articles vs thèses) expliquent ici ces dissemblances entre les deux listes. Enfin, certains éléments absents de LST-2016 (*exclusion, détail, emplacement, conservation*) et présents dans D-2007 correspondent toutefois à notre définition du LST. Leur absence dans notre liste peut s'expliquer par deux principaux paramètres, qui différent dans les deux méthodes.

D'une part, la composition des corpus d'analyse et de contraste peut expliquer certaines variations dont l'impact est important notamment pour des mots de faible fréquence. Le nom *emplacement,* par exemple, dont seulement 40 occurrences sont relevées dans notre corpus, contre par exemple plus de 3000 occurrences pour le nom *analyse,* peut difficilement remplir le critère d'apparition dans un minimum de 40 tranches dans le corpus. L'absence du nom *détail* s'explique différemment, par sa fréquence relative plus forte dans le corpus de contraste (68 occurrences par million de mots) que dans celui d'analyse (55 occurrences par million de mots).

D'autre part, les mesures adoptées peuvent expliquer la présence d'un mot dans une liste et son absence dans l'autre. Il nous faudrait alors, pour isoler clairement l'impact du choix de calcul de spécificité, appliquer intégralement la méthode utilisée par Drouin sur notre corpus d'analyse.

Au niveau des adjectifs, les éléments spécifiques à D-2007 absents de notre ressource renvoient majoritairement au LAG et à la langue générale (*éloigné, libre, prêt, retenu, viable, frappant*) ou à la thématique des articles (*numérique, mondial, mathématique, génétique, écologique*).

Si l'étude des noms et des adjectifs fait émerger des lexiques différents (LST, LAG, langue générale), l'analyse des verbes et des adverbes mène à d'autres observations. Les différences pour ces catégories ne semblent pas concerner des éléments relevant de la terminologie ou renvoyant aux objets d'études des SHS.

Ceci s'explique par la tendance moindre des verbes et des adverbes de renvoyer à des éléments terminologiques.

Les mots spécifiques à notre liste (*contraster, élargir, qualifier, concrètement, parallèlement, récemment*) ou ceux propres à la liste de Drouin (*incorporer, initier, extraire, contrairement, curieusement, présentement*) correspondent en effet aux propriétés du LST que nous avons définies. La différence s'explique alors moins par les thématiques abordées dans les différents corpus d'analyse que par le type d'écrits (thèses ou articles) pris en compte ou par un effet des mesures employées.

Si nous comparons la liste de Drouin avec notre liste du LST avant l'étape de validation manuelle, nous pouvons faire plusieurs constats. D'une part, nous retrouvons, dans les deux listes, les noms renvoyant aux objets d'études des SHS. Nous trouvons par exemple dans la liste de Drouin des éléments tels que *histoire, commune, naissance, marché, pays*, filtrés dans notre liste lors de l'étape de validation manuelle. Nous constatons qu'avant ce filtrage, notre liste contient un grand nombre de ces éléments tels que les noms *étudiant, habitant, quartier, presse, école, cité*, les adjectifs *agricole, européen, industriel, chrétien, administratif, démocratique, judiciaire,* les verbes *financer, gouverner, négocier, enseigner*. La phase de validation manuelle se révèle ainsi particulièrement importante, d'autant plus que le corpus d'analyse est homogène, tel que le nôtre. En effet, plus les thématiques des articles sont partagées, plus notre méthode produira du bruit en identifiant comme LST ces éléments renvoyant aux objets d'études.

Chaque liste incorpore des éléments résultant d'erreur d'étiquetage propre à l'outil adopté. Par exemple, dans notre liste – pour les verbes *nombrer, fondre, traire,* – ou dans la liste de Drouin – pour le « nom » *puisqu'il*. La catégorie des adverbes n'est pas concernée par ce bruit provoqué par les thématiques des articles et/ou thèses.

D'autre part, nous repérons dans la liste de Drouin des éléments absents de notre liste, même avant filtrage, correspondant à des éléments d'expressions polylexicales (*mise, long, cours*).

Les différences non négligeables (32 % de non recoupement pour l'ensemble des catégories) entre les listes extraites selon deux méthodes proches montrent tout d'abord l'influence importante du corpus d'analyse et du corpus de contraste dans les résultats de l'extraction du LST. En cela, la constitution du corpus multidisciplinaire homogène et du corpus de contraste hétérogène se révèle essentielle. De même, la définition de critères statistiques empiriquement éprouvés et l'évaluation manuelle assistée par des formulaires adaptés s'avèrent nécessaires pour aboutir à une ressource de qualité optimale.

2.2.3.2 Comparaison de méthodes et corpus

Dans le cadre du projet TermITH, le LST est utilisé pour améliorer l'indexation en termes. Afin d'analyser la qualité de la liste du LST résultant de notre méthode d'extraction, et en vue d'évaluer l'apport potentiel d'autres méthodes dans la détection du LST, nous avons mené plusieurs expérimentations en faisant varier méthode d'extraction et corpus d'analyse.

Dans un premier temps, nous avons voulu appliquer notre méthode à un corpus similaire au niveau des disciplines, mais différents en termes de genre, afin d'évaluer l'impact du corpus d'analyse dans la détection du LST. Nous avons ainsi tenté d'extraire le LST d'un corpus de notices. Ces notices sont des courts documents rédigés à l'INIST, utilisés pour indexer les articles. Nous voulions ainsi vérifier l'adéquation de notre corpus d'analyse (présenté section 2.2.1.1) pour la tâche de constitution d'une ressource du LST. Les traitements d'indexation automatique dans lesquels s'intègre le LST concernent le genre de l'article de recherche intégral. Pour cette raison notamment, notre ressource a été constituée d'après un corpus d'articles intégraux. Par cette première expérimentation, nous voulons évaluer l'influence du corpus d'analyse dans les résultats d'extraction du LST.

Dans un second temps, nous avons comparé notre méthode d'extraction à une méthode dite par mots-clés utilisée à l'INIST. Ces deux méthodes ont été appliquées sur le corpus de notices, dans le but de faire émerger les points forts de la méthode par mots-clés.

2.2.3.2.1 Effets du corpus d'analyse

Afin d'étudier la présence du LST dans un autre genre d'écrit, et également dans le but de tester notre méthode d'extraction sur un corpus différent du nôtre, nous avons expérimenté notre méthode sur un corpus composé de notices bibliographiques d'articles de disciplines des SHS[50], utilisé par l'INIST[51] dans des traitements d'indexation terminologique. Une notice est composée :

- des informations bibliographiques de l'article (éditeur, date, auteur, titre)
- d'une liste de descripteurs en français et en anglais
- d'un résumé en français

Un exemple de notice, en linguistique, est illustré ci-dessous.

```
NO : Francis 524 524-05-12284 INIST
FT : Quelques pistes pour le traitement des collocations
ET : (Some clues for collocation analysis)
AU : GROSSMANN (Francis); TUTIN (Agnès); GROSSMANN (Francis); TUTIN
     (Agnès)
AF : Université Stendhal-Grenoble 3/France (1 aut., 2 aut.); Université
     Stendhal-Grenoble 3/France (1 aut., 2 aut.)
DT : Publication en série; Niveau analytique
SO : Travaux et recherches en linguistique appliquée. Série E, Lexicologie
     et lexicographie; ISSN 1572-042X; France; Da. 2003; No. 1; Pp. 5-21
LA : Français
FA : Dans cet article introductif à un numéro spécial consacré à la
     collocation, les As. apportent des précisions sur la notion de
     collocation puis proposent des pistes pour le traitement des
     collocations à partir du modèle qui semble le plus abouti en la
     matière, celui des Fonctions Lexicales du Dictionnaire Explicatif et
     Combinatoire d'Igor Mel'cuk et ses collègues. Enfin, ils mènent une
     réflexion sur la motivation de ces associations lexicales, centrale
     d'un point de vue linguistique.
CC : 52455; 524
FD : Phraséologie; Association lexicale; Collocation; Concept linguistique;
     Modélisation; Théorie sens-texte; Motivation; Lexicologie explicative
     et combinatoire
ED : Phraseology; Lexical association; Collocation; Linguistic concept;
     Modelling; Meaning-text theory; Motivation; Explanatory and combinatory
     lexicology
LO : INIST-27516E.354000122695450010
```

Illustration 2.12: Exemple de notice en linguistique

Nous avons soumis à l'analyse syntaxique un ensemble de 11 corpus disciplinaires (philosophie, sciences de l'éducation, sociologie, histoire des sciences, archéologie, religion, géographie, littérature, préhistoire, ethnologie)

[50] Nous remercions notamment Sabine Barreaux pour la mise à disposition du corpus de notices et pour ses éclairages sur la méthode par mots-clés.

[51] L'INIST, en tant que partenaire du projet TermITH, intervient au niveau des méthodes d'indexation en termes. Voir notamment : http://www.atilf.fr/ressources/termith/partinist.php [consulté le 09/09/2016]

composés de 5000 notices. Nous avons ainsi confronté notre méthode textuelle à un corpus de résumés en appliquant les mêmes critères de spécificité et de répartition. En comparant les résultats extraits de ce corpus de notices aux résultats issus du corpus d'articles (avant la phase de validation manuelle), nous avons observé de nombreuses différences. Nous notons que le lexique présent dans les résumés et les titres (composant les notices) ne recouvre pas entièrement celui des articles. Les mots renvoyant à la structure textuelle, au méta-discours ou à la méthodologie sont ainsi absents ou peu présents dans les notices. De ce fait, les connecteurs du type *dans un premier temps, d'une part,* ou les noms intervenant dans la structure textuelle du type *conclusion, section, tableau*, ne sont pas extraits par notre méthode sur ce corpus.

De plus, les mots extraits du corpus de notices sont en majorité des termes ou des mots référant à la thématique de l'article, que nous incluons dans la catégorie des objets des SHS, tels que : *esclavage, fédéralisme, avortement, glaciation, socialisme, impérialisme,* etc.

Si nous nous intéressons aux mots extraits des deux corpus, incluant les candidats-LST non validés à l'issue de la phase d'évaluation manuelle, nous retrouvons l'ensemble lexical, mentionné précédemment (voir section 2.2.2.2), renvoyant aux objets d'études des SHS : *langue, culture, politique, enseignement, mémoire, religion, établissement,* etc.

L'étape d'évaluation manuelle, intégrée à notre méthode d'extraction, se révèle ainsi bien nécessaire, comme nous avons pu le voir sur la comparaison présentée dans la section 2.2.3.1.

Nous observons néanmoins un recoupement significatif dans les listes extraites sur les deux corpus de notices et d'articles, recoupement correspondant au LST : ainsi 483 des 786 noms extraits du corpus d'articles sont également extraits du corpus de notices.

Bien que ce dernier corpus permette l'extraction d'une majorité des noms du LST, nous notons l'absence d'un grand nombre d'éléments essentiels au genre scientifique, tels que le LST renvoyant aux relations (*rapport, condition, lien, effet*), aux objets construits par l'activité scientifique (*question, système, groupe, résultat*), à la caractérisation (*type, manière, valeur, nombre, façon, présence,*

état), aux éléments des méthodes scientifiques des SHS (*segment, effectif, questionnaire*), ou à la quantification (*somme, total, maximum, moyenne, rang, minorité*).

Ces absences mettent en évidence l'hétérogénéité entre mots-clés et LST et confortent le bien-fondé du choix de travailler sur un corpus d'articles intégraux pour une extraction optimale des éléments du LST.

2.2.3.2.2 Effets des mesures

Dans le cadre du projet TermITH, nous avons également mené une expérimentation avec nos partenaires de l'INIST visant à tester d'autres mesures afin d'évaluer leur intérêt dans l'extraction du LST. Nous détaillons dans cette section les résultats de cette comparaison entre notre méthode, ou méthode textuelle, et une méthode dite par mots-clés, reposant sur des critères statistiques différents.

En analysant les recoupements et différences de résultats selon ces deux méthodes, nous constatons que leur utilisation conjointe peut mener à deux types d'amélioration.

D'une part, les résultats de la méthode par mots-clés peuvent permettre de repérer des éléments du LST non extraits par notre méthode ou de « valider » les candidats-LST également extraits par notre méthode. Utiliser l'intersection des deux méthodes pourrait, si ce n'est permettre d'automatiser l'étape de validation manuelle, du moins limiter le bruit dans les résultats de l'extraction automatique et ainsi rendre moins coûteuse la phase d'évaluation manuelle des candidats.

D'autre part, la méthode par mots-clés peut aider à filtrer les résultats de notre méthode, en repérant les éléments clés du corpus mais n'appartenant pas au LST, tels la terminologie et les objets d'études des SHS. Il semble cependant complexe de pouvoir automatiquement identifier dans les résultats de la méthode par mots-clés ces deux ensembles lexicaux que sont le LST et les noms renvoyant aux thématiques (relevant de la terminologie et/ou renvoyant aux objets d'études des SHS).

Cette comparaison est effectuée en appliquant les deux procédés sur le même corpus de notices, préalablement détaillé (voir section 2.2.3.2.1).

La méthode par mots-clés utilise deux indices :

- indice de généralité : indice reflétant la diversité des disciplines dans lesquelles apparaît un mot (Trajtenberg, Henderson, & Jaffe, 1997) ;
- indice de Gini : indice développé pour mesurer le degré d'inégalité de distribution des revenus dans une société. Il est ici utilisé pour calculer la répartition dans les disciplines (Gini, 1921).

L'analyse des convergences et divergences entre la méthode par mots-clés, qui ne prend pas en compte le critère de spécificité par rapport à un autre genre (aucun corpus de contraste n'est utilisé), et notre méthode permet de faire plusieurs observations. Si nous nous intéressons aux éléments extraits seulement par la méthode textuelle, nous constatons une grande proportion d'éléments du LST, confirmant ainsi la pertinence de cette méthode concernant l'extraction de ce lexique.

Parmi les mots extraits par notre méthode et non extraits par la méthode par mots clés, nous identifions :

- les 30 plus fréquents : *étude, analyse, siècle, relation, forme, objet, formation, site, processus, contexte, activité, œuvre, approche, notion, science, réflexion, expérience, élément, production, ensemble, période, philosophie, nature, exemple, connaissance, donnée, territoire, perspective, méthode, terme* ;
- les 30 moins fréquents : *unification, hiérarchisation, prolongement, acception, discontinuité, dichotomie, actualisation, imagerie, reconfiguration, théoricien, optique, jalon, prisme, glissement, substitution, affect, accomplissement, répercussion, avatar, coexistence, dispersion, lacune, potentialité, fragmentation, inclusion, clivage, esquisse, appréciation, facette*.

Ces éléments, centraux dans le discours scientifique, renvoient à plusieurs ensembles sémantiques : des processus scientifiques (*étude, analyse, approche, réflexion, expérience, perspective, méthode*), des objets construits par l'activité scientifique (*objet, exemple, ensemble, terme*), des observables (*donnée, contexte, site, élément*), et des relations (*coexistence, substitution, discontinuité, hiérarchisation, unification, dispersion, répercussion, prolongement, inclusion*). Ces éléments du LST sont essentiels pour notre ressource, et ce pour les diverses

applications envisagées. D'une part, étant fréquents et spécifiques, ils peuvent être erronément extraits comme candidats-termes et constituent en cela une information utile dans la recherche terminologique. Pour ces mêmes raisons, ils se révèlent également indispensables à maîtriser dans le cadre d'activités de rédaction ou de lecture d'écrits scientifiques, dans lesquelles ils sont mobilisés pour la construction de l'argumentation et pour l'organisation textuelle. Le silence sur les mots du LST produit par la méthode par mots-clés confirme ainsi l'adéquation de la méthode textuelle.

Nous nous intéressons dans un deuxième temps aux mots ignorés par la méthode textuelle. Parmi les mots extraits via la méthode par mots clés, nous identifions dans un premier temps les 30 plus fréquents dans le corpus :

- *histoire, politique, économie, ville, société, femme, mémoire, corps, droit, environnement, christianisme, agriculture, état, peinture, pouvoir, commerce, famille, mort, enquête, enfant, historique, internet, temps, image, animal, communauté, violence, travail, musique, urbanisme.*

Hormis *enquête* et *temps*, ces mots renvoient aux objets d'étude des SHS et ne doivent donc pas être inclus, selon nous, dans le LST. Ils renvoient principalement aux thématiques de recherche. L'utilisation d'un corpus de contraste nous permet de réduire ce bruit dans les résultats de la méthode textuelle en appliquant le critère de spécificité. Nous avons cependant vu qu'une validation manuelle des résultats reste nécessaire pour totalement filtrer ces éléments.

Nous considérons dans un second temps les 30 mots extraits les moins fréquents dans le corpus :

- *boisson, ciel, classement, luxe, profane, sidérurgie, apartheid, biomédecine, repas, rumeur, système social, travailleur, famine, inspiration, marchand, peau, alcool, algorithme, cadavre, conservatisme, euthanasie, fertilité, message, nourriture, saison, blé, dictature, fidélité, mélodie, prédiction, bestiaire.*

Nous ne retrouvons pas de mots, dans cet ensemble, excepté *classement*, ayant été validés comme LST. Ces noms exclusivement extraits par la méthode par mots-clés renvoient à des termes ou à des concepts spécifiques à certaines disciplines des SHS et ne répondent donc pas à nos besoins.

Nous pouvons constater, l'intersection entre les deux méthodes étant négligeable, que l'apport de la méthode par mots-clés se situe davantage au niveau du bruit généré par notre méthode. Ainsi, les mots renvoyant aux objets d'étude des SHS pourraient être filtrés en utilisant les résultats de la méthode par mots-clés comme liste d'exclusion, et ainsi remplacer (en partie) l'étape de validation manuelle.

2.3 Conclusion : Nécessités d'une approche semi-automatique

Nous avons présenté dans ce chapitre une méthode d'extraction semi-automatique du LST, combinant critères statistiques pour un repérage automatique et validation manuelle par des juges experts.

Nous avons ainsi, à la manière de Paquot (2010) ou Drouin (2007), procédé à l'identification automatique des mots du LST en tirant parti des informations lexicométriques. Nous avons affiné la prise en compte du critère de spécificité en constituant un corpus de contraste spécialement pour cette phase d'extraction. L'ajout de la gestion des segments répétés, de la dispersion intra – (par l'utilisation de tranches de corpus) et trans – disciplinaire nous a permis de proposer une première liste de candidats LST au bruit minimisé. Le partage de thématiques communes par l'ensemble des disciplines composant notre corpus d'analyse nous a amené à identifier un ensemble lexical, les objets d'étude des SHS, répondant aux critères du LST, mais n'entrant pas dans le sujet de la présente étude. Nous avons alors opté pour une validation manuelle de cette première liste afin de nous assurer de la qualité de notre ressource. L'intégration des informations de dépendances, aux travers des propriétés lexico-syntaxiques déduites de l'utilisation d'un corpus arboré, a ensuite permis aux évaluateurs une recontextualisation des éléments du LST et donc un accès au sens facilité. Dans la mesure où nous nous intéressons aux propriétés linguistiques du LST, et notamment à ses propriétés sémantiques (voir chapitres 3 et 4) et syntaxiques (voir chapitre 5), nous nous sommes principalement basé sur ces propriétés.

Le recours au jugement humain a ainsi rendu possible la validation d'une première version du LST. L'analyse des accords inter-annotateurs a ensuite mis en évidence la difficulté à discriminer le LST du lexique abstrait général, justifiant ainsi le fait d'élargir notre définition du LST, en y intégrant le lexique abstrait

général, dont l'intérêt est similaire au LST tant pour les applications didactiques envisagées que pour les applications de recherche d'information (de termes, de routines, etc.).

Nous notons également que l'application du critère de spécificité engendre un certain silence au niveau de mots pourtant largement convoqués dans l'écrit scientifique (tels les verbes *comprendre* et *voir*) mais non sur-représentés en comparaison de notre corpus de contraste. La non-spécificité de ces unités lexicales n'implique pas qu'ils soient non-pertinents pour certaines des applications décrites en introduction de notre travail. Nous pensons cependant que l'absence de ces éléments dans notre ressource n'est pas problématique dans le cadre de l'indexation terminologique étant donné leur caractère générique qui les rend peu susceptibles d'être faussement extraits comme candidats-termes. Cependant, dans une perspective didactique, certains de ces mots filtrés peuvent se révéler intéressants à inclure dans une ressource pour l'aide à la rédaction d'écrits scientifiques, bien que leur haute fréquence dans le corpus de contraste nous incite à penser que les apprenants maîtrisent davantage ces unités lexicales que les éléments proprement spécifiques du LST.

Notre ressource du LST est ainsi constituée d'entrées définies par un lemme et une catégorie. Cette ressource, en l'état, est exploitable dans des applications d'extraction terminologique, avec par exemple une utilisation en tant que filtre d'exclusion. Nous avons cependant vu qu'une telle utilisation du LST reste limitée et que l'intégration d'informations sémantiques et syntaxiques augmenterait de façon non négligeable l'intérêt du LST dans ce type de processus.

Outre les applications dans la recherche de termes, nous avons également souligné l'intérêt de disposer d'une ressource du LST finement décrite aux niveaux sémantique et syntaxique pour l'identification de phénomènes phraséologiques tels les routines sémantico-rhétoriques. De plus, les perspectives didactiques, par l'intégration du LST dans l'enseignement/apprentissage de la rédaction scientifique, s'ajoutent à ces précédents points dans la nécessité d'élaborer une ressource de ce lexique permettant un accès au sens, à l'usage en contexte, représentatif du genre de l'écrit scientifique. Enfin, au niveau description linguistique du LST, il nous paraît essentiel de structurer ce lexique, en tissant des liens sémantiques et syntaxiques.

La suite de nos travaux participe de ces mêmes préoccupations et constitue une réponse à l'absence d'une telle ressource enrichie du LST pour le français. Nous présentons ainsi dans le chapitre suivant l'ensemble des traitements sémantiques effectués sur la catégorie nominale dans le but d'enrichir et de structurer le LST. Partant de la liste présentée dans ce chapitre, nous opérons dans un premier temps un travail d'identification, pour les noms du LST, des acceptions transdisciplinaires mobilisées dans le corpus d'analyse. Dans un deuxième temps, nous structurons ces entrées en définissant une typologie du LST, organisée en classes et sous-classes sémantiques.

L'ensemble de ces traitements se base sur les informations de combinatoires, calculées d'après les relations de dépendance présentes dans le corpus d'analyse, participant ainsi de la même approche de linguistique outillée. Les informations de combinatoire (permettant de prendre en compte les spécificités de notre corpus) nous permettent de proposer une ressource des noms du LST intégrant différents niveaux d'analyse (lexical, sémantique, syntaxique), à large couverture (531 entrées nominales), et représentative de l'usage dans les écrits scientifiques. Pour ce faire, nous combinons, à la manière de la phase d'extraction du LST, traitements automatiques et analyse manuelle dans un souci de qualité de la ressource.

Chapitre 3
Classification sémantique du LST nominal

Sommaire

À la suite de l'extraction des mots du LST, présentée dans le chapitre précédent, nous avons abouti dans un premier temps à une liste de lemmes-catégories (par exemple *analyse-NOM, examiner-VERBE, abstrait-ADJECTIF, majoritairement-ADVERBE*) répondant positivement aux critères statistiques du LST. La phase de validation manuelle, faisant appel à plusieurs annotateurs, nous a ensuite permis d'éliminer le bruit présent dans cette liste (notamment les mots renvoyant aux objets d'études des SHS) et de s'assurer ainsi de la pertinence des éléments retenus, appartenant à ce que nous nommons le LST étendu (union du LST et du LAG).

La suite de notre étude du LST se situe alors au niveau sémantique car une liste brute du LST de lemmes-catégories, sans autre indication que le lemme et la catégorie syntaxique, se révélerait limitée pour les perspectives d'utilisation que nous développons par la suite. L'affectation de traits sémantiques fins, représentatifs des divers concepts mobilisés à travers le discours scientifique, est essentielle tant pour la recherche d'information que pour la mise en place d'applications didactiques d'aide à la rédaction/compréhension de textes scientifiques.

Le présent chapitre est ainsi consacré aux traitements sémantiques que nous avons effectués, en deux temps.

Nous avons dans un premier temps, pour l'ensemble des éléments du LST, procédé à l'identification des différentes acceptions mobilisées dans le corpus d'analyse. L'appartenance au LST peut être multiple pour un même lemme, si celui-ci correspond à plusieurs acceptions effectivement transdisciplinaires. Le nom *objet* peut ainsi renvoyer à deux acceptions appartenant au LST, comme l'illustrent les deux exemples suivants :

- *Au contraire, les SN génériques nous semblent être des termes fortement classifiants et catégorisants, puisqu'ils rassemblent en une classe des objets ou des individus partageant des propriétés communes*[1] → *objet* prend le sens de 'chose concrète'.
- *Sans préjuger de la manière dont peut se répercuter l'agenda – setting auprès des publics, notre propos ici est bien d'identifier l'état de l'agenda*

[1] Godart-Wendling, B. (2000). Comment ça réfère ?. *Revue de sémantique et pragmatique*, *7*, pp-105-121.

– building en ligne, en se donnant pour objet de questionner les spécificités des différents sites d'information[2] → *objet* prend le sens de 'but d'une action'.

Dans un second temps, nous avons procédé à une classification sémantique, en classes et sous-classes, pour l'ensemble des acceptions retenues.

L'enrichissement sémantique est effectué selon une méthodologie similaire à l'extraction semi-automatique. En effet, ce travail de classification se base sur des données (principalement sur les relations lexico-syntaxiques) extraites automatiquement du corpus. Ces données sont alors utilisées afin d'opérer les regroupements manuels sur la base de propriétés lexico-syntaxiques, issues du corpus. Nous reprenons ainsi le principe de linguistique outillée, pareillement à l'extraction du LST, en combinant traitements automatiques et manuels.

Nous présentons dans cette partie notre travail d'identification des sens et de leur classification pour la catégorie des noms du LST. Les traitements sémantiques des autres catégories ont été effectués, selon la même méthodologie et à l'aide des mêmes ressources, concomitamment par nos collègues du laboratoire LIDILEM[3]. Nous n'aborderons ainsi dans cette partie la classification des autres catégories (verbe, adverbe, adjectif) que lorsque cela s'avère nécessaire (voir notamment section 3.3.3).

Le recours aux méthodes automatiques de classification sera abordé dans le chapitre suivant, avec pour objectifs de proposer une aide à la classification manuelle et à sa validation. En effet, les ambiguïtés présentes dans le corpus, dues notamment à la polysémie des éléments à classer et de leurs cooccurrents, rendent peu effective la prise en compte automatique des cooccurrences. Certaines méthodes se basent sur les définitions présentes dans les dictionnaires électroniques pour gérer les ambiguïtés, cependant cette approche est moins adaptée lorsque les unités lexicales font partie d'un lexique spécialisé (Brun, Jacquemin, & Segond, 2001). De plus, ce type de désambiguïsation automatique

---

[2] Marty, E., Rebillard, F., Pouchot, S., & Lafouge, T. (2012). Diversité et concentration de l'information sur le web. Une analyse à grande échelle des sites d'actualité français, *Réseaux*, (6), 27-72.

[3] Magdalena Augustyn pour les adjectifs, Rui Yan (voir (Yan, 2012)) pour les verbes et Thi Thu Hoai Tran pour les adverbes (pour une description approfondie des adverbes marqueurs de discours, voir (Tran, 2014)). Nous remercions également Agnès Tutin et Marie-Paule Jacques pour leur aide et nombreuses remarques tout au long de ces processus.

nécessite une ressource dictionnairique incorporant des définitions issues d'un corpus comparable au nôtre. Or, il n'existe pas, à notre connaissance, de telle ressource. Enfin, comme nous l'avons précisé ci-dessus, nous visons un enrichissement fin du lexique, ce qui passe nécessairement par des traitements validés manuellement.

La première partie de ce chapitre est consacrée à la présentation du cadre de constitution de notre ressource, en détaillant les différents objectifs poursuivis.

Nous revenons dans un second temps sur plusieurs modèles et études en sémantique lexicale, et identifions plusieurs concepts et méthodes dont nous nous inspirons.

La dernière partie s'attache à détailler la méthodologie de classification mise en place. Nous présentons les outils et ressources employés, ainsi que les méthodes utilisées lors des différents traitements sémantiques que nous effectuons sur le LST.

3.1 Cadre et objectifs pour la classification sémantique

Comme résultat de l'extraction du LST, nous disposons d'une ressource pouvant être utilisée comme liste brute d'exclusion lors de l'indexation en termes. Cet usage du LST en tant que filtre doit cependant tenir compte de la polysémie de certains éléments de notre lexique, pouvant aussi avoir une acception terminologique. Ainsi, *sujet*, nom du LST dans son acception de 'thématique', peut dans d'autres acceptions faire partie de la terminologie de la linguistique ou de la psychologie. Le statut terminologique doit alors être validé malgré l'appartenance du mot-candidat au LST. Nous visons, à travers la classification, à nous affranchir des lemmes pour atteindre un niveau de description supérieur, de l'ordre des concepts. Cette description plus fine des éléments du LST est nécessaire pour envisager des traitements plus complexes tels que :

- le repérage de termes dans des patrons lexico-syntaxiques incluant des sous-ensembles des noms du LST ;
- l'annotation sémantique des éléments du LST à des fins didactiques, dans des applications d'aide à la rédaction/compréhension de textes scientifiques (Tran & Hatier, 2015). En effet, le recours au corpus, enrichi

d'informations sur les propriétés linguistiques (sémantiques et syntaxiques) du LST, permet d'accéder aux données authentiques et favorise ainsi une meilleure prise de conscience linguistique (Cavalla & Loiseau, 2013) ;

- l'extraction de routines sémantico-rhétoriques[4] correspondant à des fonctions identifiées. Par exemple, les énoncés définitoires (Jacques, 2011) incluent des verbes spécifiques du LST : *définir, désigner, nommer*[5]. Ainsi, l'identification automatique d'une séquence du type « *nous emploierons le terme de X au sens de SN* » permettrait d'extraire un terme émergent (*X*) et sa définition (*SN*). L'identification de ces routines permettra notamment une analyse approfondie des procédés rhétoriques présents dans les écrits scientifiques : expliquer, prouver, se positionner, évaluer.

L'ajout de traits sémantiques participe à l'amélioration de l'indexation automatique par la détection de routines (intégrant ces traits) permettant de valider ou non le statut terminologique d'un candidat-terme, telles les routines définitoires incorporant un verbe de dénomination. Plus généralement, certains sous-ensembles du LST (nominal et verbal) jouent un rôle de délimiteur de termes et procurent un indice fiable du statut terminologique du candidat-terme qui co-occurre avec eux, à l'inverse de certaines classes sémantiques nominales (par exemple {temporalité} et {support de communication}) du LST peu introductrices de termes. Une « classification sémantique des noms du lexique transdisciplinaire apparaît donc indispensable pour effectuer un filtrage [des termes] plus efficace » (Jacquey *et al.*, 2013, p. 126).

Il existe également un enjeu didactique de l'enrichissement en informations sémantiques du lexique, toute information notionnelle améliorant l'appropriation de nouveaux mots, la compréhension du sens véhiculé. L'annotation sémantique permet de mieux comprendre les stratégies rhétoriques comme le montre Tutin (2010) dans son étude sur les adjectifs évaluatifs. Il est donc indiqué de proposer une ressource lexicale avec, pour chaque entrée, outre des exemples issus du

---

4 Nous reprenons ici la définition de Tutin (2014) des routines sémantico-rhétoriques, dans le cadre de l'écrit scientifique : formulations stéréotypées, à la limite de la phraséologie, renvoyant à des fonctions rhétoriques spécifiques au genre académique

5 Correspondant à la sous-classe sémantique verbale {ancrage_construction/#dénomination}

corpus, des informations sémantiques et phraséologiques, comme le recommandent Granger & Paquot (2009b).

Notre ressource du LST sera alors constituée, à la suite de la classification sémantique, d'une liste d'entrées correspondant à une acception spécifique pour un lemme dont la catégorie est précisée. Le tableau ci-dessous illustre quelques exemples d'entrées lexicales du LST.

Identifiant	**Catégorie**	**Classe**	**Sous-classe**	**Définition**
développement-1	Nom	{communication_support}	{/section}	Exposé
développement-2	Nom	{processus_évolutif}	{/amélioration_augmentation}	Croissance
développer	Verbe	{processus_évolutif}	{/amélioration_augmentation}	Donner de l'extension
élever	Verbe	{processus_évolutif}	{/amélioration_augmentation}	Augmenter
strict	Adjectif	{modalité}	{/restriction}	Limité
strictement	Adverbe	{modalité}	{/restriction}	Rigoureusement

Tableau 3.1: Exemple d'entrées lexicales du LST

Nous pouvons voir que le nom *développement* a deux entrées correspondant à deux acceptions distinctes :

- l'identifiant *développement-1* correspond à l'acception 'exposé' et appartient à la sous-classe {/section} de la classe {communication_support}, comme l'illustre l'exemple suivant :
  - *Au-delà de ces dénominations, la fréquence de déplacement ou le motif seront bien sûr pris en compte dans les développements et explications*[6] *;*
- l'identifiant *développement-2* correspond à l'acception de 'croissance' et appartient à la sous-classe {/amélioration_augmentation} de la classe {processus_évolutif}, comme dans l'exemple ci-dessous :

[6] Terrier, E. (2009). Les mobilités spatiales des étudiants internationaux. Déterminants sociaux et articulation des échelles de mobilité. *Annales de géographie* (No. 6, pp. 609-636). Armand Colin.

- *Gérard Noiriel note bien à plusieurs reprises une corrélation entre le développement des techniques d'identification et celui des grands systèmes techniques.*[7]

En matière d'évaluation de la classification sémantique, il est complexe d'avoir recours à un « gold standard ». En effet, des classifications peuvent être complètement dissemblables sans pour autant qu'une soit intrinsèquement supérieure. Ceci prévient alors une évaluation automatique par comparaison des résultats de la classification avec une ressource de référence. Cette complexité d'évaluation au niveau de la classification se retrouve au niveau des classes qui la constituent. Ainsi, Tanguy *et al.* (2015) mettent en évidence cette difficulté d'évaluation au niveau des classes (ou voisins distributionnels). La solution, coûteuse en temps, est alors le recours à des juges qui estiment la qualité des regroupements.

Il nous faut ici re-préciser que la constitution des classes (spécifiques à une catégorie ou transcatégorielles) a été faite collectivement, afin de prendre en compte noms, verbes, adjectifs et adverbes. Ce travail collaboratif, impliquant de nombreux échanges sur la méthodologie, la définition des classes, implique une évaluation collective continue de la ressource ce qui améliore ainsi la qualité et la cohérence de la ressource du LST.

La présence d'un gold standard ne résoudrait pas cette problématique de l'évaluation de la classification, car, comme le notent Poibeau & Messiant (2008), un gold standard ne peut pas forcément montrer à quel point les résultats sont exacts. En effet, une classification peut s'avérer adaptée à une application et inadéquate pour une autre. Se pose ainsi le problème d'une évaluation intrinsèque ou extrinsèque des résultats, dichotomie récurrente dans le champ de l'analyse distributionnelle, comme le rappellent Fabre & Lenci (2015, p. 11).

L'évaluation intrinsèque est effectuée par confrontation au jugement humain ou à des ressources comparables. L'évaluation extrinsèque met au centre de l'évaluation l'apport des résultats dans des applications déterminées. En l'occurrence, les résultats de la classification du LST ont plusieurs applications,

---

[7] Dubey, G. (2008). Nouvelles techniques d'identification, nouveaux pouvoirs. Cahiers internationaux de sociologie, (2), 263-279.

comme nous l'avons déjà mentionné : aide à l'indexation en termes, intégration dans une ressource didactique, utilisation des classes dans le repérage de routines.

Nous nous situerons ici dans une logique d'évaluation extrinsèque de la classification sémantique, en examinant l'apport des traits de classes et sous-classes dans les différentes applications envisagées. Nous pourrons par exemple juger de l'utilité et de la qualité des classes et sous-classes au vu des routines capturées. Une autre évaluation est envisageable au niveau de la tâche d'indexation, en mesurant le gain dans les résultats dû aux traits sémantiques. Pour reprendre le terme utilisé par Sagot (2013), nous procéderons donc à une évaluation « orientée-tâche » en l'absence de ressources comparables et compte tenu de la complexité et du coût inhérent à une évaluation manuelle de la précision et de la couverture.

Les objectifs et les spécifications de la ressource ainsi fixés, nous présentons dans la partie suivante les différents modèles existant en sémantique lexicale afin de préciser les bases de la typologie sémantique que nous souhaitons établir pour les noms du LST. Dans un second temps, nous aborderons plus en détail les questions inhérentes à la constitution d'une telle ressource lexicale.

3.2 Sémantique et typologie lexicale

Nous revenons dans cette section sur les concepts et méthodes de la sémantique lexicale afin de préciser notre méthodologie pour l'élaboration de la classification sémantique des noms du LST. Cette revue de la littérature nous permettra d'identifier quelques ensembles lexicaux déjà définis, dont certains dans le genre de l'écrit scientifique, ainsi que plusieurs points méthodologiques que nous reprenons pour aboutir à une classification sémantique correspondant au genre et lexique de spécialité que sont l'écrit scientifique et le LST.

Pour conclure, nous nous intéresserons aux ressources lexicales sémantiques existantes afin de préciser le modèle visé pour notre ressource.

3.2.1 Apports de la sémantique lexicale

Le lexique transdisciplinaire, sur lequel nous effectuons un travail d'analyse sémantique, fait partie d'une « forme particulière de langue de spécialité », selon

les mots de Pecman (2007, p. 80). Or, la classification du lexique d'une langue spécialisée est généralement centrée sur la terminologie. Ainsi, Le Pesant & Mathieu-Colas (1998) notent que dans la description des langues spécialisées, peu de travaux sont axés sur l'articulation entre unités lexicales, et les études se limitent aux termes. Nous nous proposons de combler ce manque en procédant à une étude descriptive des éléments du LST, de leur combinatoire et des relations sémantiques qu'ils entretiennent entre eux. Ainsi que nous l'avons déjà annoncé (voir section 1.3.5.2), la prise en compte de la dimension lexico-syntaxique est primordiale dans l'étude d'une langue associée à un genre, un domaine spécialisé. Ces principes sont à la base de l'élaboration de notre classification sémantique, fondée en premier lieu sur l'étude de la combinatoire des noms du LST. De même, les traitements sur les patrons verbaux et les routines sémantico-rhétoriques participent de ce même intérêt d'analyser le LST dans sa dimension syntactico-sémantique grâce aux données issues du corpus analysé en dépendance.

L'intérêt d'une telle approche est également confirmé par Tanguy *et al.* (2015, p. 103) qui montrent que « les modèles qui prennent en compte de façon raisonnable les informations syntaxiques obtiennent globalement de meilleurs résultats ». Plus précisément, les modèles distributionnels se basant sur les dépendances (filtrées et normalisées, tels que nous le faisons et détaillons dans la suite de ce chapitre) sont les plus avantageux, la nature des contextes sélectionnés étant ainsi essentielle. Ceci est d'ailleurs d'autant plus vrai que la taille du corpus d'analyse est réduite. Le choix d'élaborer notre classification à partir de propriétés lexico-syntaxiques que nous définissons est ainsi pleinement adapté si l'on considère la taille modeste de notre corpus d'analyse spécialisé (5 millions de mots).

Dans leurs travaux, Tanguy *et al.* effectuent une comparaison des méthodes par fenêtres graphiques et des méthodes syntaxiques. Ils testent différents paramètres pour extraire les voisins distributionnels des 30 mots cibles. Tanguy *et al.* (2015, p. 115) concluent que les méthodes par contextes syntaxiques obtiennent les meilleurs résultats, rejoignant les conclusions antérieures de plusieurs travaux. Plus précisément, Otero (2008) a montré que ces modèles basés sur les contextes syntaxiques sont plus à même d'identifier la relation de co-hyponymie (relation unissant les membres d'une même classe ou sous-classe dans notre classification du LST) que les modèles basés sur une fenêtre graphique.

Ce constat est corroboré par Fabre & Lenci (2015) qui notent que les modèles basés sur les dépendances tendent à repérer principalement des co-hyponymes (notre objectif) tandis que les modèles basés sur les fenêtres graphiques repèrent des mots en relation associative. Il est ainsi plus adapté, étant donné notre problématique, d'opter pour un modèle distributionnel basé sur des contextes syntaxiques, filtré et normalisé à partir des dépendances brutes.

La démarche de Tanguy *et al. (2015)* est comparable à celle de Lapesa et Evert (2014), qui éprouvent différents modèles distributionnels pour classifier automatiquement des noms, en comparant les résultats à des jugements humains. Leur objectif est l'identification des modèles optimaux pour l'analyse distributionnelle automatique, en prenant en compte la catégorie des mots cibles à traiter, un modèle pouvant s'avérer idéal pour les noms et moins recommandé pour les verbes. Notre classification se limitant aux noms, nous n'opérons pas de telles comparaisons et nous nous limitons à la définition des dépendances pertinentes dans le cadre d'une typologie des noms du LST.

Nous nous situons donc dans une tâche de sémantique lexicale qui, selon Cruse (1986), appartient aux études empiriques. L'observation des concordances et des données lexicométriques issues du corpus permet de se baser sur l'usage effectif des unités traitées. Les calculs combinatoires permettent d'avoir une vue condensée des caractéristiques sémantiques des noms. Cet apport des métriques doit néanmoins être modéré au vu de la polysémie. En effet, les acceptions associées à un même lemme sont représentées par un seul profil. Un examen manuel reste donc nécessaire pour la désambiguïsation. L'identification des sens présents et l'analyse des propriétés sémantiques s'effectuent à partir de l'observation des occurrences en contexte.

3.2.2 Typologie lexicale

Dans cette partie, nous revenons dans un premier temps sur divers travaux dont le but commun est la constitution de classes lexicales, regroupant des mots partageant des propriétés sémantiques et syntaxiques. Nous nous centrons par la suite sur les études concernant spécifiquement le lexique présent dans les écrits académiques, afin de mettre en évidence les particularités linguistiques de ce genre. Enfin, nous présentons quelques ressources lexicales incorporant des

informations sémantiques en comparant leur structure, avantages, inconvénients, et en évaluant leur utilisabilité dans notre contexte de recherche.

3.2.2.1 Classes de mots et propriétés

La définition de classes de mots est une tâche complexe pour laquelle la subjectivité est souvent de mise. Cruse (1986) remarque qu'il est parfois préférable d'opter pour une plus grande richesse descriptive (au détriment d'une plus grande rigueur théorique), plutôt que de privilégier une théorie complexe aux dépens de l'analyse descriptive. Nous voulons ici représenter les classes le plus formellement possible, cependant, notre but premier est la description sémantique du lexique nominal transdisciplinaire. Dans nos emprunts aux différentes méthodologies et typologies lexicales sur lesquelles nous revenons, nous tenons ainsi compte des particularités du genre étudié et des objectifs détaillés dans l'introduction.

Une typologie peut se faire de manière ontologique ou fonctionnelle (Huyghe, 2015). La première est composée de types incompatibles, partitionnant le réel, définis selon les propriétés référentielles et les « propriétés distributionnelles distinctives ». Les classes de noms d'objets (*Le N se trouve*) et de noms d'événements (*Le N a lieu*), décrites à l'aide de constructions linguistiques, relèvent ainsi de la caractérisation ontologique.

La seconde typologie, fonctionnelle, dont les types ne s'excluent pas, prend en compte les relations prédicatives entre arguments, telle la classe des noms partitifs, dont les méronymes *guidon, clavier, main* décrivent des parties fonctionnelles et séparables. Cette classe regroupe des éléments syncatégorématiques, dénués d'autonomie référentielle, opposés aux noms catégorématiques, comme le rappellent Kleiber *et al.* (2012). Ces classes nominales « ajoutent ainsi à la composante référentielle la description d'une relation ou d'une prédication » (Huyghe, 2015, p. 12), comme dans les exemples suivants : *le guidon de ce vélo, la main de untel.* Les classes fonctionnelles ne correspondent donc pas à un découpage du réel mais sont caractérisées différemment, par exemple au niveau de leurs rôles relationnels ou argumentaux.

Certaines classes, basées sur les rôles thématiques, combinent typologie fonctionnelle et ontologique. Ainsi, Huyghe & Tribout (2015) étudient les noms d'agents caractérisés ontologiquement (entité animée) et fonctionnellement (agent

réalisant l'action décrite par le verbe de base : un *squatteur* fait action de squatter, un *chanteur* fait action de chanter). La construction de classes de noms peut donc se fonder sur différents critères : syntaxiques, ontologiques, morphologiques, etc. Barque (2015), dans son étude sur les noms relationnels, note également qu'il n'y a pas de correspondance systématique entre les classements ontologiques et fonctionnels.

Au niveau des critères classificatoires, nombre de travaux en classification lexicale définissent les classes de mots en fonction de propriétés linguistiques. Ainsi, Flaux & Van De Velde (2000) définissent des classes de noms selon des propriétés syntactico-morphologiques vérifiées par des tests linguistiques. Leur classement distingue par exemple les noms de qualité (employés dans la construction *être d'un N* + expansion[8]) des noms d'affects (construits avec les verbes *ressentir* et *éprouver*).

Le regroupement en classes de mots sur la base de propriétés syntaxiques partagées est également à la base des travaux de Levin (1993) sur les prédicats verbaux de l'anglais. Dubois & Dubois-Charlier (1997) construisent selon une approche similaire, à partir des propriétés syntactico-sémantiques des verbes du français, la ressource *Les Verbes Français*[9] (LVF) à laquelle nous avons eu recours.

Gross (1994) élabore des classes d'objets à partir de traits (*humain, animal, locatif…*) et définit alors ces groupes en extension en listant les verbes s'y inscrivant. Comme le remarque Le Pesant (1994), le principe des classes d'objets non hiérarchisées fait que certaines classes rassemblent des éléments qui ne sont pas nécessairement en relation de co-hyponymie, ce qui rend leur représentation dans une arborescence unifiée impossible. De plus, Giry-Schneider (1994) soulève les problèmes liés à l'utilisation de ces traits, en premier lieu celui de leur inventaire, pour lequel un consensus reste à trouver.

Au niveau de la structuration des typologies lexicales, Kleiber & Lammert (2012) rappellent que les études se sont spécifiquement concentrées sur 3 domaines particuliers : la détermination (à travers l'opposition comptable/massif), la hiérarchisation des noms (par les relations d'hyponymie et d'hyperonymie) et les

---

8 *« Ce passage est d'une grande clarté »* – Orléan, A. (2005). La sociologie économique et la question de l'unité des sciences sociales. L'*Année sociologique*, *55*(2), 279-305.

9 Disponible à cette adresse : http : ://rali.iro.umontreal.ca/rali/?_q=fr/node/1237 [consulté le 30/07/2015]

correspondances entre les noms et leurs dérivés verbaux, adjectivaux ou adverbiaux. Plus spécifiquement, les recherches sur les noms sont particulièrement nombreuses, ces derniers étant des « items lexicaux privilégiés dans la réflexion [...] sur la structure du lexique » (Huyghe, 2015). La catégorie nominale, la plus fréquente dans les vocabulaires et les occurrences, est également centrale dans la réflexion sur l'organisation des concepts. Afin de réduire au maximum la subjectivité inhérente à la classification sémantique, les chercheurs élaborent des tests sur les propriétés spécifiques des membres d'une classe. Ces propriétés peuvent porter sur la complémentation, la possible association avec certains verbes supports, l'aspect duratif, etc. Ainsi, parmi les classes ayant fait l'objet d'études approfondies, nous pouvons citer, d'après Huyghe (2015) :

- Les noms d'artefacts : comptables, se construisent avec des verbes comme *fabriquer, construire*.
- Les noms de propriétés : dénotent des situations d'aspect statif, non dynamique.
- Les noms de territoires : hyponymes des noms de lieux, se construisent avec *dans* et *sur*, et ne peuvent être repris par *dedans* et *dessus*.

Différents critères sont utilisés pour définir les types, notamment ceux portant sur la description spatio-temporelle : les tests d'extension (*Le N mesure/a duré*), d'ancrage (*Il y a eu un N)*, et le repérage (*Elle est vers le N, Elle a lieu lors du N*). Les noms d'objets répondent positivement aux tests spatiaux tandis que les noms d'événements valident les tests temporels. Cette distinction permet ainsi de désambiguïser l'acception résultative ou processive d'un nom déverbal (le nom *classification* peut valider le test d'extension temporelle dans son acception processive mais non dans son acception résultative). Un autre critère, l'aspect dynamique, est également employé, via la construction *venir de / être en train de*. Des classes intermédiaires sont quelquefois définies, pour une description plus fine, tel le dégroupement fait entre objets matériels et informationnels, ou celui opéré entre agentifs dispositionnels et occasionnels (Huyghe & Tribout, 2015). La définition et l'étude de ces sous-classes sont alors justifiées par la grande hétérogénéité de leur super-classe.

Huyghe (2015) note cependant que même la dichotomie *concret/abstrait* est complexe à définir, selon que la distinction repose sur l'aspect matériel/immatériel, accessible/inaccessible aux sens, autonome/dépendant ontologiquement.

Nous nous situons ici dans une approche comparable à celle de Flaux & Van de Velde, en définissant des classes sémantiques en fonction de propriétés lexico-syntaxiques. Des tests d'appartenance sont alors associés aux classes en question, validant les traits définitoires de celles-ci. En cas de polysémie, il faut s'assurer de tester le mot sous l'acception traitée afin de ne pas intégrer à tort une acception à une classe, comme le note Cruse (1986). Considérons par exemple le nom *étude* et les acceptions qu'il recouvre selon le *DEM* :

1. 'travail intellectuel' : acception relevant du LST, appartenant à la sous-classe {processus_cognitif/examen} ;
2. 'livre sur sujet défini' : acception non intégrée au LST, du fait de la relation entre acception processive (sens précédent) et acception résultative (sens présent) comme nous le détaillons dans la section 3.3.1.2 ;
3. 'lieu pour travail individuel' : acception ne relevant pas du LST ;

Lorsque nous intégrons à la classification le nom *étude*, nous testons son appartenance à une classe dans son acception transdisciplinaire, à savoir l'acception 1. Le test lexico-syntaxique d'appartenance à la sous-classe {processus_cognitif/examen} est « *fournir/conduire un N* ». L'acception 1 du nom *étude* satisfaisant le test de cette sous-classe, cette entrée *étude-1* du LST nominal y est intégrée.

L'acception 3 répond positivement au test de la classe {espace} *:* « *Se situer dans/à N* ». Cependant, comme ce sens ne relève pas du LST, le nom *étude* n'est pas associé à cette classe.

En conclusion de cette revue des travaux antérieurs, nous pouvons distinguer, en sémantique lexicale, les travaux portant sur des classes larges et génériques, de ceux se concentrant sur l'étude de classes précises. Parmi les premiers, Fasciolo (2012) met en place des tests pour différencier les mots dont les traits définitoires sont des connaissances partagées (*tigre, ordinateur*) des noms

plus abstraits (*lieu, chose*), les sommitaux[10], comparés aux pronoms de par leurs comportements similaires. Parmi les seconds, se trouvent les études centrées sur les noms d'agents et d'instruments (Huyghe & Tribout, 2015), d'idéalités (Flaux & Stosic, 2014), relationnels (Barque, 2015), dynamiques (Haas & Gréa, 2015), etc.

De ces travaux, nous reprenons plusieurs points :

- la prise en compte de la polysémie nominale (voir l'exemple précédent avec le nom *étude*) ;
- une approche basée sur la combinatoire, en associant aux classes des tests lexico-syntaxiques déduits de l'observation des cooccurrences en corpus ;
- certaines classes déjà décrites par différents auteurs : *temporalité, espace* ;
- certains tests lexico-syntaxiques :
    - test d'extension, pour la classe {temporalité/durée} : *année, cycle, période* ;
    - test d'ancrage, pour la classe {support de communication/exposé} : *colloque, conférence.*

Nous notons cependant que ces différentes classifications ne sont pas complètement adaptées à notre lexique spécialisé (aux acceptions largement abstraites) et aux applications que nous visons. En effet, les traitements automatiques de recherche d'informations (de termes, de routines sémantico-rhétoriques) impliquent une certaine granularité dans la typologie du LST. Nous retiendrons ainsi certains éléments de ces travaux tout en s'assurant de leur adéquation avec la spécificité du genre d'écrit et de lexique qui nous intéressent ici.

Nous présentons dans la partie suivante les études portant sur l'élaboration de classifications dans les écrits scientifiques dont nous reprenons en partie la méthodologie ou la typologie.

[10] Les noms sommitaux constituent « un ensemble réduit de noms à fréquence élevée ayant une référence généralisée au sin de classes nominales majeures (e.g. noms de lieux, faits). » (Adler & Eshkol-Taravella, 2012, p. 113)

3.2.2.2 Classifications lexicales dans les écrits scientifiques

Comme nous l'avons abordé dans la section 1.2.2.2.3, plusieurs travaux ont entrepris de structurer au niveau sémantique les lexiques scientifiques, en prenant pour entrée les unités lexicales (Tutin, 2007c; Paquot, 2010) ou la phraséologie (Pecman, 2004b; Simpson-Vlach & Ellis, 2010; Tran, 2014).

Paquot (2010) utilise le logiciel *UCREL System* (qui ne traite pas le français) pour automatiquement annoter les mots simples de l'*Academic Vocabulary*. Les unités lexicales, non désambiguïsées, sont alors automatiquement affectées à une ou plusieurs grandes classes : termes génériques, le temps, la langue et communication, etc. La catégorie « termes génériques et abstraits » représente 87 % des 599 candidats que Paquot extrait.[11] Cet étiquetage sémantique automatique présente l'avantage de pouvoir regrouper plusieurs parties du discours dans une même classe. Nous verrons dans la section 3.3.3 que notre classification sémantique des noms du LST se situe sein d'une classification transcatégorielle (comprenant également adjectifs, adverbes et verbes) et autorise ainsi le regroupement de différentes parties du discours dans une même classe.

Tutin (2007c) utilise pour sa part les méthodes de la sémantique distributionnelle, en s'inspirant en partie des tests et critères définis par Flaux & Van de Velde (2000), pour élaborer une classification d'une partie des noms du LST. Nous retrouvons notamment parmi les critères utilisés l'aspect extensif, le trait abstrait/concret et les relations avec des verbes identifiés[12]. Elle distingue, à l'aide des propriétés morphologiques, sémantiques et syntaxiques : les processus, les objets construits, les observables, les supports. Tutin remarque alors que certaines associations entre mots ne sont pas pertinentes sémantiquement bien que statistiquement importantes. Le recours à une classification manuelle est alors nécessaire pour éviter la prise en compte de ces associations calculées automatiquement.

L'analyse sémantique de la phraséologie dans les écrits scientifiques a également été le sujet de plusieurs études. La distinction et le classement des

---

[11] Nous n'avons pas chiffré ce point dans notre liste mais les noms sont en grande majorité abstraits d'où la difficulté à effectuer une classification sémantique automatique assez fine pour différencier ces unités lexicales

[12] Ainsi, Tutin note que les noms d'acteurs de l'activité scientifiques sont souvent sujets de *examiner, décrire, observer*.

acceptions se font alors sur des unités polylexicales dont la polysémie, bien qu'existante, est moindre que pour les mots simples. Pecman (2007) propose une classification notionnelle fine de la phraséologie de la langue scientifique générale, en décrivant les différences sphères conceptuelles mises en œuvre dans le discours scientifique : scientificité, universalité (comparable à notre LAG), modalité et discursivité. Son ontologie, détaillée section 1.2.2.2.3, comporte au total 125 concepts spécifiques ou hyponymiques. Sa classification est issue d'une analyse entièrement manuelle, i.e. ne reposant pas sur l'extraction automatique de la combinatoire dans le corpus. Son but est alors de « dégager l'ensemble des notions à l'œuvre dans le discours scientifique général » (2007, p. 89). Certains de ces concepts, centraux dans l'écrit scientifique, font effectivement partie de la classification manuelle des noms du LST que nous élaborons, notamment au niveau des classes {état_qualité} et {relation}. À l'inverse, plusieurs des classes que nous intégrons à notre typologie sont absentes de l'ontologie de Pecman, telles les classes {communication_support}, {collectif_partitif} ou {personne}.

La problématique de la classification sémantique des éléments du LST est également au cœur des travaux de Tran (2014), qui se concentre sur les séquences lexicalisées à fonction discursive. Ces séquences, ou marqueurs discursifs, « établissent des relations entre les segments textuels » (2014, p. 30) et structurent ainsi le discours. Elle élabore ainsi une typologie de ces marqueurs (*par ailleurs, en premier lieu, par conséquent,* etc.) à travers une modélisation syntactico-sémantique. Sa typologie se fait selon une approche onomasiologique, en regroupant les séquences par fonctions (topicalisation, énumération, reformulation, concession, etc.), en prenant en compte trois critères : la position syntaxique, la portée syntaxique et la valeur sémantique du marqueur. Lors de nos travaux communs de classification transcatégorielle du LST (voir section 3.3.3), nous avons pu constater que l'ensemble de sa typologie, pleinement adaptée au genre et au lexique en question, n'est pas transposable pour la catégorie nominale. Les dimensions rhétorique et énonciative, présentes dans sa typologie des adverbes, sont également au centre des travaux de Simpson-Vlach & Ellis (2010) qui opèrent une classification des blocs lexicaux spécifiques aux écrits académiques : les expressions référentielles de spécification, les expressions de posture et les organiseurs de discours.

L'étude de ces principaux travaux de classification du lexique scientifique nous amène à identifier plusieurs points essentiels. En premier lieu, différents objectifs justifient de l'intérêt d'une telle classification manuelle. Ainsi, les enjeux didactiques (Paquot, 2010 ; Pecman, 2004 ; Tran, 2014) ou de description linguistique (Siepmann, 2007 ; Tutin, 2007) liés à une telle ressource sémantique sont prouvés par l'existence des travaux pré-cités. Au niveau de la définition des entrées lexicales, plusieurs études démontrent l'intérêt, didactique et linguistique, d'étudier la phraséologie scientifique transdisciplinaire. Nous partageons ce constat sans cependant traiter dans les présents travaux de la phraséologie transdisciplinaire au sens strict, par ailleurs abordés (Jacques, 2011; Tran, 2014; Tutin, 2014; Yan & Hatier, 2016). Nous intégrons certains éléments phraséologiques dans notre étude, à la condition qu'ils intègrent au minimum un élément de notre ressource des mots simples du LST. Nous faisons d'ailleurs l'hypothèse qu'une classification sémantique des mots simples permet de servir d'amorce dans l'identification de routines phraséologiques récurrentes spécifiques dans l'écrit scientifique.

Au-delà de la classification, se pose la question de la ressource intégrant les informations sémantiques. Nous faisons dans la partie suivante une revue des ressources lexicales sémantiques existantes, afin de définir la structure et l'organisation de notre ressource pour la rendre la plus adaptée possible aux différentes applications envisagées.

3.2.3 Ressources lexicales : objectifs et état de l'art

Nous précisons dans cette partie le type de structure que nous envisageons pour notre ressource lexicale du LST, en abordant les différents principes de constitution d'une telle ressource, au travers d'une revue des travaux du domaine.

3.2.3.1 Cadre et structure de la ressource

Les tâches de dégroupement en acceptions et de classification sémantique ont pour but d'enrichir notre ressource lexicale, en l'état constituée de lemmes-catégories avec informations de fréquence. Une entrée correspond alors à une unité formelle composée d'un lemme associé à une catégorie grammaticale (par exemple *sens* et NOM).

La polysémie des unités lexicales rend complexe la précision des informations de fréquence et de cooccurrences. En effet, lorsque nous comptabilisons le nombre d'occurrences des mots du LST, nous nous situons au niveau des lemmes-catégories et non au niveau des acceptions. Ainsi, le nom *sens* apparaît dans notre corpus 2761 fois. Nous ne savons, pour ce nombre total d'occurrences, lesquelles correspondent à sens_direction, à sens_signification, ou à toute autre acception. Puisqu'il n'est pas possible d'effectuer une désambiguïsation de l'intégralité des mots potentiellement polysémiques de notre corpus, nous optons pour un traitement manuel afin de nous assurer des acceptions effectivement présentes dans le corpus.

Immanquablement, comme le notent Manuelian *et al.* (2010), une classification sémantique effectuée automatiquement implique une phase de validation manuelle dans l'optique de traitements fins comme la détection de routines. De plus, comme précédemment explicité, l'objectif d'applications didactiques intégrant notre ressource impose que celle-ci soit la plus fiable possible.

L'utilisation de l'analyse distributionnelle automatique (avec ou sans validation manuelle) pour la constitution de ressources lexicales est néanmoins fréquente dans les travaux en TAL (Habert & Zweigenbaum, 2002), mais pose certains problèmes spécifiques. Ainsi, Fabre & Lenci (2015) notent qu'une difficulté importante dans l'utilisation des méthodes d'analyse distributionnelle est le fait qu'il est complexe de distinguer la similarité sémantique (liens de synonymie, antonymie, co-hyponymie, par exemple entre *voiture* et *van*) d'autres relations sémantiques (de méronymie par exemple entre *voiture* et *volant*). Ces méthodes ont alors pour résultats des regroupements de mots effectués sur la base de relations sémantiques diverses, aboutissant à une représentation approximative du sens des unités lexicales.[13]

Le fait d'allier extraction automatique de propriétés distributionnelles et interprétation manuelle de ces propriétés a alors pour but de veiller à ce que les regroupements soient cohérents malgré l'hétérogénéité des relations sémantiques ainsi identifiées.

---

[13] « DSM provide quite a coarse-grained representation of lexical meaning » (Fabre & Lenci, 2015, p. 13).

En termes de lexiques sémantiques généralistes, Valette (2008) distingue 2 types de ressources. Les premières ont une approche paradigmatique, liant les acceptions par relations hiérarchiques, le plus souvent sous forme d'ontologies, comme le fait *Wordnet*[14] (Fellbaum, 1998) à travers les *synsets* et les liens sémantiques tels la méronymie. Les secondes, d'approche syntagmatique, sont centrées sur les propriétés syntaxiques.

Notre ressource du LST est d'approche paradigmatique, en regroupant des ensembles de co-hyponymes, dans des classes et sous-classes sémantiques, organisées en une typologie à deux niveaux. Notre classification prend cependant également en compte la dimension syntagmatique à travers les cooccurrences lexico-syntaxiques. Le résultat de cette typologie est une classification sémantique reposant sur les propriétés sémantiques et syntaxiques des éléments, à travers l'étude de leur combinatoire dans le corpus d'analyse.

L'enrichissement sémantique du LST permettra alors pour des apprenants une double consultation de la ressource : onomasiologique et sémasiologique, deux entrées typiques pour ce type de ressource selon Gala & Zock (2013). L'approche onomasiologique du LST n'a d'ailleurs pas été entièrement résolue, comme le note Pecman (2007), d'où l'importance d'une modélisation fine de ce lexique, permettant des recherches par contenu notionnel ou fonction rhétorique, tel que conseillé par Rinck (2010).

3.2.3.2 Ressources existantes

Nous présentons ici une sélection de ressources en abordant les aspects de couverture, granularité, adaptabilité à notre corpus, et utilisabilité dans les applications envisagées.

Il existe peu de ressources sémantiques lexicales, à notre connaissance, couvrant l'ensemble des catégories sur lesquelles nous travaillons, et proposant une classification qui serait adaptée à notre problématique.

La ressource *Wordnet* se rapproche du type de ressource que nous recherchons pour le français. Cette ressource est une base de données lexicale, en anglais, à large couverture particulièrement utilisée dans les applications TAL. Les

[14] Disponible en version électronique ici : http : ://wordnet.princeton.edu [consulté le 30/07/2015]

unités lexicales sont organisées en *synsets* (groupe de quasi-synonymes). Les synsets sont reliés par des liens d'hyperonymie, d'hyponymie, d'antonymie et de synonymie. Des adaptations pour d'autres langues que l'anglais ont vu le jour, notamment pour le français avec la ressource *Wolf* (Sagot & Fišer, 2008) issu de *Wordnet*. Cependant, cette ressource générée automatiquement, sans validation manuelle, n'est pas adaptée à notre problématique et plus particulièrement au lexique spécifique qu'est le LST.

Le type d'organisation lexicale, du lexique anglais présent dans *Wordnet*, est comparable à ce que nous recherchons pour le français. Cependant, le caractère massivement abstrait de notre lexique rend l'utilisation des ressources connexes à *Wordnet* peu appropriée. De plus, en comparaison d'autres ressources, la description sémantique et syntaxique ne nous paraît pas suffisamment riche. Ainsi, comme le remarquent Hadouche & Lapalme (2010), certaines propriétés ne sont pas explicitement incluses dans *Wordnet* telles le domaine, les sens figurés, les structures syntaxiques dans lesquelles s'inscrivent les unités, etc.

FrameNet[15] (Baker, Fillmore, & Lowe, 1998) est une autre ressource lexicale, de l'anglais, à large couverture et intégrant des informations et liens sémantiques. *FrameNet* est basée sur la sémantique des cadres (Fillmore, 1982), dans laquelle la signification des mots s'étudie à travers le contexte événementiel ou situationnel. Toute acception d'une unité lexicale est représentée par sa combinatoire syntaxique et sémantique. Les cadres (environ 1000) sont associés à des rôles sémantiques, fondamentaux (*core*) ou non, instanciés par les unités lexicales en contexte. Les propriétés syntaxiques et sémantiques sont finement décrites. Cependant, outre sa non-disponibilité en français, *FrameNet* pose le problème de la subjectivité de sa construction, inhérente à la définition des rôles thématiques centraux dans ce modèle, comme le notent Messiant, Gábor, & Poibeau (2010).

Une autre limite de certaines ressources est leur restriction à une catégorie syntaxique. Ainsi, *Dicovalence* (Van den Eynde & Mertens, 2003), *LVF* (Dubois & Dubois-Charlier, 1997), et *VerbNet* (Schuler, 2005) sont trois ressources ne couvrant que la catégorie verbale.

---

[15] Disponible en version électronique ici : https : ://framenet.icsi.berkeley.edu/fndrupal/home [consulté le 30/07/2015]

La base de données *DiCo*[16] (Polguère, 2003), développée sur les principes de la *Lexicologie Explicative et Combinatoire* intègre des informations sur deux niveaux pour chaque entrée. Sont renseignées les relations sémantiques entre unités lexicales ainsi que les liens syntagmatiques au travers des collocations. Cette ressource propose en outre une riche hiérarchie d'étiquettes sémantiques associées aux acceptions des vocables. Bien que certaines de ces étiquettes, et leur arborescence, nous semblent tout à fait pertinentes pour la classification du LST (notamment les étiquettes *caractéristique*, *événement*, *lieu* et *entité informationnelle*), la limite de la couverture du *DiCo* et son inscription dans la langue générale rendent cette ressource peu adaptée à notre problématique.

La *Lexicologie Explicative et Combinatoire* est également le modèle à la source de l'élaboration du *Réseau Lexical du Français* (*RLF*). Cette ressource, en cours de développement, (Lux-Pogodalla & Polguère, 2011), se donne pour objectif une large couverture du lexique français (10 000 vocables visés) et comme applications possibles le traitement automatique de la langue. Le *RLF*, en tant que graphe lexical intégrant de nombreuses relations lexicales, nous semble une ressource adaptée pour l'élaboration d'une typologie des noms du LST.

Pour le français encore, le *Dictionnaire Électronique des Mots* (*DEM*) (Dubois & Dubois-Charlier, 2010), intègre l'ensemble des catégories et propose, pour chaque entrée lexicale, des informations catégorielles, morphologiques (dérivation, flexion), sémantiques (domaine, définition), et syntaxiques (à travers notamment les contextes et opérateurs[17]). Ressource librement disponible, le *DEM* est constitué de 140 000 entrées, chacune définie par plus de dix rubriques.

Le *DEM* propose des informations sur les axes syntagmatique et paradigmatique à travers les synonymes (donnés dans les définitions) et les étiquettes de domaines. Ceci permet de tester la cohérence lors de la construction d'une typologie en confrontant les classes créées aux regroupements présents dans le *DEM*. Ainsi, nous verrons que notre classification comprend une classe {collectif} avec notamment pour membres les noms *ensemble, groupe, totalité, échantillon, élément*. Ces noms sont tous associés dans le *DEM* à l'opérateur

---

[16] Une description complète est disponible ici: http://olst.ling.umontreal.ca/?page_id=77&lang_pref=fr [consulté le 30/07/2015]

[17] Pour les noms, les opérateurs correspondent à la classe de verbes avec lesquels le nom se combine prototypiquement.

'groupe'. Les informations morphologiques permettent d'envisager des traitements fins, par exemple au niveau des routines. Ainsi, en sachant que *analyse* et *analyser* sont reliés, un regroupement des deux séquences suivantes pourrait être fait automatiquement :

- *Nous procédons à l'analyse des déterminants*
- *Nous analysons les déterminants*

En dépit de ces avantages, le *DEM* n'intègre pas de typologie fine telle que nous le souhaitons. Notre ressource du LST vise à organiser les mots en classes et sous-classes. Le *DEM* propose lui une organisation du lexique par grands domaines tels *commerce, droit, entomologie, oiseau,* etc. Cette ressource constitue cependant une base solide, à large couverture, pour les traitements sémantiques envisagés, même si certaines acceptions particulières de mots éléments du LST n'y sont pas présentes et devront être ajoutées manuellement[18]. En adoptant le *DEM* comme ressource de référence pour l'identification des acceptions, nous nous assurons aussi de maintenir une certaine cohérence avec les travaux effectués sur les acceptions des verbes du LST par Yan qui se base sur le *LVF*, ressource élaborée de manière analogue au DEM, par les mêmes auteurs.

Cet inventaire des ressources existantes ne fait ainsi pas apparaître de lexiques sémantiques en français correspondant à nos critères de couverture, dégroupement sémantique et typologie adaptée à notre lexique abstrait. Les différents lexiques sémantiques généralistes, concernant majoritairement l'anglais, ne sont ainsi pas adéquats pour le corpus spécialisé sur lequel nous travaillons.

C'est pourquoi nous optons pour une élaboration manuelle (en partie basée sur des résultats de traitements automatiques, tels l'extraction des profils combinatoires) de notre lexique nominal sémantique du LST. À ces informations sémantiques seront ajoutées par la suite certaines informations syntaxiques, liées aux routines sémantico-rhétoriques et aux constructions verbales.

Ce choix d'une constitution manuelle est également motivé par le fait que l'utilisation de ressources sémantiques constituées automatiquement pose un problème lié à leur qualité. Sagot & Fišer (2008) confirment ainsi que la

[18] Telles *lecture* au sens de 'interprétation' ou *impact* au sens de 'conséquence'.

construction manuelle d'une classification produit de meilleurs résultats en termes de pertinence et de précision linguistiques.

Ayant défini les caractéristiques de la ressource sémantique du LST que nous envisageons, nous pouvons alors mettre en place les différents traitements sémantiques s'appuyant à la fois sur des méthodes automatiques et manuelles. Ces traitements sont ainsi fondés sur une approche de linguistique de corpus, tirant parti de notre corpus d'articles de recherche analysé en dépendance. Nous reprenons le principe de regroupements lexicaux basés sur des propriétés sémantiques-syntaxiques partagées et prenons pour amorce, pour l'identification des différentes acceptions mobilisées dans le corpus, les entrées du *DEM* pour chaque nom du LST précédemment extrait.

Nous présentons dans la partie suivante la constitution de notre typologie, à travers les deux étapes que sont le dégroupement en acceptions et la classification de ces dernières.

3.3 Méthodologie des traitements sémantiques du LST

Le traitement sémantique s'effectue en deux étapes. Nous avons procédé dans un premier temps au dégroupement en acceptions de notre liste de lemmes de noms du LST. Dans un second temps, nous avons élaboré notre classification, en partant des classes sémantiques définies par (Tutin, 2007c) pour les noms du LST, afin de regrouper les noms dans des classes et sous-classes sémantiques de co-hyponymes.

L'objectif est ainsi de disposer d'autant d'entrées lexicales que d'acceptions transdisciplinaires présentes dans le corpus. À chaque acception, et donc chaque entrée lexicale, sont ensuite associées une classe et une sous-classe sémantique.

Par exemple, le nom *sens* doit avoir deux entrées :

1. Pour l'acception 'direction', dont la classe serait {espace} et la sous-classe {orientation} et correspondant aux exemples phrastiques suivants :
    - *Les résultats obtenus par Croyle (1985) vont dans ce sens*[19] ;

[19] Martinie, M-E., & Joule, J-V. (2004). Changement d'attitude et fausse attribution effet de la centration sur le comportement de soumission. *L'année psychologique*, Vol 104, 3, 517-535.

- *Les propos de Donald Preziosi vont dans le même sens : le musée est « l'une des institutions les plus centrales et fondamentales de l'invention de la modernité »*[20] ;
- *[…] il n'est pas un accord de libre-échange mais une convention à sens unique contraire aux règles du commerce mondial […]*[21].

2. Pour l'acception 'signification', dont la classe serait {objet_scientifique} et la sous-classe {/explicatif_simple} et correspondant aux exemples phrastiques suivants :
   - *Nous entendons « politique » dans le sens large du terme, c'est-à-dire « qui a rapport aux affaires publiques*[22] *»*
   - *La syllepse fait se « rencontrer » dans une même occurrence deux sens d'un terme polysémique*[23]
   - *Pour autant, existe-t-il un territoire au sens de Le Berre (1992) ?*[24]

Les classes et sous-classes sont constituées par le regroupement d'éléments partageant des propriétés sémantiques et syntaxiques. Ainsi, les unités d'une même classe :

- sont en relation de co-hyponymie[25] ;
- valident la définition donnée pour la classe ;
  - par exemple, les éléments de la sous-classe {communication_support/graphique} sont définis ainsi : 'représentation graphique éclairant le texte' ;
- partagent des relations lexico-syntaxiques significatives ;

---

[20] Chivallon, C. (2006). Rendre visible l'esclavage. *L'homme*, 4/2006, 7-41.

[21] Grégoire, E., & Théry, H. (2007). L'Ogre et le Petit Poucet. *L'Espace géographique*. 3/2007, 267-282.

[22] Étienne A (2006), Les non-dits de la scène anglaise (xviiie-xxe siècle). *Ethnologie française* 1/2006 (Vol. 36), p. 19-26.

[23] Lecolle, M. (2000). Figures et références plurielle en corpus journalistique. *Cahiers de Grammaire*, 25, 29-52.

[24] Musard, O., Fournier, J, & Marchand, J-P. (2007). Le proche espace sous-marin : essai sur la notion de paysage. *L'espace géographique,* 2/2007, 168-185.

[25] Certains éléments de classe ont lien de méronymie, par exemple *année* et *siècle* dans la sous-classe {temporalité/durée}

    - par exemple, les éléments de la sous-classe {communication_support/graphique} (*tableau, graphique, figure, illustration, motif*) sont fréquemment modifiés par un numéral (*figure 1, tableau 4*) et sont fréquemment sujets du verbe *montrer* ;
- répondent positivement à des tests lexico-syntaxiques découlant de ces relations ;
    - par exemple les éléments de la sous-classe {communication_support/graphique} valident le test *: Dét N Numéral montre*.

Nous utilisons les mêmes ressources et outils pour l'identification des acceptions et pour leur classification. Ces deux traitements sont effectués à partir des données issues du corpus, selon les méthodes de l'analyse distributionnelle.

Dans un premier temps, nous recensons les sens présents dans le corpus à l'aide des concordances obtenues via le *Lexicoscope*, qui intègre dans la présentation des exemples l'information du sous-corpus disciplinaire. Nous nous basons parallèlement sur une ressource dictionnairique en prenant soin de ne garder que les acceptions transdisciplinaires, et en ajoutant un sens si nécessaire.

Dans un deuxième temps, nous observons les profils combinatoires, calculés d'après les associations statistiquement significatives, pour regrouper les éléments partageant des cooccurrents identiques ou sémantiquement proches. Ceci nous permet également de définir des tests lexico-syntaxiques permettant de valider ou d'invalider l'appartenance d'un nom à une classe ou à une sous-classe. La définition des classes est inspirée en partie de la classification opérée par Tutin (2007c).

3.3.1 Dégroupement en acceptions

Nous présentons dans cette section le traitement d'identification des différentes acceptions des noms du LST mobilisées dans notre corpus d'analyse. Nous détaillons tout d'abord les outils et ressources que nous utilisons afin d'inventorier ces acceptions et abordons par la suite le traitement de la polysémie des noms du LST.

3.3.1.1 Outils et ressources pour l'identification des acceptions

En nous basant sur les sens présents dans la ressource du *Dictionnaire Électronique des Mots*[26] (*DEM*) de (Dubois & Dubois-Charlier, 2010), nous dégroupons notre liste de 493 lemmes nominaux en 1099 entrées ou lemmes-catégories-acceptions. Une entrée du LST, précédemment définie de façon formelle par un doublet lemme-catégorie, par exemple *sens-NOM,* est ainsi combinée à un identifiant d'acception, par exemple *sens-NOM-1* où *1* renvoie à l'acception 'direction'.

Nous récupérons dans le *DEM* l'ensemble des acceptions décrites pour les noms du LST puis opérons un filtrage en procédant comme suit. À l'aide de *ScienQuest*[27] (Falaise, Tutin, & Kraif, 2011) et du *Lexicoscope*, nous observons les concordances et éliminons les acceptions absentes ou non transdisciplinaires et aboutissons à 531 entrées.

Nous présentons ci-dessous un exemple avec le nom *sujet* qui, outre son sens transdisciplinaire de 'thématique' peut notamment renvoyer au sujet humain d'expérience scientifique ou au sujet syntaxique en linguistique.

---

[26] Disponible à l'adresse : http : ://rali.iro.umontreal.ca/rali/? q=fr/dem [consulté le 30/07/2015]

[27] Disponible à l'adresse : http : ://scientext.msh-alpes.fr/scientext-site/spip.php ? article1 [consulté le 30/07/2015]

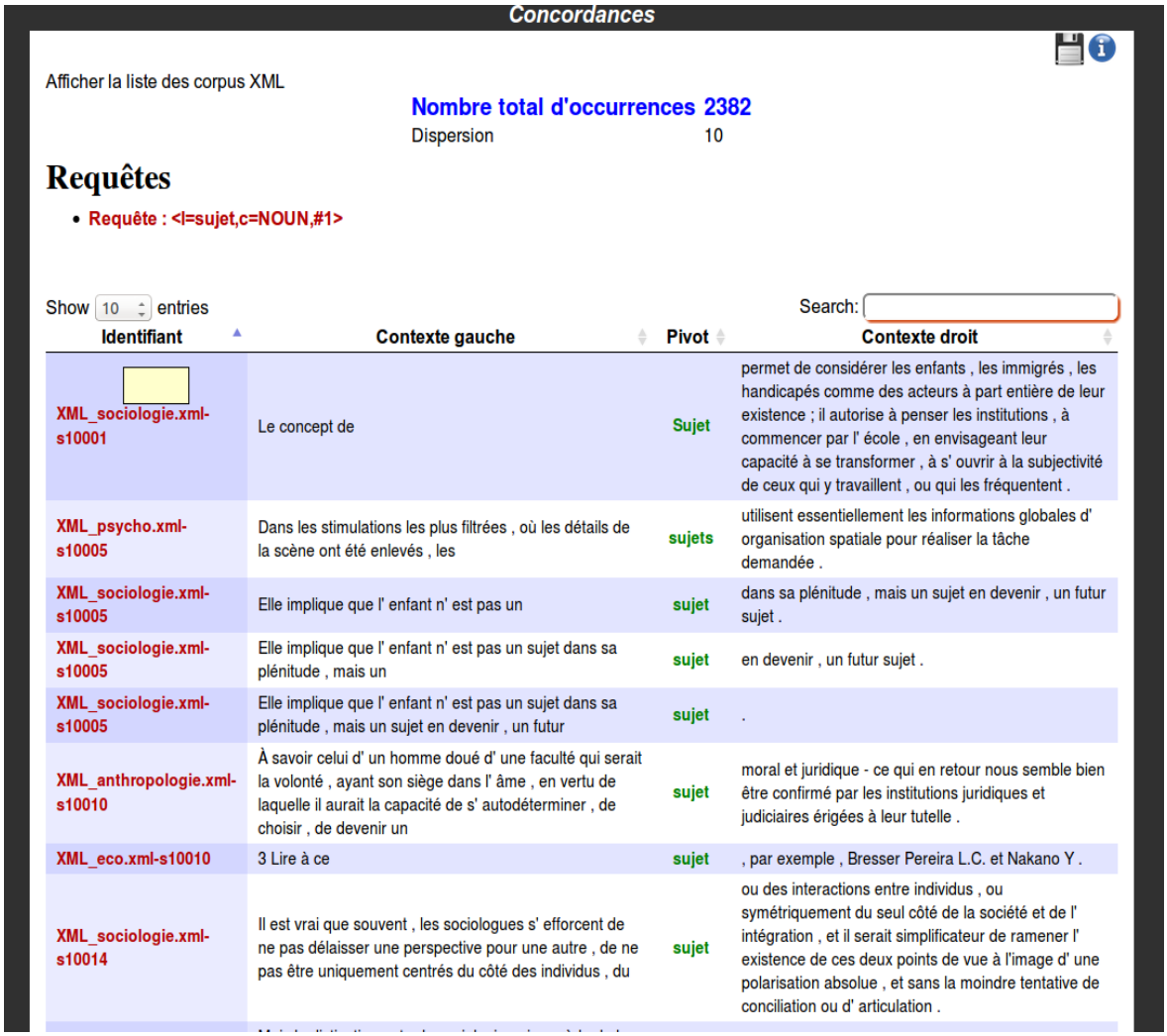


Illustration 3.1: Concordance de sujet dans le Lexicoscope

Nous pouvons ainsi vérifier les sens en contexte afin de procéder au dégroupement en acceptions.

Pour qu'une acception soit considérée comme effectivement présente et transdisciplinaire dans le corpus d'analyse, nous appliquons le seuil de fréquence absolue à 20 et celui de dispersion à 5 disciplines. Ainsi, pour l'exemple de *sujet* au sens de 'thématique', nous vérifions, en observant les concordances, que le nombre d'occurrences dans ce sens soit au moins égal à 20 et qu'il soit inscrit dans un minimum de 5 disciplines différentes, pour nous assurer de son emploi effectivement transdisciplinaire.

Le *Lexicoscope* nous donne accès, via les identifiants de texte, à la discipline concernée pour chaque concordance.

La vérification de ces données statistiques doit être faite manuellement, le corpus n'étant pas désambiguïsé. Cependant, lorsque les noms étudiés atteignent certaines hautes fréquences[28], une estimation du nombre d'occurrences par acception est effectuée d'après les 100 premières concordances issues du *Lexicoscope*[29].

Ce travail de sémantique lexicale s'effectue à partir des informations présentes dans notre ressource de référence (*DEM*), et de celles issues du corpus, via les outils cités afin de faire émerger des profils combinatoires pour les éléments que nous souhaitons classer. Ces profils combinatoires, comme l'expliquent Kraif & Diwersy (2012), donnent accès à la liste des cooccurrents les plus fréquents du mot étudié.

3.3.1.2 Acceptions et polysémie

La polysémie, présente dans la langue générale, intervient également dans les langues spécialisées. Ce fait a été montré dans de précédents travaux, notamment pour la terminologie (Bertels & Geeraerts, 2012).

Il en est de même pour le LST : le mot *identité*, élément de notre lexique, peut renvoyer (outre son sens transdisciplinaire de 'similitude') à 2[30] ou 4[31] autres acceptions (selon la référence dictionnairique), y compris celle renvoyant au 'caractère permanent et fondamental de quelqu'un, d'un groupe', particulièrement fréquente dans les écrits de SHS.

Cette polysémie, chiffrée à trois acceptions par mot par Delbecque (2006), est en moyenne de deux acceptions par mot si nous comparons notre liste de 493 lemmes aux 1099 entrées du *DEM* correspondantes, avant le filtrage des acceptions relevant du LST. Ces 1099 entrées ne sont en effet pas obligatoirement transdisciplinaires et/ou présentes dans notre corpus. La phase de filtrage manuel des acceptions effectivement mobilisées dans notre corpus nous permet d'ailleurs d'aboutir à une polysémie moindre, comme les sections suivantes le détaillent.

---

[28] Ainsi, le nom *relation* apparaît 3788 fois dans le corpus (contre 88 pour *divergence*).
[29] Il est à noter que les concordances ne sont ni triées par texte, ni par discipline.
[30] Selon le *DEM* : http://rali.iro.umontreal.ca/DEM/alphabetique/I.html#identite [consulté le 30/01/2015]
[31] Selon le Larousse en ligne : http://www.larousse.fr/dictionnaires/francais/identité/41420 [consulté le 30/01/2015]

Nous étudions donc le sens des mots dans une approche contextuelle, à l'appui du corpus, à la manière de Cruse (1986).

Les acceptions sont confrontées à des tests linguistiques en vue de les délimiter. L'analyse sémantique que nous effectuons sur les noms se base principalement sur les concordances et l'emploi des mots en contexte. Nous prenons également en compte les relations lexico-syntaxiques les plus significatives données par le *Lexicoscope* pour identifier les sens présents, tel qu'illustré ci-après.

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log
sujet_NOUN	actualité_NOUN	U3_DE_NMOD ~U3_DE_NMOD	96	11998	1050	7556564	3	599,0554	1
sujet_NOUN	aborder_VERB	~U3_DEEPOBJ NMOD ~OBJ ~U3_SUBJ ~U3_OBJ ~U3_DE_VMOD ~DEEPOBJ ~SUBJ	83	11998	1751	7556564	6	407,6283	2
sujet_NOUN	sur_PREP	PREPOBJ	165	11998	24351	7556564	10	228,1723	3
sujet_NOUN	photographier_VERB	~U3_DEEPOBJ NMOD ~U3_OBJ	26	11998	96	7556564	1	223,2962	4
sujet_NOUN	parlant_ADJ	U3_ADJMOD	21	11998	50	7556564	3	202,8077	5

Illustration 3.2: Extrait du lexicogramme de sujet

Nous pouvons observer ci-dessus les cooccurrents syntaxiques statistiquement significatifs pour le nom *sujet*. La seconde colonne renseigne le lemme et la catégorie syntaxique du cooccurrent alors que la colonne *f.deprels* permet de connaître le type de relation de dépendance reliant les deux mots. Viennent ensuite, notamment, dans les colonnes suivantes, les informations sur le nombre d'occurrences de la collocation (*f*), du mot pivot (*f1*), en l'occurrence *sujet*, puis le nombre d'occurrences du collocatif (*f2*).

Cet extrait de lexicogramme, ou ensemble de cooccurrents syntaxiques statistiquement significatifs, nous permet ainsi de visualiser rapidement les emplois les plus fréquents du nom *sujet*. Les relations lexico-syntaxiques préférentielles sont définies à travers le cooccurrent syntaxique (2e colonne, par exemple *actualité_NOUN*) et le type de la relation de dépendance (3e colonne, par exemple *PREPOBJ*).

L'expression semi-figée *sujet d'actualités* est ainsi la cooccurrence dont le rapport de vraisemblance est le plus élevé. Cette acception de *sujet* au sens de 'thème' est également représentée par la cooccurrence *aborder un sujet, le sujet abordé*. La relation objet entre *sujet* et *photographier* peut être interprétée comme renvoyant à l'acception spécifique au domaine des arts plastiques. Enfin, la collocation *sujet parlant* semble convoquer l'acception d'un « être soumis à observation scientifique ».

Ces différents indices des acceptions mobilisées en corpus offrent ainsi un résumé de la polysémie des éléments du LST.

Un aspect important dans notre méthodologie est donc le fait de se baser principalement, en plus de la ressource lexicale du *DEM*, sur le corpus et non sur notre seule intuition.

Les informations de domaine du *DEM*, reportées dans le tableau 1 en annexe, nous permettent d'éliminer les sens ne relevant pas du lexique transdisciplinaire (tels ceux des domaines de l'aéronautique, la botanique, la zoologie, etc.). Le tableau ci-après présente un extrait d'acceptions du *DEM* possibles pour les noms LST.

Lemme	**Domaine (*DEM*)**	**Sens (*DEM*)**
angle	BAT	Coin, encoignure
annexe	BAT	Dépendance
avancée	BAT	Saillie de mur
construction	BAT	Édifice, maison
degré	BAT	Marche d'escalier
dépendance	BAT	Local annexe à résidence
seuil	BAT	Entrée de maison
suite	BAT	Appartements dans hôtel

Tableau 3.2: Extrait d'entrée du DEM pour les lemmes du LST

Ces acceptions ont ainsi été rejetées du fait de leur appartenance au domaine du bâtiment (BAT).

Nous prenons en compte, pour faire la correspondance entre les entrées lexicales du *DEM* et les lemmes-acceptions du LST présents dans le corpus, le domaine et le contenu du champ *Sens*. Ce champ consiste en une courte définition ou une liste de synonymes. Ceci nous permet d'identifier les situations pour lesquelles, à une acception présente dans le corpus, ne correspond aucune entrée du *DEM*. Ainsi, *pair* n'a pas d'entrée dans le *DEM* correspondant au sens de 'personne', idem pour *lecture* dans son sens de 'manière de comprendre'. Or ces acceptions sont attestées dans le corpus[32]. Nous avons, dans de tels cas, manuellement ajouté une entrée à notre lexique, en donnant une définition succincte.

[32] Comme dans la phrase : *On voit donc se confirmer une première lecture possible du sens à donner à la participation des animaux à l'Alliance.* [Albert, J-P. (2009). Les animaux, les hommes et l'Alliance. L'Homme, 189, 81-114.]

Nous ne retenons que les acceptions renvoyant au LST, d'autres acceptions, notamment terminologiques, pouvant être présentes dans le corpus d'analyse. Dans les cas de polysémie transdisciplinaire avérée (plusieurs sens présents pour un même mot, répondant chacun aux différents critères de fréquence et de dispersion[33]), nous opérons un dégroupage seulement si la distinction entre les sens est aisément réalisable, sans devoir faire appel à une interprétation personnelle pouvant fausser l'intention originelle de l'auteur. Cruse (1986) propose alors un test, lié au concept de spectre sémantique[34], pour valider le dégroupement d'acceptions et utilise ainsi des constructions zeugmatiques :

* *Un moustique est entré dans la bouche d'égout puis dans celle de l'enfant.*

Ce test d'identité permet ainsi de conclure à l'existence d'une ou plusieurs acceptions.

Il fait partie des trois tests généralement utilisés pour distinguer la polysémie (Goossens, 2011, p. 12-17) :

- test logique : si un mot renvoie à plus d'un sens, il peut être employé dans une assertion à la fois vraie et fausse[35] ;
- test linguistique : test zeugmatique, de co-prédication (test syntagmatique coordonnant deux sens distincts d'un mot) ;
- test définitionnel : le mot est polysémique si plus d'une définition est nécessaire pour dénoter son sens.

Malgré les tests, il existe des unités lexicales pour lesquelles plusieurs sens sont difficilement différenciables[36] en contexte. Dans de tels cas, aucun besoin particulier ne justifiant la prise en compte de deux sens très proches, nous faisons le choix de n'en garder qu'un, le plus générique, i.e le premier donné dans la ressource dictionnairique de référence.

---

33 Par exemple *sens* dans ses deux acceptions : *signification* et *orientation*.

34 Sense-spectra : continuum d'acceptions permettant de représenter la polysémie.

35 Nous pouvons citer l'exemple donné par Goossens (2011, p. 12) avec louer$_{\text{donner/prendre contre loyer}}$ et louer$_{\text{admirer}}$: *Je <u>loue</u> ses services mais je ne le <u>loue</u> pas*

36 Par exemple *nécessité,* qui a pour entrées : *caractère nécessaire de quelque chose* / *obligation*

Un exemple de ce cas de figure est le nom *cadre* qui a notamment pour entrée dans le *DEM* (et dont on trouve les correspondantes dans *Le Trésor de la Langue Française Informatisé*[37] en ligne) :

- *ce qui délimite, borne* (pour le *TLFi,* niveau A : [Désigne un objet délimitant] ;
- *milieu, contexte* (pour le *TLFi,* niveau B : [Désigne un domaine délimité].

Étant donné les difficultés que nous avons éprouvées pour différencier ces deux acceptions dans les concordances, présentées dans l'illustration 3.3, nous avons opté pour un regroupement.

XML_scienceseducation.xml-s6690	Il s' agit donc d' un moment du processus où les	cadres	prennent acte d' une crise antérieure , reconnaissent que la réalité sociale résiste à la volonté institutionnelle , cherchent des idées pour fabriquer l' outil qui servira au bilan final .
XML_histoire.xml-s6695	Par conséquent , au sein même de la mouvance communiste , « l'autodéfense de masse » , reposant sur l' activisme d' une minorité de	cadres	et de militants , conserve un caractère contingent , aggravé par les méthodes conspiratives de la direction Barbé - Célor .
XML_anthropologie.xml-s6718	Je cherche à comprendre des rapports paradoxaux à la liberté d' expression , à la lumière des courants - permanents par leur présence , mais variables quant à leur poids - qui modifient les comportements , et aussi par rapport aux conséquences des structures économiques actuelles , qui dépassent le	cadre	national , d'où la nécessité de quelques rappels historiques .
XML_histoire.xml-s6721	Le rôle accordé jusque-là aux usines est donc fortement relativisé puisque l' organisation de l' autodéfense doit désormais privilégier le	cadre	de l' agglomération .
XML_histoire.xml-s6740	Ainsi , en confiant l' organisation du service d' ordre à des	cadres	appartenant aux différents échelons de l' appareil , les dirigeants communistes estiment limiter au mieux les risques politiques et répressifs posés par l' autodéfense .
XML_histoire.xml-s6752	Alors que la hantise de l' extrême droite ne cesse de croître dans les rangs des forces de gauche , l' autodéfense est désormais présentée par le PCF comme un	cadre	d' action unitaire « antifasciste » pour construire un vaste organisme de protection , baptisé « défense populaire » .
XML_scinfo.xml-s6800	Cette tendance , nommée « Bring Your Own Device » ou BYOD et qui repose sur l' utilisation par les salariés de leur matériel informatique personnel dans un	cadre	professionnel , répond à de nouveaux besoins des utilisateurs : accessibilité aux données (notamment les boîtes mail) en tout temps et tout lieu , télétravail , etc .
XML_scinfo.xml-s6806	Une étude menée par l' IFOP indique aussi que les	cadres	tirent des bénéfices de l' omniprésence des technologies de l' information , notamment pour mieux s' organiser (72 %) , être plus réactifs (81 %) et plus productifs (60 %) [7] .

Illustration 3.3: Concordances du nom cadre

Sur le modèle des prototypes conceptuels, représentants de catégorie, Delbecque (2006), parle alors de prototypicité lexicale pour définir le sens central d'un mot polysémique. Nous déterminons ce sens en identifiant les acceptions les plus fréquentes dans les concordances observées. Dans les cas où le mot recouvre

[37] Disponible ici : http://atilf.atilf.fr/tlf.htm [consulté le 07/09/2016]

effectivement plusieurs acceptions proches, nous décidons ainsi de ne garder que la plus représentative. Ainsi, le nom *composition* a 7 entrées dans le *DEM* (*composition scolaire, musicale, typographique, etc.*). Nous ne retenons que la plus générique, dont la définition est 'assemblage, format'.

Nous prenons également le parti de ne garder que l'acception processive (au détriment de l'acception résultative) pour les noms déverbaux, lorsque ce sens d'action est attesté dans le corpus. Ce glissement sémantique étant généralisé (*étude, analyse, augmentation, évaluation*), dans notre lexique, nous optons pour un traitement spécifique de cette variation lors de l'analyse des routines plutôt qu'une multiplication des sens (que nous devons traiter et classer manuellement). Nous pouvons distinguer dans les exemples 1 et 2 un emploi de *interprétation* et *classification* dans leur acception processive, contrastant avec les exemples 3 et 4 d'acception résultative.

1. *Pour faciliter l'interprétation, nous calculons un effet marginal de la manière suivante* [38] *;*
2. *Les deux premières couches de l'extension [...] rendent possibles des déductions supplémentaires telles que la vérification de la cohérence d'un schéma, la classification automatique des types*[39] *;*
3. *Nos observations et interprétations indiquent que les relations de l'"être nouvelle" avec les dispositions discursives des ONG en tant que protagonistes sont positives*[40] *;*
4. *Progressivement, il s'attelle à une révision systématique, ambitieuse et de longue haleine, des classifications linguistiques en vigueur en Amérique du Sud*[41]

Cette prédominance de l'ambiguïté action/résultat a d'ailleurs été montrée par Jacquey (2013) et est étudiée par Flaux & Stosic (2014) qui montrent par

---

38 Munier F., & Rondé P. (2000). Mimétisme rationnel et connaissance : une analyse empirique. *Économie Appliquée*, vol. 53, pp. 93-116.

39 Bachimont, B., Gandon, F., Poupeau, G., Vatant, *et al.* (2011). Enjeux et technologies : des données au sens. *Documentaliste-Sciences de l'Information*, *48*(4), 24-41.

40 Ferreira, J. (2010). La médiatisation : de la production à la circulation des nouvelles. *Les Enjeux de l'information et de la communication*, *2010*(1), 121-143.

41 Laurière, C. (2008). L'anthropologie et le politique, les prémisses. L'*Homme*, (3), 69-92.

exemple que les noms d'idéalités[42] sont souvent des déverbaux d'action correspondant à l'acception résultative. Notre lexique intègre ainsi un sous-ensemble de ces noms d'idéalités (*définition, affirmation, expression*), pour lesquels nous procédons à un traitement particulier en gardant d'une part, l'acception processive (appartenant généralement à la classe {processus_cognitif}), et d'autre part, l'acception résultative. Ce choix est guidé par plusieurs considérations d'ordre pratique. En premier lieu, au niveau de l'application du LST dans l'extraction terminologique, nous avons observé que le nom *définition* dans son acception résultative, est fréquemment introducteur de terme (au même titre que d'autres membres de la sous-classe {communication_expression/formulation} tels *terme* et *expression*), comme dans l'exemple suivant :

- *Nous adoptons ici la définition de la forêt$_{terme}$ proposée par la FAO*[43]

En second lieu, dans la perspective de l'étude des routines sémantico-rhétoriques de l'écrit scientifique, nous avons pu constater que le nom *affirmation* (comme les noms *définition* et *expression*) rentre fréquemment dans des constructions dont les fonctions métatextuelle ou métadiscursive nous paraissent dignes d'intérêt, tel que l'illustrent les exemples ci-dessous :

- *Or une telle affirmation ne relevait pas de l'évidence lorsque son auteur s'est risqué à la formuler*[44] ;
- *Reprenant, comme Sauvy, les affirmations de Colin Clark concernant les tendances d'évolution des trois grands secteurs, Fourastié affirme que l'évolution des emplois suivra en France la même voie qu'aux États-Unis*[45]

Pour les deux exemples précédents, *définition* et *affirmation*, une acception est ainsi classée dans {processus_cognitif}, dont la définition est 'processus cognitif permettant de structurer/analyser les observables et/ou objets construits'. Le test d'appartenance à cette classe est alors : *une personne fait/effectue/propose*

---

42 Les noms d'idéalités sont des entités munies d'un contenu spirituel (Flaux & Stosic, 2014) à interpréter : *sonate, roman, témoignage, expression*.

43 Fourault-Cauët, V. (2010). Le paysage, outil de territorialisation et d'aménagement incomplet pour les forêts méditerranéennes ?. *Annales de géographie* (No. 3, pp. 268-292). Armand Colin.

44 Lallement, M. (2005). Relations industrielles et institutionnalisme historique aux États-Unis. L'*Année sociologique*, *55*(2), 365-389.

45 Chapoulie, J. M. (2007). Une révolution dans l'école sous la Quatrième République ?. *Revue d'histoire moderne et contemporaine*, (4), 7-38.

un N. Ces deux noms, polysémiques au niveau du LST (i.e. ayant plus d'une acception transdisciplinaire), ont également une acception appartenant à la sous-classe {communication/formulation}, dont la définition est 'élément d'un énoncé', et dont le test d'appartenance est : *Dans l'article, l'auteur utilise/énonce un N*.

À la suite de ces filtrages, l'ensemble des noms du LST est ainsi composé de 531 acceptions correspondant à 493 lemmes. Les concepts permettant « une vision classificatoire du monde » (Pecman, 2004b), nous poursuivons ce travail d'analyse sémantique du lexique nominal en procédant à la définition de classes (ou concepts), groupes de noms du LST partageant des propriétés sémantiques et de combinatoire. Ce typage sémantique est également effectué en collaboration avec plusieurs membres de notre équipe pour les catégories des adjectifs, noms et verbes, (1500 acceptions pour l'ensemble des catégories), comme nous le détaillons dans la section 3.3.3.

3.3.2 Classification des acceptions des noms du LST

Notre ressource des noms du LST, avant la définition de la classification sémantique, prend pour entrée un lemme associé à une acception, identifiée par un numéro et une courte définition. Cette définition est issue du *DEM* ou rédigée par nos soins lorsque notre ressource dictionnairique de référence ne dispose pas d'entrée correspondant à l'acception en question.

Nous tirons parti d'autres informations du *DEM* pour effectuer les regroupements en classes : les informations de domaine et de construction (voir tableau 1) permettent ainsi de regrouper des mots renvoyant à un même domaine et partageant des constructions identifiées par le *DEM*. Nous observons notamment que dans le *DEM,* les noms *position, extérieur, source* et *rang* sont tous rattachés au domaine 'LOC' (lieu) et sont inscrits dans une construction définie comme suit : *situer N*. Ces informations sémantico-syntaxiques nous servent ainsi d'amorce pour la définition de certaines classes et sous-classes. En l'occurrence, ces quatre noms sont regroupés dans la même sous-classe {espace/localisation} définies par les propriétés suivantes :

- a pour membres : *centre, extérieur, niveau, plan, pôle, position, rang, source* ;

- a pour définition : ‘place déterminée dans un espace’ ;
- a pour test d’appartenance : *localiser le N / Au|à N de.*

L’ensemble des membres de cette sous-classe reçoivent ainsi pour attribut de classe la valeur {espace} et pour attribut de sous-classe la valeur {localisation}, en plus de leur attribut de lemme et d’acception.

L’attribution, pour chaque entrée du LST, d’une classe et d’une sous-classe implique ainsi l’élaboration d’une typologie simple des noms du LST, sur deux niveaux. Les entrées du LST sont alors organisées en groupe de co-hyponymes, à une même classe et/ou une même sous-classe. Un lien d’hyponymie/hyperonymie structure l’ensemble des entrées au travers des liens de sous-classe/super-classe.

Notre ressource ainsi enrichie sémantiquement autorisera un accès onomasiologique, via les classes et sous-classes, et sémasiologique, à partir des entrées du LST.

3.3.2.1 Cadre et fondements de la classification

Avant de présenter dans les sections suivantes la méthodologie de définition des classes et de regroupements des entrées, nous définissons dans cette partie le cadre dans lequel s’inscrit l’élaboration de notre classification. Ce cadre s’appuie sur de précédents travaux tirant parti des propriétés distributionnelles pour opérer des regroupements lexicaux.

Nous nous référons ici à la distinction entre classification et catégorisation, détaillée par Habert & Zweigenbaum (2003). La classification n’implique pas de types à priori, alors que la catégorisation se base elle sur des catégories prédéfinies. Nous nous situons dans une tâche hybride, tenant principalement de la classification, n’ayant pas un ensemble préconçu des types, et procédant au regroupement de mots partageant des propriétés syntaxiques et sémantiques. Mais ce travail participe également de la catégorisation puisque nous prenons comme point de départ la classification opérée par Tutin (2007c), dans son étude sur les méthodes de l’analyse distributionnelle appliquées aux noms du LST. Nous profitons également du travail réalisé par Magdalena Augustyn sur la classification des adjectifs du LST. Sa typologie est en partie basée sur les informations

sémantiques présentes dans GermaNet[46], adaptation pour l'allemand de *Wordnet* (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010).

Notre classification se base sur l'analyse distributionnelle, à la manière de la typologie de Flaux & Van de Velde (2000). Les classes et sous-classes réunissent des mots conceptuellement proches, s'organisant, comme le rappelle Delbecque (2006, p. 62), dans un champ lexical et renvoyant à un même domaine conceptuel. L'identification des propriétés communes découle de l'observation du corpus et des profils combinatoires, et repose ainsi sur l'usage en contexte des éléments que nous classons. L'ajout de propriétés sémantiques se fait ainsi selon une stratégie endogène, i.e. à partir des textes du domaine, selon les recommandations de Bourigault, Aussenac-Gilles, & Charlet (2004).

Outre l'aspect onomasiologique, nous prenons en compte, pour certaines classes, les propriétés de prédication, en identifiant les constructions privilégiées. Ainsi notre classe {déterminant} est définie par sa fonction de détermination à l'intérieur d'une structure du type *Dét* $N_{déterminant}$ *de SN*. Ce postulat des propriétés syntaxiques définissant les classes sémantiques (Le Pesant & Mathieu-Colas, 1998) est à l'œuvre dans le *LVF* et le *DEM* utilisés pour le dégroupement en acceptions des éléments du LST.

Notre classification intègre donc des classes relevant de deux typologies, ontologique et fonctionnelle. Ainsi, les grandes classes {objet_scientifique} et {processus_cognitif} sont d'ordre ontologique, car elles correspondent à un découpage de la réalité en catégories complémentaires (Huyghe, 2015, p. 7). À l'inverse, la classe {collectif_partitif} relève de la typologie fonctionnelle puisqu'elle est définie en termes de prédication (de la même manière que les noms partitifs). Comme nous l'avons évoqué précédemment (section 3.2.2.1), plusieurs travaux ont déjà expérimenté cette association de classes de types ontologique et fonctionnel.

Comme nous l'avons vu, la classification peut se faire selon différents critères : textuels, morphologiques, syntaxiques, sémantiques, etc. Nous utilisons d'une part les informations sémantiques issues du *DEM*, et d'autre part les propriétés lexico-syntaxiques extraites du corpus. Ces propriétés sont

[46] Disponible en version électronique ici : http : ://www.sfs.uni-tuebingen.de/lsd/ [consulté le 30/07/2015]

désambiguïsées manuellement lors de l'analyse des concordances. Ceci permet d'éviter l'écueil de la prise en compte d'associations non désambiguïsées qui peuvent mener à des regroupements hétérogènes. Ainsi, si l'on observe les cooccurrents syntaxiques de la préposition *selon*, nous retrouvons une grande diversité de noms : *conception, auteur, définition, lieu, position, valeur, état, etc.* Nous pouvons regrouper ces noms en deux classes correspondant aux acceptions de *selon* présentes dans le corpus :

- Sens 'se référant à' en cooccurrence avec *conception, auteur, définition* ;
    - *Selon cette conception, le premier pilier repose sur le régime général.*[47]
- Sens 'en fonction de' en cooccurrence avec *lieu, position, valeur, état* ;
    - *Il est équivalent de déterminer un salaire selon le lieu de travail.*[48]

Fabre & Bourigault (2006, p. 125) notent également que « le partage de dépendants identiques n'est pas toujours un critère suffisant pour détecter une proximité sémantique. » Diwersy & Kraif (2013) font le même constat et préviennent que les observations peuvent être faussées par la polysémie des mots étudiés et de leurs cooccurrents. L'étape d'analyse et de validation manuelle nous permet de gérer ces problèmes de polysémie et de regroupements incohérents. De plus, en comparaison de classifications basées sur des cooccurrences graphiques, notre méthode fondée sur les relations de dépendance permet d'éviter plusieurs problèmes récurrents, tels la définition de la taille de la fenêtre à considérer ou le choix de la direction de cette fenêtre. L'orientation de l'empan de recherche a d'ailleurs un impact non négligeable, comme le notent Tanguy *et al.* (2015, p. 119) qui remarquent par exemple que la direction de la fenêtre (gauche, droite ou bidirectionnelle) est le paramètre critique pour la classification des adjectifs (qui peuvent être postposés ou antéposés).

Notre objet d'étude étant le lexique scientifique transdisciplinaire, nous nous concentrons sur l'étude de ses acceptions, cooccurrences et propriétés transdisciplinaires. Nous veillons ainsi à ne pas prendre en compte des spécificités disciplinaires, identifiables par leur faible dispersion. Ainsi la collocation *sujet*

---

47 Rémond, A. (2007). Le rôle politique des sondages. Retour sur la réforme des retraites de 2003. *Actes de la recherche en sciences sociales* 4/2007 (n° 169), p. 48-71

48 Zenou, Y. (2000). Externalités spatiales, économies d'agglomération et formation endogène d'une ville monocentrique. *Annales d'Economie et de Statistique*, N°58, P.233-251.

syntaxique ne sera pas prise en compte lors de l'étude du nom *sujet* car elle ne concerne qu'un seul des dix corpus disciplinaires, le sous-corpus d'articles en linguistique. Nous ne tenons ainsi compte ni de cette acception terminologique de *sujet* ni de la cooccurrence *sujet syntaxique*, terminologique et disciplinaire. De même, si l'on considère le nom du LST *proximité,* un de ses cooccurrents les plus fréquents est *police,* renvoyant à l'expression *police de proximité*. Cette association, dont la dispersion est de 2[49], ne correspond pas à une cooccurrence transdisciplinaire et ne relève pas du LST. Nous ne le considérons donc pas comme une propriété pertinente pour le mot *proximité* lors de la phase de classification.

3.3.2.2 Définition des classes sémantiques

Nous avons vu que les travaux en sémantique nominale portent sur des classes multiples. Parmi ces dernières, Benninger & Theissen (2013) distinguent deux grands groupes : ceux concernant des sous-catégories identifiées (noms d'affects, d'idéalités, de qualités…) et ceux formant des classes hétérogènes, rassemblant noms sommitaux (noms référant à une catégorie : *matière, odeur, couleur*) et noms généraux (*fois, quantité, chose*). Nous intégrons dans notre classification ces deux types de classes, sans forcément systématiquement inclure les premières dans les secondes. Cette hétérogénéité est une des explications de la hiérarchie peu profonde de notre classification, à deux niveaux, et de la granularité irrégulière selon les classes (contenant de une à 13 sous-classes au final, dont l'extension varie de 1 pour {état_qualité/changement_négatif} à 20 pour {quantité/rapport}.

La définition de sous-classes est ici motivée par plusieurs raisons :

- la grande hétérogénéité de certaines classes telles que {objet_scientifique} ou {processus_cognitif} ;
- la difficulté à identifier des propriétés distributionnelles partagées par l'ensemble des membres d'une même classe (difficulté augmentant avec le nombre d'éléments appartenant à une même classe) ;
- la perspective d'utilisation des sous-classes pour l'extraction de routines. Le fait de disposer de classes et de sous-classes nous permet d'envisager des

[49] La cooccurrence *police de proximité* (76 occurrences) a ainsi 75 occurrences dans le sous-corpus de sciences politiques et une dans le sous-corpus de psychologie.

extractions à deux niveaux de granularité. D'une part, l'extraction de routines basées sur les classes nous assure un regroupement de constructions supérieur à des routines basées sur les sous-classes. D'autre part, l'extraction de ces dernières nous permet d'envisager l'identification de constructions plus précises, et plus contraintes, dont les propriétés sémantiques et rhétoriques seront plus homogènes.

Comme point de départ de la classification nous nous appuyons en partie sur la typologie de Tutin (2007c), qui définit 7 grandes classes : processus de l'activité scientifique / objets construits / observables / supports de rédaction / caractérisation / acteurs scientifiques / relation logique. Tutin définit ces classes à partir d'une liste de noms filtrés par leur fréquence absolue et leur dispersion. Sa typologie est élaborée pour les 83 noms les plus fréquents du corpus, éléments dont les nombreuses occurrences permettent d'aboutir à des profils combinatoires suffisamment riches pour être exploités. Les mots peu fréquents de notre liste posent le problème du manque de données distributionnelles. Leur trop faible fréquence ne nous permet pas d'aboutir à une liste de relations lexico-syntaxiques pouvant aider à leur classification. Dans son article, Tutin montre d'ailleurs qu'à peine la moitié des regroupements effectués sont en accord avec la classification manuelle utilisée comme étalon.

Ces observations expliquent notre choix d'une classification manuelle, fondée sur les profils combinatoires calculés automatiquement.

Nous reprenons certaines classes sémantiques du LST définies par Tutin :

- {communication_support} pour les noms de support de rédaction ;
- {état_qualité} pour les noms de caractérisation ;
- {objet_scientifique} regroupe les objets construits et observables de l'activité scientifique ;
- {personne} pour les acteurs de l'activité scientifique ;
- {processus_cognitif} pour les processus de l'activité scientifique ;
- {relation} pour les relations logiques ;

En plus de ces adaptations, nous définissons, à la suite de l'identification de groupes de mots sémantiquement proches, les classes suivantes : {communication_contenu}, {déterminant}, {espace}, {état_humain}, {collectif_partitif}, {processus_évolutif}, {processus_générique_chose}, {processus_humain}, {quantité}, {temporalité}, {temporalité_espace}.

Au final, notre typologie des noms du LST est constituée de 17 grandes classes et 80 sous-classes. Le fait de recourir à un plus grand nombre de classes est à mettre en regard des effectifs à classer : 85 noms dans la typologie de Tutin contre 531 entrées nominales pour notre typologie.

La dénomination des classes et sous-classes comporte une part de subjectivité, et ne prétend pas au consensus.

Les étiquettes que nous apposons pour les classes et sous-classes sont purement conventionnelles. Nous utilisons ainsi comme étiquettes de classes des noms renvoyant quelquefois à des concepts explicités dans des travaux antérieurs (par exemple {état}, {collectif}). Nous ne prétendons pas à une redéfinition de ces classes déjà étudiées dont nous reprenons seulement l'étiquette.

À l'inverse, il peut arriver qu'une de nos classes recouvre partiellement ou totalement une ou plusieurs catégories nommées autrement par ailleurs, outre les correspondances déjà établies avec la typologie de Tutin. Nous pouvons ainsi relever les correspondances suivantes :

- nos sous-classes {espace/domaine} et {espace/localisation} correspondent aux classes habituellement nommées *lieu* et *territoire,* toutes deux en relation d'hyperonymie comme le rappelle Huyghe (2015) ;
- notre sous-classe {collectif_partitif/partitif} est nommée dans d'autres travaux noms *partitifs*, dont une des propriétés est l'emploi possible dans les anaphores associatives[50] ;
- notre classe {déterminant} est à rapprocher de celle des noms *généraux* qui sont, selon Adler (2012), à la frontière des items lexicaux et grammaticaux. Cette classe des noms *généraux*, qui pose une condition d'homogénéité, se combine avec *se rassembler*, *affluer*, et s'emploie comme déterminant complexe ;

[50] Ainsi, la phrase suivante : *Le livre traite de linguistique. La première section débute par…*

- certaines de nos sous-classes recoupent les *shell nouns*[51] de Schmid (2000) : les noms des sous-classes {objet_scientifique/cas} : *fait, situation, condition, cas*, et {objet_scientifique/thème} : *problème, question, sujet, idée* ;
- la classe nommée noms d'objets *mentaux*[52] par Kleiber *et al.* (2012) trouve une équivalence partielle avec notre classe {objet_scientifique} (*thèse, solution, concept*) *;*
- notre classe {état_qualité} a une intersection importante avec celle des noms de propriétés, noms démunis d'extension temporelle et caractérisés par le fait de pouvoir être modifiés par des intensifs (Kleiber *et al.*, 2012).

Dans la lignée de Flaux & Van De Velde (2000), la classification sémantique prend appui sur la combinatoire des unités lexicales à classer. À la différence de leurs travaux, nos traitements sont ici guidés par le corpus. Les étapes de définition des classes et de regroupement des acceptions sont effectuées en parallèle, se nourrissant mutuellement. En effet, l'identification de nouveaux groupes homogènes, à travers leurs propriétés communes, permet d'affiner les classes existantes ou d'en créer de nouvelles. Ces redéfinitions autorisent à leur tour des regroupements plus fins.

Nous présentons dans le tableau ci-dessous un extrait de la typologie des noms du LST structurée en 17 classes et 80 sous-classes. L'intégralité de la typologie est consultable en annexe, section A.VII.

Les classes et sous-classes sont ainsi définies :

- en intension → à travers une courte définition et un ou plusieurs tests vérifiant certaines propriétés linguistiques ;
- en extension → en recensant les unités lexicales qui les composent.

---

[51] Noms abstraits généraux répondant au test : *the noun_shell be that clause / the problem was that we…*

[52] Noms répondant au test de localisation dans un espace abstractif : *cette idée se trouve dans la démonstration de X.*

Classe	Sous-Classe	Noms – Acceptions Termith
collectif_partitif 'ensemble ou élément d'un ensemble' *diviser en plusieurs N*	**collectif** 'ensemble d'éléments' *regrouper dans un N / Plusieurs SN constituent un N*	`catégorie, classe_1, ensemble_2, groupe_2, population, série_2, totalité_2`
	partitif 'élément, portion d'un tout' *être composé de N / extraire N d'un tout*	`échantillon, élément, extrait, partie_2, section_2, segment, unité_2`
communication_support : 'support de communication et de transfert des idées (selon le moyen de transmission)' *Le N est consacré à*	**document** 'document écrit' *Lire, publier un N*	`article, document, documentation, essai_1, littérature, note, ouvrage…`
	exposé 'événement scientifique' *un N a lieu / Pendant un N*	`colloque, conférence`
	graphique 'représentation graphique éclairant le texte' *Le N {numéral} montre / cf N, sur N, dans N*	`figure_1, illustration_1, image_1, motif_1, tableau`
	section 'sous-partie d'un document écrit' *Dans le N de l'ouvrage / cf N*	`annexe, bibliographie, chapitre, conclusion_1, développement_1, introduction_1, section_1`

Tableau 3.3: Extrait de la classification sémantique des noms du LST

3.3.2.3 Regroupement des acceptions

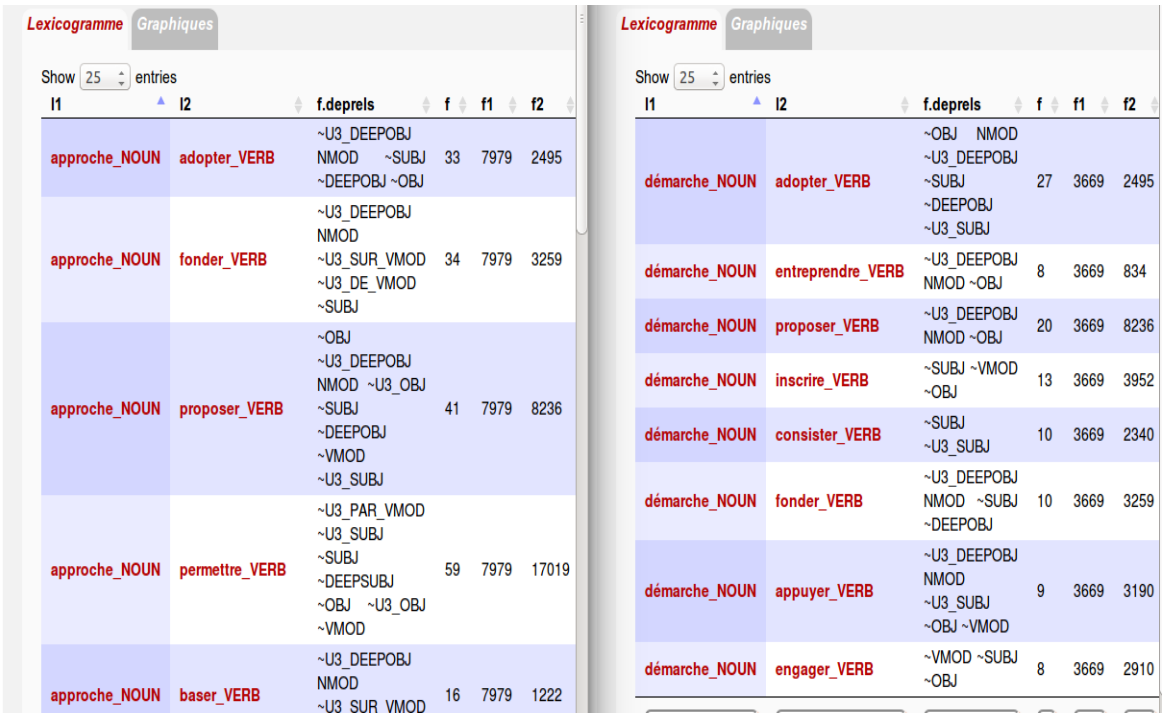
Nous présentons ici l'étape de classification sémantique du lexique nominal transdisciplinaire. Il nous faut cependant rapidement aborder l'aspect transcatégoriel des traitements et de la ressource. Les mots reliés morphologiquement (par exemple *analyse* et *analyser*), et renvoyant à une acception similaire, seront identifiés manuellement et mis en correspondance. Nous présenterons dans la partie 3.3.3 le travail fait sur l'harmonisation des classes et sous-classes sémantiques au niveau des différentes catégories traitées. Ainsi, au niveau des liens lexicaux présents dans notre lexique des noms, un mot sera en relation avec les autres membres de sa sous-classe, ceux de sa classe, et éventuellement mis en correspondance avec un mot d'une catégorie différente (verbe, adjectif ou adverbe), ou un autre nom[53].

[53] Ainsi, *méthode* et *méthodologie* sont deux noms reliés morphologiquement partageant la même sous-classe {objet_scientifique/méthode}. À l'inverse, *différence* et *différenciation* ne partagent pas de traits sémantiques, et ont pour sous-classes respectives {relation/opposition} et {processus/inclusion_séparation}).

La complexité de la phase de classification ne relève pas exclusivement du caractère abstrait du lexique étudié et de la subjectivité inhérente à ce type de tâche. Ainsi, comme « le lexique n'est pas cette mosaïque parfaite où chaque item lexical trouverait sa "place" » (Delbecque, 2006, p. 68), la recherche d'une taxonomie unifiée est intrinsèquement complexe. Huyghe (2015, p. 6) le confirme en expliquant que « les différents types nominaux ne peuvent donc coexister dans une structure unique partitionnée en classes complémentaires ».

Nous avons également pu observer que cette complexité à classer des mots est d'autant plus grande que la fréquence des éléments à classer est faible. Ainsi, quel que soit le type de modèle distributionnel envisagé, les mots ayant peu d'occurrences dans le corpus d'analyse se révèlent difficile à traiter étant donné le nombre limité de propriétés lexico-syntaxiques extraites automatiquement. Tanguy *et al.* (2015) relèvent d'ailleurs que la fréquence est le paramètre le plus important dans la qualité de la classification, suivi par la catégorie syntaxique du mot cible.

Nous nous basons une nouvelle fois sur les données issues du corpus pour ce travail de regroupement. Les concordances et les profils combinatoires nous permettent d'identifier les constructions privilégiées communes aux mots sémantiquement proches. Le *Lexicoscope* permet l'identification de verbes que nous utilisons dans nos critères de classe et autorise également la comparaison des profils de plusieurs mots. Le recours à cet outil participe de notre approche de linguistique outillée, dans laquelle nous tirons parti du corpus arboré pour accéder aux usages du LST pour le genre de l'écrit scientifique, ainsi que nous pouvons l'observer dans l'illustration ci-dessous.



Lexicogramme | Graphiques

Show 25 entries

l1	l2	f.deprels	f	f1	f2
approche_NOUN	adopter_VERB	~U3_DEEPOBJ NMOD ~SUBJ ~DEEPOBJ ~OBJ	33	7979	2495
approche_NOUN	fonder_VERB	~U3_DEEPOBJ NMOD ~U3_SUR_VMOD ~U3_DE_VMOD ~SUBJ	34	7979	3259
approche_NOUN	proposer_VERB	~OBJ ~U3_DEEPOBJ NMOD ~U3_OBJ ~SUBJ ~DEEPOBJ ~VMOD ~U3_SUBJ	41	7979	8236
approche_NOUN	permettre_VERB	~U3_PAR_VMOD ~U3_SUBJ ~SUBJ ~DEEPSUBJ ~OBJ ~U3_OBJ ~VMOD	59	7979	17019
approche_NOUN	baser_VERB	~U3_DEEPOBJ NMOD ~U3_SUR_VMOD	16	7979	1222

Lexicogramme | Graphiques

Show 25 entries

l1	l2	f.deprels	f	f1	f2
démarche_NOUN	adopter_VERB	~OBJ NMOD ~U3_DEEPOBJ ~SUBJ ~DEEPOBJ ~U3_SUBJ	27	3669	2495
démarche_NOUN	entreprendre_VERB	~U3_DEEPOBJ NMOD ~OBJ	8	3669	834
démarche_NOUN	proposer_VERB	~U3_DEEPOBJ NMOD ~OBJ	20	3669	8236
démarche_NOUN	inscrire_VERB	~SUBJ ~VMOD ~OBJ	13	3669	3952
démarche_NOUN	consister_VERB	~SUBJ ~U3_SUBJ	10	3669	2340
démarche_NOUN	fonder_VERB	~U3_DEEPOBJ NMOD ~SUBJ ~DEEPOBJ	10	3669	3259
démarche_NOUN	appuyer_VERB	~U3_DEEPOBJ NMOD ~U3_SUBJ ~OBJ ~VMOD	9	3669	3190
démarche_NOUN	engager_VERB	~VMOD ~SUBJ ~OBJ	8	3669	2910

Illustration 3.4: Comparaison des lexicogrammes de approche et démarche

L'illustration ci-dessus offre un aperçu des lexicogrammes issus du *Lexicoscope* pour les noms *approche* et *démarche*. Les cooccurrents sont ordonnés, par défaut, par valeur décroissante du rapport de vraisemblance (LLR). Nous pouvons ainsi observer que ces deux noms, appartenant à la même sous-classe {objet scientifique/méthode} partagent un certain nombre de propriétés distributionnelles statistiquement significatives :

- ils sont dépendants dans la relation objet du verbe *adopter* ;
- ils sont dépendants dans la relation objet profond du verbe *fonder*[54] ;
- ils sont dépendants dans la relation objet du verbe *proposer*.

Les lexicogrammes facilitent ainsi l'identification de propriétés lexico-syntaxiques communes, à la base des regroupements en classes et sous-classes sémantiques.

L'outil permet de trier ces cooccurrences syntaxiques par fréquence de l'association ou du cooccurrent, par la valeur de la dispersion (nombre de sous-

[54] Cette relation correspond aux constructions passives telles : *Cette approche/démarche est fondée sur…*

corpus dans lesquels la cooccurrence est présente). Ce dernier critère permet ainsi de s'assurer de la transdisciplinarité de la relation et donc de prendre en compte dans le profil seulement des cooccurrences effectivement transdisciplinaires.

Ainsi, parmi les cooccurrents du nom *approche* se trouve l'adjectif *paysagère* (au rang 7 selon le calcul du LLR). Or, la dispersion de cette association est de 1 puisque l'ensemble des occurrences concerne le sous-corpus de géographie.

Les tests peuvent être du type : tout élément de la classe *X* peut être sujet/objet du verbe *V*. Cette propriété sur la nature des verbes cooccurrant avec les éléments d'une classe particulière est d'ailleurs mise en avant par Le Pesant & Mathieu-Colas (1998). Nous reprenons ce principe en le subordonnant aux données issues du corpus : les cooccurrences lexico-syntaxiques insérées dans les tests doivent apparaître dans les profils combinatoires et/ou les concordances des éléments de la classe. Ces tests peuvent alors prendre la forme d'une relation syntaxique binaire nom-verbe, de patrons morpho-syntaxiques (par exemple *Dét N de SN* pour {déterminant}) ou de routines plus complexes (X *est considéré comme Dét N de Y* pour {objet_scientifique/représentation}).

Si un mot correspond à la définition donnée pour la classe et répond positivement au(x) test(s) associé(s), il appartient alors à la classe en question. Nous avons également ajouté une définition de la classe consistant le plus fréquemment en une courte phrase reprenant l'étiquette sémantique, ou en une liste de membres prototypiques. Les sous-classes sont définies de la même façon. Un élément doit alors satisfaire aux définitions et aux critères de la classe et de la sous-classe pour être intégré à cette dernière.

Ainsi la classe {espace} sera définie par la périphrase 'espace concret ou milieu abstrait' ainsi que par le test *se situer dans* (semblable au test de repérage rappelé par (Huyghe, 2015)), ses sous-classes étant définies comme suit :

- {espace/domaine} : 'zone, partie d'un espace, d'un milieu' *;*
  - *Le N s'étend / Dans le N → cadre, champ, domaine, territoire* ;
- {espace/limite} : 'limites définissant un espace donné' ;
  - *Le N délimite l'espace → contour, frontière, limite* ;
- {espace/localisation} : 'place déterminée dans un espace' ;

- *Localiser le N / Au N de → centre, extérieur, niveau, position* ;

- {espace/orientation} : ‘sens, direction dans un espace donné’ ;
  - *Aller vers un N / s’orienter selon un N → axe, direction, sens, voie*.

Nous tenons également compte des relations sémantiques de synonymie et d’antonymie présentes dans le *DEM* afin d’assurer davantage la cohérence de la classification. En effet, en plus des rubriques de domaine et de construction, nous prenons en compte la rubrique *Sens*, renvoyant à une courte définition, constituée régulièrement par une liste de quasi-synonymes.

Si un nom du LST a pour quasi-synonyme un autre nom du LST dans cette rubrique, nous nous assurons dans la mesure du possible de leur co-présence dans une même sous-classe. Ainsi, le nom *expansion* a pour valeur dans la rubrique *Sens* du *DEM* : « *accroissement, augmentation* ». Ces quasi-synonymes sont inclus dans la même classe que l’acception (associée à un lemme) qu’ils définissent.

Nous nous assurons ainsi d’une certaine cohérence dans notre classification, en mettant en regard la constitution des classes et les informations sémantiques issues du *DEM*, notamment dans les rubriques de domaine, de construction et de sens, comme l’illustre le tableau suivant.

Nom	**Sens (*DEM*)**	**Sous-classe**
accroissement	‘augmentation’	{amélioration_augmentation}
amélioration	‘progrès’	{amélioration_augmentation}
augmentation	‘accroissement’	{amélioration_augmentation}
développement	‘croissance’	{amélioration_augmentation}
expansion	‘accroissement, augmentation’	{amélioration_augmentation}
extension	‘allongement’	{amélioration_augmentation}
multiplication	‘accroissement’	{amélioration_augmentation}

Tableau 3.4: Éléments de la classe {processus évolutif/amélioration_augmentation}

Ce travail visant à assurer la cohérence de notre typologie n’a pas pour but de pouvoir proposer, à des fins didactiques, un ensemble de co-hyponymes

mutuellement substituables lors d'exercice de production ou de compréhension écrite. Ce faisant, nous nous exposerions au risque de circularité, contre-productif pour tout type d'application didactique. De même, nous ne considérons pas que les courtes gloses associées aux différentes acceptions du LST soient adaptées à des applications didactiques, qui demanderaient alors des définitions plus fines. L'intérêt didactique de la ressource sémantique du LST repose plus sur la typologie en elle-même, au travers des ensembles de co-hyponymes définis par les classes et sous-classes.

Un dernier filtrage des acceptions est finalement effectué, dans le but de faciliter de futurs traitements. Ainsi, si un mot comporte plusieurs acceptions renvoyant à une même sous-classe sémantique, nous ne gardons qu'une acception, la plus générique. Ceci nous permet d'éviter d'introduire une ambiguïté en plus lors de l'extraction de routines. Ainsi, lorsque nous extrayons une routine sémantico-rhétorique, les traits de sous-classes présents dans la routine représentent d'une certaine manière un paradigme lexical. Par exemple, si nous considérons le test d'appartenance de la sous-classe {espace/localisation}, '*localiser le N*'. Ce test peut s'apparenter aux routines que nous souhaitons identifier. L'élément *'N'* peut alors se réaliser par l'un des éléments de cette sous-classe, à savoir *pôle, rang, source, niveau,* etc. Si un de ces lemmes avait plus d'une acception appartenant à cette sous-classe, nous ne saurions distinguer quelle acception s'inscrit effectivement dans cette routine. En choisissant de se restreindre à une acception par lemme par sous-classe, nous nous assurons de pouvoir faire le lien entre une routine et les acceptions potentiellement mobilisées dans cette routine. De plus, au niveau du traitement sémantique du LST, il nous semble judicieux d'adopter une telle granularité, afin de ne pas avoir à gérer une trop grande quantité de classes et sous-classes. En outre, dans la perspective d'une application didactique, une granularité trop fine serait peu exploitable par les apprenants.

Cependant, la distinction, pour un même lemme, d'acceptions proches appartenant à la même sous-classe, peut s'avérer utile en vue d'applications didactiques intégrant par exemple un accès à une définition. Notre but immédiat étant ici d'élaborer une classification favorisant le repérage et l'extraction de routines, nous ne gardons alors qu'une seule entrée, correspondant (comme expliqué page 142 avec l'exemple de *cadre*) à l'acception la plus générique. Ainsi,

parmi les acceptions de *matériau* dans le *DEM*, deux entrées correspondent à des sens transdisciplinaires effectivement présents dans notre corpus :

- matière pour fabrication ;
- matière de base pour documentation.

Ces deux acceptions sont associées à la sous-classe {objet_scientifique/donnée}.

Les applications ultérieures de notre classification guident ainsi sa constitution.

Ces différents critères de filtrage d'acceptions, ainsi que la méthodologie de classification, ont été définis pour l'ensemble des catégories des mots simples du LST afin de s'assurer de l'homogénéité transcatégorielle de la ressource.

3.3.3 Correspondances sémantiques transcatégorielles

Les traitements sémantiques sur les noms, verbes, adverbes et adjectifs ont été effectués concomitamment dans un souci de garantir l'homogénéité transcatégorielle. L'utilisation combinée du *DEM* pour les noms, adjectifs et adverbes au *LVF* pour les verbes permet de traiter l'ensemble des catégories avec des ressources similaires. De plus, le fait de disposer de classes déjà établies pour les verbes permet de faciliter la classification des noms déverbaux, de même pour les adjectifs et les noms déadjectivaux.

Ces traitements communs ne doivent cependant pas occulter les différences entre catégories. La catégorie nominale, étant la plus hétérogène, a de nombreuses correspondances avec les autres catégories, principalement les verbes et les adjectifs.

Les critères d'effectifs (nombre de lemmes) et de polysémie (nombre d'acceptions) sont ainsi fonction de la catégorie envisagée, comme l'illustre le tableau 3.5.

	Adjectifs	Adverbes	Noms	Verbes
Lemmes	274	202	493	342
Acceptions	322	215	531	698
Classes	17	10	17	15
Sous-classes	39	54	80	105

Tableau 3.5: Lemmes, acceptions, classes par catégories

Nous observons que la catégorie verbale est la plus polysémique (2,04 acceptions/lemme), et ainsi, la plus complexe à traiter. Les adverbes, en majorité méta-discursifs, sont la catégorie la moins polysémique avec une moyenne de 1,06 acceptions/lemme, juste en deçà des 1,08 acceptions/lemme pour les noms.

En travaillant conjointement avec les chercheurs procédant à la classification sémantique des autres parties du discours, nous avons également tenté d'établir le maximum de correspondances de classes entre catégories. Ces correspondances se font tout d'abord au niveau de l'étiquette représentant les classes et sous-classes sémantiques, le plus souvent un nom. Ainsi, certaines classes couvrent l'ensemble des catégories étudiées :

- {relation/correspondance} : *tout autant, correspondre, comparable, convergence* ;
- {espace} : *ailleurs, se situer, proche, lieu* ;
- {temporalité} *: précédemment, succéder, actuel, période*.

Les classes {quantité} (*ampleur, abondant, au total*) et {axiologique} (*à tort, avantage, approprié*) couvrent 3 catégories : adverbe, adjectif et nom. Enfin, certaines classes sont spécifiques à une catégorie, comme la classe des adverbes {discursif} (*certes, d'ailleurs, donc,* etc.).

En disposant de classes transcatégorielles, nous avons pour perspective l'identification des reformulations : *Nous analysons les* résultats / *notre analyse des résultats*. Cette mise en correspondance entre constructions nominales et verbales a notamment été étudiée par Fabre & Bourigault (2006). qui ont comparé des verbes et leurs correspondants nominaux déverbaux en mettant en regard les constructions dans lesquelles ils s'inscrivent. La classification des différentes catégories syntaxiques du LST opérant des liens transcatégoriels, il devient

possible, en tirant parti des informations de dérivation morphologique présentes dans le *LVF*[55], d'identifier automatiquement ce type de reformulation.

Cet intérêt est confirmé par Vossen (1997, p. 7), cité dans (Fabre & Bourigault, 2006), qui précise :

> « By unifying higher-order nouns and verbs in the same ontology it will be possible to match expressions with very different syntactic structures but comparable content. »

Outre les étiquettes transcatégorielles, nous tenons également compte de familles morphologiques pour assurer l'homogénéité de notre typologie. Ainsi, comme précédemment explicité, nous nous assurons que les mots reliés morphologiquement (*dépendre/dépendance*), lorsqu'ils recouvrent une acception comparable, appartiennent à des classes sémantiques identiques. Ainsi, les éléments du LST suivants appartiennent à la sous-classe {processus_évolutif/amélioration_augmentation} : *accroître/accroissement, améliorer/amélioration, développer/développement*.

Les correspondances morphologiques les plus nombreuses ont été faites entre noms et verbes, entre noms et adjectifs, ainsi qu'entre adjectifs et adverbes, comme nous pouvons l'observer dans le tableau 3.6.

Classe/Sous-classe	**Adjectifs**	**Adverbes**	**Noms**	**Verbes**
{relation/opposition}	*différent*		*différence*	*différencier*
{processus_évolutif/amélioration_augmentation}			*accroissement*	*accroître*
{relation/association}			*combinaison*	*combiner*
{relation/correspondance}	*correspondant*		*correspondance*	*correspondre*
{réalisation}			*élaboration*	*élaborer*
{modalité/probabilité}	*éventuel*	*éventuellement*		
{complexité}	*difficile*	*difficilement*	*difficulté*	

Tableau 3.6: Correspondances sémantiques entre mots liés morphologiquement

[55] La rubrique 'DER' du *LVF* informe sur la construction des dérivés nominaux et adjectivaux des verbes. La rubrique 'N' renseigne sur le nom dont est dérivé le verbe. Par exemple, le verbe *analyser* a '1*' pour valeur pour la rubrique 'N'. Cela signifie que pour retrouver la base nominale dont est dérivé *analyser*, il suffit de supprimer le dernier caractère du verbe.

Cette préoccupation vis-à-vis de la cohérence sémantique entre les différentes catégories a également pour objectif de pouvoir traiter les variantes syntaxiques des routines, liées par exemple aux nominalisations et adjectivations. Ainsi, en tenant compte des traits sémantiques (indiqués en exposant) dans les phrases suivantes, nous pouvons imaginer que les constructions suivantes seront regroupées comme réalisation d'une même routine sémantico-rhétorique :

- *L'objectif*$_{objectif}$ *de cette étude est la classification*$_{examen}$…
- *Cet article est consacré*$_{objectif}$ *à la catégorisation*$_{examen}$…
- *Ces travaux visent*$_{objectif}$ *à classer*$_{examen}$…

Ainsi, nous conservons la logique de constitution incrémentale de la ressource en injectant les résultats d'un traitement sur le LST (ici les traits sémantiques) pour enrichir la description de ce lexique (ici par les routines sémantico-rhétoriques)

Nous détaillons dans la partie suivante les différents objectifs d'utilisation de la ressource du LST enrichi sémantiquement, inscrits dans le cadre de notre participation au projet TermITH (domaine de l'indexation en termes), dans le cadre de travaux à visée didactique ou dans le cadre de la description linguistique de notre objet d'étude.

3.4 Perspectives d'utilisation de la ressource sémantique du LST

La ressource du LST est constituée selon plusieurs objectifs.

Au niveau didactique, le fait de regrouper les éléments lexicaux renvoyant à des concepts proches permettra un accès onomasiologique du LST. De plus, la mise en évidence de routines aux fonctions rhétoriques identifiées constitue un enjeu clair de l'aide à la rédaction scientifique dans la mesure où la remise en contexte des unités lexicales améliore nettement la compréhension et l'appropriation de ce lexique spécialisé.

En termes de description linguistique, le repérage de ces patrons nous permet de mieux catégoriser le genre de l'article de recherche et de décrire finement les fonctions rhétoriques du LST convoquées à travers la phraséologie.

Enfin, en matière d'indexation automatique, les intérêts du LST et des routines sont multiples. Par la prise en compte des routines, il est alors possible d'identifier les contextes favorables ou défavorables au statut terminologique d'un candidat-terme. De plus, un nom du LST, potentiellement polysémique (et surtout possiblement terminologique), pourra être désambiguïsé dans de telles routines. Par exemple, si *sujet* apparaît dans une phrase (par exemple : *nous abordons le sujet de…*) correspondant à une routine de topicalisation, son sens renvoyant au LST sera confirmé. L'utilisation du LST en tant que filtre d'exclusion est alors améliorée, l'acception transdisciplinaire validée éliminant alors toute possibilité de statut terminologique pour l'élément en question.

Au niveau des perspectives d'utilisation à long terme du LST dans de futurs travaux, nous envisageons notamment une étude sur les déverbaux présents dans notre lexique nominal et dont la racine verbale appartient aussi au LST. Nous voudrions ainsi comparer les constructions et acceptions mobilisées entre les verbes et les noms du LST lorsqu'ils sont reliés morphologiquement et/ou sémantiquement, à la manière de travaux précédents (Condette, Marin, & Merlo, 2012; Fabre & Bourigault, 2006).

Comme le rappellent Bouillon & Viegas (2001), les applications de lexiques sémantiques en TAL sont nombreuses, et nous ne listons qu'un échantillon des possibles utilisations de notre ressource. Celle-ci a été élaborée manuellement, dans un souci de qualité et de finesse de la ressource. La construction de cette classification étant relativement coûteuse, nous souhaitons expérimenter des méthodes automatiques permettant de faciliter l'intégration potentielle de nouveaux éléments (par exemple pour des mots du LST non repérés automatiquement lors de la phase d'extraction mais finalement« rattrapés » ultérieurement).

Nous présentons donc dans le chapitre suivant deux expérimentations de classification (semi-)automatique sous-tendues par cet objectif de maintenance et de mise à jour de la ressource ainsi que par d'autres préoccupations. D'une part, nous souhaitons confronter notre typologie manuelle à des classifications intégralement automatisées. D'autre part, nous voulons tester l'intérêt de ces méthodes automatiques dans la détection de critères communs aux éléments de

nos classes afin de parfaire les tests d'appartenance ainsi que la définition en intension des classes.

Chapitre 4
Catégorisations distributionnelles automatiques

Sommaire

Nous avons détaillé dans le chapitre précédent les difficultés liées à l'élaboration d'une classification sémantique du LST. Nous expérimentons dans ce chapitre deux méthodes avec pour objectif l'évaluation de l'apport de méthodes automatiques afin d'améliorer notre classification des noms du LST. En nous appuyant sur les relations lexico-syntaxiques présentes dans le corpus, nous tentons de regrouper les noms du LST en classes d'unités lexicales partageant une certaine proximité sémantique. Cette tâche fait suite au travail de dégroupement en acceptions des noms du LST, présenté dans la section 3.3.1. Il est alors nécessaire d'attribuer autant de classes sémantiques à un élément que d'acceptions qu'il recouvre dans la ressource de référence.

Dans le but d'évaluer notre classification manuelle, obligatoirement sensible à la variation inter-individuelle, nous testons deux méthodes de classifications automatiques, pour lesquelles les résultats seront confrontés à nos classes définies manuellement. De plus, dans l'optique de la reproduction de notre méthodologie d'extraction et de classification du LST (pour un corpus de sciences exactes par exemple), l'apport d'une phase de classification semi-automatique préalable à la coûteuse phase de classification manuelle serait indéniable.

Nous testons dans un premier temps une catégorisation basée sur la théorie des prototypes en partant des classes proposées par Tutin (2007c) pour lesquelles nous définissons des prototypes qui constituent selon nous les éléments les plus représentatifs de la classe.

Nous effectuons par la suite une deuxième expérimentation de classification automatique en utilisant une méthode d'analyse de concepts formels : le treillis de Galois. Cette méthode présente l'avantage de pouvoir représenter la hiérarchie entre classes et ainsi de vérifier la validité de notre classification en classes et sous-classes.

Les classes sémantiques constituées à l'aide de ces méthodes automatiques peuvent alors être évaluées de deux manières. D'une part, dans une perspective d'évaluation extrinsèque, la pertinence de la classification automatique peut être éprouvée à l'aune de son apport pour l'identification et l'extraction de routines sémantico-rhétoriques. C'est l'optique adoptée dans les travaux d'évaluation de classification distributionnelle qui sont par ailleurs généralement orientées-tâches (Baroni & Lenci, 2011). D'autre part, l'évaluation intrinsèque des résultats de la

classification automatique est effectuée par comparaison avec les résultats de la classification sémantique manuelle du LST présentée dans le chapitre précédent, section 3.3.2. Ce dernier type d'évaluation nécessite ainsi alors d'avoir à disposition une ressource de référence.

4.1 Apports des méthodes automatiques

Nous avons ici recours aux méthodes automatiques en connaissant à priori les classes sémantiques constituant notre lexique. L'objectif est alors de les valider et d'identifier des critères supplémentaires pour les définir. Une application intéressante pour ces classes sémantiques est la détection de routines sémantico-rhétoriques. Dans ce but, il est nécessaire d'associer aux éléments du LST des étiquettes sémantiques afin de pouvoir identifier les patrons syntaxiques intégrant ces classes sémantiques. Ainsi, si nous définissons les classes suivantes :

- nom de la sous-classe {thème} : *question, problème, problématique…*
- nom de la classe {document} : *article, ouvrage, thèse, texte…*
- nom de la classe {personne} *: auteur, chercheur, expert…*
- verbe de la sous-classe {examen} : *traiter, aborder, étudier…*

Un exemple de patron permettant de présenter le thème d'un document ou de travaux pourrait être :

- ($Nom_{document}$ | $Nom_{personne}$) $Verbe_{examen}$ Dét $Nnom_{thème}$

Dans cette construction, un nom de la classe {personne} ou {document} est sujet d'un verbe de la classe {examen} prenant pour objet un nom de la classe {thème}. Ce patron pourra alors permettre d'extraire les réalisations suivantes :

- *Cependant, ces auteurs$_{personne}$ n'abordent$_{examen}$ pas la question$_{thème}$ de la structure argumentale de la base verbale […]*[1]
- *Ces articles$_{document}$ traitent$_{examen}$ bien entendu d'autres problèmes$_{thème}$ plus spécifiques aux contextes étudiés […]*[2]

---

[1] Jalenques, P. (2002). Étude sémantique du préfixe RE en français contemporain: à propos de plusieurs débats actuels en morphologie dérivationnelle. *Langue française*, (133), 74-90.

[2] Malin, E., & Martimort, D. (2001). Les limites à la discrimination par les prix. Annales d'Economie et de Statistique, 209-249.

L'identification de telles constructions a également une application directe dans le cadre du projet TermITH pour l'extraction terminologique. Ces patrons peuvent en effet être exploités pour mieux circonscrire la terminologie, dans la mesure où certaines constructions sont fréquemment introductrices de termes, ainsi que nous le verrons dans la section 5.3 sur les cooccurrences entre termes et éléments du LST.

Au-delà de leur utilisation dans la recherche de routines, les classes et sous-classes sémantiques du LST seront intégrées à la ressource lexicale. Nous envisageons plusieurs applications pour lesquelles une classification sémantique serait un apport : aide à la lecture, à la rédaction, repérage d'informations, etc. Ces traits sémantiques permettront aux utilisateurs de la ressource de percevoir les relations sémantiques liant les éléments lexicaux et ainsi faciliter l'appropriation de nouveaux mots. Ainsi, comme le note Tutin (2007b), le codage sémantique doit pouvoir répondre à plusieurs besoins, parfois contradictoires : l'aide à la rédaction (par accès onomasiologique grâce à une organisation sémantique du lexique accessible à des non-spécialistes) ou extraction et annotation de routines (nécessitant une formalisation plus fine du lexique). Cette complexité est accentuée par la nature même du LST, ensemble lexical couvrant plusieurs domaines, massivement abstrait et complexe à circonscrire dans le genre de l'écrit scientifique. D'une part, certains concepts auxquels renvoie le LST « se rapprochent de l'univers scientifique » (les hypothèses, les entités scientifiques, les théories mobilisées). D'autre part, certains éléments du LST renvoient à des concepts « qui relèvent davantage d'un cadre notionnel générique » (fait, instrument, qualité) (Pecman, 2004b, p. 305). De plus, comme nous l'avons vu lors des traitements précédents sur ce lexique, s'ajoutent à ces difficultés la polysémie du LST et sa proximité avec le lexique de la langue générale, dont il partage un sous-ensemble, le LAG.

4.2 Principes de la classification distributionnelle

Comme le rappellent Fabre & Lenci (2015), de nombreux travaux en TAL se basent sur les informations distributionnelles extraites de corpus afin de calculer des proximités sémantiques entre mots. Le principal intérêt de ces méthodes est le fait qu'il est possible de regrouper, rapprocher, des mots sémantiquement proches,

en s'appuyant uniquement sur des données issues du corpus (à travers les contextes définis par des empans ou des relations syntaxiques).

Ce principe de catégorisation sémantique se situe dans la lignée des travaux de Harris (1991) qui, comme le rappellent Habert & Zweigenbaum (2003), soutient que les relations lexico-syntaxiques conduisent à des distinctions sémantiques. Autrement dit, « le partage de contextes disposés selon une même configuration syntaxique [...] constitue un indice de proximité sémantique » (Girault, 2008, p. 263).

Dans le cadre de la théorie des sous-langages (*sublanguages*), cette approche distributionnelle est sous-tendue par l'hypothèse selon laquelle seuls des corpus spécialisés permettent la construction de catégories sémantiques claires (Habert & Zweigenbaum, 2002) cités par Fabre & Lenci (2015)). Cette observation nous conforte ainsi dans l'utilisation de telles méthodes dans le cadre de l'analyse d'un corpus spécialisé d'écrits scientifiques. Nous devons cependant souligner la difficulté à gérer la polysémie du LST avec ces méthodes.

Selon Fabre & Lenci, le modèle le plus utilisé aujourd'hui en sémantique distributionnelle est le modèle vectoriel (ou sémantique vectorielle). La proximité sémantique est alors calculée en termes de distance entre vecteurs, chaque vecteur représentant la distribution d'un mot. Dusserre (2016) expérimente ainsi le modèle *Word2vec* (Mikolov, Chen, Corrado, & Dean, 2013) en vue de classer les noms du LST. Ce modèle construit un réseau de neurones afin de représenter les éléments (unités lexicales ou autres) dans un espace vectoriel et de le regrouper ainsi par leur proximité dans cet espace. Au vu des résultats, Dusserre constate que ce modèle ne permet cependant pas de gérer la polysémie présente dans le LST et qu'il engendre certains regroupements ne reposant pas sur la co-hyponymie mais la simple appartenance à un même champ sémantique. Bien que certaines classes générées soient pertinentes, une validation manuelle reste nécessaire, d'autant plus dans la perspective d'une utilisation de ces classes dans des applications didactiques.

Nous avons choisi, pour la présente étude, d'expérimenter deux méthodes qui nous semblaient davantage adaptées à nos objectifs. Notre but n'est pas ici l'intégration d'un classifieur automatique mais l'exploitation des résultats des méthodes automatiques dans l'aide à la maintenance et/ou l'amélioration de la

ressource. Ces résultats peuvent alors être exploités de deux façons. D'un côté, certains regroupements générés automatiquement peuvent se révéler plus pertinents que la classification manuelle et peuvent ainsi nous orienter dans la création/modification de classes. D'un autre côté, les classes automatiquement générées permettent de mettre en évidence certaines propriétés définitoires pertinentes, non identifiées lors de la classification manuelle. Ces propriétés peuvent alors être associées à nos classes sémantiques du LST, en tant que test d'appartenance, pour l'intégration future de nouveaux éléments dans la ressource. Nous dressons, sur le modèle du *Lexicoscope*, la liste d'un ensemble de propriétés lexico-syntaxiques issues du corpus. Ces propriétés (relations de dépendance typées avec un mot dont le lemme et la catégorie sont renseignés) correspondent aux vecteurs dans les matrices de cooccurrences, donnant une représentation explicite de la distribution d'un mot (Fabre & Lenci, 2015, p. 10).

Nous adoptons, dans les présents travaux, une approche distributionnelle basée sur le corpus arboré en comparant les mots en fonction des relations lexico-syntaxiques dans lesquelles ils s'inscrivent. Les travaux sur la classification et/ou la recherche d'informations reposant sur l'analyse distributionnelle sont nombreux, et s'intéressent notamment aux patrons syntaxiques (Habert & Nazarenko, 1996; Morin & Martienne, 2000), aux cooccurrences avec des expressions (Grefenstette, 1994), avec des ensembles de verbes (Faure & Nedellec, 1999) ou des termes (Bourigault *et al.*, 2004).

Malgré le foisonnement des études en analyse distributionnelle, « ce champ d'investigation n'est pas épuisé » (Tanguy *et al.*, 2015, p. 103). Les travaux portent massivement aujourd'hui sur l'optimisation des paramétrages des méthodes distributionnelles, ainsi que sur l'étude de la pertinence de ces méthodes à des applications déterminées, comme le notent Fabre & Lenci (2015).

Geeraerts (2009), dans son ouvrage sur les théories actuelles de la sémantique lexicale, rappelle les trois principales caractéristiques de l'approche distributionnelle : méthodologie basée sur les usages en corpus, rôle central des collocations, aspect technique de l'approche. Notre catégorisation sémantique des mots du LST se fait ainsi sur corpus, en extrayant de celui-ci les propriétés lexico-syntaxiques, reprenant les mots de Firth (1957) : « You shall know a word by the company it keeps ». Les mots partageant des environnements lexico-syntaxiques

proches ont un lien sémantique qui peut être la synonymie, la co-hyponymie, l'antonymie, etc : l'analyse distributionnelle permet de rapprocher des mots dans de telles relations sans pour autant déterminer la nature de la relation. Manser (2012) relève qu'il est possible d'identifier des relations de proximité sémantique entre des termes, comme peuvent le faire Bourigault *et al.* (2004). Les classes sémantiques que nous avons créées manuellement regroupent des unités lexicales en relation de proximité sémantique, en l'occurrence de co-hyponymie, comme décrit dans le chapitre sur la classification manuelle.

La classification sémantique peut également s'appuyer sur des ressources externes, tels des dictionnaires. Billami, Camacho-Collados, Jacquey, & Kister (2014) exploitent ainsi les mots placés dans les définitions lexicographiques pour définir les traits sémantiques. De la même manière, certains travaux se basent sur des patrons définitoires (Malaisé, 2005), ou exprimant des relations d'hyponymie (Hearst, 1992). Par exemple, le patron — *SN1 est un type de SN2* — peut permettre d'identifier de telles relations, comme l'exemple suivant, extrait de notre corpus d'analyse, l'illustre :

- *Le slash est un type de fanfiction (récit de fans) qui subvertit les codes hétérosexuels de la culture populaire*[3]

Cette méthode, bien qu'efficace sur des corpus de dictionnaires, n'est pas adaptée aux textes scientifiques, les tournures définitoires y étant plus rares (Bendaoud, Toussaint, & Napoli, 2010).

La spécificité de notre corpus nous incite également à l'utilisation de l'analyse distributionnelle, car, comme le notent Galy & Bourigault (2005), les regroupements issus de l'analyse distributionnelle automatique sont meilleurs lorsque la classification se fait sur un corpus spécialisé plutôt que sur un corpus de langue générale. L'utilisation de la distribution à des fins classificatoires nous apparaît alors plus évidente pour un sous-langage, dans lequel la variation des constructions sera moindre que dans la langue générale. Cependant, bien que le LST soit convoqué dans un genre défini, il n'est pas pour autant assimilable à la terminologie, et de ce fait se révèle plus complexe à classifier.

---

[3] Tamagne, F. (2006). Histoire des homosexualités en Europe : un état des lieux. *Revue d'histoire moderne et contemporaine*, (4), 7-31.

Afin de s'assurer que la classification se base sur des propriétés effectivement représentatives de l'usage transdisciplinaire, nous excluons les propriétés correspondant à une seule discipline, i.e. présentes dans un seul sous-corpus disciplinaire. Nous appliquons ainsi le critère de transdisciplinarité aux dépendances syntaxiques. Le critère de répartition, appliqué pour repérer les unités lexicales du LST (voir section 2.2.2.1.2), est alors adapté au niveau lexico-syntaxique. Les observations de Biber (1993b), qui a montré que les suites de catégories morpho-syntaxiques varient avec le domaine, nous confortent dans le choix de ce critère de répartition.

Au niveau de l'efficacité de ces méthodes automatiques, des travaux ont montré que les résultats de l'analyse distributionnelle varient fortement lorsque sont modifiés les critères combinatoires, d'empan ou de nature des contextes (Fabre, Hathout, Sajous, & Tanguy, 2014). Sont également avancés comme facteurs de variation dans les résultats la catégorie des mots à classer et pour certains cas les mots eux-mêmes. À ces variations s'ajoutent les difficultés dues à la polysémie et aux expressions polylexicales qui amoindrissent la qualité des résultats, ainsi que le note Tutin (2008).

De plus, comme Grefenstette (1993) l'a montré, la qualité des regroupements dans l'analyse distributionnelle est également fonction de la fréquence. Pour les mots très fréquents, le recours aux cooccurrents syntaxiques s'avère le plus efficace. À l'inverse, pour les mots les moins fréquents, l'utilisation des cooccurrents par proximité est la plus adaptée. Nos premières évaluations ne permettent pas de confirmer ou d'infirmer cette observation mais la pauvreté d'attributs présentée par certains mots peu fréquents est assurément un obstacle à l'obtention de regroupements satisfaisants. Ces regroupements sont obtenus à la suite des trois étapes classiques de la classification distributionnelles, rappelées par Habert & Zweigenbaum (2003). En premier lieu, les attributs représentant les unités lexicales à classer doivent être définis. Ensuite, des scores de similarité sont calculés en fonction de ces attributs. Le regroupement en sous-ensembles de mots peut alors se faire à partir des scores de similarité. La catégorisation utilise également ces calculs de similarité de contexte mais en les appliquant à des classes déjà définies.

Notre expérimentation suit les mêmes étapes en procédant dans un premier temps à l'extraction des relations lexico-syntaxiques qui constitueront le profil combinatoire des noms du LST, extraction présentée section 4.4. Les comparaisons de propriétés et les regroupements sont ensuite effectués selon deux modalités. Dans l'expérimentation selon le modèle des prototypes, détaillée section 4.5, nous comparons le profil combinatoire des noms aux profils définitoires des classes, établis par intersection des profils des prototypes. Dans la seconde expérimentation, section 4.6, les regroupements sont faits par l'analyse de concepts formels. Comme nous l'avons vu dans le chapitre précédent, la structuration sémantique du LST peut être envisagée au niveau catégoriel ou transcatégoriel (i.e. les classes sémantiques sont constituées de plusieurs parties du discours). Nous expérimentons ici une catégorisation pour la catégorie nominale que nous avons organisée en classes et sous-classes sémantiques.[4]

4.3 Principes de catégorisation des noms du LST

Nos deux expérimentations, utilisant la méthode des prototypes et les treillis de Galois, partent des principes de la classification distributionnelle, détaillés dans la section 4.2. Nous nous appuyons sur les informations lexico-syntaxiques issues du corpus pour représenter et regrouper les noms du LST. Cette approche distributionnelle, comme le rappelle Geeraerts (2009, p. 166), permet de faire le lien entre sémantique et syntaxe en regroupant des ensembles lexicaux sémantiquement homogènes sur la base de leurs propriétés syntagmatiques.

La définition des critères permettant de sélectionner les relations lexico-syntaxiques influe ainsi grandement sur la qualité des regroupements. Tutin (2007c) constate d'ailleurs que la difficulté de la sémantique distributionnelle automatique est que certaines associations ne sont pas pertinentes sémantiquement alors qu'elles sont statistiquement importantes, et que la désambiguïsation des éléments analysés et/ou de leurs cooccurrents est nécessaire à la bonne exploitation de cette combinatoire. Or, nous ne pouvons procéder à la désambiguïsation sémantique intégrale de notre corpus et devons donc gérer ces difficultés.

---

[4] Nous envisageons dans de futurs travaux une expérimentation de comparaison des profils combinatoires entre substantifs déadjectivaux et leurs correspondants adjectivaux, et entre substantifs déverbaux et leurs correspondants verbaux.

À celles-ci s'ajoute la taille de notre corpus d'analyse qui, avec 5 millions de mots, peut être considéré comme modeste[5] pour les tâches d'analyse distributionnelle (Fabre *et al.*, 2014). La constitution de gros corpus en langue de spécialité est cependant complexe, l'accessibilité à de nombreux textes l'étant également, spécialement en français pour les textes scientifiques.

Une autre difficulté se situe au niveau de certaines relations lexico-syntaxiques (nous en présentons un exemple section 4.6.3 avec la préposition *selon*). Le rapprochement entre les mots se fait alors sur deux acceptions différentes d'une même relation et conduit à des regroupements peu satisfaisants. Girault (2008) soulève également ce problème de la polysémie pour, d'une part les unités lexicales à classer, et d'autre part les attributs les représentant.

Malgré ces difficultés, la classification automatique offre l'avantage d'un gain de temps précieux lorsque les résultats sont exploitables, et s'ils ne le sont pas, permet d'en identifier les raisons : définition trop stricte ou trop large des critères, taille du corpus, qualité de l'analyse syntaxique, etc. Selon Bouaud, Habert, Nazarenko, & Zweigenbaum, « construire des catégories sémantiques pour une langue de spécialité est un travail laborieux » (2000, p. 278), et le recours à des techniques de traitement automatique permet de guider les linguistes lors de la réalisation de ces catégories.

Nous voyons donc l'intérêt de développer une méthode pour la reproduction d'une telle ressource à partir d'autres corpus. De plus, outre leur utilisation dans la phase d'évaluation de notre classification manuelle, ces méthodes peuvent s'avérer utiles dans le cadre de la maintenance et de la mise à jour de notre ressource du LST.

La tâche d'acquisition automatique de classes sémantiques basée sur corpus a ainsi été de nombreuses fois explorée. La méthodologie la plus couramment employée, et que nous reprenons, implique une première phase d'extraction des cooccurrents des éléments à classer, pour ensuite comparer ces ensembles (ou profil combinatoire, que nous détaillons section 4.4) afin de rapprocher les éléments partageant un maximum de propriété.

---

[5] En comparaison des 2 milliards de mots du corpus ukWaC ou des 380 millions de mots du corpus AQUAINT 2.

La classification automatique peut se baser sur des critères de niveaux différents : les métadonnées des articles, les annotations en parties textuelles, les informations morphologiques, les dépendances, etc. Dans la continuité de nos traitements précédents, nous nous situons au niveau des informations lexico-syntaxiques, i.e les relations de dépendance avec des lemmes-catégories, que nous filtrerons et hiérarchiserons selon les critères explicités ci-après. La méthode que nous appelons par prototype inclut également des critères de distribution textuelle : grâce à l'annotation en parties textuelles, nous calculons la partie textuelle dans laquelle la fréquence relative du mot est minimale, ainsi que celle pour laquelle la fréquence est maximale. Certaines unités lexicales étant spécifiquement mobilisées[6] dans les introductions (*objectif*), notes de bas de page (*article*) ou les conclusions (*résultat*), nous souhaitons intégrer cette propriété dans la mesure où elle peut permettre le regroupement d'unités lexicales sémantiquement proche.

Les méthodes d'analyse distributionnelle automatique peuvent également différer selon la sélection des contextes pris en compte et selon les métriques utilisées pour la comparaison de contextes. Nous opérons également une sélection des relations lexico-syntaxiques mais retenons également les colligations, ou relations syntaxiques avec des mots grammaticaux. Observons pour illustrer ce point l'analyse syntaxique de la séquence « *Méthodes de la classification* ».

[6] Ces fréquences par parties textuelles peuvent également être observées à l'aide de l'interface ScienQuest.

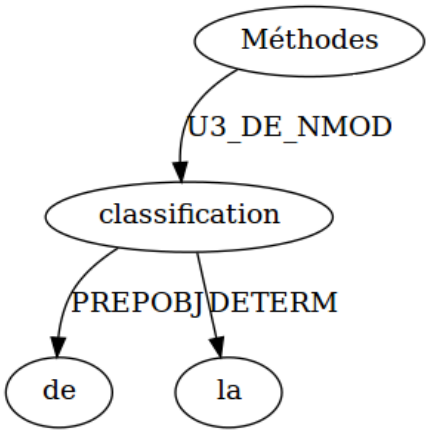


Illustration 4.1: Analyse en dépendance et propriétés lexico-syntaxiques

Nous pouvons constater, dans l'illustration ci-dessus, que le lemme-catégorie *classification-NOM* aura potentiellement comme attributs les relations lexico-syntaxiques suivantes :

1. *le_det#determ :* le nom *classification* est gouverneur dans la relation de détermination avec le dépendant *la* de la catégorie déterminant ;
2. *de_prep#prepobj :* le nom *classification* est gouverneur dans la relation prépositionnelle avec le dépendant *de* de la catégorie préposition ;
3. *méthode_nom#~u3_de_nmod* : le nom *classification* est dépendant dans la relation modifieur de nom prépositionnel avec le gouverneur *méthodes* de la catégorie nom.

4.4 Extraction des profils combinatoires

La catégorisation s'appuie donc sur la définition, pour l'ensemble des noms du LST, d'un profil combinatoire constitué des relations lexico-syntaxiques présentes dans le corpus et répondant aux critères définis ci-dessous. Les deux expérimentations présentées dans ce chapitre reprennent deux paramètres importants qui permettent de filtrer les relations.

Le premier critère est celui de transdisciplinarité : une relation lexico-syntaxique (i.e une relation de dépendance avec un lemme-catégorie défini) doit apparaître dans un minimum de trois disciplines. Nous voulions initialement poser

le seuil à 5 disciplines sur 10 mais la taille modeste du corpus, et le fait de travailler sur des informations syntaxiques nous ont convaincu de baisser le seuil de transdisciplinarité afin que les mots peu fréquents puissent avoir suffisamment de relations syntaxiques comme critères de classification. Nous conservons néanmoins ce critère de répartition en nous basant sur l'hypothèse selon laquelle l'acception transdisciplinaire d'un mot sera la mieux représentée par une distribution transdisciplinaire.

Le deuxième critère est celui de la significativité statistique des relations lexico-syntaxiques. Nous utilisons le calcul de similarité du rapport de vraisemblance afin d'ordonner les relations. Ce calcul nous permet de ne pas sur-représenter les relations impliquant un cooccurrent très fréquent (et souvent peu informatif, tels les auxiliaires *être* et *avoir*). Cette pondération des fréquences permet également de faire remonter les associations avec des cooccurrents à fréquence absolue faible mais statistiquement significatives. Nous retenons les 10 premières relations lexico-syntaxiques selon la mesure du rapport de vraisemblance[7] pour les relations suivantes :

- dépendant dans la relation sujet avec un verbe
- dépendant dans la relation objet avec un verbe
- dépendant dans la relation modifieur de nom avec un nom
- gouverneur dans la relation modifieur de nom avec un nom
- gouverneur dans une relation avec un adjectif

Nous retenons, parmi les colligations, les 4 premières relations impliquant un déterminant ainsi que les 4 premières relations impliquant une préposition.

Enfin, le cooccurrent doit être absent de la liste d'exclusion, ou stop-liste, que nous avons définie pour limiter le bruit dû aux erreurs de l'analyseur et à la présence de mots étrangers.

L'ensemble de ces contraintes a été défini suite à plusieurs expérimentations sur le choix des attributs représentant le profil combinatoire des noms du LST.

[7] Nous avions initialement appliqué le seuil de 10,83 pour le rapport de vraisemblance. Cependant, certains mots avaient alors trop peu de cooccurrents répondant positivement à ce seuil. Nous avons alors décidé de retenir les 10 premières relations dans l'ordre décroissant du score pour le rapport de vraisemblance.

Nous avons également tenu compte des observations de Tanguy *et al.* (2015) dans leur analyse de l'impact des relations syntaxiques prises en compte, ainsi que des normalisations opérées pour l'analyse distributionnelle automatique. Nous résumons ci-dessous les principaux constats émanant de cette étude :

- la relation modifieur de nom est essentielle pour le calcul de voisinage des adjectifs, et a un impact non négligeable dans le traitement des noms. Nous l'avons ainsi intégrée dans nos profils combinatoires, à la fois pour la phase de classification manuelle et pour la présente expérimentation de classification automatique ;
- l'apport de la relation sujet n'est pas significatif pour les verbes et les noms. Malgré ce constat, nous avons choisi de conserver cette relation étant donné son intérêt dans la définition des tests d'appartenance de classe. (voir tableau 2 de la classification en annexe) ;
- la relation objet est essentielle pour les verbes ;
- la prise en compte de la relation de modification adverbiale améliore les résultats pour les adjectifs.

Ces points ne concernent pas tous directement la classification des noms mais sont d'un intérêt certain dans la perspective de la mise en place d'une aide à la classification d'éléments des catégories mentionnées. Nous envisageons, pour de futurs travaux sur les adjectifs et les verbes, d'intégrer ces remarques sur les différents paramètres optimaux afin d'en évaluer l'impact sur les deux méthodes présentées dans le présent chapitre.

Nous choisissons donc de caractériser chaque mot à classer par un ensemble de relations lexico-syntaxiques. Nous déterminons à priori celles que nous incluons, suivant en cela l'observation de Tanguy *et al.* (2015) qui notent que les modèles basés sur des configurations syntaxiques brutes sont moins efficaces que ceux qui opèrent une sélection des relations de dépendance à prendre en compte.

Chaque relation retenue (après la phase de filtrage opérée selon les critères définis ci-dessus) devient dès lors un attribut, un trait de l'élément à classer. Autrement dit, nous procédons à l'identification des relations lexico-syntaxiques transdisciplinaires et statistiquement significatives pour associer à chaque mot un

ensemble de triplets[8] : relation typée et orientée (selon que le mot est gouverneur ou dépendant), lemme du cooccurrent, catégorie du cooccurrent.

Ainsi, si nous prenons pour exemple le mot *travaux* dans la phrase suivante :

- *Nous présentons dans cette partie nos travaux sur la catégorisation du LST.*

Cette phrase comprend une relation *objet* entre le gouverneur *présentons* et le dépendant *travaux*. Cette cooccurrence est utilisée dans la définition du profil combinatoire du mot *travaux* en tant que propriété lexico-syntaxique. Cette propriété, notée dans notre formalisme *présenter_verb#~obj*, est constituée de :

- la relation typée et orientée : type donné par l'étiquette (*Objet, Sujet, Attribut*) – un préfixe (le caractère ~) précède le type si le mot est dépendant dans la relation, dans le cas contraire, il est le gouverneur ;
- le cooccurrent : représenté par son lemme et sa catégorie morphosyntaxique.

Le résultat de ces extractions n'est donc pas la représentation d'occurrences phrastiques, puisque nous perdons la concomitance des relations. Nous considérons donc ces relations comme indépendantes les unes des autres dans le but de favoriser le maximum de regroupements. En effet, la taille relativement modeste de notre corpus, le fait de comparer des dépendances plutôt que des contextes par proximité, ainsi que les contraintes de transdisciplinarité des relations, nous incitent à maximiser les possibilités de recoupement de traits.

Avant d'extraire les attributs, issus des calculs de relations de dépendance, nous avons effectué un post-traitement des sorties de *XIP*[9] (distribution de la coordination, traitement des pronoms relatifs, etc.[10]) afin de regrouper par exemples les relations *subj* et *deepsubj* (sujet profond).

La définition de ces attributs constituant le profil combinatoire des noms du LST à classer est alors utilisée dans deux expérimentations de catégorisation sémantique, à commencer par celle de la méthode des prototypes.

---

[8] Les dépendances sont couramment représentées par ces triplets, voir par exemple (Tanguy, Sajous, & Hathout, 2015, p. 109).

[9] La section 2.2.1.3.1 présente le fonctionnement de cet analyseur.

[10] La section 2.2.1.3.2 détaille cette étape de post-traitement des sorties de l'analyseur.

4.5 Méthode des prototypes

Nous avons pour objectif de catégoriser les unités lexicales nominales du LST en fonction des traits qui leur sont associés. Nous avons vu, dans le chapitre sur la classification manuelle, que la constitution de classes sémantiques est un processus pour lequel la définition de critères et de frontières nettes entre catégories se révèle complexe. La méthode des prototypes nous permet d'introduire la notion de degré d'appartenance à une catégorie, et ainsi de contourner le problème de classes disjointes, dont l'appartenance se juge de façon binaire.

Cette méthode s'inscrit dans la conception de catégorisation par partage de traits, formulée par Rosch (1973) à travers la prototypie, qui fait suite aux Conditions Nécessaires et Suffisantes (CNS) d'Aristote. La définition de critères combinatoires (pour l'ensemble des expérimentations) peut être vue comme la définition de traits catégoriels, similaires aux CNS. Cependant, le caractère booléen des CNS ne fonctionne pas lorsque nous devons catégoriser un ensemble hétérogène de mots, dans lequel la polysémie ajoute encore de la complexité. Nous nous situons plus dans une logique de degré d'appartenance, de proximité par rapport à un prototype de la classe. Cette approche inductive nous permet alors de définir les classes en intension, en opérant une catégorisation par ressemblance de traits. Cette classification supervisée facilite l'ajout de nouveaux éléments aux classes définies grâce aux définitions de celles-ci en intension.

Comme l'a montré Rosch (1973), l'esprit humain n'assigne pas de frontières nettes entre les classes de concepts. Erk (2010) propose de transposer cette notion de degré d'appartenance au sens des unités lexicales. Elle ajoute qu'une occurrence peut correspondre à plusieurs sens simultanément, à des degrés différents. Le fait de pouvoir associer une unité lexicale à plusieurs catégories nous permet de représenter cette polysémie. Les résultats de catégorisations s'inspirant de ce modèle doivent, selon Erk (2010), être évalués de manière extrinsèque, dans le cadre d'une application définie, telle que la détection des routines les classes sémantiques. Les membres d'une classe n'ont donc pas un statut équivalent, certains représentant celle-ci plus centralement que d'autres. Les prototypes de la classe sont les éléments qui partagent le maximum de traits avec l'intension de la

classe. D'autres entités peuvent appartenir à la classe sans répondre à l'ensemble des critères définitoires.

Pour opérer notre catégorisation, nous partons dans un premier temps de la classification effectuée par Tutin (2007c). Elle distingue, parmi les noms du LST : les processus, les objets construits, les observables, les supports, les qualités, les acteurs, les relations. Nous testons ce protocole en sélectionnant 150 noms du LST étendu[11] parmi les plus fréquents.

Pour définir quels éléments seront utilisés en tant que prototype, nous nous basons sur leur haute fréquence et le fait qu'ils nous paraissent centraux à la classe qu'ils représentent. La tâche est effectuée sur quatre classes ayant trois prototypes chacune : la classe artefact (*critère, hypothèse, modèle*), processus (*expérience, test, comparaison*), support de communication (*chapitre, ouvrage, figure*) et relations logiques (*influence, correspondance, conséquence*). Nous calculons pour chacun de ces prototypes les propriétés suivantes :

- La partie textuelle dans laquelle la fréquence est la plus élevée. Les parties textuelles étant associées à des fonctions particulières (introduction du sujet, développement, conclusion), elles sont susceptibles de mobiliser un lexique particulier à ces fonctions.
- La partie textuelle dans laquelle la fréquence est la moins élevée.
- Le type de détermination la plus significative statistiquement : déterminant défini, indéfini, possessif ou démonstratif. Nous avons par exemple constaté que certaines classes ou sous-classes sont plus à même d'être déterminées par un possessif ({processus_cognitif/examen} avec *étude, analyse*, {objet_scientifique_méthode} avec *approche, démarche, méthode*) ou par un démonstratif ({état_qualité/caractéristique} avec *genre, aspect, propriété*).
- La présence ou absence d'introducteurs prépositionnels localisateurs (*dans, en*). Les classes {espace} (*zone, cadre, direction*) et {communication_support} (*article, annexe, ouvrage*) sont par exemple caractérisées par cette propriété.

[11] Dans le chapitre précédent, nous avons fait le choix d'intégrer au LST certains éléments du lexique abstrait général (LAG).

- La présence ou absence de relation avec des cardinaux et/ou numéraux. Cette propriété est typique des éléments des sous-classes {communication_support/graphique} (*figure, image, tableau*) et {communication_support/section} (*conclusion, chapitre, section*).
- Les relations lexico-syntaxiques les plus significatives (selon les critères définis dans la section 4.4) : relations sujet et objet, relation avec un autre nom, avec un adjectif.

L'ensemble de ces propriétés a alors pour but de dresser un profil des contextes d'emploi des noms du LST, en intégrant les relations qu'ils entretiennent avec les catégories pleines : nom, verbe et adjectif. La mesure utilisée pour définir les triplets syntaxiques (noms du LST, relation typée, lemme-catégorie du cooccurrent) les plus significatifs est celle du rapport de vraisemblance[12].

Ainsi, si nous prenons pour exemple la relation de dépendance modifieur adjectival (dans la séquence '*catégorisation automatique*) dont le gouverneur est le nom *catégorisation* et dont le dépendant est l'adjectif *automatique*, notée *automatique_adj#u3_adjmod, n*ous calculons pour le nom *catégorisation* (L1) le rapport de vraisemblance pour cette relation avec le cooccurrent adjectival *automatique* (L2) de la façon suivante.

Soit L1 le mot dont nous calculons le profil combinatoire, R1 la relation typée (par exemple *adjmod* pour modifieur adjectival), L2 le cooccurrent :

- a est le nombre d'occurrences de la relation *adjmod* entre L1 et L2 ;
- b est le nombre d'occurrences de la relation *adjmod* impliquant L1 et non L2 ;
- c est le nombre d'occurrences de la relation *adjmod* impliquant L2 et non L1 ;
- d est le nombre d'occurrences de la relation *adjmod* n'impliquant ni L1 ni L2.

Le rapport de vraisemblance (LLR) se calcule selon les formules suivantes :

[12] Formule dont le calcul est détaillé à l'adresse suivante : http://ucrel.lancs.ac.uk/llwizard.html [consulté le 19/09/2016]

$$E_1 = c * (\frac{(a+b)}{(c+d)})$$

$$E_2 = d * (\frac{(a+b)}{(c+d)})$$

$$LLR = 2 * ((a * \ln(\frac{a}{E_1})) + (b * \ln(\frac{b}{E_2})))$$

Nous créons ainsi un profil par classe en retenant l'intersection des ensembles de propriétés de ses prototypes. Nous obtenons par exemple pour la classe {communication_support} (*chapitre, ouvrage, figure*) les traits suivants :

- sujet d'un verbe : *paraître, consacrer, proposer, apparaître, décrire* ;
- modifié par un adjectif : *consacré, nouveau, dernier, majeur, premier, classique* ;
- objet d'un verbe : *analyser, évoquer, intituler, retrouver, consacrer* ;
- gouverneur dans la relation modifieur de nom : *référence* ;
- dépendant dans la relation modifieur du nom : *choix* ;
- partie textuelle à la fréquence minimale=notes ;
- partie textuelle à la fréquence maximale=introduction ;
- relation avec numéraux/cardinaux=1[13] ;
- détermination par un démonstratif=1[13] ;
- présence de localisateur=1[13] ;
- détermination par un possessif=0[13].

Nous notons cependant une grande variabilité dans le nombre de traits définissant le profil des classes : 46 pour *artefact,* 44 pour *processus,* 24 pour *support* et 30 pour *relation*. Cette différence s'explique par la fréquence des prototypes constituant une classe. Les prototypes les moins fréquents sont ceux ayant le moins de traits répondant aux critères de fréquence et de répartition. Ce manque de traits se répercute dans le profil combinatoire de la classe qu'ils représentent.

[13] Ces traits sont donc binaires : 0 si le trait n'est pas présent dans le corpus, 1 dans le cas inverse.

Nous générons un script *Perl* permettant de calculer un profil pour chaque élément à classer. La comparaison des profils des éléments du LST à ceux des classes nous donne alors un rang d'appartenance par classe pour chaque élément. Ce rang nous permet d'approcher le concept de « degré d'appartenance » catégorielle, central dans la théorie des prototypes. Les mots en tête de liste sont alors logiquement les plus représentatifs de la catégorie puisque partageant un maximum de traits avec les prototypes.

Nous présentons dans le tableau ci-dessous les 15 éléments partageant le plus de traits avec les prototypes des quatre classes étudiées. Les unités lexicales en gras sont celles pour lesquelles la classification paraît cohérente, après comparaison avec les classes manuellement définies dans le chapitre précédent.

	Support	Processus	Artefact	Relation
1	**Article**	**Étude**	**Théorie**	Aspect
2	**Section**	**Entretien**	**Idée**	**Effet**
3	Série	**Enquête**	Série	**Différence**
4	**Thèse**	Figure	**Principe**	Tendance
5	**Littérature**	Tableau	**Concept**	Distribution
6	**Texte**	**Démarche**	**Notion**	**Écart**
7	**Tableau**	Section	**Définition**	Transformation
8	Difficulté	Série	Caractéristique	Évolution
9	Variable	**Description**	**Test**	Portée
10	Référence	**Examen**	Démarche	Critère
11	Expression	Difficulté	**Exemple**	Définition
12	**Publication**	Situation	**Méthode**	Comparaison
13	Opposition	**Conclusion**	**Catégorie**	**Corrélation**
14	Cadre	Auteur	**Variable**	Diffusion
15	Formule	Statistique	Cadre	Importance

Tableau 4.1: Résultats de la classification par prototypes

Nous constatons une grande variabilité dans la qualité des résultats selon la classe envisagée. Les classes {relation} et {communication_support} ont parmi leurs 10 premiers membres une majorité de candidats non pertinents. Les classes {artefact} et {processus} sont plus homogènes, probablement du fait de leur granularité plus large et de leurs prototypes plus fréquents qui permettent la

définition de profils riches. Ces variations de la qualité de la classification selon la classe peuvent être également liées au caractère plus ou moins polysémique des membres de la classe envisagée.

Les catégorisations ainsi générées ne peuvent être entièrement validées, dans la mesure où des mots de rang 1 doivent parfois être rejetés alors qu'un mot de rang 14 peut lui être validé quant à son appartenance à la classe.

À la suite de l'observation des résultats, et au vu des critères définis, nous pensons qu'il est nécessaire de définir des sous-classes (par exemple, la sous-classe {communication_support/graphique}) pour tenir compte des différences de traits à l'intérieur d'une même classe. Ainsi, *article* et *tableau* sont deux éléments de la classe {communication_support} et tendent à fortement apparaître comme sujet de *montrer, présenter* ou objet du verbe *voir* mais seul *tableau*, comme les autres éléments de la sous-classe {communication_support/graphique}, apparaît avec des numéraux.

De plus, le problème de la polysémie des éléments à classer ainsi que de leurs cooccurrents influe sur les résultats. Les mots à fréquence basse rendent l'analyse par cooccurrents syntaxiques ineffective. Certains mots peu fréquents ont ainsi jusqu'à 5 fois moins de critères dans leur profil combinatoire que les mots fréquents, car peu de relations lexico-syntaxiques remplissent la contrainte de transdisciplinarité. Il faut également rappeler que cette méthodologie impose de définir à priori les classes et sous-classes et d'en déterminer les prototypes. Cette étape d'intervention d'experts lors de la classification supervisée est également nécessaire lors de la phase d'évaluation des résultats.

Cependant, cette expérimentation nous permet d'identifier certaines relations lexico-syntaxiques pouvant être intégrées dans les tests d'appartenance pour la classification manuelle. Ainsi, l'intersection des traits des prototypes de la classe {artefact} est constituée des relations suivantes :

- sujet d'un verbe : *appliquer, avérer, consister* ;
- modifié par un adjectif : *théorique, traditionnel, analytique, formel, explicatif* ;
- objet d'un verbe : *introduire, appliquer, retenir, rejeter, reposer*.

Nous pouvons alors formaliser, pour la classe {artefact}, les tests lexico-syntaxiques suivants :

- *Le N théorique consister* ;
- *rejeter le N traditionnel.*

Bien qu'elle ne permette pas de valider automatiquement l'appartenance d'un élément à une classe sémantique, la méthode des prototypes permet d'identifier les traits définitoires des classes par le calcul de l'intersection des traits des prototypes. Dans le cadre de notre classification sémantique des noms du LST, se pose alors le problème de la définition pour l'ensemble des classes et sous-classes de leurs prototypes puis de la validation du profil combinatoire de la classe. La notion de degré d'appartenance à une catégorie présente l'avantage de pouvoir représenter la polysémie des unités lexicales à classer en leur donnant la possibilité d'appartenir à plusieurs classes.

Nous présentons dans la partie suivante une expérimentation complémentaire, basée en partie sur les mêmes traits, la classification par analyse formelle de concepts à l'aide des treillis de Galois.

4.6 Analyse formelle de concepts

À la suite de l'expérimentation de catégorisation supervisée par la méthode des prototypes, nous avons expérimenté une autre méthode de classification non supervisée, dans le but de confronter notre classification manuelle à un traitement automatique, et également dans l'objectif d'identifier les attributs définitoires de nos classes.

Dans le cadre du projet TermiTH, nous avons collaboré avec Yannick Toussaint au sein de l'équipe Orpailleur[14] du Loria afin d'expérimenter d'autres méthodes en vue de la classification sémantique des mots du LST. L'objectif est de tester la faisabilité d'une catégorisation (semi-) automatique des mots simples du LST. Nous faisons une première expérience avec 40 noms du LST, dont 20 sont

[14] Nous remercions chaleureusement Yannick Toussaint pour son accueil, ses conseils et relectures. Site de l'équipe : http://orpailleur.loria.fr/index.php/Main_Page [consulté le 14/09/2016]

parmi les plus fréquents de ce lexique tandis que les 20 autres éléments font partie des moins fréquents.

Nous utilisons pour cela l'analyse formelle de concepts (FCA pour *Formal Concept Analysis*) et les treillis de Galois. La FCA est « une méthode de classification permettant de construire des concepts et des hiérarchies de concepts à partir d'un ensemble d'individus décrits par un ensemble d'attributs » (Toussaint, 2011, p. 37).

La FCA est donc construite à partir de deux ensembles — un ensemble d'objets et un ensemble d'attributs — et d'une relation d'incidence qui associe des attributs aux objets. La FCA construit des concepts formels qui sont structurés par un ordre partiel. Un concept formel est un couple (A,B) où A est un ensemble d'objets et B un ensemble d'attributs construit de telle façon que A est l'ensemble d'objets maximal possédant les attributs de B et B est l'ensemble d'attributs maximal possédé par les objets de A. L'ordre partiel est défini selon l'inclusion des ensembles d'objets.[15]

Les individus sont ici les noms du LST et les attributs les relations lexico-syntaxiques extraites du corpus et représentant leur profil combinatoire. Chaque concept est une paire composée d'une extension représentant un sous-ensemble des objets (des noms du LST) et d'une intension représentant les propriétés (relations lexico-syntaxiques) communes à ces objets. La FCA nous permet de comparer les regroupements opérés par les linguistes (sur la base des profils combinatoires, des définitions lexicographiques, de leur intuition) avec une classification se basant uniquement sur les relations lexico-syntaxiques extraites du corpus.

La définition des attributs des individus influe évidemment sur les résultats de la classification. Nous reprenons la méthode d'extraction des attributs définie en 4.4. Ces attributs doivent ainsi apparaître dans un minimum de trois disciplines, faire partie des 10 relations les plus significatives pour chaque dépendance considérée (sujet, objet, modifieur, etc.). Nous limitons le nombre d'attributs retenus pour éviter que la FCA ne produise un trop grand nombre de concepts peu pertinents (Godin, Mineau, Missaoui, & Mili, 1995, p. 106). Le critère

[15] Nous remercions également Yannick Toussaint pour cette explication de la base du calcul des concepts.

de transdisciplinarité, qui nous semble essentiel pour dresser un profil combinatoire correspondant au maximum aux acceptions transdisciplinaires, permet par ailleurs d'éliminer les relations spécifiques à un sous-corpus disciplinaire.

Comme le remarquent Bendaoud *et al.* (2010), les outils d'aide à la classification sont souvent guidés par les analystes. L'utilisation de ces outils d'extraction de connaissances suit généralement une phase de prétraitement des données et est suivie par une phase d'analyse et de validation des résultats. Dans notre cas, le prétraitement des ressources se fait par l'extraction du LST et la définition des critères permettant l'extraction des relations lexico-syntaxiques pour chaque nom du LST. Les méthodes de la FCA sont employées dans plusieurs travaux s'intéressant à la dimension sémantique d'unités lexicales. Ainsi, Cimiano, Staab, & Tane (2003) utilisent les cooccurrences verbales pour la classification de termes. Girault (2008) s'intéresse dans ses travaux aux contextes syntaxiques des entités nommées afin de permettre leur désambiguïsation automatique. Enfin, l'utilisabilité des treillis de Galois a également été éprouvée pour l'extraction de relations entre termes à partir de corpus (Toussaint, Simon, & Cherfi, 2000).

Godin *et al.* (1995) remarquent qu'un avantage de la structuration de données en graphes se situe au plan cognitif : il est plus facile de reconnaître quelque chose d'intéressant que de le décrire. Si nous transposons cette observation dans le cadre de la FCA, nous voyons en effet que certains regroupements nous semblent cohérents sans pour autant pouvoir décrire les critères permettant ces regroupements. L'intension des classes de mots générées, ou concepts, nous permet alors d'accéder à ces critères et ainsi de décrire nos classes sémantiques. L'utilisation de cette méthode se justifie donc par l'intérêt de la découverte de connaissances à partir de données, d'une manière non-supervisée.

De la même façon, Priss & Old (2004), en utilisant l'analyse formelle de concepts pour modéliser une base de données lexicale, ont montré les avantages de cette méthode pour la représentation et la visualisation des relations lexicales. D'ailleurs, un intérêt de la FCA est notamment qu'elle permet d'aboutir à des définitions précises de classes, permettant ensuite la classification objective de nouveaux objets (Bendaoud, Hacene, Toussaint, Delecroix, & Napoli, 2007). Dans

le cadre de la maintenance de notre ressource et de sa classification sémantique, cette méthode nous apparaît ainsi adaptée.

Ces différentes méthodes d'analyse de concepts formels impliquent une phase essentielle de validation des concepts par des experts. Bendaoud *et al.* (2010) effectuent une évaluation comparative entre la hiérarchie construite automatiquement et celle opérée manuellement et constatent que si certaines classes générées de façon automatique et non repérées manuellement peuvent être intéressantes, la classification automatique présente comme principale inconvénient le fait de générer un grand nombre de classes, dont la validation manuelle est nécessaire.

Cette phase de validation des éléments extraits nous permet d'identifier les critères pertinents pour la définition des classes sémantiques (par exemple, un élément de la classe {communication_support} entre en cooccurrence avec le verbe *présenter* dans la relation sujet).

4.6.1 Principe des treillis de Galois

Les treillis de Galois sont des treillis de concepts, dont les premières applications ont été faites en intelligence artificielle pour la représentation et l'acquisition de connaissances.

Comme l'expliquent Godin *et al.* (1995, p. 106), « chaque élément du treillis peut être considéré comme un concept formel et le graphe comme une relation de généralisation/spécialisation entre les concepts ». Ainsi, les relations de hiérarchies entre nos concepts de classes et sous-classes sémantiques peuvent être évaluées par cette méthode. Girault (2008, p. 261) précise que ces concepts « sont considérés comme des unités de sens qui sont organisées au sein d'une structure hiérarchique de treillis de Galois ». Szathmary & Napoli (2004) ajoutent que les treillis de Galois peuvent servir de fondement à l'élaboration d'une ontologie, ce qui nous serait utile en vue de la représentation et de l'organisation de notre ressource du LST.

Girault (2008) conçoit chaque concept de treillis comme étiquette sémantique potentiellement utile pour désambiguïser une unité lexicale. Notre évaluation considère chaque concept comme classe sémantique potentiellement

cohérente pour chaque objet de l'extension dudit concept. En termes d'évaluation extrinsèque des concepts résultant de la FCA, Girault (2008) propose de mesurer leur apport pour une tâche de désambiguïsation. De notre côté, nous opérons en premier lieu une évaluation intrinsèque par comparaison des concepts formés avec les classes sémantiques définies manuellement. Nous adoptons également un point de vue extrinsèque en évaluant l'apport de la FCA dans l'identification d'attributs définitoires pour les classes de notre classification. Une réelle évaluation extrinsèque pourrait se faire au niveau de l'opérabilité des concepts générés dans l'extraction des routines sémantico-rhétoriques que nous présentons dans la section 5.4.

Nous avons vu qu'un concept est composé de la définition de ses objets et de ses attributs. Pour notre expérimentation, un objet équivaut à un couple lemme-catégorie faisant partie de l'ensemble lexical que nous souhaitons classifier. Nous combinons, pour définir les attributs, les associations lexico-syntaxiques les plus significatives statistiquement et les traits sémantiques issus de notre classification manuelle[16]. La classe et la sous-classe d'un mot sont considérées comme deux attributs, ce qui nous permet de tester la validité de la hiérarchie classificatoire opérée manuellement. Le fait d'intégrer nos traits sémantiques permet de faciliter la visualisation du treillis de Galois en nous permettant de confronter notre classification manuelle aux résultats automatiquement générés. De plus, comme nous le détaillerons ci-après, ceci nous permet d'avoir des amorces pour naviguer dans le treillis et ainsi d'observer des éléments particuliers.

4.6.2 Définition des attributs

Les attributs que nous définissons correspondent donc à des relations lexico-syntaxiques, i.e à des triplets de la forme suivante : *lemme-cooccurrent _ catégorie-cooccurrent # relation*. Le treillis est généré d'après une matrice qui alloue à chaque objet la valeur 1 ou 0 pour chaque attribut selon que la relation est présente ou non dans le corpus.

Ainsi, si nous avons deux objets définis chacun par leurs attributs :

[16] Il est à noter qu'un mot ambigu dans notre lexique (i.e ayant plus d'une acception transdisciplinaire observée en corpus) sera alors défini par l'ensemble des classes et sous-classes auxquelles il appartient. Par exemple, *figure* et *symbole* ont chacun une acception correspondant à la sous-classe {objet_scientifique/représentation}. Ils ont également une acception correspondant à la sous-classe {communication_support/graphique}.

- l'objet *analyse_nom* a pour attributs : *contrastif_adj#adjmod, présenter_verb#~obj, démontrer_verb#~subj* ;
- l'objet *étude_nom* a pour attributs : *empirique_adj#adjmod, démontrer_verb#~subj*.

Le contexte formel sera représenté par la matrice suivante.

	contrastif_adj #adjmod	*présenter_verb #~obj*	*empirique_adj #adjmod*	*démontrer_verb #~subj*
analyse_nom	1	1	0	1
étude_nom	0	0	1	1

Tableau 4.2: Exemple de matrice dans la FCA

Le treillis généré comportera un concept dont l'extension E est : *analyse_nom ; étude_nom,* et dont l'intension I sera l'ensemble des attributs commun aux objets définissant l'extension, soit I = *démontrer_verb#~subj.*

La sélection des attributs se fait comme présenté dans la section 4.4. La section A.VIII en annexe détaille le profil combinatoire pour le nom *article* dans le cadre de cette expérimentation. Les types de dépendances retenues pour les noms sont les suivantes :

- dépendant d'un verbe dans une relation normalisée[17] *sujet* ;
- dépendant d'un verbe dans une relation normalisée *objet* ;
- gouverneur d'un nom dans une relation normalisée de modifieur de nom *;*
- dépendant d'un nom dans une relation normalisée de modifieur de nom ;
- gouverneur d'un adjectif dans une relation adjectivale ;
- gouverneur d'une préposition dans une relation prépositionnelle ;
- gouverneur d'un déterminant dans une relation de détermination ;

Les informations de fréquence du cooccurrent et de la relation nous permettent de calculer les relations lexico-syntaxiques les plus significatives selon le calcul du rapport de vraisemblance, pour s'assurer de ne pas sur-représenter les

[17] Par relation normalisée, nous entendons relations regroupées après post-traitement des sorties de l'analyse syntaxique, tel que détaillé dans la section 2.2.1.3.2.

mots les plus fréquents du corpus au détriment de ceux ayant une fréquence moindre. Nous utilisons pour cette expérimentation également le critère de répartition pour dresser un profil combinatoire correspondant au maximum aux acceptions transversales, transdisciplinaires. Nous posons le seuil à 3 pour la répartition dans les sous-corpus disciplinaires et espérons ainsi éviter l'intégration de phénomènes locaux qui seraient moins susceptibles de représenter le comportement transdisciplinaire des éléments du LST.

Nous avons ainsi repéré une relation lexico-syntaxique significativement fréquente mais trop peu dispersée pour être intégrée au profil combinatoire. Ainsi le nom *milieu* rentre 21 fois en cooccurrence avec l'adjectif *carcéral* dans la relation de modifieur adjectival. *milieu* est un nom polysémique du LST renvoyant à deux acceptions (des sous-classes {personne/collectitivité} et {espace/domaine}). Le fait de prendre en compte comme attribut une cooccurrence strictement disciplinaire[18] minimiserait les probabilités qu'un autre élément du LST partage ce cooccurrent dans les autres disciplines. Le filtrage des propriétés à l'aide du critère de répartition nous permet d'éviter la présence de ces attributs peu pertinents. La définition de l'ensemble des attributs pour chaque objet constitue alors le contexte formel, dont un exemple est présenté en annexe, section A.IX.

4.6.3 Visualisation des treillis

Via des logiciels dédiés[19], les treillis peuvent être visualisés graphiquement ainsi que l'illustre la copie d'écran ci-dessous.

[18] Après vérification en corpus, il s'avère que la totalité des 21 occurrences se trouvent dans le même article (du sous-corpus disciplinaire de psychologie) : Moulin, V., & Sevin, A. S. (2012). Souffrance au travail en milieu carcéral : les épreuves de l'exercice professionnel au parloir pénitentiaire. *Le travail humain*, *75*(2), 147-178.

[19] Nous avons pour notre part utilisé les logiciels *Erca* et *Gallica*.

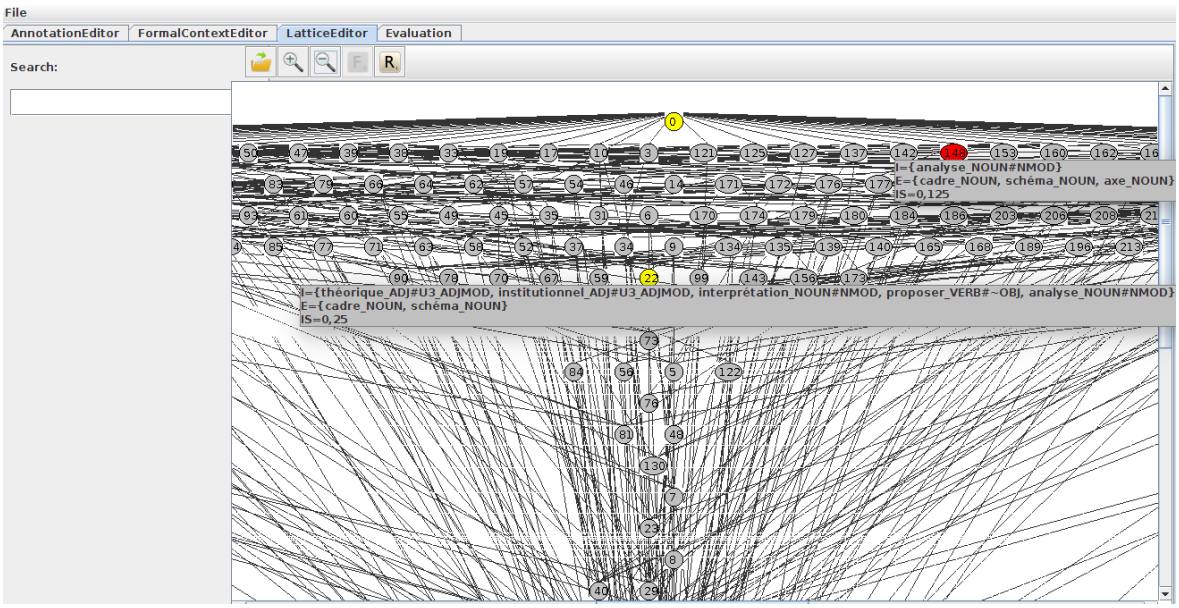


Illustration 4.2 : Visualisation d'un treillis de Gallois

Le grand nombre de concepts (représentés par les bulles numérotées) est ici immédiatement observable, de même que la profusion des liens entre concepts (représentés ici par les traits). Cette visualisation permet cependant de passer aisément d'un concept à ses concepts pères et fils, et d'accéder rapidement à son intension (I) et extension (E), via une info-bulle. L'illustration ci-dessus permet ainsi d'identifier le concept 22 dont l'extension se compose des objets *cadre* et *schéma* et dont l'intension comporte notamment les attributs suivants : *théorique_adj#u3_adjmod, interprétation_nom#nmod, proposer_verb#~obj*.

Nous générons également, à partir du fichier résultat de la FAC, des sorties au format HTML ce qui nous permet de faire des recherches de concepts par expressions régulières.[20]

L'analyse des treillis se fait verticalement :

- Le *top* est l'extension maximale (contenant tous les objets) et l'intension minimale (ensemble vide d'attributs).
- Le *bottom* est l'extension minimale (ensemble vide d'objets) et l'intension maximale (contenant tous les attributs).

[20] Ceci nous permet notamment d'identifier automatiquement les concepts « cohérents » pour lesquels l'extension est supérieure à 2 et dont l'intension intègre l'attribut de classe ou de sous-classe sémantique.

Un problème apparaît lorsque l'on observe les concepts regroupant le plus d'objets : la majorité des attributs partagés sont des colligations, des relations avec des mots grammaticaux. Ces mots sont les plus fréquents, les moins restrictifs en matière de combinatoire. L'exclusion de ces relations avec des mots grammaticaux devrait ainsi être testée pour évaluer leur intérêt, leur prise en compte générant une énorme quantité de concepts rendant complexe l'interprétation du treillis.

Afin de diminuer le nombre de concepts à évaluer, nous observons prioritairement ceux dont l'extension est au minimum de deux et dont l'intension contient au moins deux attributs, deux relations lexico-syntaxiques pleines. Nos premières observations de treillis nous ont permis d'identifier certains attributs définitoires de classe, ainsi que certaines limites.

Notre premier constat concerne la polysémie et notamment les attributs représentant les sous-classes sémantiques {objet_scientifique/représentation} et {communication_support/graphique} qui apparaissent en même temps et sont donc équivalents. Ce résultat s'explique par le lien sémantique fort existant entre ces deux sous-classes. Les étiquettes d'acceptions[21] correspondant à la sous-classe {communication_support/graphique} sont : *figure_1, illustration_1, image_1, motif_1, schéma_1, tableau*. Certains de ces mots peuvent en effet dénoter un objet concret (une représentation graphique), mais recouvrent également une acception appartenant à la sous-classe {objet_scientifique/représentation} (*équivalent, figure_2, image_2, signe, symbole*). La visualisation de l'intension des concepts permet ainsi d'identifier ce type de lien sémantique inter-classes.

Nous avons également pu observer que la classe {collectif_partitif} a pour attribut définitoire le fait d'être en relation objet avec le verbe *constituer*. Cette relation peut donc être ajoutée en tant que test d'appartenance à la classe dans notre classification sémantique. Nous avons alors fait le choix d'associer à la sous-classe {collectif_partitif/collectif}[22] (*classe, ensemble, groupe*) le test d'appartenance : *Plusieurs SN constituent un N*.

---

[21] Les numéros d'acceptions renvoient ici à ceux présents dans notre classification présentée dans le tableau 2.

[22] Nous n'avons pas appliqué le test à l'autre sous-classe {/partitif} (*partie, section, segment*) de cette classe dans la mesure où les tests identifiés précédemment pour cette sous-classe nous paraissaient plus opératoires.

A contrario, nous pouvons également identifier les attributs équivalents donc redondants : ainsi, les attributs *leur_det#determ* et *son_det#determ* (dont la co-présence peut être un argument pour leur regroupement sous une même étiquette).

Une autre observation peut se faire au niveau de la polysémie des objets à classer, les noms du LST. Pour chaque nom, nous définissons un profil combinatoire, qu'il renvoie à une ou plusieurs acceptions de notre classification. De plus, la polysémie se retrouve également au niveau des cooccurrents. Un mot ambigu pourra alors être une propriété pour un objet dans une de ses acceptions et une autre propriété pour un autre objet dans une acception différente. Cependant, dans le calcul du treillis, ces deux objets seront regroupés car partageant le même attribut.

Nous pouvons illustrer ceci avec l'exemple de *selon* (dans ses acceptions en tant que préposition) qui peut dénoter deux sens principaux : 'se référant à' et 'en fonction de'.

Si l'on considère le concept ayant comme seul attribut et donc comme intension globale *selon_prep#prepobj*, on peut dégager deux ensembles syntagmatiques disjoints :

- *conception, auteur, définition* sont en cooccurrence avec *selon* dans son acception 'se référant a' ;
    - ex : ***Selon*** *cette conception, le premier pilier repose sur le régime général […]* ;[23]
- *lieu, position, valeur, état* sont en cooccurrence avec *selon* dans son acception 'en fonction de' ;
    - ex : *Il est, en conséquence, équivalent de déterminer un salaire* ***selon*** *le lieu de travail*.[24]

Nous voyons ici la limite de représenter nos objets (entrées lexicales du LST) par des relations avec des cooccurrents recouvrant différentes acceptions. En

[23] Rémond, A. (2007). Le rôle politique des sondages. *Actes de la recherche en sciences sociales* (n° 169), p. 48-71.

[24] Zenou, Y. (2000). Externalités spatiales, économies d'agglomération et formation endogène d'une ville monocentrique. *Annales d'économie et de statistique,* 58, 233-251.

représentant non pas les relations indépendamment, mais un contexte syntaxique élargi, sorte de patron lexico-syntaxique, les risques de polysémie seraient amoindris, mais il est plus que probable que peu de patrons répondraient aux critères de fréquence et de répartition définis pour les attributs.

Le même phénomène s'observe avec le verbe *constituer* qui peut dénoter la réalisation (comme dans *nous constituons un corpus*) ou l'équivalence (comme dans *ce critère constitue un bon indice*).

A contrario, certains attributs permettent des regroupements sémantiquement cohérents et quelquefois conformes à notre classification manuelle. Ainsi, le verbe *présenter* prend pour objet des éléments de la classe {objet_scientifique} tels que *modèle, méthode, théorie, hypothèse*. De même, le verbe *utiliser* est en cooccurrence avec des noms renvoyant à notre classe {objet_scientifique}.

Nous avons pu identifier certains concepts ayant une forte cohérence sémantique.

Id du concept	**Extension**	**Intension**
1627	*analyse_nom* *enquête_nom* *observation_nom* *recherche_nom* *travail_nom* *étude_nom*	*mener_verb#~obj* *mettre_verb#~subj* *montrer_verb#~subj* *réaliser_verb#~obj*
1932	*effet_nom* *phénomène_nom* *processus_nom* *évolution_nom*	*analyse_nom#~nmod* *faire_verb#~subj* *observer_verb#~obj*
2074	*démarche_nom* *méthode_nom* *opération_nom* *stratégie_nom*	*consister_verb#~subj* *mettre_verb#~obj* *tel_adj#u3_adjmod*
447	*article_nom* *ouvrage_nom* *texte_nom*	*intituler_verb#nmod* *publier_verb#nmod*

Tableau 4.3: Exemple de concepts issus de la FCA

La visualisation peut être problématique pour cette méthode, car nous travaillons sur des treillis de grande taille composés d'environ 10 000 concepts, calculés à partir de contextes incluant 493 objets et, selon les paramètres, plus de 3100 attributs (qui correspondent à l'union des profils combinatoires de l'ensemble des objets). Nous pouvons cependant remarquer que les colligations ne génèrent pas de regroupements en concepts pertinents. Certaines relations lexico-syntaxiques permettent le regroupement d'éléments proches sémantiquement, notamment les relations avec des cooccurrents nominaux et verbaux. L'utilisation de patrons lexico-syntaxiques (ou séquences de relations lexico-syntaxiques) apparaît comme une solution potentielle pour éviter la génération de milliers de concepts peu pertinents. Cependant, comme observé par ailleurs, les concepts cohérents intégrant deux relations lexico-syntaxiques sont déjà peu nombreux, l'utilisation de patrons amoindrirait leur nombre encore davantage.

4.6.4 Évaluation avec un sous-ensemble du LST

Afin d'évaluer l'apport possible de la FCA dans le processus de classification sémantique des noms du LST, nous examinons des concepts générés pour un sous-ensemble du LST afin de quantifier le nombre de regroupements cohérents. Nous voulons également vérifier si la fréquence des objets influe sur la qualité des regroupements.

Nous définissons un jeu de tests en sélectionnant 40 noms, 20 très fréquents et 20 peu fréquents, afin d'aboutir à une liste équilibrée. Nous tenons également compte de la proportion de mots polysémiques (dans ce cas, polysémique veut dire pouvant appartenir à plus d'une classe ou sous-classe sémantique du LST). Nous avons par exemple les mots *sens, modèle, type, groupe, résultat* qui ont plus d'une acception dans notre ressource du LST. Nous représentons cette polysémie dans la FCA en attribuant à ces objets l'ensemble des traits sémantiques correspondant à leurs différentes acceptions. Les 40 noms du LST sélectionnés sont présentés dans le tableau suivant.

Lemme	Classe	Sous-classe	Freq (par million de mots)
relation	{relation}	{/association}	783.558
question	{objet_scientifique}	{/thème}	752.943
analyse	{processus_cognitif}	{/examen}	710.125
étude	{processus_cognitif}	{/examen}	703.712
modèle	{objet_scientifique}	{/explicatif_complexe}	638.347
	{objet_scientifique}	{/représentation}	
fait	{objet_scientifique}	{/cas}	630.487
situation	{objet_scientifique}	{/cas}	623.661
cas	{objet_scientifique}	{/cas}	623.247
groupe	{déterminant}	{/quantifiant}	610.008
	{collectif_partitif}	{/collectif}	
	{personne}	{/collectivité}	
niveau	{espace}	{/localisation}	606.078
type	{déterminant}	{/qualifiant}	586.634
	{objet_scientifique}	{/représentation}	
résultat	{quantité}	{/mesure}	581.669
	{relation}	{/implication}	
effet	{relation}	{/implication}	580.015
sens	{espace}	{/orientation}	556.227
	{objet_scientifique}	{/explicatif_simple}	
sujet	{objet_scientifique}	{/thème}	506.168
ensemble	{déterminant}	{/quantifiant}	468.728
	{collectif_partitif}	{/collectif}	
contexte	{espace}	{/domaine}	449.077
système	{objet_scientifique}	{/méthode}	447.008
rapport	{relation}	{/implication}	432.322
article	{communication_support}	{/document}	412.257
totalité	{déterminant}	{/quantifiant}	32.683
	{collectif_partitif}	{/collectif}	

validité	{état_qualité}	{/axiologique-positif}	32.269
renforcement	{processus_évolutif}	{/amélioration_augmentation}	31.648
contraste	{relation}	{/opposition}	30.821
thématique	{objet_scientifique}	{/thème}	29.994
singularité	{état_qualité}	{/nouveauté-positif}	29.373
généralisation	{processus_cognitif}	{/interprétation}	29.166
	{processus_évolutif}	{/amélioration_augmentation}	
liaison	{relation}	{/association}	29.166
paradoxe	{relation}	{/opposition}	28.339
extérieur	{espace}	{/localisation}	27.098
minimum	{quantité}	{/rapport}	26.684
indication	{relation}	{/implication}	25.857
rigueur	{état_qualité}	{/axiologique-positif}	23.374
appréhension	{processus_cognitif}	{/compréhension}	23.374
appréciation	{processus_cognitif}	{/évaluation}	22.961
incidence	{relation}	{/implication}	22.340
mérite	{état_qualité}	{/axiologique-positif}	21.926
avancée	{état_qualité }	{/axiologique-positif}	20.685
originalité	{état_qualité}	{/nouveauté-positif}	20.478
divergence	{relation}	{/opposition}	18.203

Tableau 4.4: Liste des noms du LST pour l'évaluation de la FCA

Nous faisons ainsi le choix d'une évaluation des regroupements automatiques par comparaison avec la classification manuelle. Une autre solution aurait été d'évaluer manuellement tout ou partie des regroupements automatiquement générés afin de leur attribuer un score représentant la qualité du concept. Ce type d'évaluation, que nous avons jugé trop coûteux, a par exemple été effectué par Tanguy *et al.* (2015) dans leurs travaux sur l'analyse distributionnelle automatique. Ils définissent ainsi un « gold standard » des meilleurs voisins pour chaque mot cible (élément à classer) en faisant appel à des juges. Les juges doivent décider, pour chaque paire de mots présentée, si les mots « sont sémantiquement proches dans le domaine du TAL » (2015, p. 107). Le critère pour évaluer une paire est le nombre de juges ayant validé le voisin. Ils obtiennent alors un accord inter-annotateurs au kappa de Fleiss à 0,55. Nous pouvons comparer ce score à celui

obtenu lors de notre tâche de validation manuelle du LST (voir section 2.2.2.2.2), pour lequel nous avons obtenu un kappa de Fleiss à 0,45 pour les noms (contre 0,56 pour les travaux de Tanguy *et al.*) 0,23 pour les verbes (contre 0,50) et 0,75 pour les adjectifs (contre 0,59).

Tanguy *et al.* font ainsi un constat similaire à celui effectué lors de l'évaluation manuelle des candidats-LST : l'annotation des adjectifs est plus simple que celles des noms et des verbes. La comparaison de ces scores doit cependant tenir compte du fait que les questions posées aux évaluateurs diffèrent dans les deux études. Dans notre cas, les juges devaient évaluer l'appartenance d'un mot à un ensemble lexical (le LST), dans le second, les juges ont évalué la similarité sémantique de paires de mots.

Nous nous concentrons par la suite sur des concepts selon différents critères, afin de dégager les attributs pertinents et de confronter notre classification manuelle aux résultats de la FCA.

4.6.4.1 Concepts définis au minimum par 2 noms et 2 attributs

Nous nous intéressons ici plus spécifiquement aux concepts intégrant un minimum de 2 objets et 2 attributs dont nous présentons des exemples dans le tableau ci-dessous.

ID	**Objets**	**Attributs**	**Extension**	**Intension**
193	4	6	*modèle_nom* *résultat_nom* *analyse_nom* *étude_nom*	*présenter_verb#~obj* *résultat_nom#~nmod* *montrer_verb#~subj* *présenter_verb#~subj* *notre_det#determ* *voir_verb#~obj*
270	4	4	*groupe_nom* *type_nom* *fait_nom* *ensemble_nom*	*constituer_verb#~subj* *apparaître_verb#~subj* *relation_nom#~nmod* *considérer_verb#~obj*
375	4	4	*modèle_nom* *question_nom* *type_nom* *fait_nom*	*utiliser_verb#~obj* *plusieurs_det#determ* *sembler_verb#~subj* *considérer_verb#~obj*
651	4	4	*modèle_nom* *question_nom* *cas_nom*	*type_nom#~nmod* *sembler_verb#~subj* *considérer_verb#~obj*

			situation_nom	*voir_verb#~obj*
27	4	4	*cas_nom* *résultat_nom* *article_nom* *étude_nom*	*précédent_adj#u3_adjmod* *montrer_verb#~subj* *présenter_verb#~subj* *voir_verb#~obj*
241	4	3	*cas_nom* *résultat_nom* *fait_nom* *étude_nom*	*présenter_verb#~obj* *d'autres_det#determ* *sembler_verb#~subj*

Tableau 4.5: Concepts définis au minimum par 2 noms et 2 attributs

L'identifiant du concept (colonne *ID*) permet de repérer ledit concept dans le treillis de Galois et ainsi d'accéder à ses concepts fils et pères. La colonne *Extension* liste les objets du concept. La colonne *Intension* détaille les attributs partagés par l'ensemble de ces objets. Nous pouvons observer plusieurs attributs pertinents. La détermination avec *notre* et la relation objet avec *présenter* peuvent être des indices d'éléments des classes {objet_scientifique} et {processus_cognitif}, que s'approprie l'auteur, ou qui le représentent implicitement. La combinaison de ces deux attributs paraît plus intéressante que la seule présence de l'un ou l'autre. Nous relevons de nouveau que la relation objet avec le verbe *constituer* est un bon critère pour la classe {collectif_partitif}, de même que la relation objet avec le verbe *utiliser* pour la sous-classe {objet_scientifique/instrument} (*instrument, moyen, outil*).

Nous remarquons également que les objets avec le plus d'attributs ont naturellement tendance à s'intégrer dans un grand nombre de concepts. La même observation peut être faite pour les attributs (relation et lemme-catégorie fréquents) qui interviennent dans un grand nombre de concepts hétérogènes (par exemple la détermination avec *d'autres*, les relations avec les verbes *sembler, permettre*).

4.6.4.2 Analyse des concepts selon la classification sémantique du LST

Nous observons ensuite les concepts dont l'intension comprend une de nos classes ou sous-classes sémantiques afin d'évaluer le recoupement entre la méthode manuelle et la FCA.

Nous remarquons qu'aucun attribut (mis à part l'étiquette sémantique) ne regroupe les éléments des classes {objet_scientifique} et {relation}. Le concept à l'extension la plus grande pour la classe {objet_scientifique} prend pour seul attribut la relation objet avec le verbe *constituer*, qui nous paraît trop large comme test lexico-syntaxique d'appartenance à cette classe. Aucun attribut ne permet de regrouper les éléments de la classe {état_qualité} qui se trouve être la seule classe représentée uniquement par des mots peu fréquents.

Le concept correspondant au meilleur regroupement pour la sous-classe {relation/implication} a dans son extension trois éléments de cette sous-classe : les trois noms les plus fréquents, *résultat, effet* et *rapport*. Ce concept a pour intension : *trouver_verb#~subj|produire_verb#~obj|compte_nom#~nmod|analyse_nom#~nmod|voir_verb#~obj*.

Nous observons également une bonne définition de la classe {déterminant}[25], qui inclut l'ensemble des éléments sauf celui à basse fréquence (*totalité*). Le concept a pour intension : *constituer_verb#~obj|constituer_verb#~subj|apparaître_verb#~subj|relation_nom#~nmod|considérer_verb#~obj*.

Pareillement, la classe {espace} voit l'intégralité de ses éléments (fréquents) réunis dans le concept ayant pour intension : *compte_nom#~nmod|donner_verb#~subj|définir_verb#~subj|définir_verb#~obj*.

La sous-classe {objet_scientifique/cas} est regroupée par les attributs : *observer_verb#~obj|considérer_verb#~obj* alors que la sous-classe {objet_scientifique/thème} est définie par l'attribut : *aborder_verb#~obj*.

Nous pouvons ainsi identifier rapidement les attributs constituant l'intension des concepts les plus satisfaisants (i.e. intégrant un maximum de noms d'une même classe). Nous notons également que certains attributs correspondent aux relations lexico-syntaxiques utilisées dans les tests d'appartenance lors de la création de la classification sémantique manuelle.

[25] La classe {déterminant} est composée d'élément rentrant dans la composition de déterminant complexe, tels *ensemble, totalité, partie, genre* …

4.6.4.3 Comparaison des concepts par fréquence des noms

Nous présentons dans le tableau ci-dessous le nombre moyen de concepts dans lesquels apparaissent les noms du LST en fonction de leur groupe de fréquence (très fréquents ou peu fréquents). Pour chaque nom du LST, nous calculons le pourcentage de voisins, ou co-membres de concepts, qui partagent effectivement un trait de classes ou sous-classes sémantiques avec lui. Ces unités, qui sont reliées dans notre classification sémantique ainsi que dans les concepts de FCA, sont appelés voisins valides pour la suite de l'expérimentation. Sont pris en compte les concepts dont l'intension n'est pas uniquement composée d'étiquettes sémantiques et dont l'extension est supérieure à 1. Nous regroupons les noms dans deux catégories représentant les tranches supérieure et inférieure de fréquence.

	Noms très fréquents	**Noms peu fréquents**
Nombre de Voisins	8115,3	227,7
% Voisins Classe OK	37,7 %	34,6 %
% Voisins Sous-classe OK	11,2 %	12,6 %

Tableau 4.6: Validité des regroupements selon la fréquence

La ligne *Nombre de Voisins* indique, pour les deux groupes de noms, le nombre moyen d'objets avec lesquels un nom du LST rentre en relation dans les concepts générés. Les lignes suivantes donnent la proportion de ces regroupements qui sont considérés comme valides, i.e. également présents dans notre classification manuelle. Ainsi la deuxième ligne donne le pourcentage de voisins ayant au minimum un trait de classe en commun avec les noms considérés. La dernière ligne est le résultat de ce calcul au niveau des sous-classes.

Nous voyons que les noms très fréquents rentrent logiquement dans un plus grand nombre de concepts et ont donc ainsi un plus grand nombre de voisins. Ceci est dû majoritairement au fait que ces noms fréquents sont ceux qui ont le plus d'attributs les représentant. Au-delà de cette observation, nous ne pouvons déduire d'incidence directe de la fréquence sur la qualité des regroupements pour notre échantillon des 40 noms à très haute et très basse fréquence. Nous constatons que, logiquement, les éléments d'une même classe sont plus fréquemment regroupés que les éléments d'une même sous-classe.

Si nous nous intéressons à la qualité des regroupements en fonction de la fréquence, nous observons, tel qu'illustré ci-après, que ce critère seul ne suffit pas à expliquer la validité ou non des regroupements.

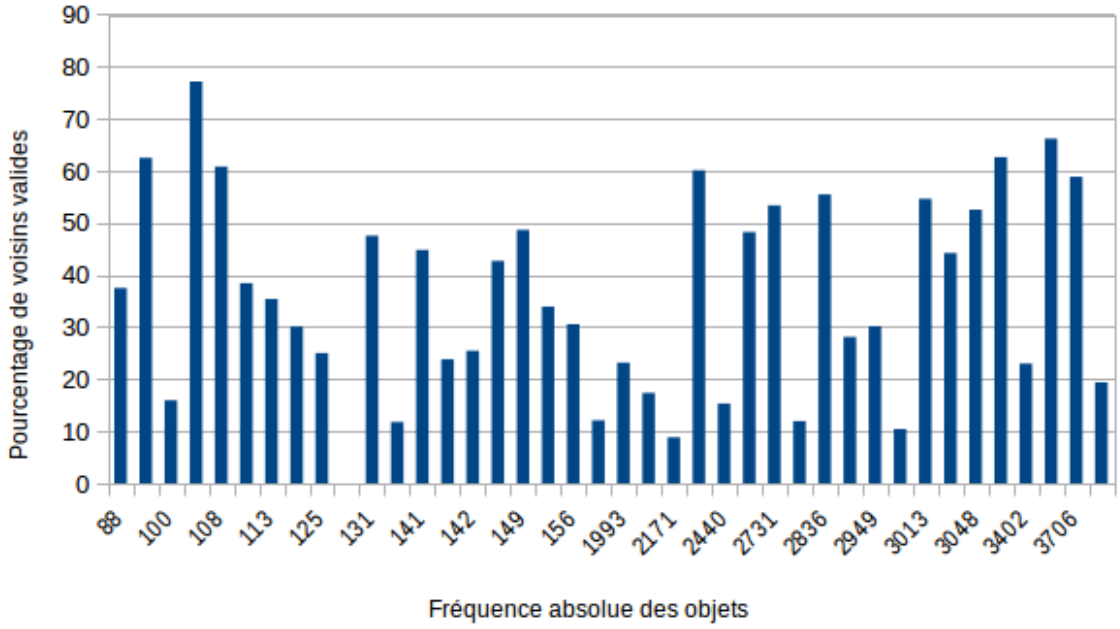


Illustration 4.3 : Validité des regroupements de la FCA selon la fréquence

Le pourcentage de voisins valides est exprimé en pourcentage, en ordonnée. En abscisse est représentée la fréquence absolue de ces noms (allant de 88 à 3788 occurrences, respectivement pour *divergence* et *relation*).

S'il existait une corrélation entre la fréquence et la qualité des regroupements, l'illustration ci-dessus représenterait ce lien par une augmentation (ou une baisse dans le cas d'une relation négative) du pourcentage de voisins valides avec la fréquence absolue des objets. Nous ne voyons ici aucune corrélation apparaître entre fréquence et validité des regroupements. La haute fréquence d'un mot, qui assure une certaine richesse du profil combinatoire, ne garantit cependant pas de meilleurs résultats dans la classification pour cette expérimentation.

Si nous prenons en compte le score de spécificité (ratio de la fréquence relative dans le corpus d'analyse sur la fréquence relative dans le corpus de contraste), les observations restent les mêmes. Nous ne pouvons dégager de corrélation entre la spécificité et la qualité des regroupements. Les noms les plus

sur-représenté du LST ne sont pas mieux regroupés que les autres, comme le montre l'illustration suivante.

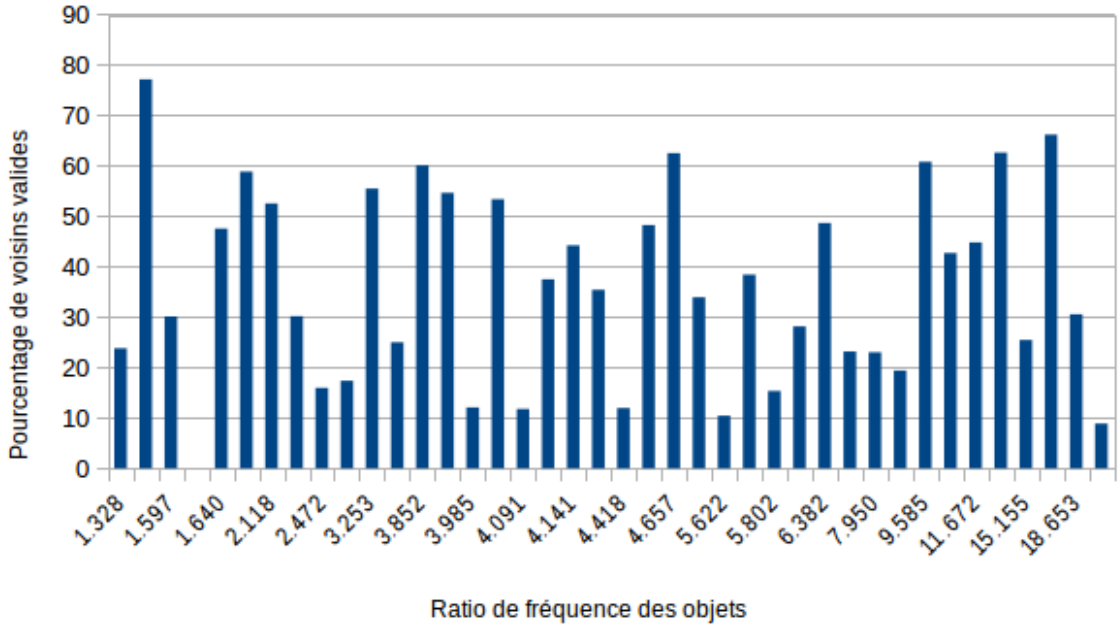


Illustration 4.4 : Validité des regroupements de la FCA selon la spécificité

Nous pouvons ici faire la même observation que pour l'illustration 4.3, dans la mesure où aucune corrélation n'apparaît entre la valeur du ratio de fréquence d'un nom du LST et la qualité des regroupements dans lesquels il est inclus. Une haute spécificité, i.e. une sur-représentation élevée dans le corpus scientifique comparé au corpus de contraste, n'assure ainsi pas une classification cohérente, dans le cadre de cette expérimentation, pour les mots du LST. Pareillement, nous n'avons observé aucun lien de corrélation en comparant le pourcentage de voisins valides en fonction de la répartition des noms (nombre de tranches du corpus d'analyse dans laquelle le mot a au moins une occurrence).

Nous avons ensuite voulu mesurer la variation de la qualité de regroupements en fonction des classes de notre classification sémantique. Nous avons pour cela calculé le pourcentage de voisins valides après avoir appliqué la méthode des treillis de Galois sur l'ensemble des noms du LST de notre classification sémantique (493 noms pour 531 acceptions). Nous nous intéressons donc à la qualité des regroupements en fonction des classes sémantiques des

éléments, autre variable pouvant expliquer ces différences. Nous pouvons observer ci-dessous les pourcentages de voisins valides selon les classes.

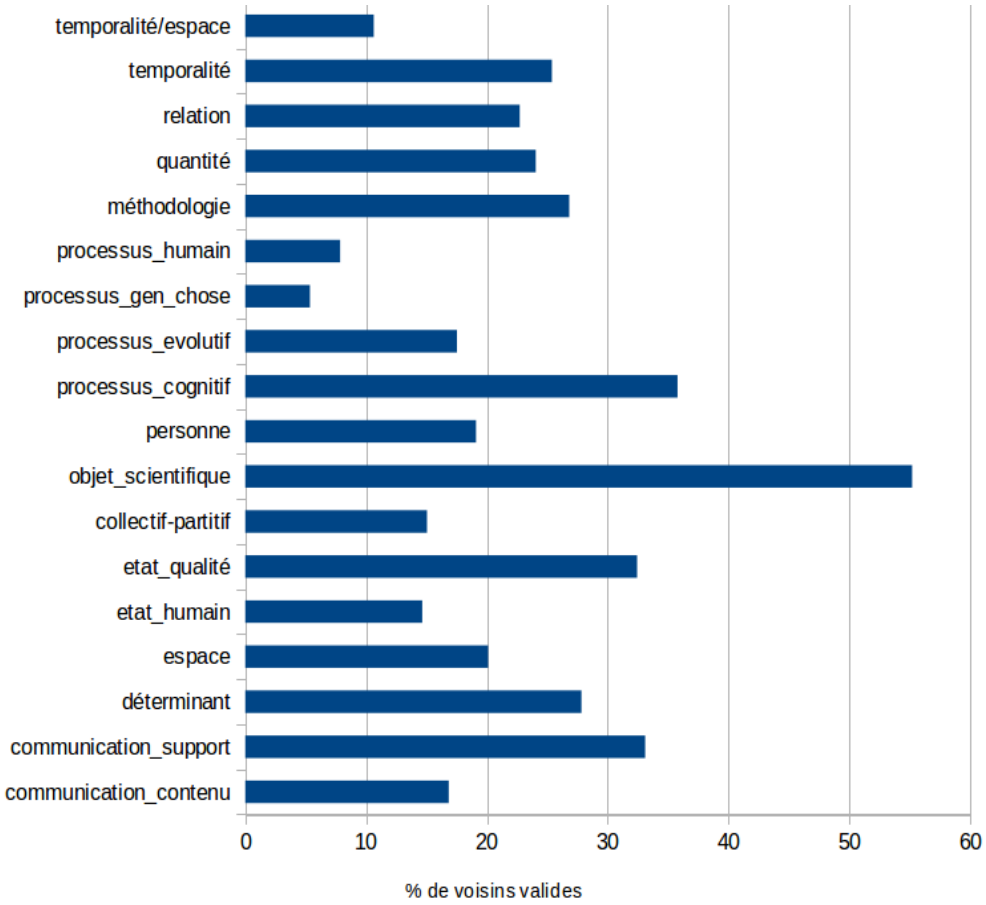


Illustration 4.5: Validité des regroupements de la FCA selon les classes

Les meilleurs scores sont obtenus pour la classe {objet_scientifique} pour laquelle nous avons pu identifier un attribut opérant : la relation objet avec le verbe *constituer*. La classe {processus_cognitif} permet également de bons regroupements. Ces deux classes sont parmi les plus larges de notre classification.

Parmi les classes dont les regroupements sont peu satisfaisants, nous observons que les classes {processus_gen_chose} et {processus_humain} ont pour point commun de renvoyer à des notions peu précises (*tâche, activité, travail, utilisation, recours*), et ont notamment pour membres des noms du LAG.

L'observation de la qualité des regroupements en fonction de la fréquence et/ou des classes sémantiques considérées ne permet pas de dégager de corrélation, si ce n'est que la classe {objet_scientifique} est celle pour laquelle la classification est la plus satisfaisante.

Les différences de scores entre classes s'expliquent par plusieurs facteurs :

- la granularité de nos classes n'est pas homogène ;
- l'extension des classes varie également (de 5 pour {état_humain} à 126 pour {objet_scientifique} dans notre ressource) ;
- les noms à classer ont des fréquences très variables et renvoient à une ou plusieurs acceptions ;
- les cooccurrents présents dans les attributs peuvent aussi renvoyer à plusieurs acceptions ;
- la richesse des profils combinatoires est également sujette à de fortes variations : le mot *groupe* a dans notre protocole 63 attributs, alors que *singularité* n'est représenté que par 3 relations lexico-syntaxiques.

Nous retrouvons ici les limites entrevues dans la première expérimentation des prototypes (essentiellement les difficultés liées au fait de traiter des mots à fréquences diverses, à la polysémie, à la grande variation du nombre d'attributs définis pour chaque élément).

Cette méthode nous a néanmoins permis d'identifier certaines propriétés définitoires de nos classes, les équivalences entre attributs (et donc de la redondance potentielle dans la définition de ces propriétés), et ainsi de mettre en lumière certains obstacles à une classification cohérente. En effet, la polysémie des cooccurrents se retrouve dans des regroupements non-homogènes résultant de l'union des contextes renvoyant à des acceptions différentes.

4.7 Conclusions sur les apports des méthodes automatiques

Les résultats des deux expérimentations ne sont pas complètement concluants dans l'optique d'une classification automatique non-supervisée des noms du LST. Cependant, l'apport de ces méthodes n'est pas négligeable dans le cadre d'une aide à la catégorisation, pour laquelle les classes sont prédéfinies. L'utilisation des traits sémantiques, dans l'identification de routines sémantico-rhétoriques, nous oblige à catégoriser le lexique finement. Or, Bouaud *et al.* (2000) notent que la classification distributionnelle génère des classes trop générales mais pertinentes en matière d'ontologie. Nous avons déjà évoqué cet intérêt en vue de la

représentation de notre ressource. De plus, nous avons pu, par l'observation des concepts créés par l'analyse formelle de concepts, identifier des attributs définitoires pour certaines classes, détecter les attributs trop vagues, générant des regroupements peu pertinents, valider certains regroupements, en questionner d'autres. Ainsi, bien que les classes automatiquement générées ne soient pas immédiatement exploitables, ces méthodes présentent l'avantage de mettre en lumière les attributs pertinents ou à l'inverse ceux qui sont à la source de regroupements moins intéressants. Une phase de validation manuelle de ces résultats est ainsi toujours nécessaire, comme le soulignent Habert & Zweigenbaum (2003), tant au niveau de la validation des classes elles-mêmes qu'au niveau des attributs la définissant. L'identification de ces attributs pouvant jouer le rôle de critères d'appartenance à une classe sémantique facilite alors la tâche de mise à jour de la ressource lors d'ajout de nouveaux mots dans la classification. Une telle expérimentation, que nous pourrons effectuer pour les futurs « repêchages » d'éléments du LST, nous permettrait alors de compléter l'évaluation de ces méthodes.

La méthode des prototypes, malgré des résultats peu probants, présente selon nous plusieurs intérêts. Elle permet dans un premier temps d'intégrer la notion de degré d'appartenance dans la définition des classes. Lors de notre travail de classification manuelle des noms du LST, présenté dans le chapitre précédent, nous avons effectivement constaté que si certains mots sont « simples » à regrouper, d'autres se révèlent plus problématiques à traiter. En intégrant une mesure de ce degré d'appartenance, nous pourrions alors distinguer les éléments « centraux » d'une classe de ceux dont l'appartenance à la classe est moins évidente. Nous pourrions dans un second temps vérifier si ces éléments « centraux » se révèlent plus ou moins opérant que les éléments « marginaux » dans la détection de routines. La notion de degré d'appartenance présente également l'avantage de pouvoir représenter la polysémie des éléments à classer, et est en ce sens adaptée aux noms du LST renvoyant à plus d'une acception. Enfin, comme nous l'avons mis en évidence, la méthode des prototypes permet de faire émerger certains traits définitoires pour les classes et donc d'améliorer la définition de nos tests d'appartenances aux classes sémantiques.

Dans la perspective d'une future extension de la ressource, ces deux méthodes pourraient ainsi être utilisées afin de guider le linguiste dans

l'affectation de classe pour les nouvelles unités lexicales de la ressource. Ces méthodes automatiques pourront également être évaluées en fonction de leur apport pour de futurs traitements, tels la détection de routines sémantico-rhétoriques ou de patrons combinant éléments du LST et éléments terminologiques. Ces éléments phraséologiques, défini par un ensemble de contraintes lexicales, syntaxiques et sémantiques, sont ainsi le sujet du chapitre suivant.

Chapitre 5
Des patrons aux routines

Sommaire

5.1 Étude syntactico-sémantique du LST

Nous présentons dans cette partie, à la suite des traitements sémantiques détaillés dans les chapitres précédents, nos travaux d'analyse du LST dans sa dimension phraséologique.

Nous avons vu dans le chapitre 1 que le genre de l'article de recherche en SHS se caractérise essentiellement au niveau des phénomènes lexico-syntaxiques. Plusieurs travaux ont ainsi montré que la phraséologie est particulièrement présente dans l'écrit scientifique (C. J. Gledhill, 2000 ; Pecman, 2004a ; Tutin & Kraif, 2016). De fait, les travaux prenant pour objet des unités phraséologiques sont nombreux, notamment dans le cadre du projet Scientext[1] (Jacques, 2011; Tran, 2014b; Tutin, 2014, 2015; Tutin, Grossmann, Falaise, & Kraif, 2009; Tutin *et al.*, 2015). De même, plusieurs études se sont concentrées sur les constructions verbales associées à ce genre d'écrits (Granger & Paquot, 2009b; Yan, 2012).

L'analyse des expressions polylexicales (collocation, routines, patrons verbaux) présente plusieurs intérêts, en particulier, l'accès à la dimension rhétorique du discours scientifique, la construction de ressources pédagogiques fonctionnelles, et les applications dans des processus de recherche d'informations.

Ces trois aspects sont abordés dans le présent chapitre qui détaille trois expérimentations s'appuyant sur la classification sémantique du LST et sur notre corpus d'analyse arboré. Nous nous intéressons ici à des ensembles polylexicaux définis par différentes contraintes lexicales, syntaxiques et sémantiques, et dont l'apport se situe à différents niveaux, ainsi que nous le détaillons ci-après.

La première expérimentation, présentée dans la section 5.2, part du constat que la maîtrise d'éléments phraséologiques s'avère complexe pour les apprenants dans le contexte de l'écrit scientifique (voir section 5.2.4). Nous avons ainsi effectué, en collaboration avec Yan, une première étude sur les constructions des verbes du LST (Yan & Hatier, 2016), en vue d'élaborer une ressource pédagogique des patrons verbaux. Dans un second temps, nous avons effectué une analyse comparative des constructions verbales dans des corpus d'experts, d'étudiants francophones natifs et d'étudiants apprenants du français, afin d'identifier les

[1] Projet ANR (2007-2010) « Corpus et outils de la recherche en sciences humaines et sociales ». http://scientext.msh-alpes.fr [consulté le 14/09/2016]

besoins particuliers des étudiants dans le cadre de la rédaction d'écrits scientifiques (Hatier & Yan, à paraître).

Dans un deuxième temps, section 5.3, nous nous intéressons aux cooccurrences entre les éléments du LST (de toutes les catégories syntaxiques : verbe, nom, adjectif, adverbe) et un élément terminologique (annoté manuellement ou automatiquement). Notre objectif, en matière de recherche d'informations, est alors d'identifier, dans la continuité des travaux du projet TermITH, des contextes favorables à l'interprétation terminologique d'un candidat à ce statut, avec pour enjeu l'amélioration de l'indexation terminologique automatique.

La dernière partie (section 5.4) est consacrée à l'étude de routines, configurations définies par des relations de dépendance entre des unités lexicales, elles-mêmes définies par leur lemme, leur catégorie ou leur classe sémantique. Nous avons pour cela appliqué un outil d'extraction d'expressions polylexicales, élaboré par Kraif (2016), sur notre corpus d'analyse arboré, enrichi en étiquettes sémantiques pour les mots du LST.

Pour l'ensemble des analyses présentées ici, nous nous basons sur les dépendances annotées dans le corpus d'articles afin d'identifier des unités lexicales liées syntaxiquement, non obligatoirement contiguës, à la différence des segments répétés, dont le processus d'identification se base sur la proximité et la définition d'un empan. L'approche syntaxique permet de faire émerger des constructions plus abstraites que la méthode des segments répétés. En effet, la présence de modifieurs (adjectivaux, adverbiaux) autour des éléments constituant la routine (verbe, noms sujets, noms objets ou compléments du verbe) ne modifie pas la nature du sous-arbre. Ainsi, dans les deux séquences suivantes, le sous-arbre minimal comprenant *faire* et *hypothèse* ne varie pas :

- *CHEN (1997) fait l'hypothèse que les parts de marché sont initialement données de manière exogène [...]*[2]
- *[...] on fait généralement aussi l'hypothèse que les préférences mêmes des acheteurs sont inconnues*[2]

[2] Malin, E., & Martimort, D. (2001). Les limites à la discrimination par les prix. *Annales d'Economie et de Statistique*, 209-249.

Outre cet avantage de disposer d'un corpus arboré, nous tirons parti des traitements précédemment effectués, en projetant les traits sémantiques du LST dans le corpus. Nous suivons ainsi le principe d'une étude incrémentale du LST, pour enrichir la caractérisation de ce lexique. Notre étude des patrons dans lesquels s'intègre le LST exploite également les propriétés distributionnelles, par l'étude de configurations communes à une classe sémantique, suivant en cela le principe que « la similarité distributionnelle reflète clairement la similarité sémantique », selon les mots de Bertels & Geeraerts (2012, p. 140). Nous nous situons ainsi dans la continuité de Dubois & Dubois-Charlier (1997) ou Levin (1993), dans une approche sémantico-syntaxique. Notre étude de la phraséologie du LST part de cette observation que les éléments partageant des comportements syntaxiques forment souvent des classes sémantiques homogènes.

5.2 Analyse des patrons du LST

Nous avons effectué, à la suite du mémoire de Yan (2012), une étude sur les patrons lexico-syntaxiques des verbes du LST. Ces patrons, définis par Yan sur le modèle *Corpus Pattern Analysis* (CPA) de Hanks (2004), ont été formalisés manuellement d'après une extraction automatique des constructions dans lesquelles s'inscrivent les verbes du LST.

Notre but était de dégager les patrons verbaux les plus fréquents dans le genre de l'écrit scientifique, avec pour perspective la constitution d'une ressource de ces patrons destinée à améliorer l'enseignement/apprentissage du Français sur Objectif Universitaire. Plusieurs travaux ont pris pour objet d'étude des constructions verbales dans l'écrit scientifique et ont constaté que les scripteurs non experts éprouvaient des difficultés à les maîtriser (Nesselhauf, 2005 ; Hyland, 2008 ; Granger & Paquot, 2009). Les besoins didactiques nécessitent la meilleure contextualisation possible, exigence à laquelle le concept de patron lexico-syntaxique adopté ici répond. La prise en compte des alternances syntaxiques et des classes sémantiques dans les patrons a également pour but d'améliorer l'appropriation et la réutilisation de ces schémas lexico-syntaxiques pour les apprenants.

En nous basant sur un corpus pour enrichir semi-automatiquement un lexique verbal en informations syntaxiques, nous reprenons une approche adoptée

dans des travaux d'acquisition automatique de cadres de sous-catégorisation en anglais (Manning, 1993; O'Donovan, Burke, Cahill, Van Genabith, & Way, 2004) ou en français (Kupsc, 2007).

Ce que nous nommons pour cette étude cadre de sous-catégorisation comprend les arguments et les compléments du verbe qui ont un rôle de modifieur récurrent. La phase de modélisation manuelle des patrons, à partir des cadres, permet à Yan de filtrer les éléments considérés comme pertinents dans la description de constructions verbales. Considérons l'exemple suivant pour le verbe *montrer* :

- *Comme le montre*$_{LST}$ *l'exemple de cet étudiant de 36 ans en troisième année de thèse en économie qui est agent de sécurité à temps partiel*[…][3]

Le verbe *montrer* est ici en relation de dépendance avec la conjonction *comme* que nous incluons dans le cadre. De même, les constructions du type *nous pouvons penser, si nous considérons* doivent pouvoir être identifiées du fait de leur récurrence dans les écrits scientifiques.

D'autres études se basent sur les stricts participants au schéma de régime, les sujets, les compléments directs et indirects. Ainsi en est-il de deux ressources élaborées manuellement, le lexique-grammaire de Gross (1975) et le *Dictionnaire Explicatif et Combinatoire* (*DEC*) (Mel'čuk, Arbatchewsky-Jumarie, Iordanskaja, Mantha, & Polguère, 1999), décrivant de façon systématique les propriétés syntaxiques et sémantiques de chaque entrée lexicale.

À la manière du lexique-grammaire et des classes d'objets (G. Gross, 2008), nous souhaitons intégrer les classes sémantiques du LST dans la sélection des arguments des verbes du LST. Nous n'avons cependant pas pour objectif, à l'inverse du lexique-grammaire, d'inventorier l'ensemble des constructions possibles pour chaque verbe mais simplement d'identifier les patrons les plus fréquents et typiques de l'écrit scientifique, en prenant en compte les restrictions de sélection sémantique et syntaxique. L'intrication de ces deux niveaux est également à l'œuvre dans le *DEC* qui décrit notamment chaque entrée à l'aide d'actants syntaxiques et sémantiques, et fournit des informations sur des collocations au travers des fonctions lexicales. La prise en compte des collocations

[3] Péroumal, F. (2009). Le monde précaire et illégitime des agents de sécurité. *Actes de la recherche en sciences sociales*, (5), 4-17.

récurrentes nous paraît importante, de même que la prise en compte de la dimension lexico-syntaxique. Nous intégrons à ces informations les modifieurs récurrents associés au verbe. Nous procédons ainsi à une extraction automatique des cadres de sous-catégorisation, au sens large, qui sont ensuite regroupés et modélisés manuellement en patrons verbaux par Yan. L'illustration suivante schématise les étapes de cette extraction.

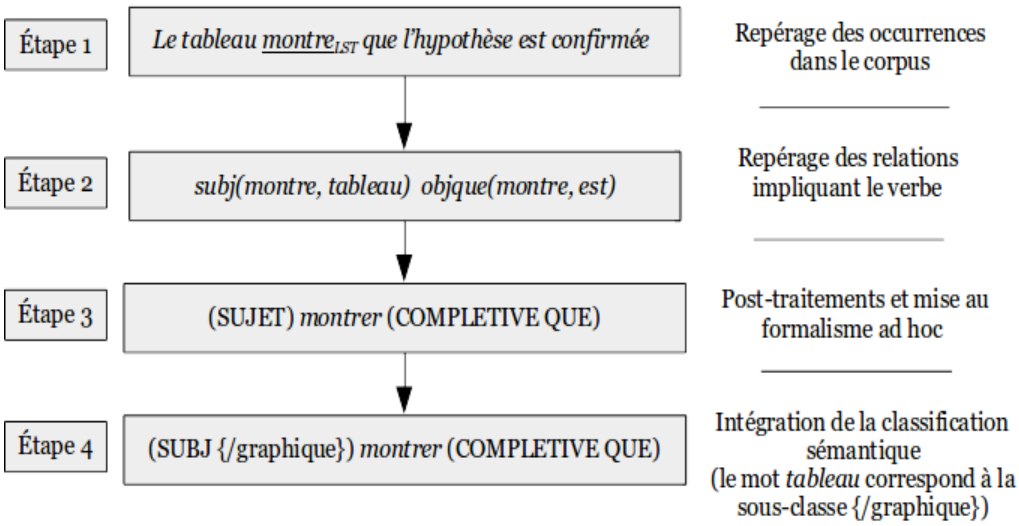


Illustration 5.1 : Méthode d'extraction des cadres de sous-catégorisation

L'étape 4 consiste à identifier les items les plus fréquents dans le cadre en question. Ces items (définis par un lemme ou un trait sémantique) sont ensuite interprétés manuellement par Yan qui valide ou non leur intégration dans les patrons. Ces derniers sont modélisés à partir des résultats de l'extraction des cadres et de l'observation des concordances pour une vérification des acceptions verbales mobilisées.

Nous détaillons dans les sections suivantes les différentes étapes menant à l'identification des patrons verbaux dans les articles de recherche.

5.2.1 Approche des constructions verbales et patrons

La présente étude des patrons lexico-syntaxiques des verbes du LST, est basée sur une approche contextualiste (Hanks, 2013; Hunston & Francis, 2000; Sinclair, 1991). Un patron est une structure syntaxique intégrant des collocations privilégiées et est notamment caractérisé par :

1. l'association entre l'acception et l'usage réel du mot ;
2. l'étiquetage sémantique au niveau des arguments.

Le modèle *Corpus Pattern Analysis* (*CPA*), centré sur les corpus et l'usage en contexte, permet de mettre en évidence les constructions typiques d'un verbe. En effet, dans ce modèle, les patrons intègrent des informations syntaxiques, lexicales

et sémantiques et sont rattachés à une acception déterminée du verbe auquel ils sont associés. Ces informations sur la combinatoire sont extraites du corpus, mettant ainsi au centre l'usage en contexte. L'usage étant alors représenté par le corpus d'analyse, la composition de ce dernier revêt une importance essentielle dans l'analyse du comportement des mots, ainsi que le rappellent[4] Messiant *et al.* (2010) dans leur étude des cadres de sous-catégorisation verbaux. En basant notre étude des constructions des verbes du LST sur un corpus d'articles scientifiques, nous nous assurons d'analyser l'usage scientifique transdisciplinaire de ces éléments.

La prise en compte de l'ensemble de ces spécificités est essentielle dans la perspective de l'élaboration d'une ressource verbale adaptée au genre de l'écrit scientifique. La pertinence de la ressource est d'autant plus importante que les verbes sont centraux dans l'expression d'un certain nombre de fonctions discursives dans l'écrit scientifique, comme le rappellent Granger & Paquot (2009b) : exprimer son positionnement, faire un état de l'art, citer, résumer, etc.

Granger et Paquot (2009b) remarquent également que certaines structures syntaxiques sont plus fréquentes dans les écrits académiques, et qu'une approche combinant lemmes et formes permet d'avoir une meilleure représentation de la diversité phraséologique présente dans les corpus. S'ensuit la nécessité d'une ressource spécialisée, inventoriant les constructions verbales du genre scientifique, afin d'améliorer l'apprentissage de la rédaction scientifique, tant au niveau de l'encodage (la production écrite) qu'au niveau du décodage (compréhension écrite).

Afin d'identifier ces structures lexico-syntaxiques, nous avons procédé à l'extraction automatique des cadres de sous-catégorisation des verbes du LST, dont les résultats sont à la base de la modélisation manuelle des patrons.

5.2.2 Extraction des cadres de sous-catégorisation

Nous nous sommes concentré, dans un premier temps, sur des verbes appartenant à une même classe sémantique. Notre objectif était alors d'observer le lien entre proximité sémantique et proximité syntaxique, à travers le partage de

[4] Les auteurs donnent pour exemple le verbe intransitif *essaimer* qui se transitive dans le genre journalistique.

constructions syntaxiques communes. Cette approche se situe notamment à la suite de Levin (1993) qui « représente le sens des verbes par le recours à des composants sémantiques [et fournit] une description systématique des alternances » (Messiant *et al.*, 2010, p. 68) Des travaux similaires ont été conduits par Dubois & Dubois-Charlier qui élaborent des classes de mots sur la base de propriétés syntactico-sémantiques partagées, de manière comparable au lexique-grammaire de Gross (travaux présentés section 3.2.2.1).

Pour cette expérimentation, nous avons sélectionné trois verbes de la sous-classe sémantique {processus/révélation}, *montrer – démontrer – indiquer*, pour analyser leurs propriétés syntaxiques. Ces trois verbes, objets de notre étude de cas, sont particulièrement fréquents (de 315 à 3053 occurrences), ce qui nous permet d'extraire un nombre important de constructions.

Le processus consiste dans un premier temps à repérer, pour chaque occurrence des verbes étudiés, l'ensemble des dépendances que nous avons choisies d'intégrer au patron, y compris certains compléments circonstanciels. La première étape permet alors d'identifier ces cadres de sous-catégorisation, ou cadres de dépendances.

Ce faisant, nous souhaitons acquérir des cadres de sous-catégorisation « sans à priori, pour faire émerger du corpus [ceux] correspondant à l'usage » (Messiant *et al.*, 2010, p. 76). Cette démarche « corpus-driven » (Tognini-Bonelli, 2001) a pour but l'identification des constructions récurrentes dans le corpus. Certains systèmes d'acquisition de constructions verbales ne sont pas adaptés pour faire émerger ces constructions, car ils reposent sur un inventaire a priori des structures visées. Notre méthode, seulement contrainte par les sorties d'analyse syntaxique, ne tombe pas dans cet écueil en intégrant automatiquement toutes les cooccurrences et constructions, présentes dans le corpus. Cependant, le fait de ne pas contraindre un certain type de construction (transitif direct, sujet exprimé, etc.) pour les verbes analysés a pour conséquence de bruiter la sortie des cadres de sous-catégorisation, en multipliant les cadres qui ne présentent qu'une seule occurrence ou en créant « artificiellement » des cadres erronés. Nous avons par exemple automatiquement extrait pour le verbe *définir* un cadre ne comportant ni sujet ni objet et renvoyant à 40 occurrences. Considérons ainsi la phrase suivante :

- *Le Répertoire pratique de Dalloz définit l'excès comme étant l'attentat de l'un des époux à la vie de l'autre*[5].

En observant l'analyse en dépendance de cet exemple, nous avons ainsi pu localiser la source de l'erreur. En effet, l'analyse syntaxique n'ayant annoté aucune relation impliquant le verbe principal *définir*, le cadre n'a pu être construit. Ces occurrences non regroupées ont alors pour cause soit une erreur d'annotation syntaxique soit une erreur dans le processus d'extraction de cadres.

La première phase correspond donc à l'extraction de l'ensemble des occurrences de chaque verbe. Considérons l'exemple suivant pour le verbe *relever* de la sous-classe {relation/appartenance} :

- *les valeurs sémantiques des prépositions relèvent essentiellement du dictionnaire*[6]

Pour aboutir à un cadre représentant cette occurrence du verbe du LST *relever,* nos traitements se basent sur les résultats de l'analyse syntaxique, tel qu'illustrés ci-dessous.

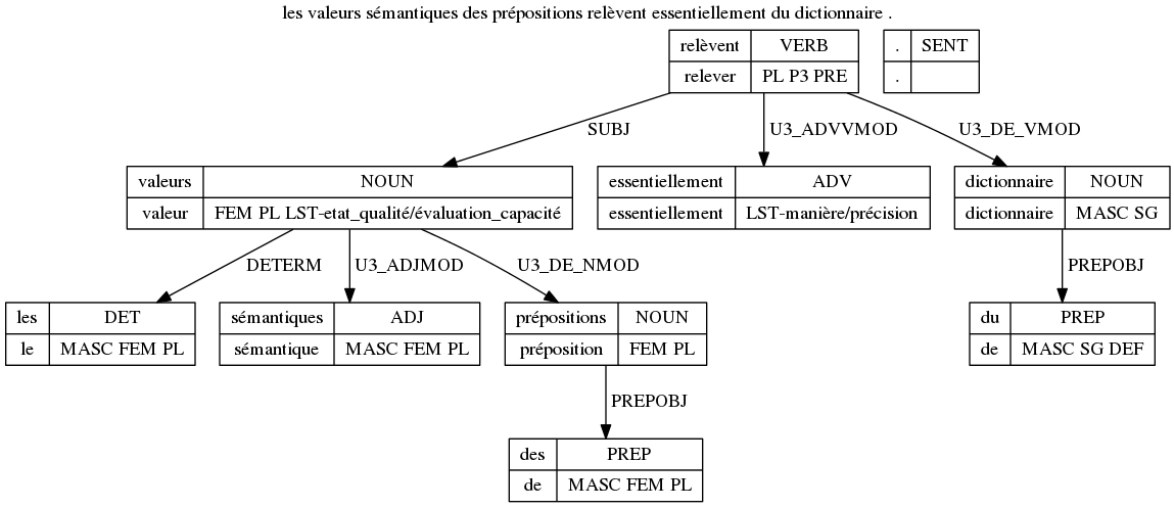


Illustration 5.2: Exemple d'analyse en dépendance pour le calcul de cadre

---

5 Vanneau, V. (2006). Maris battus. *Ethnologie française*, *36*(4), 697-703.

6 Cortier C., « Les syntagmes prépositionnels prédicatifs dans les grammaires universitaires : un observatoire de la place accordée aux prépositions », *Travaux de linguistique* 1/2001 (n°42-43), p. 121-140.

Étant donné que nous ne retenons pas les relations avec les adverbes, le verbe *relever* est, dans cet exemple, impliqué dans deux relations :

- gouverneur dans la relation sujet avec pour dépendant *valeurs* (nom du LST, dont la classe et la sous-classe sémantiques sont renseignées dans les traits) ;
- gouverneur dans la relation complément prépositionnel avec pour dépendant *dictionnaire.*

La phrase illustrée ci-dessus a alors pour cadre de sous-catégorisation :

- (SUJET) relever (DE_VMOD)

Le verbe *relever*, dans cette construction transitive indirecte prend pour complément un syntagme nominal introduit par une préposition.

Pour chaque cadre, sont indiqués les lemmes et les classes sémantiques les plus fréquents en tant qu'arguments. Le tableau suivant présente trois cadres de sous-catégorisation parmi les plus fréquents dans notre corpus, concernant le verbe *montrer,* deuxième verbe du LST le plus fréquent après *permettre*.

Cadre ***Exemple***	**Fréquence**	**Sujet** **{Classe}**	**Objet** **{Classe}**
(SUJET) montrer (COMPLETIVE_QUE) *Nos analyses montrent qu'il y a élaboration des gestes […]*	632	82 : *résultat* 54 : *analyse* 37 : *étude* 23 : *travail* 17 : *tableau* {examen} {graphique}	
(SUJET) montrer (OBJET) *Les résultats montrent un effet significatif de la présence de l'appareil […]*	118	11 : *analyse* 7 : *résultat* 5 : *exemple* 5 : *figure* {examen} {graphique}	7 : *importance* 7 : *existence* 5 : *différence* 5 : *effet* {état_qualité} {relation}
Comme le montrer (SUJET) *Comme le montre le tableau, les résultats […]*	14	2 : *analyse* 2 : *donnée* {objet} {support}	

Tableau 5.1: Exemples de cadres de sous-catégorisation du verbe montrer

Les classes sémantiques issues de notre classification nous ont permis d'ajouter un trait au niveau des actants. Ainsi, le deuxième cadre a notamment pour sujet *figure* et *tableau* (sous-classe sémantique {communication_support/graphique}), et pour objet *importance* et *rôle* (sous-classe sémantique {état_qualité/importance}). Une spécification sémantique de ce cadre serait alors :

- (SUJET {communication_support/graphique}) montrer (OBJET {état_qualité/importance})

La polysémie des éléments du LST (voir section 3.3.1.2) se perçoit aussi dans l'hétérogénéité sémantique des cooccurrents. L'approche lexico-syntaxique permet donc un traitement adapté de cette polysémie. Comme le notent Yan & Tutin (2013), les traits sémantiques permettent de représenter les restrictions de sélection des verbes selon leur acception.

Ce type de cadre sémantico-syntaxique permet par ailleurs d'envisager une étude des fonctions rhétoriques associées aux verbes présents dans l'écrit scientifique. Granger & Paquot (2009b) notent justement que les verbes du LST ont tendance à apparaître dans des structures « routinisées ». Ce constat est partagé par Teufel (1998) et Sándor (2007) qui s'intéressent à ces patrons dans l'écrit scientifique en mettant en correspondance procédés rhétoriques et réalisations linguistiques, ou par Tutin (2011) et Grossmann (2014) dans leur étude des verbes de constat dans l'écrit scientifique.

En prenant en compte les paradigmes sémantiques des arguments verbaux, nous avons pour objectif de proposer une ressource de patrons verbaux scientifiques, intégrant les récurrences observées, dont la nature diffère avec le genre d'écrits.

5.2.2.1 Regroupements des cadres et post-traitements

Nos traitements d'extraction des cadres incluent une étape de regroupement, de renommage de certaines étiquettes et de post-traitements manuels. Une telle phase est nécessaire pour éliminer le bruit, des cadres de sous-catégorisation incorrects résultant d'une « généralisation ou d'une normalisation trop importante » (Messiant *et al.*, 2010, p. 77).

L'une des difficultés majeures dans l'acquisition de cadres se situe au niveau de la distinction argument/modifieur. Nous avons opté de notre côté, que ce soit pour les cadres de sous-catégorisation ou pour les routines, pour une intégration des éléments lexicaux récurrents, qu'ils aient été annotés en tant que modifieur ou argument[7]. Nous pensons que les modifieurs récurrents sont un indice d'usage "normalisé" et doivent être enseignés. Ainsi, pour les constructions du type *dans cet article/partie/section, nous nous intéressons*, le nom tête du complément locatif est annoté comme modifieur.

Prenons pour exemple la phrase suivante :

- *Dans cet article, nous nous intéressons plus particulièrement aux changements réglementaires intervenus entre 1986 et 1996*[8].

XIP donne pour analyse, (en intégrant nos post-traitements) le résultat suivant.

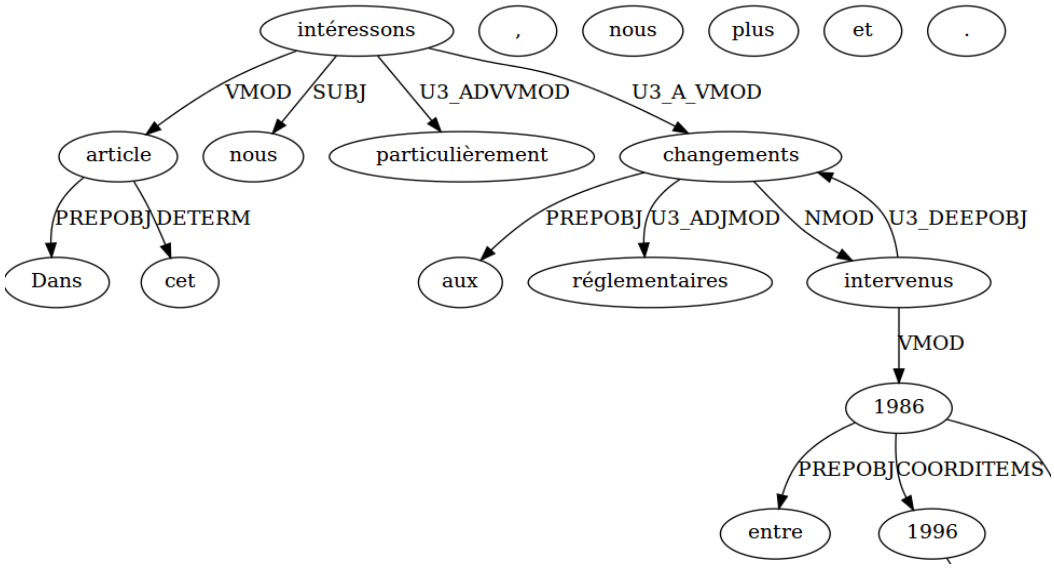


Illustration 5.3: Analyse en dépendance du locatif

Comme le montre l'illustration 5.3, la tête nominale du syntagme prépositionnel locatif, *article*, est dépendante du verbe *intéresser* dans la relation *VMOD*, ou modifieur de verbe. La relation *VMOD* peut correspondre à un

---

[7] Il est à noter que *XIP* annote les compléments prépositionnels et les compléments circonstanciels avec la même étiquette *VMOD* (Voir note 9 pour une définition de cette étiquette).

[8] Prieto, A. (2000). L'impact de la dégressivité des allocations chômage sur le taux de reprise d'emploi. *Revue économique*, 523-534.

complément indirect du verbe ou à un modifieur[9]. Cet exemple en est d'ailleurs l'illustration puisque le complément d'objet indirect *changement* a été initialement annoté comme dépendant du verbe *intéresser* dans la relation *VMOD*. L'annotation *U3_A_VMOD* résulte ici de nos post-traitements[10]. Dans la même phrase, le verbe principal est ainsi relié à un complément locatif et à un complément d'objet indirect avec la même étiquette de relation. Dans les deux cas, si le dépendant de cette relation correspond à un usage récurrent, un composant d'une routine, notre objectif est alors de l'intégrer dans la construction.

Nous avons également défini, pour l'extraction des cadres, un ensemble d'étiquettes, correspondant à une ou plusieurs relations de dépendance. Ainsi, si nous considérons la séquence suivante :

- *Enfin, nous montrons que les effets de l'assurance chômage sur le taux de chômage dépend du degré de concavité […]*[11]

Observons ensuite l'analyse donnée par *XIP* de cette phrase afin d'identifier les relations que nous retenons, modifions ou excluons.

---

[9] Ou, comme le précise la documentation de *XIP* : « This dependency attaches a modifier of a verb to the verb itself. The modifier can be an indirect complement or an adjunct of the verb. »

[10] Cette modification a notamment pour objectif d'avoir une visualisation directe de la préposition introductive dans les profils combinatoires.

[11] Lehmann, E. (1999). L'impact de l'assurance chômage et de l'assistance chômage sur le chômage d'équilibre. *Annales d'Économie et de Statistique*, 31-41.

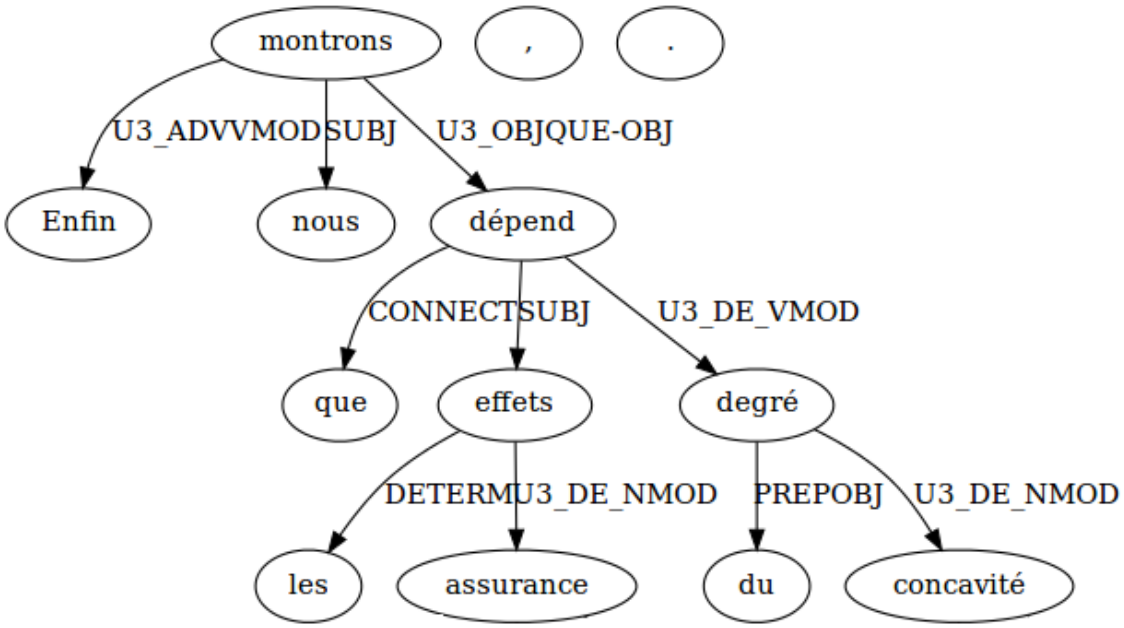


Illustration 5.4: Analyse en dépendance d'une complétive

Nous choisissons de ne pas retenir les compléments adverbiaux, identifiables par leur statut de dépendant dans la relation *U3_ADVVMOD*. L'adverbe *enfin* n'est donc pas retenu dans le cadre. Pour la relation SUBJ, nous ajoutons le trait +humain lorsque le dépendant, le sujet, est réalisé par les pronoms *je, on* ou *nous*. Enfin, pour faciliter la lecture des cadres, nous renommons la relation *U3_OBJQUE-OBJ* (ici entre *montrons* et *dépend*) par l'étiquette *COMP-QUE*. Le cadre de *montrer* serait alors ici :

- (SUJET +humain) montrer (COMP-QUE)

Nous avons opéré des reformatages similaires pour un ensemble de cas, dont les verbes modaux et semi-modaux (*pouvoir, devoir, sembler, aller, etc.*) afin d'opérer un regroupement pertinent d'occurrences sous un cadre identique. Cette phase de factorisation a ainsi pour but de limiter la dispersion d'occurrences proches au niveau de leur construction syntaxique.

Nous intégrons également des informations sémantiques concernant les cooccurrents du verbe, en listant les classes ou sous-classes sémantiques préférentiellement associées au verbe (par une relation déterminée) pour le cadre en question.

De cette première phase automatique résulte ainsi une liste des cadres de sous-catégorisation des verbes du LST, qui permet le passage à la phase de modélisation manuelle des patrons, étape essentielle pour le traitement de la polysémie.

5.2.3 Modélisation des patrons lexico-syntaxiques

Afin de modéliser les patrons verbaux, Yan est partie dans un premier temps des cadres de sous-catégorisation extraits pour repérer les acceptions verbales spécifiques à l'écrit scientifique, utilisées dans l'argumentation et la présentation de l'activité scientifique.

La phase de modélisation, présentée en détail dans une publication récente (Yan & Hatier, 2016), a consisté dans un premier temps à rattacher chaque cadre à une acception verbale déterminée, présente dans la classification du LST (voir section 3.3.2). Dans un second temps, l'observation du corpus et l'analyse des cadres de sous-catégorisation nous ont permis d'aboutir aux patrons verbaux du LST.

Considérons l'exemple du verbe *montrer,* dont les cadres les plus fréquents sont donnés dans le tableau 5.1. Le cadre le plus fréquent (*montrer* précédé d'un sujet nominal et suivi d'une complétive introduite par *que*) renvoie à deux acceptions différentes de *montrer* selon le type de sujet :

- lorsque le sujet se réalise sous forme de pronoms (*nous, il, on* et *je*), celui-ci représente essentiellement l'auteur de l'article (ou un auteur cité). Le sens mobilisé de *montrer* est alors celui de 'communiquer en faisant savoir' présent dans notre ressource. Un exemple d'occurrence de ce cadre est :
    - ***Nous montrons qu'****il est alors inutile de conserver des transferts en seconde période de vie.*[12]
- lorsque le sujet renvoie à un objet de l'activité scientifique ou à un processus, le sens mobilisé est alors celui de 'mettre en évidence'. Un exemple d'emploi de ce cadre dans notre corpus d'analyse est :

[12] Belan Pascal, « Transition vers un système par capitalisation dans un modèle de croissance endogène », *Revue économique* 6/2001 (Vol. 52), p. 1205-1226

- ***Les résultats montrent que*** *le facteur qui influence plus particulièrement la performance des enfants est l'alignement de la position finale avec un axe implicite du dispositif (vertical ou horizontal)*[13]

Yan a également constaté la présence de sujets métonymiques pour la première acception de *montrer,* par exemple dans la phrase suivante, où *études* renvoie aux auteurs de ces études :

- *Trois* ***études*** *internationales récentes* ***montrent*** *bien* ***que*** *la manière de classer la France et son école maternelle parmi d'autres systèmes d'accueil de la petite enfance repose sur ces logiques très hétérogènes*[14]

À partir de ces observations, Yan a alors modélisé les patrons lexico-syntaxiques en mettant en correspondance acception mobilisée et constructions à l'œuvre dans le corpus. L'acception *'communiquer en faisant savoir'* est ainsi rattachée au patron suivant illustré ci-dessous.

Quelqu'un (*on*, *nous*, etc.) montre que (175 occurrences)
SENS : En s'appuyant sur l'analyse et des données, quelqu'un donne des arguments pour faire savoir et justifier certains faits scientifiques en suivant un développement argumentatif.
ALTERNANCE LEXICALE : (*étude, travail, recherche, article,* etc.) *montre que*… (78 occurrences)
ALTERNANCE SYNTAXIQUE : *en montrant que*
EXPRESSIONS : *nous tentons de montrer que, nous essayons de montrer que, nous cherchons à montrer que, notre objectif est de montrer que, il s'agit de montrer que*

Illustration 5.5: Patron pour le verbe montrer

L'exemple ci-dessous illustre ce patron préférentiel avec un sujet réalisé par un pronom personnel humain. De plus, il est fréquent qu'un adverbial de lieu/de relation logique soit mobilisé pour renvoyer à un élément péri-textuel (tableau, figure, graphique), ou à une portion du texte (paragraphe, annexe, etc.) :

[13] Courbois, Y. (2005). Le développement de la « rotation mentale » : effets de la saillance de l'axe du stimulus et de son alignement avec la direction verticale. *L'année psychologique*, *105*(3), 369-386.
[14] Garnier, P. (2009). Préscolarisation ou scolarisation ? L'évolution institutionnelle et curriculaire de l'école maternelle. *Revue française de pédagogie*, 169, 5-15.

- *On* ***montre*** *toutefois, dans l'annexe 1, dans certaines configurations simples,* ***que*** *la matrice des dérivées partielles est de plein rang sous l'hypothèse nulle*[15].

En procédant à une comparaison des propriétés syntaxiques des trois verbes sémantiquement proches (*montrer, démontrer* et *indiquer*), dans le contexte scientifique, Yan a ensuite pu mettre en évidence les points communs et les spécificités entre verbes d'une même classe au niveau de ces patrons[16]. Bien qu'ils aient des propriétés sémantiques et syntaxiques comparables, Yan constate que ces verbes ne sont pas synonymes dans tous les contextes. L'analyse quantitative et qualitative des patrons[17] montre par exemple que les verbes *montrer* et *démontrer* sont employés dans le sens de 'prouver' lorsqu'ils sélectionnent un sujet humain ou métonymique. D'un autre côté, elle note que les verbes *montrer, démontrer* et *indiquer* sont synonymes, au sens de 'mettre en évidence' lorsqu'ils sélectionnent un sujet inanimé.

Ces observations montrent que l'approche guidée par le corpus permet de modéliser des patrons adaptés à des applications d'aide à la rédaction universitaire, avec pour intérêt principal un inventaire des constructions fréquemment employées, intégrant des informations sur les alternances, les groupes lexicaux en cooccurrence et les constructions spécifiques.

Cette étude sur les patrons utilisés par des scripteurs experts (chercheurs en SHS publiant dans des revues francophones à comité de lecture) nous a permis de disposer d'une base de comparaison dans un travail parallèle d'analyse des points communs et différences dans l'emploi des constructions verbales pour différents groupes de scripteurs, que nous présentons dans la partie suivante.

---

15 Jondeau, E. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. Annales d'économie et de statistique, 62, 139-174.

16 Voir (Yan & Hatier, 2016) pour le détail de l'étape de modélisation des patrons.

17 Pour une présentation détaillée de l'analyse de patron, voir (Yan & Hatier, 2016).

5.2.4 Analyse comparative des constructions verbales

Nous avons souhaité, dans des travaux en collaboration avec Yan (Hatier & Yan, à paraître), comparer les constructions verbales employées par des scripteurs experts, des scripteurs non experts natifs, et des scripteurs non experts et non natifs. Pour cela, trois différents corpus ont été exploités :

1. le corpus d'experts d'articles de recherche en SHS, utilisé tout au long de ce travail (notre corpus d'analyse du LST, présenté section 2.2.1.1) ;
2. un corpus d'écrits d'étudiants francophones natifs, de 460 000 mots, constitué de mémoires de master en spécialité didactique du français. Ce corpus est issu du corpus *Littéracie Avancée*[18] *;*
3. un corpus d'écrits d'étudiants chinois apprenants du français de 600 000 mots, constitué de mémoires de master[19] de spécialité de français à l'Université de Xi'an, dont le sujet d'étude porte sur la traduction, la littérature ou la linguistique.

Les trois corpus ont été analysés syntaxiquement avec *XIP*, ce qui a permis l'application de notre protocole d'extraction automatique des cadres de sous-catégorisation verbale.

Nous nous sommes concentrés pour cette étude sur quatre verbes du LST (*montrer, expliquer, décrire, considérer*), de fréquence élevée dans les trois corpus, et essentiels du point de vue argumentatif, car mobilisés dans le positionnement et l'analyse scientifique. Les verbes *montrer* et *considérer* sont ainsi mis à contribution pour construire la « posture réflexive » (Rinck, 2010) de scripteur (ex : *nous montrons l'importance*, *on considère que*). Les verbes *décrire* et *expliquer*, sont eux fréquemment employés dans les phases de description et d'interprétation des données.

Nous avons alors comparé la fréquence d'emploi des différentes constructions pour chacun des trois verbes, afin d'identifier les points de

---

[18] Corpus constitué par Rinck, F. Boch, F, & Jacques, M-P, disponible à l'adresse suivante : http://lidilem.u-grenoble3.fr/ressources/corpus-du-labo/article/corpus-litteracie-avancee [consulté le 11/01/2016].

[19] Mémoires de 60 pages environ, hors annexes, réunis par Rui Yan.

convergence et de divergence dans l'usage des constructions entre experts et apprenants, natifs et non-natifs.

Il nous paraît ici important de souligner le caractère non-comparable des trois corpus dans la mesure où les textes d'experts et d'apprenants ne sont pas produits dans le même contexte, et n'ont pas les mêmes buts communicatifs. Ces différences peuvent être à la source de variations, telle la présence plus importante de l'auteur dans les corpus d'experts, que nous ne devons pas faussement attribuer aux différents profils de scripteurs. Ce biais potentiel dans toute analyse contrastive de corpus est par ailleurs rappelé par Granger & Paquot (2009b).

De plus, Hyland & Milton (1997, p. 184) ont fait une mise en garde claire contre le standard irréaliste des scripteurs experts, difficilement atteignable pour les apprenants. Nous nous gardons ici de projeter sur les apprenants de telles attentes, mais proposons simplement de recenser les différents types de constructions présentes dans les trois corpus de scripteurs. Nous avons ainsi fait émerger des corpus quelques caractéristiques selon les scripteurs. Ces particularités concernent les acceptions mobilisées et les constructions employées, au niveau des ensembles lexicaux qu'elles intègrent.

Nous avons constaté des points communs entre étudiants chinois et français, par comparaison avec les experts :

- dans l'emploi des verbes *montrer* et *considérer,* les deux groupes d'étudiants ne mobilisent que très rarement un sujet de type *on* ou *nous*, à l'inverse des experts dont le positionnement est logiquement plus marqué ;
- le verbe *décrire* est peu employé au passif réduit, construction pourtant privilégiée par les experts, par les étudiants natifs et non-natifs. Nous pouvons illustrer, par l'exemple suivant, cet emploi dans le corpus d'experts :
    - *Les questions rhétoriques évaluatives rassemblent la plupart des traits des questions rhétoriques* ***décrites*** *en linguistique*.[20]

[20] Léon, J. (1997). Approche séquentielle d'un objet sémantico-pragmatique : le couple QR, questions alternatives et questions rhétoriques'. *Revue de Sémantique et de Pragmatique*, *1*, 23-50.

- la construction pronominale pour le verbe *s'expliquer* est absente chez les étudiants. Elle est cependant la deuxième la plus fréquente chez les experts, comme l'illustre l'exemple suivant :
    - *La différence **s'explique** sans doute par la proximité de ces élus avec leurs administrés, souvent plus favorables à Bonn.*[21]

Nous avons ainsi pu repérer les constructions les moins employées par les étudiants, révélatrices de certaines de leurs difficultés. Les étudiants emploient peu les alternances syntaxiques au passif et au passif pronominal, ainsi que certaines constructions spécifiques du genre scientifique, présentes chez les experts, telles que : *comme nous* + verbe, *si nous* + verbe. Ces constructions, généralement liées aux aspects textuels, au travers du guidage du lecteur ou des renvois, sont fréquentes dans les articles : *comme nous l'avons vu, comme le montre la figure*.

Les écrits d'étudiants se caractérisent également par une sur-utilisation des pronoms sujet *il* ou *cela*. Si les étudiants chinois et français se distinguent des experts sur certains points similaires, l'extraction automatique des fréquences d'emploi des constructions nous a permis d'identifier les différences d'usage entre les deux groupes d'étudiants, notamment :

- l'emploi métonymique des noms du LST de la classe {examen} est davantage présent dans le corpus d'étudiants français, comme dans l'exemple suivant :
    - *Des **travaux** en sciences de l'éducation, étudiant les pratiques enseignantes [...] **décrivent**, de façon très précise, le comportement de l'enseignant*[22] *;*
- la construction la plus employée par les étudiants chinois avec le verbe *montrer* intègre un sujet et un objet inanimés. Ceci nous a permis de mettre en évidence une différence de sens mobilisé. En effet, si les étudiants français emploient ce verbe au sens de 'prouver, indiquer', la majorité des

[21] Laporte A., Djament-Tran G., « Comment Berlin devint capitale de l'Allemagne réunifiée. Éléments pour l'analyse d'un événement territorial », *L'Espace géographique* 2/2010 (Vol. 39), p. 146-158

[22] Maurice, J. J., & Murillo, A. (2008). La distance à la performance attendue : un indicateur des choix de l'enseignant en fonction du potentiel de chaque élève. *Revue française de pédagogie. Recherches en éducation*, (162), 67-79.

constructions dans le corpus d'étudiants chinois activent le sens 'faire voir', ainsi que l'illustre l'exemple suivant :

- *Le tableau en bas peut* ***montrer*** *les diplômes différents dans les étapes différentes.* [corpus d'étudiants chinois] ;

- la construction *si l'on considère*, la plus fréquente dans le corpus d'experts, n'est pas utilisée par les étudiants chinois.

Les étudiants, français et chinois, ont des difficultés à maîtriser certaines constructions (ou acceptions) typiques du genre étudié, en particulier celles qui permettent de développer l'argumentation et d'organiser le texte. Ceci peut alors se traduire par des difficultés à établir une structure discursive cohérente, construire des opinions au sujet des savoirs d'une discipline. Ces difficultés sont encore plus grandes pour les étudiants chinois, du fait de leur statut d'apprenant du français, et entraînent des confusions de registre, des erreurs de collocations et la production d'expressions et de tournures maladroites.

Ce constat justifie ainsi l'intérêt d'une ressource de patrons verbaux adaptée, intégrant les spécificités des constructions que nous avons identifiées. La dimension phraséologique des écrits scientifiques est par ailleurs au centre d'autres travaux, mais ceux-ci ne proposent pas une description fine des contraintes syntactico-sémantiques des constructions verbales. Ainsi, le projet ARTES[23] (Kübler & Pecman, 2012) propose notamment une liste de collocations, dans plusieurs langues, pour des fonctions discursives déterminées. Par exemple, à la fonction 'exprimer le présent', sont associées les expressions suivantes : *jusqu'à présent, à ce jour, être à l'étape de, la tendance actuelle est.* La ressource permet ainsi une appropriation de formules préfabriquées mais n'est pas organisée autour du concept de patron. Notre collaboration avec Yan vise ainsi à proposer une ressource adaptée en se basant sur l'observation préalable des difficultés particulières que les étudiants rencontrent dans la production/compréhension d'écrits scientifiques.

[23] https://artes.eila.univ-paris-diderot.fr/ [consulté le 15/09/2016]

5.2.5 Perspectives

Ce travail d'analyse des cadres de sous-catégorisation des verbes du LST nous permet d'envisager une analyse des constructions pour les noms déverbaux du LST qui correspondent aux verbes analysés. Nous pourrions alors vérifier l'observation faite par Condette *et al.* (2012, p. 857) selon lesquels « la structure argumentale, dans la majorité des cas, est héritée habituellement de la structure argumentale du verbe correspondant ». Cette analyse comparative peut également intégrer des contraintes sémantiques, en confrontant les classes sémantiques en cooccurrence entre constructions nominale et verbale. Nous avons ainsi pu observer que le nom *définition* entre fréquemment en cooccurrence avec *préciser* et *identité* lorsque le verbe *définir* entre en cooccurrence lui avec *précisément* et *identité*. De même, les profils combinatoires de *élaboration* et *élaborer* ont notamment en commun *stratégie* et *projet*.

Condette *et al.* pointent cependant une difficulté dans ces correspondances nom/verbe, due au fait que l'agent habituel des verbes est absent de la réalisation en surface pour les déverbaux. Ainsi, les cooccurrents du nom *élaboration* ne renvoient que très rarement à l'agent de ce processus. Nous notons que ce problème se rencontre également pour les alternances de type passif réduit pour les verbes, comme pour le verbe *observer* dans l'exemple suivant :

- *En effet, comme l'ont souligné notamment CAMPBELL et SHILLER [1991], la pente **observée** et la pente théorique [...] présentent souvent des évolutions très proches*[24]

Un travail d'analyse comparative des constructions des verbes et déverbaux doit ainsi tenir compte de ces cadres dans lesquels l'agent n'est pas réalisé explicitement. Condette *et al.* notent ainsi que les arguments sont beaucoup plus optionnels que pour les verbes. Les variations correspondent ainsi généralement à un changement de préposition, une perte d'argument, mais jamais d'ajout argumental.

En termes de perspective d'amélioration des traitements d'extraction des cadres, nous projetons d'intégrer des informations morpho-syntaxiques sur les

[24] Jondeau, E. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. Annales d'économie et de statistique, 62, 139-174.

réalisations les plus fréquentes, par exemple pour certaines constructions qui prennent pour sujet un syntagme nominal déterminé préférentiellement par un démonstratif, ou se réalisent avec un verbe conjugué à un certain temps/mode. Par exemple, les constructions avec le verbe *aborder, s'intéresser* prennent fréquemment pour sujet des noms du type *cet article, cette étude, ces travaux*.

Nous avons présenté dans cette partie une étude des constructions verbales guidée par un objectif didactique d'aide à la rédaction. Le travail détaillé dans la section suivante est lui sous-tendu par un objectif de description linguistique dont les observations peuvent aider à l'amélioration de l'extraction terminologique.

5.3 Étude des cooccurrences entre LST et terminologie

Comme nous l'avons indiqué dans l'introduction de ce chapitre, notre intérêt pour les patrons lexico-syntaxiques du LST n'a pas pour unique objectif la constitution d'une ressource didactique. En effet, dans le cadre du projet TermITH, nous avons également souhaité évaluer l'apport potentiel de patrons lexico-syntaxiques du LST dans l'extraction terminologique. Autrement dit, nous avons pour objectif l'identification de configurations syntaxiques récurrentes combinant la présence d'au moins un élément du LST et d'un terme. Nous partons de l'observation selon laquelle le LST est un indice fort de la présence de terme, tel que l'ont montré Jacquey *et al.* (2013). Notre hypothèse est qu'une caractérisation sémantico-syntaxique des contextes de cooccurrences entre LST et terminologie peut être un outil efficace dans l'indexation terminologique. Nous souhaitons ainsi mesurer l'apport de patrons définis par des relations de dépendance et des traits sémantiques (LST, classes et sous-classes), et, ce faisant, l'intérêt de la classification sémantique du LST pour la validation du statut terminologique de certaines unités lexicales.

Nous présentons dans cette partie une expérimentation sur l'étude des cooccurrences entre élément terminologique et élément du LST. Ce travail a été effectué en collaboration avec Patrick Drouin [25], dont le logiciel *Termostat*[26] (Drouin, 2003) permet d'extraire des termes à partir d'un corpus de textes.

[25] Nous remercions Patrick Drouin pour son accueil lors de notre séjour de recherche à l'université de Montréal

[26] Disponible à cette adresse : http://termostat.ling.umontreal.ca/index.php [consulté le 05/07/2016]

Termostat relève, pour l'acquisition des termes, « les spécificités positives nominales et adjectivales » (Drouin, 2004, p. 348) dans le corpus d'analyse, en le comparant à un corpus global hétérogène. Dans cette étude, nous souhaitons observer si l'intégration de contraintes lexico-syntaxiques dans le repérage des termes permet d'améliorer cette identification.

Pour ce faire, nous mettons en place un protocole en deux temps. Nous partons d'un corpus d'articles dont les termes ont été annotés manuellement[27] dans le cadre du projet TermITH. Nous repérons dans ce corpus des patrons lexico-syntaxiques récurrents, comprenant a minima un terme et un élément du LST. À la différence des patrons sur le modèle *CPA* de la partie précédente, les patrons dont il est ici question n'intègrent pas forcément un verbe et s'ils sont de type verbal, ne comprennent pas obligatoirement la totalité de la structure argumentale du verbe. Un patron est alors défini, pour cette étude des cooccurrences LST-terme, par la présence d'au moins deux unités lexicales, contraintes sémantiquement (LST et terminologie) et syntaxiquement (ces deux unités appartiennent au même sous-arbre, résultat de l'analyse de *XIP*). Par exemple, le patron – $\text{Nom}_{\text{qualitifant}}$ Prep $\text{Nom}_{\text{Terme}}$ – correspond aux cooccurrences entre un terme et un élément du LST de la sous-classe sémantique {déterminant/qualifiant} dans lesquelles le nom du LST (*genre, type*) est gouverneur du terme dans une relation prépositionnelle. Ce patron correspond alors aux exemples repérés dans le corpus tels *type d'espace/de migration, genre de profession/de parure.*

Dans un deuxième temps, nous utilisons le logiciel *Termostat* pour extraire d'un second corpus[28] (dont les disciplines couvertes sont différentes de celles du premier corpus) les candidats-termes. Nous projetons alors les patrons précédemment identifiés et comptabilisons le nombre de fois où ils intègrent effectivement un candidat-terme. Les patrons sont linéarisés, i.e représentés comme une suite de tokens, afin de pouvoir les intégrer dans le logiciel, qui ne prend pas en charge les corpus arborés mais se base sur les sorties de TreeTagger. Nous souhaitons ainsi vérifier si les patrons relevés du premier corpus (annoté manuellement) permettent de valider des candidats-termes dans le second corpus,

---

[27] Nous remercions Éveline Jacquey et les partenaires du projet TermITH pour le partage de ces corpus, différent du corpus utilisé pour l'extraction du LST et présenté section 2.2.1.1.

[28] Nous remercions Patrick Drouin pour la mise à disposition de ces corpus.

ou du moins de servir de pondération positive pour le statut terminologique des candidats.

5.3.1 Identification des cooccurrences terme-LST

Pour identifier les cooccurrences entre termes et éléments du LST, nous constituons un corpus en regroupant des textes annotés manuellement en termes, textes distincts du corpus d'analyse utilisé jusqu'à présent. Suite à cette annotation, effectuée pour des expérimentations dans le cadre du projet TermITH, nous analysons le corpus avec *XIP* et projetons les traits sémantiques de classes et sous-classes sémantiques issus de notre classification du LST. Ce corpus d'articles porte sur trois disciplines des SHS, dont la composition est détaillée dans le tableau suivant, et comporte plus de 600 000 mots.

Discipline (nombre d'articles)	**Nombre de mots**
Archéologie (11)	78 799
Linguistique (75)	392 194
Psychologie (19)	148 139
Total (105)	619 132

Tableau 5.2: Corpus pour l'extraction des patrons de cooccurrence LST-terme

Pour l'extraction des patrons, nous avons pris pour amorce la co-présence d'un terme et d'un élément du LST dans une même phrase. Le patron correspond alors au sous-graphe minimum comprenant cette cooccurrence. Considérons par exemple la phrase suivante :

- *[34] Voir « Les conditions institutionnelles de la scolarisation secondaire des garçons entre 1920 et 1940 ».*[29]

Le résultat de l'analyse en dépendance de cette phrase est présenté ci-dessous.

[29] Chapoulie Jean-Michel, « Une révolution dans l'école sous la Quatrième République ? La scolarisation post-obligatoire, le Plan et les finalités de l'école », *Revue d'histoire moderne et contemporaine* 4/2007 (n° 54-4), p. 7-38

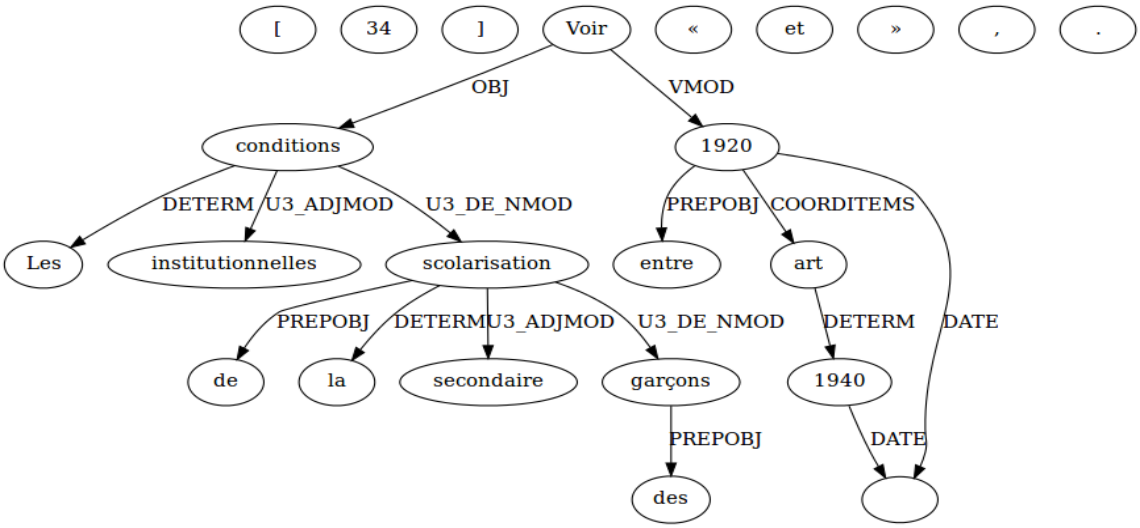


Illustration 5.6: Analyse en dépendance – Cooccurrence LST-terme

Le nom *condition* étant un élément du LST (de la sous-classe {objet/cas}, si le nom *scolarisation* est annoté en tant que terme, alors quatre patrons LST-terme correspondent à cette phrase :

- au niveau du lemme : *condition* Prep Nom_{Terme}
- au niveau de la sous-classe : Nom_{cas} Prep Nom_{Terme}
- au niveau de la classe sémantique : Nom_{objet} Prep Nom_{Terme}
- au niveau du lexique : Nom_{LST} Prep Nom_{Terme}

Nous dégageons automatiquement des patrons sur différents niveaux et prenons en compte les traits sémantiques du LST pour identifier les classes le plus fréquemment en cooccurrence avec un terme. Nous nous intéressons alors aux patrons de cooccurrence LST-Terme les plus fréquents dans ce corpus.

Nous avons par exemple repéré le patron – Nom_{cas} Prep Nom_{Terme} – réalisé dans les exemples suivants :

- *Dans une <u>condition</u>$_{cas}$ de <u>perception</u>$_{Terme}$ optimale, "habituelle", l'analyse des données obtenues ne permet pas de mettre en évidence de différence significative*[30].

[30] Giraudet, G., & Roumes, C. (2004). La signature spatiale de l'objet: une information essentielle pour la localisation de cibles dans une scène naturelle. *L'année psychologique*, *104*(1), 9-49.

- *L'objectif était d'apporter des éléments de connaissance des pratiques d'information de ces jeunes en dehors de* <u>*situations*</u>$_{cas}$ *d'*<u>*apprentissage*</u>$_{Terme}$ *déclarées comme telles*[31].

Nous avons par ailleurs identifié plusieurs sous-classes et classes sémantiques productives dans ce même patron : Nom_{LST} Prep Nom_{Terme} :

- les noms de la sous-classe {relation/implication} (*base, condition, conséquence, etc.*), comme dans les cooccurrences suivantes, extraites du corpus : *conditions de travail, effet de terreur, contrainte d'incitation*
- les noms de la classe {état_qualité} : *configuration du marché, caractéristiques des élèves, composition des écoles, capacité de production, valeur du client*

Deux patrons peuvent alors être définis selon ces contraintes sémantiques, et énoncés de la façon suivante :

- $Nom_{implication}$ Prep Nom_{Terme}
- $Nom_{état\_qualité}$ Prep Nom_{Terme}

Nous avons également identifié un patron intégrant fréquemment un terme, dans lequel un verbe de la sous-classe {état/inclusion} prend pour objet un nom de la sous-classe {état_qualité/caractéristique} et pour sujet un terme. Ce patron – (SUJET Nom_{Terme}) $Verbe_{inclusion}$ (OBJ $Nom_{état\_qualité}$) – est ainsi réalisé dans les exemples suivants :

- *Le* <u>*langage*</u>$_{Terme}$ <u>*possède*</u>$_{inclusion}$ *à la fois une* <u>*forme*</u>$_{état\_qualité}$ *subjective et une forme objective*[32] *;*
- *Les* <u>*industries*</u>$_{Terme}$ *de réseaux* <u>*présentent*</u>$_{inclusion}$ *des* <u>*caractéristiques*</u>$_{état\_qualité}$ *bien particulières*[33] *;*

---

[31] Aillerie, K. (2012). Pratiques juvéniles d'information: de l'incertitude à la sérendipité. *Documentaliste-Sciences de l'Information*, *49*(1), 62-69.

[32] Galinier, J. (2006). L'anthropologie hors des limites de la simple raison. L'*Homme*, (3), 141-164.

[33] Flochel, L. (1999). Interconnexion de réseaux et charge d'accès : une analyse stratégique. *Annales d'Economie et de Statistique*, 171-196.

- *les* <u>*scènes*</u>$_{\text{Terme}}$ *de l'avant-garde* <u>*présentent*</u>$_{\text{inclusion}}$ *des* <u>*profils*</u>$_{\text{état\_qualité}}$ *diversifiés*[34] *;*

Nous notons dans les exemples de ce patron, tels ceux présentés ci-dessus, une forte proportion de noms de la sous-classe {état_qualité/caractéristique} (*forme, profil, caractéristique, nature…*). Le patron pourrait alors être affiné de la manière suivante :

- (SUJET Nom$_{\text{Terme}}$) Verbe$_{\text{inclusion}}$ (OBJET Nom$_{\text{caractéristique}}$)

Nous avons ainsi élaboré une liste de 11 patrons sur plusieurs niveaux (sous-classe, classe, ou LST), en retenant les plus fréquents dans ce corpus annoté en termes. L'annexe A.X détaille l'ensemble des patrons en donnant pour chacun deux exemples de réalisation (extraits du corpus d'analyse de SHS présenté section 2.2.1.1).

Il nous faut souligner qu'un des problèmes récurrents dans la sélection de ces patrons est le fait que certains patrons LST-Terme correspondent en fait à un terme complexe dont la tête est l'élément du LST. Il nous faut alors ne pas considérer cet élément en tant que membre du LST mais comme composant d'un terme complexe. Tel est fréquemment le cas pour les sous-classes nominales {objet/méthode}, {état_qualité/composition} et {collectif/partitif}, avec par exemple les noms suivants : *relation (associative, asymétrique, client, conjugale), structure (centralisée, de la personnalité, des besoins), analyse (aveugle, bayésienne, contextuelle, transactionnelle), système (d'éducation, de gestion, d'exploitation, de gestion de base de données, informatique).* Dans le cadre de l'indexation terminologique, le LST ne doit alors pas être utilisé comme filtre d'exclusion pour les candidats complexes de ce type pour ne pas sur-segmenter les candidats-termes.

5.3.2 Patrons LST-Terme et validation terminologique

Suite à l'identification de ces patrons combinant termes et éléments du LST, nous souhaitons observer sur un corpus test l'apport de ces configurations. Plus précisément, nous souhaitons vérifier l'hypothèse selon laquelle plus un élément défini comme potentiellement terminologique par *Termostat* entre dans ces

[34] Lizé, W. (2010). Le goût jazzistique en son champ. *Actes de la recherche en sciences sociales*, (1), 60-87.

configurations lexico-syntaxiques, plus son statut terminologique est probable. En effet, *Termostat* offre en sortie une liste de mots potentiellement terminologiques, dont le statut doit être validé par un humain. Les mots en tête de liste, dont la taille se compte en milliers de mots, sont alors ceux dont le statut terminologique est le plus probable statistiquement. Notre hypothèse est qu'une pondération de ce statut, corrélée aux nombres d'occurrences dans des patrons préalablement identifiés, permettrait d'améliorer les candidats en tête de liste.

Nous avons lancé *Termostat* sur un second corpus, composé de deux sous-corpus disciplinaires. Nous disposons d'une part d'un ensemble de 50 articles de sociologie (issu du projet Scientext) et d'autre part d'un ensemble de thèses en informatique (mis à notre disposition par l'OLST). Ces deux sous-corpus ne sont pas tout à fait comparables en termes de type d'écrits (articles *vs* thèses) et d'ensemble disciplinaire (SHS *vs* sciences dures). Nous avons ainsi souhaité observer la généricité des patrons au travers des disciplines et des types de textes.

Sociologie (articles)	608 088 mots
Informatique (thèse)	484 969 mots
Total	1 093 057 mots

Tableau 5.3: Corpus d'entrée pour l'extraction de termes avec Termostat

Afin d'évaluer l'apport des patrons pour la recherche de termes, nous avons projeté les patrons retenus sur ce corpus de test en vérifiant si le candidat-terme (annoté automatiquement par *Termostat*) est effectivement terminologique. Nous avons pour cela utilisé plusieurs ressources de références : le Grand Dictionnaire Terminologique pour les deux sous-corpus, le *DicoInfo* (L'Homme, 2008) pour le sous-corpus d'informatique, ainsi qu'une ressource terminologique en sociologie développée à l'INIST[35]. Nous avons mené en cela une étude qualitative sur les résultats étant donné que le corpus de test n'est pas annoté manuellement et que toute occurrence d'un candidat-terme n'est pas automatiquement terminologique. Nous avons examiné les concordances correspondant aux patrons retenus et avons ainsi pu constater que si certains patrons intègrent fréquemment un terme (validé manuellement), aucun patron fréquent ne permet de valider de façon certaine le statut terminologique des candidats.

[35] Nous remercions nos partenaires de l'INIST pour le partage de cette ressource.

L'application des patrons LST-Terme, de par la haute fréquence de ces deux lexiques, a donné lieu à l'identification de concordances pour lesquelles les patrons s'avèrent utiles dans la tâche de validation terminologique. Cependant, ces patrons génériques sont également réalisés dans des cas où le candidat-terme n'est pas terminologique. De manière générale, les premiers résultats ne permettent pas de définir la granularité idéale pour ces patrons : au niveau des lemmes, des sous-classes, ou des classes. Nous avons pu observer que le simple niveau du lexique (étiquette 'LST') est trop large et génère des patrons effectivement fréquents, mais dont la présence n'est pas une indication fiable de la présence d'un terme (le candidat-terme est également souvent un élément du LST). À l'autre extrémité, les niveaux du lemme et des sous-classes entraînent beaucoup de dispersion dans les patrons, et rendent complexe le repérage des patrons opératoires pour l'indexation en termes.

Nous avons également pu constater une forte diminution de couverture des patrons, sur le corpus test. Ceci s'explique par le fait d'avoir linéarisé les patrons pour les appliquer sur un corpus non analysé syntaxiquement. L'utilisation des patrons et de la ressource du LST a néanmoins permis de mettre en évidence certaines configurations pouvant être exploitées dans l'indexation terminologique, et d'observer différents types d'interactions entre deux lexiques constitutifs de l'écrit scientifique, le LST et la terminologie.

Cette expérimentation nous ouvre ainsi plusieurs perspectives, que nous détaillerons en conclusion, notamment sur l'utilisation d'un corpus test arboré.

5.4 Routines dans l'écrit scientifique

Nous nous sommes attachés, dans la partie précédente, à identifier automatiquement des configurations sémantico-syntaxiques, ensemble défini par des relations syntaxiques entre des mots typés par leur appartenance à un certain lexique : LST et terminologie.

Dans cette section, nous tentons de faire émerger automatiquement un autre type de configurations, les routines sémantico-rhétoriques, « qui correspondent à des fonctions rhétoriques spécifiques de l'écrit scientifique » tel que le définissent Tutin & Kraif (2016, p. 120).

Nous partons de l'hypothèse que la classification sémantique peut être utilisée comme amorce pour l'identification de routines récurrentes dans l'écrit scientifique. Nous voulons également évaluer l'adéquation de la granularité de la classification à la tâche de détection de routines, porteuses de fonctions rhétoriques typiques de l'écrit scientifique telles que le positionnement et la définition de la problématique.

Gledhill (1994) insiste d'ailleurs sur l'importance du genre dans la phraséologie employée, relevant ainsi l'intérêt de l'étude des routines dans l'analyse du genre de l'écrit scientifique en SHS. Dans son étude de la phraséologie de la LSG, Pecman (2004b, p. 428) aborde ces expressions en mettant en évidence certains patrons, tel le suivant, définissant la problématique du travail présenté :

- « the goal of [the/this/our/present] [study/research/paper/contribution] is to… »

Pecman (2004b, p. 428) conclut en notant que le genre scientifique se distingue par un emploi prépondérant de ces collocations restreintes. Nous pouvons ici faire le lien entre les paradigmes entre crochets et les classes sémantiques de notre ressource du LST. Ce type de patron relevé par Pecman est semblable aux routines que nous souhaitons identifier grâce aux résultats de l'analyse en dépendance et à la classification sémantique du LST.

Nous pouvons, pour illustrer la notion de routine étudiée ici, partir de la classe sémantique (présentée chapitre 3) du LST {communication_support}, tels *article, figure, tableau*. En observant les profils combinatoires de ces noms, nous avons pu constater qu'ils sont fréquemment sujets des verbes de la sous-classe {analyse/description} tels *présenter, décrire, illustrer*.

Une ébauche de patron, correspondant alors à une fonction de présentation du sujet abordé, peut ainsi être avancée :

- (SUJET $\text{Nom}_{\text{support}}$) $\text{Verbe}_{\text{description}}$ (OBJET 'thème, sujet du support')

Ce patron est notamment réalisé dans les trois exemples suivants :

- *Les tableaux*$_{\text{support}}$ *1 et 2 ci-dessous décrivent*$_{\text{description}}$ *l'interaction de ces composantes dans la définition de la pertinence.*[36]
- *L'annexe*$_{\textit{support}}$ *2 présente*$_{\textit{description}}$ *l'approche utilisée pour estimer par simulation la distribution de la statistique du test formel et des statistiques de test affaiblies.*[37]
- *La figure*$_{\text{support}}$ *5 illustre*$_{\text{description}}$ *le pourcentage de préférence accordée à chaque format pour l'ensemble des participants.*[38]

Les travaux sur les routines dans l'écrit scientifique montrent que leur étude est essentielle dans l'analyse de genre. Ainsi, selon Grossmann (2010), le travail de description du genre académique se doit de prendre en compte les routines phraséologiques et les combinaisons libres des mots du discours.

Nous nous attachons à une telle description, en précisant dans la section suivante la nature des routines que nous voulons identifier.

5.4.1 Définition des routines sémantico-syntaxiques

La notion de routine s'entend dans le cadre d'une « conception étendue de la phraséologie » (Legallois & Tutin, 2013), où les objets d'étude ne sont pas seulement les expressions non compositionnelles. Les routines qui nous intéressent se situent ainsi, par rapport aux expressions figées, à l'autre extrémité du continuum de figement, largement utilisé dans les travaux sur la phraséologie. Plus globalement, les routines sont sous-tendues par le principe phraséologique de la langue[39], dans lequel des unités pré-construites sont étudiées à l'interface du lexique et de la syntaxe. Nous avons d'ailleurs vu que les dimensions lexicale et syntaxique sont essentielles dans l'analyse des langues de spécialité (Lerat, 1997)

---

[36] Malrieu, D. (1997). Reconnaissance de catégories discursives et repérages énonciatifs : l'exemple des hypothèses dans les textes scientifiques. *Revue de Sémantique et Pragmatique*, n° 1, pp. 147-163.

[37] Jondeau, E. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. Annales d'économie et de statistique, 62, 139-174.

[38] Paire-Ficout, L., Saby, L., Alauzet, A., & Boucheix, J. M. (2013). Quel format visuel adopter pour informer les sourds et malentendants dans les transports collectifs ?. *Le travail humain*, *76*(1), 57-78.

[39] ou *idiom principle* de Sinclair (1991, p. 110).

La prise en compte de ces dimensions dans l'étude de routines à fonction rhétorique est ainsi primordiale, ainsi que l'observe Gledhill (1994, p. 4) qui remarque que les « unités pré-construites, utilisées dans un contexte rhétorique donné sont comparables au principe de la lexico-grammaire [postulant] qu'il n'y a pas de schéma grammatical sans paradigme de lexèmes associés à cette structure. » Les fonctions rhétoriques de ces unités sont ainsi associées à un schéma (analyse en dépendance) et un paradigme de lexèmes (dont le LST).

Pour cette expérimentation, nous nous concentrons sur les routines intégrant un ou plusieurs éléments du LST (des quatre différentes catégories représentées). À l'instar de Tutin & Kraif (2016), nous délimitons les routines qui nous intéressent à l'aide de quatre critères principaux :

- les routines sont organisées autour d'un verbe, racine du sous-arbre constituant la routine ;
- les routines intègrent une ou plusieurs unités lexicales appartenant au LST ;
- les routines sont associées à une fonction rhétorique spécifique : introduire un nouveau concept, se positionner vis-à-vis d'un pair, définir la thématique d'une section, etc. ;
- les routines sont transdisciplinaires et se réalisent dans un minimum de disciplines différentes. Pour les besoins de l'expérimentation, nous avons empiriquement fixé ce seuil à trois.

Les routines sont ainsi des structures linguistiques, non-linéaires, contraintes syntaxiquement, sémantiquement et ayant une fonction rhétorique déterminée.

Nous devons ajouter à cette définition le fait que les routines sont étudiées à l'aide de la liste des mots simples du LST. Nous amorçons ainsi la recherche de routines à l'aide d'éléments de notre ressource du LST, extrait lors de la phase présentée chapitre 2. En conséquence, le silence généré lors de cette étape se répercute dans la recherche de routine. Par exemple, les verbes *voir, penser* et *comprendre* n'ont pas été extraits, car ils ne sont pas spécifiques au corpus d'analyse (*i.e* non surreprésentés par rapport au corpus de contraste) en tant qu'unités lexicales. Ce silence, au niveau des mots simples, se répercute ainsi au niveau phraséologique étant donné que nous ne pouvons extraire ces routines sans amorce lexicale.

Le critère de spécificité, mis en œuvre pour les mots simples (voir section 2.2.2.1.1), n'est pas ici appliqué aux routines. Notre approche n'a pas pour but de faire émerger l'ensemble de la phraséologie transdisciplinaire mais d'en identifier une partie, laquelle est définie par la présence d'un verbe du LST et de un ou plusieurs autres éléments de ce lexique. L'objet d'étude à caractériser reste ainsi le LST en tant qu'ensemble de mots simples, lesquels interviennent ou non dans des routines transdisciplinaires.

5.4.2 Extraction des routines via le Lexicoscope

Ayant à disposition un corpus analysé en dépendance, nous pouvons tirer parti d'une fonctionnalité particulière du *Lexicoscope* : l'extraction d'expressions polylexicales récurrentes, ou plus exactement d'arbres lexico-syntaxiques récurrents (ALR). Les ALR sont, selon Tutin & Kraif (2016, p. 7), des « sous-arbres récurrents, dont la fréquence est suffisamment élevée pour traduire un degré d'association significatif sur le plan statistique ».

La méthode mise en place par Kraif prend pour amorce un pivot, défini par un mot ou un arbre. Les résultats du calcul des cooccurrents statistiquement significatifs (visualisables par les lexicogrammes) sont intégrés au pivot pour aboutir à un arbre étendu. L'opération est réitérée tant que les arbres peuvent être étendus et que les nouveaux cooccurrents sont statistiquement significatifs. Les expressions résultant de ce traitement sont donc identifiées par la fréquence significative[40] de configurations lexico-syntaxiques définies par un ensemble d'items (Quiniou, Cellier, Charnois, & Legallois, 2012) ainsi qu'un ensemble de relations de dépendance entre ces items. L'attribut définissant un item peut être son lemme, sa catégorie morpho-syntaxique ou son appartenance à un ensemble lexical préalablement défini dans les paramètres du *Lexicoscope*. Considérons l'illustration suivante représentant une phrase et son analyse par *XIP*.

[40] Le *Lexicoscope* permet de choisir la mesure utilisée : t-score, z-score ou rapport de vraisemblance. Nous avons opté pour ce dernier avec un seuil de spécificité à 10,83.

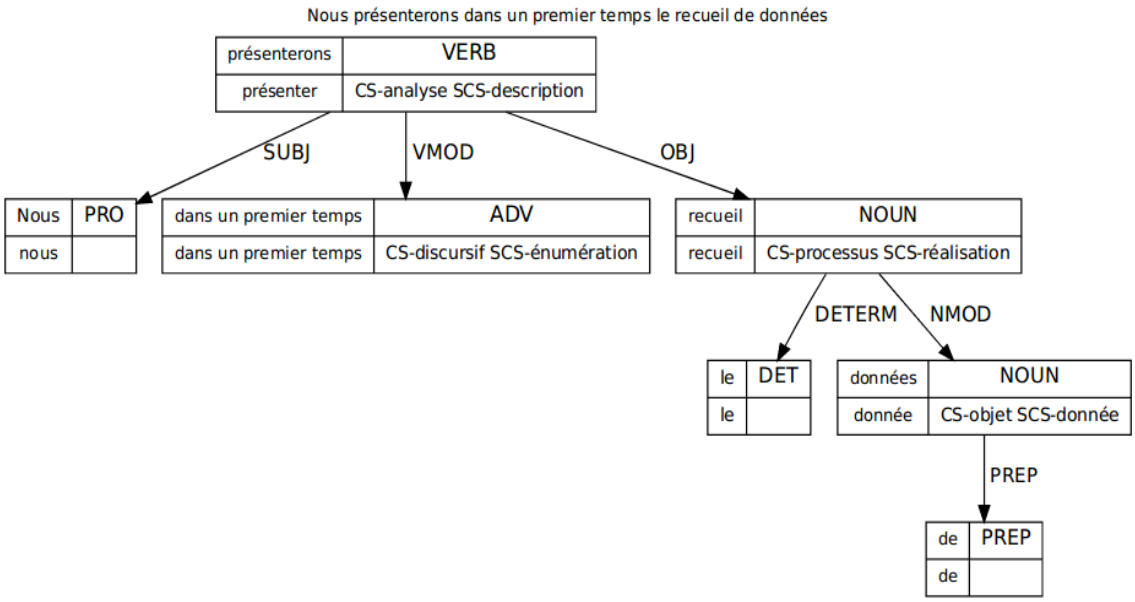


Illustration 5.7: Analyse en dépendance et traits sémantiques

Nous pouvons observer que chaque item est défini, sur cette illustration, par sa forme, son lemme, sa catégorie morpho-syntaxique ainsi que les traits de classe et sous-classe sémantiques pour les mots du LST. Chacun de ces attributs peut être utilisé pour définir les éléments de la routine. Ainsi, si nous représentons chaque item par son trait de sous-classe, nous obtenons le patron suivant :

- (SUJET) $\text{Verbe}_{\text{description}}$ $\text{Adverbe}_{\text{énumération}}$ (OBJET $\text{Nom}_{\text{réalisation}}$ Prep $\text{Nom}_{\text{donnée}}$)

Nous avons choisi d'intégrer des contraintes sémantiques, représentées par les classes et sous-classes sémantiques du LST, dans ces expressions polylexicales afin de faire émerger des routines syntactico-sémantiques. À l'instar de Hyland (1999) relevant l'importance des verbes de communication (*reporting verbs*) dans le positionnement, il est alors possible de lier classes sémantiques et routines. La classification du LST est ainsi adaptée à l'identification automatique de ce type de patrons, comme le prouvent les précédents travaux sur certaines classes de verbe (verbes de constat (Grossmann, 2014 ; Tutin & Kraif, 2016) ou d'opinion (Yan, 2014)).

Afin d'identifier les routines dans le corpus d'articles de recherche, nous nous concentrons sur neuf verbes de la classe sémantique {analyse}, appartenant à trois sous-classes différentes :

- {analyse/#examen} : *examiner, analyser, étudier*

- {analyse/#évaluation_qualitative} : *estimer, évaluer, mesurer*
- {analyse/#description} *: décrire, présenter, illustrer*

Nous sélectionnons ces verbes en raison de leur haute fréquence et de leur importance dans la présentation de l'activité scientifique. Nous définissons, à l'aide des paramètres du *Lexicoscope*, des classes lexicales, afin d'intégrer les traits sémantiques du LST dans les arbres lexico-syntaxiques extraits. Nous devons ici rappeler que ces traits sont non désambiguïsés. Ainsi, chaque élément du LST renvoyant à plus d'une acception transdisciplinaire se voit attribuer les traits sémantiques correspondant à l'ensemble de ces acceptions. Par exemple, les noms *résultat* et *différence* ont chacun une acception dans la classe {relation} et la classe {quantité}.

Nous transposons la classification sémantique des noms du LST dans le formalisme des classes lexicales du *Lexicoscope* (voir section A.XI dans l'annexe). Les autres catégories du LST (adjectifs, verbes et adverbes) ayant également été organisées en classes et sous-classes sémantiques, de futurs travaux pourront bénéficier d'un étiquetage transcatégoriel pour l'extraction d'expressions polylexicales, comme nous l'abordons dans la section 5.4.3.

Nous procédons en deux temps pour l'extraction des routines. Nous commençons par extraire les arbres lexico-syntaxiques récurrents en définissant comme pivot la classe {analyse}, représentée par les neufs verbes précédemment détaillés. Dans un second temps, nous relançons l'extraction en nous situant au niveau des sous-classes verbales afin d'observer l'effet de la granularité sur les routines capturées.

5.4.2.1 Pivot de la classe {analyse}

Nous nous intéressons ainsi en premier lieu aux routines communes à différentes sous-classes de la classe {analyse}.

La fonctionnalité d'extraction des expressions polylexicales nous permet d'extraire les arbres lexico-syntaxique récurrents (ALR) pour les éléments de la classe {analyse}. La présentation des ALR est organisée dans un premier temps selon le type de dépendance (concernant le pivot) et dans un second temps selon le collocatif, ce qui permet de naviguer dans les différents ALR, de les étendre et d'en

Illustration 5.8: Extraction des ALR pour la classe {analyse}

Les éléments dont le premier caractère est '$' sont les classes lexicales définies par l'utilisateur. : NLSTREL_NOUN équivaut à la classe nominale {relation}, NLSTEVOL à la classe nominale {processus_évolutif}. Notre classe verbale pivot {analyse} est ici représentée par ANALY.

Les expressions polylexicales, dans la présentation des résultats, sont organisées par relation de dépendance (entre le pivot et le premier cooccurrent) puis par collocatif. Par exemple, nous constatons, dans l'illustration ci-dessus, que le premier collocatif de la classe {analyse} en position objet est la classe {relation}. L'interface permet également d'étendre les expressions polylexicales (par l'ajout de cooccurrents) et d'accéder aux concordances correspondant à ces constructions.

Il est ainsi possible d'analyser en contexte si l'expression remplit une fonction rhétorique stable et de vérifier si les réalisations de l'expression correspondent à un paradigme large (plusieurs éléments de la classe peuvent l'intégrer), ou à un unique élément de la classe.

L'extraction met ainsi en évidence l'ALR illustré ci-après.

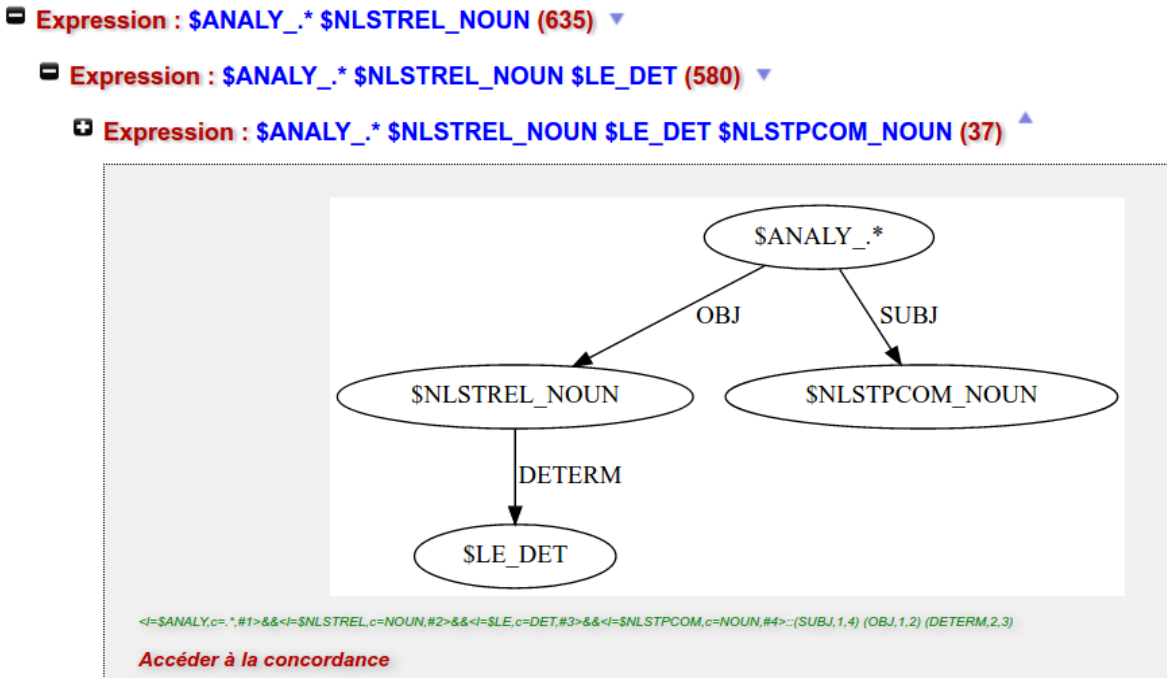


Illustration 5.9: ALR – {communication_support} {analyse} {relation}

Cet arbre peut être représenté de la façon suivante :

- (SUJET $\text{Nom}_{\text{communication\_support}}$) $\text{Verbe}_{\text{analyse}}$ (OBJET $\text{Nom}_{\text{relation}}$)

Il est notamment réalisé dans les exemples suivants :

- *La section*$_{\text{communication\_support}}$ *2 décrit*$_{\text{analyse}}$ *les implications*$_{\text{relation}}$ *de la théorie des anticipations dans le cadre d'une représentation RVAR.*[41]
- *Le tableau*$_{\text{communication\_support}}$ *II présente*$_{\text{analyse}}$ *les corrélations*$_{\text{relation}}$ *entre les scores cognitifs et scolaires des différentes épreuves.*[42]
- *les publications*$_{\text{communication\_support}}$ *de l'ONCSF examinent*$_{\text{analyse}}$ *parfois les incidences*$_{\text{relation}}$ *sociales de la territorialité animale.*[43]

---

41 Jondeau, É. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. *Annales d'Économie et de Statistique*, 139-174.

42 Barrouillet, P., Camos, V., Morlaix, S., & Suchaut, B. (2008). Progressions scolaires, mémoire de travail et origine sociale : quels liens à l'école élémentaire?. *Revue française de pédagogie. Recherches en éducation*, (162), 5-14.

43 Poinsot, Y. (2009). Protection de la grande faune et territoires : deux modèles de gestion dans la cordillère Cantabrique. *L'Espace géographique*, *38*(4), 289-302.

- *Cet article*$_{\text{communication\_support}}$ *évalue*$_{\text{analyse}}$ *l'impact*$_{\text{relation}}$ *des formations financées par les employeurs sur la mobilité et les rémunérations des salariés.*[44]

Les différents exemples nous permettent de constater que les paradigmes sont larges pour l'ensemble des classes lexicales. Le nom sujet se réalise ainsi au travers de plusieurs sous-classes de la classe {communication_support} : {/document} (*article, publication*), {/section} (section), {/graphique} (*figure, tableau*). Le verbe de l'ALR renvoie aux trois sous-classes de la classe {analyse} que nous avons choisi d'étudier. Enfin, le nom complément d'objet renvoie à deux sous-classes de la classe {relation} : {/implication} (*impact, incidence*) et {/association} (*corrélation, équilibre, relation*). L'extraction d'ALR, configurations contraintes syntaxiquement et sémantiquement, tire ainsi profit de la classification du LST et permet de rapprocher des réalisations diverses.

Cet ALR correspond dans les concordances à des segments dont la fonction est la définition de la thématique d'un document ou d'une partie d'un document. La méthode permet donc de faire émerger des constructions assurant une fonction identique. La routine de présentation de la thématique est ainsi définie à différents niveaux de granularité. En effet, cette routine associe une classe sémantique verbale qui prend pour sujet un élément défini par un trait de sous-classe nominale et pour objet un élément défini par un trait de classe nominale. Nous notons d'ailleurs que l'objet, dans cette routine, n'est pas obligatoirement contraint au niveau de la classe sémantique. Ainsi, les concordances mettent en évidence pour le nom objet du verbe un paradigme large, recouvrant un grand nombre de classes du LST : *démarche, cadre, fonctionnement, situation, cas, durée, modèle, moyenne, nombre, taux, pourcentage, estimation, interprétation, conclusion…*

En observant les concordances, nous avons constaté que l'ALR présenté dans l'illustration 5.9 correspond principalement à deux ensembles de routines :

- les routines avec un sujet de la sous-classe nominale {communication_support/graphique} et un verbe de la sous-classe {analyse/description}, que nous détaillons section 5.4.2.4

[44] Fougère, D., Goux, D., & Maurin, E. (2001). Formation continue et carrières salariales. *Annales d'économie et de statistique*, 62, 49-69.

- les routines avec un sujet de la sous-classe nominale {communication_support/document} et un verbe de la sous-classe {analyse/examen} abordées section 5.4.2.2

Un autre ALR extrait se rapproche de ces routines, la différence étant la classe lexicale objet du verbe, tel qu'illustré ci-dessous.

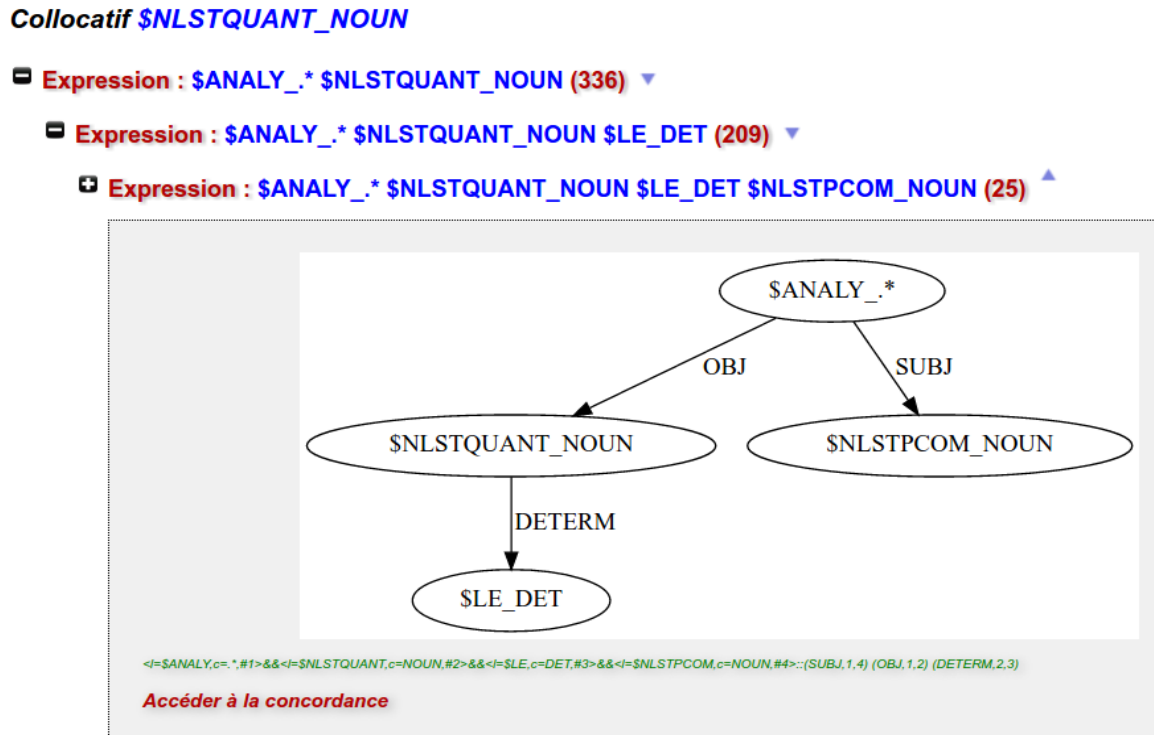


Illustration 5.10: ALR – {communication} {analyse} {quantité}

L'unique différence avec l'ALR précédent est que l'objet est ici réalisé par un nom de la classe {quantité}. Les exemples suivants correspondent alors à cet ALR :

- *Cette publication*communication_support *présente*analyse *les résultats*quantité *d'une enquête par questionnaire menée en mai et juin 2005 auprès de toutes les CAF.*[45]
- *Cet article*communication_support *présente*analyse *les résultats*quantité *de l'application d'une méthode – les personas – permettant de produire des connaissances sur l'utilisateur.*[46]

---

45 Destremau, B., & Messu, M. (2008). Le droit à l'assistance sociale à l'épreuve du local. *Revue française de science politique*, *58*(5), 713-742.

46 Brangier, E., Bornet, C., Bastien, J. M. C., Michel, G., & Vivian, R. (2012). Effets des personas et contraintes fonctionnelles sur l'idéation dans la conception d'une bibliothèque numérique. *Le travail humain*, *75*(2), 121-145.

- *La figure*$_{\text{communication\_support}}$ *1 présente*$_{\text{analyse}}$ *les moyennes*$_{\text{quantité}}$ *globales de la proportion du temps total pris pour identifier le changement à travers l'ensemble des sujets et des images.*[47]

En analysant les concordances, nous avons dégagé deux ensembles distincts. Le premier est constitué des exemples pour lesquels l'objet (de la classe lexicale {quantité}) se réalise par le nom *résultat*. Dans ces exemples, le nom *résultat* renvoie à l'acception de 'effet', appartenant à la sous-classe {relation/implication}, et non à l'acception 'calcul' de la sous-classe {quantité/mesure}. Cet ensemble est donc à regrouper avec l'ALR précédemment étudié (voir illustration 5.9). Le second correspond à une routine détaillée section 5.4.2.4 dans laquelle le verbe ne se réalise qu'à l'intérieur de la sous-classe {analyse/description}. La polysémie de certains noms du LST a ici pour effet de regrouper deux ensembles différents qui sont dégroupés lorsque le niveau sémantique se situe au niveau des sous-classes et non des classes.

Cependant, ainsi que nous l'avons évoqué pour le premier ALR (voir illustration 5.9), la classe lexicale du nom en position objet est la moins contrainte, ce qui nous incite à formaliser une routine plus générique, regroupant ces deux ALR, sous la forme suivante :

- (SUJET Nom$_{\text{communication}}$) Verbe$_{\text{analyse}}$ (OBJET Nom$_{\text{LST}}$)

Cette routine a ainsi l'avantage de correspondre à des occurrences dans lesquelles la thématique ne renvoie ni à une relation ni à une quantité, comme dans les exemples suivants, dans lesquels le nom objet appartient au LST :

- *Le présent article*$_{communication\_support}$ *examinera*$_{analyse}$ *les arguments*$_{LST}$ *avancés par les tenants de la réforme pour justifier celle-ci.*[48]
- *La figure*$_{communication\_support}$ *1 présente*$_{analyse}$ *le nombre*$_{LST}$ *moyen de dessins évalués positivement en fonction de l'âge du dessinateur et de l'émotion requise pour la méthode initiale (A) et pour la méthode des réponses libres (B).*[49]

---

47 Auvray, M., & O'Regan, J. K. (2003). L'influence des facteurs sémantiques sur la cécité aux changements progressifs dans les scènes visuelles. *L'année Psychologique*, *103*(1), 9-32.

48 Nanta, A. (2008). L'arrière-plan idéologique de la réforme scolaire au Japon. *Revue française de pédagogie. Recherches en éducation*, (165), 105-115.

- *Cet <u>article</u>*$_{\text{communication\_support}}$ *<u>décrit</u>*$_{\text{analyse}}$ *des <u>éléments</u>*$_{\text{LST}}$ *d'appréciation sur le caractère plus ou moins équitable du système de retraite par répartition en France.*[50]

- *Le <u>tableau</u>*$_{\text{communication\_support}}$ *IV <u>présente</u>*$_{\text{analyse}}$ *une <u>comparaison</u>*$_{\text{LST}}$ *de ces variables favorables à la réussite des élèves entre la France et chacun des pays considérés dans cette étude.*[51]

Nous voyons ainsi que la granularité de l'étiquetage sémantique dans les routines doit être envisagée sur plusieurs niveaux. Certains éléments de routine étant peu spécifiés, il peut être plus efficace de recourir à une simple étiquette LST, voire à aucune restriction sémantique si le paradigme est encore plus large. À l'inverse, d'autres éléments de routine sont plus restreints que le niveau des classes et méritent d'être étiquetés au niveau des sous-classes.

Les exemples correspondant à des routines ainsi définies, au-delà du partage de propriétés syntaxiques et sémantiques, ont alors pour point commun leur fonction discursive de définition de la thématique ou de l'objectif d'un document ou d'une sous-partie de ce document (chapitre, section, tableau, figure, etc.).

Nous abordons dans les sections suivantes cet aspect de la granularité, en examinant les ALR dont le pivot est défini par une sous-classe verbale.

5.4.2.2 Pivot de la sous-classe {/examen}

Nous avons réitéré la procédure d'extraction des ALR en nous intéressant à la sous-classe {analyse/examen}, représentée pour l'expérience par trois verbes : *analyser, étudier, examiner*.

Nous avons vu dans la partie précédente que la classe {analyse} s'inscrit dans une routine de définition de la thématique du travail présenté. L'extraction des ALR pour la sous-classe verbale {examen} fait apparaître une spécification de cette routine. L'ALR est en effet plus particulier puisqu'il restreint le verbe à la sous-

---

[49] Brechet, C., Picard, D., & Baldy, R. (2007). Expression des émotions dans le dessin d'un homme chez l'enfant de 5 à 11 ans. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *61*(2), 142.

[50] Dantec, A., Nauze-Fichet, E., & Pelgrin, F. (2000). Projection de trajectoires économiques par microsimulation: Quelle équité pour les retraites ?. *Revue économique*, 115-129.

[51] Meuret, D. (2003). Pourquoi les jeunes Français ont-ils à quinze ans des compétences inférieures à celles de jeunes d'autres pays ?. *Revue française de pédagogie*, *142*(1), 89-104.

classe {analyse/examen} et le nom sujet à la sous-classe {communication_support/document}. Cet ALR est notamment réalisé dans les exemples suivants :

- *Cet article*$_{\text{document}}$ *examine*$_{\text{analyse}}$ *la relation*$_{\text{LST}}$ *entre volatilité de court terme des taux de change et volume du commerce international dans un double but.*[52]
- *Ce document*$_{\text{document}}$ *étudie*$_{\text{examen}}$ *l'effet*$_{\text{LST}}$ *d'une catastrophe climatique irréversible sur la décision optimale dans un modèle de pollution cumulative.*[53]

Nous pourrions, en l'état, formaliser la routine de la manière suivante :

- (SUJET Nom$_{\text{document}}$) Verbe$_{\text{examen}}$ (OBJET Nom$_{\text{LST}}$)

Nous avons cependant noté que deux sous-classes nominales participent à cette routine, et sont de ce fait les thématiques les plus fréquemment annoncées dans le corpus : les sous-classes {relation/association} (*lien, relation, équilibre, corrélation, interaction*) et {relation/implication} (*effet, conséquence, rapport, contrainte*). Ces thématiques correspondent alors souvent à une relation dont l'auteur se propose d'expliquer la nature, la cause ou le fonctionnement. D'un autre côté, les cooccurrences fréquentes entre un sujet de la sous-classe {communication_support/graphique} et un objet de la classe {quantité} sont représentatives de la fonction de ces éléments graphiques (figure, tableau) dont le but dans le texte est de mettre en évidence une donnée quantifiable servant l'argumentation de l'auteur.

Le *Lexicoscope* donne également accès aux paradigmes lexicaux (en listant les lemmes) pour chaque ALR. Nous avons ainsi constaté la présence d'un ALR voisin de cette routine, mais réalisé avec un sujet pronom humain (*je, nous, on*) ou un nom appartenant à la classe nominale {personne} (*auteur, chercheur*). Le sujet, de type humain, peut aussi être réalisé, par métonymie, par les noms *étude* et *travail*. Les exemples suivants sont ainsi à regrouper avec la routine précédente :

---

[52] Vanelle, V. (2001). L'impact de la volatilité des taux de change sur le commerce international, l'apport des études empiriques. Économie Appliquée, tome LIV, 2, 59-90.

[53] Rouillon, S. (2000). Catastrophe climatique irréversible et politique de l'effet de serre. *Annales d'Économie et de Statistique*, 165-175.

- *Nous*$_{+humain}$ *cherchons ici à étudier*$_{examen}$ *le lien*$_{LST}$ *entre les caractéristiques des récits, leur accroche ainsi que leur niveau de rappel par les étudiants.*[54]
- *Les auteurs*$_{personne}$ *analysent*$_{examen}$ *notamment le système*$_{LST}$ *d'opposition, dans la musique dogon, entre éléments mâle (na) et femelle (tolo).*[55]
- *Les travaux analysant*$_{examen}$ *cette évolution*$_{LST}$ *privilégient le terme de « nouveaux modes de gouvernance » européens (NMG) supposés plus souples et participatifs.*[56]
- *Les études parues sur ce sujet analysent*$_{examen}$ *essentiellement la relation*$_{LST}$ *anaphorique.*[57]

Pour l'ensemble de ces exemples, les ALR extraits assurent la même fonction de définition de l'objectif des travaux présentés. Une formalisation plus générique de la routine de présentation de la thématique nous amène alors à regrouper les variantes suivantes ;

- (SUJET Nom$_{document}$) Verbe$_{examen}$ (OBJET Nom$_{LST}$)
- (SUJET Pronom$_{+humain}$) Verbe$_{examen}$ (OBJET Nom$_{LST}$)
- Nom$_{personne}$ (SUJET) Verbe$_{examen}$ (OBJET Nom$_{LST}$)
- (SUJET *étude*|*travail*) Verbe$_{examen}$ (OBJET Nom$_{LST}$)

Les éléments intégrant la routine peuvent ainsi être définis selon plusieurs attributs : l'appartenance au LST, à une classe ou sous-classe du LST, à une catégorie du discours (par exemple les verbes) ou à un lemme en particulier. En prenant en compte l'ensemble de ces propriétés, nous pouvons alors identifier des routines avec des paradigmes larges et dont la fonction dans le texte est stable.

---

54 Beaujouan, J., & Daniellou, F. (2012). Les récits professionnels dans une formation d'ergonomes. *Le travail humain*, *75*(4), 353-376.

55 Gérard, B. (2009). De l'ethnographie à l'ethnomusicologie. L'*homme*, (3), 139-173.

56 Ravinet, P. (2011). La coordination européenne « à la bolognaise ». *Revue française de science politique*, *61*(1), 23-49.

57 Godart-Wendling, B. (2000). Comment ça réfère ?. *Revue de sémantique et pragmatique*, *7*, pp-105.

5.4.2.3 Pivot de la sous-classe {/évaluation}

La sous-classe verbale {analyse/évaluation} est représentée ici par les verbes suivants : *estimer, évaluer, mesurer*. Comme nous l'avons vu section 5.4.2.1, cette sous-classe s'inscrit, comme l'ensemble de la classe {analyse} dans la routine de thématisation/définition de l'objectif. L'extraction des ALR pour cette sous-classe fait apparaître une autre configuration statistiquement significative remplissant cette fonction, que l'on peut observer dans les exemples suivants :

- *Notre approche*$_{\text{LST}}$ *consiste à évaluer*$_{\text{évaluation}}$ *l'influence*$_{\text{implication}}$ *de la morphologie urbaine sur la répartition spatiale des polluants.*[58]
- *Des enquêtes*$_{\text{LST}}$ *de socio-psychologie ont entrepris de mesurer*$_{\text{évaluation}}$ *les effets*$_{\text{implication}}$ *de ces campagnes sur les comportements électoraux.*[59]
- *Parallèlement, le nombre toujours croissant d'études*$_{\text{LST}}$ *empiriques cherchant à estimer*$_{\text{évaluation}}$ *l'impact*$_{\text{implication}}$ *de la volatilité des changes sur le commerce témoigne de l'intérêt toujours vif accordé à ce sujet.*[52]

Nous remarquons que pour l'ensemble de ces exemples, le nom sujet appartenant au LST renvoie implicitement à l'auteur, et donc à un sujet humain. Nous identifions ainsi une routine dans laquelle le sujet est un nom du LST (dont la classe n'est pas définie mais qui doit pouvoir représenter l'auteur) et l'objet est un nom de la sous-classe {relation/implication}, que l'on peut représenter de la façon suivante :

- (SUJET Nom$_{\text{LST}}$) Verbe$_{\text{évaluation}}$ (OBJET Nom$_{\text{implication}}$)

La fonction portée par cet ALR est alors de clarifier l'objectif du travail en question ou de travaux cités dans le texte. Nous remarquons également que, dans ces trois exemples, le verbe de la sous-classe {analyse/évaluation} est objet d'un verbe (*chercher, entreprendre, consister*) dans une construction infinitive. Nous pourrions ainsi inclure cet élément supplémentaire dans la routine comme suit :

- (SUJET Nom$_{\text{LST}}$) (Verbe — /*infinitif*/ Verbe$_{\text{évaluation}}$) (OBJET Nom$_{\text{implication}}$)

---

[58] Maignant, G. (2007). Dispersion de polluants et morphologie urbaine. *L'espace géographique*, *36*(2), 141-154.

[59] Braconnier, C., & Dormagen, J. Y. (2010). Le vote des cités est-il structuré par un clivage ethnique ?. *Revue française de science politique*, *60*(4), 663-689.

5.4.2.4 Pivot de la sous-classe {/description}

Nous avons enfin extrait les ALR avec pour pivot la sous-classe {description}, représentée par les verbes *présenter*, *décrire* et *illustrer*. Comme nous l'avons vu section 5.4.2.1, cette sous-classe s'intègre également dans des routines de définition du thème d'un document ou d'une partie de document :

- *La* section$_{communication_support}$ *2* présente$_{description}$ *le* modèle$_{LST}$ *et la solution de premier rang.*[60]
- *L'annexe*$_{communication\_support}$ *2* présente$_{description}$ *l'approche*$_{LST}$ *utilisée pour estimer par simulation la distribution de la statistique du test formel et des statistiques de test affaiblies.*[61]
- *Cet* article$_{communication_support}$ présente$_{description}$ *une* méthodologie$_{LST}$ *d'analyse de la variation des locutions et des termes désignés par "connecteurs discursifs" ou "mots du discours".*[62]

Nous pouvons représenter cette routine de la façon suivante :

- (SUJET Nom$_{communication_support}$) Verbe$_{description}$ (OBJET Nom$_{LST}$)

En observant les concordances, nous constatons que le paradigme du nom sujet est assuré par trois sous-classes de la classe {communication_support}. D'une part les sous-classes {/document} et {/section} qui sont alors sujets des verbes *présenter* et *décrire*. D'autre part, la sous-classe {/graphique} entre en cooccurrence avec les verbes *présenter* et *illustrer*. L'intégration des sous-classes nominales permet alors de distinguer les cooccurrences possibles de celles peu probables. Nous ne trouvons ainsi pas trace dans notre corpus des cooccurrences telles *un article illustre*[63]*, un tableau/une figure décrit*[64].

---

60 Mougeot, M. (2000). La tarification hospitalière : de l'enveloppe globale à la concurrence par comparaison. *Annales d'économie et de statistique*, 195-213.

61 Jondeau, E. (2001). La théorie des anticipations de la structure par terme permet-elle de rendre compte de l'évolution des taux d'intérêt sur euro-devise ?. Annales d'économie et de statistique, 62, 139-174.

62 Franckel, J. J., & Paillard, D. (1997). Représentation formelle des mots du discours ; le cas de 'd'ailleurs'. *Revue de sémantique et de pragmatique*, *1*, 51-64.

63 Le nom *article* n'est sujet du verbe *illustrer* que dans deux occurrences dans notre corpus d'analyse.

64 Les noms *figure* et *tableau* ne sont jamais sujets du verbe *décrire* dans notre corpus d'analyse.

Les résultats fournis par le *Lexicoscope* comportent également un ALR intégrant un nom de la classe {communication_support} en tant que complément locatif d'un verbe de la sous-classe {analyse/description}. Considérons les exemples suivants :

- *Ces <u>données</u>*$_{\text{LST.}}$ *sont <u>illustrées</u>*$_{\text{description}}$ *<u>dans</u> la <u>figure</u>*$_{\text{communication\_support}}$ *1.*[65]
- *<u>Dans</u> la <u>section</u>*$_{\text{communication\_support}}$ *5.2, nous <u>décrivons</u>*$_{\text{description}}$ *succinctement la <u>théorie</u>*$_{\text{LST.}}$ *de J.-C. Milner […]*[66]
- *Cette <u>conception</u>*$_{\text{LST.}}$ *que Ribot a <u>présentée</u>*$_{\text{description}}$ *<u>dans</u> sa <u>thèse</u>*$_{\text{communication\_support}}$ *(1872) remonte à David Hartley […]*[67]

Le verbe pivot de cet ALR peut être réalisé à la voix active, à la voix passive et au passif réduit. Nous remarquons que pour cette routine également, le nom objet est peu spécifié, à l'inverse du sujet qui se réalise par la classe {communication_support}. Nous pouvons formaliser la routine correspondant à ces exemples de la façon suivante :

- Verbe$_{\text{description}}$ (OBJET Nom$_{\text{LST}}$) *dans* Nom$_{\text{communication_support}}$

Nous constatons ainsi que les différentes sous-classes verbales de la classe {analyse} se rejoignent dans des routines communes mais intègrent également des routines qui leur sont propres.

Parmi les ALR extraits pour la sous-classe {description}, un des plus fréquents correspond de façon erronée à l'acception de *présenter* au sens de 'comporter'[68] de la classe {état/inclusion}. Dans cet ALR, le nom objet du verbe *présenter* (le paradigme verbal est composé de ce seul élément) appartient à la sous-classe {état_qualité/caractéristique}. Les cooccurrences sont alors du type : *présenter caractéristique|particularité|caractère|propriété|trait*. La polysémie fait ainsi émerger une collocation particulière concernant une autre sous-classe

---

65 Dalle, N., & Niedenthal, P. M. (2003). La réorganisation de l'espace conceptuel au cours des états émotionnels. *L'année psychologique*, *103*(4), 585-616.

66 Bourigault, D., & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, (25), 131-151.

67 Bergounioux, G. (2001). Esquisse d'une histoire négative de l'endophasie [avec une attention presque exclusive pour les productions en langue française consacrées à cette question]. *Langue française*, *132*(1), 3-25.

68 Sens 14 dans le *LVF* – « l'aventure présente des risques » : http://rali.iro.umontreal.ca/LVF+1/alphabetique/P.html#presenter [consulté le 20/08/2016]

verbale, {état/inclusion} (*présenter, comporter, posséder, contenir*). Cet ALR correspond par ailleurs, lorsque le sujet du verbe LST est un terme, à un des patrons repéré dans l'expérimentation section 5.3.1 : (SUJET Nom_{Terme}) $Verbe_{inclusion}$ (OBJET $Nom_{caractéristique}$). La granularité des sous-classes verbales permet cependant le repérage de cette routine (d'attribution de caractéristiques à un objet, généralement un terme) et offre comme perspective l'utilisation de l'extraction des ALR, couplée à la classification du LST, pour faire émerger d'autres patrons aidant à la validation terminologique.

Cette première expérimentation sur les routines du LST, basée sur notre classification sémantique, nous a permis d'identifier des constructions partagées par des membres d'une même classe et ayant une fonction rhétorique stable. L'extraction automatique nécessite cependant d'être suivie d'une observation des concordances pour valider les occurrences correspondant à une même routine. Les traits sémantiques autorisent ainsi le groupement de paradigmes lexicaux pour l'extraction de routines. Les différents ALR repérés par cette technique nous conduisent également à nous interroger sur la granularité de la classification sémantique du LST pour cette application. Il serait intéressant de pouvoir définir les classes lexicales sur plusieurs niveaux et ainsi d'extraire des routines combinant ces différents niveaux.

Nous explorons dans la section suivante l'utilisation des classes sémantiques transcatégorielle avec en perspective une analyse des correspondances nominales et verbales des routines.

5.4.3 Routines et variations

La présente partie a pour objectif de mettre en évidence quelques apports potentiels de la classification sémantique du LST dans l'étude des variantes phraséologiques. Ces variantes sont repérables par le fait qu'elles mobilisent les mêmes classes sémantiques du LST mais au travers de catégories morpho-syntaxiques différentes.

Les traits sémantiques transcatégoriels permettent d'envisager l'extraction et le regroupement de classes de routines. Ainsi, à l'aide des traits sémantiques, il est possible de faire émerger automatiquement des motifs équivalents (intégrant les

mêmes classes de mots pour les mêmes rôles thématiques) réalisés par une expression nominale et verbale.

Ce phénomène est notamment étudié par Siepmann (2007) qui met en évidence des variantes récurrentes, comme l'illustrent les deux constructions suivantes ;

- *On peut raisonnablement supposer que* P ;
- *Il est raisonnable de penser que* P.

La typologie sémantique du LST, élaborée parallèlement sur l'ensemble des catégories syntaxiques traitées (nom, adjectif, verbe et adverbe), rend possible l'identification de variantes syntaxiques. Par exemple, si nous considérons les quatre phrases suivantes :

1. *Cet article$_{support}$ vise$_{finalité}$ à voir en quoi le « marché intérieur » a constitué la base du développement d'un dialogue social à un niveau communautaire.*[69]
2. *Cet article$_{support}$ tente$_{finalité}$ de dégager ce que la gestation pour autrui […] nous apprend de la maternité.*[70]
3. *Le but$_{finalité}$ de cet article$_{support}$ était d'apporter une contribution au thème des comportements de citoyenneté organisationnelle.*[71]
4. *L'objectif$_{finalité}$ de cet article$_{support}$ est de mettre en évidence l'impact du régime d'allocation chômage.*[72]

Le fait de relier la sous-classe sémantique nominale {objet_scientifique/finalité} et la sous-classe sémantique verbale {ancrage-construction/finalité} a pour but d'attribuer des traits sémantiques identiques aux noms *but* et *objectif* et aux verbes *viser* et *tenter*. Cette identité de traits sémantiques nous permet alors d'identifier le lien entre les deux premières

[69] Didry, C. (2009). L'émergence du dialogue social en Europe: retour sur une innovation institutionnelle méconnue. L'*Année sociologique*, *59*(2), 417-447.

[70] Parseval, G. D. D., & Collard, C. (2007). La gestation pour autrui: un bricolage des représentations de la paternité et de la maternité euro-américaines. L'*homme*, 29-53.

[71] Paillé, P. (2008). Les comportements de citoyenneté organisationnelle : une étude empirique sur les relations avec l'engagement affectif, la satisfaction au travail et l'implication au travail. *Le travail humain*, *71*(1), 22-42.

[72] Prieto, A. (2000). L'impact de la dégressivité des allocations chômage sur le taux de reprise d'emploi. *Revue économique*, 523-534.

constructions et les deux dernières : celles-ci sont en quelque sorte une nominalisation des premières. De même, la co-appartenance de *viser* et *tenter* à la sous-classe sémantique {/finalité} rend possible le regroupement de deux schémas syntactico-sémantiques similaires, définis ici par une relation de dépendance entre un gouverneur élément du LST de la sous-classe {/finalité} et un dépendant élément du LST de la sous-classe {/support}. Une représentation de ce patron pourrait être comme suit :

- (SUJET $\text{Nom}_{\text{support}}$) $\text{Verbe}_{\text{finalité}}$
    - correspondant aux exemples 1 et 2 ci-dessus.

  Une variante possible pour ce patron serait alors :

- $\text{Nom}_{\text{finalité}}$ Prep $\text{Nom}_{\text{support}}$
    - correspondant aux exemples 3 et 4 données en exemple.

Nous pourrions également confronter les différentes routines organisées autour d'un verbe à leurs potentielles nominalisations correspondantes. Pour cela, nous pouvons tirer parti des informations de dérivations du *DEM* et extraire automatiquement un verbe et son dérivé nominal afin de comparer leur combinatoire, à la manière de Condette *et al.* (2012) dans leur étude des déverbaux.

5.5 Perspectives sur la phraséologie et la classification sémantique du LST

Nous avons présenté dans ce chapitre plusieurs utilisations possibles du corpus arboré et de la classification sémantique du LST, pour diverses applications.

Nous avons dans un premier temps enrichi la ressource du LST pour la catégorie verbale, en collaborant avec Yan sur la modélisation des patrons verbaux, avec pour objectif une aide dans l'enseignement/apprentissage de l'écrit scientifique. Notre approche, combinant l'extraction automatique de cadres de sous-catégorisation et l'analyse manuelle des constructions et des acceptions, a permis de proposer une ressource des patrons verbaux dans leur usage dans l'écrit scientifique. Nous avons également pu intégrer dans ces patrons des informations

sémantiques sur les arguments verbaux en tirant parti de notre classification sémantique du LST.

Notre expérimentation sur un autre type de patrons, défini par la cooccurrence entre un terme et un élément du LST, nous a permis d'identifier un certain nombre de configurations susceptibles de valider le statut terminologique d'un candidat-terme lors de l'indexation. Nous avons cependant constaté que l'apport de ces cooccurrences repose sur l'utilisation d'un analyseur syntaxique pour la projection des patrons.

En dernier lieu, nous avons montré l'intérêt de la ressource sémantique du LST en intégrant cette classification à un outil d'interrogation de corpus permettant l'extraction de routines. Les résultats sont encourageants quant à l'adéquation des classes et sous-classes dans la détection de la phraséologie scientifique transdisciplinaire. Le traitement des routines ici présenté est ainsi transposable aux autres classes sémantiques verbales, adjectivales, nominales.

Les perspectives d'études (semi-) automatisables sont donc nombreuses, ainsi que nous l'abordons dans la conclusion.

Conclusion

Au terme de notre étude sur l'identification et la caractérisation syntaxique et sémantique du LST, nous conclurons par une synthèse du travail effectué et ouvrirons plusieurs perspectives d'utilisation et d'enrichissement de la ressource lexicale du LST.

Nous avons pu constater que le LST est un lexique spécifique au genre de l'écrit scientifique et qu'il est essentiel dans l'exposé de l'activité scientifique, dans l'argumentation et la structuration du discours scientifique. En combinant traitements automatiques et analyse manuelle, nous avons pu extraire les mots simples du LST et caractériser ces éléments aux niveaux sémantique et lexico-syntaxique. En partant des textes actualisant le genre dans lequel s'inscrit le LST, nous avons élaboré une ressource lexicale du LST intégrant les particularités d'usage dans les articles de recherche en SHS.

Ce champ scientifique se caractérise par une forte ambiguïté entre lexique terminologique et langue générale. Nous avons pu mettre en évidence ce problème de délimitation de lexique au travers des traitements automatiques mis en place. D'une part, le lexique des objets d'étude des SHS et le LST n'ont pu être entièrement discriminés à l'aide des critères statistiques, mettant ainsi en relief la proximité entre un lexique associé à un genre et un lexique renvoyant à des objets partagés par un ensemble disciplinaire. D'autre part, l'intégration finale du LAG dans la ressource du LST est révélatrice de l'intrication de ces différents lexiques dans le contexte des articles en SHS.

Faisant écho à l'observation de Lerat sur la spécificité lexico-syntaxique des lexiques associés à un genre particulier, notre étude du LST s'est appuyée sur l'extraction de ses propriétés de combinatoire, lesquelles sont interprétées manuellement en dernier lieu. Les caractérisations lexico-syntaxique et sémantique du LST nous ont alors amené à raffiner la notion de transdisciplinarité du LST. Ainsi, nous pensons que ce critère, en plus d'être appliqué au niveau lexical (nécessaire pour un inventaire des éléments du lexique), doit l'être aux niveaux d'analyse supérieurs, lexico-syntaxique et sémantique. En effet, l'analyse

sémantique doit s'appuyer sur le critère de transdisciplinarité pour identifier les acceptions réellement transversales des éléments du LST. De même, l'étude des cooccurrences, des patrons et des routines doit s'effectuer sur des phénomènes répartis dans les différents sous-corpus disciplinaires. Le critère de spécificité nécessiterait aussi d'être appliqué aux niveaux sémantique et phraséologique pour faire émerger les acceptions et constructions typiques des écrits scientifiques en regard des autres genres. Le LST s'est révélé être une amorce préférentielle dans la détection des phénomènes phraséologiques transdisciplinaires, comme nous avons pu l'expérimenter dans le chapitre 5.

Nous avons par ailleurs montré l'intérêt de notre classification du LST en classes et sous-classes sémantiques dans la détection de tels phénomènes. Cette utilisation des résultats de nos traitements sémantiques pour une étude des constructions sémantico-syntaxiques nous a ainsi conforté dans notre approche de linguistique outillée, dans laquelle la validation manuelle de traitements automatiques vient enrichir les possibilités de traitements de niveaux supérieurs. Nous avons d'ailleurs pour perspective la projection de notre ressource du LST sur des corpus d'autres disciplines pour une analyse de ce lexique dans des domaines connexes.

De plus, bien que non reproductible à l'identique, nous estimons que notre démarche reste valable pour l'étude de lexiques apparentés, dits de genre ou de spécialité, mais définis par des textes ou des appartenances disciplinaires différentes. Plusieurs perspectives nous apparaissent intéressantes à explorer dans ce sens. Outre les intérêts et applications du LST déjà abordés, nous voyons ainsi, pour de futurs travaux, plusieurs (ré-)utilisations possibles de notre ressource du LST, et plus globalement du travail ici présenté. Notre approche d'un lexique spécifique peut ainsi être adaptée à des corpus dont la composition serait différente du corpus d'analyse utilisé dans notre travail.

Les objectifs et applications présentés en introduction nous ont conduit à l'étude d'un LST particulier, que l'on peut redéfinir en fonction des disciplines représentées et de la référence adoptée que constitue le corpus de contraste. Dans une perspective contrastive, la méthode ici détaillée pourrait être appliquée à des ensembles disciplinaires différents et ainsi faire émerger des lexiques spécifiques

divers. Ainsi, dans de futurs travaux, nous souhaiterions intégrer des articles d'autres ensembles de sciences (sciences formelles et de la nature).

D'une part, nous pourrions, ce faisant, identifier le lexique partagé par l'ensemble des différentes sciences, au-delà des SHS, et observer dans quelle mesure les acceptions mobilisées seraient influencées. D'autre part, il serait possible, en opposant un corpus d'articles en SHS à un corpus d'articles de sciences expérimentales et sciences de la nature, de faire émerger un lexique spécifique à chaque ensemble. Une analyse contrastive de ces lexiques permettrait alors une mise en évidence des propriétés particulières de chaque ensemble disciplinaire.

Selon la même approche pourrait être effectuée une caractérisation des sous-genres de l'écrit scientifique en étudiant les points communs et divergences entre différents types d'écrits scientifiques, tels, outre l'article de recherche, les actes de colloque, thèses ou ouvrages.

L'ensemble de ces perspectives d'études contrastives peut alors prendre pour objet d'étude le niveau lexical du LST ainsi que les autres niveaux abordés dans le présent travail : propriétés lexico-syntaxiques, sémantiques et phraséologiques.

Parallèlement à ces perspectives, nous envisageons plusieurs études pour enrichir la description du LST dans les SHS. Ainsi, une observation du comportement des éléments du LST dans les sous-corpus disciplinaires composant notre corpus d'analyse pourrait permettre de mettre en évidence des groupes disciplinaires à l'intérieur des SHS. Nous pouvons également envisager une étude centrée sur les classes sémantiques mobilisées selon les disciplines pour en analyser les différences. Il faut cependant rappeler que bien que nous disposions d'une ressource structurée sémantiquement, les corpus sur lesquels nous travaillons sont annotés au niveau des unités lexicales et non des acceptions. Un tel travail nécessiterait alors une désambiguïsation dans les textes pour permettre une comparaison des occurrences par disciplines.

Nous souhaitons également rappeler le fait que notre ressource n'est pas figée en l'état, et que nous procédons à son enrichissement au fur et à mesure, notamment afin de réduire le silence originellement produit, cause de l'absence par exemple des verbes *voir* et *comprendre* dans notre ressource. Ce rattrapage

manuel pourra alors bénéficier de la méthodologie présentée dans notre travail ainsi que de la classification et des tests d'appartenance associés pour une maintenance en cohérence avec les principes de constitution de la ressource.

Nous considérons par ailleurs des études complémentaires pour certaines expérimentations détaillées précédemment, notamment au niveau de la classification automatique. Ainsi que nous l'avons mentionné dans la section 4.6.3, il nous paraît intéressant de mener une étude sur la classification automatique des éléments du LST en prenant comme attributs non pas des propriétés lexico-syntaxiques prises isolément, mais des contextes d'occurrence, prenant en compte les différentes relations lexico-syntaxiques dans lesquelles s'inscrit chaque élément. Nous notons cependant qu'une telle expérimentation, bien qu'elle permette une meilleure représentation des contextes, nécessiterait sans doute un corpus d'une taille supérieure dans la mesure où des contextes étendus (représentés sous la forme de sous-arbres) ont moins de probabilité d'être partagés par plusieurs éléments qu'une unique relation de dépendance.

L'étude des cooccurrences LST-terme mérite par ailleurs d'être approfondie, notamment en appliquant les patrons sur un corpus arboré, afin de tirer profit des relations de dépendance composant les patrons avant leur linéarisation. Nous souhaitons ainsi mettre en place une expérimentation dont les résultats soient directement ré-exploitables dans le cadre d'un processus d'extraction terminologique.

Enfin, la piste de travail la plus prometteuse à nos yeux se situe dans la poursuite de l'étude des routines sémantico-rhétoriques en utilisant la classification sémantique des mots simples du LST comme amorce pour faire émerger ces unités phraséologiques dont la fonction est essentielle dans les écrits scientifiques.

Bibliographie

Abney, S. (1991). Parsing by chunks. In R. Berwick, S. Abney, & C. Tenny (Éd.), *Principle-based Parsing* (p. 257–278). Dordrecht: Kluwer Academic Publishers.

Adam, J.-M. (1999). *Linguistique textuelle: des genres de discours aux textes*. Paris: Nathan.

Adler, S. (2012). Trois questions relatives aux noms généraux factuels attitudinaux. *Scolia*, *26*, 11-37.

Adler, S., & Eshkol-Taravella, I. (2012). « Geste » et « démarche » en tant que noms généraux dans le langage médiatique écrit. *Revue de Sémantique et Pragmatique*, 90-132.

Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, *8*(2-3), 121–144.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley FrameNet project. In *ACL-COLING* (p. 86–90). Montréal.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Baroni, M., & Lenci, A. (2011). How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the 2011 Workshop on GEometrical Models of Natural Language Semantics* (p. 1–10). Edinburgh.

Barque, L. (2015). Les noms relationnels de type humain. *Langue Française*, *1*(185), 29-41.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison: University of Wisconsin Press.

Bazerman, C. (2011). The Disciplined Interdisciplinarity of Writing Studies. *Research in the Teaching of English*, *46*(1), 8-21.

Beacco, J.-C. (2004). Trois perspectives linguistiques sur la notion de genre discursif. *Langages*, *1*(153), 109-119.

Bendaoud, R., Hacene, M. R., Toussaint, Y., Delecroix, B., & Napoli, A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. In *IC 2007*. Grenoble.

Bendaoud, R., Toussaint, Y., & Napoli, A. (2010). L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des Nouvelles Technologies de l'Information*, 133–164.

Benninger, C., & Theissen, A. (2013). Lexique des noms, regards croisés. Présentation. *Cahiers de lexicologie*, *2*(103), 9-12.

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

Bestgen, Y. (2012). Analyse des différences lexicales entre des corpus: test ou distance du Khi-2? In *Actes de JADT 2012* (p. 150-161). Liège.

Biber, D. (1993a). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Biber, D. (1993b). Using register-diversified corpora for general language studies. *Computational linguistics*, *19*(2), 219–241.

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins.
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371–405.
Billami, M.-B., Camacho-Collados, J., Jacquey, E., & Kister, L. (2014). Annotation sémantique et validation terminologique en texte intégral en SHS. In *Actes de TALN 2014* (p. 363–376). Marseille.
Blumenthal, P. (2007). Sciences de l'Homme vs sciences exactes : combinatoire des mots dans la vulgarisation scientifique. *Revue française de linguistique appliquée*, *XII*(2), 15-28.
Blumenthal, P. (2008). Combinatoire des prépositions : approche quantitative. *Langue française*, *1*(157), 37-51.
Bouaud, J., Habert, B., Nazarenko, A., & Zweigenbaum, P. (2000). Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. In J. Charlet, M. Zacklad, G. Kassel, & D. Bourigault (Éd.), *Ingénierie des Connaissances, évolutions récentes et nouveaux défis* (p. 275-290). Paris: Eyrolles.
Bouillon, P., & Viegas, E. (Éd.). (2001). Lexiques sémantiques dans les applications du traitement automatique. *TAL*, *42*(3), 663-665.
Bourigault, D., Aussenac-Gilles, N., & Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, *18*(1), 87–110.
Brun, C., Jacquemin, B., & Segond, F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. *Traitement Automatique des Langues*, *42*(3), 667-690.
Cavalla, C., & Loiseau, M. (2013). Scientext comme corpus pour l'enseignement. In A. Tutin & F. Grossmann (Éd.), *L'écrit scientifique : du lexique au discours. Autour de Scientext* (p. 163–182). Rennes: Presse universitaire de Rennes.
Cimiano, P., Staab, S., & Tane, J. (2003). Automatic acquisition of taxonomies from text: FCA meets NLP. In *Proceedings of the PKDD/ECML'03 International Workshop on Adaptive Text Extraction and Mining (ATEM)* (p. 10-17). Cavtat/Dubrovnik.
Condette, M.-H., Marin, R., & Merlo, A. (2012). La structure argumentale des noms déverbaux: du corpus au lexique et du lexique au corpus. In *Actes de CMLF 2012* (p. 845–858). Lyon.
Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, *8*(4), 389–399.
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.
Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. *Language and Computers*, *42*(1), 73–89.
Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.
Da Sylva, L. (2009). Corpus-based derivation of a « basic scientific vocabulary » for indexing purposes. *Journal of Linguistics*, *45*(1), 167–201.
Da Sylva, L. (2010). Extraction semi-automatique d'un vocabulaire savant de base pour l'indexation automatique. In *Actes de TALN 2010*. Montréal.
Delbecque, N. (2006). Ce qu'il y a dans un mot: la sémantique lexicale. In N. Delbecque, *Linguistique cognitive* (p. 47-76). Bruxelles: De Boeck Supérieur.
Depecker, L. (2002). *Entre signe et concept: éléments de terminologie générale*. Paris: Presses Sorbonne Nouvelle.

Desmet, I. (2006). Variabilité et variation en terminologie et langues spécialisées: discours, textes et contextes. In *Septièmes journées scientifiques du Réseau « Lexicologie, Terminologie, Traduction » de l'Agence Universitaire de la Francophonie : Mots, Termes et Contextes* (p. 235–247). Bruxelles.

Diwersy, S., & Kraif, O. (2013). Observations statistiques de cooccurrents lexico-syntaxiques pour la catégorisation sémantique d'un champ lexical. In F. Baider & G. Cislaru (Éd.), *Cartographie des émotions. Propositions linguistiques et sociolinguistiques* (p. 55–69). Paris: Presse Sorbonne Nouvelle.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9*(1), 99–115.

Drouin, P. (2004). Spécificités lexicales et acquisition de la terminologie. In *Actes de JADT 2004* (p. 345–352). Louvain-la-Neuve.

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

Dubois, J., & Dubois-Charlier, F. (1997). *Les verbes français*. Paris: Larousse.

Dubois, J., & Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, (3), 31–56.

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B., & Mével, J.-P. (2001). *Dictionnaire de linguistique*. Paris: Larousse.

Dusserre, E. (2016). *Apport de la méthode distributionnelle à la constitution de classes sémantiques pour une liste de formes du lexique scientifique transdisciplinaire*. Mémoire de Master 2, Université Grenoble Alpes, Grenoble.

Erk, K. (2010). What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics* (p. 17–26). Uppsala.

Fabre, C., & Bourigault, D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Actes de TALN 2006* (p. 121–130). Louvain.

Fabre, C., Hathout, N., Sajous, F., & Tanguy, L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *Actes de TALN 2014* (p. 266-279). Marseille.

Fabre, C., & Lenci, A. (2015). Distributional Semantics Today Introduction to the special issue. *Traitement Automatique des Langues*, *56*(2), 7-20.

Falaise, A. (2014). Exploitation linguistique de corpus arborés d'écrits scientifiques à l'aide du logiciel ScienQuest. In A. Tutin & F. Grossmann (Éd.), *L'écrit scientifique : du lexique au discours. Autour de Scientext* (p. 123–143). Rennes: Presses Universitaires de Rennes.

Falaise, A., Tutin, A., & Kraif, O. (2011). Une interface pour l'exploitation de corpus arborés par des non informaticiens: la plate-forme ScienQuest du projet Scientext. *Traitement Automatique des Langues*, *52*(3), 241–246.

Fasciolo, M. (2012). Y-a-t-il un continuum entre nom et pronoms ? *Scolia*, *26*, 61-80.

Faure, D., & Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: The system ASIUM. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management* (p. 329–334). London.

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. Cambridge: MIT Press.

Fifielska, E. (2015). *Les constructions syntaxiques de l'écrit scientifique - exploration et analyses de corpus*. Mémoire de Master 2, Université Grenoble Alpes, Grenoble.
Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm* (p. 111–137). Séoul: Hanshin Publishing Co.
Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In J. R. Firth (Éd.), *Studies in Linguistic Analysis* (p. 1-32). Oxford: Blackwell.
Flaux, N., & Stosic, D. (2014). Les noms d'idéalités et la nominalisation. In J. Goes, C. Lachet, & A. Masset-Martin (Éd.), *NominalisationS* (p. 19-37). Arras: Artois Presses Université.
Flaux, N., & Van de Velde, D. (2000). *Les noms en français: esquisse de classement*. Paris: Ophrys.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378-382.
Fløttum, K., Dahl, T., & Kinn, T. (2006). *Academic voices: Across languages and disciplines* (Vol. 148). Amsterdam/Philadelphia: John Benjamins.
Fløttum, K., & Rastier, F. (2003). *Academic discourse: multidisciplinary approaches*. Oslo: Novus Press.
Gala, N., & Zock, M. (2013). *Ressources lexicales. Contenu, construction, utilisation, évaluation*. Amsterdam/Philadelphia: John Benjamins.
Galy, E., & Bourigault, D. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. In *Actes de JLC 2005* (p. 163–174). Lorient.
Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305-327.
Geeraerts, D. (2009). *Theories of lexical semantics*. Oxford: Oxford University Press.
Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, *31*(121), 124–126.
Girault, T. (2008). Exploitation de treillis de Galois en désambiguïsation non supervisée d'entités nommées. In *Actes de TALN 2008* (p. 260-269). Avignon.
Giry-Schneider, J. (1994). Sélection et sémantique: problèmes et modèles. *Langages*, *28*(115), 5–14.
Gledhill, C. (1994). La phraséologie dans l'analyse de genres textuels. L'exemple des formules rhétoriques dans Le Monde. In *Aston Papers in Languages Studies and Discourse Analysis* (Vol. 2 (Series Editor John Gaffney)). Birmingham: Aston University Publications.
Gledhill, C. J. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.
Godin, R., Mineau, G., Missaoui, R., & Mili, H. (1995). Méthodes de classification conceptuelle basées sur les treillis de Galois et applications. *Revue d'intelligence artificielle*, *9*(2), 105–137.
Goossens, V. (2011). *Propositions pour le traitement de la polysémie régulière des noms d'affect* (Thèse de doctorat). Université Grenoble Alpes, Grenoble.
Gougenheim, G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré): Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*. (Nouv. éd., refond. et aug.). Paris: Didier.
Granger, S., & Paquot, M. (2009a). In search of a General Academic vocabulary: A corpus-driven study. In K. Katsampoxaki-Hodgetts (Éd.), *Options and Practices of L.S.P practitioners Conference Proceedings* (p. 94-108).

Heraklion.
Granger, S., & Paquot, M. (2009b). Lexical Verbs in Academic Discourse: A Corpus-driven Study of Learner Use. In M. Charles, S. Hunston, & D. Pecorari (Éd.), *Academic Writing : At the Interface of Corpus and Discourse* (p. 193-214). London: Continuum.
Granger, S., & Paquot, M. (2010). Customising a general EAP dictionary to meet learner needs. In S. Granger & M. Paquot (Éd.), *eLexicography in the 21st century: New challenges, new applications. Proceedings of the eLex2009 Conference* (Vol. 6, p. 87–96). Louvain-la-Neuve: Cahiers du CENTAL.
Grefenstette, G. (1993). Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*. Columbus.
Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Dordrecht: Kluwer Academic Publishers.
Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, (115), 15–30.
Gross, G. (2008). Les classes d'objets. *Lalies*, (28), 111-165.
Gross, M. (1975). *Méthodes en syntaxe*. Paris: Hermann.
Grossmann, F. (2010). L'Auteur scientifique. *Revue d'anthropologie des connaissances*, *4*(3), 410-426.
Grossmann, F. (2014). Les verbes de constat dans l'écrit scientifique. In A. Tutin & F. Grossmann (Éd.), *L'écrit scientifique : du lexique au discours. Autour de Scientext* (p. 85-100). Rennes: Presses Universitaires de Rennes.
Haas, P., & Gréa, P. (2015). Action et événement, deux types nominaux distincts ? *Langue Française*, *185*, 85-98.
Habert, B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée*, *IX*(1), 5-24.
Habert, B., & Nazarenko, A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances: bilan critique d'une expérience. In *Journées sur l'acquisition des connaissances* (p. 137–142). Sète.
Habert, B., & Zweigenbaum, P. (2002). Contextual acquisition of information categories. In *The Legacy of Zellig Harris: Language and information into the 21st century* (Mathematics and computability of language, Vol. 2, p. 203-231). Amsterdam/Philadelphia: John Benjamins.
Habert, B., & Zweigenbaum, P. (2003). Classer les mots: sémantique à gros grain et méthodologie harrissienne. *Revue de sémantique et Pragmatique*, *12*, 101-119.
Hadouche, F., & Lapalme, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, *3*(179-180), 193-220.
Hagège, C., & Roux, C. (2003). Entre syntaxe et sémantique: Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. In *Actes de TALN 2003*. Batz-sur-mer.
Hagège, C., & Tannier, X. (2007). XRCE-T: XIP temporal module for TempEval campaign. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (p. 492–495). Prague.
Hamp, B., & Feldweg, H. (1997). Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (p. 9–15). Madrid.

Hanks, P. (2004). Corpus pattern analysis. In *Proceedings of the XI EURALEX International Congress* (p. 87-98). Lorient.
Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.
Harris, Z. S. (1991). *A theory of language and information. A mathematical Approach*. Oxford: Oxford University Press.
Harris, Z. S., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T., & Harris, S. (1989). *The form of Information in science: analysis of an immunology sublanguage*. Dordrecht: Kluwer Academic Publishers.
Hatier, S. (2013). Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. In *Actes de RECITAL'2013* (p. 138–149). Les Sables d'Olonne.
Hatier, S., Augustyn, M., Yan, R., Tran, T. T. H., Tutin, A., & Jacques, M.-P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In G. Meladze (Éd.), *Proceedings of the XVII EURALEX International congress* (p. 355-365). Tbilisi.
Hatier, S., & Yan, R. (à paraître). Analyse contrastive des constructions verbales dans l'écrit scientifique entre scripteurs étudiants et experts. *CORELA. Cognition, Représentation, Langages*.
Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational Linguistics* (p. 539–545). Nantes.
Henrich, V., & Hinrichs, E. W. (2010). Gern-EdiT - The GermaNet Editing Tool. In *Proceedings of LREC 2010* (p. 19–24). La Valette.
Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins.
Huyghe, R. (2015). Les typologies nominales : présentation. *Langue française*, *185*(1), 5-27.
Huyghe, R., & Tribout, D. (2015). Noms d'agents et noms d'instruments : le cas des déverbaux en -eur. *Langue française*, *185*(1), 99-112.
Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, *20*(3), 341–367.
Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62.
Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, *27*(1), 4–21.
Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of second language writing*, *6*(2), 183–205.
Hyland, K., & Tse, P. (2007). Is there an « academic vocabulary »? *TESOL quarterly*, *41*(2), 235–253.
Jacques, M.-P. (2011). Nous appelons X cet Y : X est-il un terme émergent ? In K. Kageura & P. Zweigenbaum (Éd.), *Actes de TIA 2011* (p. 31–37). Paris: INALCO.
Jacquey, E. (2013). Déverbaux en français, lexicographie et corpus. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*, (103), 63–84.

expérimentation en sciences humaines. In *Actes de TIA 2013* (p. 121-128). Paris.
Kayser, D. (1995). Terme et dénotation. *La Banque des mots*, (Numéro spécial 7), 19–34.
Kister, L., & Jacquey, E. (2012). Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral. In *Actes du 3ème Congrès Mondial de Linguistique Française* (p. 909-919). Lyon.
Kittredge, R., & Lehrberger, J. (1982). *Sublanguage: studies of language in restricted semantic domains*. Berlin: De Gruyter.
Kleiber, G., Benninger, C., Biermann Fischer, M., Gerhard-Krait, F., Lammert, M., Theissen, A., & Vassiliadou, H. (2012). Typologie des noms : le critère se trouver + SP locatif. *Scolia*, (26), 105-130.
Kleiber, G., & Lammert, M. (Éd.). (2012). Questions de sémantique nominale, (26).
Kocourek, R. (1991). *La langue française de la technique et de la science* (2eme édition). Zurich: Brandstetter Verlag.
Kraif, O. (2016). Le lexicoscope: un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*, (108), 91–106.
Kraif, O., & Diwersy, S. (2012). Le Lexicoscope: un outil pour l'étude de profls combinatoires et l'extraction de constructions lexico-syntaxiques. In *Actes de TALN 2012* (p. 399-406). Grenoble.
Kübler, N., & Pecman, M. (2012). The ARTES Bilingual LSP Dictionary: From Collocation to Higher Order Phraseology. In S. Granger & M. Paquot (Éd.), *Electronic Lexicography* (p. 186-208). Oxford: Oxford University Press.
Kupsc, A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Actes de TALN 2007*. Toulouse.
Labbé, C., & Labbé, D. (1994). Que mesure la spécificité du vocabulaire. *Lexicometrica*, *3*.
Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, *1*(1), 127–165.
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363–374.
Lapesa, G., & Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, *2*, 531–545.
Le Pesant, D. (1994). Les compléments nominaux du verbe lire une illustration de la notion de «classe d'objets». *Langages*, (115), 31–46.
Le Pesant, D., & Mathieu-Colas, M. (1998). Introduction aux classes d'objets. *Langages*, (131), 6–33.
Legallois, D., & Tutin, A. (2013). Présentation: Vers une extension du domaine de la phraséologie. *Langages*, (189), 3–25.
Lerat, P. (1997). Approches linguistiques des langues spécialisées. *ASp. la revue du GERAS*, (15-18), 1-10.
Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
L'Homme, M.-C. (2004). *La terminologie: principes et techniques*. Montréal: Les Presses de l'Université de Montréal.
L'Homme, M.-C. (2005). Sur la notion de «terme». *Meta: Journal des traducteurs / Meta:Translators' Journal*, *50*(4), 1112–1132.

L'Homme, M.-C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, *217*, 78–103.
Lux-Pogodalla, V., & Polguère, A. (2011). Construction of a French lexical network: Methodological issues. In *Proceedings of the First InternationalWorkshop on Lexical Resources, WoLeR 2011* (p. 54–61). Ljubljana.
Malaisé, V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels* (Thèse de Doctorat). Université Paris-Diderot-Paris VII, Paris.
Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics* (p. 235–242). Columbus.
Manser, M. (2012). État de l'art sur l'acquisition de relations sémantiques entre termes: contextualisation des relations de synonymie. In *Actes de TALN 2012* (p. 163-175). Grenoble.
Manuelian, H., Bruscand, A., Cholewka, N., & Hetzel, A.-M. (2010). Le Petit Larousse illustré de 1905 en ligne: secrets de fabrication et présentation. *Ela. Études de linguistique appliquée*, *156*(4), 453–474.
Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., & Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques*. Montréal: Les Presses de l'Université de Montréal.
Mel'čuk, I. A., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Paris/Louvain-la-Neuve: Duculot.
Messiant, C., Gábor, K., & Poibeau, T. (2010). Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français. *Traitement automatique des langues*, *51*(1), 65–96.
Meyer, I., & Mackintosh, K. (2000). L'étirement du sens terminologique: aperçu du phénomène de la déterminologisation. In H. Béjoint & P. Thoiron (Éd.), *Le sens en terminologie* (p. 198–217). Lyon: Presses universitaires de Lyon.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR 2013*. Scottsdale.
Morin, E., & Martienne, E. (2000). Using a symbolic machine learning tool to refine lexico-syntactic patterns. In *Proceedings of Machine Learning: ECML 2000* (p. 292–299). Barcelone.
Münchow, P. von. (2007). Le genre en linguistique de discours comparative. Stabilités et instabilités séquentielles et énonciatives. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (56), 109-125.
Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.
O'Donovan, R., Burke, M., Cahill, A., Van Genabith, J., & Way, A. (2004). Large-scale induction and evaluation of lexical resources from the Penn-II Treebank. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 367-375). Barcelone.
Otero, P. G. (2008). Comparing window and syntax based strategies for semantic extraction. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language* (p. 41–50). Aveiro.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.
Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Corpora: Pragmatics and Discourse*, *68*, 247-269.
Pecman, M. (2004a). Exploitation de la phraséologie scientifique pour les besoins de l'apprentissage des langues. In *Actes des journées d'étude de l'ATALA, Traitement Automatique des Langues et Apprentissage des Langues* (p. 145–154). Nice.
Pecman, M. (2004b). *Phraséologie contrastive anglais-français: analyse et traitement en vue de l'aide à la rédaction scientifique* (Thèse de Doctorat). Université de Nice-Sophia Antipolis, Nice.
Pecman, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée*, *XII*(2), 79–96.
Phal, A., & Beis, L. (1972). *Vocabulaire général d'orientation scientifique, VGOS: part du lexique commun dans l'expression scientifique...* Paris: Didier.
Poibeau, T., & Messiant, C. (2008). Do we still need gold standard for evaluation ? In *Proceedings of LREC 2008* (p. 1-6). Marrakech.
Polguère, A. (2003). Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement automatique des langues*, *44*(2), 39–68.
Poudat, C. (2006). *Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres* (Thèse de Doctorat). Université de Bretagne Sud, Lorient.
Pouliot, K. (2012). *Élaboration d'un modèle d'article de dictionnaire de collocations du lexique scientifique transdisciplinaire pour l'aide à la rédaction de textes scientifiques* (Thèse de Doctorat). Université d'Orléans, Orléans.
Priss, U., & Old, L. J. (2004). Modelling lexical databases with formal concept analysis. *Journal of Universal Computer Science*, *10*(8), 967–984.
Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). Fouille de données pour la stylistique: cas des motifs séquentiels émergents. In *Actes de JADT 2012* (p. 821–833). Liège.
Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL Semantic Analysis System. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks* (p. 7-12). Lisbonne.
Rayson, P., Archer, D., Piao, S., & McEnery, T. (2008). From key words to key semantic domains. In *International Journal of Corpus Linguistics* (Vol. 13, p. 519–549).
Rinck, F. (2006). *L'article de recherche en Sciences du langage et en Lettres. Figure de l'auteur et identité disciplinaire du genre* (Thèse de doctorat). Université Grenoble Alpes, Grenoble.
Rinck, F. (2010). L'analyse linguistique des enjeux de connaissance dans le discours scientifique. *Revue d'anthropologie des connaissances*, *4*(3), 427-450.
Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.
Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceeding of CICLing2002* (p. 1–15). Mexico.
Sagot, B. (2013). Construction de ressources lexicales pour le traitement automatique des langues. In N. Gala & M. Zock (Éd.), *Ressources Lexicales – Contenu, construction, utilisation, évaluation* (Vol. 30, p. 217-254).

Amsterdam/Philadelphia: John Benjamins.
Sagot, B., & Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008*. Avignon.
Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue française de linguistique appliquée*, *XII*(2), 97–108.
Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: Mouton de Gruyter.
Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon* (Ph.D. thesis). University of Pennsylvania, Philadelphia.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education* (Vol. 22). Amsterdam/Philadelphia: John Benjamins.
Siepmann, D. (2007). Les marqueurs de discours polylexicaux en français scientifique. *Revue française de linguistique appliquée*, *XII*(2), 123–136.
Siepmann, D. (2016). Lexicologie et phraséologie du roman contemporain: quelques pistes pour le français et l'anglais. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*, (108), 21–41.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Szathmary, L., & Napoli, A. (2004). Les treillis de Galois pour l'organisation et la gestion des connaissances. In *Actes des 11emes Rencontres de la Société Francophone de Classification* (p. 153–164). Bordeaux.
Tanguy, L., & Fabre, C. (2014). Évolutions de la linguistique outillée : méfaits et bienfaits du TAL. *L'information grammaticale*, (142), 15-23.
Tanguy, L., Sajous, F., & Hathout, N. (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement Automatique des Langues*, *56*(2), 103-127.
Teufel, S. (1998). Meta-discourse markers and problem-structuring in scientific articles. In *Proceedings of the Workshop on Discourse Relations and Discourse Markers at the 17th International Conference on Computational Linguistics* (p. 43–49). Haifa.
Thoiron, P., & Béjoint, H. (1989). Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle. *Meta: Journal des traducteursMeta:/Translators' Journal*, *34*(4), 661–671.
Tiedemann, J., & Nygaard, L. (2004). The OPUS corpus - parallel & free. In *Proceedings of LREC 2004* (p. 26-28). Lisbonne.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
Toussaint, Y. (2011). *Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances* (Habilitation à diriger des recherches). Université Henri Poincaré - Nancy I, Nancy.
Toussaint, Y., Simon, A., & Cherfi, H. (2000). Apport de la fouille de données textuelles pour l'analyse de l'information. In *Actes de IC 2000* (p. 335–344). Toulouse.

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, *5*(1), 19–50.

Tran, T. T. H. (2014). Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs (Thèse de Doctorat). Université Grenoble Alpes, Grenoble.

Tran, T. T. H., & Hatier, S. (2015). *L'annotation sémantique pour l'enseignement/apprentissage des marqueurs polylexicaux à fonction métadiscursive à l'appui du corpus*. Communication orale à un colloque international présenté à JLC 2015, Orléans.

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

Tutin, A. (2007b). Modélisation linguistique et annotation des collocations : une application au lexique transdiciplinaire des écrits scientifiques. In S. Koeva, D. Maurel, & M. Silberztein (Éd.), *Formaliser les langues avec l'ordinateur : de INTEX à Nooj* (Vol. 8, p. 189–215). Besançon: Presses Universitaires de Franche Comté.

Tutin, A. (2007c). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. In *Actes de TALN 2007*. Toulouse.

Tutin, A. (2008). Sémantique lexicale et corpus: l'étude du lexique transdisciplinaire des écrits scientifiques. *Lublin Studies in Modern Languages and Literature*, (32), 242–260.

Tutin, A. (2010). Evaluative adjectives in academic writing in the humanities and social sciences. In R. Lores-Duenas, P. Mur-Duenas, & E. Lafuente-Milan (Éd.), *Constructing Interpersonality: Multiple Perspectives on Written Academic Genres* (p. 219–239). Cambridge: Cambridge Scholars Publishing.

Tutin, A. (2011). Dans cet article, nous souhaitons montrer que…Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines. *Lidil. Revue de linguistique et de didactique des langues*, (41), 15-40.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques. In A. Tutin & F. Grossmann (Éd.), *L'écrit scientifique : du lexique au discours. Autour de Scientext* (p. 27-44). Rennes: Presses Universitaires de Rennes.

Tutin, A. (2015). Surprise routines in scientific writing: A study of French social science articles. *Review of Cognitive Linguistics, 13(2)*, p. 415-435.

Tutin, A., & Grossmann, F. (Éd.). (2014). *L'Écrit scientifique : du lexique au discours. Autour de Scientext*. Rennes, France: Presse universitaire de Rennes.

Tutin, A., Grossmann, F., Falaise, A., & Kraif, O. (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. In *Actes de JLC 2009* (p. 333-349). Lorient.

Tutin, A., & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines: l'apport des arbres lexico-syntaxiques récurrents. *Lidil. Revue de linguistique et de didactique des langues*, (53), 119–141.

Tutin, A., Tran, T. T. H., Kraif, O., & Hatier, S. (2015). French collocations of cross-disciplinary scientific lexicon. Présenté à PARSEME 5th general meeting (COST action), Iasi.

Valette, M. (2008). À quoi servent les lexiques sémantiques généralistes ? Discussion et proposition. In M. Constant, A. Dister, L. Emirkanian, & S. Piron (Éd.), *Description linguistique pour le traitement automatique du français* (p. 43-58). Louvain-la-Neuve: Presses universitaires de Louvain.

Van den Eynde, K., & Mertens, P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, *13*(1), 63–104.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the 1997 DELOS Workshop on Cross-language Information Retrieval*. Zurich.

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans.

Yan, R. (2012). *Observation du lexique verbal transdisciplinaire et modélisation des patrons dans l'écrit scientifique : construction d'un dictionnaire électronique d'apprentissage d'encodage*. Mémoire de Master 2, Université Grenoble Alpes, Grenoble.

Yan, R. (2014). Modélisations des patrons lexico-syntaxiques dans le corpus Scientext : exemple des verbes d'opinion. In *Actes de CEDIL 2014*. Grenoble.

Yan, R., & Hatier, S. (2016). L'extraction et la modélisation de patrons lexico-syntaxiques pour leur enseignement en FLE : un exemple à partir du verbe montrer. *Linguistik Online*, *78*(4), 93-112.

Yan, R., & Tutin, A. (2013). Un modèle lexicographique basé sur le corpus Scientext comparé avec d'autres ressources lexicographiques : l'exemple des constructions verbales. In *Actes de JLC 2013*. Lorient.

Annexes

Sommaire des annexes

A.I Post-traitements de l'analyseur XIP

Nous présentons dans cette section le fonctionnement des règles de post-traitements de l'analyseur *XIP*. Nous avons défini des règles pour adapter les sorties de l'analyse syntaxique à nos besoins, comme détaillée dans la section 2.2.1.3.2. Ces règles se construisent comme suit :

- Des configurations (ensemble de tokens et de dépendances) particulières sont énoncées. Par exemple, la configuration suivante (*obj*(#2[lemme : être],#3) & *subj*(#2,#1)) signifie qu'un token #1 est sujet d'un token #2, prenant lui-même pour objet un token #3.
- Certains attributs (lemme, trait morpho-syntaxique, catégorie, etc.) des tokens concernés peuvent être précisés. Dans l'exemple précédent, le token #2 a pour lemme *être*.
- Des « résultats d'applications » sont définis. Ils correspondent généralement, pour nos grammaires, à une suppression, à une modification ou à un ajout de dépendances. Dans le cas de l'exemple en cours, le résultat sera l'ajout d'une nouvelle dépendance du type 'attribut profond du sujet' entre le token #1 et le token #3. Dans le formalisme utilisé pour la définition de ces règles, le résultat d'application se traduit comme suit : *u3_deepattrsubj*(#1,#3). Ainsi, si l'analyseur rencontre les configurations données, une relation de dépendance de type « attribut du sujet profond » sera ajoutée entre le token #1 (gouverneur) et le token #2 (dépendant). Une telle relation est ainsi ajoutée entre *usines* et *propriétés* dans l'exemple suivant :
    - *La plupart des usines sont des propriétés familiales, associant les plantations, l'usine et/ou la distillerie*.[1]

[1] Grégoire E., Théry H., (2007). L'Ogre et le Petit Poucet. Le Brésil et l'île Maurice dans le négoce mondial du sucre. *L'Espace géographique* 3/2007 (Tome 36), p. 267-282.

A.II Noms du LST

À la suite de l'extraction automatique des noms du LST et de leur évaluation manuelle par 5 juges, nous aboutissons à une liste de 493-noms du LST :

absence accroissement acquisition acteur activité adoption affirmation alternative ambiguïté amélioration ampleur analyse angle année annexe apparence apparition appartenance application apport appréciation appréhension approche appropriation aptitude argument article articulation aspect association attribution augmentation auteur avancée avantage axe baisse base biais bibliographie bilan but cadre calcul capacité caractère caractéristique cas catégorie catégorisation cause centre champ changement chapitre chercheur chiffre choix cible circonstance classe classement classification clivage code cohérence collaboration collectif colloque combinaison commentaire communauté communication comparaison compétence complexité comportement composante composition compréhension compromis concentration concept conception conclusion condition conférence configuration confrontation connaissance consensus conséquence considération constat constitution construction contenu contexte continuité contour contradiction contrainte contraste contribution convention convergence corpus corrélation correspondance couple création critère critique croissance cycle date débat décision déclin découverte défaut définition degré démarche démonstration dépendance déplacement description détermination développement différence différenciation difficulté diffusion dimension diminution direction discipline discours discussion dispositif distance distinction distribution divergence diversité division document documentation domaine donnée durée dynamique écart échantillon échec échelle effectif effet égalité élaboration élément émergence enjeu enquête ensemble entité entretien environnement épreuve équilibre équivalent erreur espace essai essentiel estimation étape état étude évaluation événement évolution examen exception exemple exigence existence expansion expérience expérimentation expert expertise explication expression extension extérieur extrait façon facteur faculté faiblesse fait figure finalité flux fonction fonctionnement fondateur fondement forme formulation formule fréquence frontière gain généralisation genèse genre groupe hasard hiérarchie horizon hypothèse idée identification identité idéologie illustration image impact implication importance impossibilité incertitude incidence indépendance indicateur indication indice individu influence information instance institut instrument intégration intensité interaction intérêt intermédiaire interprétation interrogation introduction inverse investigation jugement justification laboratoire lecture liaison lien lieu limite littérature localisation logique loi maîtrise majorité manière manifestation manipulation manque marque matériau matériel maximum mécanisme membre mérite mesure méthode méthodologie milieu minimum minorité modalité mode modèle modification mot motif mouvement moyen moyenne multiplication mutation nature nécessité niveau nombre norme note notion objectif objet observateur observation obstacle opération opinion opposition option ordre organisation organisme orientation originalité origine outil ouvrage pair paradigme paradoxe paramètre participant particularité partie perception performance période perspective perte pertinence phase phénomène plan poids point point de vue pôle population portée position positionnement possibilité posture pourcentage pratique précision préférence préoccupation préparation présence présentation principe priorité problématique problème procédé procédure processus production profil programme progrès progression projet proportion proposition propriété protagoniste proximité publication qualité quantité question questionnaire questionnement

raisonnement rang rapport rapprochement réalisation recherche reconnaissance recours recueil réduction réel référence réflexion règle rejet relation remarque renforcement répartition repère réponse représentation réseau résolution résultat rigueur rôle rupture savoir schéma science secteur section segment sélection sens séquence série seuil siècle signe signification singularité situation solution somme souci source spécialiste spécificité sphère stabilité statistique statut stratégie structuration structure succession suite sujet support symbole synthèse système tableau tâche taille taux technique technologie témoignage tendance tentative terme terrain territoire test texte thématique thème théorie thèse tiers total totalité tradition traduction trait traitement transfert transformation transition transmission travail type typologie unité univers université usage utilisation utilité valeur validité variable variation variété version vision vocabulaire voie volume zone

A.III Adjectifs du LST

Les méthodes d'extraction automatique du LST suivie de la phase de validation manuelle des adjectifs candidats au statut LST nous permette d'extraire et d'identifier 274 adjectifs du LST :

abondant absent abstrait accru actuel adéquat aléatoire alternatif ambigu analogue analytique annuel antérieur apparent approfondi approprié apte attendu attentif automatique autonome bibliographique bref brut caractéristique central certain classique cohérent collectif commun comparable comparatif compatible complémentaire complet complexe conceptuel concret conforme consacré considérable constant constitué constitutif contemporain contradictoire contraire correct correspondant courant court critique croissant crucial décisif défavorable définitif dépendant descriptif détaillé déterminant déterminé différencié différent difficile direct disciplinaire disponible distinct divergent divers dominant donné dynamique effectif égal élaboré élevé emblématique empirique envisageable épistémologique équivalent essentiel établi étroit éventuel évident exclusif exemplaire exhaustif existant expérimental explicatif explicite extérieur externe extrême faible familier favorable fiable fin final fixe flou fonctionnel fondamental fondateur formel fort fréquent futur général générique global grand hétérogène hiérarchique homogène horizontal humain idéal identique immédiat implicite important incertain incompatible incontournable indépendant indirect indispensable indissociable individuel inédit inégal inférieur informel inhérent initial instrumental insuffisant intense intéressant interne interprétatif intrinsèque inverse isolé large lent lié limité linéaire logique majeur majoritaire marginal marqué massif matériel maximal mécanique même méthodologique minimal moyen multiple nécessaire négatif négligeable net normatif notable nouveau objectif observable opérationnel opératoire opposé ordinaire organisationnel original paradoxal parallèle particulier partiel pertinent porteur posé positif possible pragmatique pratique préalable précédent précis prégnant premier première prépondérant présent primordial principal privilégié probable problématique proche profond progressif propice propre qualitatif quantitatif radical rapide rare récent réciproque reconnu récurrent régulier relatif relationnel remarquable représentatif respectif restrictif révélateur rigoureux satisfaisant scientifique second semblable séparé significatif similaire simple simultané singulier sous-jacent spécifique stable statistique stratégique strict structurel subjectif substantiel successif suffisant suivant supplémentaire supposé surprenant susceptible symbolique systématique temporel théorique total traditionnel typique ultérieur utile valable variable varié vaste véritable vertical voisin

A.IV Verbes du LST

Les méthodes d'extraction automatique du LST suivie de la phase de validation manuelle des verbes candidats au statut LST nous amène à valider une liste de 343-verbes du LST :

aborder aboutir accéder accentuer accompagner accorder accroître adapter admettre adopter advenir affecter affiner affirmer affranchir agir ajuster alimenter améliorer amener amorcer amplifier analyser ancrer apparaître apparenter appartenir appliquer apporter appréhender approfondir appuyer articuler assigner assimiler associer assumer attacher atteindre atténuer attester attribuer augmenter autoriser avérer baser calculer caractériser centrer cerner choisir citer classer coïncider collecter combiner comparer compenser compléter comporter composer concentrer concerner concevoir conclure conditionner conduire conférer confirmer confondre conformer conforter confronter conjuguer consacrer conserver considérer consister constater constituer construire consulter contenir contester contraindre contraster contredire contribuer contrôler convenir converger corréler correspondre corriger critiquer débattre déboucher décliner découler décrire déduire définir dégager délimiter démarquer demeurer démontrer dépasser dépendre désigner dessiner détailler déterminer développer devenir différencier différer diffuser diminuer discuter disperser disposer dissocier distinguer distribuer diviser doter éclairer effectuer efforcer élaborer élargir élever émerger employer emprunter engendrer englober énoncer enrichir entraîner entretenir envisager équivaloir esquisser estimer établir étendre étudier évaluer évoluer évoquer examiner exclure exercer exiger exister expliciter expliquer exploiter explorer exposer exprimer faciliter favoriser figurer focaliser fonctionner fonder formaliser former formuler fournir garantir généraliser générer guider hériter heurter hiérarchiser identifier illustrer impliquer importer imposer inciter inclure indiquer induire influencer influer inscrire insister inspirer instituer intégrer intéresser interpréter interroger intervenir introduire inverser invoquer isoler juger justifier lier limiter localiser maintenir manifester marquer masquer médiatiser mener mentionner mesurer mettre mobiliser modifier montrer motiver multiplier nécessiter négliger nommer noter nuancer objectiver observer obtenir occulter offrir opérer opposer organiser paraître partager participer permettre placer porter posséder postuler pratiquer précéder préciser préconiser présenter prétendre privilégier procéder produire progresser prolonger prôner proposer provenir provoquer publier qualifier questionner raisonner rapprocher rassembler rattacher réaliser recenser rechercher reconnaître reconstituer reconstruire recourir recouvrir recueillir redéfinir réduire référer refléter regrouper rejeter relater relativiser relever relier remplacer renforcer renouveler renvoyer répartir repérer replacer reporter reposer représenter reproduire résider résoudre respecter ressortir restituer restreindre résulter résumer retenir retracer révéler revendiquer revêtir saisir satisfaire sélectionner sembler séparer signaler signifier simplifier situer souligner soumettre sous-tendre soutenir spécifier structurer subsister substituer succéder suggérer supposer symboliser témoigner tendre tenter tester traduire traiter transformer utiliser valider varier véhiculer vérifier viser

A.V Adverbes du LST

Pour la catégorie des adverbes, nous n'avons pas eu recours à des juges et avons seulement effectué un filtrage des erreurs de l'annotation automatique (erreurs de lemmatisation, de catégorisation) pour valider une liste de 202 adverbes du LST. Comme le lexique de l'analyseur intègre des adverbiaux polylexicaux, nous en repérons un certain nombre. Le statut de ces polylexicaux est différent de celui des mots simples du LST dans la mesure où il est se révéler ambigu au niveau de la segmentation. Ainsi, la séquence *dans ce cadre* peut effectivement correspondre à un adverbe polylexical mais peut également correspondre à plusieurs unités, dont des éléments terminologiques, comme dans la séquence *dans ce cadre de sous-catégorisation*. Ces adverbiaux ont cependant tout leur intérêt dans la ressource, tant au niveau des applications didactiques que des applications pour la détection de routines. Les 202 adverbes du LST sont les suivants :

à cette époque, à l'origine, à la fois, à long terme, à plusieurs reprises, a priori, à terme, actuellement, ailleurs, ainsi, aisément, au contraire, au départ, au mieux, au moins, au plus, au préalable, au-delà, auparavant, aussi, autant, autrement, autrement dit, avant tout, bien entendu, bien plus, bref, brièvement, ce faisant, cependant, certes, ci-dessous, ci-dessus, clairement, collectivement, concrètement, considérablement, constamment, couramment, d'abord, d'ailleurs, d'autant plus, d'autre part, d'emblée, d'une part, dans ce cadre, dans ce cas, dans ces conditions, dans son ensemble, dans un premier temps, davantage, de même, de plus, de plus en plus, définitivement, dès lors, désormais, deuxièmement, différemment, difficilement, directement, donc, du moins, du reste, effectivement, également, en ce sens, en commun, en conséquence, en d'autres termes, en définitive, en effet, en général, en l'occurrence, en même temps, en moyenne, en outre, en particulier, en partie, en pratique, en quelque sorte, en réalité, en revanche, en rien, en somme, enfin, ensuite, entièrement, environ, essentiellement, etc, etc., étroitement, éventuellement, évidemment, exclusivement, explicitement, extrêmement, facilement, faiblement, finalement, fondamentalement, forcément, formellement, fortement, fréquemment, généralement, globalement, guère, habituellement, ibid, ibid., implicitement, indirectement, infra, initialement, inversement, jusqu'alors, jusque-là, la plupart du temps, largement, librement, localement, majoritairement, massivement, moins, néanmoins., nécessairement, nettement, nom, non seulement, notamment, nullement, où, par ailleurs, par conséquent, par contre, par exemple, par la suite, paradoxalement, parallèlement, parfois, partant, particulièrement, partiellement, peu, pleinement, plus, plus ou moins, plutôt, positivement, potentiellement, pour autant, pour ce faire, pour la plupart, pourtant, préalablement, précédemment, précisément, premièrement, principalement, profondément, progressivement, proprement, purement, quasi, quasiment, quelque peu, radicalement, rapidement, rarement, récemment, réellement, régulièrement, relativement, respectivement, sans doute, sensiblement, seulement, significativement, simultanément,

socialement, souvent, spécialement, spécifiquement, strictement, successivement, suffisamment, surtout, systématiquement, tantôt, tardivement, totalement, tout à fait, tout d'abord, toutefois, traditionnellement, uniquement, véritablement, voire, volontairement, volontiers, vraisemblablement

A.VI Extrait du LST nominal enrichi sémantiquement

Le tableau ci-dessous présente un aperçu des informations sémantiques présentes dans notre ressource que nous avons croisée avec le *DEM* pour un premier inventaire des acceptions du LST nominal. Les colonnes issues du *DEM* sont :

- *M, C,26* : mot d'entrée, avec numérotation dans les cas de polysémie ;
- *CONT, C,10* : ébauche de phrase simple ou prototypique pour le mot d'entrée (information syntagmatique et paradigmatique ; plusieurs mots peuvent partager une même construction) ;
- *DOM, C,4* : domaine d'application du mot d'entrée ;
- *OP, C,4* : opérateur, classe d'objet correspondant à un hyperonyme ;
- *SENS, C,23* : définition ou synonyme du mot d'entrée ;
- *OP1, C,4* : classes de verbes (du *LVF*) avec lesquels le nom se combine prototypiquement.

La colonne *acception* correspond à l'étiquette arbitraire que nous associons à l'entrée en question afin de distinguer les entrées pour les éléments polysémiques du LST. Les colonnes *classe* et *sous-classe* correspondent aux résultats de la classification manuelle présentée dans la section 3.3.2.

acception	M, C,26	CONT, C,10	DOM, C,4	OP, C,4	SENS, C,23	OP1, C,4	classe	sous-classe
comparaison	comparaison 01	rli qc p N	RLA	tech	parallèle	U3a1	processus_cognitif	inclusion_séparation
compétence	compétence 01	f preuve N	PSY	car	connaissance, expérience	H2a1	état_qualité	axiologique_positif
complexité	complexité 01	rli qc p N	PHI	état	complexe, compliqué	U3a1	état_qualité	complexité
composante	composante	val x p N	MAT	élém	facteur	H3f1	objet_scientifique	explicatif_simple
composition_1	composition U3-01				nature		état_qualité	composition
composition_2	composition 02	rli qc p N	TEC	tech	assemblage, format	U3a1	processus_cognitif	réalisation
compréhension	compréhension 01	f preuve N	PSY	car	aptitude à comprendre	H2a1	processus_cognitif	compréhension
compromis	compromis	rli qc p N	DRO	tech	transaction, accord	U3a1	relation	conformité
concentration	concentration 01	rli qc p N	QUA	tech	réunion d'éléments	U3a1	état_qualité	grandeur
concept	concept				idée abstraite et générale		objet_scientifique	explicatif_simple
conception_2	conception				représentation globale		objet_scientifique	thèse
conception_1	conception 01	f preuve N	PSY	car	d concevoir, projet	H2a1	processus_cognitif	réalisation
conclusion_1	conclusion 03	écrire N	LIT	text	ce qui termine un ouvrage littéraire (tlfi)	R4a1	communication_support	section
conclusion_2	conclusion 02	f preuve N	PHI	car	déduction, conséquence	H2a1	objet_scientifique	raisonnement
condition_1	condition 02	rli qc p N	SOC	tech	circonstance	U3a1	objet_scientifique	cas
condition_2	condition 03	rli qc p N	PHI	tech	ce q ê requis pr qc	U3a1	relation	implication
conférence	conférence 02	dire p N	LOQ	dit	exposé public sr qc	C2a1	communication_support	exposé
configuration	configuration 02	rli qc p N	TEC	tech	aspect, forme, structure	U3a1	état_qualité	composition
confrontation	confrontation 01	rli qc p N	RLA	grp	comparaison d qc différ	U3a1	processus_cognitif	inclusion_séparation
consensus	consensus	dire p N	LOQ	dit	assentiment de ts	C2a1	relation	conformité
conséquence	conséquence 01	rli qc p N	RLA	tech	suite logique, résultat	U3a1	relation	implication
considération	considération 01	f preu p N	PHI	car	examen attentif	H2a1	état_humain	attitude
constat	constat 02	f preu p N	PSY	car	constatation	H2a1	processus_cognitif	constat
constitution	constitution 01	rli qc p N	TEC	tech	structuration	U3a1	processus_cognitif	réalisation

Tableau 1: Extrait du lexique sémantique des noms

A.VII Tableau de la classification sémantique des noms du LST

Le tableau ci-dessous détaille la classification en classes et sous-classes sémantiques pour les 531 acceptions nominales du LST. La colonne *Noms – Acceptions Termith* correspond à la colonne *acception* du tableau 1.

Les colonnes *Classe* et *Sous-Classe* sont composées des éléments suivants :

- de l'étiquette de classe : **communication_support** ;
- d'une définition de la classe : 'support de communication et de transfert des idées' *;*
- d'un patron lexico-syntaxique qui permet de tester l'appartenance à la classe : *Le N est consacré à*.

Classe	Sous-Classe	Noms – Acceptions Termith
collectif_partitif 'ensemble ou élément d'un ensemble' *diviser en plusieurs N*	**collectif** 'ensemble d'éléments' *regrouper dans un N / Plusieurs SN constituent un N*	catégorie, classe_1, ensemble_2, groupe_2, population, série_2, totalité_2
	partitif 'élément, portion d'un tout' *être composé de N / extraire N d'un tout*	échantillon, élément, extrait, partie_2, section_2, segment, unité_2
communication_expression : 'acte de communication et de transfert des idées (selon le type de message)' *Par/dans ce N, l'auteur précise*	**discussion** 'acte de communication en réponse à un autre acte/événement' *Ce N répond à…* / Ceci appelle un *N*	commentaire, débat, discussion, remarque, témoignage
	formulation 'élément d'un énoncé' *Dans l'article, l'auteur utilise/énonce un N*	affirmation_1, définition_1, discours expression_1, formulation, formule_1, mot, proposition, terme_1, vocabulaire
communication_support : 'support de communication et de transfert des idées (selon le moyen de transmission)' *Le N est consacré à*	**document** 'document écrit' *Lire, publier un N*	article, document, documentation, essai_1, littérature, note, ouvrage, publication, synthèse_1 texte, thèse_1, volume_1
	exposé 'événement scientifique' *un N a lieu / Pendant un N*	colloque, conférence
	graphique 'représentation graphique éclairant le texte' *Le N {numéral} montre / cf*	figure_1, illustration_1, image_1, motif_1, tableau

Classe	Sous-Classe	Noms – Acceptions Termith
	N, sur N, dans N	
	section ‘sous-partie d’un document écrit’ *Dans le N de l’ouvrage / cf N*	annexe, bibliographie, chapitre, conclusion_1, développement_1, introduction_1, section_1
déterminant ‘mot élément d’un déterminant complexe’ *un N de SN <=> un SN*	**qualifiant** ‘renvoie à une instance particulière du SN déterminé‘ *un N de SN*	genre_1, type_1
	quantifiant ‘renvoie à un ensemble ou sous-ensemble du SN déterminé‘ *un N de SN*	couple, diversité_1, ensemble_1, groupe_1, partie_1, totalité_1
	suite ‘suite ordonnée d’éléments’ *un N de plusieurs SN*	cycle_1, séquence, série_1, succession, suite
espace ‘espace concret ou milieu abstrait’ *Se situer dans/à un N*	**domaine** ‘zone, partie d’un espace, d’un milieu’ *Le N s’étend / Dans le N de*	cadre, champ, contexte, dimension_2 domaine, environnement, espace, lieu, milieu, pôle_1, secteur, sphère, terrain, territoire, univers, zone
	limite ‘limites définissant un espace donné’ *Le N délimite cet espace*	contour, frontière, limite
	localisation ‘place déterminée dans un espace’ *Localiser le N / Au N de*	centre, extérieur, niveau, plan_1, pôle_2, position, rang, source
	orientation ‘sens, direction dans un espace donné’ *aller vers un N / s’orienter selon un N*	axe_1, direction, orientation, sens_1, voie_1
état_humain ‘disposition dans laquelle se trouve un humain‘ expérienceur humain *Un humain a de N pour SN / Le N d’un humain envers SN*	**attitude** ‘attitude intérieure‘ expérienceur humain *Un humain a de N pour SN / Le N d’un humain envers SN*	considération, intérêt_2, préférence, préoccupation
état_qualité ‘état ou qualité d’une entité abstraite ou concrète’ *avoir un N / être de N*	**axiologique_négatif** ‘Jugement négatif de valeur sur la qualité/état’ *Qqc souffre de un N*	biais, défaut_1, erreur, faiblesse, manque
	axiologique_positif ‘Jugement positif de valeur sur la qualité/état’ *Qqc bénéficie de un N*	apport, avancée, avantage, cohérence, compétence, contribution, intérêt_1, maîtrise, mérite, pertinence, précision, rigueur, utilité, validité
	caractéristique ‘manière dont une personne ou une chose se présente à la vue ou à l’esprit’ *un N propre de qqc / SN*	apparence, aspect, caractère, caractéristique, état_1, forme, genre_2, nature, particularité, profil,

Classe	Sous-Classe	Noms – Acceptions Termith
	présente un N propre	propriété, qualité_1, tendance, trait
	certitude ‘évaluation des chances de réalisation d’un élément’ *Un événement a 1 degré de N*	impossibilité, incertitude, possibilité
	changement_négatif ‘absence de variation ou de modification’ *assurer un N dans le temps*	continuité, stabilité
	changement_positif ‘présence de variation, modification’ *un N débute*	rupture
	complexité ‘degré de difficulté d’analyse ou de compréhension d’une entité’ *présenter un dégré de N / rencontrer un N*	ambiguïté, complexité, difficulté, obstacle, problème_1
	composition ‘nature d’une entité en tant qu’unité’ *une entité est caractérisée par DétPoss N*	composition_1, configuration, distribution_1, diversité_2, organisation, répartition, structuration_1, structure, unité_1, variété, version
	évaluation_capacité ‘qualité d’une entité à évaluer’ *évaluer un N / développer un N*	aptitude, capacité_1, faculté_1, performance, qualité_2, valeur
	grandeur 'qualité mesurable, renvoyant à la taille, la force...' *un N total / limiter un N*	ampleur, capacité_2, concentration, dimension_1, effectif, fréquence, intensité, quantité, seuil, taille, volume_2
	importance ‘degré d’importance’ *conférer un N*	fonction, importance, poids, portée_2, priorité, rôle, statut
	manière ‘aspect particulier d’un processus ou état’ *Procéder/traiter à/selon un N*	façon, manière, modalité, mode
	nécessité ‘évaluation de l’obligation la nécessité’ *Répondre à un N*	exigence, nécessité
	nouveauté_positif 'degré d’originalité, du caractère courant, habituel (négatif) ou original, hors du commun (positif)' *un N réside dans*	originalité, singularité, spécificité

Classe	Sous-Classe	Noms – Acceptions Termith
objet_scientifique 'observables et objets construits de l'activité scientifique, *étudier/collecter/s'appuyer sur un N*	**cas** 'situation, cas' *rencontrer un N / un N survient*	alternative, cas, circonstance, condition_1, échec, événement, exception, fait, option, phénomène, situation
	donnée_observable 'donnée observable' *analyser/collecter un N*	corpus, donnée, entité, information, matériau, objet_1
	explicatif_complexe 'élément structurant, organisant plusieurs éléments par des relations, utilisé pour l'explication' *Expliquer qqc dans le cadre de N*	code, convention, loi, modèle_1, norme, paradigme, principe, règle, réseau, schéma, typologie
	explicatif_simple 'élément servant de base à une explication, un jugement' *Plusieurs N expliquent ceci*	composante, concept, contenu, critère, exemple, facteur, indicateur, notion, paramètre, sens_2, signification, variable
	finalité 'but déterminé d'une action' *Nous visons tel N*	but, cible, finalité, objectif, objet_2, projet_1
	instrument 'instrument, outil, moyen utilisé pour exécuter quelque chose' *à l'aide de N / utiliser N*	enquête, entretien, instrument, matériel, moyen, outil, questionnaire, support, technique, technologie
	méthode 'ensemble de moyens mis en place pour effectuer une recherche scientifique' *s'appuyer sur/proposer un N*	approche, démarche, dispositif, logique, méthode, méthodologie, perspective, plan_2, procédé, procédure, programme, stratégie, système, théorie
	objet_mathématiques 'objet du domaine des mathématiques' *un N mathématique*	axe_2, formule_2
	raisonnement 'élément dans un processus de raisonnement' *Proposer un N convaincant*	argument, conclusion_2, réponse, solution
	représentation 'représentation d'une entité par une autre' *X est considéré comme un N (de|pour) Y*	équivalent, figure_2, image_2, indication, indice_1, instance, marque_2, modèle_2, référence, repère, représentation, signe, symbole, traduction, type_2
	savoir 'relatif au domaine de la connaissance, des sciences' *enseigner/acquérir un N*	connaissance, discipline, recherche_1, savoir, science
	thème 'thème, sujet abordé'	enjeu, point_1, problématique, problème_2,

Classe	Sous-Classe	Noms – Acceptions Termith
	L'article, l'auteur aborde/soulève un N	projet_2, question, sujet, thématique, thème
	thèse 'Position intellectuelle, manière de voir les choses' *étudier un objet selon un N / adopter, soutenir un N*	angle, conception_2, hypothèse, idée, idéologie, opinion, positionnement, posture, thèse_2, vision, voie_2
personne 'personne ou groupe de personnes' *+humain*	**collectivité** 'groupe de personnes' *+pluriel*	classe_2, collectif, communauté, faculté_2, institut, laboratoire, université
	individu 'personne ' *+singulier*	auteur, chercheur, expert, individu, membre, observateur
processus_cognitif 'processus cognitif permettant de structurer/analyser les observables et/ou objets construits' *une personne fait/effectue/propose un N*	**attribution** 'processus par lequel on donne qqc, une propriété, à un élément' *un N de propriétés, d'élément, à un SN*	attribution
	choix 'processus de choix, sélection de qqc' *un N se justifie/est guidé par*	adoption, appropriation, choix, décision, rejet, sélection
	classement 'processus de mise en structure d'un ensemble d'entités afin d'en faciliter l'étude' *établir un N de plusieurs éléments*	catégorisation, classement, classification, structuration_2
	compréhension 'entendement, fait de comprendre un problème' *le N d'un problème*	appréhension, compréhension, résolution
	constat 'phase dans l'analyse, d'identification d'entités/ de phénomènes' *un N montre que / un N visuel*	bilan, constat, découverte, état_2, identification, localisation, observation, perception, reconnaissance
	démonstration 'processus de mise en évidence de qqc' *Un phénomène est éclairé par un N de SN*	démonstration
	description 'processus de présentation d'un élément' *proposer un N détaillé*	description, illustration_2, présentation
	détermination 'processus de délimitation des propriétés de qqc' *un N repose sur SN*	définition_2, détermination
	évaluation 'phase dans l'analyse, d'évaluation d'entités/ de phénomènes'	appréciation, estimation, évaluation, mesure

Classe	Sous-Classe	Noms – Acceptions Termith
	un N (subjectif|objectif) se fonde sur/ une méthode de N	
	examen 'observation attentive dans un but d'explication' *fournir|conduire DétPoss N (détaillé|approfondi| systématique)*	analyse, étude, examen, expertise
	inclusion_séparation 'action de chercher des ressemblances, dissemblances entre différents éléments' *opérer|établir un N entre plusieurs éléments*	comparaison, confrontation, différenciation, distinction, division, intégration, introduction_2, rapprochement
	interprétation 'explication, lecture d'une chose en fonction des faits *proposer/donner DétPoss N*	explication_1, généralisation_1, interprétation, justification, lecture, synthèse_2
	méthodologie 'processus typique de la méthodologie de l'activité scientifique destiné à vérifier une hypothèse, à établir un fait' *Le chercheur effectue un N* *Le chercheur met en place un N*	calcul, épreuve, essai_2, expérience, expérimentation, investigation, manipulation, recherche_2, tentative, test
	opinion 'réflexion évaluative de qqc' *un humain émet un N*	critique, jugement
	réalisation 'processus de réalisation' *participer à Dét N d'un ensemble / réaliser Dét N*	composition_2, conception_1, constitution, construction, création,

Classe	Sous-Classe	Noms – Acceptions Termith
		élaboration, opération préparation, production, réalisation, recueil
	réflexion **‘**considération attentive d’un fait’ *mener/approfondir DétPoss N*	interrogation, questionnement, raisonnement, réflexion
processus_évolutif ‘processus dénotant une évolution, un changement’ *constater un N depuis des années*	**amélioration_augmentation** ‘processus dénotant une évolution positive’ *conduire à / entraîner Dét N*	accroissement, affirmation_2, amélioration, augmentation, croissance, développement_2, expansion, extension, gain, généralisation_2, multiplication, progrès, progression, renforcement
	baisse ‘processus dénotant une évolution négative’ *compenser un N*	baisse, déclin, diminution, perte, réduction
	changement ‘processus dénotant une	changement, déplacement, diffusion, dynamique,

Classe	Sous-Classe	Noms – Acceptions Termith
	évolution' *un N intervient à/affecte*	`évolution, flux, modification, mouvement, mutation, processus, transfert, transformation, transition, transmission, variation`
	manifestation 'manifestation, émergence d'une entité, d'un phénomène' *un N a commencé à tel moment / Certaines conditions favorisent un N*	`apparition, émergence, expression_2, genèse, manifestation`

Classe	Sous-Classe	Noms – Acceptions Termith
processus_gen_chose ‘processus de fonctionnement d’une entité’ qqc a 1 N <=> le N de qqc	**action_chose** 'fonctionnement de qqc » *expliquer/analyser le N de qqc*	`comportement, fonctionnement, mécanisme`
processus_humain ‘processus impliquant un sujet humain’ *Le N de qqn / Le N entre X+hum et Y+hum*	**collaboration** ‘interaction entre humains’ *Un N entre X+hum et Y+hum*	`collaboration`
	usage ‘processus humain générique’ *Le N de /à SN par qqn*	`acquisition, application, pratique, recours, tradition, usage, utilisation`

Classe	Sous-Classe	Noms – Acceptions Termith
	action_humain 'action de qqun » *qqn assure/accomplit un N*	activité, tâche, traitement, travail
quantité 'renvoie à la notion de quantité, de mesure' *Calculer un N / un N s'élève à*	**mesure** 'grandeur résultant du rapport entre plusieurs valeurs' *calculer/mesurer un N entre X et Y*	différence_1, distance, écart, échelle, égalité, indice_2, majorité, maximum, minimum, minorité, moyenne, pourcentage, proportion, résultat_1, somme, statistique, taux, tiers,

Classe	Sous-Classe	Noms – Acceptions Termith
		`total`
	unité ‘unité de mesure, de calcul’ *ajouter un N au total*	`chiffre_1, degré, nombre, point_2`
relation ‘dénote une relation entre entités’ *Il existe une relation de N entre X et Y*	**appartenance** ‘relation d’inclusion ou exclusion d’une entité par rapport à un ensemble’ *un N d’un élément à un groupe*	`appartenance`

Classe	Sous-Classe	Noms – Acceptions Termith
	association ‘mise en relation de plusieurs entités’ *un N (étroit) entre deux éléments*	articulation, association, combinaison, corrélation, équilibre, indépendance, interaction, liaison, lien, ordre, relation
	conformité **‘**convergence de points de vue**’** *rechercher/dégager un N*	compromis, consensus
	correspondance ‘mise en relation mettant en évidence les ressemblances entre entités’ *un N est établi entre X et Y/ il y a un N entre X et Y*	convergence, correspondance, identité, proximité
	implication ‘mise en relation entre entités consistant en ce	base, cause, condition_2, conséquence, contrainte, dépendance, effet,

Classe	Sous-Classe	Noms – Acceptions Termith
	qu'une chose entraîne comme conséquence logique une autre' *X a/est un N pour Y*	explication_2, fondement, impact, implication, incidence, influence, motif_2, rapport, résultat_2
	opposition 'mise en relation mettant en évidence les différences entre entités' *un N est marqué entre X et Y*	clivage, contradiction, contraste, différence_2, divergence, opposition, paradoxe
temporalité	**chronologie**	date, étape, horizon,

Classe	Sous-Classe	Noms – Acceptions Termith
‘relatif à la dimension temporelle’ *avoir lieu pendant/à un N*	‘localisation sur l’axe temporel’ *a lieu à un N*	`origine, terme_2`
	durée ‘intervalle sur l’axe temporel’ *pendant/à la fin de un N*	`année, cycle_2, durée, période, phase, siècle`
temporalité_espace ’relatif à la dimension spatio-temporelle’ *un N attesté*	**temporalité_espace** ‘relatif à la dimension spatio-temporelle’ *un N attesté*	`absence, existence, présence`

Tableau 2: Classification Sémantique des noms du LST

A.VIII Représentation des propriétés lexico-syntaxiques dans la FCA

Cette section détaille l'ensemble des propriétés lexico-syntaxiques définissant un objet dans le cadre de l'analyse de concepts formels, présentée dans la section 4.6. Nous prenons l'exemple du nom *article* et listons ses propriétés lexico-syntaxiques, issues de l'analyse automatique du corpus arboré. Les propriétés, pour cet exemple, ont été retenues après application des critères suivants :

- Le triplet (relation typée, lemme et catégorie du cooccurrent) doit correspondre à des occurrences d'un minimum de trois disciplines
- Pour chaque relation de dépendance utilisée pour la catégorisation (*nmod, ~nmod, ~subj,* etc.) les cooccurrents ont été ordonnés de façon décroissante par leur score de rapport de vraisemblance. Nous retenons pour chaque relation les 10 cooccurrents en tête de liste.

Les propriétés calculées pour le nom *article* sont ainsi :

- *nmod → publier_verb | paraître_verb | intituler_verb | code_nom | presse_nom | revue_nom | figaro_nom | loi_nom | traité_nom*
- *~nmod → objectif_nom | objet_nom | partie_nom | suite_nom | auteur_nom | version_nom | nombre_nom | lecture_nom | corps_nom*
- *~subj → proposer_verb | montrer_verb | présenter_verb | examiner_verb | étudier_verb | porter_verb | viser_verb | intéresser_verb | interroger_verb| attacher_verb*
- *~obj → publier_verb | voir_verb consacrer_verb | intituler_verb | paraître_verb | citer_verb | lire_verb | présenter_verb | rédiger_verb | considérer_verb*
- *u3_adjmod → présent_adj | fondateur_adj | consacré_adj | récent_adj | précédent_adj | scientifique_adj | nombreux_adj | dernier_adj | même_adj | premier_adj*
- *determ → plusieurs_det | quelque_det | notre_det*

- *prepobj → dans le cadre de_prep | à la fin de_prep | cf._prep |au début de_prep | pour_prep*

A.IX Exemple de contexte relationnel pour la FCA

Cette section présente un exemple de contexte relationnel représentant les attributs de 10 noms du LST. Ce contexte est utilisé en tant qu'entrée dans notre expérimentation sur l'analyse en concepts formels pour générer les treillis de Galois (voir section 4.6).

Certaines étiquettes de classes ont subi des modifications depuis l'expérimentation de classification automatique. Les correspondances suivantes sont ainsi à faire :

- sous-classe {relation/équivalence} → {relation/représentation} ;
- sous-classe {quantité/rapport} → {quantité/mesure}.

Ci-dessous est présenté un exemple de contexte relationnel composé :

- d'une liste d'objets : 10 noms du LST (*schéma, signe, évaluation, interprétation, tableau, annexe, estimation, indice, symbole, figure*) ;
- d'une liste d'attributs : propriétés lexico-syntaxiques formalisées de la manière suivante :
  - *influence_nom#~nmod* : l'objet est dépendant (caractère ~) dans la relation *nmod* avec pour gouverneur le nom *influence* ;
  - *donnée_nom#nmod* : l'objet est gouverneur dans la relation NMOD avec pour dépendant le nom *donnée* ;
- de la matrice associant une valeur binaire (0 ou 1) pour chaque couple objet-attribut.

```
[Relational Context]
attributs/nomlst_10noms_5llr_5cooc.3repart.rcf
[Binary Relation]
Name_of_dataset
schéma_nom | signe_nom | évaluation_nom | interprétation_nom | tableau_nom | annexe_nom |
estimation_nom | indice_nom | symbole_nom | figure_nom |
efficacité_nom#nmod | possible_adj#u3_adjmod | degré_nom#nmod | 6_num#nmod |
visible_adj#u3_adjmod | comparable_adj#u3_adjmod | montrer_verb#~subj |
dispositif_nom#~nmod | travail_nom#~nmod | base_nom#nmod | résistance_nom#nmod |
proposer_verb#~subj | calculer_verb#~obj | situation_nom#nmod | réaliser_verb#~subj |
trouver_verb#~obj | faire_verb#~obj | cadre_nom#~nmod | autre_nom#~nmod |
récapitulatif_adj#u3_adjmod | explicatif_adj#u3_adjmod | proximité_nom#nmod |
interprétation_nom#nmod | premier_adj#u3_adjmod | 1_num#nmod | représenter_verb#~obj |
devenir_verb#~obj | compréhension_nom#coorditems | effet_nom#nmod | tangible_adj#u3_adjmod
| utiliser_verb#~obj | donnée_nom#nmod | extérieur_adj#u3_adjmod | influence_nom#~nmod |
mettre_verb#~subj | externe_adj#u3_adjmod | seul_adj#u3_adjmod | faire_verb#~subj |
mettre_verb#~obj | qualité_nom#nmod | impact_nom#nmod | subjectif_adj#u3_adjmod |
```

définir_verb#~obj | objet_nom#~nmod | représenter_verb#~subj | tutélaire_adj#u3_adjmod |
signifier_verb#~subj | faciliter_verb#~obj | l’un_pron#~nmod | pair_nom#nmod |
donner_verb#~obj | utiliser_verb#~subj | fonder_verb#nmod | valeur_nom#~nmod |
ligne_nom#~nmod | positif_adj#u3_adjmod | processus_nom#~nmod | calcul_nom#~nmod |
quantitatif_adj#u3_adjmod | 5_num#nmod | méthode_nom#~nmod | certain_adj#u3_adjmod |
géométrique_adj#u3_adjmod | constituer_verb#~obj | indiquer_verb#~subj | analyse_nom#nmod |
performance_nom#nmod | contradictoire_adj#u3_adjmod | appartenance_nom#nmod |
rendre_verb#~subj | partie_nom#~nmod | élaborer_verb#~obj | autre_adj#u3_adjmod |
activité_nom#nmod | objet_nom#coorditems | message_nom#nmod | nombreux_adj#u3_adjmod |
global_adj#u3_adjmod | effectuer_verb#~obj | cas_nom#~nmod | tête_nom#nmod | 4_num#nmod |
système_nom#~nmod | article_nom#nmod | niveau_nom#nmod | bas_adj#u3_adjmod |
utiliser_verb#nmod | proposer_verb#~obj | porter_verb#~subj | tel_adj#u3_adjmod |
synoptique_adj#u3_adjmod | réaliser_verb#~obj | montrer_verb#~obj | présenter_verb#nmod |
emblématique_adj#u3_adjmod | théorique_adj#u3_adjmod | pensée_nom#nmod | donner_verb#~subj
| différent_adj#u3_adjmod | obtenir_verb#~obj | voir_verb#~obj | produire_verb#~obj |
refléter_verb#~subj | activité_nom#~nmod | élaborer_verb#~subj | différence_nom#~nmod |
usage_nom#coorditems | obtenir_verb#nmod | prendre_verb#~obj | colonne_nom#~nmod |
présenter_verb#~obj | précédent_adj#u3_adjmod | manque_nom#nmod | coût_nom#nmod |
nombre_nom#~nmod | objectif_adj#u3_adjmod | susciter_verb#~obj | calculer_verb#~subj |
comparatif_adj#u3_adjmod | présenter_verb#~subj | crise_nom#nmod | résultat_nom#nmod |
rester_verb#~subj | conduire_verb#~subj | institutionnel_adj#u3_adjmod |
général_adj#u3_adjmod | terme_nom#nmod | fournir_verb#~obj | classique_adj#u3_adjmod |
ii_num#nmod | central_adj#u3_adjmod | reconnaissance_nom#nmod | il_pron#~nmod |
critère_nom#~nmod | outil_nom#~nmod | 2_num#nmod | reposer_verb#~subj | compléter_verb#~obj
| 3_num#nmod | illustrer_verb#~subj | schéma_nom#~nmod | majeur_adj#u3_adjmod |
dresser_verb#~obj | %communication_support | %processus_cognitif | %objet_scientifique |
%quantité | %méthode | %rapport | %équivalence | %évaluation | %graphique |
%objet_mathématiques | %section |
0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 0 0 0
0 0 0 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0
0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0
0 1 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 1 0 0 1 0 1 0 0
0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0
0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 1
1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1
0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 1 0 0 0 0 0 0
1 0 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 1 1 0 0 1 0
0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0
0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0
1 1 0 1 0 1 0 1 0 0 0 1 0 1 0 0
[END Relational Context]

A.X Patrons de cooccurrence entre LST et terminologie

Le tableau ci-dessous présente les différents patrons de cooccurrence entre un élément du LST et un élément terminologique que nous avons identifiés dans l'expérimentation détaillée section 5.3. Un patron consiste, a minima, en la présence d'un terme et d'un élément du LST, défini par son appartenance au LST, à une certaine classe ou à une certaine sous-classe sémantique du LST. Les unités lexicales composant ces patrons sont par ailleurs définies par leur catégorie syntaxique. Le tableau ci-dessous présente dans la première colonne les patrons de cooccurrence entre LST et terme. Nous proposons également des exemples de membres des classes et sous-classes intervenant dans les patrons. Les exemples, donnés dans la seconde colonne, sont extraits de notre corpus d'analyse présenté section 2.2.1.1.

Patron LST-terme	**Exemple**
Nom_{LST} Prep Nom_{Terme} $Verbe_{analyse\text{-}info}$ {analyse-info} : *opposer, décrire, détailler…*	*Les différentes formes$_{LST}$ de régulation$_{Terme}$ comportementale décrites$_{analyse\text{-}info}$ précédemment peuvent être représentées […]*[2] *Face à la crise, les networks choisissent en effet deux stratégies$_{LST}$ de programmation$_{Terme}$ opposées$_{analyse\text{-}info}$ pour leurs dramas.*[3]
Nom_{LST} Prep Nom_{Terme} Adj_{LST}	*Vu la proximité qui lie l'objet de discours et sa reprise, il est fréquent d'avoir des formes$_{LST}$ de diaphonies$_{Terme}$ implicites$_{LST}$.*[4] *Les consensus normatifs sont importants et rendent possible le processus$_{LST}$ de légitimation$_{Terme}$ globale$_{LST}$ de l'évaluation et de l'auto-évaluation.*[5]
$Verbe_{association}$ Nom_{Terme} {relation/association} : *articuler, accompagner, conjuguer…*	*Cette inscription spatiale et temporelle articule$_{association}$ en quelque sorte l'identité$_{Terme}$ civile$_{Terme}$.*[6] *Les gestes accompagnant$_{association}$ les conduites$_{Terme}$ verbales$_{Terme}$ constituent de fait des moyens cruciaux de communication.*[7]
$Verbe_{influence}$ Nom_{Terme} {relation/influence} : *influer, guider, opérer…*	*Par exemple, quatre attentes guident$_{influence}$ les délibérations$_{Terme}$ à propos du CSJ.*[8] *La présence de bois dans les chenaux influe$_{influence}$ sur la biomasse$_{Terme}$ de macro-invertébrés benthiques.*[9]

[2] Gillet, N., Berjot, S., & Paty, E. (2010). Profils motivationnels et ajustement au travail : Vers une approche intra-individuelle de la motivation. *Le travail humain*, *73*(2), 141-162.

[3] Perreur, N. (2011). La néo-série, arène d'évaluation culturelle d'une société américaine en crise. *Réseaux*, (1), 83-108.

[4] Miche, E. (2000). L'articulation entre les structures linguistique, textuelle et situationnelle du discours. Revue de Sémantique et Pragmatique, 7, 165-178

[5] Demailly, L. (2003). L'évaluation de l'action éducative comme apprentissage et négociation. *Revue française de pédagogie*, 115-129.

[6] Dubey, G. (2008). Nouvelles techniques d'identification, nouveaux pouvoirs. *Cahiers internationaux de sociologie*, (2), 263-279.

[7] Lemétayer, F. (2005). Le développement de l'autorégulation du comportement dans un contexte interactionnel au cours de la deuxième année de vie. L'*année psychologique*, *105*(4), 573-590.

[8] Vigour, C. (2010). Politiques et gouvernements fédéraux en Belgique, entre contraintes coalitionnelles et logique de compromis. *Politix*, (4), 63-86.

[9] Le Lay, Y. F., & Piégay, H. (2007). Le bois mort dans les paysages fluviaux français : éléments pour une gestion renouvelée. *L'Espace géographique*, *36*(1), 51-64.

$Verbe_{durée}$ Nom_{Terme} {temporalité/durée} : *conserver, maintenir…*	*La résolution divine de <u>maintenir</u>$_{durée}$ désormais l'<u>ordre</u>$_{Terme}$ <u>cosmique</u>$_{Terme}$ a pour contrepartie […].*[10] *Elle peut aussi être fondée sur des raisons de politique interne, <u>conserver</u>$_{durée}$ son <u>leadership</u>$_{Terme}$, répondre à une attente de la population.*[11]
Nom_{Terme} $Verbe_{observation}$ {observation} : *diviser, intégrer, englober…*	*Les <u>acteurs</u>$_{Terme}$ <u>institutionnels</u>$_{Terme}$ avec lesquels il faut négocier <u>englobent</u>$_{observation}$ l'administration et la classe politique.*[12] *La <u>posture</u>$_{Terme}$ <u>intègre</u>$_{observation}$ progressivement les perturbations posturales associées à un mouvement volontaire de façon anticipée.*[13]
Nom_{Terme} $Verbe_{processus\text{-}non\text{-}humain}$ {processus-non-humain} : *comporter, apparaître, intervenir, refléter…*	*La <u>globalisation</u>$_{Terme}$ <u>comporte</u>$_{processus\text{-}non\text{-}humain}$ ou favorise divers phénomènes migratoires qui amènent eux aussi des identités culturelles.*[14] *Ensuite, le <u>paysage</u>$_{Terme}$ nous <u>apparaît</u>$_{processus\text{-}non\text{-}humain}$ fondamental pour envisager les possibilités d'émergence d'une nouvelle gouvernance des milieux.*[15]
$Nom_{représentation}$ Prep Nom_{Terme} {objet_scientifique/représentation} : *équivalent, référence, symbole…*	*Le terme hybridity, équivalent*$_{représentation}$ *anglais de « métissage*$_{Terme\ »}$*, condense une vision idéalisée de l'identité.*[16] *Rousseau fait implicitement référence à la fameuse nuit de l'Escalade du 11 au 12 décembre 1602, devenue <u>symbole</u>$_{représentation}$ de l'<u>indépendance</u>$_{Terme}$ genevoise.*[17]

[10] Albert, J. P. (2009). Les animaux, les hommes et l'Alliance. L'*homme*, (1), 81-114.

[11] Faget, J. (2008). Les métamorphoses du travail de paix. *Revue française de science politique*, *58*(2), 309-333.

[12] Tissot, S. (2005). Reconversions dans la politique de la ville : l'engagement pour les « quartiers ». *Politix*, (2), 71-88.

[13] Marty, E., Rebillard, F., Pouchot, S., & Lafouge, T. (2012). Diversité et concentration de l'information sur le web. *Réseaux*, (6), 27-72.

[14] Wieviorka, M. (2008). L'intégration : un concept en difficulté. *Cahiers internationaux de sociologie*, (2), 221-240.

[15] Fourault-Cauët, V. (2010). Le paysage, outil de territorialisation et d'aménagement incomplet pour les forêts méditerranéennes ?. *Annales de géographie,* (3), 268-292

[16] Trémon, A. C. (2007). Fils illégitimes, affiliations conflictuelles. L'*homme*, (1), 75-101.

[17] Markovits, R. (2009). L'incendie de la comédie de Genève (1768) : Rousseau, Voltaire et l'impérialisme culturel français. *Revue historique*, (4), 831-873.

$Verbe_{LST}$ $Nom_{déterminant}$ Prep Nom_{Terme} {déterminant} : *type, ensemble, partie…*	*La seconde variable concerne le filtrage fréquentiel des images, constituant$_{LST}$ 4 types$_{déterminant}$ de stimulation$_{Terme}$.*[18] *Ainsi, si on regroupe$_{LST}$ l'ensemble$_{déterminant}$ des prépositions$_{Terme}$ spatiales.*[19]
$Nom_{réalisation}$ Prep Nom_{Terme} {processus-cognitif/réalisation} : *conception, constitution, élaboration…*	*La vague du Web 2.0 a particulièrement mis l'accent sur la figure d'un usager actif et co-responsable de la conception$_{réalisation}$ du dispositif$_{Terme}$ de communication$_{Terme}$*[20] *Cet article se propose d'analyser le rôle des outils spirituels et politiques dans l'élaboration$_{réalisation}$ de cette communauté$_{Terme}$ politique*[21]
$Nom_{évaluation}$ Prep Nom_{Terme} {processus-cognitif/évaluation} : *appréciation, estimation, évaluation…*	*La seconde limite concerne l'évaluation$_{évaluation}$ de la performance$_{Terme}$ au travail*[2]. *Les variables relatives […] à l'estimation$_{évaluation}$ des revenus$_{Terme}$ du ménage sont sans influence statistique*[22].

Tableau 3: Exemples de patrons de cooccurrence entre LST et terminologie

[18] Giraudet, G., & Roumes, C. (2004). La signature spatiale de l'objet : une information essentielle pour la localisation de cibles dans une scène naturelle. *L'année psychologique*, *104*(1), 9-49.

[19] Borillo, A. (2001). Il y a prépositions et prépositions. *Travaux de linguistique*, (1), 141-155.

[20] Casemajor Loustau, N. (2011). La contribution triviale des amateurs sur le Web : quelle efficacité documentaire ?. *Études de communication. langages, information, médiations*, (36), 39-52.

[21] Hilgers, M. (2007). La dynamique de la croyance. L'*homme*, (2), 131-161.

[22] Gonthier, F. (2008). La justice sociale entre égalité et liberté. *Revue française de science politique*, *58*(2), 285-307.

A.XI Classes lexicales pour l'extraction de routines

Ci-dessous sont présentées les classes lexicales utilisées pour l'extraction de routines à l'aide du *Lexicoscope*. Son module d'extraction des routines, présenté section 5.4.2, peut intégrer dans les calculs de cooccurrences des classes de mots. Ces classes sont définies de la manière suivante : $nomclasse:=(mot1|mot2|mot3|..). Nous avons transcrit nos classes sémantiques des noms du LST, ainsi que les classes verbales que nous étudions dans la section 5.4, au formalisme des classes lexicales pour lancer l'extraction des routines. Nous avons également réutilisé deux classes définies par défaut dans les paramètres du *Lexicoscope* qui permettent de regrouper les éléments de négation d'une part et les déterminants possessifs d'autre part.

$PAS :=(pas|plus|jamais|guère) # négation

$SON :=(son|leur|mon|ton|notre|votre) # déterminant possessif

$NLSTOBJ :=(circonstance|condition|événement|exception|phénomène|situation|corpus|donnée|entité|information|matériau|objet|code|convention|loi|modèle|norme|paradigme|principe|règle|réseau|schéma|typologie|composante|concept|contenu|critère|exemple|facteur|indicateur|notion|paramètre|sens|signification|variable|but|cible|finalité|objectif|objet|projet|enquête|entretien|instrument|matériel|moyen|outil|approche|démarche|dispositif|logique|méthode|méthodologie|perspective|plan|procédé|procédure|système|théorie|argument|conclusion|réponse|solution|équivalent|figure|image|indication|indice|instance|marque|modèle|référence|repère|représentation|signe|symbole|enjeu|point|problématique|problème|projet|question|sujet|thématique|thème|angle|conception|hypothèse|idée|idéologie|opinion|positionnement|posture|thèse|vision|voie) # nom LST classe {objet_scientifique}

$NLSTESP :=(cadre|champ|contexte|dimension|domaine|environnement|espace|lieu|milieu|pôle|secteur|sphère|terrain|territoire|univers|zone|contour|frontière|limite|centre|extérieur|niveau|plan|pôle|position|rang|source|axe|direction|orientation|sens|voie) # nom LST classe {espace}

$NLSTTEMPO :=(date|étape|horizon|origine|terme|année|cycle|durée|période|phase|siècle) # nom LST classe {temporalité}

$NLSTREL :=(articulation|association|corrélation|interaction|liaison|relation|convergence|correspondance|identité|proximité|base|cause|condition|conséquence|contrainte|dépendance|effet|explication|fondement|impact|implication|incidence|influence|motif|rapport|résultat|clivage|contradiction|contraste|différence|divergence|opposition|paradoxe) # nom LST classe {relation}

$NLSTQUANT :=(différence|distance|écart|échelle|égalité|indice|majorité|maximum|minimum|minorité|moyenne|pourcentage|proportion|résultat|somme|statistique|taux|tiers|total|chiffre|degré|nombre|point) # nom LST classe {quantité}

$NLSTEVOL :=(accroissement|affirmation|amélioration|augmentation|croissance|développement|expansion|extension|gain|généralisation|multiplication|progrès|progression|renforcement|baisse|déclin|diminution|perte|réduction|changement|déplacement|diffusion|dynamique|évolution|flux|modification|mouvement|mutation|processus|transfert|transformation|transition|transmission|variation|apparition|émergence|expression|genèse|manifestation) # nom LST classe {processus_évolutif}

$NLSTPCOM :=(article|document|ouvrage|texte|thèse|volume|figure|illustration|image|tableau|annexe|bibliographie|chapitre|conclusion|introduction|section) # nom LST classe {communication_support}

$EXAM :=(analyser|étudier|examiner) # verbe LST sous-classe {analyse_info/examen}

$EVAL :=(mesurer|estimer|évaluer) # verbe LST sous-classe {analyse_info/évaluation}

$DESCRIPTION :=(présenter|décrire|illustrer) # verbe LST sous-classe {analyse_info/description}

$ANALY :=(analyser|étudier|examiner|mesurer|estimer|évaluer|décrire|illustrer|présenter) # verbe LST classe {analyse_info}

Résumé

Cette thèse s'intéresse au lexique scientifique transdisciplinaire (LST), lexique inscrit dans le genre de l'article de recherche en sciences humaines et sociales. Le LST est fréquemment mobilisé dans les écrits scientifiques et constitue ainsi un objet d'importance pour l'étude de ce genre. Ce lexique trouve également des applications concrètes tant en indexation terminologique que pour l'aide à la rédaction/compréhension de textes scientifiques. Ces différents objectifs nous amènent à adopter une approche outillée pour identifier et caractériser les unités lexicales du LST, lexique complexe à circonscrire, situé entre lexique de la langue générale et terminologie. En nous basant sur les propriétés de spécificité et de transdisciplinarité ainsi que sur l'étude des propriétés lexico-syntaxiques de ses éléments, nous élaborons une ressource du LST intégrant informations lexicales, syntaxiques et sémantiques. L'analyse de la combinatoire à l'aide d'un corpus arboré autorise ainsi une caractérisation du LST ancrée sur l'usage dans le genre de l'article de recherche. Selon cette même approche, nous identifions les acceptions nominales transdisciplinaires et proposons une classification sémantique fondée sur la combinatoire en corpus pour intégrer à notre ressource lexicale une typologie nominale sur deux niveaux. Nous montrons enfin que cette structuration du LST nous permet d'aborder la dimension phraséologique et rhétorique du LST en faisant émerger du corpus des constructions récurrentes définies par leurs propriétés syntactico-sémantiques.

Mots-clés : linguistique de corpus, traitement automatique du langage, lexicologie, écrits scientifiques, lexique scientifique transdisciplinaire, sémantique

Abstract

In this dissertation we study the French cross-disciplinary scientific lexicon (CSL), a lexicon which fall within the genre of scientific articles in humanities and social sciences. As the CSL is commonly used in scientific texts, it is a gateway of interest to explore this genre. This lexicon has also practical applications in the fields of automatic terms identification and foreign language teaching in the academic background. To this end, we apply a corpus-driven approach in order to extract and structure the CSL lexical units which are complex to circumscribe. The method relies on the cross-disciplinarity and specificity criteria and on the lexico-syntactic properties of the CSL lexical units. As a result, we designed a lexical resource which include lexical, syntactical and semantical informations. As we analyze the combinatorial properties extracted from a parsed corpus of scientific articles, we performed a CSL study based on its genre specific use. We follow the same approach to identify cross-disciplinary meanings for the CSL nouns and to design a nominal semantic classification. This two-level typology allow us to explore rhetorical and phraseological CSL properties by identifying frequent syntactico-semantic patterns.

Keywords : corpus linguistics, natural language processing, lexicology, scientific texts, cross-disciplinary scientific lexicon, semantics