Analyse contrastive des verbes dans des corpus médicaux et création d'une ressource verbale de simplification de textes

Ornella Wandji Tchami

UNIVERSITÉ DE LILLE 3
UMR STL 8163
et
UNIVERSITÄT HILDESHEIM
IWiSt

Analyse contrastive des verbes dans des corpus médicaux et création d'une ressource verbale de simplification de textes

Ornella WANDJI TCHAMI

Thèse présentée en vue de l'obtention du grade de
Docteur en Science du Langage – spécialité Linguistique computationnelle

Le 26 Février 2018

Composition du jury

Dany Amiot	Professeur des Universités, HDR	Université de Lille (directrice)
Ulrich Heid	Professeur des Universités, HDR	Université de Hildesheim (directeur)
Natalia Grabar	Chercheur CR1, CNRS	Université de Lille (encadrante)
Anne-Laure Ligozat	Maître de conférences, HDR	ENSIIE et LIMSI, CNRS (rapporteur)
Amalia Todirascu	Professeur des Universités, HDR	Université de Strasbourg (rapporteur)
Christiane Maas	Professeur des Universités, HDR	Université de Hildesheim (Présidente)

Remerciements

Je tiens tout d'abord à remercier chaleureusement Dany Amiot ma directrice de thèse (en France) et à lui exprimer ma gratitude, non seulement pour son encadrement et son expertise en matière de Linguistique qui ont contribué à rehausser considérablement la qualité de mon travail de thèse, mais aussi pour sa présence permanente, sa patience, ses encouragements, son soutien et sa gentillesse dont j'ai grandement bénéficié tout au long de mon parcours académique depuis mon master à l'Université de Lille 3. Avec elle, j'ai appris que faire un doctorat est avant tout une aventure humaine.

Je tiens également à exprimer ma gratitude et ma reconnaissance à M. Ulrich Heid mon directeur de thèse (en Allemagne) pour la confiance à moi accordée dès le départ, en acceptant de diriger mon travail dans le cadre de la cotutelle de cette thèse. Je le remercie grandement de m'avoir accueillie dans un environnement de travail agréable et convivial, qui a contribué à mon épanouissement, tant au plan de la recherche que dans la vie de tous les jours, dans un milieu que je découvrais à peine. Je lui suis également reconnaissante pour sa grande disponibilité, malgré ses diverses occupations, pour sa bonne humeur, toujours au rendez-vous à chacune de nos séances de travail, et surtout pour son expertise en Terminologie et en TAL, ses explications, ses commentaires, qui m'ont permis d'améliorer et d'enrichir considérablement mon travail de recherche. Je ne saurais achever mon propos sans le remercier de m'avoir appris à développer plus de rigueur dans le travail et de m'avoir davantage insufflé le goût du travail bien fait.

Ma reconnaissance et mes remerciements vont également à l'endroit de Natalia Grabar sans qui ce travail n'aurait pas existé. C'est elle qui m'a proposé ce grand défi de thèse en Traitement Automatique des Langues (TAL) avec ce thème que j'ai trouvé des plus passionnants dès le départ (à travers mon mémoire de Master), tandis que je découvrais encore le monde du TAL. Merci pour ta confiance, merci de m'avoir transmis ta passion pour le TAL, merci de m'avoir fait bénéficier de ton expertise, de tes conseils et commentaires avisés sans lesquels ce travail n'aurait pu aboutir à un tel résultat.

Je remercie Mme Amalia Todirascu et Mme Anne-Laure Ligozat (les rapporteurs), sans oublier Mme Christiane Maas (Présidente du jury), qui m'ont fait l'honneur de participer à

l'évaluation de ce travail de recherche.

Je remercie Marie-Claude L'Homme avec qui j'ai eu l'honneur de faire quelques publications et grâce à qui j'ai pu effectuer un séjour de recherche productif et agréable au Laboratoire Observatoire de Linguistique Sens-Texte (OLST) de Montréal, Canada. Je ne saurais oublier Patrick Drouin avec qui j'ai également eu l'honneur de travailler sur la plate-forme TermoStat durant mon séjour. Cette expérience à Montréal m'a été très bénéfique notamment pour développer mes connaissances en ce qui concerne l'analyse des verbes en tant qu'unités terminologiques, me permettant ainsi de mieux développer ma propre méthode d'analyse des verbes de mon corpus selon les objectifs visés. Merci à toute la chaleureuse équipe du laboratoire de l'OLST qui m'a permis de passer un moment agréable en plein hiver canadien.

Je tiens également à remercier toute l'équipe de médecins et infirmiers français, belges et canadiens qui m'ont fait bénéficier de leur expertise en acceptant d'effectuer la validation de mes données. Merci particulièrement à mon oncle tonton, Faustin Nguembou et à ses collègues. Merci à ma cousine Paule Georgina et à mon amie Fany Merveille pour leur efficacité. Merci du fond du coeur à ma grande soeur Delboise dont l'expertise dans le domaine médical m'a été d'une aide indispensable pour la compréhension des expressions verbales spécialisées. Tu as plusieurs fois sacrifié ton propre travail pour voler à mon secours, mille merci pour ta grande disponibilité et ton soutien dévoué ma petite maman chérie.

Merci à l'organisme DAAD pour la bourse dont j'ai été bénéficiaire et qui m'a permis de m'installer en Allemagne dans le cadre de la cotutelle et de financer plus de deux années d'étude. Merci à l'université de Lille 3 pour les différences bourses reçues (DAI, bourse régionale) et au laboratoire STL (Savoirs, Textes, Langage) pour les différents financements reçus lors de mes déplacements pour les conférences et autres évènements dans le cadre de la recherche.

Ce travail n'aurait pas pu être réalisé sans le soutien matériel et humain dont j'ai bénéficié au sein du laboratoire STL et au sein de l'institut IwiSt (Institut für Informationswissenschaft und Sprachtechnologie) de l'université de Hildesheim. Je tiens à exprimer ma gratitude à tous mes collègues, doctorants ou non, linguistes ou non. Je remercie particulièrement Laetitia, Dodzi, Pierre, Laurence, Hanna Karpenko, Nicolas et Julie. Du côté de Hildesheim, je dis merci spécialement à Gertrud Faass, Fritz Kliche, David Lindemann, Monsieur Folker Caroli, pour leur soutien et leurs encouragements (surtout pendant la semaine de la soumission), pour l'accueil chaleureux qu'ils m'ont reservé dès mon arrivée à Hildesheim. Un clin d'oeil particulier à ma collègue Laura Giacomini, avec qui j'ai eu le plaisir de partager un bureau. Enfin, un merci spécial à ma collègue et amie Noushin pour son soutien inqualifiable, surtout durant la semaine de la soumission. Vielen Dank für Ihre Unterstützung !

Lors de mes participations à des conférences et/ou workshops, j'ai bénéficié de nombreux retours, remarques, questions et commentaires sur mes travaux. Je remercie tous ceux qui m'ont ainsi permis de faire avancer ce travail de recherche. Je remercie toutes les personnes

avec lesquelles j'ai échangé pendant mon parcours de thèse, de près ou de loin en rapport avec mon sujet. Merci particulièrement à Karolina, Milica, Alexandra et Amandine Périnet qui s'est toujours montrée disponible et prompte à m'apporter son aide.

À Jérôme Michaud, j'adresse un merci particulier pour son temps précieux, ses nombreuses explications et ses solutions qui m'ont aidée à maintes reprises à me débarrasser des blocages rencontrés avec Latex.

À chacune de mes relectrices, je formule ma profonde gratitude car n'eut été votre contribution, ce travail n'aurait pas été ce qu'il est aujourd'hui. Edwige, merci ne suffirait pas pour t'exprimer ma reconnaissance pour le magnifique travail que tu as abattu en relisant mon manuscrit avec efficacité malgré les contraintes de temps qui se présentaient. Vraiment merci pour tes commentaires et corrections. Ma grande soeur Gertrude, mes très chères soeurs et amies Marie-Thérèse, Julia, Kévine, Christelle, vous avez été bien plus que des relectrices car votre présence (chacune d'une façon particulière) m'a accompagnée tout au long de ce long voyage qui a duré quatre ans. Vous avez toujours été là, malgré la distance : du Sénégal et du Cameroun à la Finlande en passant par l'Allemagne et la France, auprès de vous, j'ai toujours trouvé une écoute attentive, un soutien morale et spirituel inconditionnel, des paroles encourageantes et la motivation pour retrouver le sourire et me remettre au travail quand il fallait. Je ne saurais jamais exprimer véritablement ma gratitude. Vous savez d'où votre récompense viendra.

Mes remerciements sont également adressés à mes frères et soeurs en Christ : Patrick Talom, Lili Joy, Paola, Audrey, Nairice, Koryn, Chrystele, ainsi que toutes les prunelles et les sentinelles, qui m'ont soutenue et encouragée de leurs prières, conseils et paroles de motivation. Nous récolterons ensemble les fruits de ce travail.

À mes chers amis Olivia, Jessica, Josiane, Eveline, Diane Morgane, Gaëlle, Olivia Samo, Franck (Prési), Stéphane Tongsi, Maurice, Mathias, Diane, Yannick, Lucrèce, Carole, Laetitia, Thyura qui malgré la distance ont toujours été présents, je dis merci de tout coeur.

Un grand merci à tous mes beaux-frères et mes belles-soeurs, particulièrement à Sophie, Fulbert, Raïssa, Dolly, grande soeur Stéphanie. Merci spécial à Cédric pour son aide dévouée en matière de génération des graphes en R, à Docteur Dimi pour m'avoir fait bénéficier de son expertise en médecine, merci à mon fils Arnold pour les moments de détente et de fou-rire. Je remercie particulièrement ma belle-soeur Manuella Yapomo, sans qui je n'aurais pas suivi ce cursus TAL, merci pour tes encouragements, conseils, et directives Manu, je te souhaite tout le meilleur. Merci à Kévine (une fois de plus) et Jocelyn mon partenaire, Marie-Chantale, Sandra, Patty et Stanley, Murielle. Maman Christiane, merci pour tout, tu es un cadeau du ciel dans ma vie, tout ce que tu as fait et ne cesse de faire signifie beaucoup pour nous. Merci à ma petite Angela pour ses gentils petits messages.

À la famille Briolet, Ludo, Véro, Julia et les garçons, j'ai trouvé en vous une véritable famille, merci pour toutes vos marques d'attention et d'affection à mon égard. Acceptez ma

reconnaissance !

Je remercie enfin toutes les personnes qui m'ont soutenue d'une façon où d'une autre, je ne saurais citer tout le monde mais c'est avec une profonde gratitude que je vous dis merci.

Je ne saurais mettre à l'écart ma grande famille qui est toujours là pour moi, je ne peux rester insensible à votre sollicitude durant ces moments, merci du fond du coeur à chacun en particulier. Merci spécialement à mon oncle et père Marcel Tchami et à ma tata chérie Arlette pour tout, à tonton Albert, Papa Tagni et mamam Magni également. Mes mots sont faibles pour exprimer ma reconnaissance à votre égard. Merci à mes grands parents (Maa Mali mon soutien, Père Tchami notre modèle et pilier, et Maa tonta), mes oncles, tantes, cousins et cousines du Cameroun, de la France, la Belgique, l'Italie, du Canada et des USA : à maman Marie, tata Blanche, tata Raméline, tonton Audifac, tonton Jovite, Mami ma mère, Merveille mon bébé, Willy, Eric, Carine, Saurelle, Olive, Igor, Japhète, Yann, Jaurel, et spécialement à mes soeurs Merciel et Suzie pour leurs prières. Je remercie particulièrement mes grandes soeurs adorées Laetitia et Delboise, et mon petit frère cheri Loïc, pour leur soutien infaillible, autant sur le plan moral, financier, que intellectuel. Vous m'êtes si précieux !

Je ne saurais oublier ceux là qui m'ont permis de venir au monde, mes parents bien-aimés, papa et maman Wanji, ainsi que mes très chers et adorables beaux parents papa et maman Tchanyou. Ces quatre personnes ont été plus que présentes tout au long de cette dernière année, me soutenant par leurs prières et paroles réconfortantes. Merci beaucoup, sans vous, je n'en serai pas là ! Vous êtes une bénédiction dans ma vie, je suis contente de faire votre fierté et de vous honorer par l'obtention de ce doctorat.

Enfin, un immense merci à mon très cher et tendre époux Patrony, mon empereur, mon Sterling à moi, mon ami, mon complice, celui- là qui est et a été mon support permanent tout au long de ce parcours. Merci pour ta patience, ta compréhension et pour ton soutien indéfectible tant moralement, spirituellement, et aussi sur le plan intellectuel. Tu as relu mon travail et tu as su m'apporter des solutions lorsque je butais sur certains problèmes que je pensais impossibles à résoudre. Tu as été avec moi jusqu'à la dernière minute. Près de toi, j'ai toujours trouvé l'assistance, l'assurance et le réconfort dont j'avais besoin dans les moments difficiles. Tu as d'ailleurs réussi à me faire rigoler dans certains de mes moments de pression les plus intenses. Il n'en y a qu'un seul comme toi ! Tu as tant donné pour cette thèse, je te la dédie.

An exact classification of scientific and technical texts on the basis of dominant distinguishing features forms one of the decisive prerequisites for the successful solution of specific communicative tasks [...]. A catalogue and later a systematic description of the different types of texts which are mainly determined by their specific functions, together with their typical features may contribute to improved comprehension of specialized information and above all to the production of texts which convey such information adequately. This, again, would be a major contribution to more efficient mono and multilingual communication.
Hoffmann 1983, p. 62

Table des matières

Abréviations **xiii**

INTRODUCTION **1**

1 État de l'Art **7**

1.1 Langues de spécialités et langue médicale 8
1.1.1 Typologies des textes en langue de spécialité 8
1.1.2 Discours scientifique ou Langue de spécialité 9
1.1.3 Langue médicale 10
1.2 Communication médecins vs. patients 13
1.2.1 Communication orale 14
1.2.2 Communication écrite 16
1.2.3 Bilan 20
1.3 Le verbe dans les travaux de recherche 20
1.3.1 Linguistique 20
1.3.1.1 Structure actantielle et valence verbale 21
1.3.1.2 Grammaire de dépendance 22
1.3.1.3 La sémantique des cadres 24
1.3.2 Terminologie 27
1.3.2.1 L'approche conceptuelle 29
1.3.2.2 L'approche lexico-sémantique 30
La structure argumentale 30
Le réseau lexical 32
1.3.3 Traitement Automatique des Langues 34
1.3.3.1 FrameNet 34
1.3.3.2 VerbNet 35
1.3.3.3 La base de données des Verbes Français 36
1.3.3.4 Dicovalence 38
1.3.4 Bilan 39
1.4 La simplification de textes 40

1.4.1 La simplification syntaxique . . . 41
1.4.2 La simplification lexicale . . . 42
1.5 Bilan . . . 44

2 Corpus, ressources et outils **47**
2.1 Corpus . . . 48
2.1.1 Types . . . 48
2.1.2 Sources . . . 49
2.1.2.1 CISMeF . . . 49
2.1.2.2 Les forums médicaux : Doctissimo . . . 51
2.1.3 Taille et Contenu . . . 52
2.1.3.1 Corpus des experts . . . 52
2.1.3.2 Corpus des étudiants . . . 55
2.1.3.3 Corpus des patients . . . 56
2.1.3.4 Corpus des forums . . . 57
2.2 Ressources et outils . . . 59
2.2.1 Cordial Analyseur . . . 59
2.2.2 La terminologie Snomed International . . . 60
2.3 Bilan . . . 63

3 Méthode **65**
3.1 Collection et pré-traitement des corpus . . . 67
3.2 Annotation des corpus et acquisition des PSS . . . 68
3.2.1 Annotation syntaxique et extraction des patrons valenciels . . . 68
3.2.1.1 Etiquetage syntaxique des corpus avec Cordial . . . 68
3.2.1.2 Pré-traitement des résultats de Cordial . . . 70
3.2.1.3 Extraction des patrons syntaxiques des verbes . . . 71
3.2.2 Annotation sémantique des arguments . . . 72
3.2.3 Traitement des têtes multicatégorielles . . . 76
3.2.3.1 Evaluation des têtes multicatégorielles . . . 78
3.2.3.2 Désambiguïsation des termes ambigus (têtes multicatégorielles) 82
La méthode fréquentielle . . . 82
Le sémantisme du verbe pivot . . . 84
Le contexte syntaxico-sémantique du terme ambigu . . . 85
3.2.3.3 Faiblesses de la méthode de désambiguïsation des termes . . . 88
3.2.4 Acquisition des PSS . . . 90
3.3 Sélection des verbes et PSS pour la validation . . . 91
3.3.1 Sélection des verbes . . . 92

3.3.2 Sélection des PSS . . . 94
3.3.3 Vérification et correction des PSS . . . 97
3.3.4 Formatage des PSS en vue de la validation . . . 98
3.4 Validation des patrons syntaxico-sémantiques par les experts . . . 99
3.4.1 Présentation de la tâche . . . 99
3.4.2 Motivation et objectifs . . . 99
3.4.3 Population . . . 100
3.4.4 Description du questionnaire et du protocole de validation . . . 101
3.5 Création de la ressource pour la simplification . . . 101
3.5.1 Alignement des PSS avec des équivalents de la langue générale . . . 102
3.5.1.1 Sélection automatique des potentiels candidats équivalents . 102
3.5.1.2 Filtrage manuel des candidats équivalents . . . 103
3.5.2 Modélisation de la ressource de simplification . . . 106
3.6 Comparaison des corpus : fonctionnement des collocations verbe-terme . . . 107
3.6.1 Extraction des collocations verbe-terme . . . 107
3.6.2 Analyse des collocations verbe-terme . . . 108
3.6.2.1 Analyse quantitative . . . 109
3.6.2.2 Analyse qualitative : préférences sémantico-lexicales des verbes 109
3.7 Bilan . . . 110

4 Résultats et discussion 113
4.1 Résultats de l'annotation des corpus et acquisition des PSS . . . 114
4.1.1 Annotation syntaxique et extraction des schémas valenciels . . . 115
4.1.1.1 Récapitulatif des résultats de l'annotation syntaxique . . . 115
4.1.1.2 Fréquence du verbe et type de corpus . . . 118
4.1.1.3 Relation entre les différents corpus . . . 127
4.1.2 Annotation sémantique et acquisition des PSS . . . 130
4.1.2.1 Approche quantitative . . . 130
4.1.2.2 Approche qualitative . . . 137
1. Variation syntaxique . . . 141
a. Alternance passif/actif . . . 141
b. Changement au niveau de la structure argumentale . . . 144
2. Variation sémantique . . . 145
3. Variation lexicale . . . 152
4.2 Difficultés de l'annotation syntactico-sémantique . . . 154
4.2.1 Annotation syntaxique . . . 154
4.2.1.1 Le choix de l'annotateur syntaxique . . . 154
4.2.1.2 Les faiblesses de l'annotateur Cordial . . . 155

4.2.2 Annotation sémantique . . . 156
4.2.2.1 La variation terminologique . . . 156
4.2.2.2 La non-exhaustivité de la ressource terminologique . . . 157
4.2.2.3 Problème de désambiguïsation des termes non procéduraux . 157
4.3 Résultats de la sélection des PSS pour la simplification . . . 158
4.3.1 Sélection des verbes . . . 159
4.3.2 Sélection et vérification des PSS pour la validation . . . 160
4.4 Résultats de la validation des PSS : analyse et interprétation . . . 162
4.4.1 Présentation générale des résultats . . . 162
4.4.2 Analyse détaillée des résultats . . . 166
4.4.3 Bilan . . . 173
4.5 Simplification des PSS . . . 174
4.5.1 Résultats de la sélection automatique des candidats PSS équivalents . . 175
4.5.2 Résultats de l'alignement . . . 176
4.6 Corpus et collocations verbe-terme . . . 182
4.6.1 Collocations verbe-terme et variations syntaxico-sémantique . . . 182
4.6.2 Collocations verbe-terme et variation lexicale . . . 184

5 Évaluation des méthodes, ressources et outils utilisés 191
5.1 Évaluation de l'annotation syntaxique des corpus avec Cordial Analyseur . . . 192
5.1.1 L'annotateur Cordial dans les campagnes d'évaluation EASY et PASSAGE 192
5.1.1.1 La campagne d'évaluation EASY . . . 192
5.1.1.2 La (pré-)campagne PASSAGE (2007) . . . 194
5.1.1.3 La seconde campagne PASSAGE (2009) . . . 195
5.1.1.4 Performance de Cordial dans les différentes campagnes . . . 195
5.1.2 Évaluation des résultats de l'annotation syntaxique des corpus . . . 197
5.1.2.1 Raison de l'évaluation et justification du choix de la méthode d'évaluation appliquée . . . 197
5.1.2.2 Présentation de la méthode d'évaluation appliquée . . . 197
5.1.2.3 Résultats de l'évaluation et comparaison avec les résultats de PASSAGE . . . 200
Délimitation des constituants . . . 201
Fonction syntaxique des constituants . . . 201
Dépendance de la structure argumentale vis-à-vis du verbe . . . 201
5.1.2.4 Quelques observations faites à l'issue de l'évaluation . . . 202
Mauvais raccordement des syntagmes prépositionnels et génération de faux COI . . . 202
Traitement des phrases longues et complexes . . . 205

Traitement des phrases du corpus des forums 206
Traitement des formes verbales complexes 206
5.2 Bilan 208
5.3 Annotation sémantique avec la Snomed 209
5.3.1 But de l'évaluation 209
5.3.2 Description de la démarche et des données évaluées 209
5.3.3 Résultats 209
5.4 Évaluation de la ressource de simplification 214
5.4.1 But et population 214
5.4.2 Démarche et justification du choix de la méthode 215
5.4.3 Résultats 216
5.4.3.1 Évaluation par les linguistes 216
5.4.3.2 Évaluation par les non-linguistes 218
5.5 Bilan 222

CONCLUSION **223**

A Les ressources pour l'annotation syntaxique des corpus **227**
A.1 Codage des fonctions grammaticales Cordial 227

B Les ressources terminologiques pour l'annotation sémantique des corpus **229**
B.1 Les formes plurielles des termes simples de la Snomed 229
B.2 Les termes mal orthographiés 230
B.3 Les 274 têtes multicatégorielles en -ment, -ion, -age et -eur 233
B.4 Fréquence des têtes multicatégorielles dans la Snomed 237

C Les ressources linguistiques pour l'annotation sémantique des corpus **243**
C.1 Les mots-outils 243
C.2 Les déterminants complexes 245
C.3 Les verbes de réalisation 246

D Les ressources conçues pour la simplification de textes **249**
D.1 Les 50 phrases (originales) utilisées dans le cadre de l'évaluation de la ressource. 249
D.2 Les 50 phrases après simplification 253
D.3 Les 243 PSS sélectionnés par les experts pour l'alignement 256
D.4 Les 230 PSS (entrées de la ressource) alignés avec leurs équivalents 259

Bibliographie **275**

Abréviations

c.-à-d.	c'est-à-dire
CatSem	Catégorie Sémantique de la Snomed
COD	Complément d'objet directe
COI	Complément d'objet indirecte
s	Sujet
FN	FrameNet
FS	Frame Semantics
GD	Grammaire de dépendance
LS	Langue de spécialité
LVF	Les Verbes Français
TAL	Traitement Automatique des Langues
PSS	Patron syntactico-sémantique
VN	VerbNet
VSD	Désambiguïsation des sens du verbe
WSD	Désambiguïsation des sens du mot

INTRODUCTION

Contexte

Grâce à l'évolution de la technologie à travers le Web, la documentation relative à la santé est de plus en plus abondante et accessible à tous, en particulier aux patients qui ont ainsi accès à une panoplie d'informations sanitaires. Malheureusement, la grande disponibilité de l'information médicale ne garantit pas systématiquement sa bonne compréhension par le public visé, en l'occurrence les non-experts, c'est-à-dire ceux qui ont très peu ou pas du tout de connaissances en médecine. Ce constat concerne particulièrement les textes électroniques (disponibles sur le Web) qui sont accessibles à un large panel de lecteurs. Un travail de recherche (McCray, 2005) réalisé en 2005 et visant à promouvoir l'accès à l'information sur la santé démontre que les patients et/ou, plus largement, les « profanes » en matière de connaissances médicales sont victimes de la complexité des informations mises à leur disposition. Cette étude trouve un prolongement dans un article (Tran *et al.*, 2009) qui étudie le rôle d'internet dans la relation médecin/patient. Les auteurs y décrivent la communication médecin/patient comme une situation d'échange inégal d'informations dans laquelle le médecin est le « détenteur du savoir médical ». Le langage médical est donc difficile à appréhender pour les non-experts. Par conséquent, la plupart du temps, ces lecteurs comprennent très peu, ou pas du tout, les informations qui leur sont adressées (Lerner *et al.*, 2000 ; Chapman *et al.*, 2003 ; Zeng-Treiler *et al.*, 2007).

La majorité des travaux de la littérature se focalise uniquement sur l'hypothèse selon laquelle l'abondance des noms de concepts médicaux (Lerner *et al.*, 2000 ; Abrahamsson *et al.*, 2014), que certains chercheurs appellent *notions opaques* (Grabar & Hamon, 2014), serait à l'origine de la complexité des textes médicaux. D'autres s'intéressent également à l'impact de la structure des textes et de la syntaxe (Callan & Eskenazi, 2007 ; Brouwers *et al.*, 2014). Les auteurs de ces travaux sont unanimes sur au moins une chose : il y a un réel besoin de développer des méthodes pour rendre l'information médicale plus compréhensible pour les non-experts (McCray, 2005 ; Zeng-Treiler *et al.*, 2006 ; Deléger & Zweigenbaum, 2008).

Face à ce constat, des chercheurs de différents domaines, y compris l'informatique médicale et le Traitement Automatiques des Langues (TAL), préconisent la simplification des textes médicaux adressés au grand public. Cette consiste à réduire la complexité linguistique d'un texte, tout en préservant son contenu sémantique (Siddharthan, 2014b). Il s'agit de cibler et supprimer les éléments susceptibles d'empêcher la compréhension aisée d'un texte, afin de faciliter l'accès à son contenu sémantique. Sur le plan lexical, les chercheurs proposent la création de vocabulaires alignant la terminologie spécialisée (noms, groupes nominaux complexes, etc.) avec la terminologie non spécialisée et/ou des définitions, des paraphrases ou des explications (Zielstorff, 2003 ; Zeng-Treiler & Tse, 2006 ; Elhadad, 2006 ; Elhadad & Sutaria, 2007 ; Deléger & Zweigenbaum, 2009 ; Kandula *et al.*, 2010 ; Grabar & Hamon, 2014). En ce qui concerne la syntaxe, les chercheurs proposent d'alléger la structure syntaxique de la phrase (raccourcir les longues phrases, segmenter les propositions coordonnées et subordonnées) à partir de règles de substitution ou de remplacement, de suppression, de modification, de division et de regroupement (Callan & Eskenazi, 2007 ; Brouwers *et al.*, 2012). Certains travaux récents encouragent la combinaison des deux approches pour un meilleur rendement (Angrosh *et al.*, 2014), l'approche lexico-syntaxique que nous proposons va dans le même sens.

Problématique et objectifs

Notre étude s'inscrit dans le cadre de la simplification des textes médicaux, et plus précisément dans le domaine de la simplification lexico-syntaxique. Dans cette thèse, nous abordons la simplification de textes à travers le verbe. À la différence des travaux existant en simplification lexicale, focalisés uniquement sur les termes nominaux, nous nous intéressons au verbe dans sa nature terminologique, à ses arguments et au rôle que le prédicat verbal est susceptible de jouer dans la tâche de simplification des textes en langues de spécialité. En effet, la manière d'utiliser le verbe dans les textes spécialisés, plus précisément les différentes constructions dans lesquelles le verbe intervient, entraîne très souvent une variation sémantique qui peut poser des difficultés de compréhension pour un non-expert. Dans de tels cas, une simplification aiderait à faciliter l'accès au sens du verbe. Dans l'exemple ci-dessous, le verbe *relever* apparaît dans des contextes variés. Ces contextes se distinguent grâce aux types sémantiques d'arguments qui se combinent au verbe, et chacune de ces combinaisons (constructions) est associée à un sens différent que nous réussissons à reconnaître, simplifier, et exprimer en des termes plus simples, grâce à notre méthode d'analyse des corpus et de simplification.

Constructions spécialisées	**Sens du verbe simplifié**
Cette PROCÉDURE *relève des* SERVICES *d'urgences*	*'fait partie de'*
Ce PATIENT *relève d'une* AFFECTION *de longue durée*	*'souffre de', 'a'*
*L'*INFIRMIÈRE *relève la* TEMPÉRATURE *du patient*	*'prend', 'note'*
*L'*EXAMEN *relève une* MALADIE	*'révèle', 'signale'*

Tel que le montre l'exemple proposé ci-dessus, la proposition de verbes synonymes à l'issue d'une simplification permettrait à un « profane » du domaine médical de mieux comprendre le sens du verbe et de la phrase en général.

Cette thèse a donc pour but la création d'une ressource d'aide à la simplification des textes médicaux écrits par des experts pour le grand public, à partir d'une analyse contrastive des verbes dans des corpus médicaux ayant des niveaux de spécialisation différents. Notre travail de recherche étant focalisé sur les verbes en contexte, la ressource résultante devra avoir comme entrées des constructions verbales (similaires à celles de l'exemple), que nous appelons *patrons syntactico-sémantiques* (PSS), provenant d'un corpus de textes produits par des experts. Ces PSS utilisés par les experts seront alignés avec des PSS équivalents, c'est-à-dire des substituts provenant d'un corpus de non-experts et exprimant des sens quasi-synonymiques des verbes simplifiés, comme dans l'exemple. Pour ce faire, nous utilisons 4 corpus médicaux différents, visant 4 types de publics. Les textes des trois premiers corpus sont écrits par des experts, respectivement à l'attention des experts, des étudiants, et du grand public. Le quatrième corpus est composé de textes tirés d'un forum médical. Pour parvenir à l'objectif final de cette thèse, différents objectifs secondaires ont été atteints :

— L'annotation syntaxique des corpus : elle est réalisée grâce à l'analyseur syntaxique Cordial (Laurent *et al.*, 2009).

— L'annotation sémantique des corpus : elle est basée sur un ensemble de catégories sémantiques provenant de la terminologie médicale Snomed Internationale (Côté, 1996). Ces catégories associent des informations sémantiques aux arguments des verbes, à travers un processus d'annotation automatique.

— L'acquisition des patrons syntactico-sémantiques : la réalisation de cette tâche passe par une méthode semi-automatique qui est appuyée par une phase de validation manuelle des PSS. Cette validation est effectuée par trois groupes d'experts en médecine.

— L'analyse contrastive du fonctionnement des verbes (à travers les patrons syntactico-sémantiques et collocations verbe-terme) dans les différents corpus.

Motivations

Plusieurs raisons ont motivé le choix de notre sujet de thèse. Premièrement, l'intérêt que nous portons à la classe grammaticale des verbes, qui ont longtemps été mis à l'écart dans les travaux en terminologie, tout comme les autres catégories grammaticales. Notre projet de thèse s'inscrit dans la vague de travaux qui, selon L'Homme (2012b), essaient de caractériser la nature spécialisée du verbe, ou encore de mieux comprendre son aptitude à exprimer des sens spécialisés en fonction de son environnement linguistique. En effet, comme L'Homme & Bodson (1997), nous pensons qu'en tant qu'unité terminologique, le verbe est un bon point de départ

pour cerner la syntaxe et la sémantique des textes spécialisés, puisqu'il permet d'exprimer les connaissances véhiculées par les termes avec lesquels il cooccurre.

De plus, de nombreux travaux de recherche (Condamines & Bourigault, 1999 ; Fang, 2005 ; Deléger & Zweigenbaum, 2008 ; L'Homme, 2012b) démontrent que les écrits scientifiques tendent à contenir plus d'entités nominales que de verbes, tandis que les « profanes » sont enclins à utiliser les formes verbales lorsqu'ils s'expriment. Ce constat nous motive davantage à nous intéresser au fonctionnement des verbes (à travers les PSS) dans les textes spécialisés, par opposition aux textes non spécialisés, car une telle analyse permettrait de cerner les similarités, les divergences et surtout les spécificités qui caractérisent l'utilisation des verbes par chacun des principaux protagonistes du domaine médical. La simplification des constructions verbales utilisées par les experts permettrait donc de les adapter au langage des non-experts, lors de la rédaction des documents qui leur sont adressés.

En ce qui concerne le domaine de la simplification des textes, il n'existe pas, à notre connaissance, de travaux consacrés aux verbes. Aucune méthode de simplification n'est axée sur le sens des verbes comme la nôtre. Les ressources existantes (Grabar & Hamon, 2014), du moins pour le français, sont restreintes et concernent surtout les entités nominales. Aucune ne s'attèle exclusivement à l'alignement des patrons syntactico-sémantiques des verbes tel que nous le faisons. On mentionnera toutefois Deléger & Zweigenbaum (2008) et Deléger & Zweigenbaum (2009) dont l'étude implique un type particulier de paraphrases verbales (*la maladie est traitée*) provenant d'un corpus pour non-experts, et alignées avec des équivalents nominaux (*traitement de la maladie*) extraits d'un corpus pour experts. Cependant, cette étude ne s'intéresse pas au fonctionnement du verbe dans les textes spécialisés et par conséquent, ne propose pas de paraphrases verbales des experts comme équivalents des paraphrases verbales tirées du corpus des non-experts.

Une ressource comme celle que nous proposons pourrait être d'un apport considérable dans un travail de simplification. Elle peut notamment être intégrée comme ressource dictionnairique dans un outil de simplification de textes. De même, tout comme elle peut servir d'outil d'aide à l'encodage pour un expert qui essaie d'adapter son langage à un public de non-experts, elle peut être utilisée chez un non-expert comme dictionnaire d'aide pour le décodage de textes médicaux spécialisés.

Annonce du plan et présentation des chapitres

Ce document est organisé en cinq chapitres.

Le Chapitre 1, intitulé *État de l'art*, est consacré à la présentation de quelques approches de description du verbe dans les domaines à l'intersection desquels se situe ce travail de thèse, à savoir la linguistique, la terminologie et le traitement automatique des langues. Dans ce chapitre, nous introduisons également la méthode mise au point pour atteindre nos objectifs,

tout en présentant les notions centrales à ce travail de thèse : verbes spécialisés, simplification des textes, langues de spécialité, langue médicale, communication médecin/patient, etc.

Dans le chapitre 2, *Corpus, ressources et outils*, nous présentons les quatre corpus sur lesquels nous avons réalisé nos expériences. Les ressources et outils qui interviennent dans la chaîne de traitement des données acquises à partir des corpus sont également présentés. Il s'agit principalement de l'analyseur syntaxique Cordial et de la terminologie médicale Snomed Internationale. Les données analysées sont extraites de quatre corpus médicaux qui se distinguent grâce au niveau de spécialisation des auteurs et surtout des publics cibles.

Le chapitre 3, *Méthode*, est consacré à la description de l'architecture de la méthode semi-automatique appliquée dans ce travail de thèse pour l'analyse contrastive du fonctionnement des verbes et la création de la ressource de simplification. Nous faisons une présentation détaillée de la chaîne de traitement qui comporte cinq étapes : collection et pré-traitement des corpus, annotation des corpus et acquisition des PSS, sélection des PSS pour la validation, validation des PSS par les experts, création de la ressource pour la simplification. Outre ces tâches, qui contribuent toutes à la mise au point de cette ressource de simplification, ce chapitre présente la démarche appliquée lors de l'analyse contrastive des PSS et des collocations verbe-terme dans les différents corpus. Cette comparaison nous permet d'aborder les corpus du point de vue des collocations verbales, suite à une expérience que nous avons effectuée récemment (Wandji Tchami *et al.*, 2016) et qui a signalé l'importance d'une telle étude, surtout pour le domaine de la lexicographie et la conception des dictionnaires spécialisés.

Quant au chapitre 4, *Résultats et discussion*, il est consacré à la présentation des résultats obtenus au cours des différentes phases de la méthode. Bien évidemment, la ressource de simplification, qui représente le principal objectif de cette thèse, est également présentée dans ce chapitre. Elle sera mise à la disposition de la communauté scientifique.

Enfin le chapitre 5, intitulé *Évaluation des méthodes, ressources et outils utilisés*, est réservé à la description de la méthode utilisée pour évaluer les résultats des tâches fondamentales de notre travail de thèse, à savoir l'annotation syntaxique et l'annotation sémantique des corpus. De même, ce chapitre décrit la méthode et les résultats de l'évaluation de la ressource de simplification.

Chapitre 1

État de l'Art

1.1 Langues de spécialités et langue médicale

1.1.1 Typologies des textes en langue de spécialité

La question de la structure interne des différentes langues de spécialités (désormais *LS*) a toute une tradition dans les travaux de recherche germaniques, qui remonte aux années 1960. Cette époque a vu naître de nombreuses théories décrivant la langue de spécialité comme une structure, un système [1] qui se divise en deux dimensions :

— la dimension horizontale : elle suppose qu'il existe différentes disciplines au sein d'un domaine de spécialité (médecine, chimie, droit, informatique, mathématique, etc). Par exemple, le domaine de la médecine est constitué de plusieurs disciplines : l'anatomie, la psychiatrie, la physiologie, etc. Chacune de ces disciplines possède une langue spécialisée et un système linguistique qui la rendent autonome par rapport aux autres disciplines du domaine médical (Roelcke, 2010) ;

— la dimension verticale : elle décrit la langue de spécialité en termes de niveaux, qui varient selon les degrés de spécialisation. En effet, elle suppose qu'il existe dans la LS différents niveaux d'abstraction et de communication (entre les acteurs impliqués dans chaque discipline) au sein de chaque langue de spécialité, et que ces niveaux de communication diffèrent selon le caractère général ou particulier de la spécialité concernée et des connaissances liées (Roelcke, 2010).

Malgré les différentes critiques [2] (Lerat, 1995) qui ont été formulées contre ce modèle de description de la LS, nous en retiendrons quelques éléments qui cadrent avec notre travail de recherche. La dimension verticale nous intéresse particulièrement car elle se focalise sur la communication entre les différents acteurs [3] d'un domaine de spécialité, ce qui représente l'un des éléments fondamentaux de notre étude. En effet, la notion de verticalité présuppose l'existence dans une langue de spécialité de plusieurs niveaux de technicité dont deux représentent les bornes : un niveau général, abstrait, avec une faible spécialisation (niveau des « profanes ») qui s'oppose à un niveau plus particulier, plus concret, très spécialisé (niveau des experts). Cette conception débouche sur différentes typologies de langues, de communications et de textes techniques (Ischreyt, 1965 ; Möhn & Pelka, 1984 ; Hoffmann, 1985 ; Gläser, 1990 ; Göpferich,

1. Cette conception systémique de la LS transparaît par exemple dans la définition suivante proposée par la norme ISO 1087 : « langue de spécialité : sous-système qui utilise une terminologie et d'autres moyens linguistiques et qui vise la non-ambiguïté de la communication dans un domaine particulier ». Citation tirée de (Gautier, 2014).

2. L'objectif de ce travail n'est pas de prendre position par rapport à ces critiques. Par conséquent, nous n'allons pas nous y attarder.

3. Ces acteurs sont présentés comme appartenant à différents niveaux de communication qui sont déterminés par leurs niveaux de spécialisation.

1995 ; Lothar *et al.*, 1998 ; Nickel, 1999), parmi lesquelles celle de Roelcke (2014) qui distingue 5 différents types de communications :

— la communication entre experts de différents domaines de spécialités ;

— la communication entre experts d'un domaine particulier ;

— la communication entre différents experts d'un même domaine mais ayant différents niveaux d'expertise (Exemple : médecin vs. infirmier) ;

— la communication entre experts d'un domaine et non-experts impliqués dans ce domaine (Exemple : médecin vs. patient) ;

— la communication entre non-experts d'un domaine particulier (Exemple : communication entre les internautes sur un forum).

Le domaine médical est utilisé pour illustrer cet aspect de la langue de spécialité. Lothar *et al.* (1998) expliquent que la communication dans ce domaine est caractérisée par 3 niveaux notamment le niveau scientifique (chercheurs vs. chercheurs), le niveau professionnel (médecins vs. médecins/personnel soignant), et le niveau patient (médecins vs. patients) ; et qu'elle est susceptible d'être sujette à des conflits causés par les différents niveaux d'expertise des acteurs.

Comme l'on pouvait s'y attendre, les différentes typologies de communication débouchent à leur tour sur différentes classifications des textes en langues de spécialités correspondants, déterminées par le degré de spécialisation. Les typologies présentent des variations, car certaines couvrent un/plusieurs domaine(s) de spécialité(s) (Göpferich, 1995), tandis que d'autres sont plus spécifiques, et portent sur les branches d'une discipline particulière (Lothar *et al.*, 1998). Trois principaux types de textes techniques résultent de cette classification :

— les textes experts (dégré de spécialisation élevé) ;

— les textes semi-experts (dégré de spécialisation moyen) ;

— les textes profanes (dégré de spécialisation faible).

Ces trois types de textes correspondent aux types de communication écrites qui caractérisent le domaine médical. À ce sujet, Feyrer (2016) souligne que la dimension verticale de la LS sert de distinction pour les différents types de textes médicaux. L'auteur cite les exemples suivants : textes de la recherche scientifique, textes de la formation, textes de la pratique médicale, et textes de vulgarisation. La typologie des textes spécialisés qu'offre la dimension verticale capture la structure de notre corpus (cf. chapitre 2, section 2.1.1), qui est constitué de 4 types de textes différenciés par le niveau de spécialisation des auteurs et des publics cibles.

1.1.2 Discours scientifique ou Langue de spécialité

Une langue de spécialité désigne « un sous-système linguistique correspondant à l'emploi de la langue dans un domaine de connaissances unique » (Pecman, 2007). Autrement dit, une

langue de spécialité comme la langue médicale permet d'exprimer les connaissances propres à un domaine scientifique, dans le cas présent, la médecine. D'après Pecman (2007), au-delà des spécificités terminologiques propres à chaque discipline, il existe un certain nombre d'invariants à travers les différents discours scientifiques. Ces éléments caractérisent le discours scientifique universel qu'elle nomme la LSG, c.-à.-d. *Langue Scientifique Générale*. Il s'agit « d'une pratique langagière spécifique à une communauté de discours composée de chercheurs en sciences exactes et sciences humaines dont les objectifs communicatifs émanent de préoccupations partagées par des scientifiques à travers le monde et indépendamment de leurs spécificités disciplinaires » (Pecman, 2004). Cavalla (2010) identifie quatre dimensions ou caractéristiques des écrits scientifiques (textes en langue de spécialité) :

1. la dimension scientifique : elle concerne le savoir scientifique, qui peut varier d'une communauté à l'autre ;
2. la dimension méthodologique : elle décrit l'architecture, la structure de l'écrit. L'auteure montre par exemple que la structure des écrits universitaires répond à des normes très précises d'un pays à l'autre, et que d'une discipline à l'autre, les normes rédactionnelles sont différentes ;
3. la dimension terminologique : elle permet d'ancrer un écrit dans une discipline donnée, à l'aide notamment d'un lexique spécialisé (la terminologie) ;
4. la dimension linguistique : elle tente de donner les structures linguistiques qui contribuent à l'élaboration du sens scientifique. Il existe des codes linguistiques dans chaque spécialité et il est indispensable de les connaître afin de mieux comprendre le système linguistique concerné. Par exemple, l'emploi du passif avec agent absent, l'utilisation du pronom personnel à la troisième personne du singulier et du pronom indéfini « on », etc. sont selon Pecman (2004) des procédés linguistiques qui caractérisent la LSG.

Ces éléments, qui constituent la LSG, se retrouvent de façon relative dans les différentes langues de spécialité à l'instar de la langue médicale qui est l'objet d'étude de ce travail de thèse.

1.1.3 Langue médicale

Chaque discipline ou profession possède une langue qui lui est propre, un système linguistique qui permet de véhiculer les connaissances du domaine. Il s'agit de ce que Christy (1979) nomme *Ingroup language*, c.-à-d. une langue parlée autant à l'oral qu'à l'écrit, au sein d'un groupe de personnes qui partagent des connaissances communes d'un domaine de spécialité. Cette langue, qui est très souvent caractérisée d'*hermétique aux non-initiés* (Faure, 2010), relève du besoin de communiquer de façon efficace et rapide. D'après la littérature, ce caractère impénétrable de la langue est particulièrement frappant en médecine, où différentes dénominations sont données à la langue parlée par les professionnels de la santé, dans le but d'indiquer cette spécificité :

Faure (2010) parle de *medspeak* ou encore de *medicalese* en ce qui concerne l'anglais. Pour Christy, la langue parlée dans les hôpitaux aux USA n'est pas l'Anglais : "What is spoken on rounds is not English". L'auteur fait cette déclaration dans un ouvrage intitulé "English is our [c.-à-d. les médecins] second language" (Christy, 1979), un titre fort révelateur.

Dans une étude qui fait la description du langage médical anglais des XIX^{e} et XX^{e} siècles, Brunt (2008) rejoint Christy (1979), en déclarant que les médecins et infirmiers parlent une langue qui est tout sauf l'anglais. Il la nomme *hospital English* c.-à-d. l'anglais de l'hôpital. Cet anglais hospitalier est caractérisé par l'utilisation de termes médicaux [4], ainsi que d'autres mots de la langue, mais avec des significations différentes de celles qui sont connues du grand public. Par exemple, d'après les investigations menées par l'auteur, chez les professionnels de la santé, les termes *incompetence* (incompétence) et *failure* (échec) sont employés dans des contextes similaires (synonymiques), ce qui n'est pas le cas dans la langue générale, où le deuxième terme (*failure*) a tendance à être la conséquence du premier (*incompetence*). De même, d'après cette étude, on dit d'un patient qu'il est *toxique* lorsqu'il souffre de la *toxaemia*, tandis que l'on définit la *décompensation* comme *l'incapacité d'un organe du corps d'assurer sa fonction*. Le constat de l'auteur ne concerne pas uniquement les entités nominales, mais les verbes aussi : on dit d'une maladie qu'elle est *documented* (documentée) lorsqu'il a été prouvé sur la base de tests et d'examens objectifs que cette maladie existe. Un patient est considéré comme ayant *échoué un traitement* s'il ne manifeste aucune réaction en réponse à ce traitement ; *to consent patients* (consentir les patients) signifie obtenir leur consentement pour une opération ; *to explore a patient* (explorer un patient) c'est lui faire subir un ensemble d'examens préliminaires ; *to follow a patient* (suivre un patient) c'est de lui offrir un suivi et des soins de façon continue.

En plus de ces pratiques linguistiques, Brunt (2008) décrit un ensemble de techniques et de figures de style utilisées par les membres du corps médical, afin de communiquer entre eux dans certaines circonstances délicates. Par exemple, des expressions relevant de l'euphémisme sont utilisées, lorsqu'ils sont face à un cas de patient souffrant extrêmement, ou lorsque la mort du patient est imminente. On parlera de *adverse event* (évènement indésirable) lorsque l'on a affaire à une blessure résultant d'une erreur médicale, plutôt que causée par la maladie sous-jacente [5]. D'après l'étude de Brunt (2008), le besoin de communiquer de façon effective et rapide pousse le personnel soignant à faire recours à d'autres figures de style, notamment la conversion et la coupure. Par exemple, *to take blood samples from a patient* devient *to blood*. En d'autres termes, le néologisme *to blood* est utilisé pour faire référence à la procédure médicale qui consiste à prélever des échantillons de sang d'un patient.

Certains éléments caractéristiques du langage médical décrit ci-dessus se retrouvent également dans une étude qui a pour but de comparer le discours médical dans différentes langues :

4. Dans cette étude, le mot *terme* désigne toute unité terminologique d'une langue de spécialité.
5. Exemple emprunté à Segen (1992), repris par Brunt (2008).

l'anglais opposé à quelques langues indo-européennes (Faure, 2010). L'auteure observe que la langue de la médecine est marquée par des procédés stylistiques (troncation, initialisme, métonymie, métaphore, etc.), qui sont dictés par des besoins de communication spécifiques à la médecine (concision, exactitude ou encore discrétion). En d'autres termes, la pratique de la médecine impose aux professionnels certaines techniques de communication pragmatiques, c.-à-d. déterminées par le contexte et la situation de communication.

En ce qui concerne le lexique, l'étude diachronique de Faure (2010) nous apprend que la nomenclature de base de la terminologie médicale résulte de la terminologie latine, qui remonte à la période de la Renaissance [6]. Les différentes nomenclatures des différents pays ont été conçues à l'origine à partir de traductions du latin. Toutefois, les terminologies médicales sont très influencées par la culture des différents pays utilisateurs. Ce constat s'observe à travers la structuration des différents systèmes de santé, et les dénominations de certaines maladies [7]. Les codes [8], très utilisés dans la langue médicale, varient d'une langue à l'autre, etc.

En plus de celles déjà mentionnées, les textes médicaux sont marqués par d'autres caractéristiques discursives, et syntaxiques. Nous avons adopté la liste de propriétés discursives que propose Fleischman (2003) [9] et qui proviennent de l'analyse des historiques médicaux de patients d'hôpitaux :

— la dépersonalisation : aucun lien explicite n'est établi entre le patient en question et le texte qui présente son historique médical. Pas de nom, mais à la place, des mots ou termes généraux du type *la patiente*, *elle*, etc.

— l'attribution de la fonction d'agent aux termes réfèrant aux procédures médicales.

— l'utilisation des prédicats non assertifs : *dire*, *signaler*, *réfuter*, etc.

— la voix : qui parle dans le texte ? et le point de vue : le texte est rédigé selon le point de vue de quel acteur : le patient ? le médecin ? les deux ?

— l'omission de l'agent : préférence pour les constructions passives avec agent absent. L'auteure souligne que ce procédé a pour effet de mettre l'accent sur ce qui a été fait, plutôt que sur qui l'a fait. En général, le discours scientifique médical est rédigé tel que les faits parlent d'eux-mêmes, c.-à-d. avec des efforts pour éliminer tout indicateur de subjectivité.

6. Pendant la Renaissance, les travaux majeurs en médecine étaient systématiquement traduits du grec ou de l'arabe vers le latin.

7. Quelques exemples proposés par Faure (2010) : en allemand, la syphilis peut être nommée *Franzosenkrankheit* (littéralement 'la maladie des Français') car elle leur fut longtemps attribuée dans le passé historique. Le rachitisme quant à lui est appelé *Englische Krankheit* (littéralement 'la maladie anglaise') car d'après l'histoire les Allemands découvrirent cette maladie pour la première fois durant la révolution industrielle en Grande-Bretagne.

8. Par exemple, le 15 permet de contacter le SAMU (service d'urgences) en France, tandis qu'aux États-Unis, c'est le 911, et en Grande-Bretagne, le 999.

9. L'auteure reprend le travail de Anspach (1988).

La forte fréquence des nominalisations, l'utilisation des faux amis, l'emploi du mode impersonnel, des mots composés, des syntagmes et phrases complexes font également partie des caractéristiques lexico-syntaxiques des textes médicaux (Kandula *et al.*, 2010). La structure de surface des textes est également concernée, avec la longueur des mots, des phrases et des paragraphes (Zeng-Treiler *et al.*, 2007). Quelques travaux récents considèrent la flexion et le mode du verbe comme éléments caractéristiques des textes spécialisés (Stein *et al.*, 1992 ; Da Cunha *et al.*, 2011). Ces travaux se limitent au niveau grammatical, aucun ne s'intéresse à la structure argumentale du verbe, encore moins à la nature des actants du verbe, qui est pourtant un critère déterminant dans la description des textes de spécialité et surtout dans le repérage des emplois verbaux spécialisés. Dans notre travail, cette approche axée sur les verbes et leur structure argumentale (cf. section 1.3.2.2) constitue la base de la méthode de sélection des patrons syntaxico-sémantiques candidats pour la ressource de simplification.

1.2 Communication médecins vs. patients

Le domaine médical est caractérisé par sa tendance à faire interagir différents acteurs, principalement les médecins et les patients. Ces protagonistes du domaine médical, bien que différents de par leurs statuts sociaux, leurs formations professionnelles, leurs cultures, leurs niveaux intellectuels et, surtout, par leurs compétences et niveaux d'expertise en matière de connaissances médicales, sont appelés à communiquer pour le succès de leur interaction, et leur satisfaction respective.

En effet, les travaux de recherche en médecine montrent qu'une communication effective et efficace avant, pendant et après un traitement est essentielle pour les deux partenaires et, surtout, pour la sécurité du patient. En plus d'instaurer un climat de confiance, elle améliore la qualité de vie du patient, le suivi du traitement, les résultats cliniques et sa satisfaction, ainsi que celle des médecins (Williams & Ogden, 2004 ; Fournier & Kerzanet, 2007). Si la communication est inexistante ou insuffisante, il peut en découler déceptions, incompréhensions et malentendus, et le patient peut se sentir livré à lui-même avec ses questions et ses problèmes de santé. Pire encore, le processus de soin peut être compromis, voire déboucher sur l'échec du traitement. Cet argument est d'autant plus important qu'il fait l'objet de plusieurs travaux de recherche dans le domaine médical. Nous faisons plus particulièrement allusion à un ensemble d'études (Berkenkotter *et al.*, 2015 ; Charpy, 2015) qui ont été menées dans le domaine de la psychiatrie, décrivant différentes situations dans lequelles les problèmes de communication, et plus précisément d'inter-compréhension entre médecins et patients ont débouché sur l'échec des diagnostics posés par les professionnels de la santé.

La communication entre les patients et les médecins fait ainsi partie des principales thématiques de recherches effectuées de nos jours dans le domaine des sciences humaines et sociales et

de la médecine. Qu'elle soit écrite ou orale, cette communication fait face à différents défis d'ordre social, culturel, professionnel, intellectuel et surtout linguistique, qui la rendent difficile et, parfois, non effective. En effet, le langage médical spécialisé (cf. section 1.1.3) est très souvent difficile à comprendre pour les non-experts. De nombreux patients en souffrent car ils ne comprennent pas toujours les informations ou instructions qui leurs sont adressées (Lerner *et al.*, 2000 ; Chapman *et al.*, 2003). Pour des raisons qui seront expliquées par la suite, nous pensons que l'impact de ces défis de communication est plus important lorsqu'il s'agit d'une communication écrite, raison pour laquelle la réussite de ce type de communication requiert une forte implication des experts en médecine qui rédigent des textes à l'attention de patients ou du grand public.

1.2.1 Communication orale

La communication orale médecin-patient est une problématique qui demeure au coeur des travaux de recherche en sciences humaines et en médecine depuis plusieurs années. De nombreuses études ont vu le jour, la plupart décrivent la communication médecin-patient comme la clé du succès du processus de soins médicaux (Buller & Buller, 1987 ; Beckman *et al.*, 1989 ; Bensing, 1991 ; Nciri, 2009). D'autres chercheurs vont plus loin en montrant que la qualité de cette communication peut avoir un impact considérable sur la vie du patient (Roter & Hall, 2006).

Dans un état de l'art des travaux portant sur la communication médecin-patient (Ong *et al.*, 1995), la relation médecin-patient est présentée comme étant l'une des plus complexes relations interpersonnelles de par les éléments qui la caractérisent : elle fait interagir des individus ayant différentes situations sociales, professionnelles, familiales, financières, etc. ; le plus souvent, elle n'est pas choisie et porte sur des sujets intimes et d'une importance vitale, précisent les auteurs. Feyrer (2016) décrit la communication médicale comme un « écosystème complexe » déterminé par des facteurs sociaux, culturels et pragmatiques, parmi lesquels on peut nommer le caractère institutionnel de la communication médicale, les connaissances spécialisées, les compétences linguistiques des participants, ainsi que leurs identités sociales. Cette combinaison d'éléments fait de la communication médecin-patient une communication par nature délicate. Pourtant, son efficacité est indispensable pour la réussite des soins et le bien-être du malade. Depuis plusieurs décennies, de nombreux travaux de recherche s'attellent à trouver des méthodes pour améliorer cette communication afin de la rendre plus efficace. Ces études s'intéressent le plus souvent au but de la communication (Pendleton *et al.*, 2003 ; Fournier & Kerzanet, 2007), aux différentes techniques de communication utilisées par les médecins pendant leurs échanges avec les patients, ainsi qu'à l'impact de ces techniques sur les résultats cliniques des patients (Blanchard *et al.*, 1983 ; Roter *et al.*, 1987 ; Hall *et al.*, 1988 ; Meeuwesen *et al.*, 1991).

Les expériences effectuées dans le cadre de ces différents travaux, plus particulièrement ceux qui portent sur les techniques de communication, montrent qu'il existe des moyens de rendre un

échange communicationnel oral médecin-patient plus aisé et satisfaisant pour les deux parties. Par exemple, le caractère interactif d'une conversation en face-à-face offre aux participants la possibilité d'avoir recours à des techniques de *négociation du sens* (Kida, 2001) (questions, signaux faciaux, gestuelle, etc.) qui leur permettent, en cas d'incompréhension, de saisir le sens du message véhiculé. Ceci n'est pas possible dans le cadre de la communication écrite, dont la prise de connaissance (par le lecteur) est souvent effectuée en mode non présentiel. D'après certains chercheurs, son succès repose en grande partie sur les connaissances du rédacteur et sur son aptitude à adapter son message au public cible (Zeng-Treiler & Tse, 2006).

Dans une étude portant sur les procédés de communication utilisés par les médecins lors des entretiens avec les patients, Grecu (à paraître) compare les dialogues de consultations médicales de deux groupes de patients : des adultes et des enfants. L'auteure observe qu'au cours des consultations avec les patients d'un groupe particulier, les médecins tendent à utiliser différentes techniques de communication (ton, intonation de la voix, choix des mots, etc.) afin d'adapter leur langage au type de patient interrogé (adultes, enfants). Les patients, quant à eux, ont tendance à manifester leur incompréhension via différents signaux verbaux et/ou non verbaux (expression faciale, mouvement de la tête, etc.) qui poussent les médecins à reformuler leurs messages de façon plus compréhensible. Ce processus de va-et-vient entre médecins et patients correspond à la technique de négociation de sens à laquelle les interlocuteurs ont recours, facilitant ainsi l'intercompréhension lors d'une communication orale.

Par ailleurs, l'avancée des recherches a permis depuis plusieurs années de mettre en place des outils d'aide à la communication médecin-patient. C'est le cas de la gamme *Outils de communication I - une meilleure communication médecin-patient pour de meilleurs résultats auprès des patients* (Canada, 1998), publiée par le groupe *SFP*[10] (Stratégie sur la formation des professionnels) dans le cadre de l'initiative canadienne sur le cancer du sein. Cette initiative met à la disposition des médecins une variété de ressources centrées sur les compétences et techniques de communication. Il s'agit de cours, destinés aux praticiens de la médecine, qui ciblent des habiletés spécifiques à acquérir pour une communication efficace avec les malades (Canada, 2001). Les bons résultats produits par l'utilisation d'*Outils de communication I* ont débouché, à la demande des praticiens, sur la mise en place d'une deuxième version appelée *Outils de communication II* (Canada, 2001). Cet outil, constitué de deux cours, se concentre sur des compétences et des techniques de communication professionnelles (à savoir : *Comment aider le patient à s'exprimer plus facilement et comment prendre des décisions ensemble ? Comment faire face aux émotions du patient et comment clore sur une bonne note ?*) qui peuvent être acquises, perfectionnées et appliquées dans le cadre des rapports médecin-patient.

Les techniques et outils présentés ci-dessus permettent d'améliorer la qualité de la commu-

10. http ://www.groupe-sfp.com/

nication en face-à-face entre médecin et patient et de la rendre satisfaisante pour les deux acteurs. Cependant, ces méthodes n'impliquent pas la communication écrite qui se confronte à des défis encore plus complexes.

1.2.2 Communication écrite

Contrairement à la communication orale, le type de communication écrite qui nous intéresse principalement, à savoir les textes à valeur didactique [11] et informative écrits par les médecins pour d'autres experts ou pour le grand public, et diffusés sur internet, se caractérise par l'absence de ce que Marcoccia (2000) appelle le *non-verbal*. Il s'agit de l'ensemble des procédés et matériaux sémiotiques utilisés dans une communication en face-à-face et qui contribuent à la rendre effective : les marqueurs personnels (l'apparence physique), les données paraverbales (ton, intonation, rythme d'un énoncé, etc.) et les données non verbales (la gestuelle, les expressions faciales, etc.). De plus, le caractère formel de ce type de communication réduit davantage les possibilités (techniques dont dispose le rédacteur) de combler le vide laissé par l'absence du verbal, comme cela se fait dans d'autres types de textes numériques tels que les e-mails, les messages instantanés, ou les messages de forums (Marcoccia, 2004).

Dans une étude portant sur la représentation du non-verbal dans la communication écrite médiatisée, Marcoccia (2000) explique qu'une bonne partie des possibilités offertes par la communication en face-à-face disparaît avec la communication écrite médiatisée par ordinateur, en l'occurrence avec les messages de type forums, e-mails, etc. L'auteur relève et analyse différentes formes de représentation du non-verbal et du paraverbal dans des textes de forums de discussion francophones (à savoir les smileys, les autoportraits, la ponctuation expressive, les lettres capitales, et les commentaires métadiscursifs) et ainsi parvient à montrer que ces procédés n'ont pas seulement une valeur ludique, esthétique ou distinctive mais qu'ils recouvrent les fonctions qu'ont le non-verbal et le paraverbal dans la communication en face-à-face (Marcoccia, 2004). Il est important de remarquer que les textes analysés dans ce travail de Marcoccia sont de type informel, un statut qui autorise l'utilisation des symboles comme les smileys et la ponctuation expressive. Or, 2/3 des textes médicaux étudiés dans le cadre de notre projet de thèse sont certes numériques, mais ont un caractère formel [12], ce qui exclut l'application des procédés proposés ci-dessus, utilisés pour combler l'absence du non-verbal.

Dans les médias écrits, tels que les journaux, les brochures, les magazines, etc., la communication n'est pas toujours évidente car elle repose principalement sur les connaissances et points

11. Dans cette étude, nous nous intéressons aussi aux textes des forums de discussion, c.-à-d. des textes informels (cf. section 2.1.3.4, chapitre 2), mais ils interviennent en seconde position, après les textes formels dont provient notre problématique, à savoir, la difficulté de lecture (cf. section 3.3.2).

12. Dans ce travail, les 2/3 de notre corpus sont composés de textes écrits par des experts en médecine soit pour des experts, soit des étudiants, soit pour des patients.

de vue de l'auteur seul, dans notre cas les experts en médecine (Zeng-Treiler & Tse, 2006). Le succès de cette forme de communication dépend donc du type de langage utilisé par le rédacteur. À ce sujet, la plupart des travaux de la littérature identifient l'utilisation des termes nominaux spécialisés comme la principale cause de difficulté de lecture des textes de spécialité (Lerner *et al.*, 2000). Or la difficulté de lecture ne concerne pas uniquement les entités nominales, mais peut aller bien au-delà, en portant non seulement sur d'autres catégories grammaticales, mais aussi sur la structure externe et interne du texte. Certains travaux de recherche effectués récemment (Da Cunha *et al.*, 2011 ; Todirascu *et al.*, 2012) proposent une grande grille de propriétés caractéristiques des textes scientifiques, parmi lesquels des textes médicaux. Cette liste prend en considération des propriétés de différents types, qui impliquent les noms, les adverbes, les adjectifs et également les verbes et d'autres catégories : propriétés statistiques (longueur des mots et des phrases), propriétés lexicales (le fonctionnement des adverbes, des adjectifs et des verbes), propriétés morphosyntaxiques (la complexité des syntagmes et phrases). Une autre étude (Tellier, 2008) propose une approche de description des sens verbes médicaux spécialisés dans des articles dictionnairiques, afin de répondre au problème de difficulté de compréhension et d'interprétation de différents emplois de verbes dans un corpus médical du domaine de l'infectiologie.

Lorsqu'il est employé dans des constructions ou contextes spécialisés, le verbe peut contribuer à rendre la compréhension des textes spécialisés, en l'occurrence des textes médicaux, difficile pour un lecteur non-expert. Ce point de vue est l'un des principaux arguments que nous défendons dans cette thèse. En effet, contrairement à beaucoup de noms utilisés dans le vocabulaire spécialisé, le verbe est un élément relationnel, en d'autres termes, sa syntaxe exige qu'il soit accompagné d'arguments qui lui sont subordonnés et qui déterminent grandement son sens (cf. section 1.3.1.1). Par conséquent, s'intéresser au verbe en tant que source de difficulté de lecture des textes médicaux signifie analyser les constructions (les différentes structures argumentales) dans lesquelles il apparaît. Ceci signifie que l'efficacité d'une communication écrite médecin-patient dépend non seulement du choix des termes nominaux utilisés par l'expert qui rédige le texte, mais aussi du choix minutieux des constructions verbales à employer dans un texte adressé à un public de non-experts.

Plusieurs travaux de recherche en médecine étudient les problèmes de la communication écrite entre les médecins et les malades ou le grand public (AMA, 1999 ; Zeng & Parmanto, 2003 ; Zeng-Treiler *et al.*, 2007 ; Kharrazi, 2009). Ils décrivent entre autres la complexité des informations adressées aux patients par les médecins via les sites Web (McCray, 2005 ; Tran *et al.*, 2009). Les recherches en linguistique, sociologie et anthropologie ont permis de relever différents éléments qui favorisent les difficultés de communication entre médecins et patients. Par exemple, dans une étude qui promeut la conception de dictionnaires médicaux pour patients, Zeng-Treiler & Tse (2006) comparent le langage médical des médecins à celui des patients. Il

en ressort que :

— Les experts en médecine et les patients ne s'expriment pas de la même façon (Zielstorff, 2003). Ils utilisent des expressions ou phrases différentes pour décrire les mêmes concepts médicaux.

— Les experts et les patients utilisent parfois des expressions similaires ou identiques mais avec des interprétations différentes :

> Among the worst case scenarios, consumers will misinterpret or (mis)associate a term with context or connotations not intended by the content provider or author. Alternatively, consumers may recognize or use a technical form, but associate it with a concept in lay usage rather than one from a professional health care domain (Zeng-Treiler & Tse, 2006).

Ce constat est également fait par Chapman *et al.* (2003) dans le cadre d'un exercice donné à des patients souffrant du cancer, visant à évaluer leurs connaissances par rapport aux termes et concepts médicaux utilisés par les médecins lors de différentes consultations.

— Face à un patient illettré ou natif d'une langue autre que celle parlée par le médecin, la difficulté est encore plus importante car le fait que le médecin ne puisse pas traduire les informations dans la langue du patient entraîne un risque d'incompréhension totale. Cette situation est très souvent observée dans les régions multilingues. C'est ce que décrit Putz (2008) dans une étude qui expose les défis linguistiques caractérisant le système des soins médicaux au Tyrol du Sud, une région qui se distingue par la coexistence de trois langues : l'allemand, l'italien et le ladin [13]. Les résultats de l'étude, effectuée sur 9 médecins, mettent en évidence les difficultés linguistiques que ces derniers affrontent en permanence dans l'exercice de leurs fonctions : une bonne partie des patients (particulièrement les personnes âgées et les enfants) parlent des dialectes de l'allemand qui diffèrent de l'allemand formel, obligatoirement enseigné aux professionnels de la santé avant leur prise de fonction en Tyrol du Sud. Le problème ici souligné concerne certes l'oral, mais il touche également la communication écrite qui souffre des différents conflits qui caractérisent l'interaction médecin-patient.

En résumé, comme l'expliquent Zeng-Treiler & Tse (2006) dans la citation ci-dessous, la communication médecin-patient oppose deux différents groupes de personnes : d'un côté, un groupe de personnes qui partagent des connaissances communes du domaine médical (apprises dans le cadre de leur formation) et qui utilisent un vocabulaire standard et normalisé par le biais de terminologies ; et de l'autre, un groupe de personnes qui diffèrent par leur origine, leur culture,

13. Le ladin est une langue romane qui est parlée en tant que langue maternelle dans certaines régions du nord-est de l'Italie.

leur éducation, etc. Ce deuxième groupe de personnes, c.-à-d. celui des patients, utilise un langage résultant d'un mélange de langue courante et de terminologie médicale, puisqu'ils sont en contact permanent avec le langage médical dans le cadre du suivi des soins. Paralèllement, ces patients appartiennent à des sociétés qui ont différentes visions du monde, différentes cultures, différentes croyances et systèmes de pensées, qui influencent considérablement leurs interprétations du discours des experts. Cette différence génère un conflit au niveau de la communication et de l'intercompréhension.

> While health care specialists share foundational domain knowledge based on formal education and professional experience, laypersons have some socially and culturally derived notions of health and illness acquired from formal and informal sources (e.g., media exposure) and unique personal experiences [...]

Les travaux décrits ci-dessus montrent qu'il existe un réel besoin d'aider les patients en trouvant des moyens pour faciliter leur compréhension du langage médical (McCray, 2005 ; Borin *et al.*, 2007a). Des recherches ont été réalisées en sociologie et en informatique médicale dans le but de cerner les spécificités de cette communication afin de la rendre plus aisée. Effectués manuellement (Kharrazi, 2009 ; Chy *et al.*, 2012) ou automatiquement (Kokkinakis & Toporowska Gronostaj, 2006 ; Chmielik & Grabar, 2011), ces travaux, centrés essentiellement sur l'étude du niveau conceptuel et lexical de la communication entre patients et médecins, ont permis de signaler une différence importante observée au niveau du lexique (Smith & Wicks, 2008). Partant de ce constat, ces auteurs préconisent une simplification du vocabulaire des médecins afin de l'adapter au savoir des patients (Bouhaddou & Warner, 1994 ; Waisman *et al.*, 2003 ; White *et al.*, 2004). D'autres recherches en Traitement Automatiques des Langues (désormais *TAL*) et en informatique médicale poussent la réflexion plus loin, en proposant et en constituant des lexiques contenant des paires de termes médicaux spécialisés et non spécialisés alignés (Zielstorff, 2003 ; Zeng-Treiler & Tse, 2006 ; Deléger & Zweigenbaum, 2008). C'est dans ce cadre de simplification de textes médicaux que s'inscrit le présent travail de recherche.

Comme les études mentionnées supra, la plupart des travaux en simplification de textes de la langue générale et/ou spécialisée se focalisent uniquement sur les entités nominales, au détriment des autres parties du discours, en l'occurrence du verbe, dont l'étude du fonctionnement en contexte spécialisé s'avère à notre avis un apport considérable, non seulement pour la simplification de textes mais également pour d'autres tâches. C'est la raison pour laquelle la tâche de simplification de textes que nous entreprenons dans ce travail de thèse se démarque des autres travaux. Nous proposons une méthode de simplification de textes médicaux basée non pas sur les termes nominaux mais sur les verbes associés à leur structure argumentale. Cette tâche sera davantage décrite dans la section suivante.

1.2.3 Bilan

Dans cette première section de notre premier chapitre, nous avons décrit la communication entre médecin et patient en présentant les types de défis auxquels les protagonistes de cette communication sont très souvent confrontés. Il en est ressorti que dans ce contexte assez particulier, le succès de la communication dépend grandement de l'habileté du médecin à adapter son discours au public cible. De façon immédiate, la conception des outils et ressources de simplification des textes médicaux, en l'occurrence la nôtre, est une solution pour participer à l'amélioration de cette communication, en assistant le patient dans la lecture de textes médicaux qui lui sont adressés. Cette assistance passera par la proposition des patrons verbaux alternatifs, substituts ou équivalents des emplois verbaux spécialisés et difficilement compréhensibles.

1.3 Le verbe dans les travaux de recherche

Cette section sera consacrée à la description des travaux de recherche en linguistique, terminologie et TAL, qui octroient une place privilégiée au verbe. N'ayant pas la prétention de parcourir tous les travaux de la littérature, nous nous focaliserons uniquement sur quelques théories, approches et méthodes susceptibles d'intervenir dans le cadre de cette étude. Un accent particulier sera mis sur les approches qui ont été sélectionnées et exploitées pour la réalisation de ce travail.

1.3.1 Linguistique

La linguistique est considérée comme notre point de départ car de nombreux travaux du TAL et certaines approches de la terminologie s'en inspirent. En effet, les théories et approches linguistiques ont une grande incidence sur les tâches du TAL (cf. section 1.3.3), et les approches de la terminologie. Par exemple, l'approche lexico-sémantique (cf. section 1.3.2.2), sur laquelle notre méthode s'appuie grandement, trouve son origine dans la sémantique lexicale qui est une théorie linguistique.

En linguistique, de nombreux cadres théoriques placent le verbe au coeur de leurs travaux. C'est ce qui explique l'existence de diverses approches de description du verbe (Tesnière, 1959 ; Vendler, 1967 ; Grimshaw, 1990 ; Jackendoff, 1990). Nous nous intéressons plus particulièrement aux théories linguistiques qui présentent le verbe comme un constituant recteur (c'est-à-dire un élément dont la réalisation syntaxique et sémantique dépend grandement de la présence d'autres constituants qui lui sont subordonnés) et qui décrivent son rapport avec les autres constituants de la phrase. N'étant pas en mesure de fournir une description exhaustive de toutes les approches existantes, nous nous limiterons à celles qui servent de bases à la réalisation de certaines tâches du TAL, et/ou de la terminologie, et plus particulièrement celles qui pourraient être nécessaires à la conception de notre ressource de simplification. Ainsi, dans cette partie,

nous parlerons de la théorie de la valence verbale (cf. section 1.3.1.1) et de la sémantique des cadres (cf. section 1.3.1.3).

1.3.1.1 Structure actantielle et valence verbale

La syntaxe structurale est une approche fondée par Lucien Tesnière (Tesnière, 1959) au début des années soixante ; elle est la première à avoir mis le verbe au centre de la phrase. En syntaxe structurale, l'ensemble des mots d'une phrase constitue une véritable hiérarchie au sein de laquelle les constituants sont liés les uns aux autres par des liens de dépendance. C'est dans cette approche qu'apparaît la notion de « dépendance » pour désigner plus proprement la relation de subordination. Tesnière le précise en ces termes dans son ouvrage : « [les] connexions structurales établissent entre les mots des rapports de dépendance » (Tesnière 1959, p. 13).

La phrase, encore appelée *stemma*, est décrite comme étant un schéma arborescent, ou un ensemble de noeuds. Le noeud, quant à lui, désigne un ensemble constitué par un régissant et tous ses subordonnés. Le subordonné *dépend* du régissant – inversement, le régissant *commande* ou *régit* le subordonné (Tesnière 1959, p. 13). Comme l'indique la figure 1.1, dans la stemma, la connexion [14] est en principe représentée par un trait vertical, reliant le régissant (supérieur), et le subordonné (inférieur).

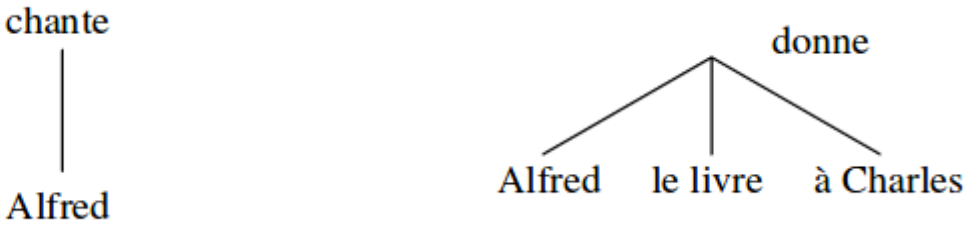


FIG. 1.1 – Représentation graphique de la stemma (tirée de Tesnière 1953, p. 4).

Cette représentation graphique de la stemma est appelée *diagramme en branches* (Tesnière 1953, p. 4). Dans cette configuration, le noeud verbal correspond en général au noeud central de la phrase simple. Le verbe, étant au centre du noeud verbal, est par conséquent au coeur de la phrase. Il est pour ainsi dire le régissant de toute la phrase. La notion de noeud verbal est définie à travers une métaphore du drame : *« le noeud verbal [...] exprime un tout petit drame. Comme un drame [...], il comporte obligatoirement un procès, et le plus souvent des acteurs et des circonstances »* (Tesnière, 1959). Les actants sont en général des syntagmes nominaux, des subordonnés immédiats du verbe. D'après Tesnière, ils complètent le verbe dont ils dépendent. Sur le plan sémantique, ils sont donc indispensables pour la réalisation du procès

14. La connexion, c.-à-d. la subordination, est la relation syntaxique par excellence en syntaxe structurale.

verbal. Les circonstants par contre expriment des circonstances de temps, de lieu, de manière, etc., dans lesquelles se déroule le procès. C'est dans ce sens que la syntaxe structurale de Tesnière postule l'existence des actants ou participants au procès verbal. L'ensemble des actants d'un verbe constitue sa structure actantielle, et la valence c'est le fait pour un verbe de régir un ou plusieurs actants. Ainsi, un verbe peut régir zéro (verbe avalent), un (verbe monovalent), deux (verbe bivalent) ou trois actants (verbe trivalent). Par exemple, dans la phrase 2, le verbe *donner* a trois actants : *Alfred*, *le livre* et *Charles*.

1) *Alfred chante*.

2) *Alfred donne le livre à Charles*.

Les actants sont ancrés dans le sens du verbe, et chacun joue un rôle sémantique bien déterminé dans le procès verbal, c'est cette propriété qui fait la différence entre eux. Le premier actant, qui sur le plan syntaxique occupe la fonction de sujet, est celui qui fait l'action (dans la phrase active), le second actant c'est celui qui subit l'activité ou l'action qu'exprime le verbe, et enfin le troisième actant est celui qui bénéficie ou souffre de l'action. Cette distinction est à l'origine des concepts de *cas profonds*, *rôle sémantiques* et *rôles thématiques* qui apparaitront plus tard dans les travaux de Fillmore (Fillmore, 1968).

La conception de Tesnière (mettant le verbe au centre de la phrase) s'oppose fermement à la grammaire traditionnelle qui postule qu'une opposition logique existe entre le sujet et le prédicat. Dans cette grammaire le sujet renvoie à « ce dont on dit quelque chose » (le thème) et le prédicat « ce qu'on en dit » (le rhème) [15]. En effet, pour les grammairiens traditionalistes, le sujet est un élément complètement externe au verbe, tandis que Tesnière place le verbe au centre de la phrase et compte le sujet parmi les actants, c.-à-d. les éléments qui complètent le sens du verbe. Sans pour autant rentrer dans ce débat, nous tenons à préciser que cette place centrale octroyée au verbe (dans l'approche de Tesnière) est un choix théorique qui sera observé tout le long de notre travail. Nous considérons effectivement le verbe comme le principal support de la phrase et, par lui, nous pouvons appréhender et comparer la syntaxe et la sémantique des textes médicaux (experts vs. non-experts) que nous analysons.

1.3.1.2 Grammaire de dépendance

Ces dernières années, un intérêt croissant s'est développé pour l'analyse des relations de dépendance comme base de la description de la syntaxe du langage naturel, dans les différentes communautés de chercheurs. Par conséquent, la notion de dépendance se retrouve dans plusieurs théories formelles modernes. Elle regroupe une famille de grammaires (désormais *GD*) basées

15. Cette hypothèse est un héritage de la philosophie d'Aristote. On la retrouvera chez les grammairiens génératives qui interviennent dans la section suivante (cf. section 1.3.1.2).

sur le postulat que les unités syntaxiques de la phrase sont connectées les unes aux autres par des relations de dépendance. Lucien Tesnière est considéré comme le père de ces théories grammaticales modernes basées sur la dépendance. En effet, la place centrale (le noeud principal, la tête) octroyée au verbe dans la syntaxe structurale de Tesnière se retrouve appliquée dans les approches dépendencielles (Nivre, 2005). Plusieurs formalismes ont ainsi vu le jour successivement :

- La description fonctionnelle et générative : dans cette approche, la phrase est traitée comme un système constitué de strates inter-connectées : phonologique, morphématique, morphonologique, analytique (syntaxe de surface) et tectographique (syntaxe profonde). Sur le plan sémantique, elle se focalise sur le sens linguistique de la phrase ainsi que sa réalisation. La phrase est ainsi structurée en deux parties : le *thème* qui renvoie à ce dont la phrase parle et le *rhème* qui désigne ce qui est dit du thème. C'est ce que les générativistes appellent *Topic-Function Articulation*. La description fonctionnelle et générative s'est développée en deux phases à travers :

  — les premières formalisations (Sgall *et al.*, 1969) : elles sont développées par Petkevic dans un système purement basé sur les dépendances (Peykevič, 1988 ; Petkevic, 1995) ;

  — le fonctionnalisme de l'école de Prague (Sgall *et al.*, 1986 ; Panevová, 1974, 1975) : s'inspirant du structuralisme dépendanciel de Tesnière, il voit le jour sous la direction de Mathesius, le fondateur du cercle de Prague. Cette approche est axée sur des relations de dépendance significatives telles que *Acteur*, *Patient*, qui pourraient être différenciées sur la base de l'analyse du comportement/fonctionnement des unités concernées en syntaxe de surface.

- La théorie Sens-Texte : créée grâce aux travaux de Igor Mel'čuk (Gladkij & Mel'čuk, 1975 ; Melčuk, 1988), elle postule que les langues sont définies par la façon dont leurs éléments (unités lexicales) sont combinés par des fonctions lexicales. Elle se rapproche de la Grammaire fonctionnelle et générative dans la mesure où elle se focalise sur la relation entre le sens et sa réalisation soit sous forme de texte, soit sous forme d'énoncé.

- Abhängigkeitgrammatik : à partir de ce formalisme, la langue allemande a vu se développer toute une tradition de la grammaire dépendancielle.

  Le domaine du TAL bénéficie depuis plusieurs années d'un grand héritage des grammaires de dépendance. En effet, plusieurs tâches du TAL sont basées sur les grammaires dépendancielles. Le *parsing* ou étiquetage syntaxique (Nilsson *et al.*, 2007 ; Debusmann & Kuhlmann, 2010) est une tâche qui consiste à identifier de façon automatique les différents éléments qui constituent une phrase, ainsi que les relations qui les lient. À l'issue de cette opération, les différents mots et constituants des phrases annotées sont

étiquetés et enrichies d'informations grammaticales et morpho-syntaxiques : catégorie grammaticale, fonction grammaticale, propriété morpho-syntaxique, rôle syntaxique et liens de dépendances. Cette analyse se fait sur la base de la conformité de la phrase analysée par rapport à une grammaire donnée. L'analyseur en flux de Vergne (1999) qui a remporté la campagne d'évaluation des étiqueteurs pour le français (*GRACE*), a été conçu selon une approche dépendancielle.

Les grammaires de dépendance trouvent un autre exemple d'application dans les technologies des systèmes d'extraction d'informations (Culotta & Sorensen, 2004). L'extraction d'informations (*EI*) consiste à remplir automatiquement des formulaires ou une banque de données, à partir d'informations pertinentes extraites de textes écrits en langue naturelle (Bessières *et al.*, 2001 ; Poibeau, 2001). Il ne s'agit pas simplement d'extraire des données textuelles de façon aléatoire, mais de mettre en oeuvre une analyse du texte, de façon à interpréter et construire une représentation formelle (acquisition de ressources) qui permettra d'apporter automatiquement des réponses précises à des utilisateurs.

La traduction automatique (Quirk *et al.*, 2005) fait également partie des tâches du TAL qui bénéficient des apports des grammaires de dépendance. L'opération de traduction automatique se veut être la traduction des textes écrits ou audios, d'une langue naturelle à une autre, à l'aide de programmes informatiques (Hutchins & Somers, 1992). Elle s'oppose donc à la traduction dite humaine, qui requiert l'intervention d'un humain.

Les grammaires de dépendance sont également utilisées dans le domaine de la génération automatique de textes. Le générateur de textes implémenté par Coch (1996) et intégré au système MultiMétéo (Coch, 1998) qui génère des bulletins météo multilingues est un système d'envergure développé sur la base des techniques de la GD.

D'autres travaux sur les GD sont décrits dans les actes du dernier atelier sur le traitement automatique par des grammaires basées sur la dépendance (Kahane & Polguère, 1998), également dans le numéro spécial de la revue TAL sur les grammaires de dépendance (Kahane, 2000).

1.3.1.3 La sémantique des cadres

Encore appelée *Frame semantics*, la sémantique des cadres (désormais *FS*) est une théorie qui remonte aux années 1980. Elle est une extension de la grammaire des cas (Fillmore, 1968), qui évoquait déjà l'existence des cas profonds ou rôles thématiques (agent, patient, thème, instrument, expérienceur, etc.) dans le sémantisme du verbe.

Selon le sémanticlopédie [16] (dictionnaire de sémantique), un rôle thématique est une étiquette abstraite qui caractérise la relation sémantique qu'un prédicat (verbe, adjectif, nom prédicatif et

16. http ://www.semantique-gdr.net/dico/index.php/Accueil

préposition) peut entretenir avec l'un de ses arguments. En d'autres termes, les rôles thématiques dénotent des fonctions de nature sémantique assurées par les participants qui interviennent dans la réalisation du prédicat verbal. Sur le plan syntaxique, ces participants sont représentés par les arguments du verbe. Autrement dit, ils sont à l'interface entre la syntaxe et la sémantique. Les rôles thématiques, tel que pensés par Fillmore (1968) ont pour intérêt de décrire les relations sémantiques entre un verbe et ses différents arguments. Plus précisément, leur rôle consiste à expliquer le comportement lexical du verbe au-delà de la sous-catégorisation syntaxique.

Cependant, de nombreuses critiques ont été émises par rapport à certaines irrégularités liées au nombre [17], à la nature (complétude/unicité, distinction), et bien d'autres propriétés de ces rôles thématiques, donnant naissance à diverses théories (Jackendoff, 1972, 1987 ; Foley & Van Valin Jr, 1984 ; Dowty, 1991 ; Levin, 1993 ; Baker, 1997 ; Màrquez *et al.*, 2008 ; Kasper, 2008 ; Palmer *et al.*, 2010). Le caractère abstrait des rôles thématiques fait partie des éléments qui ont été fortement remis en question. D'après les opposants de Fillmore (1968), les rôles thématiques ne décrivent que de manière globale les relations prédicat-arguments, puisqu'ils ne réussissent pas à capturer certaines spécificités caractérisant de façon particulière les types de participants au procès verbal. Autrement dit, cette liste prédéfinie de rôles sémantiques ne peut pas saisir certaines distinctions sémantiques pertinentes entre des situations particulières décrites par un ou différents verbes. Les exemples suivants peuvent servir d'illustrations :

3) *La dame propose la marchandise à la cliente*.

4) *Le chargé de cours propose un test aux étudiants*.

Les phrases 3 et 4 partagent la même structure argumentale (Su-COD-COI) [18], avec en commun l'infinitif *proposer* comme prédicat verbal. Si l'on décide de faire une analyse en rôle thématique de ces phrases, les arguments sujets *commerçant* et *professeur* joueraient le rôle d'agent. Toutefois, si l'on souhaite faire une analyse comparative de la relation sujet-verbe entre ces deux phrases, il serait nécessaire de caractériser de façon plus détaillée la relation entre le verbe *proposer* et chaque sujet. Ceci reviendrait à préciser que le sujet *commerçant* joue le rôle de VENDEUR, tandis que le sujet *professeur* assure la fonction d'ENSEIGNANT dans la situation que dénote le verbe.

L'avènement de la sémantique des cadres [19] de Fillmore (1982) a contribué à combler certaines lacunes identifiées dans la notion de rôles thématiques. D'ailleurs, son principal objectif, tel que le précise Fillmore & Baker (2010), est de proposer une solution aux problèmes que pose la richesse sémantique des unités lexicales (noms, adjectifs, verbes) dont elle décrit la syntaxe

17. Au fil du temps, différentes listes de rôles sémantiques ont été proposées par différentes théories. Kasper (2008) effectue une étude comparative de ces théories.

18. Su = sujet ; COD = complément d'objet direct, COI = complément d'objet indirect.

19. Elle est perçue comme une approche évoluée des rôles thématiques, puisque Fillmore fait partie des fondateurs de cette théorie.

et la sémantique. Pour ce faire, la FS propose une grille de rôles sémantiques des verbes. Contrairement aux cas profonds (rôles thématiques), les rôles sémantiques de la FS sont propres à la situation de communication décrite par l'unité lexicale analysée (*acheter* : VENDEUR, ARGENT, BIEN, ACHETEUR, etc.). Mais des généralisations sont possibles via des relations entre les rôles et les différentes situations qui sont liées.

Le principe fondamental de la FS est que le sens d'un mot ne peut être interprété que si l'on a accès aux informations (linguistiques, extralinguistiques ou encyclopédiques) essentielles faisant référence à ce mot dans une situation de communication schématisée par un frame ou cadre sémantique. Ces informations peuvent donc être accessibles grâce au cadre [20] au sein duquel les unités lexicales sont organisées.

Le *cadre* est défini comme un scénario, un schéma ou une structure conceptuelle qui sous-tend l'utilisation d'un item lexical ainsi que son interprétation (Fontenelle, 2009). Il décrit un type particulier d'évènement, de relation, d'entité, etc., ainsi que les participants qui y interviennent. Ses participants, les *Frame elements (FE)*, peuvent être obligatoires (*core elements*) ou facultatifs (*non-core elements*). Un cadre est évoqué par une unité lexicale *(LU)*. Par exemple, le frame de la transaction commerciale (Fillmore, 1976) peut être évoqué par différentes unités : *acheter*, *vendre*, *payer*, *récupérer*, etc., et plusieurs participants : obligatoires (VENDEUR, ARGENT, BIEN, ACHETEUR) et facultatifs (MOYEN, etc.). Lorsque l'unité évocatrice du cadre est un verbe, l'analyse est focalisée sur les arguments de ce dernier qui représentent les éléments du cadre. La sémantique des cadres est la théorie de base du projet FrameNet (cf. section 1.3.3.1) (Ruppenhofer *et al.*, 2006, 2013).

Depuis son apparition, la sémantique des cadres s'applique aux textes portant sur la langue générale. Certains travaux de recherche récents font appel à ce cadre théorique pour l'annotation et l'analyse des textes en langues de spécialité : dans le domaine biomédical (Dolbey *et al.*, 2006 ; Dolbey, 2009), en sport, notamment en football (Schmidt, 2008, 2009), en droit (Pimentel, 2011), en informatique et dans le domaine de l'environnement (L'Homme, 2012a).

Le domaine médical n'est pas en reste. Borin *et al.* (2007b) entreprennent une étude expérimentale dont l'objectif principal est d'examiner dans quelle mesure l'intégration d'informations syntaxiques et sémantiques dans un corpus médical suédois peut contribuer de manière significative à l'acquisition et à l'extraction semi-automatique de frames qu'ils appellent *schémas sémantiques*. L'étude porte exclusivement sur les prédicats verbaux. Les éléments qui entourent les verbes subissent un étiquetage morpho-syntaxique effectué grâce au Cass parser (Abney, 1997). Les unités nominales sont ensuite annotées en classes sémantiques, grâce au MeSH tagger [21] (pour le suédois) qui leur associe des catégories sémantiques décrivant les types

20. Màrquez *et al.* (2008) résume cette idée, en expliquant qu'un cadre relie la sémantique linguistique aux connaissances encyclopédiques.

21. http ://mesh.kib.ki.se/swemesh/

d'entités médicales qu'elles dénotent : (A) *anatomy*, (B) *organism*, (C) *chemicals and drugs*, (D) *analytical, diagnostic and therapeutic techniques and equipment*, (E) *psychiatry* and (F) *psychology*. La méthode d'annotation utilisée dans cette étude est très similaire à la nôtre. En effet, contrairement au projet FrameNet qui met beaucoup l'accent sur les rôles sémantiques joués par les participants au prédicat verbal, Borin *et al.* (2007b), tout comme nous, s'intéressent davantage aux types de concepts médicaux impliqués dans la réalisation du procès verbal. Cependant, notre étude se démarque de FrameNet dans le sens où nous faisons une analyse syntaxique dépendancielle des phrases de notre corpus avec l'analyseur syntaxique Cordial Dependency Parser (Laurent *et al.*, 2009), afin de détecter les relations de dépendance entre les verbes et leurs arguments. La deuxième différence est que notre étude s'inscrit dans le contexte de la simplification de textes [22], tandis que celui-ci a pour but l'extraction semi-automatique de schémas sémantiques des verbes. Et enfin, notre étude porte sur la langue française qui, à notre connaissance, n'a pas encore été impliquée dans un tel projet axé sur les cadres sémantiques médicaux.

La théorie des cadres sémantiques occupe une place importante dans notre travail car elle fait partie des théories de base qui servent de socle à cette étude. Toutefois, nos approches diffèrent sur certains aspects qui seront abordés dans la section suivante (cf. 1.3.3.1). Par conséquent, notre méthode est considérée comme le fruit d'une adaptation de la FS.

1.3.2 Terminologie

Les entités nominales ont longtemps occupé une place centrale dans les travaux sur les langues de spécialité au détriment des autres parties du discours, dont les verbes, mis à l'écart pour diverses raisons. En effet, les travaux en terminologie se focalisent la plupart du temps sur la description des concepts (en général dénotés par des entités nominales) et la mise au jour des relations que partagent ces concepts (genre-espèce, partie-tout, etc.). L'un des motifs principaux énoncés pour justifier l'exclusion du verbe est la place importante accordée aux objets et à leurs dénominations dans l'approche de Wüster, connu comme le fondateur de la Théorie Générale de la Terminologie (*TGT*). Évidemment, le principe fondamental de la TGT, telle qu'inventée par Wüster et ses collègues (Wüster, 1981 ; Wüster, 1985), était l'étude du terme (entité nominale portant les connaissances du domaine), de ses fonctions et de sa relation avec la notion (concept).

L'importance donnée aux noms en terminologie s'explique également par le fait que les entités nominales sont généralement utilisées pour le développement des terminologies, ontologies, thésaurus, glossaires et vocabulaires. De plus, de nombreuses tâches du TAL sont typique-

22. Comme nous l'avons indiqué dès le départ de cette étude, notre objectif final est de mettre en place une ressource de simplification des textes médicaux spécialisés.

ment centrées sur les entités nominales (l'indexation, l'extraction d'informations, les systèmes questions-réponses, etc.), d'où les besoins croissants des applications. Pour toutes ces raisons, les différentes approches théoriques et méthodologiques le plus souvent mises en place sont axées sur les entités nominales.

Néanmoins, les interrogations sur le statut des unités verbales, adjectivales, adverbiales et même prépositionnelles dans les langues de spécialité faisait déjà l'objet de travaux d'un groupe de chercheurs, dont l'un des pionniers est Picht Heribert. Dans ses travaux (Heribert, 1983, 1985) sur les unités phraséologiques et collocations en LGP (*Language for General Purpose*) et LSP (*Language for Specific Purpose*), l'auteur fait de nombreuses investigations sur les phénomènes linguistiques qui sont à la frontière entre LGP et LSP et qui constituent les « zones obscures » de ces deux disciplines. Dans une étude qui vise à proposer une méthode d'éclairage de celles-ci, Picht décrit une technique d'analyse des verbes qui sont à l'intersection entre terminologie et langue générale. Cette méthode est basée sur une grille de questions dont les réponses permettraient de déterminer les propriétés qui caractérisent un infinitif au moment où il cesse d'être un verbe général et commence à être considéré comme un verbe spécialisé ou terme (Heribert, 1987). Pour cet auteur, les unités phraséologiques et collocations verbales (ainsi que les autres parties du discours) devraient être étudiées comme unités terminologiques au même titre que les unités phraséologiques et les collocations nominales. Une vision similaire se retrouve dans les travaux de terminologues tels que Pamela Faber (Faber, 2012), Anne Condamines (Condamines, 1992 ; Condamines & Bourigault, 1999), Marie-Claude L'Homme (L'Homme & Bodson, 1997 ; L'Homme, 1998, 2012b), etc., dont les recherches ont également largement contribué à démontrer l'importance d'aborder les unités verbales comme termes dans les textes en langues de spécialité.

Différentes études s'inspirant des courants de la sémantique lexicale, de la linguistique de corpus, et d'autres approches, se focalisant sur le verbe ont commencé à voir le jour. Coulon (1972) ; L'Homme (1992) ; Cajolet-Laganière & Maillet (1995) ; L'Homme (1996) étudient le fonctionnement des verbes dans des corpus de textes scientifiques en langue française. Différentes méthodes ont été mises en place pour l'analyse du verbe et de ses arguments dans les travaux sur les langues de spécialité : dans le domaine de la finance, Condamines (1993) propose une méthode permettant d'exploiter les propriétés syntaxico-sémantiques des verbes et des noms afin d'identifier des termes inconnus constituant de potentiels candidats pour la conception d'un vocabulaire spécialisé ; en informatique, L'Homme & Bodson (1997) implémentent un modèle de description des verbes spécialisés qui combine base de connaissances et hypertexte ; en biologie, Tateisi *et al.* (2004) développent une méthode d'annotation des structures argumentales prédicat-argument, qui permet d'améliorer les performances d'un système d'extraction d'information ; en médecine, Tellier (2008) met en place une méthode de description des verbes pour la rédaction des articles terminologiques ou dictionnairiques ; en

droit, Lerat (2002) propose une classification des verbes juridiques, tandis que Pimentel (2011) développe une approche de traitement du verbe terminologique qui aide à la traduction des textes de l'anglais vers l'espagnol. Certains travaux impliquent les éléments tels que la flexion et le mode du verbe parmi les critères d'identification des textes spécialisés (Stein *et al.*, 1992 ; Da Cunha *et al.*, 2011).

Par ailleurs, certains outils sont adaptés ou mis en place pour le traitement du verbe dans les travaux en terminologie. C'est le cas de Termostat, un extracteur terminologique créé au départ exclusivement pour les entités nominales, qui a été adapté et mis à disposition pour l'extraction des verbes dans des textes de spécialité (Drouin, 2003). Plus récemment encore, c'est Sketch Engine (Kilgarriff *et al.*, 2004, 2014), un outil de travail sur corpus (généraux et spécialisés), qui a vu le jour. Il offre différentes fonctionalités applicables autant aux unités nominales qu'aux verbes : construction d'un thésaurus, d'une liste de termes clés qui caractérisent le corpus, de concordances et collocations de mots clés, etc.

Il existe différentes méthodes employées pour l'analyse du verbe terminologique. Nous nous focaliserons principalement sur deux approches parmi les quatre que proposent L'Homme (2012b) :l'approche conceptuelle (cf. section 1.3.2.1) et l'approche lexico-sémantique (cf. section 1.3.2.2). Cette dernière représente un cadre fondamental pour ce travail de thèse.

1.3.2.1 L'approche conceptuelle

Tel que décrit par L'Homme (2012b), le principe de l'approche conceptuelle voudrait que l'on ne s'intéresse au verbe que s'il a l'aptitude de désigner un « concept d'activité ». Telle serait la condition qui détermine l'intégration des verbes dans des ressources terminologiques. Autrement dit, selon cette approche, le verbe ne peut être considéré comme terme que s'il est fortement assimilable à un nom sur le plan conceptuel. Rey (1979) [23] définit clairement le statut du verbe selon la perspective conceptuelle en ces termes : « la terminologie ne s'intéresse aux signes (mots et unités plus grandes que le mot) qu'en tant qu'ils fonctionnent comme des noms dénotant des objets et comme des "indicateurs de notions" (de concepts) et dans cette optique, les verbes sont des noms de processus, d'actions » (Rey, 1979). En d'autres termes, lorsque un verbe permet de nommer un procéssus ou une action, il constitue un potentiel candidat pour une ressource terminologique. Cette conception justifierait en partie la discrimination observée entre les parties du discours traitées ou non dans un dictionnaire spécialisé, et plus précisément le faible nombre de verbe. En effet, en général, on y compte très peu de verbes et d'adjectifs mais beaucoup d'entités nominales. Les résultats d'une étude portant sur la présence des verbes dans quatre dictionnaires de spécialité évaluent à 2,44% (entre 0 et 4 verbes par dictionnaire)

23. Cité par L'Homme (2012b).

la moyenne d'apparition des verbes dans ces dictionnaires terminologiques[24] (L'Homme, 2003). L'approche conceptuelle a débouché de nos jours sur une démarche conceptuelle, incarnée par les ontologies, qui permet de distinguer les concepts d'activité exprimés par les noms ou par les verbes dans les domaines de spécialité. Ainsi, dans le domaine médical par exemple, les verbes tels que *traiter*, *observer* et *activer* peuvent devenir terminologiques puisqu'ils permettent de rendre compte de notions comme *traitement de la maladie, observation du patient* et *activation des cellules*, qui renvoient à des processus ou des activités (L'Homme, 2012b).

1.3.2.2 L'approche lexico-sémantique

Cette approche repose sur la théorie linguistique qu'on appelle la *sémantique lexicale*, qui d'après (L'Homme, 2012b) est l'une des premières théories à avoir montré l'importance de la structure argumentale du verbe et du réseau lexical auquel il appartient. Dans ce cadre théorique, la caractérisation de la nature spécialisée du verbe est basée sur la description de sa structure argumentale ou son appartenance à un ou plusieurs réseaux lexicaux, (morpho-)sémantiques ou paradigmatiques. L'observation et l'analyse des différentes occurrences du verbe en corpus est la base de cette approche qui joue un rôle déterminant dans ce travail de thèse.

La structure argumentale

Dans cette branche de l'approche conceptuelle, L'Homme (2012b) défend l'hypothèse selon laquelle l'analyse de la structure argumentale du verbe peut permettre de démontrer sa nature terminologique. En effet, comme il a été montré dans la section 1.3.1.1, la nature prédicative du verbe fait qu'il a besoin des éléments qu'il régit pour la réalisation de son sens.

5) *Le médecin* examine le *patient*.

6) *Le patient* développe la *maladie*.

Dans les exemples 5 et 6, les verbes *examiner* et *développer* véhiculent des sens spécialisés liés au domaine de la médecine. Ces sens médicaux sont déterminés par le type d'actants (*médecin, patient, maladie*) qui accompagnent les verbes. Supposons que les actants de l'exemple 6 soient remplacés respectivement par *le programmeur* et *l'outil*, on obtiendrait la phrase suivante : *le programmeur développe l'outil*. Dans cette phrase, le verbe *développer* a un sens (*'programmer', 'mettre en place', 'concevoir'*) complètement différent de celui (*'manifester', 'présenter', 'souffrir de'*) de la phrase 6, et qui de surcroît ne relève plus du domaine de la médecine mais plutôt de celui de l'informatique. À travers ce petit test, l'on peut constater combien le changement de type d'actants d'un verbe peut influencer son sens. Ce constat permet de comprendre à quel

24. Toutefois, il faudrait souligner que la taille des dictionnaires interrogés pourrait avoir un impact sur les résultats de ce type d'étude.

point le sens d'un verbe est déterminé par les propriétés sémantiques de ses arguments. Une étude visant à définir des critères d'identification des verbes terminologiques propose de prendre en considération la nature sémantique des arguments du verbe, qui détermine son degré de spécialisation (L'Homme, 1998). Ce raisonnement illustre l'hypothèse selon laquelle le verbe n'est pas spécialisé par lui-même mais grâce à la prise en compte de sa structure argumentale, et plus précisement du type d'actants qui interviennent dans sa réalisation. C'est ce critère qui permet à L'Homme (2012b) d'admettre *installer* comme verbe spécialisé dans l'exemple suivant :

7) *L'utilisateur installe la nouvelle version du traitement de texte sur son PC.*

Dans cette phrase, les termes (*utilisateur*, *version*, *PC*) qui représentent les têtes des arguments du verbe appartiennent au domaine de l'informatique. Par conséquent, dans cet emploi, *installer* peut être considéré comme verbe terminologique du domaine de l'informatique.

L'analyse des arguments des verbes constitue également un critère de poids chez Tellier (2008) qui y trouve un moyen de sélection des verbes représentant de bons candidats termes à ajouter à la nomenclature d'un dictionnaire spécialisé. Pour y parvenir, l'auteure étudie le fonctionnement des verbes dans un corpus spécialisé relevant du domaine de l'infectiologie, et en ressort avec une liste de 34 verbes sélectionnés, parmi lesquels l'infinitif *éliminer*. Le prédicat *éliminer* s'est révélé être un terme de par la nature de ses actants (*biopsie*, *infection*, *antiseptique*, *malade*, *lymphocyte*, *caseum*, *mucus*) repérés dans différentes occurrences en corpus.

Ce critère est également utilisé dans d'autres travaux de recherche du même type (Lerat, 2002 ; Pimentel, 2011). Cependant, la caractérisation des arguments du prédicat verbal n'a pas pour unique but l'identification des verbes terminologiques. D'autres objectifs peuvent être visés tels que : l'extraction d'informations (à partir de corpus spécialisés du domaine de la biologie moléculaire (Tateisi *et al.*, 2004)), et l'élaboration de dictionnaires ((Pimentel, 2011), dans cette étude, il s'agit d'un dictionnaire juridique bilingue portugais-anglais).

Cette approche, basée sur l'exploitation de la structure argumentale pour la reconnaissance des sens spécialisés des verbes, cadre bien avec la méthode que nous souhaitons implémenter dans ce travail de thèse. En effet, l'un de nos objectifs est de parvenir à identifier des patrons verbaux médicaux à simplifier, à partir des catégories sémantiques [25] (cf. section 2.2.2) des arguments des verbes. Nous nous proposons donc d'utiliser l'approche lexicale dans un contexte de simplification de textes, pour la sélection des patrons spécialisés des verbes, entrées potentielles pour la ressource de simplification finale. À notre connaissance, il n'existe pas encore d'étude dans le domaine de la simplification de textes qui propose une telle approche axée sur la structure

25. Ces catégories sémantiques seront obtenues à partir d'une terminologie médicale utilisée pour l'annotation sémantique des corpus (cf. chapitre 3, section 3.2.2).

argumentale du verbe.

Le réseau lexical

Outre la nature des actants, L'Homme (2012b) émet d'autres paramètres qui peuvent être pris en compte par les chercheurs lors du repérage des verbes terminologiques. L'un de ces paramètres, qui revient très souvent, est le lien entre un verbe et un nom. Ainsi, si le nom est terminologique, et si le verbe est sémantiquement et le plus souvent morphologiquement apparenté à celui-ci, alors il est fort possible que le verbe soit spécialisé lui aussi. Ce critère s'observe avec les couples suivants, proposés par l'auteure : *développement - développer*, *téléchargement - télécharger*, *réchauffement - réchauffer*.

Dans ces exemples, le verbe et le nom correspondant désignent tous les deux une activité. On aurait donc affaire à des noms de processus d'actions dont il est question dans l'approche conceptuelle (cf. section 1.3.2.1). Autrement dit, l'approche conceptuelle et le réseau lexical se rejoignent, lorsque le nom dérivé du verbe et le verbe lui-même expriment une activité, un processus ou encore un procès comme le désignent la plupart des travaux de la littérature. Malgré les différentes réserves [26] émises (Meinschaefer, 2003, 2005 ; Huyghe & Marín, 2007) par rapport au transfert de propriétés entre le verbe et ses noms dérivés, il est généralement présumé que les nominalisations [27] héritent des propriétés aspectuelles [28] des verbes dont elles dérivent (Gross & Kiefer, 1995 ; Fábregas *et al.*, 2012 ; Fábregas & Marín, 2012). Ce phénomène concerne particulièrement les *noms déverbaux* [29] (Grimshaw, 1990 ; Haas *et al.*, 2008 ; Villoing & Namer, 2008 ; Barque *et al.*, 2009), c.-à-d. des nominalisations qui partagent un lien morphologique avec leurs bases (*développer - développement*). Autrement dit, selon le type de procès (Grimshaw, 1990) qu'exprime le verbe (état, action, évènement), les noms déverbaux dérivés peuvent avoir différentes interprétations. Ils peuvent renvoyer soit au procès qu'exprime le verbe de base, soit à l'un ou plusieurs de ses actants (agent, instrument, moyen, etc.) (Villoing & Namer, 2008).

Ainsi, l'approche conceptuelle se distingue de l'approche basée sur le réseau lexical, car il existe des cas de figure où le sens du verbe et celui du nom dérivé sont distincts, malgré le lien morphologique existant entre eux [30]. En effet, en dehors du sens d'activité, les noms déverbaux

26. Certains travaux de recherche montrent que le degré et processus de transfert de propriétés aspectuelles entre les verbes et les noms dérivés dépend de différents paramètres selon les types de verbes.

27. Nom dérivé d'une base verbale.

28. Depuis quelques années, les recherches sur l'aspect lexical (*Aktionsart*) ne s'intéressent plus uniquement aux verbes mais également aux noms dérivés, le but étant d'étudier les relations entre ces deux catégories en se focalisant sur l'aspect.

29. Un nom déverbal est un substantif qui résulte d'une opération morpho-sémantique consistant à appliquer sur un verbe un suffixe de type *–ion*, *–age*, *–ment*, *–eur*, ou zéro(Magri-Mourgues, 2015).

30. À ce propos, Meinschäfer (Meinschaefer, 2003, 2005) soutient que la relation morphologique (verbe-nominalisation) n'induit pas systématiquement un effet aspectuel particulier sur la nominalisation. Fábregas *et al.* (2012) fait référence à ce phénomène en utilisant l'expression *Aspect Preservation Hypothesis*

peuvent exprimer d'autres types de relations avec le verbe de base (objet résultant, instrument, agent, etc.). Ces relations renvoient aux actants qui participent à la réalisation du prédicat verbal. C'est le cas du couple *programmeur* - *programmer* que propose (L'Homme, 2012b). Le nom *programme*[31] désigne le résultat de l'activité que dénote le verbe *programmer*, tout comme le nom *programmeur* désigne l'agent de l'action qu'exprime le verbe *programmer*. Comme nous pouvons le constater, dans l'approche lexico-sémantique les noms peuvent servir de point de départ à partir duquel les verbes spécialisés sont identifiés, en fonction des liens qu'ils partagent avec eux. C'est d'ailleurs cette méthode qui permet à Tellier (2008) de retenir les verbes *évoluer*, *excréter*, *infecter* et *sécréter* comme termes du domaine de l'infectiologie, de par leur parenté aux noms *évolution*, *excrétion*, *infection* et *sécrétion*. L'auteure base sa méthode sur le transfert du caractère spécialisé des unités nominales vers les verbes correspondants.

Toutefois, il est possible de déplacer le point de départ de l'analyse vers le verbe. Cette technique peut permettre de découvrir d'autres unités reliées au verbe et d'élargir ainsi le réseau lexical construit autour de ce dernier (L'Homme, 2012b). Cette démarche a été appliquée lors de la conception du *DicoInfo* (*Dictionnaire fondamental de l'informatique et de l'Internet*), une base de données lexicales contenant des termes, y compris des verbes fondamentaux appartenant aux domaines de l'informatique et de l'Internet (L'Homme, 2009). L'approche utilisée s'inspire grandement des principes théoriques et méthodologiques de la Lexicologie explicative et combinatoire (Mel'cuk *et al.*, 1995) et permet de fournir pour chaque entrée du dictionnaire différents types d'informations : la réalisation linguistique des actants, les liens lexicaux, les synonymes, les contextes d'apparition du mot-clé, etc. Pour le verbe *programmer* par exemple, DicoInfo propose divers types d'unités lexicales appartenant à son réseau lexical tels que *programmation* ('action de programmer'), *programme* ('résultat de l'action de programmer'), *informaticien* ('agent de l'action de programmer'), *langage* ('instrument utilisé pour programmer'), *logiciel* ('résultat de l'action de programmer'), *écrire* ('synonyme de programmer'), *développer* ('synonyme de programmer'), etc. Cet exemple permet d'observer que les mots repérés dans ce dictionnaire sont liés au verbe par différentes relations exprimées à travers de courtes gloses explicatives.

Comme nous l'avons déjà souligné, l'approche lexico-sémantique fait partie des cadres théoriques sur lesquelles se base notre étude. En effet, tout comme L'Homme & Bodson (1997), nous pensons que le verbe, en tant que prédicat central de la phrase, est un excellent point de départ pour cerner la syntaxe et la sémantique des phrases des textes spécialisés, puisqu'il permet d'exprimer les connaissances véhiculées par les termes avec lesquels il cooccurre. Ainsi, dans notre travail, la structure argumentale est perçue comme un canal par lequel les sens spécialisés des verbes sont identifiés. Elle permet ainsi de proposer un/des meilleur(s) substitut(s) aux verbes spécialisés, dans le contexte de la simplification de textes (cf. section 1.4).

31. Ce nom peut également faire référence à un instrument ou un outil, si l'on considère qu'un programme permet de réaliser une tâche bien précise.

1.3.3 Traitement Automatique des Langues

Dans le domaine du Traitement Automatique des Langues, les méthodes et outils dédiés aux verbes se focalisent le plus souvent sur l'analyse de leurs contextes d'apparition et sur la caractérisation de leur structure argumentale : les fonctions grammaticales et rôles sémantiques des arguments (Gildea & Jurafsky, 2002), la valence verbale (Eynde & Mertens, 2003), les possibilités combinatoires et les relations de dépendance (Marneffe *et al.*, 2006), la désambiguïsation du sens des verbes (Ide & Véronis, 1998 ; Ye & Baldwin, 2006 ; Wagner *et al.*, 2009 ; Brown *et al.*, 2011), l'acquisition de schémas de sous-catégorisation (Messiant *et al.*, 2010), les classes sémantiques des verbes à partir de patrons de sous-catégorisation (Schulte im Walde & Brew, 2002) etc. Nous allons ici présenter quelques ressources et outils du TAL qui impliquent les unités verbales, en précisant à chaque fois ceux qui sont exploités dans cette étude, ainsi que les raisons de notre choix.

1.3.3.1 FrameNet

FrameNet (désormais *FN*) (Ruppenhofer *et al.*, 2006) est une base de données lexicales [32] initialement conçue pour l'anglais. Elle contient plus de 10 000 sens [33] d'unités lexicales décrits à travers plus de 1 000 cadres sémantiques liés hiérarchiquement les uns aux autres et illustrés par plus de 170 000 phrases. Le projet FrameNet propose une description des unités lexicales prédicatives (verbes, noms et adjectifs) basée sur l'annotation en cadres sémantiques (Fillmore, 1982) des phrases dans lesquelles ces unités apparaissent (cf. section 1.3.1.3).

La méthode appliquée dans FrameNet passe par la définition de cadres conceptuels qui sont ensuite annotés dans des phrases exemples. Ces cadres sémantiques permettent de voir comment les éléments qui les constituent se réalisent syntaxiquement autour de l'unité lexicale évocatrice du cadre. Cette approche a été appliquée pour la conception du FrameNet du français [34] qui a vu le jour récemment (Djemaa *et al.*, 2016). Cette première version de la ressource couvre 4 scènes (transaction commerciale, positions cognitives, causalité et communication verbale) et contient 98 cadres sémantiques, 662 unités lexicales évocatrices des cadres, 872 sens, et environ 13 000 phrases exemples annotées provenant des corpus French Treebank (Abeillé & Barrier, 2004) et Sequoia Treebank (Candito & Seddah, 2012). Ci-dessous, un exemple de phrase annotée selon l'approche de FN :

8) *Un protocole d'accord a été signé entre SGS-Thomson et ASAT, filiale à 100 % de QPL, et la* TRANSACTION *devrait être finalisée avant la fin du premier trimestre 1993 pour un montant non précisé*.

32. https ://framenet.icsi.berkeley.edu/fndrupal/about
33. Le mot *sens* ici renvoie aux différents 'emplois' des unités lexicales traitées dans la ressource.
34. http ://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml

L'exemple 8 est une phrase annotée [35], tirée du French FrameNet, illustrant le cadre de la transaction commerciale [36] dont nous avons parlé dans la section 1.3.1.3. Chaque couleur représente un élément du cadre : bleu foncé=ACHETEUR, bleu ciel=ARGENT, rouge=VENDEUR.

Un cadre sémantique met en évidence les informations nécessaires pour capturer le sens de l'unité lexicale clé. Ainsi, pour chacune de leurs entrées, les FrameNet anglais et français ici décrits, ainsi que les autres versions existantes, sont capables de fournir (chacun dans sa langue) un cadre sémantique complet, une description de ce cadre, ses éventuelles relations avec d'autres cadres, une description des éléments du cadre et une illustration des schémas valenciels de l'entrée à l'aide d'exemples (Ruppenhofer *et al.*, 2006 ; Djemaa *et al.*, 2016).

Le modèle [37] d'annotation sémantique que nous implémentons s'inspire grandement de celui du projet FrameNet et la notion de *frame* que nous appelerons desormais *patron syntaxico-sémantique* (PSS) est également inspirée de ce projet. Toutefois, il est indispensable de souligner que les approches utilisées présentent d'importantes différences : premièrement, nous proposons une approche bottom-up (du texte vers les PSS), similaire à la méthode appliquée dans le cadre du projet SALSA [38] pour l'allemand, qui avait pour but la constitution d'un large corpus de langue générale, annoté en rôles sémantiques selon les principes de la sémantique des cadres (Burchardt *et al.*, 2006, 2009). FrameNet par contre implémente une approche top-down (des frames vers le texte). Deuxièmement, dans notre étude, l'annotation sémantique des arguments du verbe est basée sur des catégories sémantiques fournies par une terminologie existante (Côté, 1996), tandis que FrameNet propose des rôles sémantiques. Par conséquent, ce que nous appelons PSS [39] n'est pas identique au frame de FrameNet.

1.3.3.2 VerbNet

VerbNet [40] (désormais *VN*) (Kipper *et al.*, 2000 ; Kipper-Schuler, 2005) est une ressource lexicale qui propose une description des verbes basée sur la classification de Levin (Levin, 1993). Il s'agit d'une classification lexico-sémantique de verbes anglais à partir de l'analyse de leurs fonctionnements (syntaxe, classe sémantique des arguments sélectionnés, etc.). Le regroupement des verbes dans des classes sémantiquement homogènes est basée sur l'hypothèse

35. Dans cette phrase, le syntagme *un protocole d'accord* correspond à l'objet de la transaction, que FN capture grâce à l'étiquette GOODS, mais cette information n'a pas été mise en évidence. Ceci pourrait relever d'un oubli de la part de l'annotateur.

36. Cliquez sur le lien suivant pour consulter plus de phrases annotées FN :http ://asfalda.linguist.univ-paris-diderot.fr/allanno/allanno_FR_Commercial_transaction.xml (cf. phrase numéro 34).

37. Chaque frame est constitué d'un ensemble de triplets qui représentent la réalisation des FEs dans chacune des phrases annotées.

38. http ://www.coli.uni-saarland.de/projects/salsa/

39. Patron syntaxico-sémantique :structure argumentale du verbe au sein de laquelle chaque argument est associé à une catégorie de la terminologie SNOMED dénotant un concept médical.

40. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html

selon laquelle les verbes qui affichent un ensemble d'alternances de diathèses (ou schémas de valence) identiques ou similaires dans la réalisation de leurs structures argumentales partagent certaines propriétés sémantiques. L'alternance de diathèses [41] (la relation entre deux réalisations de surface d'un même prédicat), qui est le principal critère d'identification des classes verbales dans cette approche, est appuyée par des propriétés supplémentaires liées à la sous-catégorisation, à la morphologie et aux verbes ayant un sémantisme complexe. À partir de ces critères, la classification initiale propose 3 024 verbes, 4 186 sens, 240 classes de verbes construites autour de 79 alternances. Par exemple, la classe des prédicats dénotant une configuration spatiale contient les verbes suivants : *balance, bend, bow, crouch, dangle, flop, fly, hang, hover, jut, kneel, lean, lie, loll, loom, lounge, nestle, open, perch, plop, project, protude, recline, rest, rise, roost, sag, sit, slope, slouch, slump, sprawl, squat, stand, stoop, straddle, swing, tilt, tower*, qui partagent les mêmes propriétés syntaxiques et sémantiques : structure argumentale, rôles thématiques des arguments, restrictions de sélection, etc. (Levin, 1993).

VerbNet est donc un lexique hiérarchique de verbes anglais regroupés en classes, indépendamment des domaines de spécialité auxquels ils peuvent appartenir. Chaque classe est décrite à travers : (i) l'ensemble d'arguments possibles pour les verbes du groupe, présentés sous forme de rôles thématiques ; (ii) les éventuelles restrictions de sélection d'arguments (comme *animé*, *humain*, *organisation*) ; (iii) les cadres décrivant les possibles réalisations de surface de la structure argumentale (constructions transitives/intransitives, syntagmes prépositionnels/résultatifs) ; et (iv) les alternances de diathèse, c'est-à-dire les variations des différents frames. Selon le site officiel [42], après son extension (Korhonen & Briscoe, 2004), VerbNet compte 274 classes de premier niveau, 23 rôles thématiques, 94 prédicats sémantiques, 55 restrictions syntaxiques, 5 257 sens des verbes et 3 769 lemmes.

1.3.3.3 La base de données des Verbes Français

Les Verbes Français est une base de données électronique des verbes du français, comme son nom l'indique. Cette ressource a été créée à partir de la version imprimée de la ressource (Dubois & Dubois-Charlier, 1997), de taille légèrement réduite, qui a été la première à voir le jour. Il s'agit d'un thésaurus de classes verbales syntaxico-sémantiques, c'est-à-dire des groupes sémantiques de verbes définis par la syntaxe (les schémas valenciels). D'après les concepteurs de la ressource, l'élaboration du LVF est basée sur les méthodes classiques de la grammaire distributionnelle et transformationnelle de Zellig Harris. En effet, « *Les Verbes Français* repose sur l'hypothèse qu'il y a adéquation entre les schèmes syntaxiques de la langue et l'interprétation sémantique qu'en font les locuteurs » (François *et al.*, 2007). Cette hypothèse reprend l'idée

41. Elles sont souvent liées à l'alternance passif/actif.
42. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html

essentielle de la sémantique des grammaires de Harris qui stipule qu'« il existe une corrélation entre structure et signification [...] » (Harris, 1971).

La ressource contient 25 610 entrées qui couvrent 12 310 verbes, parmi lesquels plus de la moitié sont des entrées polysémiques. Pour chaque entrée, le LVF fournit différents types d'informations : classe sémantique, schéma valenciel (transitif, intransitif, pronominal, etc.), sens (un synonyme, une définition, ou une explication), domaine d'application (médecine, linguistique, droit, etc.) et exemple. Les exemples proposés par le LVF sont des phrases simples, illustrant le schéma valenciel et le sens de l'entrée verbale grâce au choix spécifique des termes qui occupent les positions d'arguments.

9) *Le médecin admet un malade dans ce service*.

La phrase 9 décrit certaines propriétés syntaxiques et sémantiques d'*admettre*, qui est un verbe transitif exigeant un complément d'objet direct et un complément prépositionnel introduit par *dans*. Comme le montre la phrase 13, les mots occupant les positions d'arguments dans les phrases exemples du LVF (*médecin, malade, service*) représentent le nombre et les types sémantiques d'arguments attendus, et respectent les restrictions de sélection qu'impose le verbe dans l'emploi proposé.

De plus, les phrases exemples à caractère « élémentaire » [43] du LVF (François *et al.*, 2007) sont conçues de façon à illustrer la possibilité d'insertion d'une complétive ou d'une interrogative indirecte, les transformations (passive, impersonnelle, réfléchie, etc.) et les variantes syntaxiques des patrons verbaux.

Grâce à ces caractéristiques, on peut envisager un mapping entre le LVF et une ressource terminologique décrivant l'une des langues de spécialité prises en considération par cette ressource verbale, de façon à extraire des patrons spécialisés des verbes. C'est ce qui a été entrepris au départ dans notre travail, comme méthode d'identification des patrons verbaux à simplifier. Dans une étude expérimentale [44], nous avons essayé d'acquérir des patrons médicaux de verbes à partir des phrases exemples du LVF, d'une terminologie médicale (cf. section 2.2.2) et des textes extraits de nos corpus (Wandji Tchami *et al.*, 2016). La méthode consistait à comparer la liste de PSS extraits du LVF à ceux extraits des corpus, afin d'identifier les PSS apparaissant dans les deux ressources, qui de ce fait constitueraient de potentiels patrons verbaux spécialisés, candidats pour la simplification. Cette méthode a produit des résultats satisfaisants sur le plan qualitatif. La technique d'acquisition des PSS à partir du LVF fonctionnait bien. Par contre, sur le plan quantitatif, l'ensemble de patrons verbaux extraits du LVF était très limité, comparé au nombre élevé de patrons extraits des corpus. Par conséquent, le nombre de patrons

43. Le mot *élémentaire* dans cette phrase signale la simplicité de la langue utilisée dans les exemples du LVF.
44. Dans une telle approche, l'utilisation du LVF est considérée comme une méthode de validation des patrons médicaux, puisque les patrons médicaux du LVF représentent des données attestées par des experts.

apparaissant communément dans les deux sources n'était pas satisfaisant (extrêmement faible) pour atteindre l'objectif final, à savoir, la création d'une ressource de simplification de textes.

1.3.3.4 Dicovalence

Dicovalence est un dictionnaire électronique qui fournit les cadres de valence de plus de 3 700 verbes simples du français, pour plus de 8 000 entrées (car de nombreux infinitifs sont polysémiques). Un cadre de valence renvoie au nombre et à la nature des arguments du verbe, y compris le sujet.

Ce dictionnaire de valence a été conçu selon l'approche pronominale, qui caractérise la valence à partir des paradigmes de pronoms proportionnels (qui décrivent les fonctions grammaticales des arguments) qu'accepte le verbe (Karel van den & Piet, 2003 ; Van Den Eynde & Mertens, 2006 ; Mertens, 2010). En d'autres termes, dans les schémas valenciels des verbes proposés par Dicovalence, les arguments des verbes sont remplacés par des pronoms, tel que le montre l'exemple de l'image 1.2, qui décrit un cadre de valence du verbe *admettre*[45]. Les positions des arguments sujet, compléments d'objet direct et complément d'objet prépositionnel sont occupées par des étiquettes du type *P+num*, où *P* est mis pour pronom et le numéro indique le type d'argument dont il s'agit : *P0* représente le sujet, *P1* le complément d'objet direct, et *PL* un complément indiquant le lieu.

```
VAL$    admettre: P0 P1 (PL)
VTYPE$  predicator simple
VERB$   ADMETTRE/admettre
NUM$    2270
EG$     il est difficile de se faire admettre dans ces milieux
TR_DU$  toelaten, binnenlaten, ontvangen, opnemen
TR_EN$  admit
FRAME$  subj:pron|n:[hum], obj:pron|n:[hum], ?loc<>:pron|n:[]
P0$     qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1$     que, qui, te, vous, la, le, les, se réc., en Q, celui-ci, ceux-ci, l'un l'autre
PL$     0, où, là, ici, là-bas
RP$     passif être, se faire passif
AUX$    avoir
```

FIG. 1.2 – Dicovalence : exemple de cadre de valence

Comme le montre l'image 1.2, pour chaque cadre de valence proposé, la ressource fournit également la fonction syntaxique des arguments et décrit en outre : les restrictions de sélection, les différentes formes de réalisation (pronominales, phrastiques) des arguments, les transformations passives possibles sur le cadre valenciel proposé, les traductions du verbe en anglais et en néerlandais, ainsi que, le cas échéant, les liens avec d'autres cadres de valence du même verbe.

45. Notez que ce cadre de valence du verbe *admettre* correspond à l'emploi décrit dans l'exemple 13 (cf. section 1.3.3.3). L'on peut ainsi observer les différences entre les modèles de description des patrons verbaux proposés par les deux ressources.

1.3.4 Bilan

Dans cette partie, nous avons présenté un état des faits en ce qui concerne l'étude des verbes dans les travaux de recherche en Terminologie, Linguistique et TAL, l'objectif étant de mieux cerner les cadres théoriques dans lesquels s'inscrit notre travail de recherche. Différentes approches théoriques, méthodes et outils ont été exposés, nous permettant de présenter ceux qui ont été sélectionnés pour la réalisation de notre travail de thèse. Force a été de constater que les travaux en terminologie, de par les principes de base de cette discipline, n'ont pas toujours donné au verbe une place de choix, contrairement à ceux effectués dans une perspective plus linguistique, et par conséquent au TAL, où le verbe est étudié au même titre que les autres catégories grammaticales, en l'occurrence les entités nominales. Cependant, nous avons noté une évolution croissante dans les travaux sur les langues de spécialité, à travers différentes approches et outils existants dédiés spécialement aux verbes.

De plus, cet état de l'art nous a permis de relever qu'il n'existe pas de travaux, dans le domaine de la simplification de textes, qui soient aussi axés que la nôtre sur le verbe au sein de sa structure argumentale. La méthode proposée dans la présente étude est centrée sur l'analyse des propriétés syntaxico-sémantiques des verbes, notamment sur l'analyse des propriétés sémantiques [46] des arguments des verbes dans des corpus médicaux. Le but de cette démarche est d'identifier les patrons syntaxico-sémantiques spécialisés des verbes [47], entrées potentielles pour la ressource de simplification qui constituera le résultat final de ce travail de thèse.

Cette approche encore non explorée [48] propose une méthode innovante qui de surcroît est à l'interface des différentes disciplines sur-citées. La nouveauté qu'apporte notre approche touche plus précisément le domaine de la simplification des textes. Les données traitées dans cette étude sont tirées de quatre corpus médicaux de niveaux de spécialisation différents basés sur le type de public visé (experts en médecine, étudiants, patients et grand public). L'annotation syntaxique de ces corpus est réalisée grâce à l'analyseur syntaxique Cordial, qui effectue une analyse dépendancielle des constituants syntaxiques des phrases, à partir des verbes pivots. Contrairement à ce qui se fait dans la plupart des travaux appliquant une *frame-based* approche [49], l'annotation sémantique de nos corpus est basée sur les catégories sémantiques d'une terminologie médicale, la Snomed International. Elle classifie les termes médicaux en

46. Dans notre étude, ces propriétés sémantiques renvoient aux catégories sémantiques acquises à partir de la terminologie médicale Snomed (cf. chapitre 2, section 2.2.2).

47. Il s'agit ici des PSS fort spécialisés qui peuvent engendrer des difficultés de compréhension chez les lecteurs non-experts.

48. À notre connaissance, il n'existe aucune étude de ce type pour le français.

49. En général, les études comme la nôtre, qui s'inspirent de l'approche FN, font l'annotation sémantique des corpus à partir des listes de rôles sémantiques adaptées ou similaires à celles du projet FN (Ruppenhofer *et al.*, 2006).

11 catégories, selon le type d'entités médicales qu'ils dénotent. Ces catégories sémantiques, associées aux arguments des verbes, nous permettent de faire la distinction entre les différentes acceptions et surtout d'identifier les sens spécialisés des verbes, qui seront par la suite alignés avec des équivalents non spécialisés. En ce qui concerne les résultats de notre travail, dans le domaine du TAL, la ressource qui résultera de ce projet de thèse pourrait être intégrée comme dictionnaire dans une machine conçue pour la simplification des textes médicaux spécialisés. Par ailleurs, dans le domaine médical, cette ressource pourrait servir d'outil d'aide au décodage, pour des patients ayant des difficultés à lire des textes médicaux. De plus, notre ressource pourrait également être utilisée comme une aide à l'encodage, pour des experts en médecine qui souhaitent adapter leur vocabulaire à celui des non-experts, dans un processus de composition des textes pour le grand public.

1.4 La simplification de textes

D'après Siddharthan, la simplification automatique de textes est une tâche du TAL qui consiste à cibler et à remplacer les éléments qui empêchent la compréhension aisée d'un texte, tout en sauvegardant le sens de ce texte (Siddharthan, 2014a,b). La simplification peut également consister à enrichir le texte cible d'informations (explications, définitions, exemples) facilitant sa compréhension par le lecteur.

Les problèmes de compréhension sont très souvent causés par la complexité linguistique des textes, qui touche principalement le lexique et la syntaxe. En effet, les complexités lexicales et syntaxiques sont reconnues comme étant de grandes causes de difficultés de lecture (Chall & Dale, 1995), en particulier chez les jeunes enfants (Belder & Moens, 2010), les apprenants d'une langue étrangère (Siddharthan, 2006 ; Petersen & Ostendorf, 2007 ; Medero & Ostendorf, 2011), les personnes présentant certaines maladies ou des déficiences intellectuelles (Carroll *et al.*, 1998, 1999 ; Inui *et al.*, 2003 ; Daelemans *et al.*, 2004 ; Huenerfauth *et al.*, 2009) et les personnes ayant un faible niveau d'éducation (Richard *et al.*, 1993 ; Williams & Reiter, 2005 ; Candido *et al.*, 2009).

Par ailleurs, la difficulté de lecture (cf. chapitre 3, section 3.3.2) est un phénomène fréquemment rencontré chez les personnes non expertes confrontées à la lecture de textes spécialisés d'une certaine discipline. Dans le domaine de la pharmacologie par exemple, Patel *et al.* (2002) étudient les erreurs observées chez des sujets de différentes origines culturelles et éducatives, au cours des processus cognitifs déployés dans le but de comprendre les textes inscrits sur des étiquettes pharmaceutiques décrivant des procédures médicales. Les auteurs démontrent que la plupart de leurs sujets rencontrent d'importantes difficultés dans la compréhension des instructions à respecter pour la bonne administration du médicament testé. En médecine, une étude réalisée en 2005 et visant à promouvoir la facilitation de l'accès à l'information en

matière de santé déplore le fait que les patients ne bénéficient pas véritablement du contenu des documents qui sont sensés les éduquer. McCray (2005) souligne entre autres le fait que les documents écrits (ordonnances, diagnostiques, orientations thérapeutiques, revues, etc.) sont souvent trop techniques pour les malades qui ne disposent pas du niveau de compétences requis pour comprendre ce type de textes fortement spécialisés.

Dès lors, le but de la simplification automatique de textes est de rendre le contenu d'un document écrit facilement compréhensible pour les lecteurs qui y sont confrontés (Elhadad & Sutaria, 2007 ; Deléger & Zweigenbaum, 2008). Par ailleurs, la simplification de textes est également utilisée comme une méthode de prétraitement de textes, dont le but est d'améliorer les performances des systèmes d'analyse syntaxique ou de traduction automatique (Raman *et al.*, 1996 ; Lucia, 2010) ; de génération automatique de questions (Heilman & Smith, 2010) ; et d'extraction de données, dans le domaine du biomédical (Lin & Wilbur, 2007 ; Jonnalagadda *et al.*, 2010 ; Jonnalagadda & Gonzalez, 2011). Malheureusement, on dénombre peu de travaux en simplification qui portent sur la langue française. Minard *et al.* (2011) fournissent une possible explication à cette situation, s'appliquant au contexte médical français, mais cette explication pourrait également s'appliquer à d'autres domaines. Ils soulignent la pénurie de corpus monolingues de documents comparables, ayant un haut degré de similitude, c.-à-d. traitant des mêmes thématiques. Par exemples des paires d'articles scientifiques [50], ou des reportages, alignés avec leurs équivalents écrits en langue courante. Pourtant, ce type de corpus alignés existe bien pour l'anglais (Barzilay & Elhadad, 2003 ; Yusuke & Satoshi, 2003).

La simplification peut porter sur les caractéristiques de surface des textes (nombre de caractères et de syllabes par mot), la capitalisation, la ponctuation, les ellipses (Tapas & Orr, 2009). Toutefois, de nos jours, la syntaxe et le lexique sont au coeur des travaux de simplification.

1.4.1 La simplification syntaxique

La plupart des travaux en simplification reposent sur la syntaxe. La grande majorité des approches de simplification syntaxique sont basées sur un ensemble de règles de transformation définies manuellement pour être appliquées aux textes à simplifier. Cette manière de procéder est un héritage des premières expériences dans le domaine de la simplification (Hoard *et al.*, 1992 ; Raman *et al.*, 1996). Il s'agit principalement de règles de substitution ou remplacement, de suppression, de modification, de division et de regroupement qui permettent par exemple de raccourcir les longues phrases, de segmenter les propositions coordonnées et subordonnées, en extrayant les appositions et les propositions relatives. Brouwers *et al.* (2012) proposent un système de simplification de textes français (portant sur la langue générale) qui repose sur 19

50. Les auteurs font référence à des articles scientifiques ayant un fort niveau de spécialisation.

règles de suppression, modification et division définies manuellement après une étude de corpus. La méthode appliquée passe par un processus de surgénération qui permet d'obtenir, dans un premier temps, un nombre important de substituts possibles pour une seule phrase. Cette étape est ensuite suivie d'une phase de sélection des meilleures simplifications produites par le système, en fonction de critères de lisibilité. D'après les résultats obtenus, 80 % des phrases générées par le système sont bonnes.

La définition manuelle de règles de grammaire est une tâche qui permet certes d'atteindre l'objectif de simplification, mais qui parallèlement requiert beaucoup de temps. C'est l'argument qu'avancent Chandrasekar et Srinivas dans une étude où ils implémentent un algorithme permettant d'induire automatiquement des règles de simplification à partir d'un corpus de textes syntaxiquement complexes, alignés avec leurs correspondants simplifiés (Chandrasekar & Srinivas, 1997).

Certains travaux récents combinent les deux approches, et proposent une méthode basée sur des règles définies manuellement et automatiquement. Angrosh *et al.* (2014) développent un système hybride de simplification de textes anglais, basé sur les grammaires de dépendance, qui utilise un petit ensemble de règles définies manuellement, associé à un important groupe de règles acquises automatiquement et appliquées aux constructions lexicalisées. D'après les auteurs, les résultats obtenus après évaluation montrent la supériorité de leur système novateur sur les autres systèmes existants.

1.4.2 La simplification lexicale

En général, la simplification lexicale porte sur les syntagmes ou sur les unités nominales de la phrase. Les travaux de la littérature montrent que la simplification lexicale a été jusqu'ici moins sollicitée que la simplification syntaxique. D'un point de vue purement technique, certains chercheurs assimilent la simplification lexicale à l'identification des paraphrases (Biran *et al.*, 2011 ; Androutsopoulos & Malakasiotis, 2010 ; Specia *et al.*, 2012). Cette approche a été appliquée dans le cadre de la tâche numéro 10 du challenge *SemEval-2007*, qui portait sur la substitution lexicale (Diana & Roberto, 2007). Cependant, une telle tâche ne prend pas en considération la complexité linguistique et la difficulté de lecture des textes, qui sont pourtant des questions centrales dans le cadre de la simplification.

Tout comme la simplification syntaxique, la simplification lexicale passe par l'utilisation de règles de transformation, principalement de substitution. Biran *et al.* (2011) implémentent un outil [51] de simplification de textes conçu selon l'approche non supervisée, qui est capable de traiter des textes de différents domaines de spécialité. Ce système génère, à partir d'un corpus

51. D'après les auteurs, l'outil devrait être téléchargeable sous le lien suivant :http ://www.cs.columbia.edu/ob2008/

de textes parallèles, des règles de simplification à appliquer à ces textes et, à partir de ces règles, forme des paires de noms synonymes (complexe vs. simple) grâce à leurs scores de similarité. La phase de simplification proprement dite consiste en l'application d'opérations de substitution automatique des formes complexes d'unités nominales par leurs équivalents simples. Les résultats de l'évaluation indiquent que le système proposé est efficace pour la simplification des mots qui apparaissent fréquemment dans les textes, tandis que les performances sont moins bonnes avec les mots peu fréquents.

En ce qui concerne le domaine médical, des recherches impliquant différentes langues ont été effectuées. Dans la plupart de ces études, la règle de substitution de synonymes représente la principale opération de simplification appliquée aux textes. Dans une étude portant sur des textes médicaux en suédois, Abrahamsson *et al.* (2014) développent un outil de simplification qui remplace les termes médicaux difficiles par des synonymes (considérés comme étant plus faciles à comprendre), en appliquant deux métriques de lisibilité suédoises (LIX [52] (Falkenjack *et al.*, 2013) et OVIX [53] (Mühlenbock & Johansson Kokkinakis, 2009)) aux textes traités. Dans ce travail, la complexité d'un terme est évaluée grâce à sa fréquence dans un corpus de langue générale, mais aussi à celle de ses sous-chaînes. Les résultats de l'étude sont partagés car ils varient selon la métrique utilisée. En effet, la métrique OVIX renvoie un texte un peu plus lisible que sa version originale, tandis que la métrique LIX retourne un texte un peu plus difficile à lire.

Plusieurs travaux s'intéressent au domaine médical en ce qui concerne la langue anglaise. Elhadad (2006) développe un outil qui identifie les termes difficiles [54] dans un texte médical et récupère sur Internet les définitions de ces termes grâce au moteur de recherche Google. Les définitions récupérées sont fournies sous forme de cible de liens vers les termes. Les résultats de ce système indiquent une amélioration de la compréhension du lecteur de 1,5 points en moyenne, sur une échelle de 5 points. Dans une autre étude, Elhadad & Sutaria (2007) présentent une méthode qui permet de construire un lexique de couples de termes médicaux (techniques et courants) sémantiquement équivalents, en utilisant un corpus parallèle constitué de résumés d'études cliniques et d'articles de presse correspondants, écrits pour un public profane. D'après les auteurs, la méthode fournit des résultats prometteurs malgré la petite taille du corpus utilisé pour cette expérience. Kandula *et al.* (2010), quant à eux, analysent un outil de simplification qu'ils ont développé en 2007 (Qing *et al.*, 2007) et qui est essentiellement consacré à l'identification et à la substitution de termes médicaux difficiles par des synonymes plus faciles, ou à l'apport d'explications en un langage facilement compréhensible. Les auteurs décrivent les améliorations apportées à leur outil, notamment la prise en compte de nouveaux types de relations lexicales (hiérarchiques et/ou sémantiques) et de paramètres syntaxiques

52. En suédois, *Läsbarhetsindex* c.-à-d. *Readability index*.
53. En suédois, *Ordvariationsindex* c.-à-d. *Word variation index*.
54. À propos de la difficulté de lecture des textes de spécialité, voir la section 3.3.2 du chapitre 3.

(segmentation des longues phrases) qui, combinés, permettent d'améliorer de façon significative les performances du système, de 35,8 % à 43,6 % sur les textes tirés de dossiers médicaux électroniques. Leroy *et al.* (2012) développent un algorithme qui utilise le degré de familiarité des termes pour identifier les textes difficiles et sélectionner des alternatives plus faciles à partir de ressources lexicales telles que WordNet, UMLS et Wiktionary. Bien que la méthode proposée n'ait pas eu un effet significatif sur les textes, les résultats de cette étude montrent clairement l'importance de la prise en compte de la familiarité des termes dans le contexte de la simplification de texte.

Le français est un peu moins représenté dans le domaine de la simplification des textes médicaux. Néanmoins, certains groupes de recherche développent de plus en plus des travaux dans ce sens. En effet, dans un projet visant à concevoir des moyens d'adapter l'information médicale spécialisée à des patients non-experts, Deléger & Zweigenbaum (2008) mettent en place une méthode basée sur la construction de corpus comparables de textes médicaux experts et non-experts. Cette méthode permet d'identifier des segments de textes similaires entre les langues courante et spécialisée, et ainsi de détecter les expressions médicales (nominalisations et noms composés) et leurs équivalents en langue courante. Les auteurs obtiennent de meilleurs résultats avec les nominalisations qu'avec les noms composés, qui se sont avérées moins productives. En effet, les résultats obtenus pour les nominalisations indiquent que leur prise en compte participe véritablement à l'évaluation de la différence entre les textes médicaux fortement spécialisés et ceux qui appartiennent à la langue courante. Dans le cadre du challenge *i2b2/VA 2010* sur l'extraction automatique de concepts médicaux et l'annotation des assertions portant sur ces concepts et des relations entre ces derniers, Minard *et al.* (2011) développent des méthodes hybrides pour améliorer l'accès à l'information dans les documents cliniques. Leurs approches reposent à la fois sur des règles et sur des méthodes de l'apprentissage automatique. Des méthodes du traitement du langage naturel sont aussi utilisées pour extraire les caractéristiques des textes d'entrée. Ces caractéristiques sont ensuite utilisées lors de l'application des techniques d'apprentissage automatique.

1.5 Bilan

Dans ce chapitre, nous avons présenté les différents cadres théoriques [55], outils et méthodes à l'interface desquelles se situe notre travail de thèse. Après avoir discuté des défis que présente la communication entre les experts en médecine et les non-experts, nous avons décrit quelques méthodes et outils disponibles en linguistique, terminologie et TAL qui s'intéressent au verbe pris dans son contexte d'apparition (structure argumentale). Nous nous intéressons dans cette thèse

55. Cet état de l'art a fait l'objet d'une publication lors de la conférence TALN (Wandji, 2014).

aux patrons syntaxico-sémantiques spécialisés des verbes dans les textes médicaux. Notre but est de proposer des équivalents en langue courante de ces patrons, afin de permettre une lecture plus aisée de textes médicaux hautement spécialisés par des patients non-experts. Comme le démontrent les études présentées dans cette partie, la simplification de textes appliquée à des textes de spécialité, et en particulier dans le domaine médical, est un champ de recherche encore très peu exploré en ce qui concerne la langue française. De plus, des méthodes comme la nôtre, c'est-à-dire une approche de simplification axée sur le verbe au sein de sa structure argumentale, sont quasiment inexistantes à notre connaissance. Nous proposons donc une approche de simplification lexicale qui utilise une terminologie médicale pour la détection des frames spécialisés des verbes et des corpus de forums pour l'extraction des variantes non spécialisées de ces frames.

Chapitre 2

Corpus, ressources et outils

Ce chapitre est consacré à la présentation des corpus que nous avons utilisés pour nos expérimentations, ainsi que les ressources et outils qui interviendront dans le processus de traitement des données extraites des corpus.

2.1 Corpus

2.1.1 Types

Les données analysées dans ce travail de thèse proviennent d'un grand corpus constitué de 4 corpus de textes de différents types. La définition et la distinction entre *type*, *genre*, et *registre* de textes se trouve au centre de nombreux débats qui perdurent depuis plusieurs années dans le domaine de la linguistique (Halliday & Hasan, 1989 ; Swales, 1990 ; Biber & Conrad, 2009). Pour éviter toutes ambiguités qui tourneraient autour de ces notions, dans le cadre de ce travail, nous utiliserons l'expression *types* de textes pour désigner des textes écrits par des experts pour différents publics cibles qui se distinguent de par leur niveau d'expertise en termes de connaissances médicales.

Les textes de nos corpus pourraient bien rentrer dans le continuum que dresse Pearson (1998). L'auteure classifie les textes spécialisés dans trois catégories, selon leur niveau de spécialisation qui est évalué en fonction de quatre critères : l'auteur du texte, le lecteur potentiel visé, la structure interne du texte (la syntaxe) et le choix des unités lexicales utilisées (le lexique). Dans la typologie de Pearson, le premier critère, c.-à-d. l'auteur, est le principal, étant donné qu'il a une certaine influence sur les autres critères. Toutefois, dans notre approche, le public cible joue également un rôle central, car il détermine voire impose une forme au contenu du texte que rédige l'auteur. Les ensembles de textes que nous analysons dans ce projet illustent cette vision, car ils se distinguent principalement par le niveau d'expertise du public visé.

En effet, les textes de nos corpus ont été collectés à partir d'un portail (*CISMeF*) qui fait la classification des textes médicaux à partir du type de public auquel s'adressent ces textes. Ce portail a été conçu selon une approche qui rejoint les principes de base définis pour la formation de nos corpus. Ces principes sont décrits dans la section 2.1.2.1.

Ainsi, les textes de notre corpus entrent dans quatre catégories différentes dont les trois premières correspondent à trois niveaux de spécialisation différents que Pearson (1998) nomme et classifie comme suit :

— *High* : niveau de spécialisation élevé, textes visant un public d'experts ;

— *Medium* : niveau de spécialisation moyen, textes visant un public de personnes ayant un certain niveau de connaissance dans le domaine, à l'instar des étudiants ;

— *Low* : niveau de spécialisation faible, textes visant un public de non-experts, c.-à-d. le grand public.

Les trois premières catégories de corpus contiennent donc des textes écrits par des experts en médecine, soit pour des experts en médecine, soit pour des étudiants en médecine, soit pour des patients et, de façon générale, le grand public.

La quatrième catégorie de textes qui a été prise en considération dans ce travail, mais qui n'est pas évoquée par Pearson (1998), contient des textes de forums, c.-à-d. écrits par des non-experts pour des non-experts. Il s'agit plus précisément d'échanges entre participants sur des forums médicaux attestés (cf. section 2.1.2.2).

2.1.2 Sources

2.1.2.1 CISMeF

CISMeF[1] (Darmoni *et al.*, 1999) signifie *Catalogue et Index des Sites Médicaux de langue Française*. Ce projet a été lancé en 1995, par le Centre Hospitalier-Universitaire (*CHU*) de Rouen.

FIG. 2.1 – Page d'accueil du CISMeF : formulaire de requêtes

Comme son nom l'indique, ce portail indexe les sites et documents médicaux français à partir du lexique que propose le thésaurus MeSH (NLM, 2001). Le but du CISMeF, tel que spécifié

1. www.chu-rouen.fr/cismef

sur sa page d'accueil, est de « faciliter l'accès à l'information de santé pour les professionnels, mais aussi les patients et le grand public, en recensant les sites et documents médicaux présents sur l'Internet répondant à certains critères de qualité ». Cette phrase [2] qui décrit la mission du CISMeF met en évidence quelques éléments de base qui rejoignent les principes établis pour la conception de notre corpus, à savoir une typologie ou catégorisation des textes axée sur le type d'audience visée, la thématique abordée par les différents textes, notamment la médecine, et la fiabilité des informations diffusées. En effet, la phrase ci-dessus stipule clairement que les sites et textes proposés par le CISMeF touchent différentes audiences, en l'occurrence les experts et le grand public ; qu'ils concernent le domaine médical, et qu'ils sont soumis à une étape de validation qui permet d'évaluer la qualité de l'information diffusée, ainsi que leur degré de fiabilité. Darmoni *et al.* (1999) l'attestent en ces termes : « CISMeF respecte le référentiel des critères de qualité de l'information de santé sur l'Internet (Net Scoring), élaboré en collaboration avec Centrale santé et APUI-Santé ».

La typologie des textes du CISMeF et surtout la rigueur du processus de sélection des sites et textes à indexer favorisent leur utilisation, non seulement par les publics visés, mais également par les chercheurs (Grabar *et al.*, 2002, 2003 ; Chebil *et al.*, 2014 ; Névéol *et al.*, 2014 ; Cabot *et al.*, 2016), et bien évidemment motivent notre choix de nous servir du portail CISMeF.

Le portail CISMeF classifie les pages Web et les documents indexés selon trois axes :

1. « la médecine factuelle » : cet axe concerne les professionnels de la santé et propose des recommandations pour la bonne pratique clinique, ainsi que des conférences de consensus.
2. « les ressources concernant l'enseignement » : elles contiennent des documents à caractère didactique, ainsi que des épreuves classantes nationales.
3. « les documents spécialement adressés aux patients et au grand public », dont le but est de favoriser l'amélioration de l'éducation sanitaire dans les pays francophones.

Tel qu'on peut l'observer à travers les zones mises en évidences sur le formulaire de requête de la page d'accueil du portail CISMeF (cf. figure 2.1), les requêtes peuvent se faire par ordre alphabétique [3]. La base de données CISMeF peut également être interrogée par thématiques (pathologies, traitement, médicaments, etc.). Ces thématiques permettent de recupérer des textes qui touchent différentes spécialités [4] de la médecine : cardiologie, pédiatrie, chirurgie, dermatologie, gynécologie, médecine générale, médecine palliative, etc. C'est de cette manière que nous avons procédé lors de la collection des textes que comportent les différents corpus. Les requêtes formulées avaient pour mots-clés les noms des différents axes ou catégories de

2. http ://www.chu-rouen.fr/cismef/Aide/
3. Index consultable via le lien suivant : http ://www.chu-rouen.fr/page/index/
4. CISMeF couvre une longue liste de spécialités, liste consultable via le lien suivant : http ://www.chu-rouen.fr/ssf/santspe.html

la terminologie médicale Snomed International (cf. section 2.2.2) utilisée dans ce travail. En procédant de cette façon, nous voulions nous assurer que les thématiques couvertes par notre corpus étaient autant que possible en accord avec les termes de la ressource Snomed, ceci dans le but de favoriser l'obtention de résultats positifs lors de l'appariement entre le corpus et la ressource médicale (cf. chapitre 3, section 3.2.2). L'éventail de thématiques considéré nous permettra d'avoir une couverture relativement large du domaine médical.

Les requêtes peuvent être affinées par le choix du type de public que visent les textes souhaités ; trois options sont proposées : *recommandations professionnelles*, *documents d'enseignement*, *documents grand public*. Il en ressort que la constitution de notre corpus n'était pas focalisée sur des spécialités ou domaines médicaux particuliers mais d'avance autour de certaines thématiques comme celles mentionnées ci-dessus. Les requêtes faites sur CISMeF ne visaient pas une liste de thématiques particulières.

Les différentes propriétés du portail CISMeF ici décrites nous ont permis d'acceder aisément aux textes qui constituent les trois premières parties du corpus : corpus des experts, corpus des étudiants et corpus des patients.

2.1.2.2 Les forums médicaux : Doctissimo

Doctissimo.fr[5] est une plateforme médicale qui appartient au groupe *Lagardère Active*[6]. Elle a été créée en mai 2000, et est dédiée au bien-être et à la santé. Doctissimo est un réseau social ouvert au grand public, à qui il offre divers services et ressources médicales : une encyclopédie médicale, un dictionnaire médical, un atlas du corps humain, un guide des médicaments, un guide des examens de laboratoire, des forums de discussion. La figure 2.2 présente quelques sujets abordés sur les forums de Doctissimo.

	Chimiothérapie	3 994	09/03/2017 à 14:10 **par Marydu40**
	Chirurgie : préparation et suites opératoires	31 431	13/03/2017 à 15:01 **par bubule**
	Cholestérol	32 159	11/03/2017 à 22:12 **par exper**
	Constipation, autres troubles du transit	94 488	13/03/2017 à 08:51 **par love-cook**
	Contraception	1 437 220	13/03/2017 à 15:30 **par keshogua**
	Cuisine minceur	78 934	13/03/2017 à 13:43 **par acerolabresil**
	Cystites et problèmes urinaires	73 476	12/03/2017 à 21:51 **par Rowca21**

FIG. 2.2 – Quelques sujets abordés sur les forums doctissimo.

Les forums de discussion *Doctissimo* se veulent un espace d'écoute et de dialogue, où les

5. http ://www.doctissimo.fr/
6. http ://www.lagardere.com/activites/lagardere-active-2610.html

participants peuvent s'exprimer anonymement et en toute confiance, afin de bénéficier en retour de l'expérience des autres internautes sur les questions qui les intéressent. Les thématiques abordées sur ces plateformes tournent autour du domaine de la santé et touchent différentes spécialités de la médecine : *médicament*, *maladie*, *procédure médicale*, *grossesse*, *nutrition*, *sexualité*, *enfant*, *psychologie*, et bien d'autres. Ces thèmes rejoignent ceux que contiennent les corpus collectés à travers le portail CISMeF, ce qui nous offre la possibilité de croiser les données qui seront extraites de part et d'autre.

2.1.3 Taille et Contenu

Le tableau 2.1 présente un récapitulatif de la description des différents corpus en termes de taille et de contenu. La longueur moyenne des phrases (*Long. moy./phr*) de chaque corpus est aussi fournie, elle est exprimée en termes de nombre de mots.

Corpus	Taille (occ. mots)	Long. moy./phr. (nb. mots)	Sources	Contenu
C_1 / expert	1,502,690	28,39	CISMeF	recommandations, rapports, publications scientifiques
C_2 / étudiant	1,755,497	24,78	CISMeF	documents d'enseignement et épreuves classantes nationales
C_3 / patient	1,627,466	20.62	CISMeF	textes de vulgarisation, brochures, conseils et consignes
C_4 / forum	1,588,697	21,18	Doctissimo	messages des participants sur des forums de santé

TAB. 2.1 – Taille et contenu des 4 corpus.

On peut remarquer que les corpus sont de tailles plus ou moins similaires, avec un peu plus d'un million et demi de mots chacun. En ce qui concerne leur contenu, les quatre corpus traitent de sujets variés qui touchent le domaine médical : maladies, procédures médicales, soins et suivi, médicaments, etc., et ces différents sujets relèvent de diverses spécialités de la médecine : cardiologie (qui est la plus représentée dans les corpus), pédiatrie, chirurgie, dermatologie, gastro-entérologie, gériatrie, gynécologie, hématologie, hépatologie, infectiologie, pneumologie, psychiatrie, radiologie, rhumatologie, médecine générale, médecine palliative, et bien d'autres spécialités médicales qui sont associées aux documents fournis par le CISMeF, comme il a été expliqué dans la section 2.1.2.1.

2.1.3.1 Corpus des experts

En tant que pôles extrêmes du continuum que constitue notre corpus, le corpus des experts et le corpus des forums jouent un rôle déterminant dans cette étude. En effet, ils représentent les

deux principaux niveaux d'expertise qui s'opposent, et ainsi, constituent des sources de données par excellence à exploiter pour la conception de la ressource de simplification. Pour cette raison, dans cette section, les corpus C_1 et C_4 bénéficieront d'une description relativement détaillée, en comparaison aux deux autres corpus qui sont considérés comme des corpus intermediaires.

Encore appelé C_1, le corpus des experts contient des textes écrits par des experts en médecine pour des experts en médecine. D'après la classification de Pearson (1998), ce type de corpus appartiendrait à la catégorie de textes qu'elle caractérise de *highly specialised texts*, c.-à-d. des textes fort spécialisés ou encore des textes relevant du *discours scientifique primaire* (Jacobi, 1993). Ce corpus regroupe des textes de rapports des organisations nationales et internationales de la santé (Haute Autorité de Santé, Organisation Mondiale de la Santé, Santé Canada, etc.) ; des recommandations et règlementations adressées aux professionnels de la santé ; des textes provenant de publications scientifiques telles que des articles, ainsi que des textes informatifs par rapport aux maladies, et procédures médicales.

Les textes du corpus des experts sont caractérisés par un niveau de langue soutenu et une très forte fréquence des termes médicaux spécialisés. L'abondance des termes spécialisés dans les textes de spécialité est une question bien connue dans la littérature, depuis les débuts de la terminologie, comme le démontrent les travaux de Jacobi (1993), qui se focalisent sur l'étude des termes que l'auteur appelle *les terminologies*. En effet, les unités terminologiques font partie des propriétés intrinsèques d'un texte spécialisé, étant donné que l'étude du terme, de ses fonctions et de sa relation avec le concept, constitue un principe fondamental de la terminologie de Wüster (Wüster, 1981 ; Wüster, 1985) (cf. chapitre 1, section 1.3.2). Ce caractère est d'autant plus connu que l'on a pendant longtemps pensé que les difficultés de lecture des textes scientifiques spécialisés tenaient exclusivement (ou presque) à l'usage d'un vocabulaire spécialisé (Guilbert, 1973).

Le corpus des experts est aussi marqué par la prédominance des unités complexes (Collet, 1997 ; Portelance, 1991) qui d'après Jacques (2003) représentent 80% des éléments constituant les textes de spécialité. En effet, le corpus des experts est caractérisé par l'abondance de termes ayant une structure complexe. Ils se caractérisent par le nombre d'unités linguistiques qu'ils mettent en jeu :

— *source d'aggravation de l'état clinique des infections urinaires*, *inhibiteur de l'enzyme de conversion d'une fréquence plus élevée d'hypotension artérielle* ; *augmentation de la concentration de la créatinine sérique, hypertension artérielle contemporaine de la phase aigüe d'un AVC ischémique* ;

— *bloc atrioventriculaire du deuxième ou du troisième degré ou un syndrome de dysfonctionnement sinusal* ;

— *nausée ou vomissement d'une thrombocytopénie ou d'une perturbation des tests hépatiques*.

La variation terminologique (au niveau des entités nominales) fait également partie des phénomènes qui décrivent les textes du corpus des experts. Elle peut désigner différents types de transformations (morphologique, lexicale, sémantique, syntaxique, etc.) qui caractérisent les textes en langues de spécialités et porter sur différents types d'unités de la langue : les entités nominales (Grabar, 2004), les collocations (Giacomini, 2015), etc. Cependant, dans cette étude, nous nous intéressons uniquement à la variation morpho-lexicale observée au niveau des termes nominaux. Ce type de variation pourrait être définie comme phénomène par lequel certains termes complexes se réalisent sous des formes diverses, que l'on appelle *variantes*. C'est ce que Haralambous & Lavagnino (2011) appellent la *polymorphie* :

— *ischémie cérébrale - ischémie focale cérébrale - ischémie focale cérébrale ou rétinienne* ;

— *ischémie myocardique - ischémie du myocarde*.

La variation terminologique est une question bien connue dans le domaine de la terminologie, et elle fait l'objet de nombreuses études (Jacquemin, 1997 ; Grabar, 2004 ; Tartier, 2006). Giacomini (2015) souligne, avec illustrations à l'appui, deux exemples de situations dans lesquelles la variation terminologique est susceptible d'intervenir : lorsque différents niveaux de spécialisation sont exprimés, comme dans l'interaction médecin (pneumologue) vs. patient (spécialiste pulmonaire) ; et lorsque des emprunts avec ou sans adaptation coexistent dans le même domaine (dans le langage médical *ECG* vs EKG). Dans le domaine du TAL, la variation terminologique constitue la base de la tâche que Grabar *et al.* (2002) définissent comme l'identification d'expressions différentes de notions identiques ou proches, c.-à-d. la détection de différentes variantes de termes exprimant une même notion (McCray *et al.*, 1994 ; Lovis *et al.*, 1995 ; Hamon *et al.*, 1998 ; Jacquemin & Tzoukermann, 1999 ; Lovis & Baud, 2000 ; Pouliquen, 2002).

Le sous-corpus des experts est également caractérisé par une grande fréquence des phrases longues et complexes (cf. exemples 1 et 2), et par l'emploi des tournures de langage très spécifiques, avec une forte préférence pour les formes verbales impersonnelles (passif, pronominalisation) :

1) *Dans l'ensemble, on considère que les données indiquent une efficacité et une innocuité acceptables de Multaq dans le traitement de patients ayant des antécédents ou présentant un épisode de fibrillation auriculaire, dans le but de réduire leur risque d'hospitalisation pour une affection cardiovasculaire due à une fibrillation auriculaire, à condition qu'il soit utilisé conformément aux conditions mentionnées dans la monographie de produit.*

2) *La présente lettre a pour but de vous faire part de renseignements importants en matière de sécurité d'emploi concernant la possibilité de perte de l'électrothérapie, pour cause de dégradation de la composante hermétique d'étanchéité, de deux sous-ensembles de stimulateurs cardiaques PULSAR MAX, PULSAR, DISCOVERY, MERIDIAN, PULSAR MAX II, DISCOVERY II, VIRTUS Plus II, INTELIS II et CONTAK TR.*

Le tableau 2.1 indique que le corpus des experts est en tête de liste avec une longueur moyenne des phrases égale à 28,39 mots. Dans la littérature, les corpus spécialisés, et les corpus médicaux en particulier, sont bien connus pour la longueur et la complexité de leurs phrases, ainsi que pour leur tendance à contenir des tournures de langage relevant de la discipline dont ils traitent. Ce phénomène est d'autant plus important que lors de la dernière campagne d'évaluation des analyseurs syntaxiques du français (PASSAGE, 2007), les différents analyseurs syntaxiques en compétition étaient également testés par rapport à leurs performances sur les textes de spécialité, en l'occurrence les textes médicaux et les textes littéraires (Laurent *et al.*, 2009). Les résultats de cette tâche ont démontré que la longueur et la complexité des phrases des textes spécialisés engendrent des erreurs d'annotation chez la plupart des analyseurs, y compris les meilleurs (Paroubek *et al.*, 2007).

En ce qui concerne l'emploi du passif (avec omission de l'agent), différents travaux de recherche le décrivent comme une technique de rédaction très sollicitée dans les écrits scientifiques, et qui a pour effet de cacher ou d'écarter volontairement le/les auteur(s), ôtant ainsi le caractère subjectif de l'énoncé (Heslot, 1983 ; Candel, 1984 ; Mortureux, 1991 ; Fleischman, 2003 ; Pecman, 2004) :

3) *L'amiodarone est également indiquée en cas de tachycardie jonctionnelle, après avoir éliminé une cause médicamenteuse.*

4) *26% des hospitalisations s'observent chez les personnes de plus de 15 ans.*

5) *En cas de patient polyimmunisé, on choisira un produit d'un donneur le plus proche du HLA du patient et on pratiquera un cross-match.*

Dans notre corpus, le passif est fréquemment utilisé et très souvent marqué par l'absence de l'agent de l'action (cf. exemple 3). Paralèllement, on observe la récurrence d'autres formes impersonnelles telles que la forme pronominale (cf. exemple 4), et l'emploi du pronom indéfini *on* (cf. exemple 5), qui est régulièrement utilisé dans le corpus des experts. Ces procédés font aussi partie des caractéristiques de la langue scientifique générale (cf. chapitre 1, section 1.1.2).

Dans le domaine médical, et donc dans notre corpus d'experts, cette technique de rédaction pourrait avoir pour fonction d'exprimer des savoir-faire qui relèvent de ce que nous pouvons appeler la *norme*, c.-à-d. des connaissances, traitements, pratiques, procédures standardisées et donc connues et partagées par toute la communauté médicale. En procédant ainsi, l'accent est davantage mis sur la procédure décrite au détriment de la personne qui l'applique.

2.1.3.2 Corpus des étudiants

Le corpus des étudiants (C_2) est un ensemble de textes écrits par des experts en médecine pour des étudiants en médecine. Ce corpus de notre continuum de textes pourrait rentrer dans la catégorie *Medium* de Pearson (1998), qui regroupe des textes ayant un niveau de spécialisation

moyen, ou encore des *discours à vocation didactique* (Jacobi, 1993). En effet, le corpus des étudiants est principalement constitué de supports didactiques, parmi lesquels des cours préparés par des experts pour un public d'étudiants en médecine, y compris des épreuves classantes nationales. La frontière entre les textes de ce corpus et ceux du corpus des experts n'est pas étanche. Nous avons observé plusieurs similitudes entre les deux types de textes : l'abondance des termes médicaux, la variation terminologique, la fréquence des termes complexes, l'emploi du passif, et la présence de phrases relativement longues qui se traduit par la deuxième position qu'occupe le corpus des étudiants dans le classement du tableau 2.1, avec une longueur moyenne des phrases égale à 24,78. Ce rapprochement pourrait avoir une explication logique si l'on considère les étudiants comme des experts en devenir, en phase d'initiation à une profession donnée. Une telle initiation ne saurait se faire sans l'intervention, à un moment de la formation, du langage spécialisé de la discipline enseignée. Dans le cas présent, on pourrait expliquer la similitude observée en supposant que les textes contenus dans notre corpus des étudiants relèvent du discours ou langage standardisé des professionnels de la santé qui est transmis aux étudiants à travers les cours qui leur sont dispensés.

2.1.3.3 Corpus des patients

Le corpus des patients (C_3) regroupe des textes écrits par des experts en médecine pour les patients et le grand public. Les textes de ce corpus pourraient être associés à la catégorie *Low* de Pearson (1998) et renvoient au type d'écrit que Jacobi (1993) catégorise comme discours *d'éducation scientifique non formelle*. Il s'agit de documents spécialement conçus pour les patients et ayant pour but d'améliorer leur savoir en matière de connaissances médicales. Plus précisement, ces textes visent à éduquer les patients par rapport aux maladies dont ils souffriraient, et aux procédures médicales applicables ou en cours d'exécution dans le processus de soin. Ainsi, l'on y trouve des extraits de revues, presses, et brochures informatives.

Les travaux de recherche font référence à ce type de textes en tant que *textes de vulgarisation scientifique* (Mortureux, 1988) ou encore *textes de diffusion* (Jacobi, 1993). Puisque leur objectif est de communiquer des informations de type scientifique à un public de non-spécialistes, les textes de diffusion renferment très souvent diverses techniques ayant pour but de faciliter l'accès au sens : des exemples, des reformulations (Jacobi, 1989, 1994), ainsi que des définitions sommaires et rapides des termes inconnus (Mortureux, 1985). Certains travaux combinent ces deux techniques afin d'obtenir un meilleur rendement (Mortureux, 1985). Dans une étude récente qui analyse les stratégies de communication mises en place par les auteurs des brochures anglaises et italiennes destinées aux patients, Maglie (2015) décrit les nouvelles techniques rhétoriques et structurales qu'utilisent les auteurs vulgarisateurs pour faciliter la lecture de leurs écrits par des publics de non-spécialistes.

Outre ces différents procédés présents dans notre corpus pour patients, on observe également

que les phrases sont de taille relativement courte, comparées à celles des autres corpus. En effet, la longueur moyenne des phrases du corpus des patients est de 20,62. Elle correspond d'ailleurs à la plus petite valeur enregistrée, d'après les données du tableau 2.1. Cette remarque pourrait elle aussi traduire le souci des auteurs des textes du corpus C_3 d'adapter la structure de leur langage au public cible, qui est a priori composé de personnes n'ayant pas de grandes connaissances en médecine. Ce faible niveau de connaissances pourrait engendrer des difficultés de compréhension, si les lecteurs se retrouvent confrontés à des phrases longues et complexes.

2.1.3.4 Corpus des forums

Le corpus des forums (C_4) est le nom donné au quatrième type de textes que nous exploitons pour l'extraction des données analysées dans ce travail de thèse. Ce corpus est exclusivement composé d'extraits de textes tirés des conversations entre participants sur des forums médicaux, notamment ceux qu'offre la plateforme *Doctissimo.fr*. Il s'agit du type de textes que certains chercheurs appellent *Computer-mediated discourse* (Herring & Androutsopoulos, 2015), c.-à-d. des textes résultant de l'interaction entre différents locuteurs qui se transmettent des messages à travers des plateformes d'échange, via des ordinateurs connectés à un réseau Internet.

Les forums médicaux sont des plateformes sur lesquelles les patients et autres usagers de la médecine discutent à propos de leur santé, des maladies, des médicaments, des procédures médicales et du processus de soins qu'eux ou leurs proches seraient en train de suivre. Selon certains chercheurs, ce type de réseau social est sollicité par de nombreux usagers qui souhaiteraient exprimer leurs émotions (Gauducheau, 2008). En effet, les forums représentent une grande source de données pour l'acquisition des corpus exploités dans les recherches portant sur l'analyse des sentiments (Tokuhisa *et al.*, 2008 ; Augustyn *et al.*, 2008). Toutefois, les données acquises grâce à ces plateformes ne se limitent pas aux sentiments et aux émotions. À côté de ces informations du type subjectif (émotions, sentiments, opinions, etc), les textes des forums fournissent aussi des informations conceptuelles, comme le signalent Grabar & Dumonet (2015). Les auteurs soulignent par exemple que, sur les forums médicaux, on retrouve des informations relatives aux problèmes d'ordre médical, aux médicaments, aux maladies, etc. dont souffriraient les participants. Ce type d'informations motive l'étude des textes de forum dans les travaux portant sur l'extraction d'informations par exemple (Abdaoui *et al.*, 2014). C'est également ce type de données conceptuelles qui nous intéressent dans le corpus des forums.

Contrairement aux trois autres groupes de textes qui composent notre corpus, le sous-corpus des forums contient des textes du type informel. Ce caractère informel se traduit par différentes caractéristiques :

— les fautes de grammaire, d'orthographe et les omissions (Murray, 1990 ; Herring, 1998 ; Cho, 2003 ; Balahur, 2013) ;

— la mauvaise ponctuation et l'absence de ponctuation (Murray, 1990 ; Herring, 1998 ; Cho, 2003) ;

— une forte tendance à utiliser des formes abrégées (non conventionnelles) des mots (Herring, 2003 ; Herring & Zelenkauskaite, 2008) ;

— la fréquence des phrases courtes et incomplètes (Maynor, 1994 ; Da Cunha *et al.*, 2011).

Plusieurs études ont été effectuées dans le but d'identifier les éléments qui favorisent le développement de ce style d'écriture informel dans les textes relevant de la communication médiatisée par ordinateur. Certains chercheurs démontrent qu'un pourcentage relativement faible de ces caractéristiques semble relever d'erreurs causées par l'inattention ou le manque de connaissance des règles de la langue standard (Herring, 1998). Dans la même optique, d'autres auteurs associent cette façon d'écrire (suppression des pronoms, des déterminants et auxiliaires, utilisation des abréviations, négligence des fautes, etc.) à la volonté des rédacteurs de faire des économies de temps et d'effort (Murray, 1990).

En plus des caractéristiques déjà mentionnées, on retrouve également dans le corpus des forums un certain nombre de stratégies compensatoires, utilisées par les rédacteurs pour remplacer les signaux sociaux normalement véhiculés par d'autres canaux dans l'interaction face-à-face (Marcoccia, 2000, 2004 ; Herring & Androutsopoulos, 2015). L'une de ces stratégies, et la plus connue, consiste en l'utilisation d'émoticônes (Reid, 1991 ; Raymond, 1993), pour représenter les expressions faciales ; l'utilisation des formes de ponctuation expressives pour exprimer certaines émotions (Marcoccia, 2000). Par exemple, une succession de points d'interrogation, ou d'exclamation pourrait traduire la surprise, l'insistance, l'impatience, la colère, ou d'autres sentiments.

Au vu de la description faite ci-dessus, l'on est tenté de remettre en question la fiabilité des données que fournissent les textes extraits des forums. De plus, leur structure interne (mauvaise ponctuation et absence de ponctuation) rend difficile leur manipulation par les outils du TAL. Toutefois, ces textes représentent une importante source d'informations, qui est de plus en plus exploitée dans le cadre des travaux de recherche comme le nôtre, qui s'intéressent au type de langage qu'utilisent les usagers non-experts du domaine médical, lorsqu'ils parlent des concepts médicaux. Dans le domaine du TAL, des méthodes de prétraitement (Grabar & Dumonet, 2015) des textes de forums sont implémentées ; certains analyseurs syntaxiques (comme l'analyseur *Cordial* cf. section 2.2.1 (Laurent & Nègre, 2006 ; Laurent *et al.*, 2009), qui a été créé pour la correction orthographique et grammaticale) sont conçus de façon à pouvoir analyser ce type de textes. Dans le domaine du TAL, la communauté scientifique encourage d'ailleurs les concepteurs d'analyseurs syntaxiques à davantage prendre en considération ce type de textes à caractère informel dans l'implémentation de leurs outils, étant donné que lors de la dernière campagne d'évaluation des analyseurs syntaxiques du français (*PASSAGE, 2007*), les différents outils en compétition étaient également évalués par rapport à leurs performances sur les textes

des courriels ou e-mails. Par ailleurs, des méthodes d'extraction de données (Abdaoui *et al.*, 2014) à partir de ces textes sont développées, ce qui contribue à démontrer qu'ils peuvent malgré tout être utilisés à des fins de recherche.

2.2 Ressources et outils

2.2.1 Cordial Analyseur

Dans ce travail de thèse, l'analyse syntaxique des phrases du corpus est effectuée grâce au logiciel Cordial Analyseur [7]. Cet outil a été choisi principalement parce que d'après les résultats de différentes campagnes d'évaluation (Paroubek *et al.*, 2007 ; De La Clergerie *et al.*, 2008 ; Laurent *et al.*, 2009), il est classé parmi les meilleurs analyseurs syntaxiques disponibles pour le français (cf. chapitre 5, section 5.1.1). Cordial (CORrecteur D'Imprécisions et Analyseur Lexico-sémantique) est un outil payant, conçu à l'origine pour la correction orthographique et grammaticale. C'est un analyseur en dépendances syntaxiques dont la méthode de conception est basée sur une association de règles générales et de méthodes statistiques, qui ont pour rôle d'effectuer la désambiguïsation grammaticale.

En entrée, Cordial exige des textes sous format UTF16. Après avoir traité les textes, l'outil fait l'étiquetage syntaxique des mots du corpus, ensuite il effectue l'analyse syntaxique et retourne des phrases annotées dans un format tabulé, similaire à celui du projet CoNLL (Buchholz & Marsi, 2006). Chaque représentation de l'analyse d'une phrase consiste en un ensemble de treize champs séparés par des tabulations, comme le montre la figure 2.3 [8]. Ces différents champs fournissent les informations suivantes, en allant de la gauche vers la droite : 1) numéro identificateur du mot dans la phrase, 2) *offset_begin* ou numéro identificateur du début de la chaîne de caractère, 3) *offset_end* ou numéro identificateur de fin de chaîne, 4) forme du mot, 5) lemme, 6) catégorie grammaticale, 7) propriété morpho-syntaxique, 8) syntagme, 9) fonction grammaticale, 10) numéro identifiant de proposition, 11) verbe pivot, 12) type de proposition, et 13) sens du mot : il s'agit en général des synonymes ou de la traduction en anglais du mot. L'annexe A.1 présente le jeu d'étiquettes utilisées par Cordial pour la description des fonctions syntaxiques, ainsi que leurs significations.

Les informations concernant la fonction syntaxique et le verbe pivot sont les principales données qui nous permettent d'extraire les verbes et leurs arguments. En effet, la composante syntaxique de l'analyseur Cordial favorise l'extraction automatique des constituants dépendants du noyau

7. http ://www.cordial.fr/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm

8. Le treizième colonne (*sens du mot*) n'apparaît pas dans cette figure, du fait de la petitesse de l'espace dont nous disposons.

#N	Offset_b	Offset_e	Mot	Lemme	Typegram	Codegra	Syn	Fon	Nu	Pivot	Prop.
Les deux formes n' ont présenté aucun risque hémolytique .											
1	1259142	1259145	Les	le	DETDPIG	Da-.p-d	3	T	1	présenter	Indép.
2	1259146	1259150	deux	deux	ADJNUM	Mc.p	3	T	1	présenter	Indép.
3	1259151	1259157	formes	forme	NCFP	Ncfp	3	T	1	présenter	Indép.
4	1259158	1259159	n'	ne	ADV	Rpn	5	Q	1	présenter	Indép.
5	1259160	1259163	ont	avoir	VINDP3P	Vaip3p	5	V	1	présenter	Indép.
6	1259164	1259172	présenté	présenter	VPARPMS	Vmpasm	5	V	1	présenter	Indép.
7	1259173	1259178	aucun	aucun	ADJIND	Dt-ms-	8	D	1	présenter	Indép.
8	1259179	1259185	risque	risque	NCMS	Ncms	8	D	1	présenter	Indép.
9	1259186	1259197	hémolytique	hémolytique	ADJSIG	Afpms	8	D	1	présenter	Indép.
10	1259197	1259198	.	.	PCTFORTE	Yps	-	-	-	-	-

FIG. 2.3 – Exemple d'annotation Cordial.

verbal. Tandis que la composante dépendancielle nous permet de capturer les relations de dépendance entre le verbe et ses arguments. Une évaluation des résultats de l'annotation des corpus par Cordial sera effectuée au chapitre 5 (cf. section 5.3).

2.2.2 La terminologie Snomed International

L'annotation sémantique des patrons verbaux requiert une ressource terminologique médicale. Dans le cadre de cette thèse, nous utilisons la terminologie *Snomed International*[9] : *Systematized Nomenclature of Medicine*[10] (Roger & Robboy, 1980 ; Côté, 1996). Cette ressource médicale a été choisie d'une part parce qu'elle est l'une des plus grandes terminologies médicales librement accessibles pour le français. D'autre part, le système de catégorisation des termes médicaux qu'elle offre nous permet d'associer aux arguments des verbes des informations sémantiques, qui jouent un rôle déterminant dans notre méthodologie.

La *Snomed International terminology*[11] est l'une des premières versions de la ressource terminologique qui est de nos jours connue sous le nom de SNOMED CT[12][13]. Elle a été développée en 1980 par des experts en médecine du *College of American Pathologists*. Cette nomenclature contient des termes décrivant différents concepts médicaux (maladies, procédures, médicaments, outils, etc.) auquels font face les usagers de la médecine pendant leurs échanges communicationnels avec les médecins. Les créateurs de cette ressource l'ont conçue comme

9. Dans cette thèse, elle sera également appelée *Snomed*.
10. https ://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNMI/
11. La terminologie Snomed International a été remplacée par une nouvelle nomenclature, *Systematized Nomenclature of Medicine – Reference Terminology* (SNOMED RT®) qui a ensuite été fusionnée avec les termes cliniques version 3 (CTV3), encore appelés *Read Codes*, résultant sur la création de la *SNOMED Clinical Terms* connue de nos jours sous le nom de *SNOMED CT* qui a été davantage développée au fil du temps et est disponible en plusieurs langues.
12. http ://www.snomed.org/
13. https ://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT/index.html

un moyen d'aide afin d'améliorer la communication entre le corps médical et les patients. Les concepts de la Snomed sont organisés en 12 catégories encore appelées *axes*, qui regroupent différents termes médicaux :

1. *Chemicals, Drugs and Biological Products* : produits chimiques et biologiques ;
2. *Diseases/Diagnoses* : maladies et diagnostic [14] ;
3. *Function* : fonctions de l'organisme, c.-à-d. tous les éléments physiologiques qui permettent à l'organisme de bien fonctionner
4. *General Linkage/Modifiers* : modifieurs ;
5. *List of Pharmaceutical Companies* : liste des compagnies pharmaceutiques, et organismes intervenant dans le domaine de la santé
6. *Living Organisms* : organismes vivants, c.-à-d. tous les organismes vivants autres que l'être humain : animaux, plantes, virus, bactéries, etc. ;
7. *Morphology* : éléments caractéristiques de la morphologie humaine ;
8. *Occupations* : métiers, les différents métiers qui rentrent dans le contexte médical ;
9. *Physical Agents, Forces, and Activities* : agents physiques, forces et activités, c.-à-d. les outils utilisés dans les activités médicales, les forces et activités ;
10. *Procedures* : procédures médicales, les différentes opérations et pratiques faites dans le contexte de soin ;
11. *Social Context* : contexte social, il s'agit des dénominations qui renvoient aux différents statuts sociaux ;
12. *Topography* : topographie ou anatomie, les parties, organes et cellules qui constituent le corps humain.

La catégorie 4 qui contient les modifieurs n'a pas été prise en considération dans cette étude. De même, afin de limiter la variabilité des patrons syntactico-sémantiques [15], certaines catégories ont dû être jumelées, il s'agit plus précisément de celles qui partagent certaines propriétés sémantiques. Ainsi, les catégories 7 et 12 ont été regroupées en une seule (*Topography*), car elles concernent toutes les deux l'anatomie humaine. Il en a été de même pour les catégories 5 et 8, désormais réunies sous la catégorie *Occupations*, car elles ont en commun le fait d'avoir des agents humains qui occupent une certaine fonction (métier) dans le contexte médical, comme le montre l'exemple suivant :

6) *Santé Canada a publié de nouvelles recommandations liées au traitement du cancer.*

14. Dans cet emploi, le mot *diagnostic* renvoie au résultat de l'action de diagnostiquer, et non à l'action elle-même qui correspond plutôt à la catégorie *Procedures*.

15. Ces patrons seront obtenus après l'annotation sémantique des corpus grâce aux catégories de la Snomed.

Dans la phrase ci-dessus, *Santé Canada* qui est le nom d'une compagnie représente le sujet du verbe *publier* ; toutefois, sur le plan sémantique, cette phrase a un message sous-entendu, à savoir : *les responsables de Santé Canada ont publié les recommandations [...]*, car le sujet *Santé Canada* n'a pas de propriétés agentives.

Les différents efforts de restructuration de la Snomed afin de l'adapter à cette étude ont permis d'obtenir 9 catégories sémantiques finales, portant des étiquettes que nous avons définies pour préparer la phase d'annotation automatique des corpus :

- $\mathcal{T}$: Anatomie (*coeur, phalange du pouce, vaisseau, muscle oblique externe de l'abdomen*) ;
- $\mathcal{S}$: Statuts sociaux (*mari, soeur, mère, ancien fumeur, donneur de sang, fille adoptive*) ;
- $\mathcal{P}$: Procédures (*césarienne, remplacement de cathéter, télé-expertise, mastectomie*) ;
- $\mathcal{L}$: Organismes vivants tels que les bactéries et virus (*Bacillus coagulans, Salmonella, virus de la rage*) ; les plantes (*fougère, pomme de terre*), et les animaux (*singe, chien, caméléon, cheval, chat*) ;
- $\mathcal{J}$: Métiers (*équipe du SAMU, anesthésiste, assureur, cardiologue, infirmiers diplômés*) ;
- $\mathcal{F}$: Fonctions de l'organisme telles que les protéïnes (*angiotensine, héparine éliminase, héparine lyase*) ; les paramètres du corpus (*pression artérielle, pouls, poids, hématurie, apport d'oxygène*), etc. ;
- $\mathcal{D}$: Maladies (*obésité, hypertension artérielle, cancer, paludisme, hépatite, anémie pernicieuse*) ;
- $\mathcal{C}$: Produits chimiques et biologiques (*médicament, héparine, bleu de méthylène, estolate d'érythromycine*) ;
- $\mathcal{A}$: Agents physiques, forces, activités (*cathéter, prothèse, contact avec les piquants d'une plante, accident, risque, ameublement hospitalier*).

Bien que la terminologie Snomed ait été élaborée dans un objectif communicationnel, nous utilisons ce système à des fins linguistiques. Les 9 axes ci-dessus sont considérés comme des catégories sémantiques pour l'annotation des arguments des verbes de nos corpus. Lorsqu'elles sont associées aux arguments des verbes, ces catégories sémantiques propres au domaine médical décrivent la nature de ces arguments, et nous permettent d'acquérir des patrons verbaux médicaux, à partir des textes pré-annotés syntaxiquement par Cordial.

La version originale de Snomed contient 144 267 entrées (principalement des unités nominales, et quelques adjectifs). Malgré sa grande couverture, la terminologie Snomed, tout comme les autres ressources terminologiques existantes, ne saurait couvrir tous les termes et notions du domaine médical dans son entièreté (Chute *et al.*, 1996). Pour cette raison, différentes méthodes ont été implémentées pour enrichir la terminologie Snomed à partir des données de nos corpus. Ces méthodes, ainsi que les ressources résultantes, seront décrites dans le chapitre suivant (cf. chapitre 3, section 3.2.2).

2.3 Bilan

Dans ce chapitre, il a été question pour nous de présenter notre corpus, ainsi que les outils de base nécessaires pour atteindre l'objectif que vise ce projet de thèse. La présentation des différents corpus et leurs sources nous a permis de décrire les éléments qui caractérisent chaque type de textes et de les analyser au prisme de la littérature.

Nous avons pu observer à partir de quelques critères de base que les corpus experts et étudiants partagent certaines similitudes, tandis que le corpus des patients semble constituer une passerelle entre les experts et le grand public. Quant au corpus des forums, il se démarque des trois autres de par son caractère informel et sa structure interne qui est la moins accessible automatiquement. Néanmoins, ce corpus joue un rôle tout aussi important que les autres dans cette étude. En effet, la subdivision de notre corpus en 4 parties représentant différents types de textes, a été un choix motivé par les objectifs visés dans ce travail. Chaque corpus joue un rôle bien déterminé dans la méthode qui sera appliquée : les corpus experts et étudiants, de par leur niveau de spécialisation élevé, sont considérés comme le point de départ. Ils fourniront les patrons verbaux spécialisés recherchés pour la ressource de simplification. Les corpus des patients et ceux des forums principalement, serviront de source d'extraction des patrons verbaux relevant de la langue des non-experts, qui serviront d'équivalents pour les patrons spécialisés.

Ce chapitre nous a également permis de décrire les ressources et outils utilisés dans le cadre de ce projet, et de préciser leurs rôles dans la chaîne de travail. Plusieurs raisons ont favorisé le choix de Cordial comme analyseur syntaxique pour nos corpus. Premièrement, ses performances (cf. chapitre 5, section 5.1.1.4) lors des campagnes d'évaluation *EASY* et *PASSAGE* (Paroubek *et al.*, 2007 ; De La Clergerie *et al.*, 2008 ; Laurent *et al.*, 2009) font de lui l'un des meilleurs analyseurs syntaxiques du français. D'autre part, le type d'analyse que propose Cordial est d'un grand intérêt pour cette étude. En effet, la grammaire des constituants, théorie linguistique de base qui a servi de fondation aux concepteurs du logiciel Cordial, favorise la mise en évidence des relations syntaxiques dans et entre les constituants de la phrase. De plus, cette approche est sensée favoriser l'annotation des relations de dépendance entre les verbes et leurs arguments. Les informations de ce type sont déterminantes pour notre étude, car l'une des tâches de base de notre travail est l'extraction de patrons valenciels des verbes, à partir des résultats de l'annotation syntaxique des phrases. De surcroît, le format des résultats de sortie que fournit l'analyseur Cordial est convivial pour la tâche d'extraction automatique des patrons.

Par ailleurs, un élément non négligeable qui a également motivé notre choix est que Cordial, comme son nom l'indique *CORrecteur d'imprécisions et Analyseur Lexico-sémantique*, a été entrainé pour traiter efficacement les textes de type informel comme ceux du corpus des forums. La méthode Cordial a été conçue de façon à ce que le logiciel puisse détecter les fautes d'orthographe et de grammaire, ce qui prédispose l'outil à appréhender ce genre d'erreurs, afin

de proposer malgré cela une meilleure analyse syntaxique du texte. Il est important de souligner que cette propriété ne se retrouve pas chez tous les analyeurs syntaxiques du français. Ainsi, en optant pour Cordial comme analyseur syntaxique, nous avons espoir que sa technologie de base sera bénéfique, non seulement pour le prétraitement des textes des corpus de type formel, mais surtout pour les textes du corpus des forums.

En ce qui concerne la terminologie Snomed, cette principale source d'informations sémantiques pour l'annotation de nos corpus a été choisie parce qu'elle est l'une des rares terminologies médicales existantes et accessibles gratuitement pour le français. De surcroît, le système de catégorisation des termes qu'elle offre nous permettra d'acquérir des patrons syntactico-sémantiques des verbes à partir des phrases annotées syntaxiquement par Cordial.

Chapitre 3

Méthode

Ce chapitre est consacré à la description de l'architecture de la méthode appliquée dans ce travail de thèse. La figure 3.1 met en évidence les principales étapes de cette chaîne de travail, à savoir le pré-traitement et l'annotation syntaxico-sémantique des corpus, l'extraction des patrons syntaxico-sémantiques (PSS), la validation par les experts de ces patrons verbaux, et enfin l'élaboration de notre dictionnaire de simplification alignant des PSS spécialisés avec leurs équivalents non spécialisés.

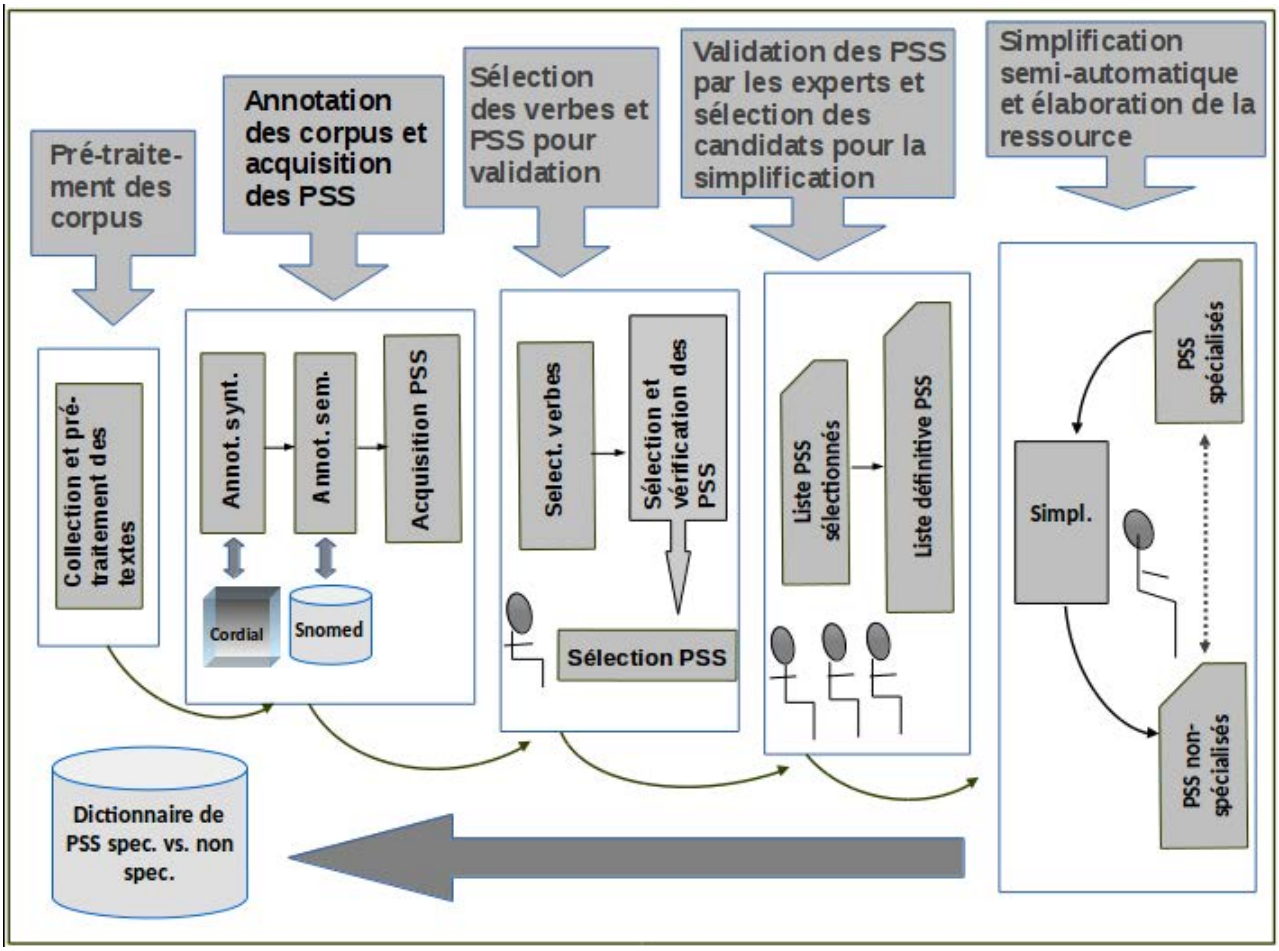


FIG. 3.1 – Schéma de la méthode.

Cette méthode semi-automatique se caractérise par une interaction fréquente entre nous et les différents automates implémentés. En effet, certaines tâches requièrent une vérification, une évaluation et/ou une validation manuelle des résultats obtenus automatiquement. Cette intervention a été particulièrement importante dans les deux dernières étapes qui sont marquées par les petits bonhommes percevables sur notre schéma récapitulatif de la méthode (cf. figure 3.1).

3.1 Collection et pré-traitement des corpus

Les quatre types de corpus présentés dans le chapitre précédent sont collectés à partir de différentes sources (chapitre 2, section 2.1.2) grâce à l'utilitaire Linux Wget [1], lancé en ligne de commande. Wget est un programme qui permet de télécharger des fichiers à partir du Web. Il exige comme attribut L'URL de la page Web souhaitée, et comme résultat, il extrait la cible du lien, c.-à-d. la page HTML telle qu'elle existe sur le Web, qu'il télécharge et enregistre localement dans un fichier.

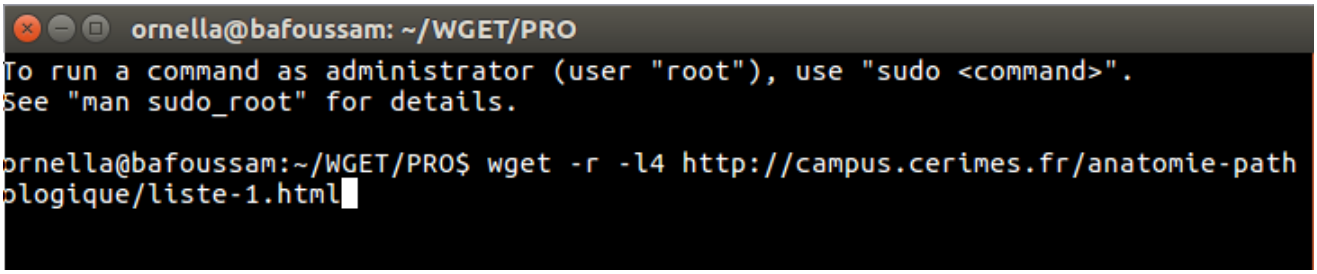


FIG. 3.2 – Exemple de ligne de commande Wget.

La figure 3.2 présente un exemple de ligne de commande lancée dans un terminal pour une requête Wget. L'option *-r* permet d'activer le téléchargement récursif des liens qui se trouveraient dans la cible de l'url de départ, tandis que *-l* permet d'indiquer la profondeur à utiliser lors d'un téléchargement récursif.

Les documents récupérés sont convertis en texte et au format UTF-8. Puis intervient une phase de nettoyage semi-automatique de ces textes, afin de faciliter leur traitement automatique. Dans cette intention, des scripts de pré-traitement sont définis et supportés par une phase de vérification manuelle. À ce stade, entre autres tâches, nous effectuons :

— le remplacement des caractères spéciaux par leurs équivalents en UTF-8 ;

— la suppression de certains caractères spéciaux pouvant créer des conflits ;

— la suppression des fragments de textes tels que les liens vers des pages Web ;

— le rétablissement de la ponctuation dans certaines phrases dont la structure a été modifiée lors des précédents traitements automatiques.

La dernière tâche de cette phase consiste en la conversion des textes au format UTF-16, afin de les rendre compatibles avec Cordial, l'outil d'analyse syntaxique que nous utilisons pour l'étape suivante.

1. https ://doc.ubuntu-fr.org/wget

3.2 Annotation des corpus et acquisition des PSS

Cette partie de notre travail de thèse est dédiée à la description des procédures qui ont été implémentées afin d'extraire et de traiter les données textuelles que fournissent les différents corpus. Ces procédures débouchent sur l'acquisition des patrons syntaxico-sémantiques des verbes, qui constituent la matière première requise pour la création de notre ressource de simplification. Nous nous focaliserons particulièrement sur les deux tâches d'annotation dont le but a été d'enrichir les textes avec les informations syntaxiques et sémantiques requises pour les prochaines étapes de cette étude. L'ensemble des méthodes ici décrites s'applique au niveau de la phrase qui est notre unité de travail. Nos analyses portent sur la phrase (et non le paragraphe ou encore le texte entier) et s'intéressent à ses constituants et sa sémantique, grâce aux méthodes semi-automatiques que nous proposons. Plus précisément, les méthodes implémentées dans ce travail de thèse se focalisent sur le verbe et ses arguments. Quant aux circonstants rattachés au verbe, ils n'interviennent que de façon accessoire dans nos analyses.

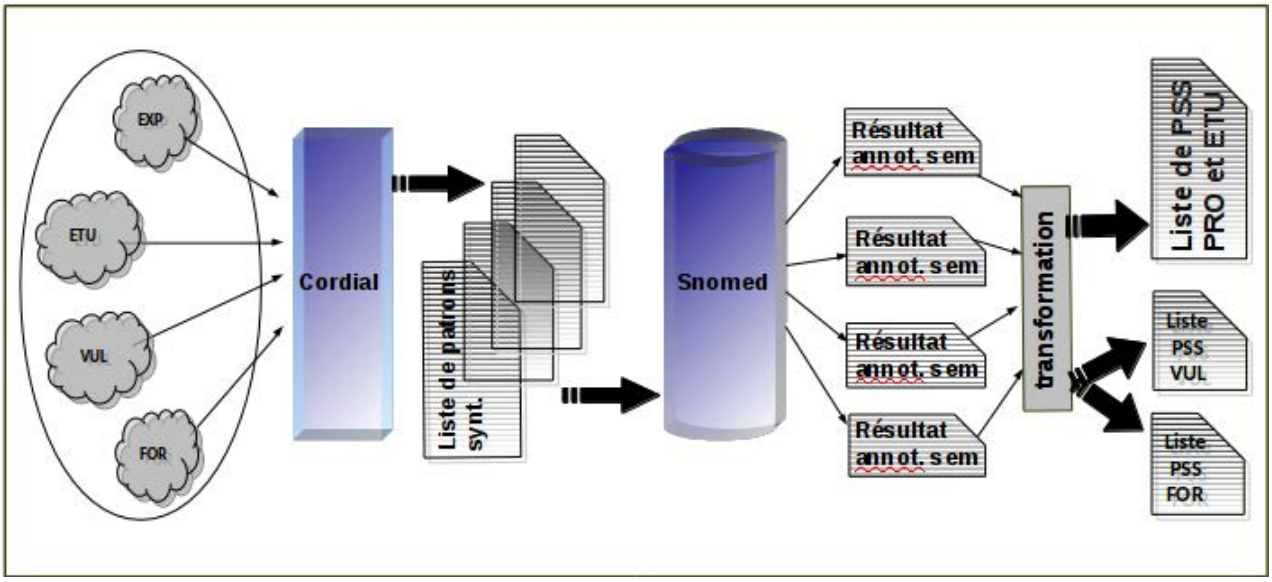


FIG. 3.3 – Processus d'annotation des corpus et acquisition des PSS.

Notre méthode d'annotation des corpus est résumée dans la figure 3.3 qui présente les principales étapes, dont l'étiquetage syntaxique et l'annotation sémantique.

3.2.1 Annotation syntaxique et extraction des patrons valenciels

3.2.1.1 Etiquetage syntaxique des corpus avec Cordial

L'annotation syntaxique des corpus a été effectuée grâce à l'analyseur syntaxique Cordial. Tel que décrit dans la section 2.2.1 du chapitre 2, cet outil propose une analyse dépendancielle des

constituants de la phrase, basée sur la syntaxe. Cette tâche permet de repérer le noyau verbal, les arguments qui en dépendent, ainsi que les circonstants qui lui sont rattachés. Ces dernières informations (verbes avec leurs arguments et les circonstants) constituent les données de base à extraire pour la suite de notre travail.

Les textes précédemment pré-traités et convertis en UTF-16 sont soumis à une analyse syntaxique par Cordial. Sous Linux, ce logiciel se présente sous la forme d'un programme Shell qui se lance en ligne de commande. L'outil opère une segmentation en phrases des textes, afin de les traiter. En effet, l'unité de travail de Cordial est la phrase [2] qu'il analyse tel qu'expliqué dans la section 2.2.1 du chapitre 2. Les résultats de l'analyse sont retournés dans des fichiers portant l'extension *.etq*. Il sont structurés comme l'indique l'image 3.4 [3] reprise ci-dessous, pour chaque phrase, le logiciel Cordial retourne la phrase elle-même, suivie de la représentation de l'analyse de cette phrase sous un format tabulé comprenant treize champs : 1) numéro identificateur du mot dans la phrase, 2) *offset_begin* ou numéro identificateur du début de la chaîne de caractère, 3) *offset_end* ou numéro identificateur de fin de chaîne, 4) forme du mot, 5) lemme, 6) catégorie grammaticale, 7) propriété morpho-syntaxique, 8) syntagme, 9) fonction grammaticale, 10) numéro identifiant de proposition, 11) verbe pivot, 12) type de proposition, et 13) sens du mot.

#N	Offset_b	Offset_er	Mot	Lemme	Typegram	Codegra	Syn	Fon	Nu	Pivot	Prop.
Les deux formes n' ont présenté aucun risque hémolytique .											
1	1259142	1259145	Les	le	DETDPIG	Da-.p-d	3	T	1	présenter	Indép.
2	1259146	1259150	deux	deux	ADJNUM	Mc.p	3	T	1	présenter	Indép.
3	1259151	1259157	formes	forme	NCFP	Ncfp	3	T	1	présenter	Indép.
4	1259158	1259159	n'	ne	ADV	Rpn	5	Q	1	présenter	Indép.
5	1259160	1259163	ont	avoir	VINDP3P	Vaip3p	5	V	1	présenter	Indép.
6	1259164	1259172	présenté	présenter	VPARPMS	Vmpasm	5	V	1	présenter	Indép.
7	1259173	1259178	aucun	aucun	ADJIND	Dt-ms-	8	D	1	présenter	Indép.
8	1259179	1259185	risque	risque	NCMS	Ncms	8	D	1	présenter	Indép.
9	1259186	1259197	hémolytique	hémolytique	ADJSIG	Afpms	8	D	1	présenter	Indép.
10	1259197	1259198	.	.	PCTFORTE	Yps	-	-	-	-	-

FIG. 3.4 – Exemple d'annotation Cordial.

Parmi ces champs, il y en a un qui identifie le/les verbe(s) pivot(s) de la phrase, c'est-à-dire les verbes têtes (selon la terminologie de Tesnière), auxquels les autres constituants de la phrase sont subordonnés. Pour chaque verbe tête, Cordial établit et matérialise des liens de dépendance avec ses subordonnés (c.-à-d. les arguments) et circonstants. Pour ce faire, il utilise une palette d'étiquettes ou codes décrits dans l'annexe A.1.

2. Ce paramètre représente un point commun entre la méthode de travail de Cordial et la nôtre.
3. Cette image a été présentée et décrite à la section 2.2.1 du chapitre 2.

Les étiquettes de Cordial jouent un rôle important dans la tâche d'extraction automatique des patrons. En effet, elles servent d'indicateurs pour le repérage des différents verbes, ainsi que leurs arguments et circonstants. En guise d'illustration, reprenons la phrase exemple donnée dans le chapitre précédent lors de la description des annotations de Cordial (cf. chapitre 2, section 2.2.1) :

1) *Les deux formes n'ont présenté aucun risque hémolityque.*

D'après la figure 3.4, le verbe *présenter* a été annoté par Cordial comme étant le pivot de la phrase 1. Il est identifié par le numéro 1 (colonne 10). Ce numéro est ensuite associé à tous les arguments (*les deux formes* et *aucun risque hémolityque*) qui se rapportent à lui, indiquant ainsi la relation de dépendance qui les lie.

3.2.1.2 Pré-traitement des résultats de Cordial

Avant de passer à l'extraction des patrons syntaxico-sémantiques des verbes, les résultats de l'annotation par Cordial subissent une phase de pré-traitement qui permet de faciliter l'extraction automatique des patrons syntaxiques des verbes :

— conversion et adaptation du format de sortie des résultats de Cordial, afin de le rendre facilement manipulable lors des extractions automatiques. Tel que cela apparaît sur l'image 3.4, pour chaque mot de la phrase, Cordial fournit diverses informations qui s'étendent sur une ligne, séparés les unes des autres par une tabulation. Pour faciliter le processus d'extraction automatique des patrons syntaxiques, il était nécessaire d'aligner les séquences de données caractéristiques des mots d'une même phrase, les unes après les autres, de façon à obtenir une seule chaîne d'informations pour chaque phrase annotée. Toutes les phrases annotées ont subi cette modification.

— restitution du verbe pivot dans les phrases où le verbe est conjugué à une forme composée (avec ou sans verbe modal) : cette tâche permet de rétablir certaines relations de dépendance (verbes-arguments) perdues pendant le processus d'annotation syntaxique.

2) *En cas de confirmation du diagnostic, l'anticoagulation sera poursuivie.*

3) *Lors de la consultation d'anesthésie, une note d'information écrite peut être remise au patient afin de renforcer l'information orale et d'en assurer la cohérence.*

Selon Cordial, la phrase de l'exemple 2 a pour verbe pivot (c.-à-d. verbe principal) *être*, ce qui est grammaticalement acceptable. Mais pour le succès de notre méthode d'analyse, l'idéal serait d'avoir le verbe *poursuivre* comme pivot. Dans un tel cas, le rôle du script de pré-traitement est de restituer *poursuivre* en tant que verbe pivot, en le substituant au verbe *être* partout où ce dernier est mentionné comme pivot. Il en est de même pour la phrase de l'exemple 3, où *pouvoir* et *remettre* sont, selon Cordial, les deux verbes pivots :

une note d'information écrite et *être* sont respectivement sujet et COD de *pouvoir*, tandis que le verbe *remettre*, pivot de la seconde proposition, a pour COI le groupe nominal *au patient*. Cette analyse est partiellement erronée : si *pouvoir* est considéré comme un semi-auxiliaire et de ce fait comme un verbe autonome, alors, s'il a un COD, ce serait la structure infinitive *être remis au patient*. *Une note d'information* serait dans ce cas le sujet syntaxique de *pouvoir* et l'objet (sémantique) de *remettre* (cf. à la voix active : X remet / peut remettre une note d'information au patient). Grâce à notre programme de prétraitement des résultats de Cordial, le résultat de l'analyse syntaxique de cette phrase est le suivant : verbe pivot : *remettre*, sujet : *une note d'information*, COI : *au patient*. Au total 18 231 phrases ont été corrigées dans l'ensemble du corpus : 5922 dans le PRO, 5346 dans le ETU, 3901 dans le VUL et 3062 dans le FOR.

— restitution des antécédents des phrases relatives. Cette tâche ne fonctionne que pour les antécédents qui sont représentés par un groupe nominal simple, comme le nom *patiente* dans la phrase *la patiente qui souffre d'un cancer [...]*.

Pour effectuer ces corrections, nous avons développé un script Perl multi-tâches, consacré exclusivement au pré-traitement des résultats de Cordial. Le programme effectue les transformations ci-dessus dans les différentes phrases concernées et ensuite retourne les phrases modifiées sous forme de chaîne de données, séparées les unes des autres par un saut de ligne.

3.2.1.3 Extraction des patrons syntaxiques des verbes

De façon basique, l'extraction automatique du patron valenciel du verbe *présenter*, à partir de la représentation syntaxique de la phrase proposée par Cordial à travers la figure 3.4 présentée supra, consisterait à extraire tous les éléments qui portent le numéro de proposition 1 dans la dixième colonne du tableau. Dans notre travail, cette tâche d'extraction automatique des patrons verbaux à partir des résultats de Cordial a été accomplie grâce à un programme Perl (*extract_patronsynt.pl*) rédigé à cet effet. Ce programme prend en entrée un fichier contenant les données phrastiques alignées résultant de l'annotation Cordial, et pour chaque phrase traitée, il renvoie le patron syntaxique correspondant. L'application du programme *extract_patronsynt.pl* à la phrase 4 fournirait un patron valenciel présenté de la façon suivante :

4) **présenter**|deux formes**_s**|aucun risque hémolityque**_cod**|Les deux formes n'ont présenté aucun risque hémolityque.

Le pipe (|) est utilisé comme séparateur de champs. Chaque argument porte une étiquette qui indique sa fonction syntaxique[4], et à la fin du patron, la phrase exemple analysée est

4. Les différentes étiquettes sont : *s* : sujet ; COD : complément d'objet direct, COI : complément d'objet indirect.

fournie. Notre programme d'extraction des patrons valenciels extrait également les circonstants des verbes. Toutefois, comme il a été souligné dès l'introduction de cette partie du travail, ces éléments n'ont pas été analysés au même titre que les arguments. Par conséquent, bien qu'ayant été extraits, les circonstants ne seront pas impliqués dans les prochaines étapes de cette étude autant que les arguments. Ainsi, dans la section suivante qui porte sur l'annotation sémantique, les analyses seront principalement focalisées sur les éléments qui sont indispensables à la réalisation du sens du verbe (c.-à-d. les arguments), et à titre accessoire sur les éléments circonstanciels qui sont eux aussi annotés sémantiquement.

3.2.2 Annotation sémantique des arguments

De façon globale, l'annotation sémantique des arguments consiste à associer automatiquement des catégories sémantiques de la Snomed aux syntagmes nominaux que contiennent les phrases illustrant les schémas valenciels tirés du corpus. Cette tâche est basée sur l'application d'un ensemble de règles visant à parcourir les syntagmes nominaux qui jouent le rôle d'arguments des verbes, à y repérer automatiquement ceux qui apparaissent dans la terminologie Snomed et à leur associer les catégories sémantiques correspondantes.

Afin de favoriser l'appariement des chaînes de caractères désignant le même concept mais ayant des formes de surface différentes, plusieurs techniques d'annotation ont été implémentées. La première est basée sur la comparaison des bigrammes et n-grammes lexicaux acquis à partir des termes du corpus et de la terminologie. Pour ce faire, notre système automatique parcourt séparément, l'un après l'autre, les termes de la Snomed et ceux du corpus. Dans chaque chaîne de caractères qui constitue un terme, l'automate supprime les mots outils (déterminants, prépositions, conjonctions, etc.) et les déterminants complexes (cf. section **??**, chapitre 4). La chaîne de caractères résultant est ensuite parcourue de la gauche vers la droite, les deux premières unités lexicales consécutives détectées sont enregistrées, formant ainsi un bigramme. De même, toute la chaîne de caractères dépourvue de mots outils est enregistrée en tant que n-gramme.

TAB. 3.1 – Exemple de n-gramme.

Terme Snomed : catégorie	**N-gramme**
Insufflation de la **trompe** de Fallope avec injection d'agent thérapeutique : P	insufflation trompe Fallope injection agent thérapeutique : P

Chaque bigramme et n-gramme formé à partir de la terminologie Snomed hérite de la catégorie sémantique du terme d'origine c.-à-d. la forme complète du terme. Les deux listes de bigrammes et n-grammes obtenues à partir des termes de la Snomed et des syntagmes nominaux du corpus sont comparées. Lorsqu'une compatibilité est détectée, la catégorie sémantique du bigramme/n-gramme concerné est associée à l'unité nominale provenant du corpus. La comparaison se fait

d'abord au niveau des n-grammes. Si elle n'est pas positive, alors l'automate procède à la comparaison des bi-grammes.

TAB. 3.2 – Exemple de bigramme.

Termes Snomed : catégories	**Bigrammes**	**Syntagmes nominaux du corpus**
Insufflation de la **trompe** de Fallope avec injection d'agent thérapeutique : P		**Insufflation** des **trompes** : P
Insufflation de la **trompe** de Fallope : P	**Insufflation trompe**	
Insufflation de la **trompe** d'Eustache : P		**Insufflation** de la **trompe** : P

Le tableau 3.2 présente un bigramme (*Insufflation trompe*) commun à plusieurs entrées de la Snomed et à des syntagmes nominaux apparaissant dans le corpus. Ce bigramme permet ainsi de mettre en relation les termes de la Snomed et les unités nominales du corpus qui se voient attribuer une catégorie sémantique grâce à la comparaison des bigrammes.

La deuxième technique mise en place consiste en la génération automatique des formes plurielles des termes simples de la Snomed, car très souvent dans les corpus, les formes plurielles des termes sont utilisées. L'automate programmé pour l'exécution de cette tâche parcourt les termes de la Snomed à la recherche des termes monolexicaux. À chaque fois qu'un candidat terme est détecté, sa terminaison est testée. Si cette chaîne de caractère finale est autre que *-aux*, *-x*, et *-s*, alors le système concatène un *-s* au terme traité et le sauvegarde dans le fichier résultat, en lui associant la catégorie du terme de départ. Cette tâche a débouché sur la mise en place d'une ressource de formes plurielles des termes Snomed.

Au-delà de la question de compatibilité entre les termes de la terminologie et ceux du corpus, les résultats de l'étiquetage sémantique dépendent également de la couverture de la ressource terminologique utilisée, dans notre cas, la terminologie Snomed International. Or comme nous l'avons vu dans la section 2.2.2 du chapitre 2, aucune ressource terminologique ne saurait couvrir tous les termes d'un domaine de spécialité. En réponse à cette question d'exhaustivité qui touche également la ressource Snomed, diverses méthodes et ressources ont été déployées, afin de favoriser l'appariement des termes de la Snomed et ceux du corpus et ainsi optimiser le processus d'annotation. Cette approche est similaire à celle appliquée par Borin *et al.* (2007a), afin de compléter la couverture de la terminologie médicale MESH [5], dans le but d'améliorer l'annotation des textes médicaux suédois, dans le cadre d'un projet portant sur la communication médecins vs. patients.

Les techniques que nous avons mises en place, et qui sont décrites ci-dessous, ont pour objectif d'enrichir la terminologie Snomed à partir des termes détectés dans les corpus :

5. http ://mesh.kib.ki.se/swemesh/

— définition de règles permettant d'attribuer une catégorie sémantique aux termes non-Snomed [6] apparaissant dans le corpus. Ces règles ont été conçues sur la base d'une propriété morphologique, qui ont permis d'identifier des termes désignant des entités médicales :

1) identification des termes désignant des maladies (catégorie D : 35 suffixes), à partir des formes suvantes : *-lepsie, -algie, -clasie, -gnosie, -gyrie, -ite, -lalie, -malacie, -manie, -nomie, -ome, -opsie*, etc.
2) identification des termes désignant des procédures médicales (catégorie P : 17 suffixes), à partir des formes suivantes : *-graphie, -ectomie, -scopie, -métrie, -plastie, -tomie, -raphie, -pexie, -stomie, -centèse*, etc.

— constitution de ressources de termes identifiés dans les corpus.

1) ressource de 205 termes mal orthographiés, conçue à partir du calcul de la distance de Levensthein (Levenshtein, 1966) de mots mal orthographiés des textes du corpus des forums. Comme nous l'avons vu dans le chapitre précédent (cf. section 2.1.3.4), ce corpus est susceptible de contenir des formes mal orthographiées des mots. D'après Balahur (2013), ce phénomène caractériserait très souvent les textes provenant des discussions de forums. La distance de Levensthein permet d'évaluer la similarité entre deux chaînes de caractères, à partir du calcul du nombre de caractères qu'il faudrait supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. Pour ce travail, nous avons pris en compte la distance jusqu'à un seuil de 3. Son application au corpus des forums a débouché sur la constitution d'une ressource de termes exploitable pour l'annotation des corpus (cf. annexe B.2).
2) ressource des formes plurielles des termes simples de la Snomed : elle contient plusieurs milliers de termes qui correspondent aux formes plurielles des entrées simples de la terminologie Snomed. Ces entrées ont été extraites de la terminologie au moyen des règles implémentées de manière à éviter l'extraction des noms dont la terminaison ne permet pas la formation du pluriel. Il s'agit des unités nominales se terminant par *-aux, -x, -s*. Le volume de données que contient cette ressource ne permet pas que son contenu soit présenté en entier dans cette thèse. Néanmoins, l'annexe B.1 fournit une cinquantaine d'exemples.

En plus de ces ressources, la chaîne d'annotation sémantique implique également quelques ressources requises pour effectuer certains tests sur les chaînes de caractères analysées :

— une liste de mots outils du français (prépositions, déterminants, pronoms, etc.), que nous avons conçue pour cette étude. Au total 146 entrées (cf. annexe C.1), qui ont servi

6. Dans cette étude, nous utilisons l'expression *termes non-Snomed* pour désigner les termes qui ne sont pas présents dans la terminologie Snomed.

pour l'extraction des bigrammes et n-grammes à partir des termes de la Snomed et des syntagmes nominaux du corpus.

  - — déterminants : *une, un, le, les,* etc.
  - — prépositions : *de, des, du, pour, avec,* etc.
  - — pronoms : *lequel, lesquels, celle-ci, ceux-là,* etc.
  - — conjonctions : *mais, et, ou, car,* etc.

— une liste de 231 déterminants complexes et d'autres syntagmes qui fonctionnent de façon similaire (cf. annexe C.2).

— la base de données Lexique.org [7] qui fournit divers types d'informations sur les mots de la langue, entre autres la fréquence de leurs lemmes dans différents types de textes de la langue générale.

L'annotation sémantique des corpus est réalisée à l'aide d'un programme que nous avons implémenté à cet effet. Ce programme nommé *annotation_sem.pl* exige plusieurs attributs en entrée : un fichier texte contenant les phrases illustrant les patrons syntaxiques ou schémas valenciels obtenus à l'étape précédente grâce au logiciel Cordial. Le programme d'annotation requiert également comme arguments la terminologie Snomed, ainsi que les ressources qui ont été décrites ci-dessus. Ce programme parcourt le fichier à traiter et analyse chacune des structures argumentales extraites du corpus. Pour chaque terme jouant le rôle d'argument du verbe, il questionne la terminologie Snomed et l'ensemble des ressources terminologiques, afin de vérifier si ce terme y est répertorié. Lorsque le terme recherché est détecté, sa catégorie sémantique est récupérée et lui est associée. Si le terme recherché n'apparaît dans aucune de nos ressources terminologiques, le programme le retourne tel quel, sans catégorie sémantique, le terme est donc considéré comme étant non-annoté. Au terme de ce processus, le programme retourne dans un fichier résultat, contenant plusieurs lignes. Chaque ligne représente un patron syntaxico-sémantique qui, à ce stade, se présente sous un format linéaire dans lequel le verbe est suivi d'une chaîne de paires argument-catégorie Snomed, puis d'une phrase exemple :

5) **présenter**|deux formes**_s**|aucun risque hémolityque**_cod**/**F** [8]/|Les deux formes n'ont présenté aucun risque hémolityque.

L'ensemble des résultats obtenus est sauvegardé dans une base de données sous un format tabulé.

Le processus d'annotation automatique des corpus a été caractérisé par plusieurs défis qu'il fallait impérativement relever afin d'améliorer la qualité des résultats. Dans nos efforts

7. http ://www.lexique.org/
8. Cette lettre représente la catégorie sémantique *fonctions de l'organisme* de la terminologie Snomed.

d'optimisation du système d'annotation sémantique, des heuristiques ont été définies afin d'attribuer, sous certaines conditions, des catégories sémantiques aux termes médicaux que nous nommons *termes inconnus*, c.-à-d. ceux qui ne sont répertoriés ni dans la Snomed, ni dans les listes de termes que nous avons conçues pour enrichir la ressource Snomed. La principale règle implémentée consiste à associer au terme inconnu la catégorie sémantique de sa tête morpho-syntaxique, si cette dernière est une entrée de la Snomed. En effet, cette règle ne s'applique que si et seulement si le terme tête concerné apparaît dans la terminologie Snomed, soit en tant qu'entrée simple, soit en tant que tête morpho-syntaxique d'une ou de plusieurs entrées complexes (termes complexes). Pour y parvenir, nous avons fait une extraction automatique de tous les termes simples Snomed ainsi que les têtes morpho-syntaxiques des entrées complexes, avec les catégories sémantiques associées.

Cette technique est basée sur la règle qui stipule que les unités lexicales portant une même tête morpho-syntaxique tendent à avoir la même nature sémantique. Prenons par exemple l'unité terminologique *implantation de la prothèse*. Supposons que ce terme apparaît dans le corpus, mais est absent de la terminologie Snomed, tandis qu'il existe dans la Snomed des variantes ou d'autres termes commençant par la tête *implantation* (*implantation de coeur atificiel ; implantation de prothèse auditive électro-magnétique ; implantation de la cornée ; implantation de prothèse cochléaire ; implantation de prothèse cochléaire, à un seul canal, etc.*). L'idée consiste à récupérer la catégorie sémantique de ces termes qui commencent par *implantation* et ensuite affecter cette catégorie au terme inconnu (*implantation de la prothèse*). Cette méthode produit des résultats de bonne qualité pour la plupart des termes, sauf exception, les termes commençant par ce que nous avons nommés les *têtes multicatégorielles*, à l'instar de la tête *manoeuvre* qui compte 3 catégories distinctes dans la Snomed : P (*manoeuvre de Heimlich, manoeuvre d'exploration et de contact corporel*), F (*manoeuvre de Barlow, manoeuvre cubitale*) et J (*manoeuvre*).

3.2.3 Traitement des têtes multicatégorielles

Avant d'entrer en détail dans le traitement des têtes multicatégorielles, il est nécessaire d'indiquer l'importance de cette tâche dans la chaîne de travail de cette étude, et de rappeler les objectifs visés. Comme il l'a été expliqué au début de cette section, le but de l'annotation sémantique est d'enrichir le corpus d'informations sémantiques provenant de la terminologie médicale Snomed. Dans cette démarche, chaque terme médical du corpus est sensé recevoir une étiquette indiquant la catégorie sémantique Snomed à laquelle il appartient. Cependant, les constats faits précédemment ont permis de découvrir qu'une poignée de termes médicaux du corpus n'est pas répertoriée dans la Snomed, ce qui exige l'implémentation d'une méthode d'annotation permettant d'associer à ces termes les catégories sémantiques correspondantes. Les techniques d'appariement mises en place à cette fin fonctionnent parfaitement bien lorsque les termes

concernés sont monosémiques c.-à-d. qu'ils ont une seule catégorie possible dans la Snomed. Toutefois, lorsque les termes [9] analysés sont polysémiques comme *manoeuvre* (c.-à-d. plusieurs catégories Snomed possibles), il se pose un problème de choix de la catégorie appropriée, d'où la nécessité d'introduire une phase de désambiguïsation des termes polysémiques. Ce phénomène s'observe précisément lorsque les termes complexes ambigus commencent par ce que nous avons nommé les *têtes multicatégorielles*.

Dans cette étude, le mot *tête* désigne un terme (mono-)lexical [10], que nous appelons encore *terme simple*, qui figure dans la Snomed en tant qu'entrée et/ou tête morpho-syntaxique d'un ou de plusieurs termes complexes [11]. En général, la tête morpho-syntaxique est aussi la tête sémantique du terme complexe c.-à-d. l'unité lexicale qui détermine la nature et la catégorie sémantique du terme. Une tête morpho-syntaxique est dite *multicatégorielle* lorsqu'elle apparaît en début d'au moins deux termes qui sont associés à différentes catégories sémantiques dans la Snomed. Le choix de cette méthode naïve a été motivé par le fait que l'implémentation d'une méthode robusque de désambiguïsation requiert davantage de connaissances du domaine, puisque l'on a affaire à des informations relevant d'une langue de spécialité. Prenons les termes *risque* et *manoeuvre*. Ce sont les têtes morpho-syntaxiques de plusieurs termes complexes dans la Snomed, qui se répartissent dans diverses catégories sémantiques : A (agent), F (fonction de l'organisme) et P (procédure). Les tableaux 3.3 et 3.4 fournissent quelques exemples de termes complexes commençant par les têtes *risque* et *manœuvre*.

TAB. 3.3 – Exemple de tête multicatégorielle : *risque*.

Tête/Cat	*Risque*
F	*risque d'hémorragie* *risque d'infarctus* *risque de non-observance*
A	*risque professionnel* *risque d'accident*

TAB. 3.4 – Exemple de tête multicatégorielle : *manœuvre*.

Tête/Cat	*Manoeuvre*
F	*manoeuvre cubitale* *manoeuvre de Barlow* *manoeuvre d'adduction*
P	*manoeuvre de Heimlich* *manoeuvre d'exploration* *manoeuvre de contact*
J	*manoeuvre*

Cette variation de catégories confère à *risque* et *manoeuvre* le statut de têtes multicatégorielles et impose l'implémentation d'une démarche spéciale pour que notre système d'annotation sémantique de corpus soit capable de désambiguïser les termes commençant par des têtes multicatégorielles, afin de leur associer les catégories sémantiques adéquates. Les exemples

9. Ce caractère polysémique s'observe plus particulièrement chez les têtes nominales des termes complexes de la Snomed.

10. Il s'agit d'un terme constitué d'un seul mot. Par exemple, le terme *dilatation* est une tête.

11. Un terme complexe est un syntagme nominal complexe c.-à-d. constitué de plusieurs mots. Exemple : *maladie d'Alzheimer* est un terme complexe qui a pour tête *maladie*.

du tableau capturent l'une des difficultés qui se présentent lorsqu'il faut faire la distinction entre les termes Snomed commençant par la même tête. Il est clair que le choix de la catégorie sémantique à associer aux termes non-Snomed commençant par une même tête multicatégorielle requiert la prise en considération de paramètres qui favoriseront la discrimination automatique de la catégorie applicable dans un contexte donné.

Avant la mise en place d'une méthode d'annotation exclusivement conçue pour les termes portant des têtes multicatégorielles, il était nécessaire d'évaluer l'ampleur de ce phénomène, afin de s'assurer de la nécessité d'entreprendre une telle démarche [12]. Pour ce faire, nous avons entrepris une évaluation automatique des têtes multicatégorielles de la Snomed.

3.2.3.1 Evaluation des têtes multicatégorielles

Cette évaluation des têtes multicatégorielles [13] consiste à compter toutes les entrées simples de la Snomed, qui jouent le rôle de têtes morpho-syntaxiques d'au moins deux termes complexes Snomed. Ce calcul prend également en compte les têtes morpho-syntaxiques de termes Snomed qui ne sont pas systématiquement enregistrées dans la terminologie en tant qu'entrées simples. À titre d'illustration, le terme *localisation* n'apparaît pas en tant qu'entrée simple dans la nomenclature Snomed, mais il est la tête morpho-syntaxique de plusieurs termes complexes Snomed qui se répartissent en deux catégories (F et P), par conséquent, il est considéré comme une tête multicatégorielle. La généralisation de cette analyse est passée par l'extraction et le décompte automatique de toutes les occurrences des termes têtes de la Snomed, en enregistrant les catégories sémantiques associées à chaque occurrence. La liste obtenue a subi une seconde analyse automatique au cours de laquelle les têtes intervenant avec différentes catégories sémantiques ont été identifées et recherchées parmi les entrées simples de la terminologie Snomed. Ainsi, pour chaque tête multicatégorielle de la Snomed, le nombre de catégories sémantiques repérées, ainsi que leurs nombres d'occurrences, sont connus.

La quantification des têtes multicatégorielles s'est avérée révélatrice et nécessaire car d'après les statistiques, on dénombre au total 1075 têtes multicatégorielles dans la terminologie Snomed, parmi lesquelles 274 (c.-à-d. 25,48%, cf. annexe B.3) sont des noms déverbaux [14] se terminant par l'un des suffixes suivants : *-ion*, *-ment*, *-age*, et *-eur*. Ce constat est intéressant, d'autant plus qu'il fournit une piste d'explication du phénomène de variation de catégories sémantiques qui caractérisent les têtes multicatégorielles. En effet, comme il a été expliqué à la section 1.3.2.2 du chapitre 1, en parlant du réseau lexical, les noms déverbaux peuvent avoir plus d'une

12. En effet, il pourrait s'agir d'un phénomène négligeable, si par exemple le nombre de têtes multicatégorielles se montrait insignifiant.

13. Pour circonscrire cette évaluation, nous avons uniquement considéré les têtes morpho-syntaxiques qui ont au moins une occurrence dans notre corpus.

14. C'est-à-dire des noms formés à partir d'une base verbale : *dilater*, *dilatation*.

interprétation. Ils peuvent renvoyer soit au procès qu'exprime la base verbale (c.-à-d. avoir un sens d'activité), soit avoir un sens résultatif, en renvoyant au résultat de l'action que dénote le verbe. En fonction de la suffixation qui est la sienne, un nom déverbal peut également désigner divers types d'actants du verbe de base : l'*agent* (*-eur*), l'*instrument* (*-oir*), le *moyen*, etc. (Villoing & Namer, 2008).

D'après la littérature, de nombreux noms liés à des verbes d'activités ont pour fonction de désigner des agents ou des instruments. Ces déverbaux ne décrivent ni des actions ni des états, et ne présentent pas les traits aspectuels de leurs bases verbales (Haas & Huyghe, 2010). Cette règle, qui s'applique plus particulièrement aux déverbaux en *-eur*, se vérifie dans la classification Snomed, où certaines têtes déverbales en *-eur* sont ambigües entre plusieurs catégories sémantiques qui désignent divers types d'entités correspondant aux sens agentif et instrumental : un métier (J : *traceur appareilleur de pierres, régulateur de trains*), un statut social (S : donneur de tissu), un agent de type instrument (A : *analyseur d'oxygène, stimulateur cardiaque, traceur de courbes*), un produit chimique (C : *stimulateur de croissance animale, régulateur de croissance des plantes*). Toutefois, au-delà de ces deux interprétations, les termes en *-eur* de la Snomed désignent dans quelques cas rares, d'autres types d'entités qui ne rentrent pas forcément dans les interprétations mentionnées ci-dessus (agentive et instrumentale). Il s'agit des classes fonction de l'organisme (F : *analyseur auditif, stimulateur utérin, régulateur du complément, récepteur de virus*), et partie du corps (T : *récepteur LTH*), dont les termes ne présentent pas a priori des propriétés agentives et/ou instrumentales (cf. chapitre 2, section 2.2.2 pour en savoir plus sur les diférentes catégories Snomed).

Quant aux têtes qui portent les trois autres suffixes (*-ion*, *-ment*, et *-age*), comme l'expliquent Haas & Huyghe (2010), leurs interprétations varient d'un terme à l'autre, selon qu'ils gardent fidèlement les propriétés aspectuelles (accomplissement, achèvement, etc (Grimshaw, 1990)) du verbe de base ou non. Certains ont un sens dynamique (activité), tandis que d'autres ont un sens plus concret et désignent l'objet résultant de l'activité que dénote le verbe de base. Certaines de ces nominalisations polysémiques gardent les deux acceptions qui se distinguent selon leurs contextes d'emploi. Les têtes déverbales d'activités se répartissent en différents types : les activités véritables, dynamiques (noms massifs), et les évènements (noms comptables) qui correspondent à la classe des noms d'événements dont parle Grimshaw (1990). Certains noms d'activités polysémiques tendent à avoir une double interprétation. Cette caractéristique fait d'eux des déverbaux spéciaux que Haas & Huyghe (2010) nomment les *N-Vact bisémiques* c.-à-d. les déverbaux d'action bisémiques. Dans la Snomed, les têtes déverbales en *-ion*, *-ment*, et *-age* rentrent dans cette catégorie. Ils oscillent principalement entre trois catégories sémantiques [15] :

15. Le fait que nous nous focalisions sur les principales catégories sémantiques n'exclut en rien la possibilité de rencontrer dans d'autres classes Snomed des termes se terminant par les suffixes ici mentionnés.

procédure (P : *interruption de grossesse, isolement d'une anse iléale, relâchement du tronc cœliaque, remplissage de dents flottantes, blocage d'un nerf intercostal*), maladie (D : *délétion d'un chromosome entier, interruption de la croissance, relâchement du diaphragme, relâchement diaphragmatique, blocage congénital*) et fonction de l'organisme (F : *délétion clonale, délétion de l'antigène, interruption de phonation, isolement sensoriel, relâchement musculaire, relâchement du sphincter, remplissage de l'estomac, blocage mental*). Très souvent, dans la terminologie Snomed, il arrive que l'une de ces interprétations l'emporte de loin sur les autres. Par exemple, *dérivation* apparaît 69 fois à la tête de différents termes dans la Snomed. Ces 69 occurrences se répartissent autour de 3 catégories sémantiques : D (1), F (3) et P (65 c.-à-d. 94,20 % d'occurrences). Il est indiscutable que l'interprétation procédurale prévaut sur les autres, vu sa fréquence dominante dans la ressource.

Les explications ci-dessus signalent que les propriétés aspectuelles des têtes déverbales sont grandement impliquées dans le phénomène de variation de catégories des termes commençant par une tête multicatégorielle [16]. Nous retenons également que parmi leurs différentes interprétations, les têtes déverbales tendent à avoir un sens de prédilection qui se démarque dans la terminologie Snomed par sa forte fréquence.

L'analyse des têtes multicatégorielles a également montré que les termes commençant par des têtes déverbales font partie des entités nominales les plus fréquentes dans le corpus. Les déverbaux *traitement* et *augmentation* font partie, entre autres, des têtes déverbales les plus fréquentes du corpus des experts, avec respectivement 261 et 105 occurrences, il s'agit de fréquences bien élevées qui signalent l'importance de ces têtes et par conséquent, celle des termes qu'elles caractérisent. D'après notre évaluation, la catégorie la plus fréquente dans les différents cas d'ambiguïté est P, celle qui regroupe les termes référant aux procédures c.-à-d. les pratiques et activités médicales. Cette catégorie caractérise exactement 347/1075 têtes multicatégorielles (soit 32,5%, environ 1/3), dont 166 cas sont des déverbaux, représentant 60,58% (sur un total de 274 déverbaux). Ces nombres signalent que la catégorie P caractérise plus de la moitié des têtes déverbales ambigües. Elle est suivie par les catégories F et D qui font elles aussi l'objet de nombreuses ambiguïtés. Ces deux catégories sont très souvent en conflit car elles cooccurrent avec différentes têtes multicatégorielles. On dénombre exactement 150 têtes déverbales ambigües entre l'interprétation F et l'interprétation D. Ce type de double interprétation fait partie des cas de polysémies dont la désambiguïsation est difficilement réalisable de façon automatique. Le tableau 3.5 fournit quelques exemples de termes illustrant ce type d'ambiguïté :

16. Cependant, de même que cette observation permet de comprendre la source des ambiguïtés, elle signale déjà les difficultés que ce phénomène pourrait causer dans le processus d'annotation automatique des corpus. L'aspect étant une propriété qui relève de l'interprétation, il n'est pas perceptible sur la forme graphique des termes (du moins pour le français) et est par conséquent difficilement prédictible.

TAB. 3.5 – Cas d'ambiguïté des termes portant les catégories D et F.

suff./cat	D	F
-ion	*délétion d'un chromosome entier*	*délétion de l'antigène*
	interruption de la croissance	*interruption de phonation*
-ment	*relâchement du diaphragme*	*relâchement du sphincter*
-age	*blocage congénital*	*blocage mental*

L'analyse des exemples que propose le tableau 3.5 permet de constater que les termes ambigus entre les catégories D et F dans la Snomed, partagent une grande similarité sur les plans morpho-syntaxique et sémantique. En effet, les patrons syntaxiques sous-jacents, en l'occurrence *N prep N* (Nom déverbal-préposition-Nom), et *NAdj* (Nom-Adjectif) sont quasiment identiques pour les deux types de termes. De plus, dans les exemples du tableau 3.5, ces patrons présentent également une grande proximité sur le plan sémantique. Les noms qui jouent le rôle de compléments des déverbaux désignent des entités qui renvoient soit à une partie du corps, soit à une fonction de l'organisme, et les adjectifs utilisés caractérisent eux aussi des propriétés physiologiques de l'organisme. Ce parallélisme traduit la proximité sémantique qui caractérise les termes des catégories D et F de la Snomed. L'analyse ci-dessus indique que la barrière entre ces deux catégories de termes Snomed n'est pas étanche. Par conséquent, la désambiguïsation des termes ayant ces deux interprétations est susceptible de requérir la prise en compte de données extra-linguistiques. Ce constat renvoie à la question de degré d'ambiguïté (entre des termes), qui constitue un critère fondamental dans les travaux de désambiguïsation du sens des mots dans le domaine biomédical. Ce paramètre permet de comprendre pourquoi la désambiguïsation de certains sens des mots est plus facile que d'autres. (Alexopoulou *et al.*, 2009) explique qu'en biomédecine, il est plus facile de distinguer entre les emplois du terme anglais «Bank», en tant que 'bâtiment' vs. 'gène', que de distinguer entre les emplois gène vs. protéine. Cet exemple contribue à montrer que ce type de désambiguïsation est compliqué. Nous avons fait un constat similaire en ce qui concerne la distinction entre les emplois F et D d'un terme médical.

L'analyse effectuée dans cette section nous a permis de mieux cerner les types d'ambiguïtés qui caractérisent les têtes multicatégorielles de la Snomed et de refléchir sur des méthodes adéquates pour leur traitement. Plus précisément, les résultats de cette évaluation nous ont permis d'envisager deux méthodes de traitement de l'ambiguïté terminologique dans la Snomed : une méthode fréquentielle axée sur la probabilité (pour les ambiguïtés du type D vs. F) (cf. section 3.2.3.2) et une méthode basée sur des heuristiques, plus adaptée pour les ambiguïtés impliquant la catégorie P (cf. section 3.2.3.2).

3.2.3.2 Désambiguïsation des termes ambigus (têtes multicatégorielles)

Dans cette section, nous implémentons différentes méthodes dans le but de désambiguïser le sens des termes ambigus afin de leur associer des catégories sémantiques correspondant au sens qu'ils ont selon leurs contextes d'apparition. Les techniques appliquées dans cette partie de la thèse nous ramènent à une tâche bien connue dans le domaine du TAL et dans la constitution des terminologies. Il s'agit de la désambiguïsation du sens des mots (désormais WSD). La désambiguïsation des sens des mots consiste à établir un lien entre l'occurrence d'un mot dans un texte et une signification spécifique, qui se distingue des autres significations que peuvent avoir ce même mot (Schuemie *et al.*, 2005). Depuis quelques années, la WSD fait partie des sujets qui dominent les recherches dans de nombreux domaines comme la biomédecine. Le défi récurrent est la croissance rapide de la littérature biomédicale, qui se manifeste par l'apparition de nouveaux termes et de leurs significations (Alexopoulou *et al.*, 2009). Cette situation qui caractérise également le domaine médical est accentuée par l'utilisation des abréviations et synonymes. Les systèmes robustes implémentés dans le domaine de la WSD sont en général basés sur des approches d'apprentissage automatique et des méthodes statistiques. Dans cette étude, nous proposons des approches axées sur des paramètres linguistiques et terminologiques.

La méthode fréquentielle

L'évaluation quantitative des têtes multicatégorielles et des termes Snomed portant ces têtes a permis d'observer la prédominance de certaines interprétations (sens dynamique par exemple (P)) qui tendent à caractériser les termes ambigus dans la Snomed. Sur la base de ce constat qui caractérise plusieurs têtes multicatégorielles (au total 274, cf. annexe B.3), nous avons mis en place une méthode de désambiguïsation des têtes déverbales que nous avons nommée *méthode fréquentielle*.

Elle repose sur un test qui permet d'évaluer la probabilité d'application d'une certaine catégorie sémantique Snomed parmi plusieurs catégories qui sont associées à une tête ambigüe. Ce test est basé sur le décompte du nombre de termes Snomed commençant par une tête ambigüe, et pour chaque catégorie sémantique associée à cette tête, le pourcentage de termes Snomed correspondant est également calculé. En effectuant ce test, on fait l'hypothèse que la catégorie sémantique qui enregistre le plus haut pourcentage (si celui-ci est ≥ 90) a la plus grande probabilité de correspondre à l'interprétation qu'a le terme dans la phrase concernée. Pour cette expérience, un pourcentage minimum de 90 a été considéré comme seuil.

Les données du tableau 3.6 permettent de mieux percevoir ce phénomène à travers l'exemple de la tête *implantation*. L'écart qui existe entre le nombre d'occurrences des catégories F et D d'une part et P d'autre part, en relation avec le terme *implantation*, pousse à penser que cette tête déverbale a une forte préférence pour le sens dynamique (qui correspond à la catégorie Snomed P). Ce constat semble indiquer que le sens procédural est l'interprétation de prédilection

de la tête *implantation*. Par conséquent, nous faisons l'hypothèse que la probabilité est grande que cette acception intervienne fréquemment dans les différents emplois de cette tête. Ainsi, pour une tête T donnée (tête multicatégorielle), qui a trois catégories Snomed possibles (A, C, D), et qui apparaît à la tête de plusieurs termes dans la Snomed, la méthode fréquentielle consiste à :

1. calculer le nombre total d'entrées Snomed ayant T comme tête.
2. calculer, pour chaque catégorie sémantique (A, C et D) que porte T, le pourcentage de termes Snomed correspondant.

Si la catégorie la plus fréquente enregistre un pourcentage $\geq$ au seuil 90 (cf. annexe B.4, tableau B.11), celle-ci domine et est donc considérée comme la catégorie par défaut du terme ambigu. Par conséquent, lors de l'annotation sémantique, elle est associée aux termes commençant par T [17], si ceux-ci ne sont pas répertoriés dans la Snomed.

Bien évidemment, il existe des cas où la catégorie dominante enregistre une fréquence dont le pourcentage est < 90. Au total, 231 têtes illustrent ce cas de figure (cf. annexe B.4, tableaux B.12, B.13 et B.14). Pour certaines têtes, il y a pas de différence notable entre la fréquence des interprétations possibles. Dans ces cas de figure, l'on a recours à des paramètres contextuels pour la désambiguïsation (cf. section 3.2.3.2), sachant que la catégorie par défaut demeure applicable en tant que dernière option pour attribuer une catégorie sémantique au terme.

La tête *implantation* peut être utilisée pour illustrer notre méthode fréquentielle d'annotation. Dans la Snomed, *implantation* apparaît en tête de 193 termes différents, qui rentrent dans trois catégories réparties comme l'indique le tableau 3.6 :

TAB. 3.6 – Exemple d'application de la méthode fréquentielle d'annotation avec la tête *implantation*.

Cat	Nb termes Snomed	%	Exemples
F	1	0.55	*implantation dans l'utérus*
D	3	1.55	*implantation tissulaire chirurgicale*
P	189	97,92	*implantation dans la peau du tronc*

Les données de la colonne *pourcentage* du tableau 3.6 montrent que la catégorie P est de loin la plus fréquemment associée aux termes lorsqu'ils commencent par la tête *implantation*. On en déduit que cette catégorie a la plus grande probabilité de correspondre au terme ambigu analysé dans le contexte concerné.

17. Sauf en cas d'exception où le contexte d'apparition du terme ambigu impose une autre catégorie sémantique (cf. section 3.2.3.2).

Le sémantisme du verbe pivot

À ce stade de notre chaîne de désambiguïsation des sens des mots, nous appliquons une technique similaire à celle de Wagner *et al.* (2009), qui consiste à analyser la cooccurrence verbe-argument en vue de la désambiguïsation. Les clarifications données à la section 3.2.3.1 sur le sémantisme des noms déverbaux ont permis de mieux comprendre et aborder la dissemblance entre des termes Snomed portant une tête morpho-syntaxique identique, mais appartenant à des catégories sémantiques distinctes. Une illustration de ce phénomène est l'ambiguïté P vs. D (ou P) qui caractérise la plupart des têtes multicatégorielles déverbales de notre corpus. C'est ce type d'ambiguïté qui oppose les termes *dilatation de l'orifice urétérovésical* et *dilatation de l'intestin*, qui portent respectivement les catégories P et D. Une analyse des propriétés aspectuelles des déverbaux en *-ion*, telle que faite à la section 3.2.3.1, permet de relever que la tête *dilatation* porte la catégorie D (maladie) lorsqu'elle a une interprétation résultative, tandis qu'elle porte la catégorie P lorsqu'elle a une interprétation d'activité. En général, lorsqu'il dénote une procédure, le nom déverbal fonctionne avec un verbe de réalisation. Par *verbe de réalisation*, nous entendons les prédicats verbaux qui ont le sens de 'faire', dans le contexte concerné. Les verbes comme *faire*, *réaliser*, *pratiquer*, *exécuter*, etc. constituent des exemples. Cette dernière propriété est déterminante car elle permet de repérer et distinguer les termes appartenant à la catégorie P (qui renvoient a priori à des actions) des autres types de termes (D et F). D'ailleurs, elle rejoint deux tests qui sont généralement effectués pour identifier les déverbaux exprimant une action dynamique (Haas & Huyghe, 2010) :

— la compatibilité des noms avec un verbe support : *procéder à une simulation*[18], *effectuer une simulation*, etc.

— l'aptitude des syntagmes verbaux formés à paraphraser les verbes d'activités correspondants : *effectuer une simulation* revient à *simuler* ; tout comme *effectuer une dilatation* revient à *dilater*.

Le premier test ci-dessus peut donc contribuer à la désambiguïsation des termes ambigus sur la base de l'analyse des verbes auxquels ils sont rattachés. Le fait que nos exemples portent uniquement sur les catégories P et D ne limite pas la méthode car le même procédé pourrait s'appliquer à d'autres couples de catégories : (P et F, P et C, P et A, etc.).

La mise en application d'un tel test exige au préalable qu'une ressource verbale soit disponible pour s'enquérir du type sémantique du verbe analysé. En effet, le seul moyen d'appliquer cette méthode automatiquement est de disposer d'une ressource de verbes de réalisation. N'étant pas au courant de l'existence d'une telle ressource, nous avons élaboré la nôtre, en nous servant principalement de la ressource DES[19] (Dictionnaires Electronique des Synonymes)

18. Exemple proposé par (Haas & Huyghe, 2010).
19. http ://www.crisco.unicaen.fr/des/synonymes/rechercher

et des données tirées des corpus. Les verbes concernés ne sont donc pas systématiquement des synonymes de *faire*, mais des prédicats ayant un sens agentif. Pour identifier les verbes candidats, un test a été effectué. Il consiste à s'assurer que dans une construction transitive directe, le verbe puisse prendre un sujet humain et le terme *procédure* en complément, ou un autre terme dénotant une procédure : *exécuter une procédure*, *procéder à une procédure*, etc. Ce travail a débouché sur une ressource contenant environ 160 entrées verbales, ayant chacune une signification relativement proche de 'faire' : *appliquer*, *pratiquer*, *envisager* et *entreprendre* font partie de ces entrées (cf. annexe C.3).

Pour chaque terme ambigu, la méthode de désambiguïsation consiste à :

1. identifier le type de phrase dont il s'agit (phrase active) ;
2. s'assurer que le terme ambigu est bien un COD ;
3. questionner la ressource verbale afin de vérifier si le verbe auquel se rapporte ce terme ambigu y figure, auquel cas, le terme ambigu se voit attribuer la catégorie P.

Cette méthode de désambiguïsation des termes s'est très vite montrée faible pour deux raisons. Premièrement, son application est limitée aux phrases à la forme active et aux termes ambigus jouant le rôle de COD. Or très souvent, dans les corpus des experts et des étudiants en particulier, les termes ambigus interviennent en tant que sujets dans des phrases passives. Dans certains contextes, les termes ambigus peuvent être sujets de phrases à la forme active comme dans l'exemple 6. La seconde faiblesse de ce test réside dans le fait qu'il ne prend pas considération le contexte d'apparition du terme ambigu, et pourtant il s'agit d'une condition essentielle pour la désambiguïsation du sens des mots.

6) *Des réactions d'ossification perpendiculaires à la base interne réalisent l'aspect en « poil de brosse » (complications qui survient chez l'enfant peu transfusé).*

Ainsi, face à une phrase comme celle de l'exemple 6, ce test n'est pas susceptible de fonctionner. Or dans cet emploi, la tête *réaction* qui a trois catégories possibles dans la Snomed (P, D et F) requiert une désambiguïsation.

La principale faiblesse de la méthode de désambiguïsation des termes ici présentée est donc le fait qu'elle est uniquement axée sur le sémantisme du verbe et ne tient pas compte du contexte syntaxico-sémantique qui accueille le terme ambigu.

Le contexte syntaxico-sémantique du terme ambigu

Les réflexions sur d'autres techniques possibles de désambiguïsation des termes ambigus ont débouché sur une méthode basée sur la prise en considération des propriétés syntaxiques et sémantiques qui caractérisent le contexte d'apparition du terme ambigu. D'un point de vue linguistique, cette méthode de désambiguïsation rejoint les techniques de la WSD qui sont très axées sur la prise en considération de paramètres contextuels dans la distinction entre les

différents sens d'un mot. Sur le plan sémantique, cette méthode repose principalement sur la présence (parmi les catégories sémantiques du terme ambigu) de la catégorie P, qui constitue le pivot de ce processus de désambiguïsation. En effet, notre méthode est fonctionnelle pour les cas d'ambiguïté impliquant la catégorie P, qui correspond à une interprétation d'activité. Puisque cette interprétation se démarque des autres, elle permet de faire un contraste avec les autres acceptions du terme ambigu. Elle facilite ainsi le choix de la catégorie Snomed correspondant au terme ambigu, selon les éléments qui constituent le contexte. De façon sommaire, voici comment fonctionne la méthode :

(i) au départ, le terme ambigu a une catégorie Snomed par défaut. Elle correspond à celle qui enregistre la plus grande fréquence dans la terminologie Snomed, comme nous l'avons vu dans la méthode fréquentielle à travers l'exemple de la tête *implantation* (cf. 3.2.3.2). Par exemple, la tête *compression* est associée à trois catégories dans la Snomed : D (88,88% d'occurrences), F (7,47%) et P. La catégorie par défaut est donc D.

(ii) analyse du contexte d'apparition et identification des éléments qui induiraient une autre interprétation du terme ambigu : (1) fonction syntaxique du terme à désambiguïser, (2) rôle syntaxique et type sémantique des autres arguments, (3) type sémantique du verbe.

7) *Pour minimiser ces interruptions lors de la RCP, on ne vérifie plus la présence d'un pouls et immédiatement après le choc et on reprend les compressions thoraciques.*

(iii) a priori, le terme garde son interprétation par défaut sauf si, dans le contexte, figurent certains des éléments identifiés à l'étape (ii), qui orientent vers une autre interprétation. C'est ce qui s'observe à travers l'exemple 7. Puisque les conditions sont remplies, le terme *compressions thoraciques* porte la catégorie P, qui est sa seconde interprétation.

- (1) la phrase à désambiguïser est à la forme active ;
- (2) le terme à désambiguïser est un COD ;
- (3) la phrase a un sujet (*on*) qui porte la catégorie *j*, en d'autres termes, il s'agit d'un humain ;
- (4) le verbe est un verbe de réalisation ayant pour agent le sujet.

Comme on peut le constater, cette méthode d'annotation des termes ambigus est basée sur un certain nombre d'informations contextuelles qui touchent la syntaxe et la sémantique de la phrase :

- le type de phrase (active, passive) ;
- la fonction syntaxique du terme ambigu et des autres arguments du verbe (sujet, COD, COI, complément d'agent) ;

— la catégorie sémantique des autres éventuels arguments du verbe ;

— le type sémantique du verbe pivot dont le terme ambigu dépend : cette condition requiert la ressource des verbes de réalisation utilisée dans le test précédent (cf. section 3.2.3.2).

Ces quatre informations permettent de mieux cerner le contexte syntaxico-sémantique au sein duquel intervient le terme ambigu. L'analyse du contexte d'apparition du terme ambigu facilite son interprétation et permet l'induction automatique d'une catégorie sémantique. En effet, en fonction du rôle syntaxique et de la catégorie sémantique des arguments, certaines catégories Snomed sont plus susceptibles d'intervenir que d'autres. Grâce à la ressource de verbes de réalisation et aux heuristiques définies sur la base des 4 paramètres ci-dessus, le processus d'annotation sémantique des termes ambigus se déroule comme suit.

Étant en présence d'un terme ambigu entre les catégories P et F (où F est la catégorie par défaut) par exemple [20], la désambiguïsation commence par l'identification du type de phrase (active vs. passive, etc.). Dès qu'on sait de quel type de phrase il s'agit, on identifie ensuite la fonction syntaxique du terme ambigu. S'il joue par exemple le rôle de COD dans la phrase, alors il faut ensuite vérifier la catégorie sémantique du sujet de la phrase. S'il s'agit d'un sujet de type humain (J ou S), ou encore d'un sujet qui désigne un instrument (A), il faut procéder à l'interrogation de la ressource des verbes de réalisation. Si le verbe analysé y apparaît, alors la catégorie P sera associée au terme ambigu, mais si le verbe n'apparaît pas dans la ressource, alors il garde sa catégorie par défaut (dans le cas présent, F). Si par contre P est la catégorie par défaut du terme et que toutes les conditions sur les critères contextuels sont remplies, le terme gardera sa catégorie par défaut, ceci même si le verbe n'existe pas dans la ressource verbale. Nous procédons ainsi parce que nous sommes consciente du fait que notre ressource des verbes de réalisation n'est pas exhaustive. Elle ne couvre sans doute pas tous les prédicats verbaux de la langue française qui rentrent dans cette catégorie.

Un autre cas de figure est celui des phrases passives. Si le terme ambigu est le sujet d'une phrase passive, si le verbe est présent dans la ressource verbale et qu'il a accessoirement un complément d'agent ou non, alors le terme ambigu est catégorisé P, comme dans la phrase suivante : *la destruction_*P *est opérée par le cathéter d'ablation [...]*. Le terme destruction est catégorisé P et non F.

Dans une phrase pronominale réflexive, si la tête déverbale ambigüe joue le rôle de sujet dans un cas d'ambiguïté entre P et F par exemple, la catégorie P sera retenue si le verbe apparaît pas dans la ressource des verbes de réalisation : *l'ablation_*P *se fait par radiofréquence, l'abord est veineux ou artériel*.

La désambiguïsation automatique des termes dont la tête déverbale n'est pas associée à un

20. Ce test prend en considération toutes les combinaisons de catégories impliquant la catégorie P : P et D, P et A, P et T, etc.

sens d'activité (c.-à-d. à la catégorie P) n'est pas une tâche évidente ; par conséquent, elle a été principalement basée sur la fréquence. Plus précisément, le terme ambigu est associé à la catégorie qui couvre la majorité de ses apparitions dans la Snomed. Supposons par exemple, qu'au cours du processus d'annotation sémantique, l'un des termes *délétion d'un chromosome entier* (D) et *délétion de l'antigène* (F) doive être annoté. Sachant que l'induction automatique de la catégorie des termes ambigus entre les interprétations D et F est difficilement réalisable par notre système automatique, le terme ambigu se verra attribuer la catégorie la plus fréquemment employée (dans Snomed) avec la tête *délétion*.

L'application de cette méthode de désambiguïsation nous a permis d'identifier quelques lacunes qui seront présentées dans les résultats de l'annotation sémantique (cf. chapitre 4, section 3.2.3.3).

3.2.3.3 Faiblesses de la méthode de désambiguïsation des termes

Le modèle de désambiguïsation des termes proposé précédemment ne couvre pas correctement certains phénomènes linguistiques rencontrés en corpus. Il présente quelques failles qui sont causées par trois principaux types de paramètres :

— le sémantisme du verbe (verbe polysémique vs. monosémique) ;

— le type de construction dans laquelle il apparaît (Exemple : la forme pronominale appelée *se moyen*) ;

— le type d'ambiguïté qui caractérise les termes portant une tête polysémique (terme ambigu entre D et F).

Notre modèle de désambiguïsation des termes fournit des résultats de bonne qualité lorsque les verbes impliqués sont monosémiques ou plutôt, ont un faible degré de polysémie. Les verbes polysémiques repertoriés dans notre ressource favorisent également l'obtention de bons résultats. Toutefois, le mode de réalisation de leurs arguments dans le texte pourrait entraîner des difficultés d'analyse dans le procesuss de désambiguïsation automatique.

En effet, notre méthode de désambiguistion s'appuie fortement sur la présence explicite des éléments contextuels (les arguments en l'occurrence) qui caractérisent et déterminent le sens du verbe. Or certaines constructions font intervenir les arguments des verbes de façon implicite, ce qui engendre des difficultés dans la discrimination du sens des termes ambigus. Ce phénomène s'observe particulièrement avec la construction pronominale appelée *se moyen*, lorsqu'elle est marquée par l'omission de l'agent. Selon le sémantisme du verbe, elle peut donner lieu à une double interprétation.

Dans certains emplois, la structure pronominale réflexive fonctionne comme équivalent du passif avec omission de l'agent, pourtant, dans d'autres contextes, cette construction a un sens non agentif. C'est ce qu'illustrent les exemples 8 et 9, dans lesquels les termes déverbaux

ambigus jouent le rôle de sujets du verbe *réaliser*, qui oscille entre deux interprétations possibles dans ces emplois.

8) *La dilatation de l'intestin se réalise chez des sujets souffrant d'un cancer de l'intestin.*

9) *La dilatation de l'orifice urétérovésical se réalise très souvent chez les patients atteints de cette maladie.*

Dans les phrases 8 et 9, la tête *dilatation* rend les termes *dilatation de l'intestin* et *dilatation de l'orifice urétérovésical* ambigus entre les catégories P et D. Les règles de base de notre méthode de désambiguïsation déboucheraient sur l'attribution de la catégorie P (procédure) aux deux termes, puisque les éléments caractéristiques des contextes convergent vers cette interprétation.

La présence[21] du pronom réflexif « se », associée à un verbe de réalisation, signalent que le terme désigne une action ou une activité. Cette interprétation correspond à la catégorie P, par opposition à l'autre acception (D) qui s'appliquerait en l'absence du verbe de réalisation. Malheureusement, cette analyse est partiellement erronée, car elle est vraie pour la phrase 9 mais pas pour la phrase 8.

Selon la terminologie Snomed, les termes *dilatation de l'intestin* et *dilatation de l'orifice urétérovésical* appartiennent à deux classes distinctes, respectivement D et P. Cette différence peut être davantage mise en évidence par un test de synonymie. Dans la première phrase, le verbe *se réaliser* peut avoir pour synonyme *se produire*, *survenir*, tandis que dans la deuxième phrase, il signifie 'effectuer' et peut être paraphrasé comme suit : *On effectue très souvent la dilatation de l'orifice urétérovésical chez les patients atteints de cette maladie*.

De plus, si on procède à un test de passivation, la phrase 8 sera sémantiquement erronée, car le patron sémantique MALADIE *est réalisée chez* PATIENT est sémantiquement erroné. Tandis que le patron qui correspond à la phrase 9, PROCEDURE *est réalisée chez* PATIENT, fonctionne très bien : *la dilatation de l'orifice urétérovésical est très souvent réalisée chez les patients atteints de cette maladie*

Dans cette phrase, la pronominalisation réflexive relève donc de la technique d'omission de l'agent qui est très fréquente dans les textes médicaux, comme cela a déjà été expliqué dans la section 3.2.3.3. Tandis que dans la phrase 8, cette construction permet de décrire un phénomène qui se manifeste.

Par ailleurs, notre méthode de traitement des têtes multicatégorielles ne couvre que partiellement les cas d'ambiguïté entre les catégories F et D. Comme nous l'avons signalé dans la section 3.2.3.1, ces deux classes de termes sont très proches et la plupart de leurs termes ont les mêmes propriétés linguistiques, par conséquent, la désambiguïsation des termes polysémiques entre ces deux catégories ne peut pas reposer uniquement sur des informations linguistiques,

21. Lorsque la phrase n'a pas de marque explicite de l'agent introduit, par la préposition *par*.

mais requiert des connaissances encyclopédiques dont notre système automatique ne saurait disposer.

Ce constat est cohérent avec le principe fondamental de la terminologie Snomed qui a été expliqué dans la section 2.2.2 du chapitre 2, étant donné que, cette terminologie n'a pas été créée à des fins linguistiques, mais plutôt pour encoder les dossiers patients. C'est la raison pour laquelle les 12 axes de la Snomed ne regroupent pas les termes sur la base de leurs propriétés linguistiques mais plutôt selon le type d'entités médicales qu'ils dénotent. Ces entités sont indiquées par les noms des différentes catégories.

Les cas de polysémies dont la désambiguïsation requiert des données encyclopédiques nous ont poussée à entreprendre une phase de vérification manuelle des résultats de l'annotation sémantique des corpus. N'étant pas capable d'effectuer cette tâche sur l'entièreté des phrases annotées du corpus, nous nous sommes limitée à la vérification de l'annotation d'une partie des phrases dont les verbes ont été sélectionnés pour la suite de l'étude (cf. section 3.3.3).

3.2.4 Acquisition des PSS

Les chaînes d'informations obtenues suite à l'annotation sémantique des termes, illustrant les schemas valenciels des verbes (cf. section 3.2.2), sont transformées automatiquement, de façon à en tirer de véritables patrons syntaxico-sémantiques. Le programme implémenté à cet effet a pour but de changer la structure des patrons sémantiques extraits à partir des corpus. Deux transformations sont opérées l'une après l'autre et les résultats de chaque type de transformation sont sauvegardés dans des fichiers différents.

Premièrement, à partir des chaînes d'informations issues de l'annotation sémantique, on extrait des patrons sémantiques spécifiques à chaque verbe selon le contexte, c.-à-d. les structures argumentales sous-jacentes. L'exemple ci-dessous illustre la transformation opérée.

10) **accompagner**|hypertension_D/**s**/|de céphalée violent, pulsatile et rétro orbitaire_D/**coi**/|
Cette hypertension s'accompagne de céphalées violentes, pulsatiles et rétro-orbitaires.

$\rightarrow$ s_D s'accompagner de coi_D.

La transformation des PSS selon le format décrit ci-dessus permet de regrouper les phrases du corpus illustrant le même patron. Ainsi, à l'issue de cette étape, chaque PSS est associé à la liste de phrases exemples qui l'illustrent.

Dans un premier temps, les schemas valenciels des verbes tirés des corpus PRO et ETU sont transformés, débouchant sur une seule liste de PSS à partir de laquelle des candidats sont sélectionnés pour la validation (cf. section 3.3.2). En second lieu, les données des corpus VUL et FOR sont transformées séparément. Nous procédons ainsi parce que les PSS résultant des corpus PRO et ETU contiennent les candidats pour la simplification. Cette liste est ensuite triée dans le but de sélectionner les patrons à proposer aux experts pour validation, en vue de la

création de la ressource de simplification. Les PSS tirés du corpus VUL et principalement du corpus FOR constituent les données de base pour l'alignement des PSS provenant des corpus des experts. Afin de rendre leur utilisation plus optimale pour le processus d'alignement, une seconde transformation est effectuée. Cette transformation qui a pour but d'acquerir des PSS génériques s'applique à tous les PSS du corpus, y compris ceux tirés des corpus des experts. L'exemple ci-dessous donne un aperçu du résultat de la transformation opérée :

s_S (statut social) relève de coi_D (maladie)
$\rightarrow$ s_S $\sim$ *coi*_D

Les PSS présentés sous ce format ont pour avantage de regrouper des verbes (même s'ils ont des sens différents) sous une même construction générique qui dépasse le niveau lexico-sémantique. Par exemple, la construction ci-dessus peut accepter les verbes tels que *souffrir, admettre, bénéficier,* etc.

11) *Cette dame souffre d'un cancer*.
12) *Le patient est admis aux soins intensifs*.
13) *Le patient bénéficie de soins supplémentaires*.

Au terme de cette phase de travail, nous disposons de deux importants ensembles de PSS :

— les listes de PSS spécifiques à chaque verbe, elles sont tirées des différents corpus.
— la liste de PSS génériques, chaque PSS est associé à tous les verbes qui l'instancient dans les différents corpus.

Les listes PRO et ETU servent de matériaux de base pour la sélection des PSS à simplifier, tandis que les listes VUL et FOR sont gardées telles quelles pour la phase de simplification. Quant à la liste des PSS génériques, elle constitue notre source de données pour l'extraction des équivalents verbaux requis pour la simplification.

3.3 Sélection des verbes et PSS pour la validation

Le volume de notre corpus et des données qui en ressortent rend difficile la possibilité de traiter la totalité des verbes (2859 lemmes) qu'il contient dans le cadre de cette étude. De surcroît, le type d'analyse que nous effectuons, ainsi que la méthode semi-automatique de sélection des entrées de notre ressource de simplification finale ne permettent pas d'envisager un tel travail sur l'ensemble des verbes des quatre sous-corpus. De plus, ce travail de thèse ne s'applique qu'à une catégorie particulière de verbes que nous nous devons de sélectionner.

Une phase de sélection des verbes et des PSS à traiter pour la conception de la ressource de simplification a donc été nécessaire. Cette tâche de sélection est passée par quatre étapes

principales : sélection des verbes, sélection des PSS, vérification et correction des PSS, formatage des PSS. Ces étapes sont décrites et identifiées dans la figure 3.5 par des chiffres.

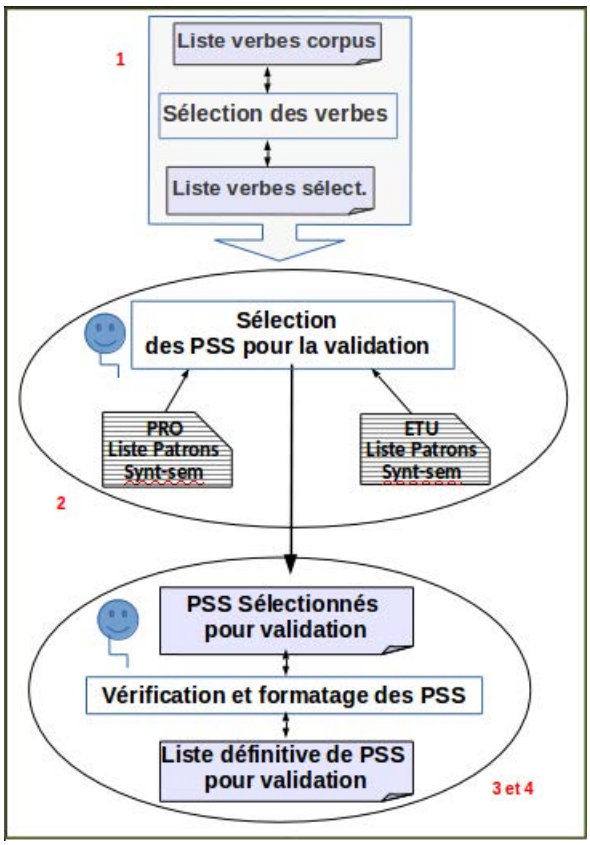


FIG. 3.5 – Étapes du processus de sélection des PSS.

3.3.1 Sélection des verbes

Dans cette phase de travail, notre objectif est de sélectionner principalement une catégorie de verbes que Lerat (2002) appelle les *verbes polysémiques*, qui correspondent à ce que Lorente (2002) appelle les *verbes phraséologiques*. D'après (L'Homme, 2012b), ce sont des verbes (*relever, associer, suivre, accompagner*, etc.) qui se combinent avec des termes du domaine et exprime, avec ces termes, des connaissances spécialisées. Autrement dit, il s'agit des verbes du lexique général mais qui ont un rôle spécifique en contexte spécialisé.

Le processus de sélection des verbes candidats pour cette étude commence par l'extraction de tous les infinitifs du corpus. Cette tâche est faite grâce au programme d'annotation sémantique qui extrait les lemmes et les fréquences (nombre d'occurrences) des verbes à partir des résultats de l'annotation syntaxique des quatre sous-corpus.

Les quatre listes de verbes obtenues sont ensuite jumelées de façon à ne former qu'une seule liste de verbes pour l'ensemble du corpus. Cette grande liste est pré-traitée afin de supprimer

les doublons et les candidats intrus qui s'y seraient infiltrés, entre autres, des participes passés (*su, connu, décrit*) et des infinitifs erronés (*subjectiver, submettre, substanter*).

La liste définitive subit ensuite un filtrage manuel qui permet de supprimer les verbes tels que :

— des verbes de mouvement et déplacement : *bouger*, *gesticuler*, *sursauter*, *marcher*, *courir*, *partir*

— des verbes de parole : *parler*, *dire*, *expliquer*, *répondre*, etc.

— des verbes de perception : *voir*, *regarder*, *apercevoir*, etc.

— des verbes d'état : *être*, *sembler*, *paraître*, etc.

— des verbes de sentiment : *aimer*, *détester*, *haïr*, etc.

— des verbes de modalité : *falloir*, *pouvoir*, *vouloir*, *devoir*, etc.

— les verbes qui n'appartiennent à aucune des catégories ci-dessus mais qui ne sont pas susceptibles d'avoir des emplois médicaux : *mentir*, *jouer*, *attendre*, *prévoir*, *abandonner*, etc.

Puisque notre étude s'intéresse particulièrement aux verbes dont le contexte d'apparition (type de mots qui les entourent) détermine le caractère spécialisé, la majorité des verbes par nature médicaux ont été éliminés (*thermoformer, anastomoser, ioder, anastomoser, anesthésier, chimiquer*, etc.). Néanmoins, quelques-uns (*disséminer, inhiber, sécréter, etc.*). ont été retenus sur la base de la variation de sens observée suite à une analyse de leurs occurrences dans les corpus.

Après avoir été pré-traitée, la liste des verbes (associés à leurs nombres d'occurrences) est triée selon deux critères dont l'application de l'un, de l'autre ou des deux, résulte en la sélection du verbe concerné. Ces critères s'appuient sur certains travaux de l'état de l'art (Tellier, 2008 ; L'Homme, 2012b) et sur les observations faites dans nos corpus.

— critère terminologique : la décision en ce qui concerne l'élimination de certains verbes a été confortée par la confrontation de notre liste de verbes à celle de Tellier (2008), qui identifie et décrit un ensemble de verbes que l'auteure considère comme des verbes médicaux, par le biais d'articles terminologiques. Ce critère a permis de retenir plusieurs verbes tels que *affecter*, *induire*, etc.

— critère linguistique : le mode de fonctionnement du verbe dans le corpus a été également pris en considération dans le choix des verbes candidats pour cette étude. Tel que suggéré par L'Homme (2012b), les contextes d'apparition des verbes ont été observés dans le but de ne retenir que des unités verbales qui sont susceptibles de changer de sens en fonction des types d'arguments qui les accompagnent. Dans cette démarche, les verbes qui tendent à intervenir dans des contextes variés sont retenus en première ligne. C'est le

cas du verbe *associer* qui est capable de se combiner aux compléments appartenant à 7 catégories Snomed : P, D, C, F, L, S, J. Cette diversité de contextes d'apparition du verbe pourrait cacher des emplois spécialisés, qui constituent le type d'informations recherchées dans ce travail de thèse.

— la fréquence cumulée ($f \geq 30$) c.-à-d. dans les quatre sous-corpus : cette condition permet de s'assurer qu'on a un nombre raisonnable de phrases à analyser pour chaque verbe sélectionné, tout en évitant que des verbes peu fréquents (dans certains types de corpus) mais spécialisés soient également repérés et pris en considération dans cette étude. Il arrive très souvent que ces prédicats aient des emplois spécialisés, c'est le cas du verbe *évaluer* qui fait partie des verbes sélectionnés pour cette étude, pourtant il a moins de 10 occurrences dans le corpus des forums, tandis que le corus des experts compte 147 occurrences.

Le critère linguistique identifie les verbes dont nous avons déjà la confirmation que le fonctionnement en corpus présente des phénomènes intéressants. Nous sommes parvenue à cette conclusion grâce aux différentes expériences effectuées dans le cadre de nos travaux de recherche [22] (Wandji Tchami *et al.*, 2013 ; Wandji Tchami & Grabar, 2014 ; Wandji Tchami *et al.*, 2014, 2015 ; Wandji Tchami, 2016). C'est par cette méthode que les verbes tels que *développer*, *observer*, *détecter* ont été retenus pour une analyse détaillée dans ce travail de thèse.

Toutefois, il n'est pas exclu qu'en plus de ces verbes, d'autres verbes du corpus interviennent lors de l'étude des collocations verbe-terme. En effet, la sélection des collocations verbe-terme analysées dans cette étude n'a pas été basée sur les verbes, mais plutôt sur la fréquence relativement élevée ($f \geq 5$) des collocations en question. Nous avons procédé ainsi parce que tous les verbes n'entrent pas systématiquement en cooccurrence régulière avec des termes particuliers dans notre corpus.

3.3.2 Sélection des PSS

Cette section porte sur la phase de sélection des PSS qui ont été soumis aux experts en médecine pour une validation qui fait office d'évaluation des PSS candidats pour la simplification. Cette tâche correspond à l'étape numéro 2 du schéma 3.5.

Les différents patrons syntaxico-sémantiques (avec fréquences) des verbes sélectionnés à l'étape précédente sont extraits à partir de la liste des patrons tirés des corpus PRO et ETU (cf. section 3.2.4). Ils subissent ensuite une sélection manuelle, qui a pour but de préparer la liste des patrons sémantiques qui seront soumis aux experts en médecine pour une validation

22. Notre mémoire de Master proposait déjà une étude comparative du fonctionnement de certains de ces verbes dans les 4 types de corpus ici étudié. Il faut cependant souligner que ces corpus ont été agrandis pour la présente étude.

définitive [23]. La liste des patrons syntaxico-sémantiques extraits pour chaque verbe est analysée manuellement, afin d'éliminer les emplois relevant de la langue générale :

14) STASOCIAL *accompagne* STASOCIAL *à* LIEU : *Maman a accompagné ma soeur à l'hôpital.*

À ce niveau de notre démarche de sélection des PSS, la fréquence n'est plus considérée comme un critère discriminant, car comme nous avons pu le constater (cf. chapitre 4, section 4.4.2), la faible fréquence d'un PSS en corpus ne signifie pas systématiquement que ce patron n'est pas spécialisé, de même que la forte fréquence d'un PSS ne signale pas forcément qu'il s'agit d'un patron spécialisé.

Les différentes combinaisons (types) de catégories Snomed apparaissant autour du verbe ont joué un rôle déterminant dans la sélection des PSS, car nous partons de l'hypothèse que le changement de catégories Snomed peut entraîner un changement de sens :

15) STASOCIAL (patient) pratique PROCEDURE : *Malheureusement, seulement 40% des femmes belges pratiquent ce dépistage.* → 'subir'

16) METIER (médecin) pratique PROCEDURE : *Dans ce cas, le chirurgien pratiquera plutôt une courte incision dans le thorax [...].* → 'faire'

Dans les travaux de recherche qui visent à proposer un modèle de simplification de textes, la réalisation de la simplification exige qu'au moins deux conditions indispensables soient remplies. Ces conditions sont résumées dans deux questions, à savoir :

1. Que faut-il simplifier ? Dans un texte, quelles phrases constituent des candidats potentiels pour la simplification ? Parmi les PSS extraits automatiquement de nos corpus, quels sont ceux qui devraient être simplifiés ?
2. Quel est le meilleur candidat possible pour la substitution ? En d'autres termes, dans le cadre d'une simplification lexicale par exemple, quel est l'équivalent le mieux adapté pour l'item simplifié ? Dans notre cas, la question serait quel verbe, quel PSS constitue un substitut compréhensible pour le PSS spécialisé selon le public cible ?

La première question [24] est celle qui est abordée dans cette partie de la thèse. La réponse à cette question correspond à la phase de sélection des phrases (et dans le cadre de ce travail des PSS) qui méritent d'être simplifiées. Dans notre étude, cette phase de sélection se fait en deux

23. Cette validation définitive par les experts en médecine permettra de retenir les PSS qui constitueront la nomenclature de notre ressource de simplification.

24. La deuxième question sera abordée dans la section 3.7 car elle sera focalisée sur la phase de simplification proprement dite.

temps : tout d'abord une phase de sélection manuelle, suivie d'une validation par les experts. La sélection manuelle qui fait l'objet de cette sous-section est effectuée par nous-mêmes, telle que décrit supra, sur la base de critères linguistique et terminologique, couplés à notre intuition linguistique. Elle débouche sur une liste de patrons verbaux qui sont ensuite validés. La phase de validation (cf. section 3.4) quant à elle est basée sur les connaissances et compétences médicales partagées par les experts en médecine qui la réalisent. Elle s'achève par le dépouillement des résultats qui permet de constituer la liste définitive des PSS candidats retenus pour la création de la ressource de simplification.

La question 1 évoque la notion de lisibilité et/ou la difficulté de lecture des textes spécialisés. Cette notion regroupe un certain nombre de paramètres permettant d'évaluer à quel point un texte est lisible et compréhensible pour un public cible. Selon certaines approches, l'évaluation du niveau de lisibilité d'un texte repose sur des caractéristiques de surface telles que la longueur des mots et des phrases, l'identification de termes clés dénotant les connaissances du domaine (Kincaid *et al.*, 1975 ; Chall & Dale, 1995 ; DuBay, 2007). Certains travaux plus récents démontrent que la prise en considération des propriétés internes du texte permet d'obtenir de meilleurs résultats que dans l'approche basée uniquement sur les paramètres de surface (Nelson *et al.*, 2012). En TAL, la plupart des approches d'évaluation du degré de lisibilité d'un texte sont basées sur des méthodes statistiques ou d'apprentissage automatique, qui permettent de classifier les textes selon leurs scores de lisibilité. Ces travaux se focalisent sur une variété de caractéristiques : les propriétés syntaxiques des textes (Callan & Eskenazi, 2007) ; les propriétés sémantiques propres à la langue de spécialité étudiée (Vor der Brück *et al.*, 2008), les propriétés lexicales (fréquence des termes du domaine) (Abrahamsson *et al.*, 2014), les propriétés morphologiques qui caractérisent la langue de spécialité concernée (François & Watrin, 2011 ; Hancke *et al.*, 2012), la cohérence et la cohésion (Graesser *et al.*, 2011), etc. La plupart des travaux que nous venons de mentionner proposent des méthodes d'évaluation du degré de lisibilité des textes, mais avec des objectifs autres que la simplification de textes. Vajjala & Meurers (2014) font partie des pionniers à proposer une approche de l'évaluation du degré de lisibilité des textes ayant pour but de détecter les phrases et textes potentiellement candidats à une opération de simplification automatique.

Notre étude s'inscrit également dans le cadre de la simplification. Toutefois, nous développons une approche différente de celle proposée par Vajjala & Meurers (2014). La particularité de notre méthode est que la sélection des patrons verbaux candidats pour la simplification n'est pas basée sur des algorithmes de l'apprentissage automatique, qui en général sont utilisés dans le but d'éviter les difficultés qu'engendre une sélection manuelle. La détection des PSS à simplifier passe par un processus semi-automatique qui repose sur une phase de validation rigoureuse effectuée par trois groupes distincts d'experts en médecine (cf. section 3.4). Cette méthode a pour avantage non seulement de proposer une solution pour la constitution de la

nomenclature de notre ressource de simplification finale, mais, elle permet d'écarter tous doutes et interrogations en ce qui concerne la validité et le caractère spécialisé des PSS sélectionnés pour la simplification. En effet, la sélection finale effectuée par les trois équipes de médecins et d'infirmiers fait office de validation des données de base de cette étude.

Les PSS retenus pour la validation par les experts proviennent des textes qui remplissent les principaux critères de sélection pris en considération dans certains travaux récents sur l'évaluation du degré de spécialisation des textes techniques. Ces critères sont : l'auteur du texte, le public cible, la structure du texte et le type d'unités lexicales utilisées (Castellví, 2002 ; Da Cunha *et al.*, 2011).

La section 2.1 du chapitre 2 a fourni d'amples informations sur le corpus utilisé dans ce travail de thèse. Cette description concourt à montrer que la structure de notre corpus est favorable à la sélection des patrons verbaux spécialisés. Le continuum qu'il représente contient des textes ayant différents degrés de spécialisation, parmi lesquels des textes du corpus des experts (textes écrits par des experts en médecine pour des experts) qui constitue l'une des extrémités du continuum, celle qui représente le plus haut niveau de spécialisation. De par le caractère spécialisé qui lui est attribué via le type d'auteur et de public qu'il vise, le corpus des experts est prédisposé à contenir des entrées potentielles pour la ressource de simplification. La forte présence dans ce corpus de termes de la Snomed confirme le niveau de spécialisation des textes, ceci ajouté aux observations faites concernant la longueur des phrases et l'emploi fréquent du passif (avec agent absent) au lieu de l'actif (cf. chapitre 2, section 2.1.3.1). Tous ces éléments rendent le corpus des experts compatible avec les critères d'évaluation du degré de spécialisation d'un texte que définit Castellví (2002). Le corpus des étudiants, décrit dans le chapitre 2 (cf. section 2.1.3.2) comme étant très similaire au corpus des experts, a également été exploité comme source pour l'extraction des PSS à simplifier. La deuxième extrémité du continuum, notamment le corpus des forums, constitue elle aussi une source de données pour la simplification. De par leur faible degré de spécialisation, les textes de ce corpus fournissent de potentiels candidats équivalents pour l'alignement des PSS provenant des corpus des experts. Dans le processus d'alignement, on aura éventuellement recours au corpus VUL pour l'extraction des PSS équivalents, lorsque aucun candidat équivalent convenable n'aura été trouvé dans le corpus des forums.

La liste des patrons syntaxico-sémantiques sélectionnés dans cette étape fera l'objet d'une validation par les experts, ce processus de validation est décrit dans la section 3.4.

3.3.3 Vérification et correction des PSS

Nous effectuons une vérification manuelle des PSS, en adéquation avec les phrases qui les illustrent, afin de s'assurer que ces patrons décrivent fidèlement les phrases sur les plans syntaxique et sémantique. Le choix méthodologique que nous avons fait, notamment celui

de ne corriger qu'une partie des phrases et patrons sémantiques des verbes sélectionnés pour la simplification, est une solution face à la contrainte temporelle qu'imposerait la correction manuelle de l'ensemble des résultats. En effet, le volume des données (environ 60 000 phrases annotées) résultant de l'annotation sémantique des corpus est tel qu'une vérification manuelle de toutes les phrases et patrons était inenvisageable dans le cadre de cette étude. La vérification manuelle des patrons porte sur deux critères :

— les étiquettes syntaxiques attribuées aux arguments des verbes ;

— les catégories sémantiques attribuées aux différents arguments.

Les patrons sémantiques présentant des erreurs liées à l'étiquetage syntaxique et/ou l'annotation sémantique sont corrigés et sauvegardés pour la phase de sélection des PSS.

3.3.4 Formatage des PSS en vue de la validation

Les patrons sémantiques sélectionnés et corrigés subissent un traitement automatique ayant pour but d'améliorer le formatage, afin de le rendre plus convivial pour les experts qui sont chargés d'effectuer la validation :

— les étiquettes sémantiques (P, D, J, etc.) qui représentent les catégories de la Snomed sont remplacées par leurs formes détaillées (PROCÉDURE, MALADIE, MÉTIER, etc).

— les fonctions syntaxiques (*s,* COD*,* COI*, etc.*) indiquant les rôles des arguments sont supprimées, car cette information est peu pertinente pour les médecins qui effectuent la validation.

— dans les phrases exemples, les termes jouant les rôles d'arguments sont mis en évidence grâce au gras.

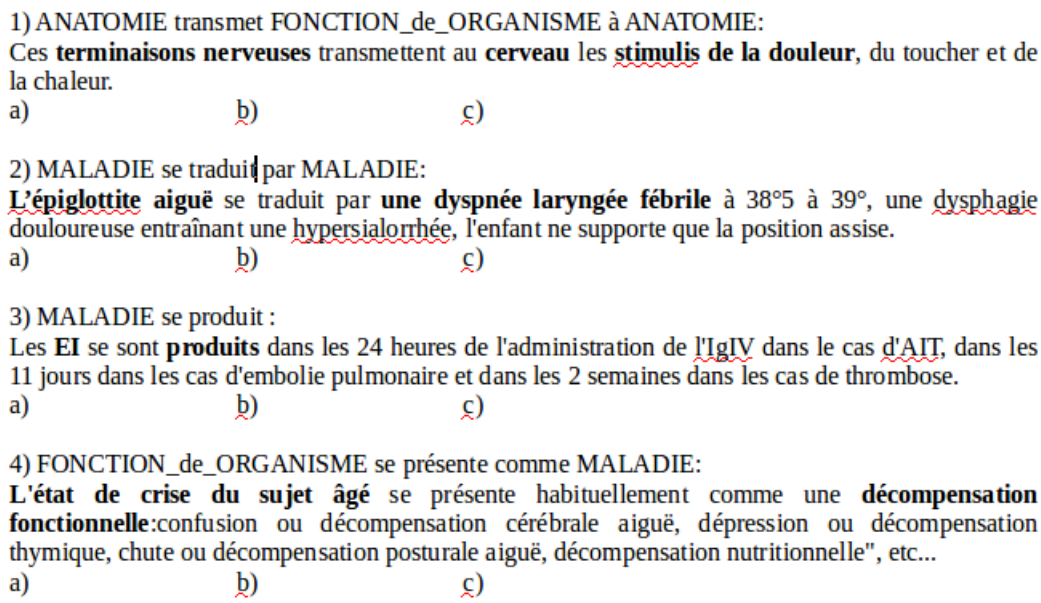
— la forme infinitive du verbe est remplacée par une forme fléchie, en accord avec la catégorie Snomed qui joue le rôle de sujet du verbe dans le PSS.

— des lettres (*a, b, c*) représentant les différentes possibilités de réponses offertes aux évaluateurs sont associées aux différents PSS.

L'application de ces règles sur le PSS *s_*D *s'accompagner de coi_*D donne le résultat suivant : MALADIE s'accompagne de MALADIE.

Le programme de formatage des PSS a également permis de changer la mise en page, afin d'obtenir des patrons syntaxico-sémantiques présentés sous un nouveau format plus adéquat pour la validation. Plus précisément, nous sommes passée du format Excel au format Word, afin que la validation puisse aisément se faire sous format imprimé, à la convenance des experts. La figure 3.6 donne un aperçu du format final des formulaires, tel qu'ils ont été remis aux experts.

Dans ces formulaires, chaque PSS est introduit par un numéro identificateur. Les positions d'arguments du verbe sont occupées par les différentes catégories Snomed des termes qui les


1) ANATOMIE transmet FONCTION_de_ORGANISME à ANATOMIE:
Ces **terminaisons nerveuses** transmettent au **cerveau** les **stimulis de la douleur**, du toucher et de la chaleur.
a) b) c)

2) MALADIE se traduit par MALADIE:
L'épiglottite aiguë se traduit par **une dyspnée laryngée fébrile** à 38°5 à 39°, une dysphagie douloureuse entraînant une hypersialorrhée, l'enfant ne supporte que la position assise.
a) b) c)

3) MALADIE se produit :
Les **EI** se sont **produits** dans les 24 heures de l'administration de l'IgIV dans le cas d'AIT, dans les 11 jours dans les cas d'embolie pulmonaire et dans les 2 semaines dans les cas de thrombose.
a) b) c)

4) FONCTION_de_ORGANISME se présente comme MALADIE:
L'état de crise du sujet âgé se présente habituellement comme une **décompensation fonctionnelle**:confusion ou décompensation cérébrale aiguë, dépression ou décompensation thymique, chute ou décompensation posturale aiguë, décompensation nutritionnelle", etc...
a) b) c)

FIG. 3.6 – Format de présentation des PSS pour la validation.

instancient. Une phrase exemple est donnée pour chaque patron. Cette phrase met en évidence les termes clés en position d'arguments. Les données des formulaires sont validées (cf. section 3.4) selon un processus d'analyse qui débouche sur le choix des PSS qui forment la nomenclature de notre ressource de simplification.

3.4 Validation des patrons syntaxico-sémantiques par les experts

3.4.1 Présentation de la tâche

Dans cette section, nous faisons une description détaillée de la procédure de validation des patrons syntaxico-sémantiques des verbes. Cette description passe par la présentation des éléments caractéristiques de cette étape de travail, à savoir :

— la raison qui a motivé la réalisation de cette tâche ;

— les objectifs visés ;

— la population interrogée ;

— les principes de la validation proprement dits.

3.4.2 Motivation et objectifs

Comme il a été souligné dans la partie introductive de cette thèse, le but de notre étude est de mettre en place une ressource pour la simplification de textes médicaux. Cette ressource

devra contenir des patrons syntaxico-sémantiques [25] spécialisés des verbes, alignés avec leurs équivalents provenant de la langue générale. La conception d'une telle ressource exige une identification préalable des PSS considérés comme spécialisés. Par souci d'objectivité, nous avons opté pour une méthode de validation manuelle de nos PSS par les personnes les mieux indiquées pour le faire, c'est-à-dire les membres du corps médical, en l'occurrence des médecins et infirmiers, puisqu'il s'agit à ce stade d'identifier des constructions qui font partie du langage spécialisé de la médecine. Ainsi, le questionnaire proposé avait pour but la validation des données extraites des corpus à partir des méthodes semi-automatiques que nous avons implémentées et décrites dans le chapitre précédent.

3.4.3 Population

La population à laquelle nous avons adressé le questionnaire est constituée de 3 groupes comportant chacun 6 experts en médecine, hommes et femmes, dont 10 infirmiers et 8 médecins, repartis tel que décrit dans le tableau 3.7.

TAB. 3.7 – Répartition des experts et des PSS.

Expert	France	Belgique	Canada	Nombre de pss
Expert 1	médecin	infirmier	infirmier	1-50
Expert 2	médecin	infirmier	infirmier	51-90
Expert 3	médecin	infirmier	infirmier	91-130
Expert 4	médecin	médecin	infirmier	131-170
Expert 5	médecin	médecin	infirmier	171-207
Expert 6	infirmier	médecin	infirmier	208-243

Comme l'on peut le constater, l'équipe canadienne est exclusivement composée d'infirmiers, tandis que les équipes française et belge sont constituées d'infirmiers et de médecins. La composition [26] des équipes d'experts a été déterminée principalement par la disponibilité des experts à qui nous avons proposé une collaboration dans le cadre de cette étude.

L'âge des personnes interrogées varie de 23 à 55 ans. Ils exercent la médecine dans l'un des trois pays francophones déjà mentionnés : la France, la Belgique et le Canada (Montréal), et appartiennent à différents domaines de spécialité qui sont : médecine générale, psychiatrie, cardiologie, pédiatrie, anesthésie, soins palliatifs, chirurgie, et réadaptation. Le choix de ces

25. Dans ce chapitre, plusieurs termes seront utilisés pour faire référence aux patrons syntaxico-sémantiques : l'abréviation PSS, ainsi que les termes *construction* et *patron*.

26. L'absence d'équilibre dans la composition des équipes n'a pas été un choix délibéré, mais plutôt une situation imposée par les circonstances de travail. Bien qu'étant consciente du fait que ce déséquilibre pourrait avoir un impact sur les résultats du questionnaire, nous restons confiante car nous croyons que les infirmiers aussi bien que les médecins, en tant que professionnels de la santé, sont des utilisateurs du langage spécialisé de la médecine, et de ce fait, ont le minimum de compétences requises pour effectuer une tâche de validation de ce type.

pays a été fortement motivé par la langue qui y est parlée. De plus, ces pays cadrent bien avec notre étude car les textes que nous analysons ont été collectés à travers des sites internet qui y sont hébergés. Les experts interrogés ont des années d'expérience situées dans un intervalle de 1 à 15 ans.

3.4.4 Description du questionnaire et du protocole de validation

L'ensemble des 6 formulaires remis aux experts comptent au total 243 patrons verbaux (cf. annexe D.3). Chaque patron est illustré par un exemple tiré du corpus. Chaque participant (Expert 1 à 6) a reçu un formulaire contenant entre 30 et 50 patrons, dont la répartition est décrite dans le tableau 3.7. Chaque formulaire est traité par trois experts provenant des trois pays impliqués dans cette étude. Autrement dit, chaque PSS a été validé par 3 experts différents. Un délai minimum d'un mois a été donné aux experts pour la validation.

Pour chaque patron sémantique figurant sur le formulaire, la validation consiste à dire si ce PSS est spécialisé ou non. Pour cela les experts ont trois options de réponses :

a) Vous et/ou vos collègues avez tendance à utiliser cette construction dans l'exercice de votre fonction.

b) Vous n'utilisez pas ou très peu cette construction mais vous la reconnaissez parce qu'elle est utilisée par des intervenants lors des conférences ; OU par les enseignants en médecine ; OU dans la littérature médicale.

c) Vous n'avez jamais entendu cette construction.

Les résultats ont été retournés sous différents formats : papier et électronique. Les réponses collectées ont été analysées, ce qui nous a permis d'obtenir les résultats que nous présenterons et analyserons dans le chapitre suivant (cf. section 4.4).

3.5 Création de la ressource pour la simplification

Cette section est consacrée à la description de la méthode semi-automatique qui a été mise en place pour la conception du dictionnaire de simplification qui constitue le principal résultat de ce travail de thèse. Cette méthode, essentiellement basée sur des critères de classification sémantiques, a pour but la recherche de verbes synonymiques pouvant substituer les verbes des PSS candidats à la simplification.

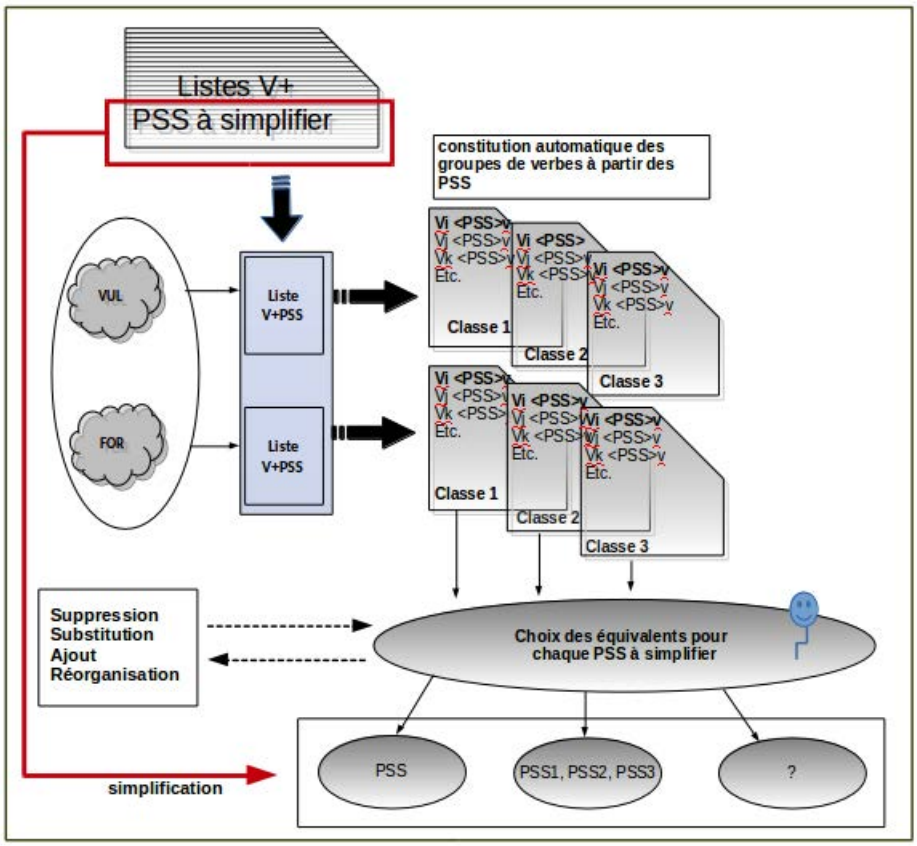


FIG. 3.7 – Étapes du processus d'alignement des PSS à simplifier.

3.5.1 Alignement des PSS avec des équivalents de la langue générale

3.5.1.1 Sélection automatique des potentiels candidats équivalents

La détection des verbes candidats pour l'alignement est réalisée par un programme implémenté à cet égard. Pour chaque patron syntaxico-sémantique retenu suite à la validation par les experts (cf. section 3.4), ce programme effectue les tâches suivantes qui sont décrites dans la figure 3.7 :

1. récupération du lemme verbal du PSS à simplifier,
2. recherche automatique de tous les lemmes verbaux qui instancient le PSS à simplifier et constitution des groupes de verbes : cette requête se fait à partir des données tirées du corpus des forums principalement et éventuellement dans le corpus $C3$ (corpus grand public). Pour chaque lemme verbal détecté, le nombre d'occurrences au sein du patron à simplifier est calculé. Ces informations numériques sont extraites grâce à un système automatique implémenté en Perl, qui s'applique sur la base de données de PSS génériques acquis à l'étape 3.2.4.

3. sélection manuelle des équivalents verbaux pour la simplification (cf. section 3.5.1.2) : à ce stade, notre compétence linguistique nous permet de trier la liste des candidats équivalents/substituts et d'éliminer les intrus, pour ne retenir que celui ou ceux qui expriment le mieux le sens véhiculé par le verbe simplifié dans le PSS traité.

Les groupes générés par notre programme à l'étape 2 sont supposés contenir des verbes qui partagent des liens sémantiques avec le verbe à simplifier. Par exemple, à l'étape 2 du processus de simplification du PSS STATUTSOCIAL *relève de* MALADIE, les verbes *avoir, faire* et *souffrir* figurent dans le groupe de verbes sélectionnés automatiquement comme potentiels candidats équivalents de *relever*. Chacun de ces lemmes intervient au sein du PSS STATUTSOCIAL_verbe_MALADIE et dans cet emploi, ils ont le sens de 'subir', 'endurer', 'éprouver quelque chose', de même que le verbe source (*relever de*). Il n'est cependant pas exclu que dans les groupes générés par notre système de sélection automatique de verbes, l'on retrouve également des formes n'ayant aucun rapprochement sémantique avec le verbe à simplifier. C'est le cas des verbes *craindre, éviter, fuir, échapper*, etc. qui apparaissent eux aussi dans le groupe de verbes retenus comme candidats pour la simplification de *relever*, au même titre que *avoir, faire*, et *souffrir*. Pourtant *craindre, éviter* et *fuir* n'ont aucun rapport de similarité sémantique avec *relever* lorsqu'ils instancient le PSS STATUTSOCIAL_verbe_MALADIE. Ils véhiculent une autre idée ('avoir peur de'), qui correspond à une autre acception que la construction sémantique impose à certains verbes qui l'instancient. Ce phénomène relève de la productivité de la construction de base qui peut être instanciée par des verbes sémantiquement hétérogènes, formant ainsi diverses sous-groupes sémantiques au sein d'un groupe de verbes intervenant dans la même construction sémantique. Ce type d'ensembles hétérogènes exige une phase de tri manuel qui correspond à l'étape numéro 3 du processus de simplification. Ce choix manuel (décrit dans la section suivante) des verbes équivalents pour l'alignement requiert non seulement des connaissances linguistiques mais également des connaissances encyclopédiques qui permettent de faire la distinction entre les différentes acceptions des verbes intervenant au sein d'un même PSS qui relève du discours médical.

3.5.1.2 Filtrage manuel des candidats équivalents

Dans cette étape qui consiste en la conception de la ressource pour la simplification, nous effectuons un tri manuel des candidats verbaux sélectionnés automatiquement (sur la base des patrons sémantiques) comme équivalents des verbes pour l'alignement des PSS. Ce tri, qui permet de choisir les meilleurs équivalents pour l'alignement, s'appuie sur les données fréquencielles des candidats équivalents dans les corpus non-experts et en grande partie sur notre compétence linguistique qui permet d'éliminer les candidats intrus (sémantiquement incompatibles) et de proposer un verbe équivalent lorsqu'aucun des candidats sélectionnés

automatiquement n'est adéquat pour l'alignement [27].

À la base, l'alignement est axé sur le remplacement ou la substitution du verbe spécialisé par un équivalent verbal compréhensible. Cet équivalent peut être un verbe, une périphrase verbale *continuer de + verbe*, une locution verbale (*faire partie de, mettre en place*), etc. l'essentiel étant que le sens du PSS de départ soit maintenu.

- Le remplacement ou substitution : STASOCIAL présente MALADIE → STASOCIAL a/fait/souffre de/manifeste MALADIE.

Dans certains cas, cette règle de base peut être accompagnée d'autres techniques de simplification, parmi lesquelles certaines, qui, comme la substitution, sont très souvent utilisées dans les travaux de la littérature portant sur la simplification de textes (Siddharthan, 2002 ; Brouwers *et al.*, 2012) :

- L'ajout : le PSS résultant a un ou plusieurs constituants de plus ; il s'agit des actants qui étaient implicitement présents dans le PSS de départ.
  PRODUIT CHIMIQUE est administré → METIER (MÉDECIN) donne PRODUIT CHIMIQUE à STASOCIAL (PATIENT)
  Ce médicament est administré en cas de thrombose veineuse.
  → *Le médecin/on donne ce médicament au patient en cas de thrombose veineuse.*
- La conversion : le verbe est remplacé par une périphrase verbale ou une autre forme verbale.
  PCHIMIQUE (MÉDICAMENT) est poursuivi → (STASOCIAL) continue de prendre PCHIMIQUE
  Si le nadolol est poursuivi jusqu'à l'accouchement, en informer l'équipe de la maternité pour lui permettre d'adapter la surveillance du nouveau-né.
  → *Si la patiente continue de prendre le nadolol jusqu'à l'accouchement, en informer l'équipe de la maternité [...].*

Comme nous l'avons déja signalé, ces techniques de simplification s'appliquent de façon à sauvegarder le sens de la construction de base. Notre méthode d'alignement des PSS est axée sur la sémantique, de ce fait, elle ne prend pas en considération les différences syntaxiques, grammaticales ou lexicales qui peuvent intervenir entre les patrons verbaux alignés. En d'autres termes, les patrons sémantiques simplifiés peuvent être associés à des patrons syntaxiquement distincts, l'essentiel étant que les verbes alignés soient sémantiquement substituables dans le contexte proposé. En effectuant l'alignement des PSS, des cas d'équivalence au delà de la syntaxe, du lexique et de la grammaire ont été rencontrés :

27. Dans ce cas de figure, le travail s'est fait en collaboration avec des médecins dont le rôle était de nous aider à bien comprendre le sens des constructions de base afin de proposer les équivalents convenables adaptés au grand public.

- Au niveau de la syntaxe :
  - STASOCIAL relève de MALADIE → STASOCIAL a/fait MALADIE ;
  - MALADIE s'accompagne de MALADIE → MALADIE entraîne MALADIE ;
- Au niveau grammatical et lexical :
  - STASOCIAL est dépisté → METIER recherche MALADIE chez STASOCIAL ;
  - PRODUIT CHIMIQUE est administré → METIER administre PRODUIT CHIMIQUE à STASOCIAL.

Le groupe des verbes qui cadrent avec le PSS à simplifier peut contenir un seul bon candidat synonyme, équivalent pour la simplification ou plusieurs. Dans ce dernier cas de figure, les verbes sont retenus comme équivalents si et seulement si ils ont le même sens que le verbe apparaissant dans le patron sémantique à simplifier. En effet, par *verbe synonyme* ou *bon candidat synonyme*, nous entendons un verbe qui a le même sens que le verbe du PSS à simplifier, au-delà des éventuelles différences grammaticales, syntaxiques, ou lexicales décrites *supra*. Si le groupe des candidats verbaux tirés du corpus des forums ne propose aucun synonyme pour le verbe du PSS en cours de simplification, alors la recherche des équivalents est effectuée sur l'ensemble des verbes automatiquement regroupés à partir des données du corpus grand public. On peut constater que notre démarche d'alignement des PSS débouche sur des phénomènes similaires à ceux qui sont généralement rencontrés dans les travaux de traduction. Pour un PSS candidat à la simplification, la recherche des équivalents débouche sur un ensemble constitué soit d'un seul élément équivalent, soit de plusieurs éléments (synonymes), soit un ensemble vide (dans le corpus des forums), auquel cas nous avons recours aux données tirées des textes de vulgarisation (corpus grand public). Si ces données ne contiennent pas non plus de bon substitut pour le verbe simplifié, alors nous en proposons un sur la base de notre compétence linguistique.

La démarche de simplification présentée ci-dessus est soutenue par des données résultant d'une approche d'alignement reposant sur une analyse manuelle du fonctionnement des verbes dans les différents corpus. En éffet, certaines expériences et travaux effectués sur les corpus, tout au long de ce projet de thèse, ont permis de mettre en place des petits [28] groupes de synonymes (experts vs. forums), à partir des tendances préférentielles qui caractérisent l'apparition des verbes dans les différents corpus. Contrairement aux groupes de verbes sémantiquement hétérogènes extraits automatiquement à l'étape précédente, les paires et/ou groupes de verbes dont nous parlons à ce stade sont exclusivement synonymiques, car leur formation résulte d'un processus d'analyse de corpus essentiellement manuel. Toutefois, il faut souligner que cette étude n'a pas porté sur tout le corpus mais uniquement sur un ensemble réduit de verbes, d'où l'impossibilité de faire reposer tout le processus de simplification des PSS sur cette approche. À l'issue de cette étude, plusieurs constats ont été faits, parmi lesquels les cas suivants :

28. Il s'agit pour la majorité de paires de verbes.

— Dans le corpus experts, le verbe *régresser* a moins de 20 occurrences dans le patron FONCTION DE L'ORGANISME *régresse*, tandis qu'il en a 28 dans le corpus étudiants. Pourtant, ce verbe est totalement absent du corpus forums (aucun patron) et n'a que 11 occurrences dans le corpus grand public. Cependant,

— Le synonyme *baisser* a plus de 100 occurrences dans le corpus des forums, dont la majorité correspond au patron FONCTION DE L'ORGANISME/MALADIE *baisse*. Dans le corpus des experts par contre, *baisser* n'a que 11 occurrences.

Ce type d'analyse nous a permis de mettre en place plusieurs groupes de verbes synonymes (experts vs. non experts) exploités pendant la réalisation de la tâche de simplification. Ces groupes de verbes synonymes constituent la principale ressource utilisée pour effectuer l'alignement des PSS.

Au terme du processus de simplification, les patrons verbaux spécialisés, qui ont été validés par les experts, sont tous alignés avec des équivalents compréhensibles, provenant des textes pour non-experts. Comme le montrent les exemples proposés, notre modèle de simplification vise principalement la sémantique des PSS, mais il passe par des modifications qui s'appliquent au niveau de syntaxe et du lexique.

3.5.2 Modélisation de la ressource de simplification

Les patrons syntaxico-sémantiques alignés à l'étape précédente sont sauvegardés dans un fichier texte, dans un format lisible par un tableur. Ce fichier comporte deux champs principaux séparés par une tabulation : un pour les entrées, un second pour leurs équivalents non spécialisés. Les éventuelles relations de synonymie entre deux ou plusieurs entrées sont également mises en évidence. En effet, pour chaque entrée simplifiée, on peut avoir un ou plusieurs équivalents extraits du corpus des forums ou du corpus grand public.

Cette ressource qui représente le principal résultat de notre travail de thèse constitue une base de données lexicales utilisable pour la simplification de textes médicaux. Cette base de données est lisible et donc exploitable, non seulement par un lecteur humain, mais aussi par une machine. Cette caractéristique fait d'elle une ressource dictionnairique potentiellement utilisable dans divers types de travaux de recherche qui s'intéressent au lexique. Par exemple, dans le cadre de la création d'un outil de simplification des textes médicaux, notre ressource verbale pourrait être d'un apport significatif, en proposant des équivalents pour les expressions verbales spécialisées. Par ailleurs, dans le domaine de la lisibilité des textes, les patrons sémantiques qu'elle propose pourraient servir de données de base pour l'implémentation de règles de détection et/ou d'évaluation de la difficulté de lecture dans un texte médical.

3.6 Comparaison des corpus : fonctionnement des collocations verbe-terme

Cette section de notre travail de thèse est dédiée à l'analyse contrastive des quatre sous-corpus, du point de vue des collocations verbe-terme. Dans cette thèse, nous abordons la notion de collocation selon une approche généralement utilisée dans les travaux d'analyse de textes axés sur des méthodes statistiques (*clustering*) : il s'agit de la fréquence. Nous désignons par *collocation verbe-terme* la cooccurrence régulière entre un verbe et un nom qui est son COD ou COI. En effet, la sélection des collocations verbe-terme analysées dans cette étude a été basée sur la fréquence (f$\geq$5). En faisant cette analyse, nous nous intéressons également aux affinités entre les verbes et les catégories Snomed des noms qu'ils prennent pour objet. Cette étude contrastive s'est avérée importante, suite à une expérience que nous avons effectuée recemment (Wandji Tchami *et al.*, 2016) et qui a signalé l'importance de la description des collocations, surtout pour les domaines comme la lexicographie et la conception des dictionnaires spécialisés. Ce point de vue se retrouve dans de nombreux travaux déjà existants sur les collocations (Heid & Freibott, 1991 ; Heid, 1994, 2001, 2009).

Cette tâche est exclusivement focalisée sur les collocations du type verbe-COD (*développer une leucémie*) et verbe-COI (*relever d'une anomalie du système immunitaire*). Ces collocations sont examinées d'un point de vue quantitatif et qualitatif, grâce à une approche automatique. Sur le plan qualitatif, la méthode d'analyse appliquée est une forme de gradation descendante qui se fait sur deux niveaux d'analyse différents. Elle commence par la sémantique, pour s'achever sur le lexique. Ces niveaux d'analyse nous renseignent sur les propriétés sémantiques des verbes, à travers l'analyse de la distribution des collocations verbe-terme observées dans les corpus. En procédant de la sorte, notre principal objectif est d'identifier les préférences sélectionnelles des verbes, en fonction des différents corpus, en ce qui concerne les catégories sémantiques des arguments et les unités lexicales qui instancient ces catégories sémantiques Snomed.

3.6.1 Extraction des collocations verbe-terme

Les collocations analysées dans cette partie de la thèse sont acquises grâce au programme d'annotation sémantique décrit à la section 3.2.2, qui permet également d'extraire des patrons de sous catégorisation verbaux (désormais *PSC*) sous le format suivant : Verbe+COD+CatSnomed (*développer une leucémie*_D), verbe+COI+CatSnomed (*relever d'une anomalie du système immunitaire*_D). Pour chaque verbe sélectionné à l'étape 3.3.1, ces deux types de collocations sont extraits.

3.6.2 Analyse des collocations verbe-terme

L'analyse comparative des collocations verbe-terme a pour but de mettre en évidence la distribution (syntaxico-sémantique et lexicale) des arguments des verbes dans les différents corpus. Cette répartition permet d'examiner les simulitudes et les disparités entre les quatre types de textes médicaux qui constituent notre corpus, en ce qui concerne le choix des collocations verbe-terme utilisées sur les plans syntaxique, sémantique et lexical. La figure 3.8 décrit la démarche appliquée dans cette analyse, en présentant le type de données qu'elle permet d'obtenir. Prenons le verbe *relever* à titre d'illustration [29]. Supposons qu'il a 50 arguments dans l'un de nos corpus.

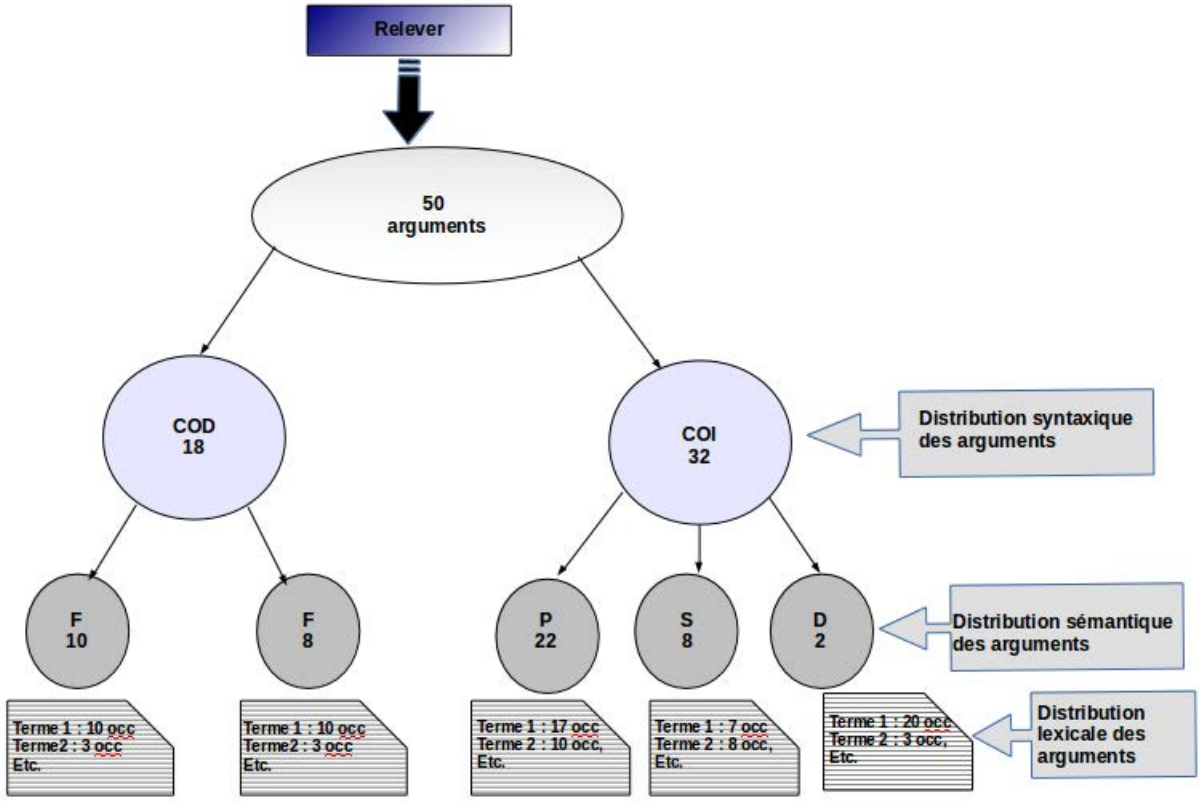


FIG. 3.8 – Modèle d'analyse de la distribution des arguments du verbe dans un corpus donné.

La figure 3.8 décrit la distribution syntaxique, sémantique et lexicale des arguments du verbe dans le corpus en question. Cet exemple met en évidence les diverses configurations auxquelles donne accès une telle analyse distributionnelle. Pour chaque verbe étudié, l'analyse qualitative consiste à contraster ces figures dans les quatre sous-corpus sur lesquels porte cette étude.

29. Attention : cet exemple a été conçu à titre d'illustration. Nous tenons donc à signaler que la distribution d'arguments et les nombres proposés dans la figure 3.8 sont tous imaginaires.

3.6.2.1 Analyse quantitative

L'analyse quantitative des collocations consiste principalement en une évaluation numérique des types de collocations (syntaxique, préférences sémantiques et lexicales) observées en corpus, pour chacun des verbes retenus pour cette étude. Cette démarche d'analyse permet d'étudier la fréquence des cooccurrences entre le verbe, les catégories syntaxiques, et sémantiques de ses arguments, ainsi que les unités lexicales qui les instancient.

Le programme implémenté pour la réalisation de cette tâche effectue une étude contrastive de la fréquence des multiples patrons de sous catégorisation des verbes, à partir des données résultant de la phase d'extraction de ces patrons de sous catégorisation. Pour chaque verbe, ce programme exécute les tâches suivantes :

- identification des arguments COD et COI cooccurrents dans chaque sous-corpus et calcul de leur somme totale ;
- calcul du nombre de catégories Snomed associées à chaque type d'argument ;
- extraction des termes occupant les positions d'arguments ;
- calcul pour chaque terme de :
  - sa fréquence cumulée dans le corpus ;
  - son nombre d'occurrences lorsqu'il est associé à chaque catégorie Snomed.

Pour chaque corpus, le nombre total d'arguments portant une catégorie Snomed vs. nombre total d'arguments non annotés sémantiquement est également calculé. Les problèmes de l'étiquetage sémantique des arguments des verbes (cf. section 4.2) pourraient avoir un impact sur les chiffres résultant de ces opérations. Nous fournirons plus de détails relatifs à ces phénomènes lors de l'évaluation de l'annotation sémantique (cf. chapitre 5).

3.6.2.2 Analyse qualitative : préférences sémantico-lexicales des verbes

À partir des données statistiques obtenues grâce à l'analyse quantitative, on étudie la distribution en corpus des termes qui jouent les rôles de COD et COI des verbes, en relation avec les catégories sémantiques Snomed. En effet, les calculs effectués à l'étape précédente sur la base des fréquences mettent en évidence les préférences sélectionnelles des verbes dans chaque sous-corpus, en ce qui concerne les catégories sémantiques Snomed des compléments et les unités lexicales qui illustrent ces catégories. Ainsi, pour chaque verbe, on peut détecter les catégories sémantiques associées aux termes qui fonctionnent en tant que objets directs ou indirects, de même que les unités lexicales qui ont tendance à porter le plus souvent ces catégories sémantiques. Toutes ces informations correspondent aux deux niveaux d'analyse mentionnés précédemment :

- niveau sémantique : pour un verbe donné, les catégories sémantiques dominantes identifiées dans chaque sous-corpus à l'étape précédente sont enregistrées et comparées à celles des

autres corpus. Les résultats obtenus font transparaître les similitudes et les spécificités entre les différents sous-corpus, en termes de choix de types sémantiques d'arguments, sur la base des catégories Snomed. Par exemple, d'après les données extraites et les calculs effectués, dans le corpus des experts, le verbe *subir* s'associe très régulièrement aux termes de la catégorie P (procédure) qui occupe la tête de liste, tandis que dans le corpus des forums, ce verbe a tendance à s'associer fréquemment aux compléments de type D (maladie).

- niveau lexical : une démarche similaire est appliquée pour la comparaison des corpus dans le but d'identifier des préférences des verbes dans le choix des unités lexicales qui jouent le rôle de compléments. Toutefois, contrairement aux préférences sémantiques des verbes, les collocations lexicales ne sont pas systématiquement identifiées sur la base de la fréquence des catégories sémantiques auxquelles elles correspondent. En effet, il a été constaté que certaines collocations lexicales sont très fréquentes au détriment de la faible fréquence (en corpus) d'association entre le verbe et les termes de la catégorie Snomed concernée. C'est le cas de la collocation *suivre un patient* qui est fréquente dans le corpus des experts, pourtant l'association entre le verbe *suivre* et les compléments de type STATUTSOCIAL ne fait pas partie des plus fréquentes dans ce corpus. Ce résultat nous pousse à déduire que les préférences lexicales des verbes (verbe-terme) ne sont pas déterminées par leurs attirances sémantiques (verbe-catégorie sémantique).

3.7 Bilan

Dans ce chapitre, nous avons présenté la méthode semi-automatique implémentée dans ce travail de thèse pour la création d'une ressource alignant les patrons syntaxico-sémantiques spécialisés des verbes avec leurs équivalents non spécialisés. Cette méthode s'applique sur des données tirées d'un corpus constitué de quatre types de textes médicaux ayant des auteurs et publics cibles différents. L'étiquetage syntaxique des corpus est effectué grâce à l'analyseur syntaxique Cordial, tandis que l'acquisition des patrons sémantiques est basée sur les catégories sémantiques de la terminologie médicale Snomed International. Notre méthode se distingue des approches existantes sur différents aspects : tout d'abord, dans le domaine de la simplification de textes, il n'existe pas d'approches autant focalisées sur le verbe que la nôtre. Ce constat signale le caractère novateur de la méthode proposée dans ce travail de thèse. De surcroît, le principal résultat visé (la création d'une ressource de simplification alignant les patrons verbaux spécialisés vs. non spécialisés) est une innovation dans le domaine de la simplification des textes où la plupart des ressources similaires portent sur les entités nominales. Sur le plan linguistique, l'annotation sémantique des corpus à partir des catégories de la terminologie Snomed remplace la phase traditionnelle d'annotation des corpus en rôles sémantiques. Les patrons sémantiques

des verbes sont acquis grâce aux catégories sémantiques Snomed qui sont propres au domaine médical. Notre méthode permet également d'analyser les propriétés sémantiques des verbes à travers la comparaison des collocations verbe-terme observées dans les différents corpus (cf. chapitre 4). Dans cette démarche, les catégories Snomed fréquemment associées aux arguments des verbes permettent de détecter les choix préférentiels des verbes sur les plans sémantique et lexical. Les résultats de cette analyse contrastive seront présentés dans le chapitre suivant.

En ce qui concerne son positionnement par rapport aux travaux de l'état de l'art, certains aspects de notre méthode rejoignent les approches existantes dans le domaine de l'annotation des rôles sémantiques (*Semantic Role Labelling*) et en WSD, mais se démarque sur certains aspects. Dans la section 1.3.4 du chapitre 1, nous avons établi un parallélisme entre la méthode FrameNet et la nôtre, tout en mettant l'accent sur certains éléments qui distinguent ces deux approches. La description détaillée de la méthode appliquée dans ce travail de thèse a permis de comparer la méthodologie de FrameNet et la nôtre d'un point de vue global. C'est ce que décrit la figure 3.9 :

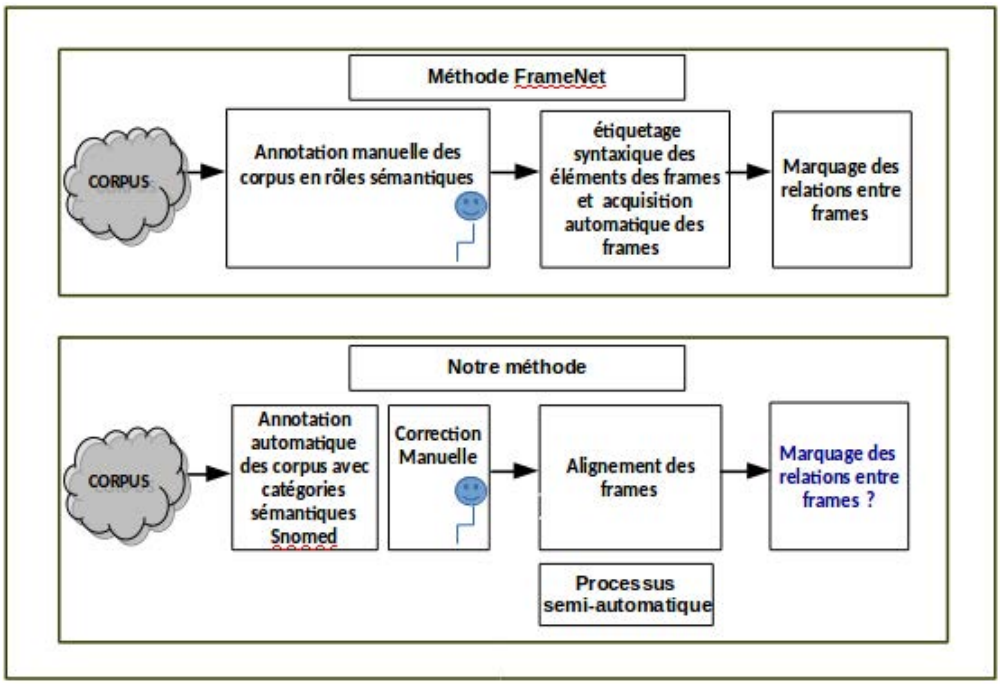


FIG. 3.9 – Méthode FN vs. notre méthode.

Comme le montre la figure 3.9, la principale différence se situe au niveau de la démarche d'annotation sémantique des textes. Notre chaîne de travail passe par un processus automatique d'attribution des catégories sémantiques aux arguments des verbes. Dans le projet FrameNet (désormais *FN*) par contre, l'annotation sémantique des arguments des verbes est entièrement effectuée par des annotateurs humains. Comme l'illustre la figure 3.9, notre méthode de travail et celle de FN présentent des similitudes, mais se distinguent au niveau de l'intervention humaine,

qui dans notre méthodologie se fait en guise de correction, après l'attribution automatique des catégories sémantiques. Par contre, dans la démarche FrameNet, l'annotation sémantique des corpus est faite manuellement, tandis que la phase d'acquisition des frames est réalisée automatiquement. Ces deux approches, qui relèvent de choix méthodologiques différents, produisent les résultats souhaités. Toutefois, la démarche que nous proposons a pour avantage de réduire la charge de travail manuel des annotateurs, favorisant ainsi le gain de temps sur l'ensemble de la chaîne de travail. De surcroît, notre méthode de travail peut être exploitée pour la création d'un FrameNet médical (cf. figure 3.9). En effet, la méthode d'annotation de corpus que nous implémentons aboutit à l'extraction des PSS qui sont des formes de frames. La mise en relation de ces frames peut se faire après la phase d'extraction automatique des groupes de verbes instanciant un même PSS. Les verbes ainsi extraits partagent un type varié de relations entre eux. La phase de tri manuel des différents groupes de verbes permet de former des classes plus restreintes et sémantiquement homogènes ; à ce niveau il est possible de regrouper les PSS selon les relations utilisées dans la méthode FN, à savoir les relations *inheritage, subframe et using* afin de générer un FrameNet médical.

La méthode décrite dans ce troisième chapitre de la thèse implique également une technique de désambiguïsation du sens des termes médicaux qui rejoint les pratiques caractérisant l'état de l'art dans ce domaine. Toutefois, contrairement à la plupart des travaux existants, notre méthode n'est pas basée sur les techniques de l'apprentissage automatique (supervisées et non supervisées). Cette démarche a pour avantage d'éviter les contraintes qu'impose la constitution d'un corpus d'entraînement manuellement annotés. De surcroît, la majorité des méthodes statistiques appliquées dans le domaine de la désambiguïsation des noms se basent sur les cooccurrences nominales du terme ambigu afin de distinguer ses différents sens. Notre démarche de désambiguïsation du sens des termes ambigus se base plutôt sur la cooccurrence verbale (verbe-terme) et d'autres paramètres contextuels pour la distinction des sens. Les différents sens à distinguer sont identifiées grâces aux catégories sémantiques de la terminologie Snomed. La coocurrence verbe-terme est parfois exploitée dans le domaine de la WSD, mais très souvent dans un but qui se distingue du nôtre. Certains travaux comme celui Wagner *et al.* (2009) y ont recours pour la désambiguïsation du sens des verbes, or nous appliquons cette technique pour la désambiguïsation du sens des termes susceptibles de porter plusieurs catégories dans la Snomed.

Chapitre 4

Résultats et discussion

Dans le chapitre précédent, nous avons décrit les étapes de la chaîne de travail qui constitue la méthode appliquée dans ce travail de thèse. Ce chapitre est consacré à la présentation des résultats obtenus au terme de chaque étape. Ces résultats sont accompagnés d'une discussion qui permet de mettre en évidence les différentes observations faites. La figure 4.1 rappelle les principales phases qui constituent la méthode, avec à chaque fois une mise en exergue de l'étape qui est décrite dans la section concernée.

4.1 Résultats de l'annotation des corpus et acquisition des PSS

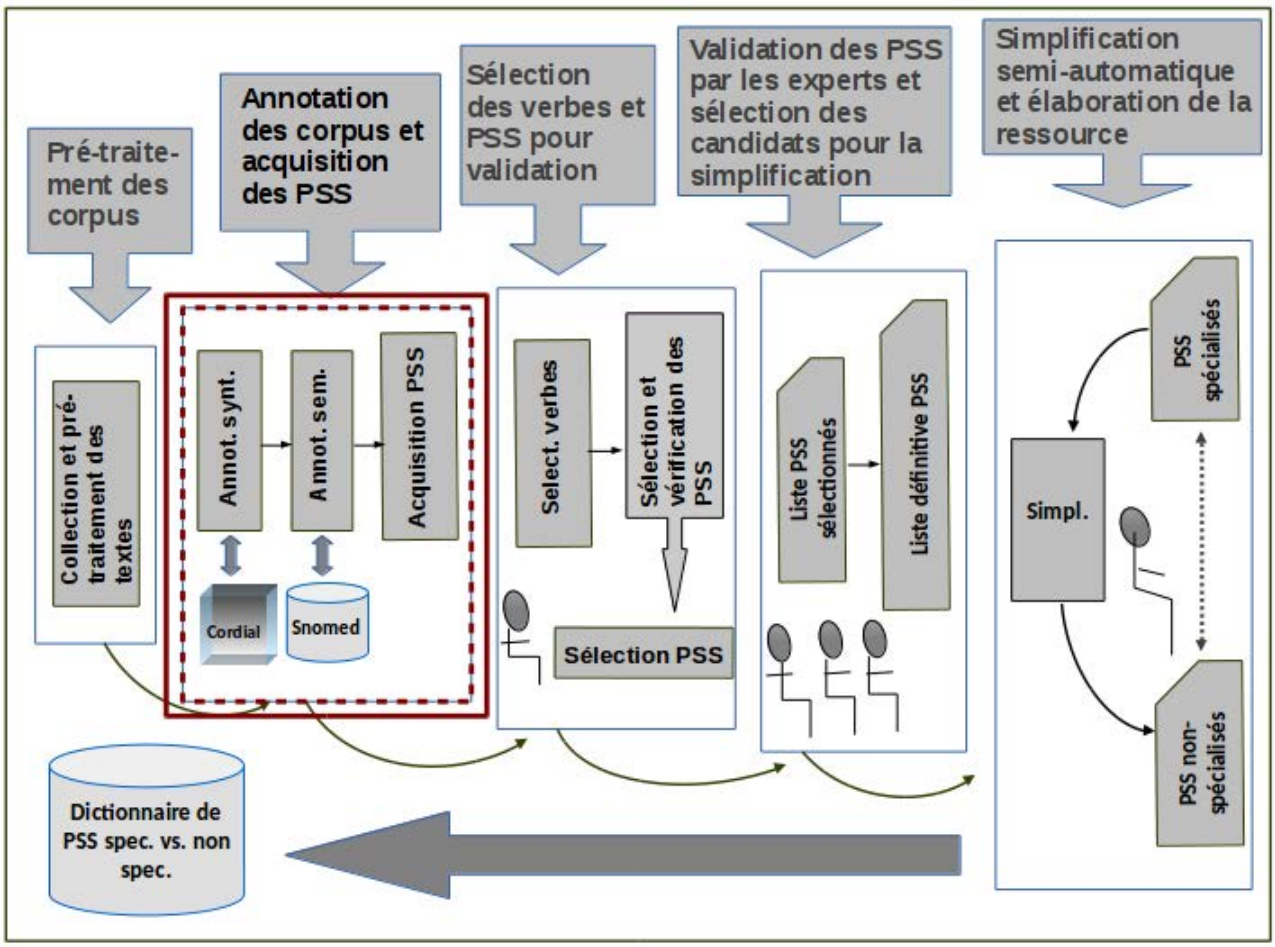


FIG. 4.1 – Schéma de la méthode : l'annotation syntaxique.

4.1.1 Annotation syntaxique et extraction des schémas valenciels

De façon globale, l'annotation syntaxique des corpus consiste à identifier automatiquement les différents syntagmes constituant la phrase, en leur associant des fonctions syntaxiques. Dans ce travail de thèse, cette tâche a été réalisée grâce à l'analyseur syntaxique Cordial. Les résultats obtenus à partir des quatre types de corpus nous ont permis d'extraire les schémas valenciels des verbes sous le format suivant *verbe|argument1|argument2|etc.|phrase exemple*, comme l'illustre l'exemple ci-dessous :

1) **présenter**|deux formes**_s**|aucun risque hémolytique**_cod**|Les deux formes n'ont présenté aucun risque hémolytique.

Les patrons syntaxiques résultants subissent ensuite un processus automatique (cf. section 4.1.2) visant à associer des catégories sémantiques aux termes-arguments, de façon à acquérir des patrons syntaxico-sémantiques.

La discussion des résultats de l'annotation syntaxique des sous-corpus sera subdivisée en deux parties : une présentation sommaire, qui sera suivie d'une phase d'analyse et d'interprétation. L'analyse sera axée sur deux principaux aspects : dans un premier temps, l'on se penchera sur la relation entre la fréquence des verbes pris individuellement et les différents types de corpus dans lesquels ils figurent. Dans la seconde phase d'analyse, les résultats seront abordés d'un point de vue général qui permettra de décrire les relations entre les différents sous-corpus (proximité vs. distance) à partir des verbes qui les caractérisent.

4.1.1.1 Récapitulatif des résultats de l'annotation syntaxique

Le tableau 4.1 propose un résumé des résultats de l'annotation syntaxique des corpus avec l'outil Cordial. De la gauche vers la droite, ce tableau fournit pour chaque corpus : le nombre d'occurrences de phrases annotées par Cordial (*occ. phr. cord.*) ; le nombre d'occurrences de phrases verbales (c'est-à-dire contenant au moins un verbe) ; le pourcentage de ces phrases par rapport aux phrases annotées (*pourcent.*) ; le nombre d'occurrences de tous les verbes du corpus (*nbocc vb*) et sa valeur normalisée en termes de ppm (*part par million*) par rapport au nombre de mots du corpus (cf. chapitre 2, section 2.1.3) ; le nombre total de lemmes verbaux (*lemmes vb*) et sa valeur normalisée (*ppm*) par rapport à l'ensemble des mots du corpus ; le nombre de verbes dont la fréquence est inférieure à la moyenne des quatre corpus ; le nombre de verbes dont la fréquence est supérieure à la moyenne des quatre corpus. Le calcul des deux dernières valeurs concerne uniquement les verbes dont la fréquence dans chaque sous-corpus est supérieure à 0, c'est-à-dire ceux apparaissant au moins une fois dans chacun des quatre corpus. On en dénombre au total 977 sur 2859 verbes que contient l'ensemble des corpus.

Les données de ce tableau, associées aux résultats du tableau 4.2, nous permettront de faire une analyse de la fréquence des verbes dans notre corpus médical.

TAB. 4.1 – La fréquence verbale dans les 4 types de corpus.

	occ. phr. Cord	occ. phr vb	pourcent.	nbocc. vb	ppm	lemmes vb	ppm	freq ≤moy	freq ≥moy
PRO	102208	53873	52,70	63910	42530	1374	914,36	739	238
ETU	131986	69475	52,63	89017	50707	1509	859,58	622	355
VUL	123317	83421	67,64	101692	62484	1586	974,52	482	495
FOR	146952	101310	68,94	173903	109462	2012	1266,44	568	409

À la lecture du tableau 4.1, nous pouvons constater que le corpus des forums enregistre le plus grand pourcentage de phrases verbales. De surcroît, dans ce corpus, le nombre d'occurrences des verbes en termes de ppm représente plus du double de celui du corpus des experts. Ces résultats en chiffres normalisés nous poussent à écarter toute hypothèse relative à la taille des corpus. D'ailleurs, les informations fournies dans le chapitre 2 indiquent que les corpus PRO et FOR ont des tailles sensiblement identiques, avec un peu plus d'un million et demi de mots chacun (cf. chapitre 2, section 2.1.3).

Le corpus des forums est également en tête de liste en ce qui concerne le nombre de verbes. Le nombre de verbes (valeur en ppm) par corpus confirme cette observation : 914,36 verbes dans le corpus PRO contre 1266,44 pour le corpus FOR. Ce constat peut s'expliquer si l'on considère le fait que les textes de forums relèvent de la langue générale, étant donné qu'ils sont rédigés par des non-experts du domaine médical. Par contre, les autres corpus, bien que visant des publics différents, tombent dans la catégorie des textes scientifiques, puisqu'ils ont tous été rédigés par des experts en médecine. En effet, plusieurs travaux de recherche (Fang, 2005 ; Tellier, 2008) soutiennent l'hypothèse de la faible fréquence des unités verbales dans les écrits scientifiques. Ces textes sont plutôt reconnus pour la prépondérance des unités nominales (Condamines & Bourigault, 1999 ; Fang, 2005 ; Deléger & Zweigenbaum, 2008 ; L'Homme, 2012b), par comparaison avec les textes littéraires ou ordinaires (non scientifiques) qui ont tendance à être riches en verbes. Les données provenant de nos corpus illustrent cette observation. Le nombre d'occurrences de phrases verbales, ainsi que le nombre d'occurrences verbales, décroît lorsqu'on évolue du corpus des forums vers le corpus des experts. Cette dissemblance transparaît plus clairement à travers la figure 4.2 qui met bien en évidence la gradation descendante caractérisant le nombre d'occurrences des verbes normalisé (sous forme de ppm) dans les quatre corpus.

La disproportionnalité qui caractérise le corpus FOR par rapport aux autres, en ce qui concerne le nombre de phrases et occurrences verbales, pourrait également être une conséquence de la forte présence de phrases verbales courtes dans le corpus. À ce stade, il faudrait rappeler que les textes du corpus des forums consistent en des extraits de conversations entre personnes intervenant sur des plateformes d'échanges spécialisées dans le domaine médical. Ces conversations informelles

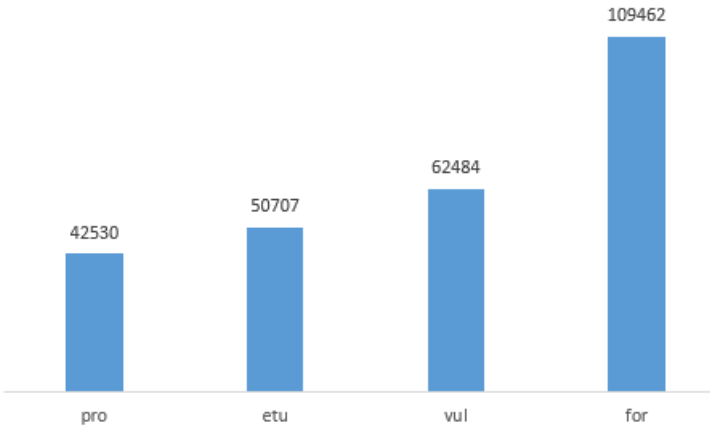


FIG. 4.2 – Comparaison de la fréquence des verbes dans les corpus.

contiennent très souvent des phrases courtes, comme celles ci-dessous, tirées du corpus [1] :

2) *je flippe comme pas possible.*

3) *ça veut dire ? ?*

4) *c'est quoi ?*

5) *je peux...*

6) *je peux marcher...*

Ce type de phrases, parfois inachevées, sont récurrentes dans les textes de forums car elles font partie des caractéristiques de la communication instantanée via les médias. Il ne serait donc pas surprenant que leur présence ait contribué à augmenter le nombre de phrases verbales du corpus FOR. Néanmoins, l'on ne saurait associer le nombre important de verbes du corpus des forums uniquement à ce phénomène de phrases courtes. Tel qu'il a été souligné précédemment, nous pensons que l'utilisation fréquente des constructions verbales par les non-experts, au détriment des constructions nominales, découle d'un choix délibéré des rédacteurs qui, par ce procédé, donnent un caractère personnel à leurs écrits et expriment des faits qui relèvent de leurs expériences quotidiennes. La transformation des constructions verbales en constructions nominales renvoie à un procédé appelé *nominalisation* (Fang, 2005), qui fait partie des pratiques caractéristiques des textes scientifiques. Ci-dessous quelques exemples :

examiner un patient → examen

traiter un patient → traitement

En effet, la nominalisation contribue à rendre les textes des experts plus abstraits, comme le souligne Christie (2002, p. 66) en ces termes : *[nominalised phrases] abstract away from immediate,*

1. Dans ce travail, les phrases extraites du corpus des forums sont présentées telles quelles, avec les éventuelles fautes qu'elles peuvent contenir.

lived experiences, to build instead truths, abstractions, generalizations, and arguments. Fang (2005)[2] donne davantage d'explications en précisant que la nominalisation permet à l'auteur de créer des termes techniques ou des entités nouvelles, d'établir des relations de cause à effet entre des phénomènes disparates, de synthétiser et de systématiser des informations données précédemment. L'auteur va encore plus loin en rappelant les propos de Halliday (1998) qui souligne que la nominalisation implique plus que le remodelage de la grammaire, qu'il s'agit d'un processus de transformation du sens de la phrase, puisque la nominalisation d'une construction verbale débouche sur l'omission de différentes informations, ce qui génère très souvent des ambiguïtés.

4.1.1.2 Fréquence du verbe et type de corpus

Le tableau 4.1 a permis d'avoir une vision panoramique des résultats de l'annotation syntaxique à travers le nombre de phrases, de verbes, ainsi que leurs fréquences cumulées dans chaque corpus. Le tableau 4.2, quant à lui, fournit des données plus précises qui permettront de décrire la relation entre la fréquence verbale et les différents types de sous-corpus. Il s'agit d'une liste dans laquelle les verbes (principalement les 45 verbes sélectionnés pour la simplification) sont associés à leurs fréquences respectives dans chacun des sous-corpus. Ces fréquences sont associées à une valeur que nous appelons la fréquence normalisée (*freqN*). Il s'agit de la proportion (sur 10 000) que chaque fréquence verbale représente par rapport au nombre total d'occurrences du verbe dans chaque type de corpus. Cet échantillon[3] tiré des résultats obtenus sur l'ensemble des verbes du corpus va nous permettre d'introduire et de décrire un certain nombre de phénomènes qui ont été observés.

Lorsqu'on analyse la variation de fréquence des verbes du tableau 4.2 d'un corpus à l'autre, trois catégories de verbes émergent : les verbes qui semblent se rapprocher du corpus PRO, ceux qui semblent se rapprocher du corpus FOR et ceux dont le fonctionnement entre les deux grands types de sous-corpus (experts vs. non-experts) ne présente pas une grande différence. Ces catégories se distinguent par le mode de variation de la fréquence des verbes.

1. La première catégorie regroupe les verbes qui semblent se rapprocher du corpus PRO. Il s'agit des verbes ayant une fréquence élevée dans les corpus de textes rédigés pour les experts (médecins et étudiants), tandis que dans les corpus adressés au grand public (VUL et FOR), leurs valeurs fréquentielles se situent à l'opposé. *Présenter*, *évaluer* et *dépister* appartiennent à cette catégorie. Cette tendance récurrente (cf. tableau 4.2) pousse à s'interroger sur la nature de ce phénomène : serait-ce juste une coïncidence ? Ou alors y

2. Il reprend les termes de Veel (1997, p. 184).

3. Les verbes qui constituent cet échantillon ont été sélectionnés selon des critères définis dans le chapitre précédent, section 3.3.1.

TAB. 4.2 – Les verbes et leurs fréquences dans les 4 sous-corpus.

	PRO		ETU		VUL		FOR	
lemmes	freq	freqN	freq	freqN	freq	freqN	freq	freqN
abaisser	26	4.06	26	2.92	26	2.55	6	0.34
accompagner	163	25.5	352	39.5	274	26.9	74	4.25
activer	20	3.12	26	2.92	25	2.45	15	0.86
administrer	161	25.1	84	9.43	220	21.6	15	0.86
affecter	72	11.2	59	6.62	117	11.5	8	0.46
altérer	16	2.50	63	7.07	19	1.86	7	0.40
associer	481	75.2	700	78.6	164	16.1	37	2.12
coloniser	11	1.72	8	0.89	11	1.08		
contrôler	88	13.7	59	6.62	65	6.39	79	4.54
diagnostiquer	43	6.72	42	4.71	85	8.35	62	3.56
disséminer	7	1.09	15	1.68	8	0.78		
dépister	13	2.03	17	1.90	12	1.18	1	0.05
détecter	34	5.31	31	3.48	31	3.04	31	1.78
développer	91	14.2	114	12.8	249	24.4	38	2.18
envisager	104	16.2	100	11.2	47	4.62	36	2.07
exposer	66	10.3	110	12.3	78	7.67	17	0.97
impliquer	66	10.3	75	8.42	53	5.21	15	0.86
imposer	94	14.7	157	17.6	31	3.04	35	2.01
indiquer	205	32.0	169	18.9	101	9.93	81	4.65
induire	51	7.97	57	6.40	27	2.65	12	0.69
inhaler	10	1.56	14	1.57	8	0.78		
nécessiter	133	20.8	220	24.7	95	9.34	14	0.80
observer	193	30.1	166	18.6	77	7.57	11	0.63
poursuivre	34	5.31	25	2.80	48	4.72	11	0.63
pratiquer	52	8.13	50	5.61	50	4.91	49	2.81
produire	79	12.3	48	5.39	139	13.6	124	7.13
provoquer	38	5.94	79	8.87	181	17.7	146	8.39
préconiser	27	4.22	10	1.12	6	0.59	6	0.34
présenter	289	45.2	161	18.0	224	22.0	75	4.31
recommander	195	30.5	68	7.63	96	9.44	39	2.24
relever	36	5.63	29	3.25	16	1.57	35	2.01
requérir	39	6.10	21	2.35	13	1.27		
réaliser	204	31.9	187	21.0	95	9.34	70	4.02
signaler	75	11.7	27	3.03	27	2.65	29	1.66
subir	43	6.72	28	3.14	35	3.44	97	5.57
synthétiser	2	0.31	12	1.34	2	0.19	1	0.05
sécréter	5	0.78	4	0.44	2	0.19	1	0.05
traduire	22	3.44	82	9.21	15	1.47	9	0.51
traiter	72	11.2	55	6.17	58	5.70	93	5.34
transmettre	26	4.06	13	1.46	97	9.53	14	0.80
éliminer	19	2.97	20	2.24	10	0.98	5	0.28
évaluer	139	21.7	47	5.27	22	2.16	7	0.40
évoquer	35	5.47	84	9.43	8	0.78	5	0.28

aurait-il une explication plausible à ce fonctionnement commun à plusieurs verbes ?

Pour répondre à cette question de façon objective, une expérience a été effectuée avec pour but de repérer les verbes dont la variation de fréquence signalerait une tendance préférentielle du verbe vis-à-vis d'un type de corpus particulier. Ce test passe par le calcul de la fréquence moyenne de chaque verbe (dans l'ensemble du corpus), qui sert de mesure pour la détection des prédicats se démarquant de par leur nombre d'occurrences. Pour chaque verbe, la fréquence dans les différents sous-corpus est comparée au nombre qui sert de mesure (la moyenne). En procédant ainsi, il est possible d'identifier les verbes qui se distinguent, notamment avec une fréquence supérieure ou inférieure à la mesure. Les deux dernières colonnes du tableau 4.1 proposent des chiffres qui résument les résultats de ce test. Ces chiffres révèlent que les corpus VUL et FOR comptent le plus grand nombre de verbes ayant une fréquence au dessus de la moyenne, respectivement (495/977) et (409/977), tandis que les corpus PRO et ETU enregistrent moins de prédicats verbaux fonctionnant de la sorte.

Le tableau 4.3 donne plus de précisions sur ce phénomène en présentant un extrait plus détaillé des données résultant de l'analyse de la fréquence des verbes en corpus. Il propose une liste de verbes avec leurs fréquences dans chaque corpus. Ces fréquences sont associées à leurs valeurs normalisées sur 10 000 (*freqN*). Cette fréquence normalisée joue un rôle déterminant à ce stade car elle permet de comparer, pour chaque verbe, les valeurs qu'enregistrent les différents corpus par rapport à la moyenne (somme des fréquences normalisées divisée par 4). Ainsi, les verbes que contient le tableau 4.3 sont ceux dont la fréquence normalisée dans les corpus PRO et/ou ETU est supérieure à la moyenne.

Les verbes du tableau 4.3 tendent à être fréquemment utilisés dans les textes écrits à l'attention des professionnels et/ou étudiants (PRO et/ou ETU). Si l'on considère la fréquence comme indicateur de spécificité d'un verbe dans le discours typique d'un corpus (et de son public), alors les verbes du tableau 4.3 peuvent être considérés comme des prédicats associés aux corpus des experts. Les lexèmes *présenter*, *évaluer* et *dépister*, avec des fréquences bien au dessus de la moyenne dans le corpus des experts et/ou celui des étudiants, se confirment comme prédicats liés aux textes des experts, tandis que *conseiller* (freqN PRO = 7,04 vs. freqN FOR = 22,30), *faire* (86,90 vs. 330), *prescrire* (5,47 vs. 9,31), et *souffrir* (3,91 vs. 16,80) semblent ne pas occuper une place de choix dans ce type de textes. Pourtant, dans le corpus des non-experts, ces derniers se démarquent considérablement (cf. tableau 4.4).

L'analyse des données du tableau 4.3 nous permet d'émettre l'hypothèse selon laquelle la haute fréquence de certains verbes dans le corpus des experts et/ou des étudiants pourrait signaler leur nature terminologique. Autrement dit, les verbes très fréquents dans ces

TAB. 4.3 – Verbes avec fréquence supérieure à la moyenne dans les corpus experts (PRO et/ou ETU).

		PRO		ETU		VUL		FOR	
lemmes	moy.	freq	freqN	freq	freqN	freq	freqN	freq	freqN
abaisser	2,46	26	4,06	26	2,92	26	2,55	6	0,34
accompagner	24,03	163	25,5	352	39,5	274	26,9	74	4,25
activer	2,3375	20	3,12	26	2,92	25	2,45	15	0,86
améliorer	15,35	178	27,8	173	19,4	103	10,1	71	4,08
analyser	10,07	113	17,6	130	14,6	50	4,91	55	3,16
appliquer	10,84	103	16,1	80	8,98	168	16,5	31	1,78
associer	43,01	481	75,2	700	78,6	164	16,1	37	2,12
augmenter	38,89	278	43,4	436	48,9	545	53,5	170	9,77
comporter	14,12	112	17,5	243	27,2	109	10,7	19	1,09
définir	11,41	146	22,8	143	16	63	6,19	11	0,63
dépister	1,29	13	2,03	17	1,9	12	1,18	1	0,05
développer	13,40	91	14,2	114	12,8	249	24,4	38	2,18
diminuer	17,87	171	26,7	218	24,4	154	15,1	92	5,29
effectuer	15,37	159	24,8	187	21	139	13,6	36	2,07
envisager	8,52	104	16,2	100	11,2	47	4,62	36	2,07
évaluer	7,38	139	21,7	47	5,27	22	2,16	7	0,4
évoluer	3,14	24	3,75	40	4,49	26	2,55	31	1,78
évoquer	3,99	35	5,47	84	9,43	8	0,78	5	0,28
exposer	7,81	66	10,3	110	12,3	78	7,67	17	0,97
exprimer	3,31	25	3,91	59	6,62	19	1,86	15	0,86
favoriser	8,83	53	8,29	145	16,2	103	10,1	13	0,74
impliquer	6,20	66	10,3	75	8,42	53	5,21	15	0,86
imposer	9,34	94	14,7	157	17,6	31	3,04	35	2,01
inclure	7,15	112	17,5	51	5,72	52	5,11	5	0,28
indiquer	16,37	205	32	169	18,9	101	9,93	81	4,65
induire	4,43	51	7,97	57	6,4	27	2,65	12	0,69
inhiber	1,42	12	1,87	31	3,48	3	0,29	1	0,05
isoler	1,42	8	1,25	31	3,48	8	0,78	3	0,17
nécessiter	13,91	133	20,8	220	24,7	95	9,34	14	0,8
observer	14,23	193	30,1	166	18,6	77	7,57	11	0,63
poursuivre	3,365	34	5,31	25	2,8	48	4,72	11	0,63
pratiquer	5,365	52	8,13	50	5,61	50	4,91	49	2,81
préconiser	1,57	27	4,22	10	1,12	6	0,59	6	0,34
présenter	22,38	289	45,2	161	18	224	22	75	4,31
réaliser	16,57	204	31,9	187	21	95	9,34	70	4,02
recommander	12,45	195	30,5	68	7,63	96	9,44	39	2,24
relever	3,12	36	5,63	29	3,25	16	1,57	35	2,01
survenir	11,54	85	13,2	189	21,2	88	8,65	54	3,1
traduire	3,66	22	3,44	82	9,21	15	1,47	9	0,51
transmettre	3,96	26	4,06	13	1,46	97	9,53	14	0,8

deux corpus pourraient être des sortes de termes médicaux, c'est-à-dire des lexèmes ayant des emplois qui relèvent du langage spécialisé de la médecine.En guise d'exemples, les lemmes *relever* (5,63/3,12 et 3,25/3,12), *présenter* (45,2/22,38) et *envisager* (16,2/8,52 et 11,2/8,52), qui apparaissent dans le tableau 4.3, font partie des prédicats dont certains emplois en corpus ont été validés par des professionnels de la santé comme faisant partie du jargon utilisé régulièrement dans leurs écrits et dans l'exercice de leur fonction. Lorsqu'ils se retrouvent dans des textes adressés au grand public, ce type d'emplois verbaux, propres aux experts, risquent d'avoir un sens différent de celui qu'ils ont habituellement en langue vernaculaire. Ce changement de sens pourrait engendrer des difficultés de lecture et de compréhension chez les non-experts, d'où la nécessité de les simplifier, afin de garantir l'intercompréhension entre les experts en médecine et le grand public.

2. La deuxième catégorie regroupe les verbes qui semblent être plus utilisés dans les textes destinés aux non-experts que dans les textes pour experts. Ce groupe concerne les verbes ayant une fréquence plus élevée dans les corpus pour non-experts (FOR et/ouVUL) par contraste avec ceux des experts. Les verbes *souffrir*, *consulter* et *faire* illustrent très bien ce mode de fonctionnement. *Souffrir* a une moyenne normalisée de 7,65 occurrences dans le corpus entier. Le corpus des forums occupe la tête du classement avec une fréquence normalisée (16,8) qui représente plus du double de la moyenne (7,65), tandis que ceux des étudiants et des experts cumulés enregistrent une valeur (5,25) inférieure à la moyenne. Par contre, les verbes *évoquer, recommander* et *associer* semblent être peu utilisés dans les corpus pour non-experts, plus particulièrement dans le corpus FOR (cf. tableau 4.4). Le tableau 4.4 fournit plusieurs autres exemples de verbes appartenant à cette catégorie. La fréquence élevée des prédicats de cette catégorie (dans les corpus pour non-experts) pourrait signaler qu'il s'agit d'unités lexicales dotées d'un faible degré de spécialisation et qui correspondraient donc au faible niveau d'expertise des rédacteurs, si ces derniers sont des « profanes » du domaine médical. La forte présence de verbes tels que *souffrir, consulter, avoir, hospitaliser, soigner, supporter,* etc. qui évoquent la maladie, la douleur, pourrait signifier que les non-experts (malades ou entourage des malades) sont davantage concernés par la souffrance, qui serait un sujet de préoccupation plus secondaire pour les experts dont les textes traitent certes des maladies, mais davantage des méthodes et procédures médicales réalisées ou à entreprendre dans le cadre des soins. Les résultats de l'analyse des collocations verbe-terme (cf. section 4.6.1) vont dans le même sens en ce qu'ils signalent la prédominance des termes de la catégorie P (et, en second lieu, D) au sein des corpus experts et étudiants. Pour le corpus des forums, par contre, la catégorie D est en tête dans les collocations (verbe-catégorie Snomed).

TAB. 4.4 – Verbes avec fréquence supérieure à la moyenne dans FOR et/ou VUL.

		PRO		ETU		VUL		FOR	
lemmes	moy.	freq	freqN	freq	freqN	freq	freqN	freq	freqN
accoucher	0,98	3,00	0,46	11,00	1,23	3,00	0,29	34,00	1,95
allonger	3,20	15,00	2,34	37,00	4,15	24,00	2,36	69,00	3,96
attaquer	2,09	3,00	0,46	5,00	0,56	50,00	4,91	42,00	2,41
avoir	703	2601	406	3588	403	4012	394	28025	1611
baisser	3,88	7,00	1,09	17,00	1,90	49,00	4,81	134,00	7,70
calmer	2,05	1,00	0,15	7,00	0,78	9,00	0,88	111,00	6,38
causer	20,01	63,00	9,85	49,00	5,50	495,00	48,60	280,00	16,10
chuter	1,46	4,00	0,62	16,00	1,79	9,00	0,88	44,00	2,53
commander	1,89	7,00	1,09	6,00	0,67	37,00	3,63	38,00	2,18
conseiller	11,38	45,00	7,04	39,00	4,38	120,00	11,80	388,00	22,30
consommer	7,25	7,00	1,09	17,00	1,90	132,00	12,90	228,00	13,10
consulter	18,22	106,00	16,50	56,00	6,29	360,00	35,40	257,00	14,70
contracter	2,60	2,00	0,31	13,00	1,46	78,00	7,67	17,00	0,97
créer	6,35	39,00	6,10	70,00	7,86	82,00	8,06	59,00	3,39
donner	32,48	152,00	23,70	229,0	25,70	371,00	36,40	767,00	44,10
découvrir	8,02	25,00	3,91	45,00	5,05	185,00	18,10	87,00	5,00
faire	150,95	556,0	86,90	774,0	86,90	1026	100,0	5742	330,0
forcer	1,73	15,00	2,34	3,00	0,33	7,00	0,68	62,00	3,56
fumer	4,28	5,00	0,78	10,00	1,12	44,00	4,32	191,00	10,90
hospitaliser	3,31	20,00	3,12	39,00	4,38	13,00	1,27	78,00	4,48
naître	1,87	8,00	1,25	9,00	1,01	20,00	1,96	57,00	3,27
nourrir	1,08	1,00	0,15	1,00	0,11	15,00	1,47	45,00	2,58
nuire	3,83	9,00	1,40	5,00	0,56	20,00	1,96	199,00	11,40
opérer	4,89	4,00	0,62	10,00	1,12	55,00	5,40	217,00	12,40
penser	18,65	19,00	2,97	38,00	4,26	80,00	7,86	1035,00	59,50
peser	1,02	7,00	1,09	4,00	0,44	8,00	0,78	31,00	1,78
placer	4,35	13,00	2,03	34,00	3,81	82,00	8,06	61,00	3,50
poser	12,61	64,00	10,00	74,00	8,31	93,00	9,14	400,00	23,00
prescrire	6,97	35,00	5,47	58,00	6,51	67,00	6,58	162,00	9,31
produire	9,61	79,00	12,30	48,00	5,39	139,00	13,60	124,00	7,13
provoquer	10,23	38,00	5,94	79,00	8,87	181,00	17,70	146,00	8,39
soigner	2,04	3,00	0,46	0,00	0,00	15,00	1,47	108,00	6,21
souffrir	7,65	25,00	3,91	12,00	1,34	87,00	8,55	293,00	16,80
soulager	0,91	4,00	0,62	2,00	0,22	15,00	1,47	23,00	1,32
soutenir	1,63	7,00	1,09	2,00	0,22	29,00	2,85	41,00	2,35
supporter	1,86	6,00	0,93	5,00	0,56	5,00	0,49	95,00	5,46
tester	1,79	12,00	1,87	7,00	0,78	13,00	1,27	56,00	3,22
transformer	1,19	4,00	0,62	20,00	2,24	13,00	1,27	11,00	0,63
trouver	19,37	87,00	13,60	48,00	5,39	149,00	14,60	764,00	43,90
venir	11,96	9,00	1,40	21,00	2,35	62,00	6,09	662,00	38,00
être	483,2	4260	666,0	4241	476,0	3130	307,0	8432	484,0

La disproportionnalité des fréquences verbales, observée entre les corpus experts et non-experts, peut certes être interprétée comme un indicateur du statut spécialisé de certains verbes, mais il est important de souligner que ce principe ne s'applique pas de façon systématique. En effet, il advient parfois que certains verbes aient une fréquence élevée dans un corpus sans pour autant avoir un lien particulier avec ce dernier (cf. section 4.4.2). De la même manière, un verbe peut avoir une faible fréquence dans un corpus, mais avoir un sens/emploi très spécialisé dans ce corpus. Les verbes *rapporter* et *couvrir* ont chacun dans le corpus PRO une fréquence normalisée supérieure à la moyenne (respectivement 18,3/6,94 et 3,12/2,5), alors qu'ils ne sont pas spécifiques aux textes des experts au même titre que des verbes tels que *observer, relever* et *détecter* par exemple. D'après nos observations, dans le corpus des experts, les verbes *rapporter* et *couvrir* opèrent dans des constructions syntaxico-sémantiques qui relèvent de la langue générale, tandis que les trois autres tendent à intervenir dans des contextes spécialisés.

Le contraste que font ressortir les tableaux 4.3 et 4.4 lorsqu'ils sont comparés, permet de percevoir clairement ce qui pourrait être l'indication des choix préférentiels de certains verbes pour certains corpus particuliers. En effet, aucun des verbes du tableau 4.3 ne se retrouve dans le tableau 4.4. La tendance est telle que les verbes fréquemment utilisés dans les corpus des experts ont une faible fréquence dans les corpus des non-experts, en particulier le corpus des forums. Un constat similaire est effectué lorsque l'on part des corpus des non-experts vers ceux des experts. Par exemple, l'apparition des verbes *évaluer*, *présenter*, *observer* et *recommander* dans le tableau 4.3 marque leur forte implication dans les textes pour experts et leur faible intervention dans les corpus VUL et FOR. Toutefois, les résultats de cette étude ont permis de découvrir dans nos corpus une troisième catégorie de prédicats dont il est difficile de cerner les choix préférentiels sur la base de la fréquence.

3. Le troisième ensemble est constitué de verbes dont les fréquences entre les deux grands types de sous-corpus (experts vs. non-experts) ne présentent pas une grande différence. Elles sont soit inférieures, soit supérieures à la moyenne. Le tableau 4.5 fournit quelques exemples.

D'après nos observations, cette catégorie contient des verbes de diverses natures qui correspondent à des typologies déjà existantes (Lerat, 2002 ; Lorente, 2002) :

— des prédicats appartenant au vocabulaire scientifique, que Lorente (2002) appelle les *verbes terminologiques*, c'est-à-dire qu'ils ont un sens spécifique dans un domaine spécialisé, pour le cas d'espèce le domaine médical : *contaminer injecter, doser, alimenter,* etc. Contrairement à Lorente (2002), nous pensons que ce type de verbes ont un sens lié au domaine de connaissance dont ils proviennent.

TAB. 4.5 – Les verbes ayant un schéma fréquentiel similaire dans PRO et FOR.

		PRO		ETU		VUL		FOR	
verbe	moy.	freq	freqN	freq	freqN	freq	freqN	freq	freqN
absorber	2,34	6	0,93	6	0,67	62	6,09	29	1,66
alimenter	1,37	5	0,78	10	1,12	23	2,26	23	1,32
apparenter	1,09	13	2,03	8	0,89	9	0,88	10	0,57
attarder	0,41	5	0,78	2	0,22	3	0,29	6	0,34
avancer	2,81	24	3,75	15	1,68	23	2,26	62	3,56
avertir	1,71	6	0,93	6	0,67	42	4,13	19	1,09
causer	20,01	63	9,85	49	5,5	495	48,6	280	16,1
cesser	3,85	16	2,5	14	1,57	79	7,76	62	3,56
contaminer	1,39	5	0,78	10	1,12	30	2,95	12	0,69
doser	4,56	43	6,72	52	5,84	5	0,49	90	5,17
déceler	1,42	15	2,34	10	1,12	8	0,78	25	1,43
déposer	1,87	8	1,25	5	0,56	56	5,5	3	0,17
détruire	1,76	2	0,31	12	1,34	46	4,52	15	0,86
encourir	0,49	3	0,46	8	0,89	4	0,39	4	0,23
engendrer	1,87	9	1,4	17	1,9	32	3,14	18	1,03
former	5,79	23	3,59	74	8,31	76	7,47	66	3,79
influer	0,96	5	0,78	8	0,89	17	1,67	9	0,51
injecter	1,05	5	0,78	8	0,89	21	2,06	8	0,46
mettre	35,75	264	41,3	222	24,9	413	40,6	631	36,2
mériter	1,01	10	1,56	6	0,67	4	0,39	25	1,43
protéger	4,11	9	1,4	11	1,23	129	12,6	21	1,2
provoquer	10,23	38	5,94	79	8,87	181	17,7	146	8,39
présumer	0,24	4	0,62	0	0	1	0,09	4	0,23
prévenir	2,36	15	2,34	20	2,24	39	3,83	18	1,03
subir	4,72	43	6,72	28	3,14	35	3,44	97	5,57
surveiller	2,38	21	3,28	9	1,01	28	2,75	43	2,47
toucher	5,49	26	4,06	36	4,04	97	9,53	75	4,31
transformer	1,19	4	0,62	20	2,24	13	1,27	11	0,63
transmettre	3,96	26	4,06	13	1,46	97	9,53	14	0,8
équiper	1,04	15	2,34	2	0,22	8	0,78	14	0,8

— des prédicats utilisés dans la langue générale mais ayant des emplois spécialisés dans un ou plusieurs domaines (*verbes phraséologiques* (Lorente, 2002)) : *former, subir, transmettre, surveiller, prévenir, transformer, progresser, déceler,* etc.

— des prédicats de la langue de tous les jours, qui peuvent difficilement avoir des emplois spécialisés : *toucher, cesser, encourir, avancer,* etc.

La diversité des verbes de cette catégorie permet de comprendre que les corpus des non-experts, celui des forums en particulier, peuvent eux aussi contenir des verbes médicaux et des verbes généraux en emploi spécialisé. Cette remarque, qui de prime abord pourrait surprendre, est tout à fait explicable. En effet, les forums font interagir par écrit des personnes (patients ou non) ayant des niveaux de connaissances médicales différents. Ce paramètre rend manifeste un phénomène très actuel. Grâce à l'évolution de la technologie à travers le Web, la documentation relative à la santé est de plus en plus abondante et accessible. De nombreux usagers du domaine médical développent ainsi une connaissance croissante de la médecine. Dans la littérature sur le sujet, différentes expressions sont utilisées pour faire référence à ce nouveau type de patients informés : *informed patients* (Parker, 2006 ; Gardiner, 2008), *informed consumers* (Hibbard & Peters, 2003 ; Norman & Skinner, 2006).

Les tableaux présentés dans cette section permettent également de noter la présence (parfois prédominante) dans le corpus FOR de prédicats par nature spécialisés, que l'on s'attendrait à rencontrer principalement dans les corpus des experts. Il s'agit de verbes qui ont généralement tendance à être sollicités dans le langage des patients. Les verbes *prescrire*, *injecter*, *opérer*, *diagnostiquer*, *hospitaliser*, qui tombent dans ce cas de figure, sont constamment utilisés par la plupart des locuteurs du français, même ceux n'ayant pas de connaissances particulières en médecine. Pour le verbe *hospitaliser*, le corpus des forums a une fréquence normalisée (4,48) supérieure à la moyenne (3,31), tandis que le corpus des experts enregistre une fréquence en dessous de la moyenne (3,12). Les données du tableau 4.4 indiquent un fonctionnement similaire avec le verbe *prescrire*.

À force d'être employés pour des besoins de communication quotidienne, ces verbes finissent par être progressivement intégrés à la langue de tous les jours et sont utilisés par tous, bien que certains patients utilisent très souvent les expressions et termes médicaux de façon erronée, comme l'expliquent Zeng-Treiler & Tse (2006).

Les résultats présentés dans cette partie de la thèse ont permis d'analyser les relations entre les verbes pris individuellement et les différents corpus abordés de façon singulière, à partir de la fréquence verbale. Dans la section suivante, nous allons présenter les résultats d'un point de vue plus général qui permettra de comparer les quatre corpus sur la base des groupes de verbes qui y figurent.

4.1.1.3 Relation entre les différents corpus

Les données du tableau 4.2 décrit précédemment, couplées à la figure 4.3, permettent d'étudier les rapports existant entre les quatre types de corpus à partir des occurrences des verbes. En observant ces données, l'on perçoit un rapprochement apparent entre les corpus PRO et ETU, et une distance entre PRO et FOR. La plupart des verbes qui ont une fréquence élevée ou une fréquence faible dans PRO ont un fonctionnement similaire dans ETU. Ces remarques sont matérialisées dans la figure 4.3 qui décrit la courbe fréquentielle de l'ensemble des verbes du corpus dans chaque variété de textes. La proximité entre les corpus PRO et ETU est symbolisée par la similarité de leurs courbes. De même, l'écart existant entre PRO et FOR (ainsi que ETU et FOR), qui constituent les deux extrémités du continuum de textes de notre corpus, est capturé par la dissemblance des courbes qui décrivent leurs fréquences verbales : observons le fonctionnement des verbes *provoquer, procurer* et *présager*. Ces constats ne sont pas surprenants, d'autant plus que les textes des corpus PRO et ETU sont écrits par des experts, respectivement pour des experts et futurs experts, tandis que les textes des forums sont en principe écrits par des non-experts pour des non-experts.

Le corpus de textes de vulgarisation, en tant que corpus orienté vers le grand public, a quant à lui un comportement spécial vis-à-vis du corpus FOR, comparé à l'harmonie qui caractérise les corpus des experts. Bien que sa courbe fréquentielle nous donne des indices par rapport à son orientation, le contenu du tableau 4.2, de même que la figure 4.3, montre que le fonctionnement des verbes du corpus VUL est très variable : les verbes sont tantôt proches de ceux des corpus des experts, tantôt de ceux du corpus des non-experts. Cette irrégularité débouche sur la catégorisation des verbes de VUL en deux classes :

— les verbes dont la fréquence se rapproche de celles des corpus PRO et ETU : *observer, recommander, relever, révéler, présager*, etc.

— les verbes dont la fréquence se rapproche de celle du corpus FOR : *soigner, subir, conseiller, évoquer, provoquer*, etc.

Cette divergence suscite une question fondamentale en ce qui concerne le statut du corpus VUL, qui regroupe des textes écrits par des experts pour le grand public. Du point de vue des occurrences des verbes, ce corpus est-il proche des textes de forums (FOR) ou des corpus pour experts (PRO et ETU) ? Pour répondre à cette question, nous avons réalisé un test qui permet de calculer la proximité entre les corpus PRO et FOR, à partir des verbes et de leurs nombres d'occurrences. Dans cette démarche, le corpus VUL est utilisé comme référence pour l'évaluation de son degré de proximité avec les différents corpus qui lui sont comparés. Ce test est effectué sous Excel, à partir de la liste des verbes et de leurs fréquences dans chaque corpus :

— Pour chaque verbe, dans chaque corpus (PRO, ETU et FOR), on calcule la distance

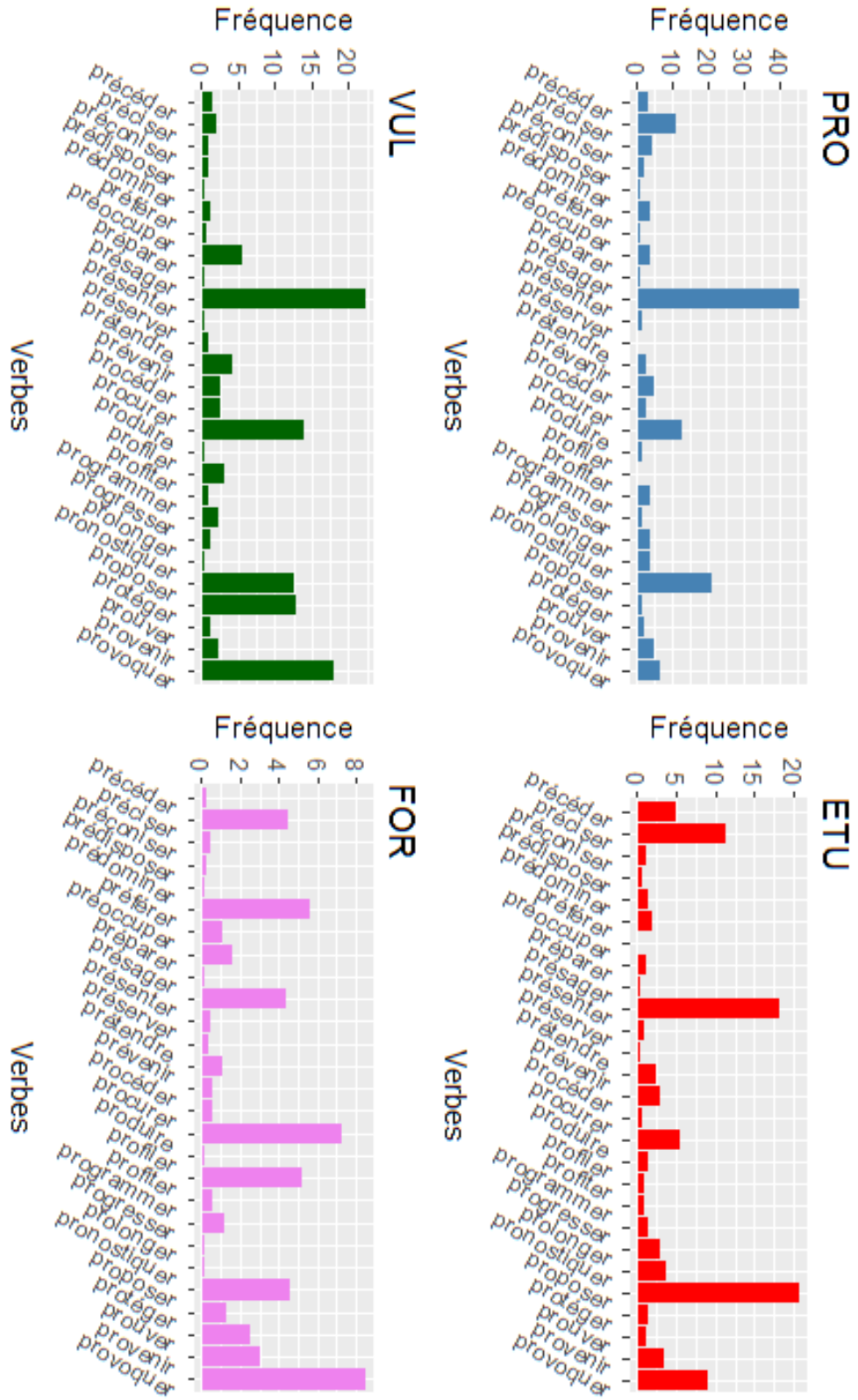


FIG. 4.3 – Distribution de 30 verbes dans les différents corpus.

(différence entre les fréquences) par rapport à la fréquence dans le corpus de référence (VUL). Cette distance est exprimée en valeur absolue.

— Les valeurs exprimant la distance des corpus par rapport à la référence sont ensuite comparées de façon à déterminer, pour chaque verbe, quel corpus est le plus proche du corpus de référence.

— Finalement, pour chaque corpus, la somme des verbes identifiés (par rapport à la référence) comme étant proches soit de PRO, soit de FOR, est calculée. Nous l'appelons *proximité* et nous la définissons comme le nombre de verbes ayant un mode de fonctionnement (fréquentiel) similaire dans le corpus de référence et dans le corpus qui lui est comparé. Ainsi, pour connaître le degré de proximité entre PRO et FOR, il faudrait faire la différence entre les valeurs qui représentent leurs *proximités* respectives par rapport à la référence.

Les résultats de cette expérience sont présentés dans les tableaux 4.6 et 4.7. La ligne *Nb vb proches de* VUL fournit le nombre de verbes (sur 2859 verbes que contient le corpus) qui expriment la proximité entre le corpus de référence (VUL) et le corpus qui lui est comparé. Le tableau 4.6 permet de comparer les corpus PRO et FOR au corpus de réference, tandis que le tableau 4.7 permet de comparer les corpus ETU et FOR au corpus de réference.

TAB. 4.6 – Proximité entre les corpus PRO, FOR et VUL, en termes de nombre de verbes (1).

Corpus	PRO	FOR
Nb vb proches de vul	1693	817
Égalité	349	
Total	2859	

TAB. 4.7 – Proximité entre les corpus ETU, FOR et VUL, en termes de nombre de verbes (2).

Corpus	ETU	FOR
Nb vb proches de vul	1652	779
Égalité	428	
Total	2859	

Les chiffres obtenus confirment les observations faites précédemment, notamment en ce qui concerne le haut degré de proximité que partagent les corpus PRO et ETU. Au total, au moins 1207/2859 verbes ont un fonctionnement similaire dans les deux corpus. Les valeurs de la proximité avec le corpus VUL sont très similaires : 1693 verbes pour PRO contre 1652 verbes pour ETU, soit une différence de 41 verbes seulement.

Par contre, la distance est grande lorsqu'on compare PRO et ETU au FOR, en prenant une fois de plus le corpus VUL comme référence. 817/2859 et 779/2859 expriment respectivement la faible proximité entre FOR-PRO et entre FOR-ETU, ce qui confirme l'analyse faite précédemment sur la distance entre les corpus des experts (PRO et ETU) et le corpus FOR. En ce qui concerne le corpus de référence (VUL), les résultats du test révèlent qu'il est, de loin, plus proche du corpus des experts que de celui des forums. Les chiffres confirment clairement cette proximité, avec 1693 et 1652 verbes marquant le rapprochement entre PRO et ETU respectivement et le corpus VUL, contre 817 et 779 avec le corpus FOR.

Vu sous un certain angle, ce résultat pourrait paraître surprenant car l'on s'attendrait à ce que le corpus VUL se rapproche du FOR, étant donné que les textes de vulgarisation sont censés s'adresser au grand public. En effet, le caractère spécial du corpus VUL réside dans le fait qu'il regroupe des textes écrits par des experts pour le grand public. L'on s'attendrait à ce que ce corpus soit une sorte de passerelle entre les principaux protagonistes du domaine médical mis en avant dans ce travail de thèse, ce qui ne semble pas être le cas. Ce constat est néanmoins très intéressant, car il met davantage en évidence l'intérêt d'une étude comme la nôtre, en ce qu'elle permettrait d'établir un véritable pont pour une meilleure communication entre les experts et les non-experts d'un domaine de spécialité.

Bien que n'étant pas négligeable, la fréquence à elle seule n'est pas un paramètre suffisant pour tirer des conclusions définitives en ce qui concerne le fonctionnement du corpus VUL et ses relations avec les autres corpus. De ce fait, cette question sera à nouveau abordée dans la section suivante à la lumière des données textuelles, en l'occurrence les patrons syntaxico-sémantiques des verbes qui décriront davantage le contenu de chaque type de corpus.

4.1.2 Annotation sémantique et acquisition des PSS

Dans cette section, nous présentons les résultats de l'annotation sémantique, c'est-à-dire l'association des catégories sémantiques de la Snomed aux structures argumentales extraites à l'étape précédente. Les patrons obtenus constituent ce que nous appelons les PSS (*Patrons Syntaxico-Sémantiques*).

Les résultats seront présentés selon deux approches : quantitative et qualitative. En procédant ainsi, nous serons à même de faire ressortir les similitudes, les divergences et les spécificités (syntaxiques et sémantiques) qui caractérisent les différents types de textes, en nous appuyant sur le fonctionnement des PSS des verbes. Des tableaux contenant des extraits représentatifs des résultats seront proposés dans cette section. Ne pouvant pas fournir dans ce document l'ensemble des résultats de notre travail de thèse, nous avons fait en sorte que les données présentées dans les tableaux de cette section puissent illustrer les différents phénomènes observés à partir des résultats obtenus. La ressource créée au terme de ce travail de thèse sera fournie sous format CD-ROM et mise à la disposition de la communauté scientifique.

4.1.2.1 Approche quantitative

Dans cette sous-section, nous présentons les résultats de l'annotation sémantique en termes de nombre. Cette quantification des données permettra de décrire et d'analyser les résultats tirés des corpus d'un point de vue numérique.

Dans un premier temps, nous allons analyser l'association terme-catégorie Snomed (désormais *terme-catsnomed*). Combien de termes ont été associés à une catégorie sémantique de la Snomed

pendant le processus d'annotation ? Combien de termes n'ont pas de catégorie Snomed ? Les chiffres obtenus sont-ils similaires dans les quatre corpus ?

Le tableau 4.8 offre une présentation sommaire des résultats de l'annotation sémantique des termes jouant le rôle d'arguments des verbes. Pour chaque corpus, il propose le nombre total d'occurrences de ces termes-arguments, le nombre d'occurrences des termes ayant été associés à une catégorie sémantique (*termes annotés, c'est-à-dire figurant dans la Snomed*), ainsi que le nombre d'occurrences de ceux n'ayant pas été associés à une catégorie sémantique (*termes non annotés, c'est-à-dire ne figurant pas dans la Snomed*). Pour chaque valeur obtenue, la proportion (en pourcentage) par rapport au nombre d'arguments nominaux (unités nominales) de chaque corpus a été calculée. Ce tableau fournit également le nombre d'occurrences des termes portant une tête multicatégorielle de la Snomed (*Occ. term. annot.*) et sa proportion par rapport au nombre d'occurrences des termes annotés. Ces informations permettront d'évaluer le résultat obtenu grâce aux méthodes de traitement des têtes multicatégorielles implémentées dans le chapitre précédent (cf. chapitre 3, section 3.2.3). Les données de ce tableau n'impliquent pas les arguments pronominaux, qui ont été mis à l'écart pour une meilleure évaluation de l'annotation sémantique des unités nominales.

TAB. 4.8 – Résultat de l'annotation des termes-arguments.

corp.	**occ. term. annot.**	**pourcent**	**occ. term. non annot.**	**pourcent**	**occ. term. annot. avec tête**	**tot**
PRO	62 852	62,85	37 137	37,14	31 845 (50%)	99 989
ETU	79 510	61,05	50 726	38,94	35 898 (45%)	130 236
VUL	86 900	58,56	61 476	41,43	38 531 (44%)	148 376
FOR	69 419	46,00	81 484	53,99	29 877 (43%)	150 903

La première remarque faite à la lecture du tableau 4.8 est que de façon générale, les scores de l'association terme-catsnomed sur l'ensemble du corpus sont raisonnables, le plus petit pourcentage de termes annotés s'élevant à 46,00% et appartenant au corpus des forums. Ces résultats évoquent une courbe qui évolue comme suit : plus on avance vers les textes des non-experts, moins on rencontre de termes Snomed.

En effet, les corpus PRO, ETU et VUL enregistrent plus de termes annotés que de termes non annotés, contrairement au corpus des forums qui fonctionne autrement, avec une prédominance d'unités nominales (arguments) non annotées. Autrement dit, les corpus des experts, étudiants et du grand public couvrent une plus grande portion de termes de la terminologie Snomed par rapport au corpus des forums qui, d'après les résultats, compte moins de termes Snomed. La similarité qu'exposent ces résultats permet de déduire que les trois premiers corpus sont plus riches en termes médicaux que le corpus des forums. Cette interprétation est tout à fait logique et a du sens, dans la mesure où ces trois corpus regroupent des textes rédigés par des experts en médecine qui sont censés maîtriser la terminologie médicale, ce qui transparaît

dans leurs écrits. Les non-experts, par contre, optent très souvent pour des noms qui leur sont familiers et qui appartiennent à la langue générale, ainsi que des expressions qui relèvent de leur parler journalier. Les deux phrases ci-dessous fournissent des exemples à travers les éléments soulignés :

7) *je tiens à dire que j'ai été un bon sportif et qu'en famille personne n'a jamais eu des problèmes de coeur mais [...] je vais ajouter que je souffrais des problèmes de digestion.*

8) *j'ai changé de cardio (et j'ai bien fait) car je voyais très bien qu'il ne m'écoutait plus et j'avais perdu confiance.*

Le constat qu'illustre ces exemples se confirme grâce au pourcentage de termes non annotés (53,99%) qu'enregistre le corpus FOR. C'est le seul corpus à avoir une telle proportion d'unités nominales non annotées, une proportion qui marque une grande distance par rapport au corpus des experts.

Les données du tableau 4.8 permettent de constater que selon le corpus, 43 à 50% des termes portant une catégorie Snomed sont des unités nominales ayant une tête multicatégorielle. Ce résultat montre que nos différentes techniques de traitement des termes avec tête multicatégorielle ont permis d'améliorer les résultats de l'annotation sémantique des corpus.

La présence d'unités nominales sans catégorie Snomed, au terme du processus d'annotation, a pour conséquence de favoriser la génération des PSS que nous caractérisons de partiels, c'est-à-dire des PSS dont certains termes-arguments demeurent sans catégories sémantiques, à l'instar de *s détecte* COD_D. Ce PSS a un argument qui ne porte pas de catégorie sémantique, en l'occurrence le sujet. Plusieurs raisons peuvent expliquer l'absence de catégorie sémantique : soit le syntagme nominal ne fait pas partie de la nomenclature Snomed, parce qu'il ne s'agit pas d'un terme médical ou du fait de n'y avoir pas été inclus ; soit le terme apparaît dans la Snomed, mais avec une différence morphologique.

En plus de l'association terme-catsnomed, l'annotation sémantique a également débouché sur l'acquisition des patrons syntaxico-sémantiques. Ces PSS ont été extraits automatiquement (cf. section 3.2.4), au moyen d'un programme qui transforme des chaînes terme-catsnomed (constituant les schémas valenciels des verbes) en patrons syntaxico-sémantiques bien structurés. Suite aux transformations effectuées à la section 3.2.4, deux ensembles correspondant à deux types de PSS ont été acquis : les PSS spécifiques (s_D affecte cod_S) et les PSS génériques (s_D cod_S). Un PSS générique renvoie à la forme non instanciée d'un patron syntaxico-sémantique. Il laisse la possibilité à différents verbes d'y figurer, d'où le terme *générique*, tandis qu'un PSS est dit spécifique lorsque la position réservée au verbe est instanciée. Le tableau 4.9 fournit les résultats obtenus pour chaque catégorie. Afin de normaliser les chiffres obtenus pour qu'ils soient comparables, nous avons calculé la proportion de PSS spécifiques par rapport au nombre de phrases verbales de chaque corpus (cf. tableau 4.1), ainsi que la proportion des PSS génériques par rapport au nombre de PSS spécifiques dans chaque corpus. Ces valeurs sont exprimées en

termes de ppm :

TAB. 4.9 – Types et nombres de PSS selon les corpus.

	nb pss spéc.	**proportion** (en ppm)	**nb pss gén. gen**	**proportion** (en ppm)
PRO	12775	237131,77	603	47201,56
ETU	13803	198675,78	559	40498,44
VUL	16366	196185,61	655	40021,99
FOR	15210	150133,25	432	28402,36

Au total, 836 PSS ont été extraits à partir des 4 corpus. Parmi ces PSS, 243 ont été sélectionnés pour être validés par les experts (cf. section 3.4). D'après les données (en valeurs absolues) du tableau 4.9, le corpus des experts et celui des textes de vulgarisation comptent plus de PSS que les deux autres. Par contre, les chiffres normalisés révèlent que dans ce classement, le corpus des experts est plutôt suivi par le corpus des étudiants. Cette prédominance de PSS dans les textes des experts transparaît également à travers les proportions que représente le nombre de PSS spécifiques dans chacun des corpus concernés. Le corpus des experts contient près du double du nombre de PSS du corpus des forums. Il est intéressant de remarquer que ce constat frappant entre un contraste avec les résultats de l'annotation syntaxique des corpus, qui signalent que les corpus VUL et FOR comptent le plus grand nombre de prédicats verbaux (cf. tableau 4.1). Ces résultats, qui de prime abord pourraient sembler contradictoires, traduisent le fait que les experts disposent d'un large panel de constructions verbales diversifiées que symbolisent nos PSS génériques, tandis que les non-experts sont limités en termes d'emplois verbaux, ceci en dépit de la grande variété de verbes qu'ils utilisent dans leurs écrits.

En outre, les résultats de l'annotation sémantique ont permis d'étudier le rapport entre, d'une part, les verbes et les types de PSS et, d'autre part, les corpus et les PSS. Le tableau 4.10 présente, pour chaque verbe, le nombre de PSS dans chacun des quatre corpus (*nbPSS*), ainsi qu'une valeur normalisée de ce nombre (*prop*), à savoir la proportion (sur 10 000) qu'il représente par rapport au nombre total de PSS du corpus entier. Cette proportion va permettre de comparer et d'étudier le rapport entre les types de corpus et le nombre de PSS par verbe.

Les verbes enregistrent un nombre de PSS variable, allant de 1 à plus de 100 pour les verbes les plus récurrents dans les corpus. Les données du tableau 4.10 sont un extrait représentatif (sur le plan qualitatif) des résultats. À partir de ces données, les principales observations faites sur les corpus seront présentées.

À la lecture de ce tableau, nous pouvons tout d'abord constater que de façon générale, les verbes tendent à avoir un nombre de PSS élevé dans l'ensemble du corpus. Le verbe *causer* par exemple a en moyenne 20 PSS par corpus. La présence de structures argumentales partiellement annotées (décrites à la section cf. 4.1.2.1) a certes un impact sur les chiffres que fournit le

TAB. 4.10 – Quelques verbes avec leurs nombres de PSS.

	PRO		ETU		VUL		FOR	
verbes	nbPSS	prop.	nbPSS	prop.	nbPSS	prop.	nbPSS	prop.
admettre	30	23,48	21	15,21	8	4,89	14	9,2
analyser	49	38,36	44	31,88	28	17,11	19	12,49
associer	116	90,8	137	99,25	67	40,94	17	11,18
augmenter	55	43,05	64	46,37	75	45,83	41	26,96
baisser	7	5,48	8	5,8	22	13,44	38	24,98
causer	27	21,14	20	14,49	59	36,05	44	28,93
conseiller	29	22,7	27	19,56	46	28,11	63	41,42
consulter	31	24,27	22	15,94	42	25,66	27	17,75
contrôler	33	25,83	27	19,56	27	16,5	22	14,46
créer	26	20,35	29	21,01	30	18,33	25	16,44
diagnostiquer	18	14,09	19	13,77	25	15,28	25	16,44
donner	67	52,45	86	62,31	116	70,88	107	70,35
déceler	13	10,18	10	7,24	7	4,28	16	10,52
découvrir	14	10,96	20	14,49	39	23,83	28	18,41
définir	54	42,27	47	34,05	27	16,5	6	3,94
dépister	8	6,26	14	10,14	5	3,06	1	0,66
développer	38	29,75	38	27,53	65	39,72	18	11,83
envisager	35	27,4	32	23,18	20	12,22	17	11,18
faire	114	89,24	133	96,36	159	97,15	189	124,26
hospitaliser	9	7,05	10	7,24	5	3,06	11	7,23
impliquer	35	27,4	30	21,73	29	17,72	6	3,94
imposer	26	20,35	46	33,33	21	12,83	15	9,86
inclure	48	37,57	23	16,66	29	17,72	5	3,29
indiquer	61	47,75	48	34,78	40	24,44	22	14,46
induire	22	17,22	28	20,29	14	8,55	9	5,92
mettre	86	67,32	87	63,03	109	66,6	94	61,8
montrer	48	37,57	43	31,15	37	22,61	34	22,35
nécessiter	38	29,75	46	33,33	27	16,5	12	7,89
observer	65	50,88	47	34,05	27	16,5	8	5,26
prescrire	23	18	30	21,73	37	22,61	46	30,24
produire	22	17,22	17	12,32	35	21,39	28	18,41
proposer	48	37,57	61	44,19	51	31,16	31	20,38
provoquer	12	9,39	33	23,91	51	31,16	37	24,33
préconiser	16	12,52	8	5,8	5	3,06	6	3,94
présenter	82	64,19	59	42,74	63	38,49	26	17,09
relever	22	17,22	19	13,77	12	7,33	20	13,14
souffrir	16	12,52	10	7,24	27	16,5	34	22,35
soulager	4	3,13	2	1,45	12	7,33	14	9,2
subir	19	14,87	16	11,59	21	12,83	22	14,46
évaluer	54	42,27	24	17,39	15	9,17	6	3,94

tableau 4.10, mais la multiplicité des PSS provient en grande partie du degré de granularité de la classification que propose la terminologie Snomed.

Lors de la description de la Snomed (cf. section 2.2.2), nous avons vu que cette ressource compte 11 classes sémantiques distinctes grâce auxquelles les termes des corpus sont catégorisés selon leurs natures. Pendant le processus d'annotation sémantique, ces catégories sémantiques ont la possibilité de se combiner à au moins 4 rôles syntaxiques différents (sujet, COD, COI, complément d'agent), ce qui débouche sur de nombreuses combinaisons sémantiques possibles pour un seul verbe. Prenons par exemple le patron valenciel *sujet provoque* COD. Nous allons ci-dessous énumérer les différentes combinaisons de catégories sémantiques obtenues à partir des catégories Snomed et des fonctions syntaxiques sujet et COD uniquement. Les exemples proposés proviennent des corpus :

1. D *provoquer D* : *Les troubles de digestion et les brûlures à l'oesophage peuvent provoquer des douleurs*.
2. D *provoquer F* : *Cette maladie peut provoquer l'élévation de la température du patient*.
3. F *provoquer D* : *L'activité électrique du coeur provoque les contractions cardiaques*.
4. F *provoquer F* : *La stimulation nociceptive provoque une vasoconstriction intense musculo-cutanée et une hypertension artérielle*.
5. P *provoquer D* : *Le sondage au stylet provoque un écoulement purulent [...]*.
6. P *provoquer F* : *Une ablation par inadvertance de tissu cardiaque ou vasculaire peut provoquer des saignements [...]*.
7. L *provoquer D* : *Ce virus peut provoquer une maladie grave*.
8. L *provoquer F* : *Le pollen provoque des allergies chez certains sujets*.
9. C *provoquer F* : *Ce médicament peut provoquer des étourdissements*.
10. C *provoquer D* : *Le dorzolamide provoque des anomalies vertébrales et sternales chez le lapin*.
11. A *provoquer D* : *Les objets rouillés peuvent provoquer un tétanos*.
12. A *provoquer F* : *Une seringue non stérilisée peut provoquer des réactions chez le patient*.
13. J *provoquer F* : *Le médecin a dû provoquer l'accouchement*.
14. S *provoquer* COD : *Cet homme me provoque sans cesse*.
15. P *provoquer A* : *Le test de 1 millivolt que l'on envoie dans l'appareil doit provoquer un signal rectangulaire [...]*.
16. F *est provoqué par* J : *L'accouchement est provoqué par la sage-femme en cas de risque de décès de l'enfant*.

Au total, on compte jusqu'à 16 PSS acquis à partir de 6/11 catégories Snomed (A, C, D, F, P, S) et 2 fonctions syntaxiques seulement. Remarquons également que les exemples proposés illustrent une seule structure syntaxique parmi plusieurs constructions possibles : il s'agit de la construction transitive directe. Ces chiffres reflètent bien la forte propension des PSS à se multiplier autour d'un seul verbe. Cet exemple permet également de mieux cerner le phénomène d'abondance de PSS (cf. tableau 4.10) qui, comme l'on peut le constater, est en grande partie la conséquence de la forte granularité de la classification Snomed. Cette particularité des catégories Snomed est d'un apport déterminant dans notre étude. Elle offre la possibilité de fournir une description profonde des sens des verbes du corpus, à travers l'étiquetage sémantique de leurs arguments.

La granularité de la Snomed nous a ainsi permis de capturer les subtilités qui font la différence entre des PSS syntaxiquement identiques, régis par le même verbe, comme le montrent les exemples suivants :

9) *Mon mari a subi quatre opérations.*

10) *Le patient a subi un AVC.*

Les phrases 9 et 10 ont la même structure syntaxique (sujet-verbe-COD). Sur le plan sémantique, les termes qui instancient la position de sujet ont les mêmes propriétés. Il s'agit de sujets animés, plus précisément humains. Les termes qui jouent le rôle de COD, notamment *opération* et AVC partagent également les mêmes propriétés sémantiques : non humains, non animés. En se basant sur toutes ces informations, il est impossible de différencier le sens du verbe entre les phrases 9 et 10. En revanche, l'association de catégories Snomed aux arguments du verbe fait transparaître la différence :

— *Mon mari_*S *a subi quatre opération_*P.

STATUT SOCIAL (c.-à-d. patient) subit PROCÉDURE → 'passer'

— *Le patient_*S *a subi un AVC_*D.

STATUT SOCIAL subit MALADIE → 'souffrir'

D'après la Snomed, *opération* appartient à la classe PROCÉDURE, tandis que AVC correspond à la catégorie MALADIE. Cette distinction permet de discriminer le sens du verbe. Dans la phrase 9, *subir* signifie 'passer', tandis que dans la phrase 10, il signifie 'souffrir'. Ainsi, les PSS résultant de cette étude (cf. tableau 4.13), acquis à partir des catégories Snomed, peuvent fonctionner comme des règles décrivant le sémantisme des verbes. D'ailleurs, c'est de cette façon que nous avons procédé afin d'identifier les emplois verbaux spécialisés candidats pour la simplification. Grâce à cette méthode, plus de 200 PSS ont ainsi été sélectionnés pour la simplification (cf. section 4.5).

Par ailleurs, la multiplicité des PSS reflète la richesse des corpus. Les PSS mettent en relation les concepts médicaux illustrés par les termes du corpus. Si les textes d'un corpus sont riches,

alors, logiquement, il devrait en ressortir divers types de patrons qui témoignent de cette richesse. Dans l'exemple proposé ci-dessus, 16 combinaisons différentes ont été effectuées entre des concepts médicaux grâce à un seul verbe, *provoquer*.

Au-delà de l'abondance de PSS, les données du tableau 4.10 montrent une forte variation du nombre de PSS des verbes selon les corpus. Autrement dit, pour certains verbes, le nombre de PSS est similaire entre le corpus des experts et celui des forums (*déceler, hospitaliser, mettre*), tandis que d'autres verbes enregistrent plus de PSS dans un certain type de corpus par rapport aux autres (*faire, donner, associer, présenter, indiquer*). Dans certains cas, l'écart entre les corpus est réellement important. Cette tendance est particulièrement marquante dans le corpus PRO qui occupe la tête du classement avec plus de 130 prédicats verbaux dont le nombre de PSS enregistrés présente un écart de plus de 10 PSS par rapport au corpus des forums.

Les verbes *augmenter, associer, observer, présenter, évaluer* illustrent bien ce phénomène. *Associer* est un cas extrême, ses PSS représentant une proportion importante (90,8) dans le corpus PRO, tandis que dans le corpus FOR, ils représentent une faible proportion (11,18) sur l'ensemble du corpus.

Si l'on considère que les PSS décrivent le langage de la médecine de par leur aptitude à mettre en relation des termes désignant différents types de concepts médicaux grâce aux verbes, alors l'abondance des PSS dans le corpus des experts pourrait être perçue comme un indicateur du haut niveau de connaissances des médecins, en ce qui concerne le domaine médical. Le langage médical étant en constante évolution, de nouvelles constructions verbales voient le jour et sont utilisées au sein de la communauté des professionnels de la santé. Dans un état de l'art sur l'évolution du langage médical anglais entre le XIX^{e} et le XX^{e} siècle, Brunt (2008) décrit un ensemble d'expressions nominales et verbales très spécialisées qui naissent des échanges entre les membres d'une même équipe médicale. Au fil du temps, ces expressions, qui étaient au départ utilisées pour des besoins de communication spécifiques, se propagent et intègrent le discours médical, ce qui contribue au développement et à l'expansion du langage des experts.

Au-delà de tout ce qui a été présenté dans cette partie de notre travail, si l'on aborde les résultats de l'annotation sémantique d'un autre point de vue, force sera de réaliser que l'abondance des PSS dans un texte peut également fournir des renseignements sur le sémantisme des verbes en emploi. Cet argument fera l'objet de la section suivante.

4.1.2.2 Approche qualitative

Dans cette sous-section, il est question de présenter les résultats de l'annotation sémantique d'un point de vue qualitatif. Pour ce faire, nous allons effectuer une analyse comparative du fonctionnement des verbes sur la base du type et de la fréquence de leurs PSS dans chacun des quatre sous-corpus. En procédant de la sorte, nous serons à même de faire ressortir les différents types de variation (syntaxique, sémantique et lexicale) qui caractérisent le fonctionnement des

verbes dans les corpus faisant l'objet de cette étude. Comme dans les sections précédentes, la sélection des verbes et des PSS qui seront présentés en exemples a été déterminée par leur aptitude à illustrer les différents phénomènes observés à travers les résultats de l'annotation sémantique.

Les tableaux 4.11 et 4.12 fournissent une liste de PSS associés à leurs nombres d'occurrences dans les différents corpus. Pour chaque verbe, la répartition des PSS à travers les sous-corpus suit une courbe particulière et la fréquence des PSS n'est pas toujours semblable d'un corpus à l'autre. À partir de la variation de fréquence des PSS, l'on peut distinguer deux groupes principaux de verbes :

1. Les verbes dont les PSS tendent à avoir un comportement homogène dans un groupe de corpus. La majorité de leurs PSS ont tendance à avoir une fréquence élevée dans les corpus concernés. En général, l'homogénéité s'observe soit dans les corpus des experts (PRO et ETU), soit dans celui des forums. Les verbes du corpus de vulgarisation oscillent entre ces deux groupes, la courbe de fréquence des verbes étant tantôt similaire à celle des verbes du corpus des experts, tantôt à celle des verbes des forums.

   Si nous observons de près la distribution des PSS des verbes *imposer*, *administrer*, et *induire* dans les tableaux 4.11 et 4.12, force sera de constater que ces PSS tendent à être plus fréquents dans les corpus ETU et PRO, par rapport au corpus FOR dans lequel ils ont une fréquence moins importante. À l'inverse, avec des verbes comme *faire* et *souffrir*, les PSS tendent plutôt à avoir une fréquence élevée dans le corpus des forums. L'écart entre les corpus est parfois très marquant. Par exemple, le PSS *s_S fait de* COD_*D* est présent 580 fois dans le corpus des forums, contre 13 occurrences seulement chez les experts. Ceci représente respectivement 10,23% et 2,43% par rapport au nombre total d'occurrences des PSS du verbe *faire* dans chacun de ces corpus.

   Ce comportement est le prolongement de ce qui transparaissait déjà dans les résultats de l'annotation syntaxique, qui indiquaient une tendance de rapprochement entre certaines unités verbales et certains corpus. Cette tendance, manifestée à travers la fréquence des verbes en corpus, a permis de constater qu'entre autres, les verbes tels que *administrer*, *imposer*, *relever* et *évaluer* ont une proximité avec le corpus PRO, tandis que les verbes comme *faire* et *souffrir* se rapprochent du corpus des forums.

   Ci-dessous quelques exemples de PSS de *imposer* et *faire* dont les plus hautes fréquences proviennent respectivement des corpus PRO et FOR :

   11) s_P s'impose : *Dans cette situation, un traitement médicamenteux s'impose d'emblée puisque les risques d'hémorragies, cérébrales notamment, sont importants.*

   12) s_P impose s_P : *Le maintien d'une ventilation spontanée efficace impose une simple sédation du patient, par exemple par le propofol à objectif de concentration, et la réalisation d'une anesthésie locale du tractus respiratoire.*

TAB. 4.11 – Quelques PSS fréquents (1).

PSS	PRO	ETU	VUL	FOR
s_F accompagne de COI_F	35	42	31	0
s_D s'accompagne de COI_D	24	23	22	0
s_P s'accompagne COI_D	4	11	12	0
s_D s'accompagne de COI_F	3	25	18	5
s_D accompagne COD_D	4	7	16	23
s_J administre COD_C	14	3	21	0
s_P administre COI_S	12	4	12	0
s_P est administrée	29	6	24	1
s_C est administré	12	8	20	0
s_J administre COD_C COI_S	8	1	7	6
s_D affecte COD_S	9	6	22	5
s_D affecte COD_F	7	10	18	0
s_F est affectée	1	5	2	0
s_L affecter COD_T	5	2	19	1
s_F augmente COD_F	4	14	18	5
s_D augmente COD_F	18	24	41	3
s_F augmente	49	84	98	17
s_P augmente COD_F	15	30	57	6
s_F augmente COD_D	42	64	47	45
s_S est diagnostiqué	7	5	2	0
s_J diagnostique COD_D	4	3	9	16
s_J diagnostique COD_S	8	6	5	0
s_D est diagnostiquée	12	7	13	2
s_J diagnostique COD_D COI_S	4	2	10	27
s_F est diagnostiquée chez COI_S	6	2	2	8
s_D est diagnostiquée chez COI_S	9	1	4	7
s_D guérit	0	3	9	4
s_P guérit COD_D	0	8	18	18
s_S est guéri	0	0	3	4

TAB. 4.12 – Quelques PSS fréquents (2).

PSS	PRO	ETU	VUL	FOR
s hospitaliser COD	0	1	0	4
s_S est hospitalisé	4	10	8	52
s_J est hospitalise s_S	7	13	2	5
s_J faire cod_P coi_S	6	11	16	99
s_S faire cod_F	10	26	37	294
s_F se fait	20	43	17	31
s_J faire cod_P	19	24	25	260
s_P faire cod_F	3	15	4	9
s_S faire cod_D	13	26	62	580
s_P faire cod	8	6	3	28
s_S faire cod	5	1	6	31
s_S impose COD_P	8	9	4	9
s_P s'impose	16	15	4	1
s_P impose COD_P	33	44	6	9
s_D impose COD_P	10	25	1	0
s_S indique qqchse	7	4	8	20
s_P indique COD_F	11	3	7	5
s_C est indiqué	32	7	11	12
s_P indique COD_D	14	11	5	4
s_P est indiquée	24	38	12	1
s_A est indiqué	8	3	1	0
s_D induit COD_D	7	4	6	2
s_F induit COD_P	3	2	0	0
s_D induit COD_F	7	7	0	1
s_D est induite	5	11	1	0
s_S souffre	0	0	1	6
s_S souffre de COI_F	3	2	9	65
s_S souffre de COI_D	8	3	26	76
s_L souffre de COI_F	4	1	2	6
s souffre	2	2	6	92

13) s_S souffre de COD_D : *j'ai 27 ans et depuis 6 ans environ je souffre d'extrasystoles qui sont apparues d'un coup*.

14) s_S souffre de COD_F : *Depuis 3 ans je souffre des douleurs thoraciques qui remontent vers le dos et irradient au bras gauche*.

D'après nos observations, les verbes dont la fréquence signale une proximité avec le corpus PRO sont généralement des prédicats qui ont des emplois spécialisés dans le discours médical. Quant à ceux qui se rapprochent du corpus des forums, il s'agit principalement de prédicats de la langue générale (*faire, souffrir*). De façon inattendue, certains verbes par nature spécialisés (*hospitaliser, guérir*) présentent une forte fréquence dans le corpus des forums. En général, il s'agit de verbes ayant intégré la langue courante à force d'être utilisés dans le langage quotidien. La forte présence de ce type de verbes traduit aussi le type de préoccupations qu'ont les personnes intervenant sur les forums.

2. Les verbes dont les PSS présentent un fonctionnement hétérogène à travers les différents corpus. Pour chacun de ces verbes, la répartition des PSS à travers les sous-corpus est très souvent irrégulière. De par les changements observés au niveau de la fréquence, certains PSS tendent à se rapprocher des corpus experts, tandis que d'autres semblent être liés à celui des non-experts. D'autres PSS ont des fréquences qui tendent à être équitablement réparties entre le corpus PRO et le corpus FOR. C'est le cas du PSS s_F augmente COD_D qui a 42, 64, 47 et 45 occurrences dans les corpus PRO, ETU, VUL et FOR respectivement (cf. tableau 4.11).

Un prototype de cette deuxième catégorie de verbe est *diagnostiquer*. Il a au total 10 PSS communs aux corpus PRO et FOR, parmi lesquels 6 sont plus fréquents dans FOR et 4 dans PRO. Le tableau 4.11 contient quelques uns de ces PSS.

L'analyse des PSS des verbes à travers les différents corpus a permis de relever trois types de variations qui caractérisent le fonctionnement des verbes : la variation syntaxique, qui est très souvent associée à une variation sémantique, qui passe généralement par une variation lexicale.

1. Variation syntaxique

La variation syntaxique porte sur les différentes structures argumentales qui caractérisent les corpus, ainsi que les alternances de ces structures et les éventuels changements sémantiques qu'elles imposent au verbe. Cette partie de la thèse s'intéresse donc aux préférences des corpus en termes de constructions syntaxiques des verbes.

a. Alternance passif/actif

Le tableau 4.13 donne un récapitulatif de la répartition des PSS dans le corpus, en mettant en évidence l'alternance des constructions syntaxiques qui y sont retrouvées. Pour chaque corpus,

le nombre de PSS passifs et actifs est fourni, de même que le pourcentage de ce nombre par rapport à l'ensemble des constructions que compte le corpus en question.

TAB. 4.13 – Distribution syntaxique des PSS dans les corpus.

	Forme passive		**Forme active**			
			PSS en "on"		autres	
	nb PSS	prop (%)	nb PSS	prop	nb PSS	prop
PRO	3159	23,98	1237	9,39	8777	66,63
ETU	2050	16,25	1220	9,67	9346	74,08
VUL	1075	7,87	1496	10,95	11092	29,74
FOR	338	2,81	3597	29,95	8077	67,24
Tot	6622		7550		37292	

Comme l'indiquent les données du tableau 4.13, les PSS obtenus comme résultat de l'annotation sémantique des corpus se répartissent principalement autour de deux types d'alternance de structures syntaxiques : la voix active et la voix passive.

Le corpus des experts et celui des étudiants enregistrent les plus grands nombres de constructions passives, qui représentent respectivement 23,89% et 16,25% de l'ensemble des PSS de chacun de ces corpus, tandis que le corpus des forums ne contient que 2,81% de constructions passives sur l'ensemble de ses PSS. Ces chiffres traduisent l'inclination des experts pour la forme passive. Ceci est une caractéristique bien connue des textes scientifiques (Biber & Conrad, 2009 ; Todirascu *et al.*, 2012). Cette forme passive est très souvent marquée par l'omission de l'agent, comme le montrent les exemples suivants :

15) *En cas de cécité unilatérale ou d'énucléation, le champ visuel est évalué sur l'oeil indemne.*

16) *Zeftera est indiqué pour le traitement des infections suivantes lorsqu'elles sont causées par des souches sensibles des micro-organismes désignés chez des patients de 18 ans et plus.*

17) *Après 6 mois un relais par AVK peut être envisagé en fonction de l'évaluation bénéfice-risque.*

18) *Les essais cliniques contrôlés non randomisés ont été examinés aux seules fins d'évaluation des effets néfastes.*

En plus des constructions passives, les corpus rédigés par des experts (c'est-à-dire PRO, ETU et VUL) sont caractérisés par la présence d'une forme particulière de construction active ayant un sujet indéfini, le pronom *on*.

19) *En conséquence, si l'on veut développer l'ETP, cela doit impérativement se faire dans le cadre d'une stratégie globale visant à rendre cohérents les différents vecteurs possibles de l'offre d'ETP et à garantir la qualité de l'ETP dispensée.*

20) *La tolérance est excellente, les effets secondaires très faibles et fort peu différents de ce qu'on observe sous placebo.*

21) *On a évalué l'ÉCG comme méthode pour prévenir la mort cardiaque subite chez des athlètes des États-Unis, à un coût estimatif de 44 000 $US par année de vie sauvée (36).*

22) *Les modèles rajustés ont également révélé que le risque d'issue défavorable était significativement plus grand chez les sujets qu'on n'avait pu examiner pour détecter une anomalie de la démarche.*

Ce constat n'est cependant pas propre aux textes des experts uniquement. Bien au contraire, d'après les résultats du tableau 4.13, le corpus des forums contient le plus grand nombre de constructions en *on*, au total 3597, qui représentent 29,95% de l'ensemble de ses PSS. On constate que l'écart est énorme entre le pourcentage de PSS à la forme passive et celui des PSS en *on* dans le corpus des forums. Cette remarque n'est pas surprenante étant donné que le pronom *on* est communément utilisé dans les registres courant et familier, qui correspondent aux styles de prédilection dans la plupart des discussions de forums, d'autant plus que ces dernières font interagir des personnes de statuts socio-professionnels différents ayant, en général, pour principal objectif d'obtenir des réponses à leurs préoccupations et/ou de partager leurs expériences.

Nos analyses permettent de retenir que l'utilisation de la construction en *on* joue un rôle différent selon qu'on a affaire aux textes rédigés par des experts pour des experts, ou à ceux écrits par des non-experts pour des non-experts. Dans le premier type de textes, le recours à la construction en *on* et/ou à la construction passive avec agent omis semble avoir un but principal, celui de maintenir l'information véhiculée impersonnelle, d'octroyer un caractère objectif aux écrits (Heslot, 1983 ; Candel, 1984 ; Mortureux, 1991 ; Fleischman, 2003), afin de leur donner une valeur de vérité générale. Par ailleurs, à travers les exemples proposés, nous constatons également que certaines structures avec *on* renvoient à des agents collectifs ou à la communauté médicale.

Dans les textes de forums, la plupart du temps, l'emploi des constructions en *on* relève du registre familier, comme le montrent les exemples 23 et 24.

23) *Chez moi, on ne fume pas, si on veut fumer, on va dehors, même sous la pluie, aucune pitié.*

24) *Moi je trouve que c'est vraiment important d'écouter et de faire des conseils qu'on te donne dans ces centres.*

25) *On m'a détecté y a 4 ans un WPW et j'ai quelques mois après subit une ablation paradiofréquence.*

Par contre, dans certains contextes comme dans l'exemple 25, le *on* tend à jouer plus ou moins le même rôle que dans les corpus experts, celui d'omettre l'agent et de rendre le discours

impersonnel. Ici, par contre, le but de l'impersonnalisation n'est pas de rendre le discours plus objectif mais, plutôt, de focaliser le discours ou encore de mettre l'accent sur une information particulière que l'on voudrait donner. Dans le cas d'espèce, il semblerait que l'information importante pour le locuteur (le patient) est ce qu'on lui a détecté et non qui l'a détecté. L'omission du sujet n'a donc aucun impact sur la transmission du message qu'il souhaite véhiculer, mais contribue à faire une focalisation sur l'objet principal de son message. Cette action de focaliser le discours joue un rôle important dans notre étude, d'autant plus qu'elle rejoint la notion de point de vue qui caractérise nos corpus et qui a joué un rôle central pendant l'alignement (cf. section 4.5.2). Ce paramètre, encore appelé *la voix* (Fleischman, 2003), fait partie des propriétés discursives des textes médicaux dont nous avons parlées dans le premier chapitre de notre travail (cf. section 1.1.3).

Toutes ces informations viennent remettre en question une hypothèse avancée au début de cette thèse, selon laquelle l'une des principales sources de divergences entre les données extraites de nos deux corpus principaux serait le style qui les caractérise : style informel dans les forums vs. style d'écriture formel chez les experts. En effet, les différents éléments que nous avons présentés jusqu'ici montrent que beaucoup de paramètres sont en jeu :

— le rôle social : le médecin est celui qui sait, qui apporte l'information, tandis que le patient est celui qui est principalement concerné par l'information et qui la reçoit.

— le niveau d'implication : le médecin n'est pas concerné directement mais doit se protéger (sensibilité, erreurs de diagnostic ou de prescription, etc.). Le patient est concerné (directement ou indirectement) par l'information donnée ; il se focalise sur ce qui est le plus important pour lui et transmet l'information de façon à mettre cela en évidence, etc.

Ces paramètres se retrouvent en quelque sorte dans les différences ou choix stylistiques que présentent les textes des deux groupes de protagonistes : les constructions impersonnelles vs. certains types de *on*, les choix lexicaux spécifiques (emploi de verbes comme *souffrir, guérir, hospitaliser* vs. non-emploi ou emploi d'autres verbes), etc. Les résultats de la section 4.5.2 fournissent davantage d'éléments allant dans ce sens. La syntaxe est également concernée par ces différences, ce qui fera l'objet de la partie suivante.

b. Changement au niveau de la structure argumentale

L'analyse de la syntaxe des PSS a également permis d'avoir une vision panoramique de la variation des structures argumentales des verbes dans l'ensemble du corpus. Le recensement des PSS de chaque verbe dans l'ensemble du corpus permet de relever les variantes de la structure argumentale sous-jacente à plusieurs PSS. Le PSS J *verbe* D *chez* S[4] peut servir d'illustration.

4. J : Job (métier : médecin, dentiste, etc.) ; S : Statut social (patient, femme, enfant, etc.)

Lorsqu'elle accueille les verbes *diagnostiquer* et *découvrir*, la structure argumentale de ce PSS permet d'acquérir respectivement 7 et 6 variantes à partir de notre corpus. Ci-dessous les variantes de ce PSS avec des exemples, pour le verbe *découvrir*.

— J découvre S D : *Aprés avoir appelé le médecin a domicile (tachycardie + + +) ou il m'a découvert une hypertension a 20*.

— J découvre D chez S : *Je m'appelle tony, j'ai 30 ans et l'on a découvert chez moi à 19 ans une insuffissance mitrale par prolaspsus valvulaire (maladie de barlow).*

— J découvre D : *Les médecins ont découvert une dilatation de l'aorte ascendante*.

— D est découvert chez S : *La CMD est souvent découverte chez un sujet en IC grave évoluant vers l'apparition d'une fuite mitrale fonctionnelle, de troubles du rythme ventriculaire et d'un tableau de bas débit cardiaque*.

— D est découvert par J : *Les cardiopathies seront alors découvertes par un médecin, lors de l'examen du patient ou lors de la réalisation d'un examen complémentaire*.

— D est découverte : *la glycémie à jeun pour les sujets dont le diabète est découvert lors de l'EPS*[5].

Les exemples ci-dessus montrent la productivité des PSS qui génèrent différents schémas valenciels représentant les variantes syntaxiques du patron de base. Cette variation de la structure argumentale est souvent accompagnée d'une variation au niveau du sens du verbe : c'est ce que nous appelons *variation sémantique*. Dans notre pipeline de travail, l'étude de la variation syntaxique a été d'une importance capitale pour l'identification du type de constructions syntaxiques dont l'utilisation par les experts pourrait participer à créer ou amplifier la difficulté de compréhension du sens des verbes par des lecteurs non experts. Dans cette démarche, la forme passive a été identifiée comme une source potentielle de difficulté de lecture, surtout lorsqu'elle est utilisée de façon régulière dans le texte. Cette construction constitue donc l'un des éléments qui seront pris en considération dans le processus de simplification des PSS.

2. Variation sémantique

La barrière n'est pas étanche entre la variation syntaxique et la variation sémantique qui opèrent dans le fonctionnement des verbes en corpus. En effet, d'après nos observations, la syntaxe joue un rôle important dans le sémantisme des PSS verbaux. Elle détermine très souvent leurs sens, comme nous allons le voir dans cette partie de notre travail.

Les résultats de l'annotation sémantique permettent de mettre en évidence la variation sémantique autour d'une structure argumentale. Grâce aux catégories sémantiques de la

5. Dans cet exemple, le syntagme prépositionnel *lors de l'EPS* fonctionne comme un circonstanciel temporel, mais sur le plan sémantique, il est très proche d'un complément en *par*.

Snomed, il est possible d'acquérir plusieurs patrons sémantiques différents (c'est-à-dire des PSS) à partir d'un seul schéma syntaxique ou d'une seule structure argumentale. Le verbe *diagnostiquer* par exemple a 43 PSS (types + variantes) dans l'ensemble du corpus, parmi lesquels les trois cités ci-dessous. Les deux premiers interviennent dans le corpus des experts, tandis que le dernier est fréquent dans le corpus des forums :

26) S *est diagnostiqué (c.-à-d.* J *diagnostique* S*)* : *L'intérêt d'un traitement précoce de l'HTAP n'est pas démontré et la majorité des patients est diagnostiquée tardivement, en classe fonctionnelle III-IV.*

27) D *est diagnostiquée (c.-à-d.* J *diagnostique* D*)* : *Les malformations cardiaques sont présentes dans 50-75 % des patients et sont généralement diagnostiquées tôt dans la petite enfance.*

28) J *diagnostique* D S *(c.-à-d.* J *diagnostique* D *chez* S*)* : *Une des raisons pour laquelle la MTE n'est probablement pas diagnostiquée chez la personne âgée est la présentation clinique souvent atypique dans cette population.*

Comme on peut le remarquer, les deux premiers PSS ci-dessus sont caractérisés par une présence implicite de l'argument sujet et d'un complément que nous considérons comme un COI. Nos analyses nous ont permis d'identifier un schéma valenciel spécial, assez fréquent, autant dans les textes des experts que ceux des non-experts ; il s'agit de *sujet verbe complément1 chez complément2*. Ce schéma est particulièrement employé avec certains verbes appartenant à la classe des verbes d'observation, qui à la base sont de valence 2 : *diagnostiquer, découvrir, observer, détecter*, etc. Le second complément est particulier car il est parfois introduit par la préposition *chez* qui pousse à croire qu'il s'agit d'un circonstanciel. Mais, en réalité, ce n'est pas le cas puisque sa présence (même implicite) est requise pour la réalisation du sens du verbe dans cet emploi. Prenons quelques exemples[6] :

29) *Le SIDA a été découvert dans les années 1970.*

30) *Mon médecin m'a découvert une hépatite.*

31) *On découvre ce type de cancer chez les sujets âgés.*

Dans les phrases 30 et 31, le verbe a une interprétation différente ('détecter') de celle de la phrase 29 ('apparaître'), ceci de par la présence du second complément (*m', les sujets âgés*). Nous considérons ce complément comme un COI. Par conséquent, dans ce schéma valenciel, les verbes sont considérés comme ayant une valence 3 (*sujet verbe cod chez coi* ou bien *sujet verbe coi cod*).

6. Ces phrases ne proviennent pas de notre corpus mais elles illustrent les schémas valenciels observés en corpus.

Pour revenir aux PSS susmentionnés, ceux des exemples 27 et 28 représentent des variantes d'un seul schéma syntaxique de base (*sujet-verbe*-COD-COI) qui est illustré par le PSS de l'exemple 28. Cependant, sur le plan sémantique, il est intéressant de remarquer qu'il existe une différence de sens non négligeable entre le PSS de l'exemple 26 et les deux autres. Cette différence est marquée par le changement de catégorie sémantique de l'argument COD. Lorsqu'on dit qu'un *médecin diagnostique un patient*, cela signifie que *le médecin examine et/ou identifie le patient (comme étant porteur d'une maladie)*. En d'autres termes, il ausculte le patient et lui découvre une maladie. Tandis que *médecin diagnostique maladie* veut dire que le médecin *détecte, décèle, découvre cette maladie* (chez le patient). Le test de synonymie permet de mettre en évidence la nuance qui existe entre ces deux emplois du verbe *diagnostiquer*. En effet, une maladie peut être identifiée par un médecin, mais il n'est pas naturel, à la place de *diagnostiquer* de dire qu'un patient est détecté/décelé/découvert (comme souffrant d'une maladie) par un médecin. Et lorsqu'on dit qu'*un patient est examiné par un médecin*, en général, c'est pour exprimer autre chose.

Il arrive très souvent que la variation sémantique se manifeste entre les différents corpus, c'est-à-dire que des variantes d'un PSS véhiculant des sens différents du verbe interviennent dans différents corpus, avec des fréquences divergentes. C'est d'ailleurs le cas des PSS S *est diagnostiqué* et D *est diagnostiquée* qui sont fréquents respectivement dans les corpus PRO et FOR. Le PSS D *est diagnostiquée* enregistre quelques occurrences dans le corpus des experts, mais il est prédominant dans le corpus des forums où il se réalise à travers un schéma syntaxique plus explicite, à la forme active (J *diagnostique* D *chez patient*), tandis que le corpus des experts privilégie le PSS S *est diagnostiqué* qui est à la forme passive (cf. tableau 4.11).

Le tableau 4.14 présente quelques cas illustrant la variation sémantique autour d'une structure argumentale de base. Le verbe de chaque structure argumentale (construction syntaxique) est associé à plusieurs interprétations déterminées par les catégories Snomed qui caratérisent ses arguments. Les catégories Snomed sont abrégées comme suit : D : maladie, F : fonction de l'organisme, J : métier (de manière générale, médecin), P : procédure, et S : statut social (de manière générale, patient).

Le nombre de variantes sémantiques d'un PSS change d'un verbe à l'autre. Certains PSS peuvent avoir jusqu'à 4 variantes sémantiques (c'est-à-dire 4 interprétations distinctes), voire même plus, selon le degré de polysémie du verbe. C'est le cas du verbe *suivre* (cf. tableau 4.14) qui présente 4 interprétations distinctes réparties principalement entre le corpus des experts et celui des non-experts. En général, le corpus des experts compte le plus grand nombre de variantes, comme nous l'avons vu grâce aux tableaux 4.11 et 4.12.

Le changement de fréquence des variantes sémantiques des PSS d'un corpus à l'autre qui est un phénomène récursif chez plusieurs verbes, fait penser à un choix délibéré des rédacteurs d'utiliser des patrons verbaux précis pour exprimer des sens particuliers du verbe. La prédominance

TAB. 4.14 – Variation sémantique autour d'une structure argumentale de base.

Construction syntaxique	Interprétations	Exemples
su diagnostiquer **cod**		
J *diagnostique* S	'examiner'	*[...] la personne qui diagnostique et suit le patient ayant le TDAH.*
J *diagnostique* D	'découvrir'	*Mon cardio a diagnostiqué de rares extrasystoles auriculaires.*
su évoquer **cod**		
P *évoque* D	'faire penser à'	*La constatation d'hématomes [...] évoque un trouble de la coagulation associé.*
J *évoque* P	'suggérer', 'penser à'	*On évoque, devant ce diabète bronzé [...], le diagnostic d'hémochromatose génétique.*
su évaluer **cod**		
J *évalue* S	'examiner', 'contrôler'	*Le lendemain matin, le médecin B évalue la patiente et décide de lui donner congé.*
J *évalue* F	'mesurer'	*On a évalué le risque de saignements chez le rat.*
su suivre **cod**		
J *suit* S	'traiter', 'surveiller'	*Nous avons suivi pendant deux ans une cohorte de sujets atteints de démence.*
S *suit* P	'prendre'	*La solution à long terme serait que les voyageurs suivent un traitement spécialisé.*
P *suit* P	'respecter'	*La réanimation cardiorespiratoire suit les mêmes principes pour ce qui concerne la ventilation artificielle et la défibrillation.*
TEMPS suit P	'contrôler', 'venir après'	*Les patients doivent être suivis au moins une fois par an pendant les 5 ans qui suivent le traitement par radiothérapie.*

d'une variante sémantique d'un PSS dans un type de corpus, au détriment des autres variantes, signalerait donc le caractère spécifique du sens qu'a le verbe dans ce PSS, vis-à-vis du corpus en question.

On a donc affaire à des verbes employés par les experts et les non-experts, mais dans chaque communauté, au moins un sens ou une interprétation particulière du verbe prévaut sur les autres. Dans le cas de *diagnostiquer*, les experts utilisent les deux interprétations du verbe en question ('examiner' et 'découvrir'), tandis qu'un seul semble être privilégié chez les non-experts, ce qui fait toute la différence. Le fonctionnement du verbe *diagnostiquer* illustre deux phénomènes que nous nommons *polysémie intratextuelle* vs. *monosémie intratextuelle*. Par polysémie intratextuelle, nous désignons le caractère d'un verbe qui a plusieurs sens dominants dans un type de corpus. La monosémie intratextuelle, quant à elle, renvoie au caractère d'un verbe qui ne manifeste qu'un seul sens dans un corpus donné.

Les verbes dont les PSS illustrent la variation sémantique, surtout l'opposition polysémie intratextuelle vs. monosémie intratextuelle, doivent être abordés avec délicatesse. Ce caractère particulier fait d'eux des candidats potentiels pour la simplification car la polysémie crée facilement des difficultés de compréhension, le plus souvent du côté des non-experts qui, en général, sont familiers de sens bien précis du verbe, liés à leur contexte socio-culturel. Quant aux experts, ils sont susceptibles de connaître et d'utiliser les verbes en fonction de l'interprétation qu'ils souhaitent leur conférer, et ceci grâce à leurs connaissances médicales. Cette inégalité du niveau de connaissance se traduit également par le choix de constructions verbales privilégiées dans les différents textes. Tandis que les experts font usage de constructions qui octroient au verbe un sens technique et spécialisé, les non-experts ont tendance à utiliser des constructions verbales à caractère général, bien connues de la plupart des locuteurs du français.

La variation sémantique peut également intervenir entre différentes structures syntaxiques (par exemple transitive vs. intransitive). Autrement dit, le verbe peut changer de sens selon le type syntaxique de PSS au sein duquel il figure. En effet, l'analyse comparative des résultats de l'annotation sémantique a permis d'observer un phénomène assez révélateur entre les verbes, les structures argumentales et les corpus. Certaines structures argumentales semblent imposer une certaine interprétation aux verbes qui les instancient. Ce phénomène est particulièrement marquant avec les constructions transitives indirectes *sujet verbe* COI (introduit par *de*) et *sujet se verbe* COI (introduit par *de*), qui sont prisées dans le corpus des experts. Les verbes *relever* et *accompagner* peuvent servir d'illustration. Les tableaux 4.17 et 4.16 présentent quelques exemples.

Le tableau 4.15 décrit la répartition en corpus des PSS de ces verbes autour des structures syntaxiques dominantes, notamment la structure transitive directe et la structure transitive indirecte, qui est parfois couplée à la forme pronominale appelée le « se-moyen » (Zribi-Hertz, 1982), connue comme une variante du passif :

TAB. 4.15 – Répartition syntaxique des PSS de *relever* et *accompagner* dans les corpus PRO et FOR.

Constructions	**relever**		**accompagner**	
	direct	indirect	direct	indirect
EXP	9	26	4	81
FOR	15	3	5	11

Le cas du verbe *relever* est particulièrement frappant. Dans le corpus PRO, sur 36 occurrences, 26 correspondent à la structure syntaxique transitive indirecte *sujet relève de* COI, qui est pratiquement absente du corpus des forums, où l'on compte uniquement 3 occurrences. Le mode de fonctionnement du verbe *relever* est intéressant car dans cette construction transitive indirecte, il fait montre d'une forte polysémie, particulièrement dans le corpus des experts. Le tableau 4.16 (ainsi que le tableau 4.17) donne quelques PSS illustrant cette polysémie, avec les synonymes du verbe dans chaque emploi. Le symbole *x* indique que le PSS en question est utilisé dans un corpus :

TAB. 4.16 – *Relever* dans les constructions transitives directe et indirecte.

PSS	synonyme	PRO	FOR	exemples
S relève qqch	noter		x	Aujourd'hui on m' a posé un MAPA pendant 24 heures et les résultats me sembles un peu élevé , j' ai relevé plusieurs fois les chiffres.
J relève F	prendre	x	x	L'infirmière a relevé ma tension.
S se relève	se tenir debout		x	[...] il pouvait plus se relever il sentait plus ses jambes, [...].
P relève de P	être lié à	x		L'usage des techniques de RTC-3D relève donc d'une bonne pratique médicale.
S relève de D	souffrir de	x		L'exonération du ticket modérateur peut être donnée [...] lorsque le patient relève d'une affection de longue durée.
S relève de P	requérir	x		Les patients hypertendus à haut risque cardio-vasculaire relèvent d'une prise en charge globale, justifiant la prescription d'un antihypertenseur et d'une statine.

Dans le corpus des forums, par contre, la structure argumentale transitive directe est dominante, comme le montre le tableau 4.15. Dans le PSS pronominal réflexif qui est le plus fréquent chez les non-experts (15 occurrences), le verbe fonctionne comme un verbe de mouvement. Dans le PSS transitif direct, qui intervient également dans le corpus des experts (avec une faible fréquence), le verbe oscille entre les sens : 'prendre des notes', 'remarquer', 'hausser', 'souligner',

etc.

Comme l'indiquent les tableaux 4.16 et 4.17, dans le corpus des experts, le verbe *accompagner* intervient aussi fréquemment dans la construction transitive indirecte, qui s'associe très souvent à la construction « se-moyen ». Les exemples fournis dans ces tableaux peuvent appuyer ce constat :

TAB. 4.17 – *Accompagner* dans les constructions transitives directe et indirecte.

PSS	synonyme	PRO	FOR	exemples
J/S accompagne S	amener		x	Le médecin a directement accompagnée ma soeur au scanner.
D accompagne D	survenir avec	x	x	Rassure toi, les extrasystoles accompagne cette maladie un peu chiante .
F accompagne F	suivre	x	x	Chez moi, les maux de tête accompagnent toujours la fièvre.
F s'accompagne de F	entraîner	x		La prise de poids s'accompagne d'une élévation de la pression artérielle.
D s'accompagne de D	entraîner	x		Les troubles bipolaires peuvent s'accompagner d'épisodes de dépressions.

Dans le PSS *J/S accompagne S*, qui est fréquent dans le corpus des forums, le verbe a un sens de déplacement, tandis que dans la construction *F/D accompagne F/D*, le verbe exprime la succession ou la précédence entre deux éléments.

Une analyse du sémantisme des verbes *accompagner* et *relever* dans les phrases des tableaux 4.16 et 4.17 permet de se rendre compte que, dans la construction transitive indirecte, il existe un élément commun à leurs sens, qui n'intervient pas lorsque ces verbes entrent dans la construction transitive directe. Il s'agit d'un sens d'association (entre deux éléments), qui est l'interprétation sous-jacente imposée aux verbes par la construction intransitive. *Un patient qui relève d'une maladie* est un patient qui a, qui souffre de cette maladie, de même qu'une *procédure qui relève d'une autre procédure* est liée à cette dernière. Il peut s'agir d'un lien de dépendance ou d'association, de cause-effet, etc. Il en est de même pour *Une fonction de l'organisme qui s'accompagne d'une autre fonction de l'organisme ou une maladie*. Cette fonction est en quelque sorte associée à la seconde car l'une implique l'autre. De ce qui précède, nous remarquons que malgré les différents rapports de synonymie que peuvent avoir les verbes dans cette construction transitive indirecte, l'idée d'association demeure présente.

Les verbes illustrant ce type de variation sémantique s'avèrent être d'excellents candidats pour la simplification. En effet, les prédicats verbaux qui apparaissent régulièrement dans la construction transitive indirecte au sein du corpus des experts sont susceptibles d'être difficiles à interpréter, vu la forte polysémie que cette construction peut leur imposer et sachant qu'elle n'est pas fortement sollicitée par les non-experts.

3. Variation lexicale

La variation lexicale concerne les différentes unités verbales qui instancient un PSS selon les types de corpus. Le tableau 4.18 présente un ensemble de PSS équivalents (PRO vs. FOR) tirés des corpus PRO et FOR. Les verbes apparaissant au sein de ces PSS fonctionnent comme des synonymes mais qui sont cependant propres à différents corpus.

TAB. 4.18 – La variation lexicale au sein des PSS.

Verbes pro	**équivalents for**
MEDECIN dépiste MALADIE	MEDECIN découvre MALADIE
PCHIMIQUE est administré	MEDECIN donne PCHIMIQUE à PATIENT
MALADIE se traduit par MALADIE	MALADIE se manifeste par MALADIE
PATIENT présente MALADIE	PATIENT souffre de/fait MALADIE
PROCEDURE dépiste MALADIE	PROCEDURE découvre MALADIE
PROCEDURE relève de PROCEDURE	PROCEDURE dépend de PROCEDURE
PCHIMIQUE est poursuivi	PATIENT (continue de) prendre PCHIMIQUE
PATIENT relève de MALADIE	PATIENT a MALADIE
PROCEDURE est réalisée	MEDECIN fait PROCEDURE
MALADIE s'accompagne de MALADIE	MALADIE entraîne MALADIE
FONCTION s'accompagne de FONCTION	FONCTION entraîne FONCTION
MEDECIN réalise PROCEDURE	MEDECIN fait PROCEDURE

Les données du tableau 4.18 illustrent la préférence lexicale des experts, d'une part, et des non-experts, d'autre part, lorsqu'ils veulent exprimer des réalités similaires en rapport avec le domaine médical. Les phrases ci-dessous instancient certains des PSS proposés dans le tableau 4.18. Chaque phrase exemple tirée du corpus des experts est suivie d'une phrase exemple tirée du corpus des non-experts, montrant ainsi la variation au niveau du choix du verbe. L'association entre les 2 registres a été faite selon une démarche basée sur les données des corpus des non-experts et sur notre compétence linguistique (cf. section 3.5.1.2) :

32) s_C est administré : *Natrecor devrait être administré selon les conditions décrites dans la monographie de produit, en tenant compte des risques potentiels associés à l'administration de ce produit pharmaceutique*.

33) s_J donne C_COD à S : *Mon docteur voulait me donner des medicaments pour me calmer des nerfs*.

34) s_S présente D_COD : *Un des enfants a présenté une bradycardie pendant le travail.*

35) s_S fait D_COD : *j'ai 16 ans et je fais de la tachycardie type bouveret depuis 2 ans et je vais certainement me faire opérer*.

36) s_F s'accompagne de COI_F : *La prise de poids s'accompagne d'une élévation de la pression artérielle et la perte de poids d'une réduction de celle-ci*.

37) s_F est suivie de COI_F : *Je ressens comme une contraction qui est suivie de battements très rapides pendant quelques secondes et puis plus rien*.

L'analyse des exemples ci-dessus permet de comprendre que le rapprochement entre les PSS et les types de corpus relève de la préférence lexicale, qui serait à l'origine de l'attirance observée entre certains groupes de PSS verbaux et certains corpus. En effet, les résultats de cette étude nous ont permis de noter que la plupart des PSS liés au corpus des experts trouvent leurs équivalents dans le corpus des non-experts mais avec une différence au niveau du verbe utilisé : *administrer* vs. *donner* ; *administrer* vs. *faire* ; *développer* vs. *avoir* ; *relever* vs. *souffrir* ; etc. Ce constat expose les choix lexicaux des différents auteurs experts vs. non-experts. Leurs niveaux de connaissances médicales se traduisent à travers les choix variés de verbes employés dans le but d'exprimer les mêmes concepts. Tandis que les experts utilisent un lexique de verbes spécialisés spécifiques et standards pour la communauté scientifique, les non-experts utilisent des verbes généraux.

Les différences entre le corpus des experts et celui des non-experts impliquent également des variations au niveau du sens des verbes et de la syntaxe, c'est-à-dire de la valence verbale.

L'étude de la variation lexicale dans les corpus s'exprime également à travers le type de constructions syntaxiques utilisées dans les textes. Dans le corpus des forums, il arrive que des constructions verbales soient utilisées dans un contexte où les experts opteraient pour une nominalisation.

Dans ce corpus des forums, le verbe *ablater* illustre ce phénomène :

38) *J'ai été ablatée du WPW en 1993... on suspectait ton Bouveret... non....*

39) *J'ai été ablatée 2 fois en 2002 à Toulouse, 1 fois en 2003 à Bordeaux (professeur Haissaguere)*.

40) *[...] j'ai été ablatée d'un wpw il y an an et demi et depuis, plus rien... sauf quelques ES*.

41) *L'ablation de la muqueuse sinusale est pratiquée, les canaux naso-frontaux obturés et les défects osseux sont comblés avec des patchs osseux, et de la poudre d'os*.

42) *[...] L'actrice Angelina Jolie, porteuse d'un des gènes mutés, a annoncé sa double ablation des seins **réalisée** à titre préventif*.

43) *[...] chaque fois qu'un patient devait subir une ablation de la tête du pancréas pour un cancer, il lui a été proposé de participer à l'étude comparative [...]*.

Le verbe *ablater* qui est utilisé dans les trois premières phrases (tirées du corpus FOR) signifie *effectuer une ablation*[7]. Il est sollicité 39 fois dans le corpus des forums, tandis que dans le corpus des experts, autant que dans les autres corpus, aucune occurrence n'a été relevée. Par

7. http ://dictionnaire.reverso.net/francais-definition/ablater

contre, les constructions à verbe support du type *pratiquer une ablation* et *réaliser une ablation* (dans PRO et ETU), et *subir une ablation* (dans VUL et FOR également), sont rencontrées dans les autres corpus et fonctionnent comme équivalents du verbe *ablater* qu'utilisent les non-experts. Au lieu de parler d'*un patient qui a été ablaté*, les experts parleraient plutôt d'*un patient sur qui l'on a pratiqué/réalisé une ablation* ou encore d'*un patient qui a subi une ablation*.

Ce type de variation (construction à verbe support vs. verbe) reflète l'inclination des experts pour les expressions nominales afin de décrire des concepts médicaux, par opposition aux non-experts qui penchent davantage vers l'emploi de verbes (Fang, 2005). Ce constat fournit une explication supplémentaire aux résultats de l'annotation syntaxique qui signalaient une forte présence des verbes dans les corpus des forums (cf. section 4.1.1.1).

L'étude comparative des corpus autour de la variation lexicale a débouché sur la détection des paires de verbes équivalentes entre le corpus des experts et le corpus des forums. Cette tâche a ainsi contribué à la collecte des données de base pour la réalisation du travail de simplification.

4.2 Difficultés de l'annotation syntactico-sémantique

Telle que présentée dans la section précédente, l'annotation syntaxico-sémantique des corpus a été un long processus basé sur une méthode semi-automatique. L'application de cette méthode passe par l'utilisation des outils sélectionnés et une chaîne de programmes Perl qui utilisent des ressources disponibles et/ou conçues à cet effet. La mise en place de la chaîne d'annotation a été confrontée à différents défis que nous présentons dans cette section.

4.2.1 Annotation syntaxique

4.2.1.1 Le choix de l'annotateur syntaxique

Le choix de l'annotateur syntaxique a été une étape importante de notre travail de thèse. Puisque l'outil d'annotation syntaxique détermine le format et surtout le type des données à manipuler lors des différentes extractions, il était impératif de faire un choix qui prenne en considération ces deux critères. Ces exigences nous ont motivées à entreprendre un changement d'annotateur syntaxique.

Au départ, l'annotation syntaxique se faisait avec le logiciel Bonsai parser (Candito *et al.*, 2010). Cet outil propose une analyse syntaxique (des phrases) basée sur des méthodes statistiques. Comme résultat, il retourne une représentation syntagmatique de la phrase sous un format parenthésé. Les différents syntagmes constituant la phrase sont capturés grâce aux multiples

étiquettes[8], dont celles que contient la phrase suivante, qui présente un exemple de résultat de l'annotation syntaxique avec Bonsai :

> *Le traitement repose sur les dérivés thiazidiques, plus accessibles, disponibles sous forme de médicaments génériques.*
> *((SENT (NP (DET Le) (NC traitement)) (VN (V repose)) (PP (P sur) (NP (DET les) (NC dérivés) (AP (ADJ thiazidiques) (COORD (PONCT ,) (NP (DET les) (ADV plus) (ADJ accessibles)) (PONCT ,) (AP (ADJ disponibles)))) (PP (P sous_forme_de) (NP (NC médicaments) (AP (ADJ génériques)))))))))*

D'après nos observations, les résultats de l'annotation syntaxique avec Bonsai mettent bien en évidence les différents constituants de la phrase, ce qui rend possible l'extraction des syntagmes nominaux par la suite appariés aux termes de la Snomed. Ces résultats présentent toutefois d'importantes lacunes vis-à-vis des besoins que nous imposent les objectifs de ce travail de thèse. Nous parlons plus particulièrement de l'objectif de cette étape, à savoir l'extraction des structures argumentales des verbes.

- Les unités lexicales ne sont pas lemmatisées ; pourtant dans notre chaîne d'analyse des textes, cette information est indispensable pour le repérage des lemmes verbaux.
- Les relations de dépendance entre les constituants de la phrase ne sont pas explicitement marquées[9], or cette information est primordiale dans cette étape. En effet, avec le modèle d'analyse syntaxique de Bonsai, l'extraction des patrons verbaux des phrases complexes devient une tâche pénible à réaliser, puisque dans les résultats obtenus, rien ne signale l'appartenance d'un ou de plusieurs arguments à un verbe particulier.
- L'annotation de certains syntagmes nominaux complexes est problématique, ce qui conduit à des erreurs d'analyse sémantique : médicaments génériques → deux syntagmes différents : *médicaments* et *génériques*.

Toutes ces raisons[10] ont entraîné un changement d'outil d'annotation syntaxique et, comme annoncé dans la section 2.2.1 du chapitre 2, nous avons opté pour l'analyseur Cordial, dont les difficultés sont brièvement présentées ci-dessous.

4.2.1.2 Les faiblesses de l'annotateur Cordial

Dans cette partie, nous ne présenterons que de façon sommaire les principales difficultés rencontrées avec l'analyseur Cordial car le chapitre 5, qui est entièrement dédié à l'évaluation de nos outils et méthodes, fournira plus de détails sur cette question.

8. NP : syntagme nominal, PP : syntagme prépositionnel, VN : syntagme verbal
9. Cette remarque indique la principale lacune qui a motivé notre décision de changer d'analyseur syntaxique.
10. Les performances de cet outil lors de différentes campagnes d'évaluation ont également motivé notre choix. Les résultats de ces évaluations seront discutés dans le chapitre 5.

— Erreur de délimitation de certains syntagmes nominaux : ce phénomène est plus fréquent lorsque l'analyse porte sur des syntagmes nominaux complexes contenant des prépositions, plus particulièrement *de*, *à* et *avec* ;

— Mauvais raccordement de groupes nominaux soit vers d'autres groupes nominaux, soit vers d'autres verbes, soit vers aucun élément. Ce dernier cas de figure s'observe très souvent lors du traitement des groupes infinitifs et des verbes introduits par des locutions impersonnelles ;

— Erreur d'annotation des relations de dépendance lorsque les verbes sont conjugués à un temps composé (Exemple : *cette maladie a été provoquée par un virus*) ;

— Difficulté de traitement des phrases longues et complexes ;

— Confusion entre adjectifs et participes passés ;

— Problème de segmentation des syntagmes coordonnés.

Les phrases (verbes et arguments) illustrant certains phénomènes listés ci-dessus ont été corrigées à l'étape de pré-traitement des résultats de l'annotation syntaxique fournis par Cordial (cf. chapitre 3 section 3.2.1.2). Des heuristiques ont été mises en place pour améliorer les cas problématiques remédiables par des méthodes automatiques. Ces initiatives ont permis d'améliorer les résultats de l'annotation syntaxique, comme l'indiquent les données présentées à la section 3.2.1.2 du chapitre 3. Au total 18 231 phrases ayant des formes verbales complexes ont été corrigées. Le processus d'annotation sémantique a lui aussi été marqué par divers types de difficultés qui seront décrites dans la section suivante.

4.2.2 Annotation sémantique

4.2.2.1 La variation terminologique

Le terme *variation terminologique* peut renvoyer à différents types de variations observées dans les langues de spécialité. Grabar & Hamon (2004) énoncent plusieurs raisons qui peuvent être à l'origine de ce phénomène : les facteurs géographiques, la différence entre les locuteurs, la diachronie, etc. La variation terminologique peut se manifester à des niveaux variés : morphologie, lexique, etc. Lors de la description du corpus dans la section 2.1.3.1 du chapitre 2, nous avons parlé de la variation morpho-lexicale qui touche les termes nominaux du corpus. Elle renvoie au fait que certains termes complexes se réalisent sous des formes diverses que l'on appelle *variantes*. Cette variation ne concerne pas uniquement les unités nominales du corpus mais également celles de la terminologie, car cette dernière contient des termes et/ou leurs variantes comme les exemples suivants :

— *auricule du coeur - auricule cardiaque*

— *extension du foetus - extension foetale*

La variation terminologique, très fréquente, cause des problèmes d'incompatibilité entre le corpus et la terminologie. L'annotation des termes illustrant ce phénomène a été effectuée par l'application de différentes techniques.

Dans un premier temps, une ressource a été créée à partir de certains termes présents dans le corpus mais non répertoriés dans la Snomed. De plus, un ensemble d'heuristiques a été défini afin de rendre possible l'appariement des variantes d'un même terme, apparaissant séparément dans la terminologie et dans le corpus : appariement des termes à partir de leur tête lexicale, appariement des bigrammes et n-grammes (cf. section 3.2.2).

4.2.2.2 La non-exhaustivité de la ressource terminologique

À travers la description de notre méthode d'annotation, il apparaît clairement que la terminologie Snomed ne couvre pas tous les termes du corpus. Une explication à ce constat est le phénomène de variation terminologique (morpho-lexicale) évoqué dans la section 2.2.2 du chapitre 2, et décrit supra. Il faudrait toutefois souligner que les termes non capturés par la Snomed ne relèvent pas tous de la variation terminologique. En effet, il a été constaté que plusieurs termes en position d'argument des verbes analysés n'étaient pas enregistrés dans la Snomed, et leurs variantes non plus (*substrat arythmogène, psychopathologie, télécardiologie, télithromycine, immunomodulateur, cardiotoxicité, ocréotide*).

Comme il a été expliqué dans la section 2.2.2, la problématique de la faible couverture des ressources terminologiques existantes (pas seulement en médecine mais dans tous les domaines de spécialité) est bien connue en recherche sur les textes en langues de spécialité (Delpech, 2011 ; Grivel, 2011 ; Charlet *et al.*, 2012). Cette réalité nous a poussée à mettre en place des méthodes et à définir des ressources (cf. section 3.2.2) qui ont permis de pallier tant bien que mal ce manque, dont l'incidence aurait sinon été pénalisante pour les résultats de cette étude.

4.2.2.3 Problème de désambiguïsation des termes non procéduraux

Dans la section 3.2.3.3, il a été souligné que notre système automatique de désambiguïsation des termes polysémiques ne peut couvrir qu'un certain type d'ambiguïté que présentent les termes de la Snomed. Il s'agit des cas d'ambiguïté qui impliquent la catégorie P (procédure). Cette catégorie regroupe les termes ayant une interprétation d'activité. Il s'agit en général des noms désignant une procédure médicale : *intervention chirurgicale*, *ablation*, *rééducation du patient*, *hospitalisation*. En effet, tel qu'il a été expliqué lors de la description du processus d'annotation sémantique, l'interprétation procédurale nous sert de pivot pour la distinction et la désambiguïsation des sens d'un terme, car son intervention impose la présence d'un certain nombre de paramètres linguistiques (présence d'un verbe de réalisation, d'un agent, etc.) sur lesquels notre système automatique s'appuie afin d'opérer la désambiguïsation. L'absence de

ces paramètres indique que le terme polysémique n'a pas d'interprétation procédurale dans le contexte concerné et, par conséquent, une autre catégorie sémantique lui est attribuée.

Les cas d'ambiguïté n'impliquant pas la catégorie P (comme l'opposition D vs. F) sont donc difficilement analysables par notre système automatique, car la désambiguïsation, pour la plupart d'entre eux, repose sur des informations extralinguistiques que ce système ne possède pas. Lorsqu'il est confronté à de tels cas, le système attribue au terme ambigu ce que nous considérons comme la catégorie par défaut de la tête, c'est-à-dire la catégorie la plus fréquemment utilisée en combinaison avec la tête du terme dans la terminologie Snomed.

4.3 Résultats de la sélection des PSS pour la simplification

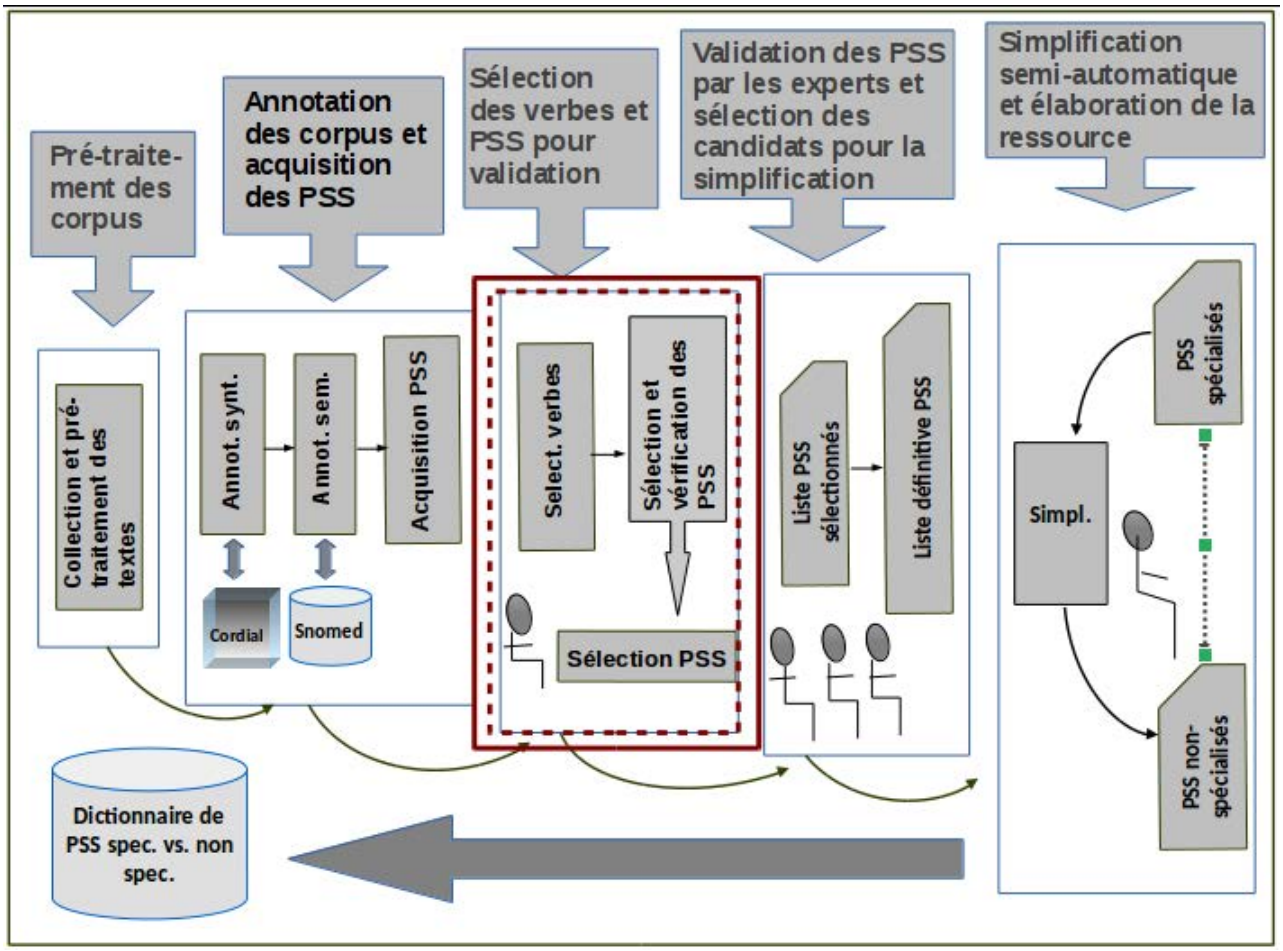


FIG. 4.4 – Schéma de la méthode : sélection des verbes et PSS.

4.3.1 Sélection des verbes

La sélection des verbes a consisté en un processus composé de trois étapes, au cours desquelles certains verbes ont été retenus et d'autres supprimés, tel qu'indiqué lors de la description de la méthode (cf. section 3.3.1). Le tableau 4.19 présente le récapitulatif du processus de sélection avec le nombre de verbes obtenus à chaque étape du tri, tandis que le tableau 4.20 présente la liste des verbes retenus pour la simplification.

TAB. 4.19 – Récapitulatif du processus de sélection des verbes.

Nombre tot. verbes	**Suppression verbes** (f<30)	**Suppression verbes et sélection finale** (de la langue générale)
2859	617	45

TAB. 4.20 – Liste des verbes sélectionnés pour la simplification

accompagner	détecter	imposer	réaliser
abaisser	développer	indiquer	relever
activer	diagnostiquer	induire	révéler
administrer	disséminer	inhaler	sécréter
affecter	éliminer	nécessiter	signaler
altérer	envisager	observer	synthétiser
analyser	évaluer	poursuivre	subir
associer	évoquer	pratiquer	suivre
coloniser	exposer	présenter	traduire
contrôler	exprimer	produire	traiter
dépister	impliquer	recommander	transmettre
manifester			

La liste des verbes retenus n'est pas exhaustive en ce qu'elle ne présente que le résultat de notre sélection, c'est-à-dire le groupe de verbes dont les patrons syntaxico-sémantiques seront analysés dans le cadre de la constitution de notre ressource de simplification. D'après nos observations, appuyées par les données du tableau 4.19, notre corpus comporte davantage de prédicats qui pourraient être de bons candidats pour la simplification. N'étant pas en mesure de traiter tous ces verbes dans le cadre de ce travail de thèse, les différentes tâches relevant de la simplification de textes porteront exclusivement sur les verbes du tableau 4.20. Cependant, comme nous allons l'observer dans la section 4.6, l'analyse des collocations verbe-terme ne portera pas systématiquement sur cette liste de verbes, d'autres verbes du corpus seront mis à contribution. De la même façon, plusieurs verbes de cette liste n'interviendront pas dans l'étude des collocations (cf. chapitre 3, section 3.6). En effet, le choix des collocations verbe-terme analysées dans cette étude n'a pas été basé sur la même sélection de verbes, mais plutôt sur la fréquence plus ou moins élevée (f≥5) des collocations en question.

Peu de verbes supports figurent dans la liste des candidats sélectionnés. Mais tout au long de cette étude, différents verbes supports (*faire, donner, mettre, prendre, etc.*) interviendront dans nos analyses, notamment lors de la comparaison des corpus et dans la phase d'alignement des PSS. En effet, plusieurs verbes supports sont utilisés comme équivalents lors de l'alignement, car ils sont fréquemment employés au sein du corpus des non-experts dans des constructions qui pourraient être instanciées par n'importe quel verbe (STASOCIAL (patient) fait/subit MALADIE).

4.3.2 Sélection et vérification des PSS pour la validation

À partir des résultats obtenus à l'étape 4.1.2, les PSS comportant les verbes du tableau 4.20 sont extraits. Au total, 836 PSS ont été extraits, parmi lesquels 243 (cf. annexe D.3) ont été sélectionnés pour la validation par les experts.

Deux critères linguistiques ont servi de socle à cette tâche. Le premier est le caractère spécialisé du sens du verbe dans le PSS : lorsque dans un PSS, un verbe semble avoir un sens qui ne relève pas de la langue courante, il est sélectionné. Il va de soi que ce critère repose en grande partie sur notre intuition linguistique. Les patrons suivants ont été retenus pour la validation grâce à ce critère :

— MALADIE est évaluée : *Les autres addictions doivent être évaluées.*

— MALADIE est évoquée (chez STASOCIAL) : *Un asthme sera aussi évoqué et écarté, particulièrement chez un sujet jeune, alors que la possibilité d'une affection neuromusculaire débutante, pouvant aussi se manifester par une dyspnée d'effort isolée, ne doit pas être négligée.*

— MALADIE est isolée : *L'absence d'accélérations pendant le travail était de signification incertaine si elle restait isolée.*

La variété des contextes d'apparition du verbe a également été prise en considération lors du choix des PSS qui ont été mis à part pour la validation. Lorsqu'un verbe a tendance à opérer dans des contextes variés, impliquant différentes combinaisons de catégories Snomed, nous supposons que certains contextes peuvent cacher des sens spécialisés, et par conséquent, les PSS illustrant ces différentes combinaisons sont retenus. Les verbes *accompagner, évaluer, évoquer, présenter, relever* enregistrent plusieurs patrons qui ont été sélectionnés sur la base de ce critère :

— MALADIE accompagne MALADIE ('associer à') : *La perlèche ou chéilite accompagne volontiers les candidoses oropharyngées.*

— PROCÉDURE accompagne PROCÉDURE ('suivre') : *Le soin s'accompagne d'une évaluation de la continence du blessé entre les sondages (6 à 8 par 24 heures en début de séjour).*

— PROCÉDURE s'accompagner de FONCTION_DE_ORGANISME ('entraîner') : *Chez les patients coronariens connus, indépendamment d'une revascularisation récente, l'arrêt de*

l'aspirine s'accompagne d'une augmentation du risque thrombotique dans la période post opératoire.

Les 243 PSS ont ensuite subi une vérification manuelle qui a permis d'identifier un certain nombre d'erreurs, dont certaines ont été décrites lors de la description des problèmes rencontrés pendant l'annotation sémantique (cf. section 4.2) :

1. Erreur d'étiquetage syntaxique des arguments des verbes : ce type d'erreur provient de l'analyse syntaxique effectuée par l'annotateur Cordial. Les résultats de l'évaluation de cet outil (cf. chapitre 5) fournissent d'amples informations à ce sujet.

   — Mauvaise délimitation de certains syntagmes : *Proposer une technique d'anesthésie et d'analgésie adaptée, dans une situation de geste programmé, en urgence, précautions / mise en garde.*

   Dans cette phrase exemple, le groupe nominal *une technique d'anesthésie et d'analgésie adaptée* a été segmenté en deux syntagmes distincts, jouant des rôles différents : *une technique d'anesthésie* (COD) ; *et d'analgésie adaptée* (complément circonstanciel).

   — Les composantes des constituants de la phrase sont parfois détachées du syntagme dont ils dépendent : *Une limite importante des cathéters ventriculaires couplés à une mesure par un système électronique est l'erreur sur la mesure lors du drainage ventriculaire.*

   Dans cette phrase, l'adjectif *ventriculaire* a été détaché du nom (*drainage*) qu'il caractérise et a été annoté comme un complément circonstanciel. Ce type d'erreur, plus précisément la confusion entre syntagme adjectival et syntagme nominal, fait partie des erreurs bien connues des concepteurs du logiciel Cordial, qui les ont exposées lors de la campagne d'évaluation PASSAGE (Laurent *et al.*, 2009).

   — Présence de faux arguments dans le schéma valenciel du verbe. Ce type d'erreur est particulièrement fréquent avec les syntagmes prépositionnels introduits par *de* et *à* : *onde E exclusive traduisant une élévation importante de la pression auriculaire gauche.*

   Dans cette phrase, le syntagme prépositionnel *de la pression auriculaire gauche* est détaché de sa tête et annoté comme étant un COI, ce qui débouche sur la présence d'un faux argument dans la structure argumentale. Ce problème fait également partie des erreurs signalées dans les résultats du logiciel Cordial lors de la campagne PASSAGE (Laurent *et al.*, 2009).

2. Attribution de catégories sémantiques inadéquates aux termes : la majorité des cas d'erreurs de ce type concernent les catégories F et D. Comme il a été souligné à la

section 3.2.3.3, ce type d'erreur relève de la performance de notre système d'annotation sémantique.

44) *Une ischémie douloureuse*_F ou silencieuse peut être induite lors d'une épreuve d'effort.

Au terme de l'annotation de la phrase proposée dans l'exemple ci-dessus, notre système a attribué la catégorie F au terme *ischémie douloureuse* qui n'existe pas dans la Snomed. Théoriquement, cette catégorisation n'est pas erronée car le terme *ischémie* désigne bel et bien une fonction de l'organisme selon la Snomed. Le système d'annotation a donc induit la catégorie F à partir des informations dont il dispose. Or, en combinaison avec l'adjectif *douloureux*, le terme *ischémie* renvoie à une maladie (D). Cependant, une telle discrimination entre les interprétations d'un terme requiert des connaissances extralinguistiques dont le système ne dispose pas.

Les patrons sémantiques présentant ces différents types d'erreurs sont corrigés et également retenus pour la phase de validation par les experts.

4.4 Résultats de la validation des PSS : analyse et interprétation

Le processus de dépouillement des résultats de la validation consiste en différentes phases et critères pris en considération lors de la formation de la liste finale des PSS (à partir des résultats de la validation) qui constitueront les entrées de la ressource de simplification. La figure 4.5 situe la tâche de validation dans notre chaîne de travail. La figure 4.6 quant à elle propose un schéma descriptif de cette étape de travail qui va de la validation jusqu'à l'obtention de la liste définitive des PSS retenus, en passant par le dépouillement des résultats de la validation. Les abréviations (*Fr*, *Be* et *Ca*) symbolisent les trois groupes d'experts de différents pays (France, Belgique, Canada) qui valident les PSS. Les cercles représentent les réponses des experts et les différentes intersections des cercles symbolisent l'accord entre les experts.

4.4.1 Présentation générale des résultats

Le tableau 4.21 présente un aperçu global des réponses sélectionnées par les experts qui ont répondu au questionnaire. Trois lettres sont utilisées, chacune correspond à une catégorie de réponse. Pour chaque groupe d'experts représentant un pays, le nombre total de réponses obtenues par lettre est fourni :

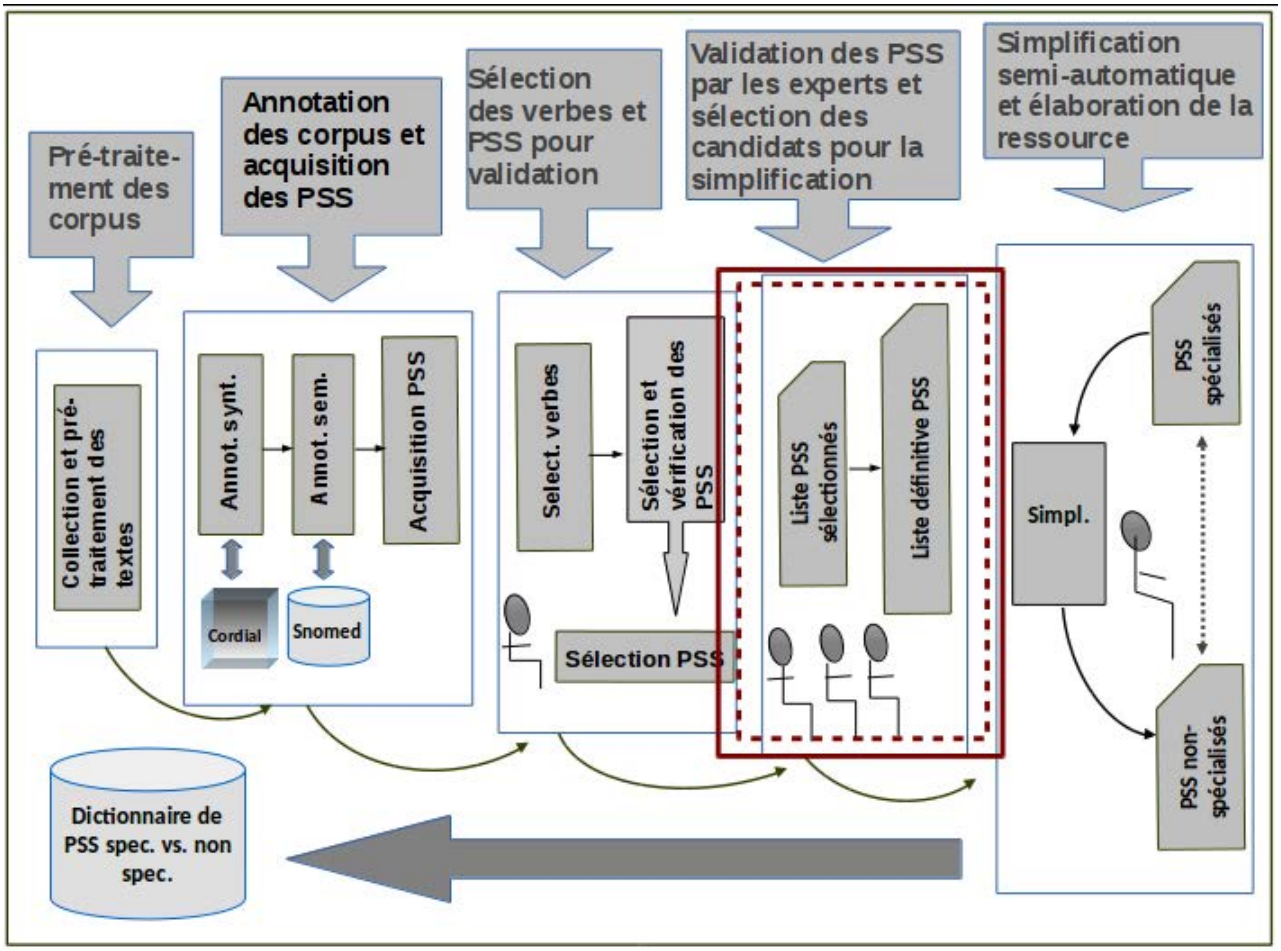


FIG. 4.5 – Schéma général de la méthode : validation des PSS.

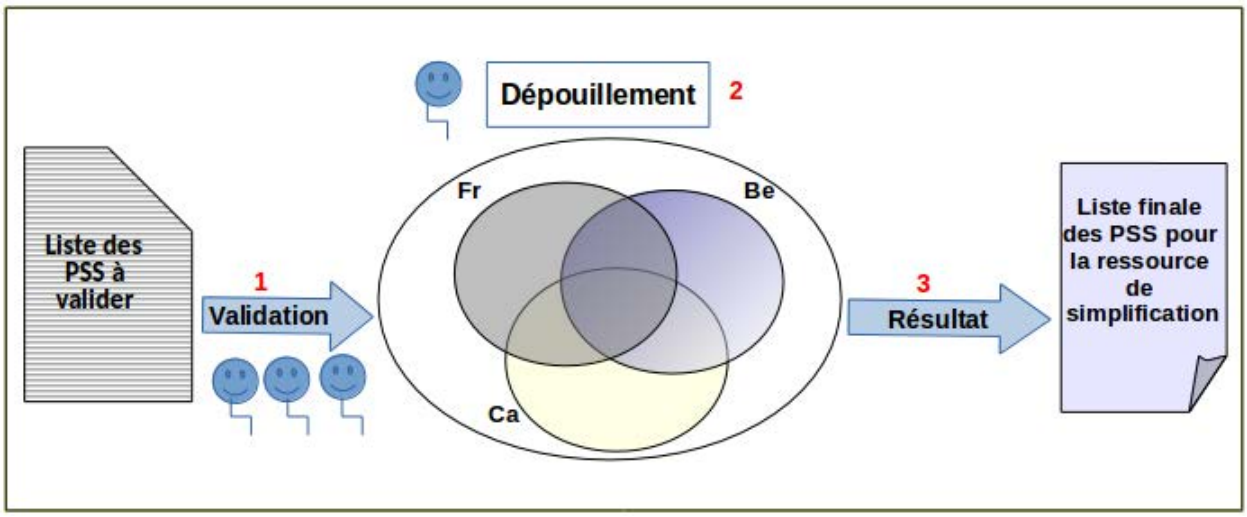


FIG. 4.6 – Processus de validation des PSS.

a) Vous et/ou vos collègues avez tendance à utiliser cette construction dans l'exercice de vos fonctions.

b) Vous n'utilisez pas ou très peu cette construction mais vous la reconnaissez parce qu'elle est utilisée par des intervenants lors des conférences ; OU par les enseignants en médecine ; OU dans la littérature médicale.

c) Vous n'avez jamais entendu cette construction dans l'exercice de vos fonctions.

TAB. 4.21 – Résultats obtenus classés par catégories.

Réponses	France	Belgique	Canada
a	143	144	81
b	74	79	103
c	26	20	59
total	243		

D'après les résultats du tableau 4.21, la catégorie *(a)* enregistre la majorité des réponses (367/729), et il y a en moyenne 122 réponses *(a)* par équipe, ce qui signifie que plus de la moitié des constructions proposées sont reconnues par les experts des trois équipes comme étant spécialisées. Les patrons verbaux classés à l'unanimité dans cette catégorie pourraient donc être considérés comme étant attestés en tant qu'emplois verbaux spécialisés de façon internationale, par des experts représentant au moins trois communautés différentes de praticiens de la médecine. Par conséquent, ces patrons constituent de bons candidats pour une ressource comme celle que nous proposons d'implémenter.

La catégorie *(b)*, qui regroupe les constructions verbales que les experts déclarent rencontrer le plus souvent dans la littérature et/ou lors de manifestations (conférences, séminaires, etc.), occupe la deuxième place après la catégorie *(a)*. Ce constat est plus particulièrement évident chez les Canadiens qui enregistrent 42,38 % de constructions associées à la catégorie *(b)*. Cette remarque est davantage soutenue par les données du tableau 4.22 qui montrent que 49 constructions sont associées à la catégorie *(b)* par des experts de deux pays différents à la fois, ceci combiné aux 21 constructions qui se voient attribuer à l'unanimité cette même catégorie. Cette observation pourrait signaler le caractère collocationnel ou idiomatique des patrons concernés. En effet, il pourrait s'agir de constructions qui tendent à être utilisées dans les écrits de type scientifique plus qu'à l'oral, ce qui les rendrait susceptibles de causer des problèmes de compréhension chez les lecteurs non experts. Ces patrons verbaux ayant le profil parfait pour notre étude, leur présence est indispensable dans la ressource qui en résultera.

La catégorie *(c)* a clairement été la moins utilisée par les experts, comparée aux deux autres. D'après le tableau 4.21, elle totalise uniquement 105/729 réponses, c'est-à-dire qu'en moyenne 35 patrons syntaxico-sémantiques ne seraient pas attestés par un répondant sur 243 patrons traités. De plus, d'après les données du tableau 4.22, 2/243 constructions seulement ont été identifiées à l'unanimité comme n'étant pas connues. De même, 19/243 constructions ont été annotées de façon partielle (c'est-à-dire par deux des trois groupes d'experts) comme n'étant

pas connues. Ces résultats démontrent que les experts des trois pays partagent des avis similaires à propos des patrons verbaux qu'ils classifient comme inconnus. Les 2 constructions portant la catégorie *c* à l'unanimité (c'est-à-dire pour les trois groupes d'experts) sont d'office éliminées, tandis que celles qui font l'objet d'une attestation unanime ou partielle sont retenues pour la suite de ce travail. Par contre, les 19 patrons qui enregistrent deux avis négatifs, c'est-à-dire annotés comme n'étant pas spécialisés par deux des trois groupes d'experts, seront analysées à la lumière de critères qui déboucheront sur leur sélection ou leur élimination (cf. section 4.4.2).

À travers les données du tableau 4.21, l'on observe par ailleurs que l'équipe canadienne enregistre le plus grand nombre de candidats annotés comme non spécialisés (réponse *(c)*), au total 59/243, contre 26 et 20 respectivement pour les équipes française et belge. Inversement, l'on remarque que les experts français et belges enregistrent le plus grand nombre de constructions validées comme étant spécialisées (respectivement 217 et 223/243), avec moins de 30 constructions non attestées pour chacune de ces deux équipes. Les experts français et belges apparaissent donc comme ceux ayant attesté le plus grand nombre de constructions comme faisant partie du langage qu'utilisent les praticiens de la médecine. Ce constat concernant la différence des réponses des experts du Canada par rapport aux deux autres groupes suscite au moins deux hypothèses qui seront exposées dans la section suivante, en nous appuyant sur des résultats plus détaillés qui permettent de mieux faire ressortir le contraste entre les réponses des équipes en jeu.

Le tableau 4.22 fournit un récapitulatif des résultats obtenus, organisés en termes de combinaisons de réponses. Dans ce tableau, chaque jeu de lettres de la ligne *combinaison* (*aaa, aab,* etc) représente une combinaison de réponses obtenues pour un certain nombre de constructions spécifié dans la cellule *Nombre*. Chaque lettre correspond à une catégorie de réponse donnée par les experts (cf. section 4.4.1).

TAB. 4.22 – Récapitulatif de la validation.

Validation	**Unanimité (3/3)**					**Validation partielle (2/3)**				
Combinaisons	aaa	bbb	ccc	aab	bba	aac	bbc	ccb	cca	abc
Nombre	42	21	2	61	37	23	12	8	11	26
Totaux	163					80				
Tot. rejet	2/163					19/80				
Total	243									

L'ordre d'apparition des lettres n'est pas significatif dans le tableau 4.22. En effet, l'accent est mis sur la cooccurrence des catégories (lettres) symbolisant l'attestation ou le rejet d'un PSS par les experts. Par exemple, la combinaison *aaa* veut dire que les 42 constructions concernées sont attestées par les experts des trois équipes qui ont sélectionné la même réponse *(a)*, alors que la combinaison *aac* signale que deux des experts (c'est-à-dire France-Belgique ou Belgique-Canada ou Canada-France) ont attesté les 23 PSS concernés en sélectionnant l'option *(a)* tandis qu'un

seul expert n'a pas attesté les PSS (option *(c)*). La combinaison *aab* indique que tous les experts ont attesté les 61 PSS concernés, mais en optant pour des réponses différentes : deux experts ont sélectionné l'option *(a)* et un seul l'option *(b)*.

Cette première partie de la discussion des résultats du questionnaire permet de faire une remarque d'ordre général : les réponses unanimes sont les plus fréquentes dans les résultats obtenus grâce au questionnaire. La majorité des PSS sont attestés par au moins deux experts sur trois. Cette unanimité transparaît encore plus à travers les données que présente la figure 4.7 proposée dans la section suivante.

4.4.2 Analyse détaillée des résultats

En plus des observations d'ordre général présentées dans la section 4.4.1, les réponses obtenues au moyen du questionnaire nous ont également permis d'analyser de façon détaillée les résultats de chaque équipe d'experts et de les comparer aux résultats des autres équipes.

La figure 4.7 décrit la répartition des réponses au questionnaire entre les trois groupes d'experts, en mettant en évidence les points d'unanimité (entre deux ou trois experts). Les trois cercles symbolisent les trois groupes d'experts interrogés pour chacun des trois pays. Les différentes intersections des cercles représentent les points d'unanimité entre les experts validant un PSS. L'unanimité peut concerner les trois experts ou uniquement deux sur trois experts (*partialité*). À l'intérieur de chaque intersection est inscrit (en rouge) le nombre de réponses communes fournies par les groupes d'experts des pays concernés.

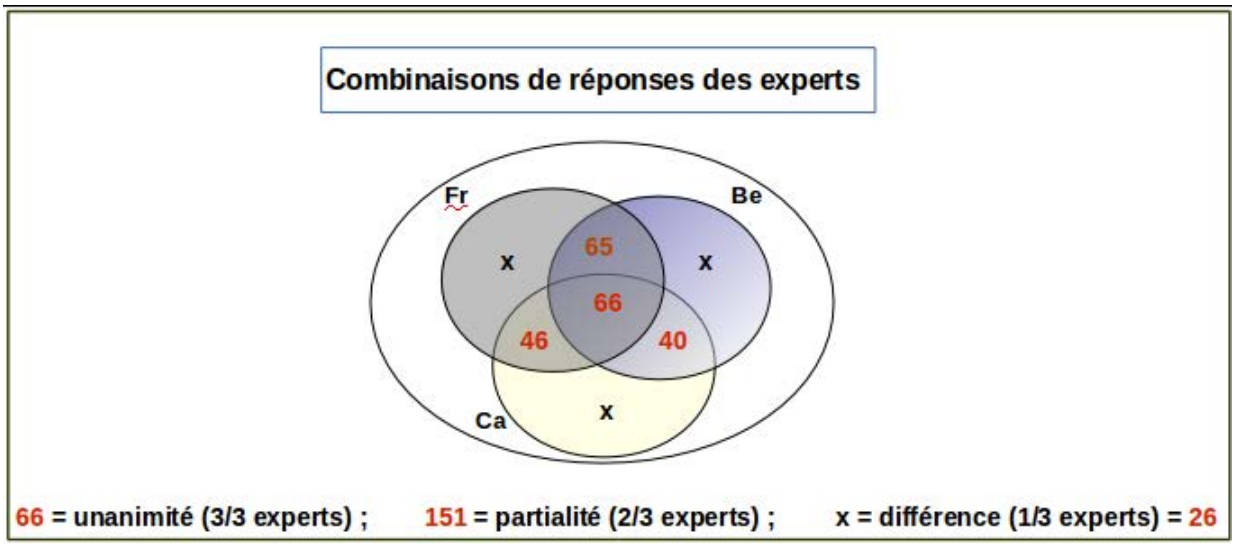


FIG. 4.7 – Les combinaisons de réponses.

Les données de cette figure montrent qu'au total 217 constructions (c'est-à-dire 89,30%) ont reçu des réponses identiques attribuées par au moins deux experts sur trois. Parmi ces

constructions, 66 (30,40%) ont reçu des combinaisons de réponses identiques (*aaa*, *bbb* ou *ccc*) de la part des trois experts et 151 (69,58%) de la part de deux experts sur trois. Ces chiffres démontrent clairement que les représentants des trois communautés d'experts partagent des avis communs en ce qui concerne l'usage (par les experts) de la majorité des constructions proposées. Ces constructions validées à l'unanimité ou par deux experts sur trois sont immédiatement retenues comme entrées de la ressource de simplification, car l'unanimité ou la majorité des experts est le premier critère que nous prenons en considération pour la constitution de la nomenclature de la ressource.

En outre, les informations que présente la figure 4.7 appuient l'observation faite dans la section 4.4.1 en ce qui concerne la similarité des réponses obtenues entre les équipes française et belge, par opposition aux réponses de l'équipe canadienne. On observe ainsi une prédominance de résultats similaires entre les experts français et belges, qui se manifeste à travers le nombre élevé de réponses qu'enregistre l'intersection France-Belgique (65 c'est-à-dire 43% des réponses communes obtenues).

Deux hypothèses sont émises pour expliquer l'écart qui existe entre les résultats des équipes française et belge, par opposition à ceux de l'équipe canadienne. Cette distance pourrait traduire la discordance qui caractériserait les niveaux d'expertise des personnes interrogées. Cette remarque est cohérente avec les données du tableau 3.7 qui indiquent que les répondants des équipes française et belge sont en majorité des médecins (5 et 3 respectivement), tandis que l'équipe canadienne est la seule constituée exclusivement d'infirmiers. La divergence de niveaux d'expertise entre médecins et infirmiers pourrait donc être à l'origine de l'écart observé. Cette première hypothèse s'applique bien au cas présent car les médecins ont une plus grande aptitude à reconnaître les constructions verbales spécialisées, puisqu'ils sont plus susceptibles d'utiliser et d'entrer fréquemment en contact avec ce type d'expressions langagières dans l'exercice de leurs fonctions.

La divergence observée entre les résultats des équipes européennes et de l'équipe canadienne pourrait également être l'expression de la différence qui existerait entre le français parlé en Europe (France, Belgique, Suisse, etc.) et le français canadien. La question de la variation entre le français de l'Europe (de la France en particulier) et celui du Canada fait l'objet de nombreux travaux de recherche depuis longtemps (Gendron, 1966 ; Ashby, 1988 ; Mougeon, 1995 ; Simoni Aurembou, 2000 ; Bagola *et al.*, 2007). Du fait de sa large diffusion à travers le monde, notamment en Amérique du Nord et dans la région québécoise, le français, langue vivante, a été et est encore de nos jours soumis à une grande variation géographique. Cette diffusion géographique de la langue française a un impact non seulement aux niveaux lexical (à travers l'élargissement du lexique) et phonétique, mais aussi au niveau de l'usage de la langue. Un bon nombre d'usages différents de part et d'autre de l'océan Atlantique ont ainsi vu le jour, ce qui favorise le développement des variantes géographiques entre le français de

l'Amérique du Nord, notamment du Canada, et celui de l'Europe. Ce phénomène est d'autant plus important que des dictionnaires (Bélisle, 1974 ; Cajolet-Laganière, 2009) sont créés pour la description et surtout la mise en relation de ces différentes variétés du français. Certains colloques internationaux (Simoni Aurembou, 2000 ; Bagola *et al.*, 2007) sont exclusivement dédiés aux travaux portant sur l'étude contrastive du français canadien et du français de France. Bien que les travaux cités dans cette argumentation portent sur la langue générale, nous pensons que certains langages de spécialité comme le langage médical sont susceptibles de subir des phénomènes de variation similaires à celui que nous avons décrit ci-dessus.

Les résultats obtenus grâce au formulaire nous ont également permis de prendre du recul par rapport aux choix de réponses qu'ont fait les experts. Dans un premier temps, nous allons discuter des constructions qui n'ont pas fait l'unanimité entre les experts. Il s'agit plus particulièrement des constructions qui ont été attestées par un seul expert sur trois. D'après les données du tableau 4.22, on en dénombre 19 (sur 80 cas de validation partielle), qui sont présentées dans le tableau 4.23. Chaque construction est précédée du numéro qu'elle porte dans le formulaire qui a été remis aux experts et est suivie des lettres représentant les réponses fournies par les experts interrogés pour chaque pays. La dernière colonne du tableau propose les fréquences de ces constructions dans le corpus.

TAB. 4.23 – Les 19 patrons validés partiellement (1/3 experts).

Num.	Constructions	Fr.	Be.	Ca.	Freq.
243	STASOCIAL présente MALADIE	c	b	c	25
95	MALADIE est associée à FONCTION_de_ORGANISME	a	c	c	22
7	STASOCIAL pratique PROCEDURE	c	a	c	20
23	PROCEDURE implique PROCEDURE	b	c	c	15
37	METIER envisage PROCEDURE	b	c	c	12
206	MALADIE est associée à PROCEDURE	c	b	c	12
107	METIER évoquer MALADIE	a	c	c	10
153	PROCEDURE abaisse FONCTION_de_ORGANISME	a	c	c	10
4	ANATOMIE contrôle FONCTION_de_ORGANISME	c	b	c	9
177	MALADIE est traitée par PROCEDURE	c	a	c	8
151	MALADIE se développer	b	c	c	7
26	MALADIE se manifeste chez STASOCIAL	a	c	c	5
141	ORGVIVANT induit MALADIE	b	c	c	3
191	METIER diagnostique STASOCIAL	c	a	c	2
38	PCHIMIQUE présente FONCTION_de_ORGANISME	b	c	c	1
200	ANATOMIE synthétise FONCTION_de_ORGANISME	c	a	c	1
207	AGENT affecte ANATOMIE	c	a	c	1
226	PROCEDURE dépiste FONCTION_de_ORGANISME	c	c	b	1
238	MALADIE se réalise	c	c	a	1

Comme on peut le remarquer, les données du tableau 4.23 confirment le constat fait à la

section 4.22 selon lequel l'équipe canadienne totalise le plus grand nombre de réponses *(c)*, qui renvoient à la non-attestation d'un patron par les experts interrogés. En effet, dans la colonne qui présente les résultats des experts canadiens, l'on compte 17/19 patrons (contre 10 et 11 pour la France et la Belgique respectivement), c'est-à-dire 89% des patrons, associés à *(c)*. Il faudrait souligner que ces résultats ont été fournis par différents experts. Ce constat évoque une fois de plus l'hypothèse portant sur la différence de niveaux de spécialisation entre les répondants médecins (pour la plupart) des équipes française et belge et les infirmiers qui constituent l'ensemble de l'équipe canadienne.

En ce qui concerne la fréquence des patrons listés dans le tableau 4.23, notre attention est attirée par le fait que 8/19 constructions ont une fréquence en corpus moyenne ($f \geq 10$), tandis qu'un peu plus de la moitié (11/19) de ces constructions ont une très faible fréquence ($f < 10$). Cette remarque laisse penser qu'il peut y avoir un lien entre la fréquence des PSS dans le corpus et leur utilisation par les experts en médecine.

Par ailleurs, l'on pourrait émettre l'hypothèse que la faible fréquence dans le corpus des 11 dernières constructions est un indicateur du fait qu'elles ne font pas vraiment partie du vocabulaire des membres du corps médical. Une telle réflexion impliquerait que les constructions concernées soient aussitôt mises à l'écart dans la suite de ce travail de recherche. Cependant, une observation faite sur les constructions sélectionnées à l'unanimité par nos experts, et qui par conséquent sont d'office retenues pour la suite de cette étude, nous pousse à ne pas exclure immédiatement les 11 constructions concernées mais à les garder pour une autre phase de tri. En effet, parmi les 161 PSS (cf. tableau 4.22) qui ont été attestés collectivement par les experts des trois pays, nous avons détecté plus d'une dizaine de PSS (cf. tableau 4.24) qui s'avèrent eux aussi avoir une très faible fréquence dans le corpus ($f < 10$). Pourtant, ces patrons ont bel et bien fait l'unanimité chez les experts en ce qui concerne leur caractère spécialisé.
Cette remarque suscite au moins deux questions qui nous semblent importantes : qu'est-ce qui pourrait motiver la sélection de telles constructions par trois experts de différents pays ? Qu'est-ce qui pourrait expliquer que des constructions soient reconnues comme faisant partie du langage spécialisé de la médecine alors qu'elles sont très peu fréquentes dans un corpus médical ?

Par exemple le patron STASOCIAL *relever de* MALADIE a été attesté à l'unanimité, pourtant il n'a que deux occurrences dans le corpus. Le même constat a été fait pour les patrons PROCÉDURE *évoquer* MALADIE et FONCTION_DE_ORGANISME/MALADIE *est isolée*[11] qui comptent moins de 10 occurrences dans le corpus mais qui ont été attestés à l'unanimité. Il

11. *L'absence d'accélérations pendant le travail était de signification incertaine si elle restait isolée.* Dans cet exemple tiré de notre corpus, le verbe *isoler* signifie « qui n'est pas associé à ... ». Il s'agit donc d'un symptôme (*L'absence d'accélérations*) qui se manifeste seul, sans être accompagné d'autres symptômes qui seraient attendus dans ce genre de situation.

TAB. 4.24 – Quelques patrons sélectionnés à l'unanimité mais ayant moins de 10 occurrences dans le corpus.

Num	Constructions	Fr	Be	Ca
9	ORGVIVANT associer à MALADIE	a	a	a
34	ANATOMIE traduire FONCTION_de_ORGANISME en FONCTION_de_ORGANISME	a	a	b
42	MALADIE réaliser MALADIE	b	b	b
65	ANATOMIE présenter MALADIE	a	a	b
68	FONCTION_de_ORGANISME présenter FONCTION_de_ORGANISME	b	b	b
75	AGENT est envisagé	a	b	a
81	ANATOMIE est augmentée	a	b	b
85	PROCEDURE dépister MALADIE	a	a	a
103	MALADIE évoquer MALADIE	a	a	b
113	METIER évaluer STASOCIAL	b	a	a
134	STASOCIAL relever de MALADIE	a	a	b
172	STASOCIAL est évalué	b	a	b
237	FONCTION_de_ORGANISME relever de FONCTION_de_ORGANISME	b	a	b

faudrait signaler que dans cet emploi, le verbe *isoler* a un sens très spécialisé n'ayant rien à voir avec celui de « mettre à l'écart par mesure de protection » (*isoler un malade*) qui est bien connu et même cité par des dictionnaires de la langue générale comme le dictionnaire Larousse consultable en ligne [12].

Les questions posées ci-dessus soulèvent le problème d'exhaustivité du corpus et nous renvoie au rôle de la fréquence. En effet, bien qu'étant un extrait représentatif, un corpus ne saurait être exhaustif au point de couvrir de façon équitable tous les phénomènes linguistiques qui caractérisent la langue de spécialité concernée. Cette question d'exhaustivité est très fréquente dans les travaux sur les terminologies spécialisées et les ressources terminologiques (Borin *et al.*, 2007b). Il en ressort que la fréquence à elle seule n'est pas un critère suffisant pour décider du caractère spécialisé d'un patron syntaxico-sémantique. L'avis des experts du domaine est un critère de poids et, d'après nous, l'un des plus importants. Le type de catégories Snomed impliqué dans les patrons joue aussi un rôle déterminant dans la sélection des PSS, d'autant plus que ces catégories déterminent le sens de la construction. Nous pensons que ces catégories doivent avoir une incidence dans le jugement des experts. Si l'on se fie uniquement à la fréquence, l'on risquerait de passer à côté de constructions spécialisées mais très peu récurrentes comme les 11 que présente le tableau 4.23. De même, en s'appuyant exclusivement sur la fréquence,

12. http ://www.larousse.fr/dictionnaires/francais/isoler/44471 ?q=isoler44405

l'on aurait tendance à retenir uniquement les patrons très fréquents, mais qui ne seraient pas pour autant spécialisés. Cette réflexion est d'ailleurs l'un des facteurs ayant motivé la décision d'entreprendre une validation manuelle de nos patrons par des experts en médecine.

En ce qui concerne les 11 constructions les moins fréquentes du tableau 4.23, deux autres hypothèses pourraient être émises par rapport à leur faible fréquence dans le corpus et à leur attestation partielle par 1/3 experts. D'une part, ces patrons pourraient décrire des emplois très spécialisés des verbes concernés mais spécifiques à des branches particulières de la médecine, d'où l'attestation par un seul expert qui, dans ce cas, exercerait dans le sous-domaine médical concerné, puisque nos experts appartiennent à différentes branches de la médecine.

Ce type d'emplois verbaux pourrait donc cacher des phénomènes intéressants qui mériteraient d'être analysés dans une étude linguistique telle que la nôtre. Les exemples suivants visent à illustrer les deux hypothèses émises ci-dessus :

45) MALADIE *évoquer* MALADIE *: Appendicite pelvienne évoquant une infection urinaire ou gynécologique, mais avec des signes francs au TR. appendicite méso-caeliaque retentissant plus volontiers sur le transit (diarrhée ou sub-occlusion).*

46) MÉTIER *évoquer* MALADIE *: On évoque ici une maladie de Kawasaki, devant l'association d'une fièvre avec : un exanthème, des adénopathies cervicales, une conjonctivite, une atteinte de la muqueuse buccale et une atteinte de la paume des mains (desquamation en lambeaux) ou de la plante des pieds.*

47) PROCÉDURE *dépister* MALADIE *: L'appréciation de la symptomatologie fonctionnelle (claudication intermittente, douleurs de décubitus, troubles trophiques, ulcères et gangrènes), la palpation et l'auscultation des artères assurent, dans la majorité des cas, le diagnostic positif de l'artérite, renseignent sur la sévérité de l'ischémie, sur la topographie des lésions selon les sites d'audition des souffles et le niveau d'abolition des pouls, et peuvent dépister une lésion anévrismale (aor-tique, iliaque ou poplitée).*

48) PROCÉDURE *dépister* FONCTION_DE_ORGANISME *: Un interrogatoire alimentaire détaillé dépistera les consommations d'aliments riches en sel caché (fromage, pain, charcuterie, pizza) bouillons cubes..*

Lorsqu'on observe le patron MALADIE *évoque* MALADIE qui a moins de 10 occurrences dans le corpus mais qui a été attesté à l'unanimité, et le patron MÉTIER *évoque* MALADIE (10 occurrences) qui n'a été attesté que par 1/3 experts, force est de constater qu'il s'agit de deux variantes sémantiques d'une même construction syntaxique du verbe *évoquer*, la construction transitive directe. On remarque également que la principale disparité entre ces deux patrons est la catégorie sémantique de l'argument sujet qui est d'un côté non humain, non animé (MALADIE) et de l'autre animé et humain (MÉTIER). Cette différence confère un caractère particulier à chacun de ces emplois du verbe *évoquer*, car la variation de types sémantiques

des sujets implique également une nuance ou une différence au niveau de l'interprétation du sens du verbe. Dans la phrase 46, le verbe *évoquer* signifie 'fait penser à' et dans la suivante, il signifie 'soupçonner', 'penser à'. Une remarque similaire s'applique aux patrons des exemples 47 et 48, où le verbe *dépister* veut dire respectivement 'permet de détecter' et 'permet d'exposer'.

Les analyses faites ci-dessus permettent de remarquer que la plupart des constructions du tableau 4.23, en particulier les constructions les moins fréquentes, expriment des sens peu courants et même spécialisés des verbes. Le fait qu'elles aient été attestées par un expert peut être considéré comme une preuve. Mettre toutes ces constructions à l'écart parce qu'elles n'ont pas été attestées à l'unanimité n'est donc pas la solution idéale dans cette étude dont le but est de rendre compréhensible les emplois verbaux peu courants, spécialisés, susceptibles de créer des difficultés de lecture chez des personnes non expertes. Ainsi, la sélection des constructions du tableau 4.23 s'est effectuée de la façon suivante :

— 5 constructions (246, 7, 23, 37, 151) ont été sélectionnées du fait d'avoir déjà fait l'objet d'une étude dans nos travaux précédents.

— 7 (95, 206, 107, 153, 4, 177) ont été retenues sur la base de la fréquence : les constructions ayant un nombre d'occurrences $\leq$ à 5 dans le corpus ont été retenues, à l'exception du patron *MALADIE se manifester chez STASOCIAL* qui a été exclu car le verbe y a un sens appartenant au registre standard.

— 6 (38, 200, 207, 226, 191, 141) ont été retenues de par le caractère spécial (rare) du sens des verbes.

— la dernière construction (238) a été éliminée car il semble manquer un élément au patron. La phrase qui l'illustre donne la même impression.

> MALADIE réalise : *Cette maladie réalise en véritable handicap avec une déficience d'origine respiratoire qui se complique d'incapacité puis de désavantage social avec une participation rapidement systémique*.
>
> Le verbe *réaliser* est transitif direct, mais dans cette phrase, aucun COD n'est identifié. On a l'impression qu'il s'agit d'une forme pronominale (se réaliser), mais l'absence du pronom réflexif *se* remet cette hypothèse en question. Il pourrait s'agir d'une erreur de l'énonciateur qui aurait oublié le pronom *se*. Ces différentes hypothèses sans réponses ont débouché sur la suppression de ce patron.

Au terme du processus de sélection, une dixaine de PSS a été écartée car les emplois verbaux concernés semblent compréhensibles (donc pas de nécessité d'être simplifiés) pour des personnes n'ayant pas d'expertise en médecine (cf. tableau 4.25).

Les phrases exemples peuvent être consultées à travers le CD-ROM qui sera livré avec cette thèse. Ce support contiendra notre ressource de simplification et l'ensemble des phrases exemples illustrant les PSS.

TAB. 4.25 – Les PSS ne nécessitant pas de simplification.

ID	PSS
25	MALADIE se manifester comme MALADIE
26	MALADIE se manifester chez STASOCIAL
38	MEDECIN suit STASOCIAL
61	STASOCIAL montre MALADIE
65	ANATOMIE présente MALADIE
67	PCHIMIQUE présente FONCTION
76	PROCEDURE est proposée dans/comme PROCEDURE
79	PROCEDURE est proposée chez STASOCIAL
193	PROCEDURE montre MALADIE
220	STASOCIAL présente ANATOMIE (affectée par quelque chose)

4.4.3 Bilan

Dans cette section, il était question, d'une part, de présenter les résultats du questionnaire qui a été soumis aux experts et, d'autre part, de fournir les critères qui nous ont permis, à partir des patrons validés par les experts, de constituer la liste des patrons qui figureront dans la ressource de simplification. Ci-dessous ces critères :

— L'unanimité des experts sur le statut spécialisé des patrons a été le critère dominant et donc le principal critère considéré dans la prise de décision. Au total, 222 patrons syntaxico-sémantiques ont été ainsi retenus.

— La fréquence, le fonctionnement et les propriétés syntaxico-sémantiques des PSS ont facilité la prise de décision pour les 19 patrons partiellement validés par les experts. Au total, 18 ont été retenus grâce à cette combinaison de paramètres.

  — La fréquence des patrons a joué un rôle non négligeable de par sa double fonction. D'un côté, les fréquences ($f \geq 10$) ont permis de confirmer la sélection d'un certain nombre de constructions qui avaient déjà été retenues par un expert sur trois. Tandis que de l'autre, certaines faibles fréquences ont mis en évidence les emplois verbaux qui pourraient être très spécialisés, ou qui seraient rarement utilisés, et ceux qui pourraient être le fruit d'une erreur de l'énonciateur.

  — Le contraste des patrons syntaxiquement identiques, mais présentant une dissimilitude au niveau de la sémantique, a également joué un rôle important dans le processus de sélection des patrons partiellement attestés par les experts.

Sur la base de ces éléments, nous avons mis en place la liste définitive (cf. annexe D.4) des 230 patrons verbaux spécialisés qui constituent les entrées de la ressource de simplification.

Le dépouillement des résultats de la validation a également permis de faire certaines remarques et de prendre du recul par rapport au travail de validation qu'ont effectué nos experts en médecine.

Nous avons ainsi pu observer que les résultats obtenus à travers le questionnaire reflètent la cohérence qui existe entre les données extraites de notre corpus et les avis des experts (cf. section 4.4). De même, les résultats du questionnaire permettent de voir que le verbe peut fonctionner comme pivot pour l'extraction de connaissances spécialisées, puisque les PSS extraits des corpus traduisent les connaissances du domaine médical.

Par ailleurs, le fait d'avoir entrepris une méthode de validation tripartite, associant des évaluateurs français, belge et surtout canadien, a permis de mettre en évidence un aspect qui n'a pas amplement été abordé dans ce travail (car il relève d'une autre thématique), mais qui constitue une piste de recherche intéressante. Il s'agit de l'utilisation des constructions verbales au sein des différentes communautés de médecins francophones. Une étude autour de cette thématique pourrait consister à analyser et comparer le français parlé par les professionnels de la santé exerçant au Canada et celui parlé par ceux qui exercent en Europe, en mettant un accent particulier sur les constructions verbales que chaque communauté utilise pour s'exprimer et décrire des situations médicales précises.

4.5 Simplification des PSS

La simplification correspond à une phase au cours de laquelle les PSS spécialisés des verbes sont alignés avec des équivalents provenant du langage des non-experts. La réalisation de cette tâche passe par une méthode semi-automatique dont les résultats pour ce travail de thèse seront fournis dans les sections suivantes. Au terme de ce processus, une ressource alignant les PSS experts et non experts est créée.

Dans cette étude, la simplification porte sur le verbe dans son contexte d'apparition. En effet, dans les textes médicaux rédigés par les experts, très souvent, les verbes de la langue générale tels que *relever, évoluer, évoquer*, etc., sont employés dans des contextes spécialisés et ont pour arguments des termes médicaux. Dans ce type d'emplois, les verbes peuvent avoir des sens très spécialisés, contribuant ainsi à rendre le texte difficilement compréhensible pour des lecteurs non experts. La simplification intervient donc comme moyen de rendre le texte aisément lisible pour le public visé.

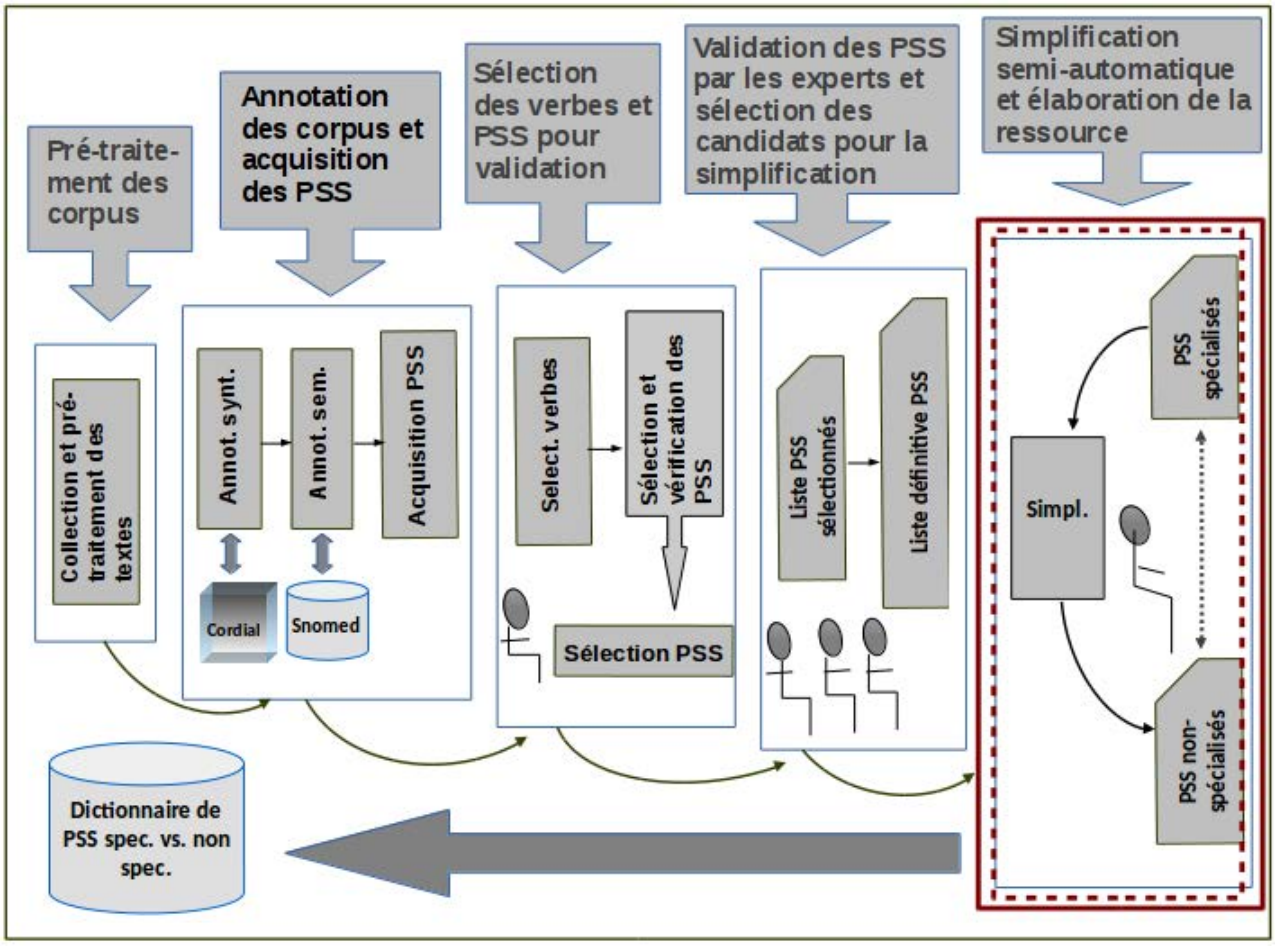


FIG. 4.8 – Schéma de la méthode : la simplification des PSS.

4.5.1 Résultats de la sélection automatique des candidats PSS équivalents

Le programme implémenté pour la détection des verbes candidats équivalents pour l'alignement des PSS génère des groupes de verbes intervenant dans les différents PSS génériques[13]. Un PSS générique est un PSS qui peut être instancié par différents verbes. Pour chaque PSS, l'automate compte et liste tous les verbes qui instancient ce PSS dans les corpus PRO, ETU, VUL et FOR. Le nombre d'occurrences de la paire verbe+PSS est également fourni. Le tableau 4.26 propose un extrait de ces résultats.

Au total, 836 PSS génériques ont été extraits, formant 836 ensembles de verbes, avec un nombre d'éléments variant entre 1 et 984. Le nombre élevé de PSS génériques est la conséquence de la présence des PSS partiellement annotés, c'est-à-dire des patrons ayant certains arguments qui ne portent pas de catégorie sémantique. Tel qu'expliqué à la section 4.1.2.1, les arguments

13. Ce type de PSS bénéficie d'une annotation différente avec le symbole tilde (∼) à la place du verbe.

TAB. 4.26 – Exemples de groupes de verbes candidats équivalents pour l'alignement.

Verbes	PSS génériques	PRO nb occ	ETU nb occ	VUL nb occ	FOR nb occ	**Synonymes**
accompagner	s_P∼COI_F	57	52	32	6	suivre, associer
révèle	s_P∼COD_D	69	101	36	8	montrer, signaler
associer	s_D∼COI_D	71	132	52	2	suivre
évoquer	s_J∼COD_P	94	36	38	15	prévoir
subir	s_S∼COD_D	386	547	687	1008	faire, avoir
diagnostiquer	s_D	1127	1958	1313	452	découvrir, déceler
développer	s_P	2278	2359	1172	619	créer
réaliser	s∼COD_P	558	607	611	1191	faire
présenter	s∼COD_F	602	823	1057	2820	avoir, manifester
relever	s_P∼COI_P	136	122	63	3	passer, contribuer

sans catégorie sémantique entraînent la formation de variantes des patrons dont tous les arguments sont annotés sémantiquement. Prenons le cas des PSS s_J∼COD_D et s∼COD_D. Le premier correspond à la forme souhaitée de PSS, tandis que le second en est une variante, du fait de l'annotation sémantique partielle des arguments. En dehors de l'augmentation de la charge de travail manuel, ce phénomène n'a pas d'impact important sur la phase de constitution de la ressource de simplification.

Le nombre élevé de verbes dans certains groupes est la conséquence du caractère permissif de la construction de base. Les PSS ayant une construction de base restrictive sont instanciés par moins de verbes.

4.5.2 Résultats de l'alignement

Le tableau[14] 4.27 (cf. page suivante) fournit un extrait des résultats de l'alignement qui sont présentés en entier dans l'annexe D.11. La ressource, qui compte 230 entrées au total, est sauvegardée dans un fichier sous un format tabulé. Le verbe de chaque PSS est aligné avec un équivalent tiré de l'un des corpus non experts (FOR ou VUL), et éventuellement avec un ou plusieurs verbes synonymes qui n'ont pas pu rentrer dans le tableau 4.27 mais apparaissent en annexe. Cette synonymie a été établie sur la base de notre compétence linguistique, à partir de la fréquence des verbes en question (dans les corpus des non-experts) au sein des PSS concernés (cf. chapitre 3, section 3.5.1.2).

Le constat effectué à la section 4.1 par rapport aux choix préférentiels entre verbe et corpus (grâce à la fréquence verbale) se confirme à travers les résultats de l'alignement des PSS. Certains verbes très récurrents dans le corpus des experts (cf. tableau 4.2) ont pour équivalents

14. Pour faciliter la lecture du tableau, les étiquettes des catégories METIER et STATUT SOCIAL ont été remplacées par le nom de leurs principaux référents, respectivement MÉDECIN et PATIENT.

TAB. 4.27 – Quelques PSS alignés avec leurs équivalents.

PSS	Équivalents
MEDECIN dépiste MALADIE	MEDECIN identifie MALADIE
MEDECIN administre PCHIMIQUE	MEDECIN donne PCHIMIQUE (à PATIENT)
MALADIE se traduit par MALADIE	MALADIE se manifeste par MALADIE
MALADIE se produit	MALADIE se manifeste (chez PATIENT)
FONCTION se présente comme MALADIE	FONCTION se manifeste comme MALADIE
PATIENT pratique PROCEDURE	PATIENT fait PROCEDURE
FONCTION est associée à FONCTION	FONCTION est liée à FONCTION
PCHIMIQUE est poursuivi	PATIENT (continue de) prendre PCHIMIQUE
PROCEDURE associe PCHIMIQUES	PROCEDURE combine PCHIMIQUE
PCHIMIQUE est administré	(MEDECIN) donne PCHIMIQUE (à PATIENT)
MALADIE s'accompagne de MALADIE	MALADIE est suivie de MALADIE
PROCEDURE dépiste MALADIE	PROCEDURE détecte MALADIE
MALADIE se développe à partir de ANATOMIE	MALADIE commence dans ANATOMIE
ANATOMIE contrôle ANATOMIE	ANATOMIE commande ANATOMIE
PROCEDURE relève de PROCEDURE	PROCEDURE dépend de PROCEDURE
ORGVIVANT développe ANATOMIE	ORGVIVANT forme ANATOMIE
FONCTION est observée chez PATIENT	(MEDECIN) constate FONCTION chez PATIENT
PATIENT relève de MALADIE	PATIENT souffre de MALADIE
PROCEDURE révèle MALADIE	PROCEDURE montre MALADIE
PROCEDURE indique MALADIE	PROCEDURE signale MALADIE
ANATOMIE traduit FONCTION en FONCTION	ANATOMIE transforme FONCTION en FONCTION
MALADIE est traitée par PCHIMIQUE	PCHIMIQUE soigne MALADIE
PROCEDURE impose PROCEDURE	PROCEDURE passe par PROCEDURE
MEDECIN envisage PROCEDURE	MEDECIN prévoit PROCEDURE
PROCEDURE est réalisée	(MEDECIN) fait PROCEDURE
PATIENT présente FONCTION	PATIENT manifeste/a FONCTION
MALADIE se développe	MALADIE évolue
PROCEDURE abaisse FONCTION	PROCEDURE baisse FONCTION
MALADIE est observée chez PATIENT	MALADIE est rencontrée chez PATIENT
AGENT est développé	AGENT est créé (par qqun)
PROCEDURE est réalisée chez PATIENT	PROCEDURE est faite (par MEDECIN) sur PATIENT
PROCEDURE est développée	(MEDECIN) crée PROCEDURE
MALADIE s'accompagne de FONCTION	MALADIE est suivie de FONCTION
PROCEDURE est pratiquée	(MEDECIN) fait PROCEDURE
PROCEDURE est observée	(MEDECIN) fait PROCEDURE
PCHIMIQUE est indiqué dans PROCEDURE	(MEDECIN) conseille PCHIMIQUE dans PROCEDURE
MALADIE est observée chez PATIENT	(MEDECIN) découvre MALADIE chez PATIENT
FONCTION est évaluée par PROCEDURE	FONCTION est testée par PROCEDURE

des prédicats dont la plus haute fréquence est enregistrée dans les corpus pour non-experts. Les couples de verbes *dépister-découvrir*, *administrer-donner*, et *présenter-souffrir* illustrent cette remarque. Ainsi, le verbe *présenter*, qui a une fréquence de 646 dans le corpus des experts, apparaît le plus souvent dans le PSS PATIENT *présenter* MALADIE. Dans cet emploi, il a pour équivalent non spécialisé le verbe *souffrir*, qui figure dans la liste des prédicats identifiés en section 4.1 comme faisant partie des verbes les plus fréquents du corpus des non-experts (cf. tableau 4.4).

En outre, l'alignement nous a permis d'observer la façon dont les verbes supports sont utilisés dans le corpus des forums. Nous avons en effet constaté que les non-experts, eux aussi, utilisent fréquemment les verbes supports, non pas dans des constructions à verbes supports comme le font les experts, mais plutôt dans des contextes où les experts optent pour d'autres verbes :

— STASOCIAL développe MALADIE vs. STASOCIAL fait/subit MALADIE

— STASOCIAL poursuit MÉDICAMENT vs. STASOCIAL prend MÉDICAMENT

— METIER administre MÉDICAMENT vs. METIER donne MÉDICAMENT à STASOCIAL

La phase d'alignement des PSS a également permis de découvrir un phénomène non identifié dans les résultats de la section 4.1, qui pourtant signalait déjà l'attirance observée entre certains corpus et des verbes spécifiques. En effet, les résultats de l'alignement fournissent des éclaircissements par rapport aux différents modes d'utilisation des verbes par les non-experts et les experts lorsqu'ils veulent exprimer des situations propres au domaine médical. Au-delà de ce qui a été vu précédemment en ce qui concerne l'attraction corpus-verbe, la manière d'utiliser le verbe différencie également le langage des experts de celui des non-experts. Certains prédicats peuvent ainsi intervenir dans les deux corpus, mais la façon dont ils sont employés dans chaque type de corpus marque la différence. L'intervention de certains verbes dans un PSS, et surtout dans un corpus particulier, n'est donc pas le fruit du hasard. En général, le choix du prédicat verbal dépend non seulement du locuteur, mais surtout de ce qu'il désire exprimer, c'est-à-dire la situation qu'il souhaite décrire. Or, les deux communautés que constituent les experts et les non-experts utilisent des systèmes linguistiques différents. Chez les premiers, nous avons affaire à un langage standardisé et partagé par l'ensemble de la communauté tandis que chez les autres, le langage utilisé est très influencé par les paramètres socio-culturels qui caractérisent le locuteur.

C'est la raison pour laquelle un verbe peut avoir des acceptions qui tendent à être privilégiées au sein d'un corpus par rapport aux autres acceptions. Les choix préférentiels des corpus sont capturés grâce au(x) PSS dominant(s) dans les différents types de corpus. Cette remarque générale est valable pour tous les verbes, autant les fort/peu fréquents dans un corpus particulier que ceux ayant une fréquence relativement similaire dans l'ensemble des sous-corpus.

Prenons par exemple *diagnostiquer* et *détecter*. Ces verbes interviennent dans les deux corpus principaux (PRO et FOR), avec des fréquences différentes. Dans la section 4.1.2, nous avons vu

que certains PSS dominés par ces verbes sont très fréquents dans le corpus des experts, tandis que dans le corpus des forums d'autres PSS dominent. Cette variation fréquentielle au niveau des PSS des verbes selon le corpus montre que les experts et les non-experts ont chacun recours au verbe *diagnostiquer* d'une façon spécifique, qui est illustrée par des PSS particuliers. Les experts ont tendance à dire PATIENT *est diagnostiqué*. Dans cet emploi, le verbe a pour synonyme *identifier*, *repérer*. Alors que dans le corpus des forums, la plupart des occurrences du verbe *diagnostiquer* correspondent à la construction suivante : MÉDECIN *diagnostique* MALADIE *chez* PATIENT. Dans ce PSS, le verbe peut être remplacé par *détecter*, *découvrir*, *observer*, qui instancient très souvent cette construction dans le corpus des forums. De la même façon, les experts ont tendance à utiliser le PSS PATIENT *est détecté par* PROCÉDURE. Chez les non-experts, par contre, s'il faut automatiquement utiliser le verbe *détecter* pour décrire une telle situation, ils seront plus enclins à dire PROCÉDURE *détecte* MALADIE *(chez* PATIENT*)*. Mais dans la réalité, pour décrire cette situation dans leur langage quotidien, les non-experts opteraient plutôt pour un verbe comme *montrer* sous la syntaxe suivante : PROCÉDURE montre MALADIE chez PATIENT.

Nous avons donc affaire à un cas de figure où des verbes sont utilisés dans les deux corpus principaux (expert et forum) mais avec des significations/syntaxes différentes, mises en évidence grâce aux PSS qui les accueillent. Cette remarque est d'une importance capitale car elle expose l'une des principales divergences qui caractérisent les discours des experts et des non-experts en ce qui concerne l'utilisation de verbes et d'autres unités linguistiques pour parler de concepts médicaux. Chacun de nos deux groupes de protagonistes du domaine médical a un système linguistique et terminologique qui lui est propre et qui est capturé dans ce travail à travers les ensembles de PSS verbaux qui constituent le discours de chaque type de locuteur. Ce constat rejoint l'argument de Zeng-Treiler & Tse (2006) selon lequel les experts et les patients utilisent parfois des expressions similaires ou identiques mais avec des interprétations différentes.

TAB. 4.28 – Récapitulatif du résultat de la simplification.

Nombre de pss	**Forme active**	**Forme passive**	**construction en « se »**
230	112	92	26

Les données du tableau 4.27, qui ne présentent qu'un extrait des résultats de l'alignement, permettent déjà de remarquer la forte présence des constructions passives parmi les PSS à simplifier, c'est-à-dire ceux qu'utilisent les experts. Le tableau 4.28 présente une répartition du contenu de la ressource de simplification autour des types de constructions répertoriées. D'après ce tableau, l'on dénombre au total 92 PSS au passif, c'est-à-dire 40% de la ressource.Cette précision concorde avec les résultats obtenus à la section 4.1.2 qui signalaient déjà la forte inclination des experts pour la forme passive, plus particulièrement celle avec omission de l'agent. De même, dans le deuxième chapitre de cette thèse (cf. section 2.1.3.1), nous avons vu que

l'emploi de la forme passive est une caractéristique des textes scientifiques et techniques. Nous avons également vu que ce procédé est particulièrement prisé dans les textes médicaux. La simplification de ce type de PSS avec agent omis passe par la restitution de l'entité qui joue le rôle d'agent et, si nécessaire, par une transformation de la construction vers la forme active (comme dans l'exemple 49), pour une compréhension plus aisée. Lorsqu'un constituant est restitué, nous le mettons entre parenthèse pour indiquer qu'il s'agit d'un ajout. Ce système d'annotation est appliqué dans la ressource.

PCHIMIQUE est poursuivi → (STASOCIAL) continue de prendre PCHIMIQUE

49) *Si le nadolol est poursuivi jusqu'à l'accouchement, en informer l'équipe de la maternité pour lui permettre d'adapter la surveillance du nouveau-né.*
→ Si la patiente continue de prendre le nadolol jusqu'à l'accouchement, en informer l'équipe de la maternité [...].

Au terme de l'alignement, 54 PSS à la forme passive ont subi un ajout (restitution de l'agent et/ou d'un autre élément), et plus de 60 ont subi une transformation vers la forme active.

Une analyse des résultats de l'alignement des PSS permet de remarquer que dans certains PSS, plus particulièrement ceux impliquant les catégories comme FONCTION DE L'ORGANISME et *organisme vivant*, le sens du verbe, et par conséquent l'équivalent (proposé pour la substitution), dépend des termes qui instancient les catégories Snomed en position d'arguments. Plus précisément, le sens du verbe dépend du sous-type d'entités auxquelles ces termes font référence. Par exemple, les PSS[15] FONCTION *est activée* et FONCTION *est abaissée* impliquent tous les deux la catégorie des fonctions de l'organisme. Cependant, ils font référence à différents sous-types de fonctions :

50) FONCTION est activée : *le système rénineangiotensinealdostérone est activé et cette activation limite la baisse de la pression artérielle.*

51) FONCTION est abaissée : *dans ce cas, la pression artérielle doit être abaissée en dessous de 185 / 110 mmHg avant de débuter le traitement fibrinolytique.*

Comme le montrent les exemples 50 et 51, le premier PSS ne peut être instancié que par des noms renvoyant à des paramètres du corps tels que le poids, la tension, la pression artérielle, le débit sanguin, etc. tandis que dans le deuxième PSS, il s'agit des molécules ou des systèmes qui participent au bon fonctionnement du corps. Cette distinction qui caractérise les termes d'une même catégorie Snomed est liée à la structuration interne des classes de la terminologie. Comme nous l'avons expliqué lors de la description de la terminologie Snomed, certaines catégories (organisme vivant, fonction de l'organisme, etc.) contiennent des termes dénotant différents sous-types d'entités. Par exemple, la catégorie des organismes vivants contient des

15. Ces PSS correspondent aux entrées 102 et 120 de la ressource de simplification.

noms d'animaux, de plantes, de bactéries et de virus. Cette absence de sous-catégorisation débouche sur le constat que nous avons fait lors de l'alignement des PSS. La sous-catégorisation des termes des différentes classes de la terminologie Snomed permettrait à nos PSS d'offrir plus de précision sur le plan sémantique.

La tâche d'alignement des PSS nous a permis de faire une observation essentielle en ce qui concerne la simplification de textes spécialisés pour un public de non-experts. La clé de la simplification, c'est-à-dire la mise en relation des textes des experts avec ceux des non-experts d'un domaine, est le point de vue sur lequel on focalise le texte simplifié. Fleischman (2003), reprenant les travaux de Anspach (1988), considère le point de vue ou la voix comme une caractéristique discursive des textes médicaux. En effet, les analyses faites sur les différents corpus révèlent qu'au-delà des spécificités syntaxiques et sémantiques qui opposent les textes écrits par les experts aux textes écrits par les non-experts, une différence fondamentale est le point de vue de l'énonciateur. Par point de vue nous entendons la position dans laquelle se place celui qui parle par rapport à ce qu'il dit : est-ce du point de vue du médecin ? Est-ce de celui du patient ?

Différents types de marqueurs peuvent indiquer le point de vue dans un texte, entre autres les pronoms personnels. Sans pour autant faire une étude de tous ces marqueurs dans notre corpus, nous tenons à souligner que le corpus des forums dénombre 39 939 occurrences de pronoms personnels, plus particulièrement les première et deuxième personnes du singulier, tandis que chez les experts (PRO), à peine 1000 occurrences ont été dénombrées. L'hypothèse généralement émise en ce qui concerne les pronoms personnels est qu'ils sont spécifiques aux documents pour non-experts. En effet, dans ces documents, les énonciateurs s'adressent directement à leurs destinataires, alors que dans les documents scientifiques les auteurs privilégient un style impersonnel. Dans une étude dont le but est la catégorisation des pages Web médicales selon qu'elles distribuent des informations pour les experts ou pour les non-experts, Grabar *et al.* (2007) constatent que les pronoms personnels à la première personne du singulier *je, tu* et du pluriel (*nous, vous*) sont spécifiques aux textes des non-experts, tandis que dans les documents scientifiques, l'utilisation de certains de ces pronoms relève de l'expression des questions qui seraient directement posées aux patients.

Dans le corpus des forums, le point de vue transparaît également à travers la prédominance des PSS complets du type sujet-verbe-complément(s), qui déploient tous les éléments syntaxiques nécessaires pour que le lecteur puisse mieux cerner le sens du verbe et son contexte d'utilisation. Les verbes *découvrir* et *diagnostiquer*, qui comptent respectivement 38 et 25 PSS du type MÉDECIN *verbe* MALADIE *chez* STASOCIAL dans le FOR, peuvent servir d'illustration grâce à la récurrence de phrases telles que :

52) *mon médecin m'a découvert/diagnostiqué une appendicite*.

Pourtant, dans le corpus des experts, le point de vue est tout autre. Il est marqué de façon

différente (types de pronoms), entre autres par la prépondérance de PSS au passif (Biber & Conrad, 2009 ; Todirascu *et al.*, 2012) et d'autres constructions qui se caractérisent par l'omission d'éléments syntaxiques tels que l'agent. La notion de point de vue est fortement liée au type de texte. Ceci est d'autant plus vrai que certains éléments ici soulignés font partie des caractéristiques mentionnées dans le chapitre 2 lors de la description des différents corpus (cf. section 2.1.3.1).

Pour relier tout ce qui précède au travail de simplification, nous pouvons dire qu'au terme de l'alignement, nous avons constaté que tous les procédés mis en application dans ce processus ont contribué à changer le point de vue des PSS spécialisés qui constituent les entrées de la ressource. L'ensemble des transformations appliquées aux PSS, sur les plans syntaxique et lexical, ont contribué non seulement à réécrire ces PSS dans une syntaxe facilement compréhensible, mais surtout à réorienter leur point de vue vers les principaux bénéficiaires de la ressource, les non-experts. Dans le cadre de cette démarche, le choix des unités lexicales et/ou des constructions a contribué à ce que des PSS équivalents obtenus puissent décrire les situations selon le point de vue des non-experts, afin que ces derniers puissent s'y identifier et comprendre aisément les équivalents proposés comme résultats de la simplification.

4.6 Corpus et collocations verbe-terme

Dans cette section, il est question de présenter les résultats de l'analyse contrastive des collocations verbe-terme dans les quatre sous-corpus. Les collocations verbe-terme concernent uniquement les cooccurrences entre le verbe et les compléments COD et COI. Ces collocations ont été étudiées selon les points de vue syntaxico-sémantique et lexical.

4.6.1 Collocations verbe-terme et variations syntaxico-sémantique

Le tableau 4.29 propose un prototype des résultats obtenus au terme de l'analyse comparative des collocations. Grâce aux extractions effectuées automatiquement, nous avons été à même d'identifier, pour chaque verbe, la fonction syntaxique favorite entre COD et COI, et surtout la/les catégorie(s) sémantique(s) préférée(s) des termes qui jouent ces rôles syntaxiques, dans chaque corpus. Le verbe *accompagner* est présenté en guise d'exemple.

D'après le tableau 4.29, D et F représentent les catégories sémantiques préférées du verbe *accompagner* dans les quatre corpus. La fonction syntaxique COI est la plus sollicitée dans les corpus PRO, ETU et VUL, ce qui est cohérent avec le constat fait grâce au tableau 4.15 qui indiquait la prédominance de la construction transitive indirecte (*sujet est accompagné de* COI) dans le corpus des experts.

L'analyse des collocations verbe-terme a permis d'identifier les catégories sémantiques dominantes dans chaque corpus. De façon générale, les catégories P, D (et en troisième position F)

TAB. 4.29 – Exemples de résultats de l'analyse des collocations verbe-terme : cas du verbe *accompagner*.

cat. Sno.	PRO			ETU			VUL			FOR		
	nbocc	**cod**	**coi**	**nbocc**	**cod**	**coi**	**nbocc**	**cod**	**coi**	**nbocc**	**cod**	**coi**
A	2	-	2	2	-	2	5	-	5	-	-	-
C	1	-	1	0	-	-	-	-	-	2	-	2
D	12	-	12	66	6	60	26	-	26	4	3	1
F	35	-	35	67	-	67	46	-	46	9	7	2
J	0	-	-	0	-	-	-	-	-	1	-	1
L	0	-	-	0	-	-	1	-	1	-	-	-
P	29	5	24	26	26	-	8	-	8	1	-	1
S	0	-	-	0	-	-	3	-	3	1	-	1
T	1	-	1	3	-	3	0	-	-	2	-	2

occupent la tête du classement dans l'ensemble du corpus. Le tableau 4.30 donne des exemples de verbes qui sollicitent régulièrement ces deux catégories dans l'ensemble du corpus.

TAB. 4.30 – Quelques verbes autour des catégories fréquentes du corpus : P et D.

	Catégorie D	Catégorie P
Verbes	diagnostiquer, présenter, découvrir, développer, dépister, évoquer, causer, signaler	relever, réaliser, indiquer, proposer, administrer, envisager, envisager, subir, appliquer, pratiquer, évaluer

La catégorie P se démarque dans les corpus PRO et ETU, tandis que dans les corpus des forums, la catégorie D est la plus sollicitée. Ce qui signifierait que les non-experts parlent beaucoup plus des maladies que des procédures médicales, qui semblent être le sujet principalement abordé dans les textes des experts, plus précisément dans ceux qui constituent notre corpus PRO.

La catégorie favorite du verbe peut changer lorsqu'on passe d'un corpus à l'autre. Par exemple, le verbe *évoquer* a pour catégorie favorite P dans les corpus PRO et ETU, tandis que dans le corpus VUL, c'est la catégorie F qui occupe la première place de son classement. Dans le corpus des forums, le verbe n'a que 5 occurrences parmi lesquelles aucune des catégories ci-dessus n'intervient.

De même, certains verbes peuvent avoir deux catégories favorites en compétition dans un corpus alors que dans d'autres corpus, cette compétition ne se manifeste pas. Ce constat concerne les verbes comme *diagnostiquer*, qui en position de COD a deux catégories de prédilection dans les corpus PRO et ETU (D et S), tandis que dans les corpus VUL et FOR, seule la catégorie D domine en position de COD. Le verbe *évaluer* illustre également ce mode de fonctionnement, en oscillant entre les catégories S, F et P dans le corpus PRO, tandis que dans le corpus des forums, seule la catégorie F se démarque sur les 7 occurrences du verbe.

Comme nous l'avons vu dans la section précédente, sur le plan sémantique, *diagnostiquer un patient* n'est pas l'équivalent de *diagnostiquer une maladie*, de même qu'*évaluer quelqu'un* ('examiner'/'contrôler') n'est pas l'équivalent d'*évaluer une fonction de l'organisme* ('mesurer'), et encore moins d'*évaluer une procédure* ('tester'). Le faible nombre d'occurrences du verbe *évaluer* pourrait remettre en question notre argumentation sur le choix préférentiel de catégories sémantiques. Cependant, les cas de figure tels que celui du verbe *diagnostiquer*, qui enregistre 85 occurrences dans le corpus des forums, parmi lesquelles aucune ne correspond à la notion *diagnostiquer quelqu'un*, soutiennent notre analyse.

— *évaluer* S : *Le Docteur A évalue le patient à 5 h et ne découvre aucune anomalie physique.*

— *évaluer* F : *Quel que soit le support nutritionnel adopté, on doit évaluer régulièrement la tolérance et l'efficacité par la surveillance régulière du poids, de la tension artérielle [...].*

— *évaluer* P : *Le médecin de Centre d'examens de santé (CES) évalue l'indication de la consultation d'oncogénétique par le score d'Eisinger.*

Comme le signalent les exemples ci-dessus, la variation de termes-objets traduit la polysémie des verbes et surtout le caractère spécifique de certains emplois des verbes pour un type de texte particulier. Dans le cas présent, il s'agit du corpus des experts. Nous avons en effet affaire à des choix préférentiels de catégories sémantiques donnant accès à des notions qui semblent être propres aux experts. Ces choix préférentiels des verbes reflètent ce que Heid (1994) appelle les *propriétés sélectionnelles des unités lexicales*. Ces dernières déterminent le sens du verbe en contexte (comme dans les exemples ci-dessus), selon la nature de son complément. Une étude approfondie de la cooccurrence verbe-terme permettra d'en savoir plus à propos de ces notions. Il sera question d'analyser les cooccurrences entre les verbes et des termes portant les catégories favorites de ces verbes. Dans cette démarche, une attention particulière sera portée à la fréquence des paires (verbe-terme) dans les quatre types de corpus. En procédant ainsi, nous serons à même d'identifier les corpus qui manifestent un lien vis-à-vis d'une paire verbe-terme particulière, ceci grâce à la fréquence obtenue par cette paire.

4.6.2 Collocations verbe-terme et variation lexicale

En analysant les collocations verbe-terme d'un point de vue syntaxico-sémantique, nous avons vu que les verbes montrent parfois une attirance pour les compléments appartenant à des catégories sémantiques spécifiques, ceci dans l'ensemble des corpus ou bien dans certains corpus uniquement. Le plus souvent, cette affinité corpus-catsnomed se manifeste à travers le lexique utilisé. C'est ce qui sera étudié dans cette sous-section. Pour y parvenir, nous allons analyser certaines cooccurrences verbe-terme fréquemment observées dans l'un des deux corpus principaux, à savoir le corpus des experts ou celui des forums. Ainsi, nous pourrons contraster les rapports que les différents corpus entretiennent avec ces collocations. Le tableau 4.31 propose

quelques collocations, chaque verbe est suivi du terme collocatif. Le nombre d'occurrences de chaque collocation est founi, pour chacun des quatre corpus :

TAB. 4.31 – Quelques collocations verbe-terme dans le corpus.

Verbe	**Cooccurrences nominales**				
	Termes	PRO	ETU	VUL	FOR
prescrire	*traitement*$\mathcal{P}$	3	0	1	7
	examen$\mathcal{P}$	0	0	4	7
	médicament$\mathcal{C}$	3	1	15	26
subir	*ablation*$\mathcal{P}$	0	0	1	39
	intervention$\mathcal{P}$	6	2	1	30
	AVC$\mathcal{D}$	0	0	2	12
augmenter	*tension*$\mathcal{F}$	0	0	7	14
	risque/risque de$\mathcal{F}$	26	33	5	7
baisser	*tension*$\mathcal{F}$	0	0	4	18
consulter	*médecin*$\mathcal{F}$	4	2	7	41
exposer	*à+risque*$\mathcal{F}$	14	11	0	3
	patient$\mathcal{S}$	23	5	1	0
suivre	*apparition de symptômes*$\mathcal{F}$	5	0	0	0
	patient$\mathcal{S}$	10	5	1	0
	régime$\mathcal{F}$	1	0	1	5
	conseil	0	0	15	10
	traitement$\mathcal{P}$	4	2	1	13
évaluer	*patient*$\mathcal{S}$	13	7	0	0
	indication	6	0	0	0
	risque$\mathcal{F}$	9	2	0	1
évoquer	*diagnostic (de* $\mathcal{D}$*)*	11	7	0	0
appliquer	*recommandation*$\mathcal{P}$	12	5	1	1
	méthode$\mathcal{P}$	7	0	0	6
	règle$\mathcal{P}$	8	4	0	0
	traitement$\mathcal{P}$	4	5	1	0

Les collocations que propose le tableau 4.31 ont été repérées à partir des catégories favorites des verbes concernés. D'après les données de ce tableau, les corpus PRO et ETU convergent, ayant des fréquences très souvent similaires. Le corpus VUL maintient sa position intermédiaire entre les corpus PRO et ETU et le corpus des forums. Les fréquences le rapprochent soit des deux premiers, soit du corpus FOR. Par contre, la fréquence de ces paires verbe-terme varie grandement lorsqu'on va d'un bout du continuum (corpus PRO) à l'autre (corpus FOR). Cette variation de fréquence n'est pas anodine car l'écart entre les fréquences de deux corpus pourrait signaler une différence en ce qui concerne le niveau de spécificité de la paire vis-à-vis de chacun de ces corpus. Une fréquence élevée indiquerait un haut degré de proximité, tandis qu'une faible fréquence indiquerait un faible degré de proximité.

Par exemple, d'après les données du tableau 4.31, les non-experts utilisent fréquemment

la collocation *suivre conseil*, tandis que les experts tendent plutôt à utiliser régulièrement les collocations *appliquer recommandation*, *appliquer méthode* et *appliquer règle*. La variation lexicale intervient à deux niveaux dans ces exemples : au niveau du choix des verbes et au niveau du choix des termes. Ces collocations impliquent des termes COD différents mais sémantiquement proches (*conseil, recommandation*). Les verbes sélectionnés sont également différents mais apportent une nuance qui marque la dissemblance dans l'interprétation de ces expressions. Il s'agit de la divergence de points de vue (entre les textes PRO et FOR) dont il a été question à la section 4.5.2. En effet, le contraste lexical (choix du verbe) qui caractérise ces collocations exprime la divergence de points de vue entre les textes des experts et ceux des forums. D'un côté, nous avons des experts qui parlent des recommandations, méthodes et règles qu'ils appliquent ou doivent appliquer dans l'exercice de leurs fonctions. À l'opposé, les patients parlent des conseils qu'ils cherchent, reçoivent et/ou suivent au cours du processus de soin.

Afin de découvrir d'autres collocations et de pousser l'analyse plus loin, nous avons extrait les verbes qui cooccurrent fréquemment avec les termes de la colonne "termes" du tableau 4.31. Les tableaux 4.32 et 4.33 fournissent un extrait des résultats de cette expérience, respectivement dans les corpus PRO et FOR. Dans ces tableaux, la couleur bleue met en évidences les données tirées du corpus des non-experts et la couleur rouge, celles tirées du corpus des experts. Pour chaque unité nominale proposée dans la colonne "termes", qui d'après le tableau 4.31 cooccurre régulièrement avec le verbe de la dernière colonne du tableau, nous fournissons une liste de verbes tirés du corpus. Ces verbes représentent les principaux cooccurrents du terme dans l'autre corpus. Le nombre d'occurrences de la paire verbe-terme est également fourni.

TAB. 4.32 – Illustration de la préférence lexicale des verbes dans le corpus PRO.

Argument COD	**Cooccurrence verbale**	
	Expert	**Forum**
médicament	indiquer(15), recommander(13) proposer(2)	
traitement	indiquer(20), envisager(10) nécessiter(7), recommander(3)	prescrire
examen	imposer(1), proposer(1) recommander(1), autoriser(1)	
intervention *AVC*	nécessiter(6),bénéficier(2) présenter(4), faire(2), avoir(2)	subir
tension	constructions nominales	baisser
conseil	proposer(9), bénéficier(2) considérer(1)	
traitement	recevoir(12), bénéficier(10) faire(6), poursuivre(6),	suivre

Les données du tableau 4.32 permettent d'observer de près quelques choix préférentiels verbes-

termes qui caractérisent les corpus PRO et FOR. Selon les types de corpus, les termes tendent à se combiner à des verbes spécifiques. Autrement dit, les rédacteurs opèrent des choix précis de verbes. Un cas particulièrement frappant est celui de *traitement*. Les traitements, qui dans les textes des non-experts sont en général prescrits, sont très souvent, chez les experts, indiqués, proposés ou envisagés. Par ailleurs, la variation lexicale des collocations peut également se manifester au niveau de la syntaxe, à travers les types de constructions favorites. En effet, les résultats de nos extractions montrent qu'à la place des collocations verbales, les experts optent parfois pour des constructions nominales. C'est ce qui est observé pour la collocation *baisser tension* qui, chez les experts, se réalise à travers les expressions nominales *abaissement tensionnel*, *abaissement du niveau tensionnel* et *baisse de tension*, fréquemment utilisées. Cette aptitude des experts à utiliser les collocations nominales à la place des collocations verbales est cohérente puisque, par nature, les textes des experts sont en majorité constitués d'entités nominales. Les résultats obtenus à la section 4.1 contribuent à confirmer la validité de cet argument pour notre corpus.

Nous pouvons percevoir, à travers ce phénomène, deux façons de parler distinctes, deux façons d'exprimer les mêmes concepts, chacune étant propre à un type de locuteurs. Les « profanes » préfèrent les expressions verbales, tandis que les experts oscillent entre les expressions verbales et nominales.

TAB. 4.33 – Illustration de la préférence lexicale des verbes dans le corpus FOR.

Argument	**Cooccurrence verbale**	
	Forum	**Expert**
patient	traiter(1), voir(1) rencontrer(2) recevoir(3)	suivre
risque *patient* *indication*	mesurer(4), juger(3) - apprécier(1)	évaluer
risque	accroître(3), multiplier(2) élever(1)	augmenter
méthode *règle*	utiliser(5) respecter(6)	appliquer
diagnostic	faire(10), donner(6) signaler(8), poser(3) avoir(5)	évoquer

À la lecture du tableau 4.33, la première remarque est la suivante : la correspondance n'est pas toujours totalement garantie entre les corpus PRO et FOR en ce qui concerne les collocations. En effet, pour certains types de collocations identifiées dans le corpus PRO, un faible nombre de correspondants existe dans le corpus des forums, et ces derniers interviennent très rarement.

Par exemple, les expressions *évaluer un patient* et *évaluer le risque* ont respectivement 13 et 9 occurrences dans le corpus des experts, contre une absence quasi-totale d'occurrences dans le corpus des forums (cf. tableau 4.33). Le verbe *mesurer* (*mesurer le risque*), possédant à peu près le même sens qu'*évaluer* lorsqu'il est associé à *risque*, n'intervient qu'une fois en cooccurrence avec le terme *risque*. Quant au terme *patient*, il ne s'adjoint à aucun verbe synonyme d'*évaluer* dans le corpus des forums.

Ce constat pourrait traduire le faible degré de corrélation qui caractérise les deux corpus (et donc les experts et non-experts) en ce qui concerne certains concepts médicaux, dans le cas présent *évaluer un patient*. L'absence de collocations équivalentes dans le corpus des non-experts pourrait signifier qu'il s'agit d'une notion pas/peu connue ou pas familière à ces acteurs du domaine médical. En effet, la collocation *évaluer un patient* renvoie à un ensemble de procédures médicales que les experts rassemblent généralement sous le terme *l'évaluation du patient*. L'évaluation du patient joue un rôle déterminant dans la prise en charge et le suivi de ce dernier. C'est sans doute l'une des raisons qui expliquent la récurrence de cette notion dans les écrits des experts. Cette pratique effectuée par les professionnels de la santé consiste à contrôler de façon globale l'état physique, psychologique et l'environnement social du patient. Elle a plusieurs finalités correspondant aux formes d'évaluations qui, d'un point de vue linguistique, peuvent être capturées à travers au moins deux collocations verbe-terme :

— *examiner un patient* : connaître le patient et reconnaître sa maladie via un ensemble d'examens.

Bien que le résident en chirurgie générale ait évalué la patiente et documenté ses antécédents de TVP, le patron chirurgien n' était pas informé de ce renseignement clinique.

— *contrôler un patient* (c'est-à-dire vérifier son état) : assurer la continuité des soins en contrôlant l'évolution de l'état du patient et l'effet du traitement qu'il suit.

Le Docteur A évalue le patient à 5 h et ne découvre aucune anomalie physique significative.

Ce deuxième type d'évaluation ne s'effectue que lorsque le patient est bien connu du médecin, c'est-à-dire qu'il a déjà été sujet au premier type d'évaluation, qui a permis à son médecin de le connaître, de savoir ce dont il souffre et d'obtenir d'autres informations relatives à sa santé.

L'expression *évaluer un patient* est donc ambiguë car elle peut avoir diverses interprétations, selon le contexte dans lequel elle est utilisée. Dans le corpus des forums, la notion *examiner un patient* est bien présente. Elle est d'ailleurs bien connue des non-experts, puisqu'ils en parlent. Mais sur le plan linguistique, cette notion se réalise d'une façon différente, notamment à travers l'emploi du PSS J *examine S*. Tandis que les experts expriment ce concept au moyen de la

collocation verbe-terme [16] (*évaluer un patient*), les non-experts tendent plutôt à utiliser le PSS J *examine S*, qui se réalise à travers les phrases du type *mon médecin m'a examiné*.

L'absence de l'expression *examiner patient* chez les non-experts a tout son sens dans la mesure où le patient, qui écrit sur un forum, parle de lui-même, de son expérience ou de l'expérience d'un proche et de ce fait, utilise le verbe (en l'occurrence *examiner*) en combinaison avec des pronoms personnels qui renvoient à lui-même ou aux personnes concernées. Ce constat nous ramène au phénomène de divergence de points de vue mainte fois évoqué : tandis que l'expert utilise des termes tels que *le patient, le malade* pour faire référence au patient, le non-expert tend à utiliser un pronom (*je, me, moi*) ou un nom (*ma soeur, ma mère, mon ami*) qui décrit le ou les protagoniste(s) impliqué(s) dans les faits qu'il relate.

Toutefois, les résultats de notre étude mettent également en évidence des cas qui traduisent une sorte d'unanimité entre les experts et les non-experts. Il s'agit des collocations verbe-terme utilisées communément par les deux groupes, pour exprimer les mêmes concepts médicaux. C'est le cas de *appliquer méthode* qui apparaît avec des fréquences relativement proches dans les deux corpus principaux ici comparés (4.31). Ce type de collocations renverrait aux connaissances partagées par les experts et les patients.

La fréquence des collocations verbes-objet et le mode de fonctionnement qu'elles déploient dans les corpus font d'elles des entités linguistiques dont l'extraction (Heid, 2001), la description (Heid, 2009), et même l'intégration dans les ressources dictionnairiques existantes, s'avère nécessaire (Heid & Freibott, 1991 ; Heid, 1994). C'est dans cette démarche que nous avons entrepris une enquête portant sur quatre dictionnaires français existants : le Larousse [17] (en ligne), le Petit Robert (2009), le TLFi, et le Dictionnaire Larousse Médical. L'objectif de cette étude (Wandji Tchami *et al.*, 2016) était de questionner et de comparer le contenu de ces principaux dictionnaires français, en portant une attention particulière à la présence des collocations et surtout au type de description qui leur était attribuée. À l'issue de cette étude, force a été de constater que les collocations verbe-terme sont quasiment absentes des ressources questionnées. Certains dictionnaires ne les intègrent pas du tout, tandis que d'autres en prennent uniquement quelques-unes en considération, et ceci non pas à la nomenclature mais à des endroits quelconques de la microstructure. En ce qui concerne la définition de ces entités, il a été observé qu'elles ne bénéficient pas d'un modèle de description homogène et cohérent dans les dictionnaires. Enfin, sur le plan méthodologique, nous avons constaté que dans la plupart des cas, le lexicographe n'attire pas l'attention du lecteur sur la nature spécialisée de l'expression verbale décrite. Ces résultats montrent que l'analyse des collocations verbe-terme est importante et peut être d'un apport considérable pour la constitution de ressources adéquates non seulement en lexicographie et en terminographie, mais aussi dans le domaine de la simplification de textes.

16. Ce qui n'exclut en rien l'utilisation du PSS J *examine* S dans le corpus des experts.
17. http ://www.larousse.fr/dictionnaires/francais

Chapitre 5

Évaluation des méthodes, ressources et outils utilisés

Dans ce chapitre, nous abordons la méthode et/ou les outils utilisés pour évaluer l'annotation syntaxique et l'annotation sémantique des corpus. Dans un travail de recherche requérant l'utilisation de ressources externes, il est indispensable d'avoir une bonne connaissance du type de résultats que fournissent les outils et surtout des conséquences qu'entraînerait l'utilisation de ces outils. Ceci permet de prendre des mesures afin d'anticiper, dans la mesure du possible, certains problèmes, de façon à s'assurer que les résultats obtenus permettent d'atteindre facilement les objectifs visés, tout en faisant montre d'un certain degré de fiabilité.

Le contenu de la ressource de simplification que nous proposons est aussi évalué. En effet, cette évaluation a pour but de nous permettre d'apprécier la qualité, la fiabilité et surtout de juger le caractère ergonomique des informations fournies par notre ressource de simplification. Ainsi, nous présentons l'évaluation menée et les démarches appliquées pour ce faire.

5.1 Évaluation de l'annotation syntaxique des corpus avec Cordial Analyseur

5.1.1 L'annotateur Cordial dans les campagnes d'évaluation EASY et PASSAGE

5.1.1.1 La campagne d'évaluation EASY

EASY [1] (Évaluation d'Analyseurs Syntaxiques) est le sigle qui fait référence à l'une des 8 campagnes d'évaluation des technologies de la langue du projet EVALDA du programme TECHNOLANGUE. « Le projet EVALDA a pour objectif la constitution d'une infrastructure d'évaluation des systèmes d'ingénierie linguistique du français, pérenne et permanente, et son exploitation par la mise en oeuvre de plusieurs expérimentations » [2]. Il est organisé à l'échelle nationale et concerne tout acteur qui voudrait évaluer son propre système par rapport à une norme et à d'autres systèmes en compétition.

La campagne EASY est l'une des premières campagnes d'évaluation de l'analyse syntaxique du français à grande échelle. Cette campagne ouverte a pour but de concevoir une méthodologie d'évaluation des analyseurs syntaxiques du français et, à partir des résultats obtenus, créer une ressource linguistique validée. Elle permet non seulement « d'évaluer la valeur des analyseurs mais aussi d'évaluer l'aptitude de leurs développeurs à se conformer à une norme » (Laurent *et al.*, 2009). 15 analyseurs, proposés par 13 participants, ont été mis en compétition. Le tableau 5.1 présente les différents participants.

1. http ://www.technolangue.net/article.php3 ?id_article=198
2. http ://www.technolangue.net/imprimer.php3 ?id_article=20

TAB. 5.1 – Liste des participants à la campagne EASY.

Nom	Signification	Location
ERSS [3]	Équipe de Recherche en Syntaxe et Sémantique	Toulouse
FT R&D [4]	France Télécom Recherche & Développement	Lannion
INRIA [5]	Institut National de Recherche en Informatique et en Automatique	Rocquencourt
LATL [6]	Laboratoire d'Analyse et de Technologie du Langage	Genève
LIC2M [7]	Centre d'Intégration des Systèmes et des Technologies	Île-de-France
LIRMM [8]	Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier	Montpellier
LORIA [9]	Laboratoire Lorrain de Recherche en Informatique et ses Applications	Nancy
LPL [10]	Laboratoire Parole et Langage	Aix-en-Provence
SYNAPSE [11]	Éditeur de logiciels spécialisé en Intelligence Artificielle appliquée au texte	Toulouse
SYSTAL [12]	Solutions de Recherches PERTinentes et IMMédiates	Asnières-sur-Seine
TAGMATICA [13]	Société spécialisée en développement logiciel en informatique documentaire et linguistique	Paris
VALORIA [14]	Laboratoire de Recherche en Informatique et ses Applications de Vannes et Lorient	Vannes
XRCE [15]	Xerox Research Centre Europe	Grenoble

Les participants sont appelés à utiliser la même segmentation en mots et en énoncés et le même formalisme, qui permet de délimiter les constituants et d'annoter leurs relations (fonctions syntaxiques). 14 relations de dépendance ont été définies pour 6 types de constituants : nominal, adjectival, prépositionnel, adverbial, verbal et prépositionnel verbal (pour les infinitives introduites par une préposition). Ces relations sont : (1) sujet-verbe, (2) auxiliaire-verbe, (3) COD, (4) complément-verbe, (5) modifieur de nom, (6) modifieur de verbe, (7) modifieur d'adjectif, (8) modifieur d'adverbe, (9) modifieur de préposition,(10) complémenteur, (11) attribut du sujet/objet, (12) coordination, (13) apposition, (14) juxtaposition. La relation complément-verbe concerne les autres compléments du verbe, exprimés sous forme de groupes prépositionnels ou de préposition-verbe, que ce soit les circonstants ou les COI. Ces constituants et relations sont amplement décrits dans Vilnat *et al.* (2004).

Les textes ont été fournis par 5 fournisseurs de corpus (L'ATILF, le LLF, le DELIC, le STIM et ELDA) dont le rôle a été de collecter des textes de différents genres et les annoter de façon à constituer de vastes corpus qu'il serait très difficile voire impossible d'annoter manuellement. Ainsi, le corpus contient différents types de textes : des articles de journaux (fournis par *Le Monde*), des textes littéraires (tirés de la ressource frantext de L'ATILF), des textes médicaux (pathologies et traitements), des questions (issues de la campagne EQUER de TECHNOLANGUE), des transcriptions de débats parlementaires (Sénat français, Parlement européen), des pages

Web du site ELDA, des courriers électroniques et des transcriptions de la parole.

L'évaluation porte sur chaque type de constituant, relation et corpus, pris individuellement. Les mesures d'évaluation sont :

— la précision (nombre d'éléments pertinents sur l'ensemble des résultats proposés par un système) ;

— le rappel (nombre d'éléments pertinents sélectionnés par un système sur l'ensemble des résultats pertinents disponibles) ;

— la f-mesure, qui combine les deux autres.

Les résultats de la campagne EASY ont été publiés en respectant l'anonymat des répondants. Les résultats de l'analyseur Cordial que nous présentons ont été en partie tirés du rapport de Laurent *et al.* (2009) qui révèle que lors de la compétition, Cordial portait l'étiquette P10. Grâce à cette information, nous avons pu collecter davantage de résultats décrivant les performances de l'outil dans Paroubek *et al.* (2007).

5.1.1.2 La (pré-)campagne PASSAGE (2007)

PASSAGE [16] (Produire des Annotations Syntaxiques à Grande Échelle) est une campagne qui est partie des résultats de l'évaluation EASY, en se basant sur plusieurs protocoles de cette dernière, mais avec des corpus différents et de taille plus grande.

Objectifs principaux :

— évaluer les analyseurs du français ;

— améliorer leur exactitude et leur robustesse sur des corpus de grande taille (270 millions de mots) ;

— exploiter les résultats pour la création d'un treebank pour le français.

Les types de constituants (6 au total) et les relations (14) sont identiques à ceux de EASY cités dans la section précédente. Cette campagne a mis en compétition 10 analyseurs proposés par 9 participants dont la plupart, y compris le laboratoire SYNAPSE, ont participé à la campagne EASY : ERSS, INRIA, LIC2M, LIRMM, LORIA, LPL, SYNAPSE (développeur de CORDIAL), TAGMATICA et XRCE.

Comme ceux de EASY, les résultats de la campagne PASSAGE 1 ont été publiés en anonymat. Néanmoins, grâce à (Laurent *et al.*, 2009), nous avons pu obtenir les résultats de l'analyseur Cordial et identifier les performances de l'outil dans les résultats officiels de l'évaluation, publiés par De La Clergerie *et al.* (2008).

16. http ://atoll.inria.fr/passage/

5.1.1.3 La seconde campagne PASSAGE (2009)

La seconde campagne d'évaluation est celle qui a clôturé le projet PASSAGE. Contrairement à la première campagne pour laquelle les participants étaient appelés à utiliser un corpus de test contenant des documents dont la segmentation en mots et en énoncés était fixée à l'avance, aucune segmentation, ni en mots ni en énoncés, n'a été imposée aux participants. Le calcul des mesures est passé par l'utilisation de l'algorithme de réalignement testé lors de la première campagne sur tout le corpus de test pour comparer les données des participants sur deux corpus :

— Le corpus de référence, dont la segmentation en mots et en phrases ainsi que les annotations syntaxiques ont été faites manuellement ;

— Le corpus ROVER, obtenu en combinant les annotations des deux participants ayant eu les meilleures performances lors de PASSAGE 1. Cordial fait partie des annotateurs dont les résultats ont été exploités à cette fin.

5.1.1.4 Performance de Cordial dans les différentes campagnes

Le choix de l'analyseur syntaxique pour ce travail s'est porté sur le logiciel Cordial parce qu'il fait partie des meilleurs analyseurs syntaxiques disponibles pour le français. Comme l'indiquent ses performances lors des plus récentes campagnes d'évaluation (PASSAGE), il occupe la première ou la deuxième position suivant le critère d'évaluation. Selon les corpus, Cordial obtient les meilleurs scores (Précision : entre 0,84 et 0,92, moyenne=0,89 et f-mesure : entre 0,79 et 0,92) en délimitation des constituants de la phrase, face aux autres concurrents lors de la campagne EASY (Paroubek *et al.*, 2007). Il obtient également la meilleure f-mesure, toutes relations et tous genres de corpus confondus. Le tableau 5.2 résume les performances globales de Cordial lors de la campagne EASY.

TAB. 5.2 – Performances de Cordial dans EASY : mesures en précision (p) et f-mesure (f) par type de corpus pour tous les constituants et toutes les relations (Paroubek *et al.*, 2007).

	lemonde	littéraire	médical	oral_delic	parlement	questions	web
constituants	p=0.904 f=0.904	p=0.910 f=909	p=0.909 f=0.902	p=0.849 f=0.794	p=0.921 f=0.917	p=0.913 f=0.902	p=0.924 f=0.922
relations	p=0.610 f=0.599	p=0.640 f=0.624	p=605 f=0.597	p=0.522 f=0.502	p=0.582 f=0.568	p=0.635 f=0.622	p=0.595 f=0.573

L'analyseur Cordial présente des performances encore meilleures lors de la campagne PASSAGE 1, où il voit son taux d'erreur se réduire considérablement (taux inférieur à 4% lors de PASSAGE contre 10% pour EASY). Les tableaux 5.3 et 5.4 présentent les résultats de Cordial dans PASSAGE 1.

TAB. 5.3 – Performances globales de Cordial dans PASSAGE 1 (Laurent *et al.*, 2009).

constituants			relations		
Précision	Rappel	f-mesure	Précision	Rappel	f-mesure
0,96	0,97	0,96	0,69	0,65	0,67

TAB. 5.4 – Performances de Cordial sur les textes médicaux et courriers électroniques (*mail*) (Laurent *et al.*, 2009).

	constituants			relations		
Corpus	Précision	Rappel	f-mesure	Précision	Rappel	f-mesure
Med.	0,91	0,95	0,93	0,65	0,57	0,60
Mail	0,95	0,96	0,95	0,67	0,67	0,67

Cordial occupe la première place en annotation des relations de dépendance et la deuxième en délimitation des constituants de la phrase, lors de la première campagne PASSAGE (Laurent *et al.*, 2009). En ce qui concerne l'annotation des relations entre les constituants, Cordial fournit les meilleures performances dans l'annotation de certaines relations : *Suj-V* (sujet-verbe), *Aux-V* (auxiliaire-verbe), *Comp* (subordonnées conjonctives principalement), *Mod-A* (modifieurs de l'adjectif) et *Coord* (coordination), et oscille entre la deuxième et la troisième place pour les autres relations. De façon globale, il fait partie des analyseurs syntaxiques en tête de classement en précision (0,96), en rappel (0,97), et en f-mesure (0,96) dans cette campagne.

Malheureusement, comme l'indiquent les données du tableau 5.4, dans PASSAGE 1, Cordial enregistre ses plus faibles performances en annotation des relations sur les textes de mails (f-mesure 0,67) et les textes médicaux (f-mesure 0,60) en particulier, pour des raisons que nous allons préciser ultérieurement. Cependant, ce constat ne concerne pas uniquement l'analyseur Cordial. En effet de façon générale, pour l'ensemble des systèmes d'annotation, les scores les plus faibles sont ceux obtenus sur les corpus de mails, suivis par les corpus littéraires et journalistiques (*Le Monde*), et enfin les corpus de textes médicaux, qui eux aussi enregistrent de moins bons résultats (De La Clergerie *et al.*, 2008).

Lors de PASSAGE2, dans les différentes étapes de l'évaluation, Cordial reste en tête du classement, oscillant entre la première et la deuxième place. Selon les corpus, il présente des performances relativement bonnes pour les relations sujet-verbe et verbe-COD, et de façon globale, il connaît une légère amélioration de ses performances de 0,002 (Paroubek, 2009).

5.1.2 Évaluation des résultats de l'annotation syntaxique des corpus

5.1.2.1 Raison de l'évaluation et justification du choix de la méthode d'évaluation appliquée

Comme nous l'avons souligné dans l'introduction de ce chapitre, dans un travail de recherche, l'évaluation des performances des outils utilisés est une tâche importante. L'évaluation du logiciel Cordial avait pour but de juger jusqu'à quel point cet analyseur pouvait contribuer à atteindre nos objectifs, sachant que notre objectif immédiat, après l'analyse syntaxique des corpus, était l'extraction des patrons syntaxiques des verbes.

Au lieu de focaliser l'évaluation sur la relation verbe-constituant comme dans les campagnes EASY et PASSAGE, nous avons opté pour une méthode qui porte sur le verbe pris dans son contexte, notamment sur sa structure argumentale. Deux raisons justifient le choix de la méthode d'évaluation appliquée.

Puisque ce travail est directement axé sur l'étude de la structure argumentale des verbes, notre principale préoccupation (au sujet de l'analyse syntaxique) porte sur la qualité de l'analyse de l'ensemble des arguments du verbe, en relation avec ce dernier. Ceci requiert bien évidemment une bonne analyse individuelle de chaque constituant de la structure argumentale : sa délimitation et sa dépendance vis-à-vis du verbe pivot. Autrement dit, nous nous intéressons plus à l'analyse de la structure argumentale dans sa globalité qu'à ses constituants pris individuellement, car l'extraction des différents patrons syntaxico-sémantiques des verbes est une tâche fondamentale dans ce travail. Certes, à un moment donné dans la chaîne de travail, nous avons étudié les paires verbe-terme. Cela n'entre toutefois pas en contradiction avec la méthode d'évaluation ici décrite. Bien au contraire, notre démarche a été pensée de façon à ce que l'on parte des patrons verbaux entiers pour ensuite descendre au niveau des paires verbe-terme. Ceci est d'autant plus vrai que le succès de l'annotation de la structure argumentale repose sur une bonne analyse des paires verbe-argument qui la constituent. En d'autres termes, l'analyse de la relation verbe-structure argumentale implique que l'étude de la relation individuelle entre le verbe et chacun de ses arguments sujet, COD, COI, etc., ait été effectuée au préalable, avec succès. En résumé, notre méthode d'évaluation se base sur le modèle de PASSAGE, qui est plus détaillé et propice lorsqu'on veut évaluer les performances des outils en profondeur, comme cela a été fait lors des différentes campagnes d'évaluation susmentionnées.

5.1.2.2 Présentation de la méthode d'évaluation appliquée

Le processus d'évaluation consiste à caractériser les propositions des phrases, en associant à la structure argumentale de chaque verbe pivot une des étiquettes suivantes :

— *ok* : bon (la structure argumentale a été bien annotée conformément au critère concerné) ;

— *par* : partiellement (une partie de la structure argumentale a été bien annotée conformément au critère concerné) ;

— *bad* : mauvaise annotation (l'annotation de l'ensemble de la structure argumentale est erronée vis-à-vis du critère concerné) ;

— *no* : pas d'information fournie par rapport au critère concerné.

Cette caractérisation de la structure argumentale des verbes pivots se fait relativement à 3 critères sur lesquels porte notre évaluation :

— la frontière des unités (*délimitation*) : est-ce que les syntagmes sont bien délimités ?

— le rôle des constituants (*fonctions syntaxiques*) : est-ce que les fonctions syntaxiques sont bien attribuées ?

— la relation verbe-structure argumentale (*pivot*) : est-ce que les relations de dépendance entre le verbe et ses arguments sont bien exprimées ? Chaque argument a t-il bien été rattaché au verbe (pivot) dont il dépend ?

Certaines phrases comportent plusieurs propositions et donc plusieurs verbes pivots. Dans de tels cas, la structure argumentale de chaque proposition est évaluée individuellement par rapport aux critères. Un élément très important à souligner est que dans le cadre de cette évaluation, nous nous sommes limités à l'analyse de la structure argumentale (le verbe, son sujet et ses compléments) de façon stricte. En d'autres termes, nous n'avons pas tenu compte de l'annotation des circonstants et autres éléments n'appartenant pas à la structure argumentale. Par exemple, si dans une phrase le verbe a deux arguments bien annotés par Cordial et un circonstant dont l'annotation est erronée, nous ne tenons pas compte de ce circonstant. Nous considérons uniquement l'analyse des constituants de la structure argumentale, qui dans le cas d'espèce serait évaluée positivement (*ok*). La figure 5.1 présente un exemple d'évaluation. Les trois dernières colonnes contiennent le résultat de l'évaluation en guise de délimitation, fonction syntaxique et association avec le verbe pivot. L'étiquette *ok* indique que l'analyse de la structure argumentale a été jugée correcte. Le sujet et son attribut ont bien été délimités et identifiés, respectivement par les lettres T et B. Ils ont été associés au verbe pivot *être*.

Un échantillon[17] de 200 phrases (50 par corpus) choisies aléatoirement a été utilisé pour effectuer l'évaluation.

Nos critères d'évaluation sont similaires à ceux utilisés dans les campagnes d'évaluation EASY et PASSAGE, à la seule différence que dans PASSAGE, les deux derniers critères ne forment qu'un seul nommé *relation de dépendance*. Dans le cadre de notre évaluation, nous avons tenu à faire une distinction entre la relation de dépendance verbe-argument et les autres, parce

17. Ces phrases seront délivrées sur une version CD-ROM.

N°	Mot	Lemr	Typegran	Codegre	Synta	Fc	N	Piv	Type P	délimitation	fonction sy	pivot
	La déplétion sodée est un facteur de risque d IRA postopératoire									ok	ok	ok
1	La	le	DETDFS	Da-fs-d	2	T	1	être	Indép.			
2	déplétion	déplé	NCFS	Ncfs	2	T	1	être	Indép.			
3	sodée	sodé	ADJFS	Afpfs	2	T	1	être	Indép.			
4	est	être	VINDP3S	Vmip3s	4	V	1	être	Indép.			
5	un	un	DETIMS	Da-ms-i	6	B	1	être	Indép.			
6	facteur	facteu	NCMS	Ncms	6	B	1	être	Indép.			
7	de	de	PREP	Sp	8|6	B	1	être	Indép.			
8	risque	risque	NCMS	Ncms	8|6	B	1	être	Indép.			
9	d	de	PREP	Sp	10|8|6	B	1	être	Indép.			
10	IRA	ira	NCI	Nc..	10|8|6	B	1	être	Indép.			
11	postopératoi	posto	ADJSIG	Afp.s	11|6	B	1	être	Indép.			

FIG. 5.1 – Exemple d'évaluation d'une phrase.

que ce type de dépendance joue un rôle déterminant dans notre travail de thèse. En effet, la dépendance verbe-complément nous permet d'évaluer les performances de Cordial en ce qui concerne l'identification du verbe pivot, c'est-à-dire le verbe principal dans une proposition. Dans les résultats que retourne Cordial après l'analyse syntaxique d'un texte, cette information apparaît dans la 11e colonne (*Pivot*) comme le montre la figure 5.2. Cette figure présente un extrait de l'analyse syntaxique que propose Cordial pour la phrase qui apparaît sur l'image.

#N°	Mot	Lemme	Typegram	Codegram	Syntagme	Fonction	Num Prop.	Pivot	Prop.
	La déplétion sodée est un facteur de risque d IRA postopératoire et peut passer inaperçue .								
1	La	le	DETDFS	Da-fs-d	2	T	1	être	Indép.
2	déplétion	déplétio	NCFS	Ncfs	2	T	1	être	Indép.
3	sodée	sodé	ADJFS	Afpfs	2	T	1	être	Indép.
4	est	être	VINDP3S	Vmip3s	4	V	1	être	Indép.
5	un	un	DETIMS	Da-ms-i	6	B	1	être	Indép.
6	facteur	facteur	NCMS	Ncms	6	B	1	être	Indép.
7	de	de	PREP	Sp	8|6	B	1	être	Indép.
8	risque	risque	NCMS	Ncms	8|6	B	1	être	Indép.

FIG. 5.2 – Repérage du verbe pivot par Cordial.

Notre démarche d'évaluation se démarque également de celles des campagnes EASY et PASSAGE en ce qu'elle porte sur la structure argumentale uniquement (c'est-à-dire strictement sur les arguments des verbes : sujet, COD, COI et complément d'agent), prise dans son ensemble, contrairement à la méthode d'évaluation appliquée lors des campagnes mentionnées précédemment, qui porte sur les paires verbe-constituant prises individuellement. Malgré ces points de dissemblance au niveau de la méthode utilisée, nous verrons dans les prochaines sections que les résultats obtenus à l'issue de cette évaluation sont cohérents avec ceux des campagnes PASSAGE.

5.1.2.3 Résultats de l'évaluation et comparaison avec les résultats de PASSAGE

Les tableaux 5.5, 5.6, 5.7 et 5.8 contiennent le récapitulatif des résultats de l'évaluation que nous avons réalisée. Pour chacune des 50 phrases analysées par corpus (nous avons 4 corpus au total), 3 paramètres ont été évalués : la délimitation des arguments (exclusivement), l'attribution des fonctions syntaxiques à ces arguments et la relation de dépendance avec le verbe pivot de la proposition.

— chaque structure argumentale bien analysée par Cordial, c'est-à-dire portant l'étiquette *ok*, enregistre un score égal à 1.

— chaque structure argumentale dont la qualité de l'analyse est partielle, c'est-à-dire portant l'étiquette *par*, enregistre un score égal à 0,5.

— les deux autres étiquettes correspondent au score 0.

La précision pour chaque corpus a été calculée comme suit :

$$P = \frac{somme\ des\ scores\ par\ critère}{nombre\ total\ de\ structures\ argumentales\ évaluées}$$

TAB. 5.5 – Résultats de l'évaluation de l'annotation syntaxique (PRO).

	délimitation	fonction	relation
ok	52	36	55
par	22	32	5
no	1	1	3
bad	0	6	12
P	0,84	0,69	0,76

TAB. 5.6 – Résultats de l'évaluation de l'annotation syntaxique (ETU).

	délimitation	fonction	relation
ok	48	30	44
par	15	24	4
no	0	5	6
bad	1	5	10
P	0,86	0,65	0,71

TAB. 5.7 – Résultats de l'évaluation de l'annotation syntaxique (VUL).

	délimitation	fonction	relation
ok	52	32	59
par	19	38	2
no	0	0	0
bad	0	1	10
P	0,86	0,71	0,84

TAB. 5.8 – Résultats de l'évaluation de l'annotation syntaxique (FOR).

	délimitation	fonction	relation
ok	95	77	113
par	33	46	2
no	0	1	1
bad	1	5	13
P	0,86	0,77	0,88

Comme cela a été souligné précédemment, plusieurs phrases évaluées sont des phrases complexes, c'est-à-dire ayant plus d'une proposition et donc plus d'un verbe pivot. C'est la raison

pour laquelle dans les tableaux 5.5, 5.6, 5.7 et 5.8, le nombre total d'éléments (structures argumentales) évalués dépasse 50 (nombre de phrases évaluées), et ce nombre varie d'un corpus à l'autre.

Les résultats ci-dessus seront comparés aux résultats de Cordial dans la campagne PASSAGE 1 car celle-ci porte entre autre sur les textes médicaux et les textes des mails, qui correspondent aux types de textes que contiennent notre corpus FOR. De façon générale, nos résultats ne s'éloignent pas tellement de ceux que Cordial a obtenus lors de PASSAGE 1. Dans les sections suivantes, nous donnerons plus de précisions sur les résultats de notre évaluation en les contrastant à ceux de Cordial dans PASSAGE 1.

Délimitation des constituants

La lecture des tableaux 5.5, 5.6, 5.7 et 5.8 permet d'observer que de façon générale, en délimitation des constituants, Cordial fournit des résultats satisfaisants sur nos corpus. Pour chaque corpus, plus de la moitié des phrases évaluées obtiennent le score 1 et la plus faible précision est égale à 0,84. Les corpus enregistrent chacun une précision qui est en dessous de la valeur obtenue lors de la campagne PASSAGE 1 (0,96). Néanmoins, la précision cumulée des 4 corpus produit un score de 0,85, ce qui correspond à une valeur tout de même élevée.

Fonction syntaxique des constituants

En ce qui concerne l'annotation des *relations* (que nous appelons *fonctions syntaxiques*), les résultats de notre évaluation sont moins bons, ce qui n'est pas surprenant, puisque les résultats de PASSAGE 1 (cf. tableau 5.3) sont caractérisés par un constat similaire. Néanmoins, notre précision cumulée (0,70) est légèrement supérieure à celle obtenue dans PASSAGE 1 autant sur les textes médicaux que sur les textes de mails (cf. tableau 5.4). D'après les tableaux 5.5, 5.6, 5.7 et 5.8, pour les corpus PRO, ETU et VUL, un peu plus de la moitié des structures argumentales portent soit l'étiquette *bad*, soit l'étiquette *par* qui signale que l'analyse syntaxique de la phrase est partiellement problématique. D'après nos observations, cette faible performance de Cordial en attribution de rôles syntaxiques est en grande partie due à l'échec de l'analyse des syntagmes prépositionnels qui provoque du bruit dans les résultats. La section 5.1.2.4 fournit de plus amples informations à propos des problèmes liés à l'attribution des fonctions syntaxiques.

Dépendance de la structure argumentale vis-à-vis du verbe

D'après les résultats de notre évaluation, en termes d'annotation des relations de dépendance verbale, c'est-à-dire le repérage du ou des verbes pivots en rapport avec leurs arguments, le logiciel Cordial enregistre ses meilleurs scores de précision (entre 0,71 et 0,88). D'après les tableaux 5.5, 5.6, 5.7 et 5.8, les performances de Cordial sont bien au-dessus de la moyenne en

matière de repérage du verbe pivot ; cette remarque est valable pour les extraits des quatre corpus.

5.1.2.4 Quelques observations faites à l'issue de l'évaluation

Certaines erreurs observées suite à l'évaluation des résultats de l'analyse syntaxique font partie des principales faiblesses de l'analyseur syntaxique Cordial, bien connues par les concepteurs de l'outil et signalées également lors des campagnes d'évaluation.

Mauvais raccordement des syntagmes prépositionnels et génération de faux COI

Certains syntagmes prépositionnels en *à, de, par, avec,* etc. se voient attribuer la fonction de COI alors qu'ils ne jouent pas véritablement, voire jamais (*avec*), ce rôle dans la phrase. Les cas les plus problématiques sont ceux où des compléments circonstanciels, ou encore des locutions adverbiales, sont annotés comme COI. La figure 5.3 fournit un exemple. Cette figure présente le résultat de l'analyse syntaxique de la phrase 1 par Cordial. Dans cette figure, les informations concernant le syntagme dont l'analyse syntaxique est problématique sont identifiées en gras.

1) *Il faut favoriser au contraire les sports avec activité progressive ou stable.*

#N	Offset_	Offset_	Mot	Lemme	Typegram	Codegr	Syn	Fon	Nu	Pivot	Type Pro
Il faut favoriser au contraire les sports avec activité progressive ou stable .											
1	648938	648940	Il	il	PPER3S	Pp3msn	1	S	1	falloir	Indép.
2	648941	648945	faut	falloir	VINDP3S	Vmip3s	2	V	1	falloir	Indép.
3	648946	648955	favoriser	favoriser	VINF	Vmn--	3	V	2	favoriser	Infinitive
4	**648956**	**648958**	**au**	**à le**	**DETDMS**	**Da-ms-**	**5**	**F**	**2**	**favorise**	**Infinitive**
5	**648959**	**648968**	**contraire**	**contraire**	**NCMS**	**Ncms**	**5**	**F**	**2**	**favorise**	**Infinitive**
6	648969	648972	les	le	DETDPIG	Da-.p-d	7	D	2	favoriser	Infinitive
7	648973	648979	sports	sport	NCMP	Ncmp	7	D	2	favoriser	Infinitive
8	648980	648984	avec	avec	PREP	Sp	9	H	2	favoriser	Infinitive
9	648985	648993	activité	activité	NCFS	Ncfs	9	H	2	favoriser	Infinitive
10	648994	649005	progressive	progressif	ADJFS	Afpfs	9	H	2	favoriser	Infinitive
11	649006	649008	ou	ou	COO	Cc	9	H	2	favoriser	Infinitive
12	649009	649015	stable	stable	ADJSIG	Afpfs	9	H	2	favoriser	Infinitive
13	649015	649016	.	.	PCTFORTE	Yps	-	-	-	-	-

FIG. 5.3 – Exemple de faux COI.

La figure 5.3 présente un tableau Excel contenant le résultat de l'analyse syntaxique qu'a effectuée Cordial pour la phrase de l'exemple ci-dessus. Chaque ligne du tableau correspond à 12 champs d'informations qui sont : identifiant du mot dans la phrase, *offset_begin* ou identifiant du début de la chaîne de caractère, *offset_end* ou identifiant de fin de chaîne, forme du mot, lemme, catégorie grammaticale, propriété morpho-syntaxique, syntagme, fonction grammaticale, numéro identifiant de proposition, verbe pivot, type de proposition, et sens du mot. L'annexe A.1 décrit le jeu d'étiquettes (lettres) utilisées par Cordial pour l'annotation des catégories

grammaticales, ainsi que leurs significations. Dans cet exemple, la lettre F mise en gras dans la neuvième colonne identifie le COI, S identifie le sujet ; V, le verbe ; D, le COD et H, le complément circonstanciel.

La phrase exemple de la figure 5.3 illustre un cas d'attribution d'une fonction syntaxique erronée à un syntagme, qui débouche sur ce que nous avons nommé un *faux* COI. Dans cette phrase, *au contraire* est un syntagme prépositionnel qui fonctionne comme locution adverbiale, pourtant Cordial l'a annoté comme COI. Ce phénomène de faux COI est fréquemment rencontré dans les résultats de Cordial. Afin d'avoir une idée de l'impact qu'il a sur nos résultats, nous avons quantifié le nombre de phrases évaluées qui en sont affectées parmi les 50 phrases extraites de chaque corpus. Le tableau 5.9 contient les résultats de cette expérience.

TAB. 5.9 – Nombre de phrases contenant des faux COI (N = 50).

PRO	ETU	VUL	FOR
13	6	11	12

D'après les résultats du tableau 5.9, les phrases touchées par le problème de faux COI représentent plus de 10% de chacun des extraits évalués. D'après nos investigations, cette erreur d'analyse provient de différentes causes qui relèvent soit des choix théoriques, soit des méthodes appliquées lors de la conception du logiciel Cordial.

Selon la démarche d'analyse appliquée par Cordial, la détermination des relations syntaxico-sémantiques commence par le traitement des relations *sujet-verbe* et *verbe-attribut*, qui d'après les concepteurs de Cordial, sont les plus importantes sources de fautes de grammaire. Ce n'est qu'après cette étape que l'analyseur Cordial effectue le traitement des relations COD, puis COI, dont l'analyse exige au préalable « une détermination de la nature réelle des déterminants du type « de la » et « des » qui conditionne la catégorisation en objet direct ou indirect » des syntagmes nominaux (Laurent *et al.*, 2009). Cela signifie que Cordial considère les syntagmes nominaux introduits par la préposition « de » comme des candidats potentiels à la fonction de COI et que les compléments du verbe autres que le *cod* et le sujet sont tous traités ensemble (le complément d'agent y compris). Ce choix théorique s'avère problématique, car les syntagmes introduits par « de » sont très souvent ambigus entre l'expression de l'appartenance et l'introduction d'un COI.

Les conséquences principales de ce choix sont, premièrement, le risque d'attribuer le rôle d'argument (COI, agent) à un circonstanciel. Il y a également le risque de caractériser comme COI des syntagmes prépositionnels qui ne jouent pas ce rôle syntaxique (un complément de nom par exemple), provoquant ainsi le phénomène de faux COI observé à travers les résultats de l'évaluation que nous avons effectuée. D'après Laurent *et al.* (2009), lors de l'évaluation PASSAGE 1, Cordial a fait de bonnes performances sur les relations liées aux accords, notamment *verbe-sujet*, *verbe-attribut*, *auxiliaire-verbe*, *verbe-cod*, etc. Les relations *verbe-coi* et *verbe-compl. d'agent* font partie de ce que les évaluateurs appellent relation *compléments-verbe*, qui

est décrite comme suit : « cette relation concerne les autres compléments du verbe exprimés sous forme de GP ou de PV, que ce soit les circonstants ou les compléments indirects (introduits par une préposition) » (cf. guide ou protocole d'annotation [18] pour PASSAGE 1). Dans PASSAGE 1, le résultat cumulé de l'évaluation de ces relations *compléments-verbe* donne une faible précision (0,57) et un rappel de 0,63, ce qui produit une f-mesure de 0,60.

Le phénomène de génération des faux COI est également lié à un problème plus important, celui de l'échec de la mise en relation (raccordement) de certains syntagmes nominaux avec le verbe ou le nom dont ils dépendent. Le plus souvent, les arguments des infinitives et des verbes introduits par des locutions impersonnelles tombent dans cette catégorie. Comme le montre l'exemple de la figure 5.4, qui présente une partie de l'analyse de la phrase ci-dessous :

2) *Ces deux drogues inhibent spécifiquement l'agrégation plaquettaire induite par l'ADP, en s'opposant à la liaison de ce dernier à son récepteur.*

14	118412	118414	en	en	PREP	Sp	15	H	1	inhiber	Indépendante
15	118415	118416	s	s	NCMIN	Ncm.	15	H	1	inhiber	Indépendante
16	118417	118425	opposant	opposer	VPARPRES	Vmpp--	16	-	1	inhiber	Indépendante
17	118426	118427	à	à	PREP	Sp	19	F	1	inhiber	Indépendante
18	118428	118430	la	le	DETDFS	Da-fs-d	19	F	1	inhiber	Indépendante
19	118431	118438	liaison	liaison	NCFS	Ncfs	19	F	1	inhiber	Indépendante
20	118439	118441	de	de	PREP	Sp	22|19	F	1	inhiber	Indépendante
21	118442	118444	ce	ce	DETDEM	Dd-ms-	22|19	F	1	inhiber	Indépendante
22	118445	118452	dernier	dernier	NCMS	Ncms	22|19	F	1	inhiber	Indépendante
23	118453	118454	à	à	PREP	Sp	25|19	F	1	inhiber	Indépendante
24	118455	118458	son	son	DETPOSS	Ds3.ss	25|19	F	1	inhiber	Indépendante
25	118459	118468	récepteur	récepteur	NCMS	Ncms	25|19	F	1	inhiber	Indépendante
26	118468	118469	.	.	PCTFORT	Yps	-	-	-	-	-

FIG. 5.4 – Problème de raccordement des arguments.

Comme le montre la figure 5.4, le COI *la liaison de ce dernier à son récepteur* a été rattaché au verbe *inhiber* que Cordial a identifié comme unique verbe pivot (principal) de la phrase. Or en réalité, ce COI se rapporte à *s'opposer*, pivot de la proposition participiale que constitue la deuxième partie de la phrase. Il faut reconnaître que cet exemple est particulier, premièrement du fait de la présence d'une proposition participiale, deuxièmement parce que le COI de *s'opposer* est lui-même complexe. *Liaison* est un nom déverbal, qui a plus ou moins la même structure argumentale que le verbe LIER dont il est issu (lier quelque chose à quelque chose). L'erreur d'analyse se situe donc à deux niveaux : l'échec d'identification du deuxième verbe pivot et le mauvais raccordement du COI qui, dans le cas présent, est la conséquence de la première erreur.

18. https ://perso.limsi.fr/anne/Guide/PEAS_reference_annotations_v2.2.htmlrelation_cpl_verb

Cette remarque fait partie des difficultés majeures que rencontre le logiciel Cordial et que Laurent *et al.* (2009) signalent en déclarant qu'« une analyse par type de constituants des erreurs commises par Cordial montre que les principales erreurs sont liées à la confusion entre adjectifs et participes ainsi qu'à de mauvais raccordements de groupes nominaux soit vers d'autres groupes nominaux, soit vers d'autres verbes [...] ».

En ce qui concerne les méthodes statistiques, Cordial est un analyseur basé sur une approche essentiellement probabiliste, ce qui signifie que les règles les plus importantes sont très souvent induites automatiquement (Laurent *et al.*, 2009). Le risque de production des erreurs de sur-génération, comme celle qui touche les COI, est en quelque sorte inévitable.

Traitement des phrases longues et complexes

Les résultats de notre évaluation indiquent que Cordial enregistre ses plus faibles performances en attribution des fonctions syntaxiques dans le corpus des experts (cf. tableau 5.5). D'après nos observations, la longueur et la complexité des phrases du corpus des experts pourraient faire partie des causes de ce faible résultat. Les données du tableau 2.1 du chapitre 2 indiquent que les phrases du corpus des experts ont en moyenne 28,39 mots. Sur 50 phrases analysées dans le cadre de notre évaluation, 16 ont au-delà de 30 mots, ce qui est au-dessus de la moyenne. Grâce aux résultats de l'évaluation, nous avons observé que 12/16 phrases longues (c'est-à-dire 75%) ont un problème de délimitation et/ou d'attribution de fonctions syntaxiques. À notre avis, ceci n'est pas une simple coïncidence. La longueur et la complexité des phrases semblent provoquer des complications dans le processus d'analyse syntaxique réalisé par Cordial. Une consultation des résultats de l'évaluation PASSAGE 1 montre que Cordial a une précision et un rappel faibles (0,67) dans le corpus médical, comparé à ses performances dans les autres types de textes. En guise d'explication, Laurent *et al.* (2009) soulignent que le langage médical (comme tout autre langage de spécialité) contient des tournures spécifiques, ce qui rendrait la tâche difficile à l'analyseur. Les résultats de Cordial sur les textes littéraires ont permis de confirmer que la longueur et la complexité des phrases sont à l'origine des différents types d'erreurs de délimitation et d'analyse des relations, qui ont causé un mauvais score.

Malgré ce qui a été dit précédemment, les résultats de notre évaluation indiquent que Cordial a produit ses meilleurs performances en délimitation des constituants avec une précision cumulée de 0,85. Et malgré ses longues phrases, le corpus PRO obtient une précision élevée (0,84). En effet, parmi les phrases évaluées dans le PRO, nous avons détecté de longues phrases pour lesquelles les résultats de l'analyse syntaxique (plus précisément la délimitation des constituants) sont de bonne qualité, bien au dessus de ce à quoi l'on pouvait s'attendre. La phrase ci-dessous en est un exemple :

3) *Alors que certaines études comme la Rotterdam Study montrent une augmentation de la prévalence des cardiopathies ischémiques chez des patients en hypothyroïdie infraclinique*

[2], d'autres essais ne retrouvent pas une telle association, et une étude a même trouvé que l'élévation de la TSH était associée à une diminution de la mortalité chez les patients âgés de plus de 85 ans [3].

Malgré la longueur (61 mots) de cette phrase, les arguments et circonstants des différents verbes pivots ont été convenablement délimités par Cordial, ce qui nous pousse à considérer que la difficulté rencontrée par le système de Cordial avec les phrases longues ne l'empêche pas d'être efficace, même dans les cas où l'on s'y attend le moins.

Traitement des phrases du corpus des forums

Les performances de Cordial sur les phrases du corpus des forums qui sont, pour la plupart, caractérisées soit par la mauvaise ponctuation, soit par les fautes d'orthographe et/ou de grammaire, s'avèrent plutôt positives, contrairement à ce que l'on pouvait craindre. De façon totalement inattendue, au terme de l'évaluation, plus de la moitié des phrases portent l'étiquette « ok » en délimitation des constituants et en relation de dépendance (respectivement 73,64% et 87,59%). L'attribution des rôles syntaxiques est la tâche qui présente le score le plus bas (0,77 de précision). Néanmoins, ce score est légèrement au dessus de la précision (0,67) obtenue dans le cadre de PASSAGE 1 sur les textes de mails. À notre grande surprise, le corpus des forums enregistre les meilleurs scores de notre évaluation, ce constat s'appliquant aux trois critères d'évaluation (cf. tableau 5.8).

D'après les résultats de PASSAGE 1, la mauvaise performance de Cordial sur les phrases des mails en délimitation des constituants a entraîné un fort taux d'erreurs lors de l'analyse des relations de dépendance dans ce corpus (Laurent *et al.*, 2009). Chez nous, par contre, Cordial a obtenu un score élevé en délimitation des constituants, ce qui a certainement favorisé l'obtention d'une meilleure précision en attribution des fonctions syntaxiques.

Traitement des formes verbales complexes

Une autre remarque qui a été faite et que nous jugeons nécessaire de souligner est la méthode d'analyse de certains types de syntagmes verbaux. Un cas de figure est celui des formes verbales composées commençant par un verbe modal (*elle doit être opérée*). Dans le système d'analyse de Cordial, les verbes modaux (*pouvoir, devoir*, etc.) sont considérés comme des verbes principaux, c'est-à-dire des verbes qui fonctionnent de façon autonome, ce qui est grammaticalement acceptable. Lorsque des syntagmes verbaux ayant en tête un verbe modal contiennent un passif, ce passif est décomposé ; le participe passé du verbe au passif devient un verbe principal (c'est-à-dire pivot), tandis que l'auxiliaire *être* est considéré comme complément du verbe modal, ce qui est erroné. L'analyse du syntagme verbal *peut être remise* dans la phrase suivante peut servir d'exemple :

4) *Lors de la consultation d'anesthésie, une note d'information écrite peut être remise au patient afin de renforcer l'information orale et d'en assurer la cohérence.*

La figure 5.5 présente le résultat de l'analyse de cette phrase par Cordial.

#N	Offset_	Offset_e	Mot	Lemme	Typegram	Codegram	Synta	For	Vof	Pivot	Type Pr
Lors de la consultation d anesthésie , une note d information écrite peut être remise au patient afin de renforce											
1	396405	396412	Lors de	lors de	PREP	Sp	3	H	1	pouvoir	Indép.
2	396413	396415	la	le	DETDFS	Da-fs-d	3	H	1	pouvoir	Indép.
3	396416	396428	consultation	consultation	NCFS	Ncfs	3	H	1	pouvoir	Indép.
4	396429	396430	d	de	PREP	Sp	5|3	H	1	pouvoir	Indép.
5	396431	396441	anesthésie	anesthésie	NCFS	Ncfs	5|3	H	1	pouvoir	Indép.
6	396441	396442	,	,	PCTFAIB	Ypw	-	-	1	pouvoir	Indép.
7	396443	396446	une	un	DETIFS	Da-fs-i	8	T	1	pouvoir	Indép.
8	396447	396451	note	note	NCFS	Ncfs	8	T	1	pouvoir	Indép.
9	396452	396453	d	de	PREP	Sp	10|8	T	1	pouvoir	Indép.
10	396454	396465	information	information	NCFS	Ncfs	10|8	T	1	pouvoir	Indép.
11	396466	396472	écrite	écrit	ADJFS	Afpfs	8	T	1	pouvoir	Indép.
12	396473	396477	peut	pouvoir	VINDP3S	Vmip3s	12	V	1	pouvoir	Indép.
13	396478	396482	être	être	VINF	Van--	13	D	2	remettre	Infinitive
14	396483	396489	remise	remettre	VPARPFS	Vmpasf	13	V	2	remettre	Infinitive
15	396490	396492	au	à le	DETDMS	Da-ms-d	16	F	2	remettre	Infinitive
16	396493	396500	patient	patient	NCMS	Ncms	16	F	2	remettre	Infinitive
17	396501	396508	afin de	afin de	PREP	Sp	18	H	2	remettre	Infinitive
18	396509	396518	renforcer	renforcer	VINF	Vmn--	18	H	2	remettre	Infinitive
19	396519	396520	l	le	DETDFS	Da-ms-d	20	D	2	remettre	Infinitive
20	396521	396532	information	information	NCFS	Ncfs	20	D	2	remettre	Infinitive
21	396533	396538	orale	oral	ADJFS	Afpfs	20	D	2	remettre	Infinitive
22	396539	396541	et	et	COO	Cc	-	-	2	remettre	Infinitive
23	396542	396543	d	de	PREP	Sp	25	-	2	remettre	Infinitive
24	396544	396546	en	en	PPER3S	Pp3..-	25	I	2	remettre	Infinitive
25	396547	396554	assurer	assurer	VINF	Vmn--	25	I	2	remettre	Infinitive
26	396555	396557	la	le	DETDFS	Da-fs-d	27	D	2	remettre	Infinitive
27	396558	396567	cohérence	cohérence	NCFS	Ncfs	27	D	2	remettre	Infinitive
28	396567	396568	.	.	PCTFORTE	Yps	-	-	-	-	-

FIG. 5.5 – Analyse syntaxique d'une phrase avec forme verbale complexe.

D'après les résultats de l'analyse syntaxique par Cordial, la phrase aurait deux verbes pivots donc deux propositions. Le verbe *pouvoir* est le premier pivot. Il a pour sujet le groupe nominal *une note d'information écrite* et pour COD *être*, tandis que le verbe *remettre* est pivot de la seconde proposition et a pour complément d'objet indirect le groupe nominal *au patient*. Cette analyse est partiellement erronée : si *pouvoir* est considéré comme un semi-auxiliaire et de ce fait comme un verbe autonome, alors, s'il a un COD, ce serait la structure infinitive *être remis au patient*. *Une note d'information* est le sujet syntaxique de *pouvoir* et l'objet (sémantique) de *remettre* (cf. à la voix active : X remet / peut remettre une note d'information au patient).

Dans ce cas de figure, le modèle d'analyse de Cordial, qui relève en partie d'un choix théorique délibéré des concepteurs de l'outil, n'est pas entièrement compatible avec l'objectif que nous visons après l'annotation syntaxique des corpus. En effet, ce type d'analyse n'est pas favorable à l'extraction des structures argumentales des verbes et à l'identification des relations de dépendance (verbe-structure argumentale). Or, ces tâches font partie des tâches fondamentales

de ce travail de thèse. Pour atteindre nos objectifs, il serait plus bénéfique qu'au terme de l'analyse syntaxique d'une phrase comme la phrase 4, le verbe *remettre* soit identifié comme le seul pivot et que les groupes nominaux *une note d'information écrite* et *au patient* soient reconnus comme ses arguments.

Consciente de ce fait dès le départ et sachant que ce type d'analyse aurait un impact considérable sur les résultats de l'analyse syntaxique des corpus, après la réalisation de l'annotation syntaxique des corpus avec Cordial, nous avons entrepris une phase de pré-traitement automatique des résultats avant de procéder à l'annotation sémantique. Ce pré-traitement a été réalisé par un programme Perl dont l'un des buts principaux est la restitution du verbe pivot des formes verbales composées (cf. chapitre 3, section 3.2.1.2).

5.2 Bilan

Les résultats de l'évaluation de l'analyse syntaxique nous ont permis d'apprécier le travail de Cordial sur la base de critères bien précis. Les résultats de cette évaluation ont été analysés en comparaison avec les résultats obtenus par le logiciel Cordial lors de la campagne PASSAGE 1, qui évaluait les performances de plusieurs analyseurs syntaxiques du français, dans le cadre d'une compétition organisée à cet effet. Les trois critères pris en considération sont :

— la délimitation des syntagmes jouant le rôle d'arguments ;

— l'attribution des fonctions syntaxiques aux arguments ;

— la relation verbe-structure argumentale, c'est-à-dire l'établissement de la relation entre un verbe pivot et ses arguments.

D'après nos observations, de façon générale, Cordial produit des résultats similaires à ceux qu'il a obtenus dans PASSAGE 1. Comme dans PASSAGE 1, la meilleure précision est obtenue en délimitation des constituants. De même, les performances de Cordial sont moins bonnes lorsqu'il analyse les phrases longues. Toutefois, selon les résultats de notre évaluation, ses performances sur le corpus des forums se démarquent positivement de celles obtenues sur les textes des mails dans PASSAGE 1. Néanmoins, un certain nombre d'erreurs non négligeables est enregistré. Les deux plus marquantes étant le problème de faux COI et le mauvais raccordement de certains constituants vers le verbe ou le nom dont ils dépendent. Par ailleurs, cette évaluation nous a permis de prendre connaissance des choix théoriques qui font la différence entre l'approche appliquée par Cordial et la méthode d'extraction de données que nous appliquons. La connaissance de ces points de divergence nous a permis de prendre les mesures nécessaires pour s'assurer que nos objectifs puissent malgré tout être atteints.

5.3 Annotation sémantique avec la Snomed

5.3.1 But de l'évaluation

L'évaluation de l'annotation sémantique des corpus consiste à analyser un certain nombre de phrases annotées par notre système d'annotation, en se focalisant sur les catégories sémantiques associées aux groupes nominaux qui représentent les arguments des verbes. Cette évaluation questionne également les cas de non-attribution de catégories sémantiques aux noms, son but étant d'apprécier la qualité de l'annotation sémantique et d'exposer les points forts et les faiblesses de notre système.

5.3.2 Description de la démarche et des données évaluées

La démarche d'évaluation s'applique aux données acquises suite à l'appariement entre les noms-arguments des verbes et les catégories sémantiques de la terminologie Snomed (cf. chapitre 3, section 3.2.4). Ces données (verbe+argument+catégorie-sémantique) représentent donc les structures argumentales des verbes, chaque nom-argument étant associé à une catégorie sémantique provenant de la terminologie Snomed, à l'exception des noms qui, pour une raison ou une autre, sont retournés sans catégorie. Au total, 200 structures argumentales (50 par corpus) ont été évaluées comme suit : pour chaque argument de la chaîne, si le syntagme nominal correspondant porte une catégorie sémantique, nous vérifions s'il s'agit bien de la catégorie adéquate, ce qui correspond à l'étiquette *ok*. Si la catégorie sémantique n'est pas correcte, alors l'étiquette *bad* est associée au terme. Si le syntagme nominal en position d'argument ne porte aucune catégorie sémantique et que cela est correct (par exemple pour un nom non médical), l'étiquette *ok* est utilisée. Dans le cas contraire, l'étiquette *no* est appliquée. Dans le cadre de cette évaluation, nous ne nous préoccupons pas de l'annotation sémantique des circonstants et autres éléments de la phrase.

5.3.3 Résultats

Le tableau 5.10 contient les résultats de l'évaluation. La cinquième ligne (*%no+bad*) fournit le pourcentage d'erreurs d'annotation sur les 50 structures argumentales évaluées dans chaque corpus. Quant à la dernière ligne, elle donne les différents scores de précision obtenus. Cette précision correspond au ratio du nombre de termes bien annotés (total de *ok*) sur le nombre total de termes à annoter.

À la lecture du tableau 5.10, la première observation est que dans chacun des extraits évalués, plus de 50% de noms ont été correctement associés à la catégorie sémantique correspondante. Les résultats de l'évaluation semblent également indiquer que plus on tend vers les textes des

TAB. 5.10 – Résultats de l'évaluation de l'annotation sémantique des corpus.

	PRO	ETU	VUL	FOR
ok	55	49	45	35
no	3	6	8	8
bad	15	10	11	19
tot.	73	65	64	62
%no+bad	24,65	24,61	29,68	43,54
Précision	0,75	0,75	0,69	0,56

non-experts (VUL et FOR), plus la qualité de l'annotation se détériore. En effet, le corpus des étudiants et celui des experts occupent la tête du classement, avec 0,75 de précision tous les deux, tandis que le corpus des forums arrive en dernière position avec seulement 0,56 de précision. Ce constat n'est pas surprenant, d'autant plus que les résultats de l'annotation sémantique indiquaient clairement que les textes des experts (PRO et ETU) contenaient plus de termes médicaux que ceux adressés au grand public (cf. chapitre 4, tableau 4.8). De plus, le type de textes (textes en général caractérisés par des fautes d'orthographe) que contient le corpus des forums est favorable à la production des erreurs d'annotation, ce qui pourrait aussi être une explication à la faible précision obtenue pour le corpus FOR.

Une analyse plus approfondie des résultats du tableau 5.10, à travers le questionnement des types d'erreurs détectées dans les corpus, permet de voir que notre système d'annotation tend à générer plus de bruit que de silence. D'après le tableau 5.10, dans les 4 extraits de phrases évalués, le nombre minimal de groupes nominaux sans catégories sémantiques est de 3, tandis que le nombre minimal de groupes nominaux ayant une catégorie sémantique erronée est de 10. Le corpus FOR compte le plus grand nombre de groupes nominaux portant une catégorie erronée (19/62). Une analyse des noms concernés par le problème de bruit (catégorie erronée) et de silence (catégorie absente) permet de faire plusieurs remarques qui sont énumérées ci-dessous :

1. Le problème de catégorie erronée ou bruit touche particulièrement les termes ambigus, c'est-à-dire les termes susceptibles d'avoir plus d'une interprétation : *site* (A, F, T), *régulation* (F, P), *sujet* (S, thème), etc.
2. Cette erreur (bruit) provient de différentes sources, notamment des règles implémentées dans le but d'améliorer la qualité de l'annotation sémantique. Le tableau 5.11 présente les différentes sources de bruit qui ont été identifiées, avec le nombre de termes illustrant chaque cas dans les différents extraits de données évaluées :

   — Certaines erreurs relèvent de la règle qui permet au système de récupérer la catégorie sémantique de la tête lexicale d'un groupe nominal et de l'associer aux autres occurrences de cette tête dans le corpus (cf. chapitre 3, section 3.2.2). Cette règle devient problématique lorsque le mot ambigu a une seule catégorie

TAB. 5.11 – Les différentes sources du bruit.

corpus	**ambiguïté** non signalée	**ambiguïté** signalée	**têtes déverbales** multicatégorielles	**autres**	**total**
PRO	2	10	2	1	15
ETU	2	8	1	0	10
VUL	4	6	0	1	11
FOR	4	12	0	3	19

possible dans la Snomed ; c'est ce que nous appelons *ambiguïté non signalée*. 12 cas ont été recensés, principalement dans les corpus VUL et FOR (cf. tableau 5.11). *Régularisation* illustre ce cas de figure. Dans la terminologie Snomed, ce terme n'apparaît qu'une seule fois, notamment en tête du groupe nominal *régularisation de l'appétit*, qui porte la catégorie F. À partir de cette occurrence, notre système d'annotation retient qu'il peut attribuer la catégorie F [19] aux termes commençant par *régularisation*, et il applique désormais cette règle. Par conséquent, s'il advient que *régularisation* intervienne dans un contexte où il a un autre sens, le système ne saura pas le distinguer, ce qui produit le bruit. Dans la phrase ci-dessous (tirée du corpus d'évaluation ETU), le nom *régularisation* a été annoté F, pourtant, dans ce contexte, il ne renvoie pas à une fonction de l'organisme (F) mais plutôt à une procédure (P) :

5) *Dans les autres cas, on tentera une régularisation après traitement anticoagulant efficace 3 semaines*.

— D'autres formes de bruit relèvent de la règle qui permet au système de sélectionner une catégorie sémantique (parmi plusieurs) à appliquer aux termes à partir de la forte fréquence de cette catégorie dans la terminologie Snomed (*ambiguïté signalée*). Cette règle a été appliquée à certains noms déverbaux en *-ion, -ment et -age*, mais elle a été principalement utilisée pour le traitement des termes ambigus non déverbaux (*site, zone, forme*). Contrairement aux déverbaux, qui ont des propriétés internes communes et qui de ce fait ont bénéficié d'une analyse particulière, les non-déverbaux n'ont pas reçu un traitement spécial car ils ne partagent pas systématiquement des propriétés communes. D'après le tableau 5.11, le bruit qui en résulte serait la principale erreur de notre système, car elle cumule le plus de candidats. Au total, 38/200 cas d'erreurs de ce type ont été identifiés à partir des données évaluées. Le terme *site* dans la phrase 6 (tirée du PRO) constitue un exemple. La catégorie F lui a été associée au lieu de A :

19. F : fonction de l'organisme, P : procédure, A : agent, T : topographie ou anatomie.

6) *Ce site_F comprend une première partie destinée aux usagers indiquant les conditions d'accès aux consultations ou à l'hospitalisation et une deuxième partie professionnelle concernant l'enseignement et les activités de recherche et de publication*.

Dans la terminologie Snomed, *site* apparaît à la tête des termes de différentes catégories (A, F, T) dont la plus fréquente est F. Cette prédominance de la catégorie F sur les autres fait qu'au cours du processus d'annotation, elle est associée aux termes inconnus commençant par le nom *site*, d'où la génération des erreurs comme celle de la phrase 6.

— Certains cas d'attribution de mauvaises catégories sémantiques proviennent de la mauvaise analyse des noms déverbaux *-ion, -ment, -age*. D'après les données du tableau 5.11, ce type de bruit intervient très peu. Au total, seulement 3 cas ont été identifiés dans les 4 extraits de données évalués.

— Malgré tous nos efforts de catégorisation des types d'erreurs, il y a quelques termes (au total 4) portant chacun une catégorie sémantique erronée dont nous n'avons pas su déterminer la source. Pour la plupart, il s'agit de termes ayant été associés à une catégorie totalement inattendue. Ce problème a été rencontré dans le corpus des experts (1 cas) et dans celui des forums (3 cas). Un exemple tiré du corpus des forums est le syntagme *pouregleres apnée du sommeil* qui a été associé à la catégorie S, malgré que ce syntagme commence par un mot qui n'existe pas. Ce mot est certainement le résultat d'une faute de saisie.

En ce qui concerne les silences, c'est-à-dire les groupes nominaux retournés sans catégories à l'issue de l'annotation sémantique, le corpus des forums et le corpus VUL sont les premiers de la liste, avec 8 exemples chacun (cf. chapitre 4, tableau 5.10). L'évaluation nous a permis de relever les différentes sources de silence. Le tableau 5.12 contient un récapitulatif.

TAB. 5.12 – Les différentes sources du silence.

corpus	**termes inconnus**	**incompatibilité de terme**	**variation terminologique**	**fautes**	**total**
PRO	2	1	0	0	3
ETU	0	5	1	0	6
VUL	4	3	0	1	8
FOR	3	1	0	5	8

D'après nos observations, *l'incompatibilité des termes*, c'est-à-dire l'incompatibilité entre les formes lexicalisées des groupes nominaux du corpus et celles des termes Snomed, est la principale cause de la non-attribution de catégories sémantiques à certains termes. Au total, 10 cas ont été identifiés dans le tableau 5.12.

En effet, les résultats de l'évaluation ont permis de constater que de nombreux groupes nominaux provenant du corpus ont une forme complexe (*des successions de blocages, l'existence d'une artériopathie*) qui diffère de celle rencontrée dans la Snomed. En général, le terme clé est précédé d'un groupe déterminatif (déterminant-nom-préposition), tandis que dans la terminologie, le terme clé sera lexicalisé de façon différente. Cette lexicalisation est parfois dépourvue de déterminant, parfois précédée d'un groupe déterminatif distinct. Cette divergence entraîne l'échec de l'appariement automatique entre la forme du terme répertoriée dans la Snomed et celle provenant du corpus, ceci malgré les efforts déployés afin de prévenir ce type de problème. La ressource mise au point à cet effet s'est montrée limitée (cf. annexe C.2). Dans la phrase suivante, le terme *blocage* a été retourné sans catégorie sémantique, pourtant il est bien répertorié dans la terminologie Snomed.

7) *Le patient est gêné par des successions de blocages et de doigts à ressaut.*

La seule raison plausible pouvant expliquer la non-reconnaissance du terme *blocage* par notre système d'annotation est la présence du syntagme *successions de*, qui crée une incompatibilité entre la chaîne de caractères provenant du corpus et son correspondant enregistré dans la terminologie. Dans certains cas, il s'agit de groupes déterminatifs. La résolution de ce problème de divergence de lexicalisations n'est pas une chose aisée, d'autant plus qu'il s'agit d'un phénomène qui relève du caractère vivant de la langue et du degré de productivité des auteurs des textes. En effet, l'impuissance de notre ressource à reconnaître les groupes déterminatifs (et autres syntagmes de ce type) a permis de comprendre qu'il est impossible de recenser la totalité de ces syntagmes dans la langue française car diverses combinaisons sont possibles selon le rédacteur et dépendant de ce qu'il veut exprimer. Quelques exemples tirés des données évaluées sont : *un index de, un épisode de, une succession de, une étendue de*, etc.

Les données du tableau 5.12 indiquent également qu'un bon nombre de silences sont la conséquence immédiate de l'absence des noms concernés dans la terminologie Snomed. 9 cas de ce type ont été dénombrés et signalés dans la colonne nommée *termes inconnus*. Nous les caractérisons comme *inconnus* puisqu'il s'agit de termes médicaux non répertoriés dans la nomenclature Snomed. Le nom *désagrégation des caillots* fait partie de ces termes inconnus figurant dans les phrases évaluées (cf. corpus PRO).

8) *D'autres médicaments accélèrent la désagrégation des caillots sanguins déjà formés par injection intraveineuse d'agents activeurs du processus de thrombolyse.*

Les noms de médicaments tombent aussi dans les silences car la terminologie Snomed ne les recense pas. Néanmoins, grâce aux enrichissements que nous avons apportés à la Snomed, certains noms de médicaments ont été annotés dans nos textes. Ils ont été associés à la catégorie C (produits chimiques) de la Snomed.

L'absence de catégorie sémantique peut également être liée aux fautes d'orthographe qui affectent les noms annotés. Ce type d'erreur intervient principalement dans le corpus des forums

où l'on rencontre un nombre non négligeable (5) d'exemples (cf. tableau 5.12). Le corpus de vulgarisation n'est pas épargné. Le mot *procedure*, retourné sans catégorie dans une phrase évaluée de ce corpus VUL peut servir d'illustration :

9) *La procedure est répétée pour toutes les veines perforantes malades.*

Bien qu'étant très peu représentée dans les résultats de notre évaluation, la variation terminologique qui caractérise les textes du corpus et la terminologie peut également être à l'origine du silence. Ce phénomène, qui a été amplement décrit à la section 4.2.2.1 du chapitre 4, ne touche qu'un seul terme (du corpus des experts) parmi les données évaluées (cf. tableau 5.12).

10) *épisode d'instabilité* vs. *instabilité articulaire, instabilité émotionnelle, instabilité moléculaire, instabilité musculosquelettique*

11) *L'hydarthrose accompagne volontiers un épisode d'instabilité.*

Dans la phrase ci-dessus, le terme *épisode d'instabilité* a été retourné sans catégorie sémantique. Alors que dans la terminologie Snomed sont répertoriés les termes *instabilité articulaire, instabilité émotionnelle, instabilité moléculaire, instabilité musculosquelettique*, qui désignent différentes formes d'instabilités, *épisode d'instabilité* est en quelque sorte un terme générique, une forme d'hyperonyme de tous ces termes.

5.4 Évaluation de la ressource de simplification

5.4.1 But et population

Cette évaluation a pour but de nous permettre d'apprécier la qualité et la fiabilité des informations fournies par la ressource de simplification que nous proposons comme principal résultat de ce travail de thèse.

L'évaluation a été effectuée en deux phases, par deux groupes constitués de 10 évaluateurs chacun :

— une équipe de linguistes : elle est composée de chercheurs, d'enseignants et de quelques étudiants en linguistique, vivant en France. 8/10 d'entre eux sont des locuteurs natifs du français. Les deux autres ont pour langue maternelle l'anglais.

— une équipe de jeunes étudiants de niveau licence (première et deuxième année) appartenant à différentes filières scientifiques : banque et finances, marketing, informatique et statistique. 9 résident en France et 1 en Allemagne, mais tous sont des locuteurs natifs du français, qui s'expriment couramment dans cette langue. Il faudrait néanmoins souligner que nos répondants ont des rapports différents avec la langue française : 5 font des études en français, 4 sont bilingues et font des études en français et en anglais, tandis que le dernier fait des études en anglais et en allemand.

Malgré les différentes caractéristiques linguistiques des répondants (autant les non-linguistes que les linguistes), nous avons délibérément choisi ces personnes premièrement parce qu'elles peuvent comprendre et s'exprimer couramment en français, ce qui correspond à notre principal critère. Le fait que les personnes interrogées soient exposées à d'autres langues que le français n'a pas freiné la sélection. Au contraire, cela pourrait apporter un paramètre en plus à l'évaluation [20], celui d'avoir une idée jusqu'à quel point le contenu de notre ressource pourrait être compréhensible et donc exploité par un non-natif du français, un bilingue, ou encore par un natif qui vit dans un milieu non francophone et/ou poursuit ses études et pratique au quotidien une langue autre que le français.

5.4.2 Démarche et justification du choix de la méthode

L'évaluation porte sur un ensemble de phrases illustrant les formes simplifiées des patrons syntaxico-sémantiques qui apparaissent en entrées de notre ressource de simplification. Autrement dit, les 50 phrases évaluées (cf. Annexe D.1) sont des phrases simplifiées qui instancient des patrons syntaxico-sémantiques, choisis au hasard dans notre ressource. Nous avons effectué la simplification en remplaçant les verbes des phrases d'origine [21] (tirées du corpus des experts) par des verbes équivalents que propose notre ressource. De même, afin de permettre aux évaluateurs de se focaliser uniquement sur le sens des verbes (qui constituent l'objet de l'évaluation), certaines longues phrases ont été raccourcies.

12) *L'exonération du ticket modérateur peut être donnée de façon ponctuelle conformément aux dispositifs réglementaires ou de manière continue lorsque le patient souffre d'une affection de longue durée.*

13) *Le patient souffre d'une affection de longue durée.*

La phrase de l'exemple 12 a été tronquée de façon à obtenir celle de l'exemple 13.

Les phrases évaluées ont été présentées aux répondants sous forme de formulaire avec des cases à cocher. Pour chaque phrase, l'évaluation consiste à dire si oui ou non, le sens du verbe est compréhensible par le répondant. Au total, chaque phrase est annotée par 20 personnes différentes, dont 10 experts en linguistique et 10 non-linguistes.

Le choix de cette méthode (double évaluation) a été motivé par le besoin d'avoir des retours non seulement de personnes expertes en linguistique mais aussi et surtout, de personnes n'ayant aucune prédisposition à comprendre facilement le langage médical. Ce deuxième groupe de personnes est la principale cible de notre évaluation. En effet, les non-linguistes représentent des patients potentiels n'ayant aucune connaissance du domaine médical mais étant tout de

20. Néanmoins, pour aboutir à des conclusions fiables basées sur ce paramètre, il serait nécessaire d'augmenter le nombre d'évaluateurs exposés à d'autres langues que le français.

21. L'annexe D, cf. section D.2 présente les 50 phrases de départ (avant la simplification).

même susceptibles d'être confrontés à la lecture de textes médicaux complexes publiés via des pages Web. Notre ressource a été mise en place principalement pour ce type de personnes qui consultent des pages Web à la recherche d'informations médicales.

5.4.3 Résultats

5.4.3.1 Évaluation par les linguistes

Le tableau 5.13 contient les résultats de l'évaluation de notre ressource par les linguistes. Les répondants non natifs sont identifiés par l'indice *NN*.

TAB. 5.13 – Résultats de l'évaluation par les linguistes.

	R1	**R2**	**R3**	**R4-NN**	**R5**	**R6-NN**	**R7**	**R8**	**R9**	**R10**
oui	45	50	50	50	50	49	47	43	48	50
non	2+3A	0	0	0	0	1	3	7	2	0

Les données du tableau 5.13 permettent de remarquer que sur 10 répondants, 5, parmi lesquels 1 non natif, déclarent comprendre le sens des verbes de toutes les phrases proposées. Les 5 autres linguistes indiquent ne pas comprendre le sens du verbe dans un certain nombre de phrases qui varie de 1 à 7 selon les répondants. Le répondant numéro 1 identifie deux verbes comme étant difficilement compréhensibles et indique que dans 3 autres phrases les verbes proposés sont ambigus (*3A*). Dans l'analyse qui suit, ces 3 phrases seront considérées comme des phrases portant l'étiquette *non*.

D'un certain point de vue, ce constat pourrait surprendre, puisque l'on s'attendrait à ce qu'en tant que linguistes, les répondants du premier groupe soient à même de comprendre, sans aucune difficulté, le sens de tous les verbes des phrases simplifiées (phrases aux constructions verbales spécialisées dans lesquelles les verbes ont été remplacés par leurs équivalents). La compétence de l'expert en linguistique le rend prompt à identifier et à questionner toute tournure, structure, phrase, tout mot, verbe de la langue dont l'emploi semble se détacher de ce qui peut être considéré comme la norme. Les quelques cas d'ambiguïté ou d'incompréhension signalés par les linguistes face à nos verbes simplifiés pourraient donc être le fruit de questionnements face à des phénomènes qui paraîtraient étranges.

Par ailleurs, il est important d'attirer l'attention sur le fait que dans cette évaluation, nous avons affaire à des phrases d'un type un peu particulier. Elles sont certes en langue française, mais elles relèvent d'un langage de spécialité. Bien qu'elles aient des verbes simplifiés, ces phrases appartiennent au langage technique de la médecine, ce qui implique la présence de termes techniques et de tournures propres au domaine médical (*évaluer un patient, un symptôme qui est isolé*).

La compétence des linguistes leur permet d'analyser les phrases et d'essayer de comprendre

le sens des verbes d'un point de vue linguistique. Or, les types de tournures qui caractérisent les textes médicaux ne sont pas systématiquement équivalents à ceux de la langue générale. Par conséquent, il peut y avoir des dissemblances entre les deux systèmes, ce qui peut provoquer, même chez les linguistes, des conflits d'interprétation et de compréhension. Pour preuve, les phrases dont les verbes ont été signalés comme étant incompréhensibles enregistrent toutes des avis divergents de la part des linguistes. Le tableau 5.14 présente les 14 phrases auxquelles les évaluateurs ont attribué l'étiquette *non*. Chaque phrase est précédée d'un numéro qui l'identifie dans la ressource de simplification et est suivie du nombre de *non* enregistrés.

TAB. 5.14 – Les phrases portant l'étiquette *non*.

ID	Phrases	non
13	La salpingite est **suivie** de la fièvre et d'une vive douleur.	2
14	Ces différences moléculaires entre cellules cancéreuses **se manifestent** par des pronostics et des réponses aux traitements très variables.	3
85	L'appréciation de la symptomatologie fonctionnelle, la palpation l' auscultation des artères [...] peuvent **déceler** une lésion anévrismale.	4
15	Les adénocarcinomes **commencent** le plus souvent à partir des cellules des canaux galactophores ou carcinome canalaire.	1
16	Le système nerveux **commande** les muscles, bien que certains muscles peuvent fonctionner de façon autonome.	1
34	Le cerveau **transforme** les stimulis en sensations.	1
36	La protection rénale **passe** par un contrôle strict de la pression artérielle.	1
153	L'utilisation de l'écarteur de Parks ouvert de 4 cm baisse significativement la pression de repos à 6 et 12 semaines.	1
187	Les patients concernés **bénéficient** alors d'une surveillance renforcée.	1
191	Il faut entreprendre la médication contre le TDAH à la recommandation de la personne qui **découvre** et suit le patient ayant le TDAH.	1
43	L'utilisation de dabigatran n'est pas **conseillée** chez les patients atteints d'une insuffisance hépatique.	1
52	Une hépatite auto-immune est **suivie** d'anticorps anti-muscle lisse.	1
53	Cet examen est **fait** depuis l'introduction de l'échographie transvaginale.	1
55	L'échelle de probabilité des interactions médicamenteuses **signale** une relation probable entre la chute soudaine des concentrations et l'introduction du méropénem chez cette patiente.	1

Une analyse des 14 phrases du tableau 5.14 permet de constater qu'aucune n'a été caractérisée d'incompréhensible à l'unanimité par les 10 linguistes. De plus, d'après nos résultats, 3 phrases seulement comptent plus d'un *non* et le maximum de *non* obtenus pour une phrase est de 4. Autrement dit, la tendance générale est telle que les phrases sont comprises par la majorité des linguistes. Nous pouvons en déduire que les prédicats verbaux que notre ressource fournit comme substituts des verbes spécialisés sont convaincants pour un public de linguistes. La question est de savoir si cette conclusion est également applicable aux non-linguistes qui représentent des

utilisateurs potentiels de notre ressource.

5.4.3.2 Évaluation par les non-linguistes

Le tableau 5.15 contient les résultats de l'évaluation de notre ressource par les non-linguistes. Les répondants bilingues sont identifiés par l'indice *B*, et le répondant bilingue qui réside en Allemagne porte un indice supplémentaire (*A*).

TAB. 5.15 – Résultats de l'évaluation par les non-linguistes.

	R1-BA	**R2**	**R3**	**R4**	**R5-B**	**R6-B**	**R7-B**	**R8-B**	**R9**	**R10**
oui	37	50	49	48	45	50	50	50	50	50
non	13	0	1	2	5	0	0	0	0	0

À notre grande surprise, les résultats de l'évaluation effectuée par les non-linguistes sont meilleurs que ceux des linguistes. Les données du tableau 5.15 indiquent que sur 10 répondants, 6, dont 3 bilingues (français-anglais) comprennent le sens des verbes dans l'ensemble des phrases proposées. Un constat plutôt positif car on pourrait en déduire que les verbes de notre ressource sont compréhensibles non seulement par des locuteurs natifs n'ayant pas de connaissances particulières du domaine médical, mais aussi par des natifs qui ne vivent pas dans un pays francophone et qui sont exposés à d'autres langues au quotidien. Bien entendu, il s'agit là d'une interprétation à relativiser, en particulier en tenant compte du nombre de personnes bilingues prises en compte (3). La confirmation d'une telle conclusion requiert que soit interrogé un plus grand nombre de personnes ne vivant pas dans un pays francophone.

4/10 répondants signalent avoir rencontré des difficultés avec 18 phrases parmi lesquelles 13 proviennent d'une seule personne qui semble être un « outlier ». Comme le montre le tableau 5.15, il s'agit du répondant bilingue qui réside en Allemagne. Le tableau 5.16 présente ces 13 phrases, ainsi que toutes celles qui ont reçu un avis négatif (*non*) de la part des répondants. Chaque phrase est précédée d'un numéro qui l'identifie dans notre ressource de simplification et est suivie du nombre de *non* enregistrés.

Comme nous l'avons souligné précédemment, 72% des phrases du tableau 5.16 sont le produit de l'évaluation du répondant 1 du tableau 5.15. Le cas de ce répondant est assez spécial et particulièrement intéressant car il se démarque considérablement des autres. En effet, au terme de l'évaluation, il enregistre 13 verbes incompris, contre 37 compris, ce qui est pour nous un bon résultat, vu le type de rapport qu'il entretient avec la langue française. Sa pratique du français, bien qu'étant courante, est limitée au cadre familial. Vivant en Allemagne et étudiant en anglais, cette personne est fréquemment en contact avec d'autres langues qu'elle pratique couramment. Cette situation pourrait restreindre sa capacité de compréhension et surtout d'interprétation du français, sachant de surcroît que dans nos phrases, il s'agit d'un français technique, qui décrit un domaine de spécialité. Par contre, les autres répondants bilingues ont l'avantage de résider

TAB. 5.16 – Les phrases portant l'étiquette *non*.

ID	Phrases	non
60	Dans cette cliniques externes, l'on **donne** aux patients des médicaments dangereux par voie parentérale.	1
7	Malheureusement, seulement 40 % des femmes belges **font** le dépistage.	1
8	Les changements squelettiques sont **liés** à des modifications radiologiques.	1
116	Son traitement actuel **combine** le Coversyl et le Fludex.	1
13	La salpingite est **suivie** de la fièvre et d'une vive douleur.	2
14	Ces différences moléculaires entre cellules cancéreuses **se manifestent** par des pronostics et des réponses aux traitements très variables.	1
85	L'Electrocardiogramme **décèle** les troubles de conduction.	1
17	Le traitement précoce des IST **dépend** de la prévention primaire.	1
19	Les virus sauvages ont **formé** des systèmes très efficaces de transfert de gènes.	1
131	Le patient **souffre** d'une affection de longue durée.	1
29	Les analyses sanguines **signalent** une leucocytose légère.	1
151	Le cancer de la paroi **évolue** à bas bruit.	1
153	L'utilisation de l'écarteur de Parks ouvert de 4 cm **baisse** significativement la pression de repos à 6 et 12 semaines.	1
51	Si le malade **fait** une infection pulmonaire, il faut la traiter rapidement.	1
52	Une hépatite auto-immune est **suivie** d'anticorps anti-muscle lisse.	3
128	Quatre patients ont **fait** un arrêt cardiaque d'issue fatale.	1
55	L'échelle de probabilité des interactions médicamenteuses **signale** une relation probable entre la chute soudaine des concentrations et l'introduction du méropénem chez cette patiente.	1
114	La tolérance et l'efficacité doivent être régulièrement **testées** par la surveillance régulière du poids.	1

en France et d'étudier en français. Même si ces études se font en parallèle avec l'anglais, il est indiscutable que ces personnes résidant en France ont un meilleur contact avec le français comparés au répondant 1.

Le croisement des données des tableaux 5.14 et 5.16 permet de faire une remarque intéressante. Un certain nombre de phrases sont identifiées comme incompréhensibles à la fois par les linguistes et les non-linguistes. Le tableau 5.17 présente ces 5 phrases avec le nombre d'avis négatifs enregistrés pour chacune. L'étiquette *NL* désigne les non-linguistes, tandis que *L* désigne les linguistes.

Les phrases du tableau 5.17 totalisent chacune au minimum 2 avis négatifs. Elles se répartissent comme suit :

— une phrase totalise 2 *non* : phrase 55 ;

— 3 phrases totalisent 4 *non* chacune : il s'agit des phrases 13, 14, et 52 ;

— la phrase 85 semble être la plus problématique puisqu'elle enregistre le maximum de *non*, au total 5.

TAB. 5.17 – Les phrases portant l'étiquette *non* chez les linguistes et les non-linguistes collectivement.

ID	Phrases	NL	L
13	La salpingite est **suivie** de la fièvre et d'une vive douleur.	2	2
14	Ces différences moléculaires entre cellules cancéreuses **se manifestent** par des pronostics et des réponses [...] très variables.	1	3
85	L'électrocardiogramme **décèle** les troubles de conduction.	1	4
52	Une hépatite auto-immune est **suivie** d'anticorps anti-muscle lisse.	3	1
55	L'échelle de probabilité des interactions médicamenteuses **signale** une relation probable entre la chute soudaine des concentrations et l'introduction du méropénem chez cette patiente.	1	1

Une analyse de ces phrases permet d'émettre quelques hypothèses qui expliqueraient le fait qu'elles causent des difficultés de compréhension autant chez les linguistes que chez les autres personnes interrogées :

De prime abord, nous avons pensé à la longueur des phrases. Mais comme 2 d'entre elles sont relativement longues et 3 plutôt courtes, ceci nous pousse à éliminer ce facteur dans le cadre de cette évaluation. De plus, comme nous l'avons signalé au début de cette évaluation, certaines phrases ont été raccourcies afin de pousser les évaluateurs à se focaliser uniquement sur les verbes.

La forte présence des termes médicaux pourrait être un facteur de la difficulté de compréhension de ces phrases. En effet, comme nous l'avons vu à la section 3.3.2 du chapitre 3, l'évaluation du degré de lisibilité d'un texte en langue de spécialité passe aussi par l'évaluation de sa teneur en termes spécialisés qui contribuent à rendre la compréhension difficile pour des lecteurs non experts du domaine. Ce paramètre est d'ailleurs pris en considération dans plusieurs études (Vor der Brück *et al.*, 2008 ; Abrahamsson *et al.*, 2014 ; Vajjala & Meurers, 2014). L'extrait évalué (cf. Annexe D) compte plusieurs phrases (3, 40, 44, 60, 114, 119, etc.) dont les positions d'arguments des verbes sont occupées par au moins un groupe nominal utilisé d'une façon relativement fréquente dans le langage quotidien. Lors de l'évaluation, la plupart de ces phrases ont fait l'unanimité des répondants des deux groupes, en recevant des avis positifs en ce qui concerne la compréhension du sens du verbe. Le tableau 5.18 présente quelques exemples.
Par contre, à l'opposé, les phrases (cf. tableau 5.17) ayant reçu des avis négatifs venant à la fois des linguistes et des non-linguistes ont en moyenne deux termes médicaux spécialisés en position d'arguments. La phrase 55 compte à elle seule au moins deux termes médicaux qui correspondent aux arguments du verbe : *l'échelle de probabilité des interactions médicamenteuses*, *la chute des concentrations*, *l'introduction du méropénem*. Dans la phrase 85, le verbe a deux termes médicaux comme arguments : *électrocardiogramme* et *troubles de conduction*. Il en est de même pour les trois autres phrases qui comportent chacune au moins deux termes spécialisés comme arguments. Ce paramètre est d'autant plus important que le sens du verbe dépend

TAB. 5.18 – Quelques phrases ayant des groupes nominaux compréhensibles en position d'arguments.

ID	Phrases
3	Les effets indésirables se sont manifestés dans les 24 heures de l'administration de l'IgIV [...]
40	L'enfant insuffisamment transfusé va avoir un retard de croissance [...]
60	Dans cette cliniques externes, l'on donne aux patients des médicaments dangereux par voie parentérale.
114	La tolérance et l'efficacité doivent être régulièrement testées par la surveillance régulière du poids.

de son contexte d'apparition, plus précisément, de ses arguments. En d'autres termes, si un verbe simplifié est entouré de termes difficilement compréhensibles, alors il est fort probable que l'interprétation du sens de ce verbe soit difficile, puisque ses compléments ont un sens opaque pour le lecteur.

Par ailleurs, nous pensons que la construction sémantique de base, qui repose sur le type de catégories Snomed impliquées, pourrait être une source d'ambiguïtés. En effet, les verbes intervenant dans les phrases du tableau 5.17, c'est-à-dire les verbes utilisés comme substituts lors de la simplification, interviennent au sein d'un patron sémantique dont le sens reste inchangé, même après la simplification. Lors de l'alignement en vue de la simplification, des changements peuvent être opérés au niveau du lexique, de la syntaxe et de la forme de surface de la phrase, mais le sens de base de la construction (qui dépend des catégories Snomed impliquées) reste le même.

Dans la phrase 14, la construction de base, non instanciée est FONCTION_DE_ORGNANISME se verbe par FONCTION_DE_ORGNANISME. La catégorie Snomed impliquée est FONCTION_DE_ORGNANISME, caractérisant à la fois le sujet et le complément du verbe. Dans la phrase d'origine, cette construction est instanciée par le verbe *se traduire*, qui a été remplacé par le verbe *se manifester*, mais le sens est resté le même puisque les catégories Snomed de la construction sont restées les mêmes. Dans la phrase de l'exemple 14, qui ne figure dans aucun des tableaux car ayant été attestée par tous les évaluateurs, la construction de base a connu une modification au niveau de la syntaxe. Le verbe de base (*poursuivre*) a été remplacé par une périphrase verbale *continuer de prendre* portant un autre verbe (*prendre*) ; néanmoins les types de catégories Snomed (S : STATUT SOCIAL et C : PRODUIT CHIMIQUE) sont ceux de la construction de départ. Le rôle de la périphrase est de maintenir le sens de base de la construction, celui d'une consommation continue du médicament dont le nom joue la fonction de sujet de la construction.

14) *Si la patiente continue de prendre le nadolol jusqu'à l'accouchement, en informer l'équipe de la maternité pour lui permettre d'adapter la surveillance du nouveau-né.*

Ces constructions ou combinaisons de catégories Snomed, ainsi que leurs interprétations, sont propres au langage médical et déterminent le sens des verbes impliqués (*se manifester, prendre*). Ce qui signifie que le choix des verbes et de leurs formes est fortement lié aux sens qu'imposent les patrons sémantiques de base. Par conséquent, une mauvaise analyse et interprétation de la construction de base (par un lecteur) pourrait déboucher sur une mauvaise compréhension du sens du verbe. Toutefois, ce constat implique également que la substitution des verbes par des équivalents plus transparents ne suffit pas toujours pour assurer la compréhension effective du sens du verbe.

5.5 Bilan

La réalisation des différentes évaluations nous a permis de juger de la robustesse des outils et méthodes utilisés dans le cadre de ce travail de thèse.

L'évaluation des différentes tâches d'annotation nous a permis de détecter et de décrire les paramètres qui peuvent restreindre la validité des différents résultats de cette thèse. Nous avons ainsi pu prendre du recul afin d'identifier les éléments à améliorer ou à changer dans notre chaîne de traitement, en partant de l'annotation syntaxique des corpus avec Cordial, jusqu'à l'acquisition des PSS, en passant par l'annotation sémantique avec la Snomed.

En ce qui concerne la ressource de simplification, l'évaluation nous a permis de constater que la substitution du verbe par un verbe non spécialisé aide bel et bien les lecteurs à mieux comprendre les phrases. Les différents cas de figure observés grâce aux phrases évaluées ont permis de remarquer que notre méthode tend à fonctionner parfaitement sur des phrases de préférence courtes, dans lesquelles les verbes ont parmi leurs arguments au moins un groupe nominal provenant de la langue générale et idéalement, uniquement des arguments de ce type.

Nous retenons donc qu'une simplification des termes-arguments des verbes amplifierait considérablement les résultats de l'étude que nous avons proposée dans ce travail de thèse. Autrement dit, les données de notre ressource de simplification sont susceptibles d'avoir un meilleur impact sur des phrases où les termes jouant le rôle d'arguments sont eux aussi adaptés au niveau de connaissances des non-experts. Ce constat est tout à fait pertinent, d'autant plus que la compréhension ou l'interprétation du sens d'un verbe dépend également de la connaissance de ce à quoi réfèrent les termes qui jouent le rôle de compléments de ce verbe. En conclusion, nous retenons que pour un meilleur rendement, le travail de simplification à faire sur les verbes requiert une simplification complémentaire des noms qui les entourent.

CONCLUSION

Au terme de ce travail, nous proposons un bilan des principaux résultats obtenus. Ce bilan soulève de nouvelles interrogations et ouvre la voie à de nouvelles perspectives de recherches.

Principaux résultats obtenus

Dans ce travail de thèse, il était question d'effectuer une étude comparative du fonctionnement des verbes dans quatre corpus médicaux ayant différents niveaux de spécialisation, afin d'extraire les données de base pour la création d'une ressource de simplification de textes axée sur les verbes. La réalisation de ce travail a requis différentes étapes : l'annotation syntaxique des corpus avec l'analyseur syntaxique Cordial ; l'annotation sémantique des arguments des verbes grâce aux catégories de la terminologie Snomed Internationale ; l'acquisition des patrons syntaxico-sémantiques (PSS) à partir d'une méthode semi-automatique achevée par une phase de validation manuelle de ces PSS par des experts en médecine ; l'analyse contrastive du fonctionnement des PSS et des collocations verbe-nom dans les quatre types de corpus. Cette chaîne de travail a débouché sur la conception d'une ressource contenant 230 patrons syntaxico-sémantiques spécialisés, alignés avec des équivalents non spécialisés convenables pour un public de non-experts.

Notre ressource de simplification a été conçue en première intention comme un dictionnaire à intégrer dans un outil de simplification automatique. Afin de juger la qualité et la valeur de cette ressource, nous avons réalisé une évaluation de son contenu à deux niveaux : par des linguistes et par des non-linguistes, tous « profanes » du domaine médical. Les résultats de l'évaluation montrent que la substitution des verbes par des équivalents non spécialisés et, lorsque nécessaire, la révision de la structure de la phrase (voix passive vers voix active, restitution des arguments implicites, etc.) aident bel et bien les lecteurs à mieux comprendre les phrases simplifiées. Notre ressource obtient de meilleurs résultats sur les phrases dont les arguments du verbe sont des groupes nominaux provenant de la langue générale. Grâce à cette

évaluation, nous retenons donc qu'une simplification complémentaire des termes qui jouent le rôle d'arguments améliorerait considérablement le rendement du contenu de notre ressource.

Ce travail de thèse a permis de souligner l'importance de la prise en compte et de l'analyse de la structure argumentale du verbe dans des textes de spécialité et dans le contexte de la simplification de textes. En effet, comme nous l'avons vu, le verbe peut constituer une source de difficultés de lecture dans des textes spécialisés (pour les non-experts), lorsqu'il instancie des constructions spécialisées.

Ce PATIENT *relève d'une* AFFECTION *de longue durée* | *'souffre de', 'a'*

Les résultats de l'analyse contrastive des PSS et des collocations verbe-nom ont contribué à la mise en évidence des différences, des similitudes et surtout des spécificités qui caractérisent le langage des médecins (experts) et celui des patients (non-experts), notamment en ce qui concerne l'utilisation des verbes. Plusieurs formes de variations ont été observées :

— La variation lexicale : selon les types de corpus, le choix d'unités verbales qui instancient les PSS varie. Les experts utilisent les verbes dans des constructions spécialisées, tandis que les non-experts ont tendance à transposer les emplois verbaux de la langue générale vers la langue de spécialité. Les experts utilisent un langage spécialisé, standardisé et propre à une communauté scientifique, tandis que les non-experts utilisent un langage hybride, qui est un mélange de termnologie médicale et de jargon familier, influencé par d'autres paramètres socio-culturels et personnels.

— La variation syntaxique : pour une verbe donné, certaines constructions et structures argumentales sont plus sollicitées que d'autres, en fonction des types de corpus. Il en est de même pour les alternances lors de la réalisation de ces structures argumentales. Par exemple, les experts préfèrent le passif (avec omission de l'agent) à la forme active qui est dominante chez les non-experts. Ce constat est tout à fait logique puisqu'on a affaire à des textes formels en langue de spécialité. En effet, l'utilisation de la forme passive sans agent est l'une des caractéristiques de la langue scientifique générale (Pecman, 2004) dont fait partie la langue médicale qu'utilisent les experts. De même, la forme impersonnelle est privilégiée chez les experts, tandis que les non-experts ont tendance à produire des écrits assez personnels, très axés sur eux-mêmes. Il arrive que les deux groupes de protagonistes du domaine médical utilisent les mêmes procédés mais leur associent des fonctions différentes. C'est le cas de l'emploi du pronom « on », qui chez les experts contribue à effacer la personne singulière qui écrit, pour privilégier l'objectivité, tandis que chez les non-experts, l'impersonnalisation a pour but d'aider l'énonciateur à focaliser son discours c.-à.d. à effacer le sujet grammatical de la phrase afin de mettre l'accent sur un message particulier qu'il voudrait véhiculer. Dans d'autres contextes, l'emploi du pronom « on », dans les textes tirés des forums, relève d'un registre familier.

— La variation sémantique : le choix préférentiel des constructions syntaxiques (selon les types de corpus) s'accompagne très souvent d'une variation sémantique. Par exemple, certaines constructions privilégiées chez les experts imposent aux verbes une certaine interprétation. Cette interprétation est parfois spécifique au langage médical spécialisé et ne se retrouve pas systématiquement chez les non-experts. La variation sémantique implique une variation au niveau de l'utilisation des PSS. En effet, certains patrons syntaxico-sémantiques n'interviennent que chez les experts car ils évoquent des concepts et/ou des pratiques propres au corps médical et que ne partagent pas les non-experts. Ce type de PSS représente de sources potentielles de difficultés de compréhension pour des lecteurs ayant peu de connaissances en médecine.

L'analyse contrastive des verbes et le travail d'alignement des PSS ont ainsi permis de remarquer que de façon générale, la différence entre le discours des experts et celui des non-experts se manifeste dans les textes à travers le point de vue, c.-à-d. la position de l'énonciateur par rapport à ce dont il parle. Cette notion de point de vue se matérialise à travers les différents choix lexico-syntaxiques et sémantiques que font les experts et les non-experts lorsqu'ils s'expriment. Ces choix sont illustrés par les différents types de variations décrits précédemment.

Perspectives de travail

Ce travail de thèse ouvre plusieurs perspectives dont les principales concernent les améliorations pouvant être apportées à la méthode que nous avons implémentée pour l'acquisition des données et la création de la ressource de simplification. L'automatisation totale de certaines tâches jusqu'ici réalisées de façon semi-automatique permettrait d'optimiser la chaîne de travail.

En ce qui concerne l'annotation sémantique des corpus, l'enrichissement de la terminologie Snomed et l'amélioration de la méthode de traitement des termes polysémiques (multicatégoriels) permettraient d'élargir considérablement la couverture de l'annotation sémantique et d'alléger le travail de correction manuel des résultats de cette tâche. L'enrichissement de la Snomed peut se baser sur l'exploitation de corpus médicaux de grande taille grâce auxquels de nouveaux termes pourront être détectés. La principale difficulté qu'implique cette démarche c'est l'association des catégories sémantiques Snomed aux termes nouveaux qui seront tirés des corpus. Une telle tâche exigerait donc une phase d'évaluation des résultats par des experts en médecine. L'avantage de cette démarche réside dans le fait que chaque corpus exploité pourrait apporter ses propres variations de termes.

À propos de la sélection automatique des candidats verbes équivalents (substituts) pour l'alignement, il serait intéressant de voir dans quelle mesure il est possible de créer un modèle qui permettrait de réduire les ensembles de verbes candidats équivalents, de façon automatique,

en utilisant des corpus alignés suffisamment larges, et des méthodes statistiques d'analyse distributionnelle. Cette méthode permettrait d'aborder avec plus d'aisance la sélection manuelle des verbes équivalents pour l'alignement mais aussi d'envisager une automatisation de la tâche d'alignement des patrons syntaxico-sémantiques.

En ce qui concerne les nouvelles perspectives de recherche, l'agrandissement de notre ressource constitue une première piste de travail. Pour cela, une augmentation de la taille des corpus sera nécessaire. Il serait également intéressant de voir dans quelle mesure il est possible de créer un FrameNet médical à partir des données résultant de notre modèle d'analyse.

L'analyse détaillée des PSS avec compléments circonstanciels constitue un autre pôle d'investigation. En effet, comme nous l'avons précisé au début de cette étude, les compléments circonstanciels ont été traités uniquement de façon accessoire car notre focalisation était sur les arguments. Nos analyses nous ont cependant permis de faire quelques observations qui méritent d'être approfondies, en ce qui concerne le fonctionnement de certains syntagmes prépositionnels circonstanciels (SP) au sein des PSS :

— étude du rôle des constituants introduits par *chez* dans certains PSS : MÉDECIN diagnostique/découvre MALADIE chez PATIENT ;

— détection des PSS dans lesquels la présence d'un SP marque une nuance ou une précision dans l'interprétation du sens du verbe ; MÉDECIN administre PCHIMIQUE dans AGENT : *On ne doit pas administrer PrecedexMC dans un cathéter intraveineux par lequel on administre du sang, du sérum ou du plasma, car leur compatibilité physique n'a pas été établie.*

  En l'absence du SP « dans un cathéter » dans cette phrase, on aurait pas la précision en ce qui concerne la voie par laquelle le produit est donné au patient. Cette information aide à mieux comprendre le sens du verbe et permet d'être plus précis dans la réalisation d'une tâche comme la simplification de textes. Dans cet emploi, le SP fonctionne comme un indicateur de précision. Nous pensons qu'il pourrait exister des cas où la présence du SP joue un rôle encore plus important, notamment celui d'indiquer un changement de sens. Cette technique qui consiste à indiquer un changement de sens du verbe grâce à un SP est d'ailleurs utilisée dans le projet FrameNet. La détection de ce type de données dans des corpus spécialisés pourrait alors permettre la création d'une ressource verbale qui serait d'un apport considérable pour la description, ou mieux encore, la désambiguïsation des sens des verbes, non seulement dans les textes médicaux mais aussi dans les textes d'autres domaines de spécialité.

Annexe A

Les ressources pour l'annotation syntaxique des corpus

A.1 Codage des fonctions grammaticales Cordial

TAB. A.1 – Liste des étiquettes utilisées par Cordial pour l'annotation des fonctions grammaticales.

Code	Signification	Code	Signification
A.	Constitue l'attribut du sujet	S.	Constitue le SUJET
B.	Appartient à l'attribut du sujet	T.	Appartient au SUJET
C.	Constitue le COD	U.	Pronom personnel de pronominalisation
D.	Appartient au COD	V.	Verbe de base de la proposition
E.	Constitue le COI	W.	Introduit le subordonné objet indirect
F.	Appartient au COI	Y.	Constitue le sujet réel
G.	Appartient au complément d'agent	Z.	Appartient au sujet réel
H.	Circonstanciel	a.	Ajout à l'adjectif
I.	Complément infinitif	c.	Reprise de l'attribut du COD
K.	Circonstanciel de temps	d.	Reprise du COD
L.	Circonstanciel de lieu	e.	Reprise du COI
M.	Constitue une apposition	h.	Reprise du circonstanciel
N.	Appartient à une apposition	n.	Ajout au nom
O.	Constitue une apostrophe	p.	Ajout au pronom
P.	Appartient à une apostrophe	s.	Reprise du sujet
Q.	Complément de négation	t.	Ajout au verbe

Annexe B

Les ressources terminologiques pour l'annotation sémantique des corpus

Cette annexe contient des ressources que nous avons conçues pour l'annotation sémantique des corpus, à partir des entrées de la terminologie Snomed Internationale. Elle se focalise particulièrement sur les données exploitées pour le traitement de ce que nous avons appelé les *termes multicatégoriels* ou les *têtes multicatégorielles* car ils figurent très souvent en position de tête syntaxique des unités nominales polylexicales. Il s'agit des unités nominales simples (constituées d'un seul mot) qui ont plus d'une catégorie sémantique dans la terminologie Snomed.

B.1 Les formes plurielles des termes simples de la Snomed

TAB. B.1 – Quelques exemples de termes pluriels avec leurs catégories.

abdomino-utérotomies	P	cations	C	gigantocotyloses	D
acanthocéphales	L	céphaloniums	C	gliales	T
acétabulectomies	P	cephenemyias	L	gnathostomas	L
acétonémies	D	chromatopsies	D	grébifoulques	L
acétylcholines	F	cicatrisés	D	hémiglossectomies	P
acidophiles	D	coccidioïdomycoses	D	hépatopexies	P
actinomaduras	L	confucianismes	S	Hilarias	L
allergologues	J	crotales	L	histidines	F
alligators	L	dasytrichas	L	hyperkaliémies	D
alopécies	D	délinquances	F	hyperuricémies	D
amoxicillines	C	deslanosides	C	hypo-uricémies	D

B.2 Les termes mal orthographiés

Cette annexe contient une liste de termes mal orthographiés extraits automatiquement du corpus des forums grâce au calcul de la distance de Levensthein. Au total 205 mots, chacun associé à la catégorie sémantique du terme Snomed correspondant.

TAB. B.2 – Les termes mal orthographiés du corpus des forums (1)

terme mal orthographié	**terme Snomed**	**catégorie Snomed**
abalation	ablation	F
accupuncteur	acupuncteur	A
accupuncteur	acupuncture	P
acident	accident	D
acident	accidents	D
activitees	activités	F
acumulation	accumulation	F
aiselle	aisselle	T
allarme	alarme	A
angioplasti	angioplastie	P
antécedants	antécédents	D
antorse	entorse	D
antorse	entorses	D
anxiétée	anxiété	D
arteriographie	artériographie	P
arteriographie	artériographies	P
arthére	artère	T
athéromateuse	athéromatose	D
atherosclerose	athérosclérose	D
attension	attention	F
auricul	auricule	T
belladons	belladone	L
blockage	blocage	P
cafaeine	caféine	C
cardiolgue	cardiologue	J
cardiomyopatjie	cardiomyopathie	D
cardiomyopatjie	cardiomyopathies	D
cardiopathis	cardiopathies	D
cardiopathi	cardiopathie	D
cardiopathis	cardiopathies	D
cardiopatie	cardiopathie	D
cardiopatie	cardiopathie	D
cardomyopathie	cardiomyopathie	D
cardomyopathie	cardiomyopathies	D
utgence	urgence	P
vagabon	vagabonds	S
venticule	ventricule	T

TAB. B.3 – Les termes mal orthographiés du corpus des forums (2)

terme mal orthographié	terme Snomed	catégorie Snomed	terme mal orthographié	terme Snomed	catégorie Snomed
cariologue	cardiologue	J	extremitées	extrémités	T
cellulles	cellule	T	fibbrilation	fibrillation	F
cellulles	cellules	T	gastro-anthérite	gastro-entérite	D
chevile	cheville	T	genioplastie	génioplastie	P
cholesterol	cholestérol	F	habdomen	abdomen	T
cholesterole	cholestérol	F	habdomin	abdomen	T
choléstérol	cholestérol	F	hemiplégie	hémiplégie	D
cholesténol	cholestérol	F	hemostase	hémostase	F
condcuteur	conducteur	J	heparine	héparine	C
consience	conscience	F	hepatites	hépatite	D
cresage	creusage	A	homones	hormone	F
defebrillateur	défibrillateur	A	homones	hormones	F
defense	défense	T	homoplate	omoplate	T
deplacement	déplacement	P	hpertension	hypertension	D
deplacement	déplacements	P	hupertension	hypertension	D
dépreseur	dépresseur	A	hypoplasie	hyperplasie	D
détériotation	détérioration	D	insuffisence	insuffisance	F
diabête	diabète	D	intervale	intervalles	F
diabête	diabètes	D	intervale	intervalle	F
diahrées	diarrhée	D	l'éffort	effort	F
diaphgramme	diaphragme	T	*conseil	conseils	P
douluers	douleurs	F	l'hopital	hôpital	S
echelle	échelle	A	l'ablation	ablation	P
échograaphie	échographie	P	l'hérédité	hérédité	F
echograhie	échographie	P	lomoplat	omoplate	T
echographies	échographies	P	lomoplate	omoplate	T
eintervention	intervention	P	labdomen	abdomen	T
electricité	électricité	A	linfarctus	infarctus	D
entrainements	entraînement	P	lombargie	lombalgie	D
enverloppe	enveloppe	T	machoir	mâchoire	T
épiglotite	épiglotte	T	machoire	mâchoires	T
epreuves	épreuve	P	matiere	matière	C
epreuves	épreuves	P	matiere	matières	C
evolutions	évolution	F	mèdecin	médecin	J
exercise	exercice	P	médeçins	médecins	J
exerçise	exercice	P	médiacalcose	médiacalcinose	D
exercise	exercices	P	medicamant	médicament	C
exercises	exercice	P	medicamnent	médicament	C
exmanen	examens	P	mycrocytose	macrocytose	D
exstrasystoles	extrasystoles	F	myocard	myocarde	T
naisance	naissance	F	pyrenées	périnée	T
naisance	naissances	F	ralentissament	ralentissement	F
nathuropate	naturopathe	J	reanimation	réanimation	P

TAB. B.4 – Les mots mal orthographiés du corpus des forums (3)

mot mal orthographié	terme Snomed	catégorie Snomed	mot mal orthographié	terme Snomed	catégorie Snomed
naturopathie	naturopathe	J	récidiv	récidive	D
necroser	nécroses	D	rééducation	rééducation	P
operationa	opération	P	réeduction	réduction	F
operationa	opérations	P	réeduction	rééducation	P
opration	opération	P	rempacement	remplacement	P
optitien	opticien	J	reponse	réponse	F
optitien	opticiens	J	reponse	réponses	F
ostrogene	estrogène	F	reponsse	réponse	F
pace-maker	pacemaker	A	resection	résections	P
pacmeker	pacemaker	A	resultat	résultat	F
palpitaion	palpation	P	resultats	résultat	F
paralizie	paralysie	D	scéphalées	céphalée	D
paresthesie	paresthésie	D	seignement	saignement	D
paupiere	paupière	T	seignement	saignements	D
paupiere	paupières	T	senssation	sensation	F
paupiéres	paupière	T	senssation	sensation	F
pericardite	péricardite	D	senssations	sensations	F
pericardite	péricardites	D	senstion	sensation	F
pesonne	personne	S	simptome	symptôme	F
pesonne	personnes	S	spasticite	spasticité	F
pharmatien	pharmacien	J	specialiste	spécialiste	J
picottements	picotement	F	specialiste	spécialistes	J
plusations	pulsations	F	sqpasmes	spasmes	D
pneumoccoque	pneumocoques	L	ssymptomes	symptômes	F
precardite	péricardite	D	synptomes	symptôme	F
pré-eclampsie	prééclampsie	D	synptomes	symptômes	F
prééclamsie	prééclampsie	D	syptomes	symptômes	F
prélevement	prélèvement	P	syxtoles	systole	F
prélevement	prélèvements	P	tacchycardie	tachycardie	D
premature	prématuré	S	tachicardi	tachycardie	D
preoccupations	préoccupation	F	tachychardie	tachycardie	D
presssion	pression	F	tchycardie	tachycardie	D
proffesseur	professeur	J	temperamment	tempérament	F
proffesseur	professeure	J	tenssion	tension	F
prolasus	prolapsus	D	thromobose	thrombose	D
pusation	pulsations	F	thromoboses	thrombose	D
pyocanique	pyocyanine	F	tyroides	thyroïdes	T
utgence	urgence	P	vertebres	vertèbres	T
vagabon	vagabonds	S	poissons	poisson	L
venticule	ventricule	T	poisssons	poissons	L
venticule	ventricules	T	rasins	raisins	L
vertebres	vertèbre	T	recidive	récidive	D

B.3 Les 274 têtes multicatégorielles en -ment, -ion, -age et -eur

TAB. B.5 – Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (1).

Têtes	**Catégories Snomed**	**Têtes**	**Catégories Snomed**
abduction	F,P	communication	D,P
aberration	D,F	complication	D,F
abrasion	D,P	comportement	S,F,D
absorption	F,P	compression	D,P,F
accouchement	F,D,P	concentration	C,F,P
accélérateur	A,F	condensation	F,D
accélération	D,F	condition	D,F,S
adaptation	D,S,F	conditionnement	P,F
adhésion	T,F,S	conducteur	D,J
affaissement	F,D	conduction	F,D
agglutination	F,D,P	congestion	D,F
agitation	F,D	constipation	D,F

TAB. B.6 – Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (2).

Têtes	**Catégories Snomed**	**Têtes**	**Catégories Snomed**
agrégation	D,F	constriction	D,P,T,F
alimentation	P,D	construction	P,A
allaitement	F,P	contraction	F,D
allongement	D,P,F	contusion	F,D
altération	D,F	conversion	F,P
amplification	P,F	crépitation	F,D
amputation	D,P	description	P,D
analyseur	F,A	desquamation	F,D
apprentissage	P,F	destruction	F,P
articulation	D,T,F	dilatation	D,P,F
aspiration	F,D,P	diminution	F,A,P,D
assemblage	P,F	discrimination	F,S
attrition	D,F	disjonction	D,F
augmentation	F,A,L,D,P	disparition	D,F,S
avortement	P,D,F	dissociation	F,D
avulsion	P,D	distension	F,D
bifurcation	T,P	distribution	D,P,F
blocage	D,P,F	division	L,T,P
broyeur	J,A	drainage	P,D
calcification	F,D	dysfonction	D,F
captation	P,F	dysfonctionnement	D,F
changement	F,S,P,D	décollement	D,F
chevauchement	F,D,P	dédoublement	D,P
circulation	T,F,P,D	défloration	F,P
claudication	F,D	déformation	D,F
clignement	D,F	dégagement	F,P
coagulation	P,D,F	déglutition	F,P,D
coloration	D,P,F	délétion	D,F

TAB. B.7 – Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (3).

Têtes	**Catégories Snomed**	**Têtes**	**Catégories Snomed**
dénervation	P,F	hybridation	P,F
déplacement	P,D	hypoventilation	D,F
déplétion	D,P	identification	P,D
dépolarisation	F,D	implantation	P,D,F
dépression	F,T,D	inadaptation	F,S
déraillement	A,F	incision	P,D
dérivation	F,P,D	inclusion	T,D
désarticulation	D,P	incoordination	F,D
détachement	F,P	induction	P,F
détecteur	F,A	infiltration	D,F
développement	D,F	inhalation	P,D
déviation	F,D	inhibiteur	P,D,F,C
embolisation	P,D	inhibition	P,D,F
engagement	F,S,D,P	injection	D,P,C
engorgement	D,F	insertion	F,T,P,D
engourdissement	F,D	installation	P,A
excavation	D,F,T	insémination	P,F
excitation	F,D	interaction	D,P,F
exclusion	S,F	interruption	F,P,D
excrétion	P,F	intervention	D,P
exfoliation	F,D,P	introduction	F,P
expansion	D,T	invagination	D,P,T
exploration	P,F	inversion	F,P,D
exposition	D,P,F,A	irrigation	P,F
expression	P,F	isolement	F,S,P
extension	D,P,F	jonction	T,F
extraction	D,P	kératinisation	F,D
facteur	C,J,F,S	lactation	F,D
fenestration	P,T	lavage	P,T
fibrillation	F,D	lavement	C,P
fixation	F,P	limitation	D,F
flexion	F,D	liquéfaction	P,D
formation	F,T,D,P	localisation	P,F
fraction	C,F	macrophage	T,F,D
fragment	D,A,F	malformation	D,P

TAB. B.8 – Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (4).

Têtes	**Catégories**	**Têtes**	**Catégories**	**Têtes**	**Catégories**
nutrition	F,P	restauration	P,D	tension	A,F,D,P
oblitération	D,P	rotation	F,P,D	torsion	D,F
observation	F,P	roulement	P,F	traceur	J,A
obstruction	D,F	réaction	F,S,P,D	traitement	A,F,P
occlusion	D,P,F	récepteur	T,F	transformation	F,P
oesophage	T,D	récession	P,D	transfusion	P,D
opération	P,D	réduction	P,D,F	transposition	D,P
orientation	T,F,P	régulateur	F,J,C	tremblement	A,F,D
ossification	D,F	régulation	P,F	tuméfaction	D,F
passage	T,F,P	régurgitation	D,F	variation	D,F,A
perception	D,F	réimplantation	D,P	vascularisation	D,F
perforation	D,P	réparation	D,P,F	ventilation	P,F
perturbation	D,F,S	répartition	D,F	vision	F,D
perversion	D,F	réplication	F,D	vomissement	F,D
pigment	F,C	résorption	D,F	vérification	A,P
pigmentation	D,P,F	rétablissement	P,F	écoulement	D,T,F
pincement	P,D,F	rétention	F,D	écrasement	D,P,A
position	F,D	rétraction	D,P,F	éducation	S,P
pression	A,F	rétroaction	P,F	élargissement	F,P,D
privation	S,F,D	rétrécissement	P,D,F,T	élimination	P,F
production	F,D	saignement	D,F	élongation	P,D,F
protection	P,A	saturation	F,P	élévation	F,P
prélèvement	P,D,T	section	P,D	émission	D,A,F
préparation	P,C	segment	D,F,T	énucléation	P,D
présentation	D,F	sensation	D,F	épaississement	D,F
pénétration	D,F	sevrage	P,D	épuisement	D,F
pêcheur	L,J	simulation	S,F	érection	F,D
raccourcissement	P,D,F	stimulateur	F,A,C	éruption	D,A,F
radiation	A,T	stimulation	F,P	établissement	S,F,P
rage	D,F	stockage	F,P	étirement	P,F
raideur	F,D	subinvolution	F,D	étranglement	T,D
ralentissement	F,D	succion	F,P	évacuation	P,F
rayonnement	A,P	suppression	D,P,F	évaluation	P,D
reconstitution	P,F	suspension	P,C	éversion	F,D

TAB. B.9 – Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (5).

Têtes	Catégories	Têtes	Catégories	Têtes	Catégories
fragmentation	D,P	marqueur	F,J		
frustration	F,S	mobilisation	P,F		
fusion	A,F,P,D	modification	F,P,D		
glissement	D,A	moniteur	J,A		
gonflement	D,F	mouvement	P,D,F		
respiration	F,P	teneur	F,J		
reconstruction	D,P	synchronisation	F,P		
recouvrement	P,T	sécrétion	F,T,D		
relation	T,F	sélection	F,P		
relâchement	P,D,F	séparation	F,D,P		
remplissage	F,P	séquestration	F,P,D		

B.4 Fréquence des têtes multicatégorielles dans la Snomed

TAB. B.10 – Têtes dont la catégorie la plus fréquente enregistre un pourcentage ≥ 90.

Têtes	Catégorie	Nb occ.	Pourcentage
opération	P	694	99,86
section	P	222	99,55
réparation	P	855	99,53
extraction	P	422	99,53
identification	P	147	99,32
reconstruction	P	112	99,12
évaluation	P	167	98,82
obstruction	D	83	98,81
complication	D	149	98,68
réduction	P	342	98,56
intervention	P	179	98,35
drainage	P	102	98,08
traitement	P	95	97,94
implantation	P	189	97,93
production	F	46	97,87

TAB. B.11 – Têtes dont la catégorie la plus fréquente enregistre un pourcentage ≥ 90.

Têtes	**Catégorie**	**Nb occ.**	**Pourcentage**
production	F	46	97,87
injection	P	272	97,84
construction	P	35	97,22
articulation	T	156	96,89
jonction	T	28	96,55
introduction	P	27	96,43
irrigation	P	52	96,30
respiration	F	47	95,92
congestion	D	23	95,83
destruction	P	181	95,77
réimplantation	P	21	95,45
insertion	P	228	95,40
déformation	D	61	95,31
sensation	F	37	94,87
calcification	D	36	94,74
décollement	D	34	94,44
lavage	P	17	94,44
perforation	D	51	94,44
comportement	F	50	94,34
dérivation	P	65	94,20
exploration	P	107	93,86
rétrécissement	D	57	93,44
épaississement	D	14	93,33
préparation	C	512	92,92
désarticulation	P	13	92,86
énucléation	P	13	92,86
diminution	F	263	91,64
transposition	P	61	91,04
avortement	D	265	90,44

TAB. B.12 – Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (1).

Têtes	**Catégories et freq. Snomed**	**Têtes**	**Cat+Fréquence**
abduction	F=1,P=1	condensation	F=1,D=1
aberration	D=1,F=3	condition	D=1,F=3,S=5
abrasion	D=112,P=6	conditionnement	P=1,F=3
absorption	F=19,P=7	conducteur	D=1,J=144
accélérateur	A=1,F=1	conduction	F=13,D=3
accélération	D=3,F=1	constipation	D=9,F=2
accouchement	F=3,D=15,P=41	constriction	D=3,P=1,T=1,F=4
adaptation	D=3,S=1,F=6	contraction	F=31,D=4
adhésion	T=1,F=1,S=1	contusion	F=2,D=162
affaissement	F=1,D=7	conversion	F=1,P=2
agglutination	F=1,D=2,P=1	crépitation	F=7,D=2
agitation	F=2,D=1	dédoublement	D=3,P=1
agrégation	D=1,F=4	défloration	F=1,P=1
alimentation	P=8,D=1	dégagement	F=4,P=2
allaitement	F=4,P=1	déglutition	F=3,P=1,D=1
allongement	D=2,P=25,F=1	délétion	D=3,F=4
altération	D=100,F=53	dénervation	P=6,F=1
amplification	P=2,F=1	déplacement	P=2,D=38
amputation	D=42,P=109	déplétion	D=8,P=1
analyseur	F=1,A=14	dépolarisation	F=2,D=1
apprentissage	P=1,F=3	dépression	F=3,T=4,D=41
aspiration	F=14,D=2,P=85	déraillement	A=1,F=1
assemblage	P=2,F=1	description	P=1,D=1
attrition	D=4,F=2	desquamation	F=3,D=1
augmentation	F=259,A=1,L=1,D=24,P=12	détachement	F=1,P=1
avulsion	P=26,D=4	détecteur	F=1,A=2
bifurcation	T=4,P=1	développement	D=4,F=12
blocage	D=6,P=7,F=5	déviation	F=14,D=17
broyeur	J=1,A=3	dilatation	D=42,P=75,F=16
captation	P=2,F=1	discrimination	F=4,S=2
changement	F=10,S=3,P=34,D=6	disjonction	D=4,F=1
chevauchement	F=1,D=2,P=4	disparition	D=1,F=1,S=1
circulation	T=5,F=4,P=2,D=1	dissociation	F=6,D=4
claudication	F=4,D=6	distension	F=1,D=7
clignement	D=1,F=1	distribution	D=2,P=1,F=1
coagulation	P=11,D=2,F=8	division	L=4,T=5,P=9
coloration	D=34,P=54,F=3	dysfonction	D=10,F=9
communication	D=7,P=1	dysfonctionnement	D=27,F=5
compression	D=24,P=1,F=2	écoulement	D=21,T=1,F=12
concentration	C=2,F=8,P=4	écrasement	D=53,P=12,A=1

TAB. B.13 – Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (2).

Têtes	**Catégories et freq. Snomed**	**Têtes**	**Cat+Fréquence**
éducation	S=1,P=4	hybridation	P=3,F=2
élargissement	F=1,P=11,D=1	hypoventilation	D=1,F=4
élévation	F=3,P=12	inadaptation	F=1,S=5
élimination	P=2,F=2	incision	P=540,D=3
élongation	P=1,D=3,F=2	inclusion	T=15,D=4
embolisation	P=11,D=2	incoordination	F=2,D=1
émission	D=4,A=1,F=6	induction	P=6,F=2
engagement	F=3,S=1,D=7,P=1	infiltration	D=19,F=3
engorgement	D=3,F=3	inhalation	P=2,D=5
engourdissement	F=2,D=1	inhibiteur	P=1,D=1,F=46,C=10
épuisement	D=7,F=3	inhibition	P=3,D=3,F=18
érection	F=4,D=4	insémination	P=5,F=1
éruption	D=30,A=1,F=6	installation	P=2,A=1
établissement	S=2,F=1,P=1	interaction	D=7,P=1,F=1
étirement	P=5,F=1	interruption	F=1,P=3,D=19
étranglement	T=1,D=4	invagination	D=7,P=1,T=5
évacuation	P=20,F=3	inversion	F=9,P=6,D=8
éversion	F=1,D=3	isolement	F=3,S=2,P=1
excavation	D=1,F=1,T=1	kératinisation	F=4,D=2
excitation	F=3,D=1	lactation	F=2,D=2
exclusion	S=1,F=1	lavement	C=1,P=7
excrétion	P=1,F=3	limitation	D=3,F=4
exfoliation	F=2,D=1,P=5	liquéfaction	P=2,D=1
expansion	D=1,T=3	localisation	P=7,F=1
exposition	D=10,P=2,F=2,A=47	macrophage	T=2,F=2,D=2
expression	P=4,F=3	malformation	D=8,P=1
extension	D=1,P=3,F=5	marqueur	F=7,J=1
facteur	C=1,J=5,F=187,S=4	mobilisation	P=7,F=5
fenestration	P=18,T=3	modification	F=2,P=3,D=26
fibrillation	F=2,D=4	moniteur	J=1,A=5
fixation	F=10,P=67	mouvement	P=1,D=1,F=47
flexion	F=11,D=1	nutrition	F=5,P=2
formation	F=23,T=5,D=30,P=41	oblitération	D=4,P=20
fraction	C=2,F=3	observation	F=2,P=6
fragment	D=1,A=1,F=3	occlusion	D=40,P=15,F=3
fragmentation	D=9,P=7	oesophage	T=10,D=4
frustration	F=1,S=1	orientation	T=3,F=6,P=1
fusion	A=1,F=4,P=60,D=4	ossification	D=8,F=11
glissement	D=3,A=1	passage	T=1,F=3,P=1
gonflement	D=3,F=1	pêcheur	L=2,J=3

TAB. B.14 – Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (3).

Têtes	**Catégories et freq. Snomed**	**Têtes**	**Cat+Fréquence**
pénétration	D=4,F=4	saignement	D=13,F=13
perception	D=2,F=12	saturation	F=2,P=1
perturbation	D=2,F=12,S=8	sécrétion	F=68,T=4,D=17
perversion	D=2,F=2	segment	D=2,F=2,T=47
pigment	F=2,C=1	sélection	F=1,P=1
pigmentation	D=35,P=1,F=3	séparation	F=1,D=3,P=8
pincement	P=2,D=1,F=1	séquestration	F=3,P=1,D=8
position	F=87,D=10	sevrage	P=1,D=1
prélèvement	P=43,D=14,T=82	simulation	S=1,F=3
présentation	D=23,F=25	stimulateur	F=2,A=3,C=1
pression	A=1,F=85	stimulation	F=13,P=5
privation	S=2,F=2,D=3	stockage	F=1,P=2
protection	P=2,A=2	subinvolution	F=3,D=3
raccourcissement	P=22,D=2,F=1	succion	F=1,P=1
radiation	A=4,T=1	suppression	D=2,P=8,F=8
rage	D=3,F=1	suspension	P=11,C=3
raideur	F=5,D=1	synchronisation	F=1,P=1
ralentissement	F=4,D=1	teneur	F=2,J=1
rayonnement	A=6,P=2	tension	A=1,F=25,D=1,P=1
réaction	F=125,S=1,P=17,D=93	torsion	D=21,F=2
récepteur	T=2,F=92	traceur	J=4,A=1
récession	P=2,D=3	transformation	F=2,P=5
reconstitution	P=3,F=1	transfusion	P=18,D=1
recouvrement	P=3,T=1	tremblement	A=1,F=28,D=10
régulateur	F=1,J=1,C=1	tuméfaction	D=13,F=2
régulation	P=2,F=12	variation	D=12,F=3,A=1
régurgitation	D=18,F=5	vascularisation	D=2,F=1
relâchement	P=7,D=7,F=7	ventilation	P=6,F=9
relation	T=10,F=12	vérification	A=2,P=13
remplissage	F=9,P=1	vision	F=24,D=2
répartition	D=2,F=3	vomissement	F=11,D=4
réplication	F=4,D=1		
résorption	D=13,F=5		
restauration	P=26,D=11		
rétablissement	P=2,F=1		
rétention	F=8,D=57		
rétraction	D=9,P=3,F=7		
rétroaction	P=24,F=1		
rotation	F=8,P=2,D=1		
roulement	P=1,F=4		

Annexe C

Les ressources linguistiques pour l'annotation sémantique des corpus

C.1 Les mots-outils

Cette liste contient un ensemble de mots outils du français collectés à partir d'Internet. Au total 146 mots, appartenant principalement aux catégories : déterminants, prépositions, pronoms, conjonctions.

TAB. C.1 – Liste des mots outils utilisés (1).

au	chacune	le leur	puis
à	chaque	le mien	qu'
à laquelle	chez	le nôtre	que
aucun	ci	le sien	quel
aucune	comme	le tien	quelle
aucunes	d'	le vôtre	quelles
aucuns	dans	lequel	quels
auquel	de	les	qui
autre	de laquelle	les leurs	quoi
autres	dernier	les miennes	se
aux	dernière	les miens	ses
auxquelles	dernières	les nôtres	si
auxquels	derniers	les siennes	soi
avec	des	les siens	son
beaucoup	desquelles	les tiennes	sur
c'	desquels	les tiens	surtout
ça	deux	les vôtres	te
ce	dont	lesquelles	toi

TAB. C.2 – Liste des mots outils utilisés (2).

ceci	du	lesquels	tous
cela	duquel	leur	tout
celle	durant	leurs	toutes
celle-ci	elle	lors	très
celle-là	elles	lui	tu
celles	en	mais	un
celles-ci	ensemble	me	une
celles-là	et	moi	vous
celui	eux	n'	
celui-ci	il	ne	
Celui-là	ils	non	
certain	je	nous	
certaine	l'	on	
certaines	la	ou	
certains	la leur	où	
ces	la mienne	par	
cet	la nôtre	parmi	
cette	la sienne	pas	
ceux	la tienne	pendant	
ceux-ci	la vôtre	peu	
ceux-là	laquelle	plusieurs	
chacun	le	pour	

C.2 Les déterminants complexes

TAB. C.3 – Liste des déterminants complexes utilisés (1).

certain nombre d'	différents niveaux d'	dans la
certains nombres de	différentes formes	dans les
certains nombres d'	épisode de	avec la
ces différents	épisodes de	avec le
ces différentes	ensemble des	avec les
ce différent	ensemble de	sur le
cas d'une	éventail de	sur la
cas d'un	éventail d'	sur les
cas de l'	équivalent de	en dehors du
cas de	grand nombre de	en dehors de
cas des	grand nombre des	en dehors des
cas d'	grand nombre d'	
une autre	grande variété d'	
catégorie de	grande variété de	
catégorie d'	grandes variétés de	
ce dernier	grandes variétés de	
cette dernière	grande proportion d'	
cas particuliers	grandes proportions de	
cas particulier de	grandes proportions des	
cas particulier d'	gamme des	
cas particuliers des	gamme de	
cas particuliers d'	gammes des	
ce présent	genre de	
le présent	genres de	
cohorte de	gammes de	
chacun de ces	grille de	
chacun de	ensemble de	
chacuns des	genre de	
d'importants	à l'occasion de	
d'importantes	aucun de ces	
de récente	aucun des	
de récentes	avec une	
de récents	à la	
de nombreuses	à le	
de nombreux	à les	
différent type de	à l'	
différent type d'	après le	
différents types de	après la	
différents types d'	après les	
différents niveaux de	dans le	

C.3 Les verbes de réalisation

TAB. C.5 – Liste des 147 verbes de réalisation.

élaborer	assouplir	confondre	établir	permettre
abandonner	assumer	conformer	étudier	perpétrer
abréger	assurer	connaître	évaluer	perpétuer
abroger	attester	consacrer	examiner	pratiquer
accélérer	attribuer	consigner	exiger	préparer
accepter	autoadministrer	contester	fabriquer	prescrire
accomplir	autoriser	continuer	façonner	prévoir
achever	booster	contourner	faire	procéder
acquérir	brouiller	contraster	former	proposer
activer	brusquer	contribuer	formuler	prospecter
actualiser	certifier	contrôler	imiter	provoquer
administrer	changer	corriger	impliquer	publier
adopter	choisir	créer	importer	réaliser
ajuster	circonscrire	débuter	imposer	recommander
ajuster	clôturer	déclencher	inclure	refaire
alléger	coadministrer	décliner	indiquer	reproduire
améliorer	cofinancer	déconseiller	instaurer	s'adonner
amplifier	combiner	découvrir	instituer	s'appuyer
analyser	commencer	dédramatiser	interdire	satisfaire
annoncer	comparer	définir	interpréter	simuler
annuler	compléter	démarrer	interrompre	suivre
appliquer	composer	désapprouver	inventer	utiliser
apposer	compresser	disposer	juger	entamer
apprécier	comprimer	donner	justifier	
apprendre	compromettre	effectuer	mener	
apprêter	compter	encourager	mettre	
approuver	concevoir	entreprendre	monter	
arrêter	conclure	envisager	nécessiter	
assigner	condamner	essayer	nommer	
assister	conduire	exécuter	omettre	
associer	considérer	exercer	ordonner	

TAB. C.4 – Liste des déterminants complexes utilisés (2).

la majorité des	intérêt d'	les meilleure
la majorité de	l'un des	les mêmes
majorité des	l'un de	les nombreux
majorité de	leur principal	les nouveaux
les autres	majorité de	les nouvelles
l'autre	majorité d'	les prochains
le maximum de	majorité des	les prochaines
les maximum des	le majorité de	les rares
maximum de	le majorité d'	leur prochain
maximum des	le majorité des	leurs prochains
même	la majorité de	leurs prochaines
meilleur	la majorité d'	leurs propres
million de	la majorité des	aux mêmes
millier de	la même	au même
million d'	le même	aux présents
millier d'	le nouveau	au présent
minimum de	la nouvelle	au seul
minorité des	le premier	aux seuls
minorité de	la première	aux seules
minorité d'	la présente	aux rares
ml de	la principale	au rare
ml d'	le principal	gramme de
modalité de	les principales	pourcentage des
moitié de ces	les principaux	pourcentage de
moitié de la	le seul	le pourcentage des
moindre	la seule	le pourcentage de
moitié de	le même	un pourcentage de
moitié des	la même	grande majorité des
moitié d'	le nouvel	grande majorité de
moitié de la	la nouvelle	d'autres
moitié de le	le meilleur	d'autre
moitié de les	la meilleure	en extrême
de nombreux	le plus	de multiple
nombre de	la plus	de multiples
nombre des	les autres	de fort
le nombre de	les derniers	certains de nos
le nombre des	les différentes	certaines de ces
intérêt de	les différents	certain degré d'
intérêt des	les autre	certain degré de
intérêt du	les meilleures	certain nombre de

Annexe D

Les ressources conçues pour la simplification de textes

Cette annexe présente les différentes ressources que nous avons conçues dans le cadre de la création et/ou de l'évaluation de la ressource de simplification de textes proposée comme résultat de ce travail de thèse.

D.1 Les 50 phrases (originales) utilisées dans le cadre de l'évaluation de la ressource.

TAB. D.1 – Les 50 phrases (originales) évaluées (1).

125	C' est pourquoi il faut absolument que les équipes du SAMU soient à même de dépister précocement les cas d' AVC et connaissent les pré-requis de la thrombolyse, afin d' activer au plus tôt la filière d' urgence AVC.
60	Cliniques externes où l'on adminsitre des médicaments dangereux par voie parentérale, infirmière à domicile.
2	L'épiglottite aiguë se traduit par une dyspnée laryngée fébrile à 385 à 39, une dysphagie douloureuse entraînant une hypersialorrhée, l'enfant ne supporte que la position assise.
3	Les effets indésirables se sont produits dans les 24 heures de l'administration de l'IgIV dans le cas d'AIT, dans les 11 jours dans les cas d'embolie pulmonaire et dans les 2 semaines dans les cas de thrombose.
4	L'état de crise du sujet âgé se présente habituellement comme une décompensation fonctionnelle : confusion ou décompensation cérébrale aiguë, dépression ou décompensation thymique, chute ou décompensation posturale aiguë, décompensation nutritionnelle", etc...
7	Malheureusement, seulement 40 % des femmes belges pratiquent ce dépistage.

TAB. D.2 – Les 50 phrases (originales) évaluées (2).

10	Si le nadolol est poursuivi jusqu' à l'accouchement, en informer l'équipe de la maternité pour lui permettre d'adapter la surveillance du nouveau-né.
116	Son traitement actuel associe Coversyl (périndopril), Fludex (indapamid et Isoptine (vérapamil).
12	L'antibiotique doit être administré en concentration suffisante au moment de la colonisation par les bactéries potentiellement pathogènes.
13	La salpingite s'accompagne d'une fièvre plus importante, d'une vive douleur à la mobilisation du col utérin, alors que la douleur spontanée est plus diffuse.
14	Ces différences moléculaires entre cellules cancéreuses se traduisent par des pronostics et des réponses aux traitements très variables, même lorsque les cancers sont à priori classés dans une seule et même catégorie.
85	L'appréciation de la symptomatologie fonctionnelle (claudication intermittente, douleurs de décubitus, troubles trophiques, ulcères et gangrènes), la palpation et l' auscultation des artères assurent, dans la majorité des cas, le diagnostic positif de l' artérite, renseignent sur la sévérité de l' ischémie, sur la topographie des lésions selon les sites d' audition des souffles et le niveau d' abolition des pouls, et peuvent dépister une lésion anévrismale (aor-tique, iliaque ou poplitée).
15	Les adénocarcinomes se développent le plus souvent à partir des cellules des canaux galactophores ou carcinome canalaire (40 % à 75 % des cas).
16	Le système nerveux contrôle les muscles, bien que certains muscles peuvent fonctionner de façon autonome..
17	Le dépistage, tout comme le traitement précoce des IST, relève de la prévention primaire.
19	Les virus sauvages naturels ont développé au cours de millions d'années d'évolution des systèmes très efficaces de transfert de gènes, puisque le cycle viral implique le transfert du génome viral dans le génome de la cellule hôte.
21	Une absence d'augmentation du chiffre de plaquettes n'a été observée que chez 12 % des malades mais il faut souligner qu' un arrêt des signes hémorragiques a été obtenu chez tous les patients témoignant de la possibilité d'une réponse clinique, même en l'absence de correction de la thrombopénie.
22	Ce sont, avant tout, des conjonctivites et des épisclérites, mais, outre des paralysies oculo-motrices dues à une neuropathie périphérique, on peut observer une vascularite rétinienne.
131	l'exonération du ticket modérateur peut être donnée de façon ponctuelle conformément aux dispositifs réglementaires ou de manière continue lorsque le patient relève d'une affection de longue durée.

TAB. D.3 – Les 50 phrases (originales) évaluées (3).

28	L'autopsie a révélé une coronaropathie et une sténose aortique valvulaire..
29	D'autres analyses sanguines indiquent une leucocytose légère et des marqueurs cardiaques normaux, alors qu' une deuxième radiographie des poumons révèle un petit épanchement pleural du côté gauche.
34	Le cerveau traduit ensuite ces stimulis en sensations.
35	Les rhinites intermittentes modérées-sévères et persistantes légères sont traitées par anti-histaminique oral ou local et / ou décongestionnant ou corticoïdes locaux.
36	La protection rénale impose un contrôle strict de la pression artérielle et une lutte agressive contre l'ensemble des facteurs de risque vasculaire.
37	La restauration des lactobacilles au niveau vaginal peut aider à éviter la rechute, pour cela on envisage une cure de Geiofil pendant une semaine à la fin du traitement.
39	En fonction du site de la thrombose veineuse, des tests de laboratoire complémentaires devraient être réalisés pour exclure une dysmyélopoïèse.
40	L'enfant insuffisament transfusé va présenter un retard de croissance et une pâleur cutanéo-muqueuse associée à un degré variable d'ictère et à une pigmentation brune de existe un syndrome hypermétabolique avec réduction de la masse musculaire, du tissu adipeux.
151	Le cancer de la paroi se développe à bas bruit.
153	L'utilisation de l'écarteur de Parks ouvert de 4 cm abaisse significativement la pression de repos à 6 et 12 semaines, par rapport à la même intervention sans cet écarteur (– 23 % versus- 8 %).
187	L'HTA de consultation pourrait cependant précéder l'HTA permanente, les patients concernés relevant alors d'une surveillance renforcée, et l'HTA masquée pourrait être un extrême de distribution lié à la variabilité tensionnelle qui est positivement corrélée au risque cardiovasculaire.
191	Il faut entreprendre la médication contre le TDAH à la recommandation de la personne qui diagnostique et suit le patient ayant le TDAH.
243	Les patients présentant une décompensation cardiaque terminale, avec NYHA III-IV semblent tirer moins d'avantages d'une CRT-D.
42	La forme extrême de l'apragmatisme réalise la catatonie telle qu' on peut l'observer au cours de la schizophrénie et qui consiste en une suspension totale de l'activité motrice., repli sur soi, etc.

TAB. D.4 – Les 50 phrases (originales) évaluées (4).

45	Normalement, il existe un équilibre entre d'une part, les facteurs activant la coagulation et, d'autre part, les inhibiteurs physiologiques de la coagulation et les activateurs de la fibrinolyse..
49	Des dispositifs électroniques aidant les professionnels de la santé à mesurer l'observance du traitement ont été développés.
88	Lorsqu' une coronarographie est réalisée chez les patients avec IMS, elle met en évidence des sténoses coronaires significatives dans un tiers à deux tiers des cas.
89	On recommande une prise en charge prudente comprenant, au besoin, une assistance circulatoire consistant en des transfusions et en une diurèse.
50	Des stratégies de dépistage afin de proposer un diagnostic anté-natal aux couples à risque sont développées dans certaines régions du monde.
51	Si le malade développe une infection pulmonaire, il faut la traiter rapidement.
52	Une hépatite auto-immune de type I s'accompagne d'anticorps anti-muscle lisse, une hépatite auto-immune de type II, d'anticorps anti-LKM-1.
53	Cet examen indolore, sûr et relativement bon marché, est souvent pratiqué depuis l'introduction de l'échographie transvaginale.
54	Même si les vérifications indépendantes des préparations finales étaient requises dans tous les sites soumis à l'étude terrain de CAPCA, dans quatre sites sur six, aucune vérification indépendante des médicaments reconstitués n'a été observée.
128	Lors de l'utilisation de DEFINITY suivant sa commercialisation, quatre patients ont subi un arrêt cardiaque d'issue fatale soit en cours d'administration soit dans les 30 minutes suivant l'administration.
55	L'échelle de probabilité des interactions médicamenteuses indique une relation probable entre la chute soudaine des concentrations plasmatiques d'acide valproïque et l'introduction du méropénem chez cette patiente.
56	Le clopidogrel est indiqué dans la prévention secondaire des complications thrombotiques après infarctus du myocarde, AVC ou artériopathie périphérique.
91	Le carcinome médullaire (1 % des cancers du sein infiltrants) est observé chez la femme âgée de moins de 50 ans.
114	Quel que soit le support nutritionnel adopté, la tolérance et l'efficacité doivent être régulièrement évaluées par la surveillance régulière du poids, de la tension artérielle, de l'état d'hydratation, du transit digestif, de la position de la sonde gastrique.
43	L'utilisation de dabigatran n'est pas recommandée chez les patients atteints d'une insuffisance hépatique modérée ou grave.
44	Les paralysies du sommeil peuvent débuter à n'importe quel âge, mais sont particulièrement observées chez l'adolescent et l'adulte d'âge moyen.

D.2 Les 50 phrases après simplification

TAB. D.5 – Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (1).

125	Il faut que les équipes du SAMU soient capables de découvrir précocement les cas d' AVC.
60	Dans cette cliniques externes, l'on donne aux patients des médicaments dangereux par voie parentérale.
2	L'épiglottite aiguë est une maladie qui se manifeste par une dysphagie douloureuse entraînant une hypersialorrhée.
3	Les effets indésirables se sont manifestés dans les 24 heures de l'administration de l'IgIV.
4	L'état de crise du patient se manifeste habituellement comme une décompensation fonctionnelle.
7	Malheureusement, seulement 40 % des femmes belges font le dépistage.
8	Les changements squelettiques sont liés à des modifications radiologiques.
10	Si la patiente continue de prendre le nadolol jusqu' à l'accouchement, il faut en informer l'équipe de la maternité.
116	Son traitement actuel combine le Coversyl et le Fludex.
12	On doit donner l'antibiotique au patient en quantité suffisante.
13	La salpingite est suivie de la fièvre et d'une vive douleur.
14	Ces différences moléculaires entre cellules cancéreuses se manifestent par des pronostics et des réponses aux traitements très variables.
85	L'Electrocardiogramme décèle les troubles de conduction.
15	Les adénocarcinomes commencent le plus souvent à partir des cellules des canaux galactophores ou carcinome canalaire.

TAB. D.6 – Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (2).

16	Le système nerveux commande les muscles, bien que certains muscles peuvent fonctionner de façon autonome.
17	Le traitement précoce des IST dépend de la prévention primaire.
19	Les virus sauvages ont formé des systèmes très efficaces de transfert de gènes.
21	Une diminution du chiffre de plaquettes a été constatée chez 12 % des malades.
22	On peut constater une vascularite rétinienne chez le patient.
131	Le patient souffre d'une affection de longue durée.
28	L'autopsie a montré une coronaropathie et une sténose aortique valvulaire.
29	Les analyses sanguines signalent une leucocytose légère.
34	Le cerveau transforme les stimulis en sensations.
35	L'anti-histaminique oral soigne les rhinites intermittentes modérées-sévères.
36	La protection rénale passe par un contrôle strict de la pression artérielle.
37	On prévoit une cure de Geiofil pendant une semaine à la fin du traitement.
39	Des tests de laboratoire complémentaires devraient être faits pour exclure une dysmyélopoïèse.
40	L'enfant insuffisament transfusé va manifester un retard de croissance et une pâleur cutanéo-muqueuse.
151	Le cancer de la paroi évolue à bas bruit.
153	L'utilisation de l'écarteur de Parks ouvert de 4 cm baisse significativement la pression de repos à 6 et 12 semaines.
187	Les patients concernés bénéficient alors d'une surveillance renforcée.
191	Il faut entreprendre la médication contre le TDAH à la recommandation de la personne qui découvre et suit le patient ayant le TDAH.
243	Les patients souffrant d'une décompensation cardiaque terminale semblent tirer moins d'avantages d'une CRT-D.
42	La forme extrême de l'apragmatisme provoque la catatonie.

TAB. D.7 – Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (3).

43	L'utilisation de dabigatran n'est pas conseillée chez les patients atteints d'une insuffisance hépatique.
44	Les paralysies du sommeil sont particulièrement rencontrées chez l'adolescent.
45	Il existe un équilibre entre les facteurs accélerant la coagulation et les inhibiteurs physiologiques de la coagulation.
49	Des dispositifs électroniques aidant les professionnels de la santé ont été crées.
88	Une coronarographie est faite chez les patients avec IMS.
89	On prescrit une prise en charge prudente comprenant, au besoin, une assistance circulatoire consistant en des transfusions et en une diurèse.
50	Des stratégies de dépistage sont créées dans certaines régions du monde.
51	Si le malade fait une infection pulmonaire, il faut la traiter rapidement.
52	Une hépatite auto-immune est suivie d'anticorps anti-muscle lisse.
53	Cet examen est fait depuis l'introduction de l'échographie transvaginale.
54	Aucune vérification indépendante des médicaments reconstitués n'a été faite.
128	Quatre patients ont fait un arrêt cardiaque d'issue fatale.
55	L'échelle de probabilité des interactions médicamenteuses signale une relation probable entre la chute soudaine des concentrations et l'introduction du méropénem chez cette patiente.
56	Le clopidogrel est conseillé dans la prévention secondaire des complications thrombotiques.
91	Le carcinome médullaire est diagnostiqué chez la femme âgée de moins de 50 ans.
114	La tolérance et l'efficacité doivent être régulièrement testées par la surveillance régulière du poids.

D.3 Les 243 PSS sélectionnés par les experts pour l'alignement

Dans cette section et dans la section suivante, certains noms de catégories ont été abrégés pour un meilleur affichage des données. ANA :anatomie (partie du corps), MAL : maladie ; ORG : organisme vivants, PRO (procédure médicale) ; STAS : statut social (c.-à-d. patient, de façon générale). La catégorie METIER renvoie de façon générale aux professionnels de la santé ; PCHIMIQUE : produit chimique, FONCTION : fonction de l'organisme ; AGENT : agent physique, artéfact, instruments, etc.).

TAB. D.8 – Liste des 243 PSS sélectionnés pour validation par les experts (1).

ID	PSS	ID	PSS
1	ANA transmet FONCTION à ANA	40	STAS présente FONCTION
2	MAL se traduit par MAL	41	FONCTION réalise MAL
3	MAL se produit	42	MAL réalise MAL
4	FONCTION se présente comme MAL	43	PROC est recommandée chez/pour STAS
5	PROC est associée à MAL	44	MAL est observée chez STAS
6	FONCTION est inhibée par PCHIMIQUE	45	QQCHSE active FONCTION
7	STAS pratique PROC	46	FONCTION est analysée
8	FONCTION est associée à FONCTION	47	MAL est évaluée
9	ORG associer à MAL	48	PROC évalue FONCTION
10	PCHIMIQUE est poursuivi	49	AGENT est développé
11	PROC est indiquée	50	PROC est développée
12	PCHIMIQUE est administré	51	STAS développe MAL
13	MAL s'accompagne de MAL	52	MAL s'accompagne de FONCTION
14	FONCTION se traduit par FONCTION	53	PROC est pratiquée
15	MAL se développe à partir de ANA	54	PROC est observé
16	PROC montre MAL	55	PROC indique FONCTION
17	PROC relève de PROC	56	PCHIMIQUE est indiqué dans PROC
18	FONCTION est exprimée dans ANA	57	PROC est indiquée chez STAS
19	ORG développe ANA	58	PCHIMIQUE provoque MAL
20	FONCTION est évaluée	59	METIER administre PCHIMIQUE dans AGENT
21	FONCTION est observée chez STAS	60	METIER administre PCHIMIQUE
22	METIER observe MAL	61	STAS montrer MAL
23	PROC impliquer PROC	62	PCHIMIQUE produit FONCTION
24	MAL relever de PROC	63	MAL est signalée
25	MAL se manifester comme MAL	64	MAL est signalée chez STAS
26	MAL se manifester chez STAS	65	ANA présente MAL
27	PROC nécessiter PROC	66	MAL présente FONCTION
28	PROC révèle MAL	67	PCHIMIQUE présente FONCTION
29	PROC indique MAL	68	FONCTION présente FONCTION
30	PROC est proposée	69	ORG présente FONCTION

TAB. D.9 – Liste des 243 PSS sélectionnés pour validation par les experts (2).

31	MAL est découverte chez STAS
32	METIER évoque PROC
33	FONCTION est recherchée
34	ANA traduit FONCTION en FONCTION
35	MAL est traitée par PCHIMIQUE
36	PROC impose PROC
37	METIER envisage PROC
38	METIER suit STAS
39	PROC est réalisée
79	PROC est proposée chez STAS
80	FONCTION est augmentée
81	ANA est augmentée
82	PROC est préconisée
83	FONCTION est induite
84	MAL induit MAL
85	PROC dépiste MAL
86	MAL induite par PCHIMIQUE
87	PROC est réalisée au moyen de/sous AGENT
88	PROC est réalisée chez STAS
89	METIER recommande PROC
90	METIER élimine MAL
91	MAL est observée chez STAS
92	MAL accompagne MAL
93	STAS se développe
94	FONCTION est associée à ANA
95	MAL est associée à FONCTION
96	STAS est dépisté
97	STAS exposé à PROC
98	MAL est associée à MAL
99	STAS inhale PCHIMIQUE
100	PCHIMIQUE est poursuivi
101	PROC est activée
102	FONCTION est activée
103	MAL évoque MAL
104	MAL est évoquée chez STAS
105	FONCTION fait évoquer MAL
106	MAL évoque FONCTION
107	METIER évoque MAL
108	FONCTION est synthétisée dans ANA
109	MAL est détectée par METIER
110	FONCTION relève de MAL
111	FONCTION est évaluée
112	METIER évalue FONCTION
113	METIER évalue STAS
70	PROC est indiquée pour/dans/en PROC
71	MAL indique QQCHSE
72	MAL se révèle en cours de FONCTION
73	PCHIMIQUE est envisagé
74	PROC est envisagée
75	AGENT est envisagé
76	PROC proposée dans/comme PROC
77	FONCTION est associée à MAL
78	MAL associe MAL à FONCTION
120	FONCTION est abaissée (avant PROC)
121	PROC s'accompagne de PROC
122	FONCTION s'accompagne de FONCTION
123	FONCTION se poursuit
124	PROC expose à MAL
125	METIER dépiste MAL
126	FONCTION est observée chez STAS
127	METIER observe MAL dans MAL
128	STAS subit MAL
129	STAS subit PROC
130	PCHIMIQUE s'associe à FONCTION
131	STAS relève de MAL
132	PROC relève de PROC
133	METIER relève MAL
134	STAS relève de PROC
135	FONCTION relève de PROC
136	MAL est relevée chez ORG
137	PCHIMIQUE est éliminé par ANA
138	MAL est relevée (chez STAS)
139	PROC s'observe chez STAS
140	FONCTION induite par PROC
141	ORG induit MAL
142	AGENT induit PROC
143	METIER observe ANA
144	MAL s'observe au cours de MAL
145	PROC est analysée
146	AGENT est activé
147	PCHIMIQUE active FONCTION
148	MAL se manifeste
149	ORG développe FONCTION
150	AGENT est développé (à partir de ORG)
151	MAL se développe
152	METIER développe PROC
153	PROC abaisse FONCTION
154	STAS présente FONCTION

TAB. D.10 – Liste des 243 PSS sélectionnés pour validation par les experts (3).

115	ORG est éliminé par FONCTION
116	PROC associe PCHIMIQUES
161	MAL évoque PROC
162	PROC évoque MAL
163	PROC est envisagée
164	PROC est envisagée chez STAS
165	FONCTION subit FONCTION
166	FONCTION est réalisée par PCHIMIQUE
167	METIER relève STAS
168	PCHIMIQUE est développé
169	FONCTION réalise FONCTION/MAL
170	STAS analyse MAL
171	AGENT analyse FONCTION
172	STAS est évalué
173	STAS est détecté par PROC
174	ORG produit FONCTION
175	MAL affecte STAS
176	MAL est traitée chez STAS
177	MAL est traitée par PROC
178	METIER signale MAL chez STAS
179	MAL est signalée chez STAS
180	MAL se présente
181	MAL se présente sur PROC
182	PROC est recommandée chez STAS
183	MAL relève de FONCTION
184	PROC associe PROCEDURES
185	FONCTION est associée à FONCTION
186	FONCTION s'observe
187	STAS relève de PROC
188	ANA est activée
189	PCHIMIQUE active FONCTION
190	STAS développe ANA
191	METIER diagnostique STAS
192	MAL évolue en MAL
193	PROC montre MAL
194	ANA contrôle FONCTION
195	FONCTION contrôle FONCTION
196	PROC survient chez STAS
197	FONCTION est synthétisée par ANA
198	MAL colonise ANA
199	FONCTION est synthétisée par AGENT
200	ANA synthétise FONCTION
201	PROC est relevée
117	PCHIMIQUE est associé à PROC
118	PROC est évaluée
119	MAL relève de FONCTION
156	FONCTION est réalisée par ANA
157	METIER relève AGENT
202	MAL affecte FONCTION
203	QQCHSE affecte PCHIMIQUE
204	PROC est évoquée chez STAS
205	STAS inhale PCHIMIQUE
206	MAL est associée à PROC
207	AGENT affecte ANA
208	STAS traité à l'aide de PCHIMIQUE
209	PROC se traduit par PROC
210	PROC s'accompagne de FONCTION
211	MAL traduit MAL
212	MAL traduit FONCTION
213	METIER pratique PROC
214	PROC est associée à PROC
215	PROC est pratiquée (par ANA)
216	FONCTION s'accompagne de AGENT
217	FONCTION isolée
218	ANA transmet ANA
219	FONCTION est transmise
220	STAS présente ANA
221	ANA sécrète FONCTION
222	FONCTION exprime MAL
223	MAL s'exprime par FONCTION
224	FONCTION est altérée
225	FONCTION altère FONCTION
226	PROC dépiste FONCTION
227	FONCTION est observée
228	PROC induire MAL
229	METIER élimine ANA
230	ANA élimine ANA
231	MAL se dissémine
232	ANA se dissémine dans ANA
233	MAL associe MALADIES
234	PCHIMIQUE inhibe FONCTION
235	ORG est excrété
236	ANA est éliminée
237	FONCTION relève de FONCTION
238	MAL est réalisée
239	PCHIMIQUE est évalué
240	ANA élimine PCHIMIQUE
241	MAL est éliminée par FONCTION
242	PROC relève de STAS
158	ORG affecte ANA
159	PROC s'observe chez STAS
160	PROC est évoquée

D.4 Les 230 PSS (entrées de la ressource) alignés avec leurs équivalents

TAB. D.11 – Liste des PSS alignés (1).

ID	PSS	PSS équivalents
1	ANA transmet FONCTION à ANA	ANA communique FONCTION à ANA
2	MAL se traduit par MAL	MAL se manifeste par MAL
3	MAL se produit	MAL se manifeste chez STAS
4	FONCTION se présente comme MAL	FONCTION se manifeste comme MAL
5	PROC est associée à MAL	PROC cause MAL
6	FONCTION est inhibée par PCHIMIQUE	FONCTION est bloquée par PCHIMIQUE
7	STAS pratique PROC	STAS subit PROC
8	FONCTION est associée à FONCTION	FONCTION est liée à FONCTION
9	ORGVIVANT associer à MAL	ORGVIVANT est lié MAL
10	PCHIMIQUE est poursuivi	STAS (continue de) prendre PCHIMIQUE
11	PROC est indiquée	(METIER) indique PROC
12	PCHIMIQUE est administré	(METIER) donne PCHIMIQUE (à STAS)
13	MAL s'accompagne de MAL	MAL entraîne MAL
14	FONCTION se traduit par FONCTION	FONCTION se manifeste par FONCTION
15	MAL se développe à partir de ANA	MAL naître dans ANA
16	ANA contrôle ANA	ANA commande ANA
17	PROC relève de PROC	PROC fait partie de PROC
18	FONCTION est exprimée dans ANA	FONCTION est présente dans ANA
19	ORGVIVANT développe ANA	ORGVIVANT forme ANA
20	FONCTION est évaluée	FONCTION est contrôlée
21	FONCTION est observée chez STAS	(METIER) constate FONCTION chez STAS
22	METIER observe MAL	METIER remarque MAL (chez STAS)
23	PROC impliquer PROC	PROC nécessite PROC
24	MAL relever de PROC	MAL bénéficie de PROC
27	PROC nécessiter PROC	PROC exige PROC
28	PROC révèle MAL	PROC montre MAL
29	PROC indique MAL	PROC signale MAL
30	PROC est proposée	(METIER) entreprend PROC
31	MAL est découverte chez STAS	(METIER) découvre MAL chez STAS
32	METIER évoque PROC	METIER mentionner PROC
33	FONCTION est recherchée	(STAS) souhaite FONCTION
34	ANA traduit FONCTION en FONCTION	ANA transforme FONCTION en FONCTION
35	MAL est traitée par PCHIMIQUE	PCHIMIQUE soigne MAL
36	PROC impose PROC	PROC passe par PROC
37	METIER envisage PROC	METIER prévoit PROC
39	PROC est réalisée	(METIER) fait PROC
40	STAS présente FONCTION	STAS avoir FONCTION
42	MAL réalise MAL	MAL provoque MAL

TAB. D.12 – Liste des PSS alignés (2).

43	PROC est recommandée chez/pour STAS	(METIER) conseille PROC à STAS
44	MAL est observée chez STAS	(METIER) constate MAL chez STAS
45	QQCHSE active FONCTION	QQCHSE accélère FONCTION
46	FONCTION est analysée	FONCTION est considérée
47	MAL est évaluée	MAL est testée
48	PROC évalue FONCTION	PROC mesure FONCTION
49	AGENT est développé	(METIER) crée AGENT
50	PROC est développée	(METIER) créé PROC
51	STAS développe MAL	STAS fait MAL
52	MAL s'accompagne de FONCTION	MAL entraîne FONCTION
53	PROC est pratiquée	(METIER) fait PROC
54	PROC est observé	(METIER) fait PROC
55	PROC indique FONCTION	PROC signale FONCTION
56	PCHIMIQUE est indiqué dans PROC	(METIER) conseille PCHIMIQUE dans PROC
57	PROC est indiquée chez STAS	(METIER) conseille PROC chez STAS
58	PCHIMIQUE provoquer MAL	PCHIMIQUE cause MAL
59	METIER administre PCHIMIQUE dans AGENT	METIER injecte PCHIMIQUE dans AGENT
60	METIER administre PCHIMIQUE	METIER donne PCHIMIQUE (à STAS)
62	PCHIMIQUE produit FONCTION	PCHIMIQUE provoquer FONCTION
63	MAL est signalée	(METIER) annonce MAL
64	MAL est signalée chez STAS	(METIER) détecte MAL chez STAS
66	MAL présente FONCTION	MAL entraîne FONCTION
68	FONCTION présente FONCTION	FONCTION manifeste FONCTION
69	ORG présente FONCTION	ORG manifeste FONCTION
70	PROC est indiquée pour/dans/en PROC	(METIER) conseille PROC pour/dans/en PROC
71	MAL indique QQCHSE	MAL montre QQCHSE
72	MAL se révèle en cours de FONCTION	MAL se manifeste en cours de FONCTION
74	PROC est envisagée	(METIER) prévoit PROC
75	AGENT est envisagé	(METIER) pense à AGENT
77	FONCTION est associée à MAL	FONCTION liée à MAL
78	MAL associe MALADIES à FONCTION	MAL combine MALADIES et FONCTION
80	FONCTION est augmentée	FONCTION est accélérée
81	ANA est augmentée	ANA est gonflée
82	PROC est préconisée	(METIER) conseille PROC
83	FONCTION est induite	FONCTION est provoquée
84	MAL induit MAL	MAL entraîne MAL
85	PROC dépiste MAL	PROC permet de détecter MAL
86	MAL induite par PCHIMIQUE	MAL causée par PCHIMIQUE
87	PROC est réalisée au moyen de/sous AGENT	(METIER) fait PROC au moyen de/sous AGENT
88	PROC est réalisée chez STAS	(METIER) fait PROC sur STAS
89	METIER recommande PROC	METIER conseille PROC
90	METIER élimine MAL	METIER détruit MAL
91	MAL est observée chez STAS	(METIER) diagnostique MAL chez STAS
92	MAL accompagne MAL	MAL (peut être) associée à MAL

TAB. D.13 – Liste des PSS alignés (3).

93	STAS se développe	STAS grandit
94	FONCTION est associée à ANA	FONCTION se manifeste au niveau de ANA
95	MAL est associée à FONCTION	MAL touche FONCTION
96	STAS est dépisté	(METIER) identifier STAS
97	STAS exposé à PROC	STAS étant sous PROC
98	MAL est associée à MAL	MAL liée à MAL
99	STAS inhale PCHIMIQUE	STAS aspire PCHIMIQUE
100	PCHIMIQUE est poursuivi	(STAS) continue de prendre PCHIMIQUE
101	PROC est activée	PROC est lancée
102	FONCTION est activée	FONCTION est déclenchée
103	MAL évoque MAL	MAL fait penser à MAL
104	MAL est évoquée chez STAS	(METIER) diagnostique MAL chez STAS
105	FONCTION fait évoquer MAL	FONCTION fait penser à MAL
106	MAL évoque FONCTION	MAL provoque FONCTION
107	METIER évoque MAL	METIER soupçonne MAL
108	FONCTION est synthétisée dans ANA	FONCTION est produite dans ANA
109	MAL est détectée par METIER	METIER découvre MAL
110	FONCTION relève de MAL	FONCTION être liée à MAL
111	FONCTION est évaluée	METIER mesure FONCTION
112	METIER évalue FONCTION	METIER contrôle FONCTION
113	METIER évalue STAS	METIER examine STAS
114	FONCTION est évaluée par PROC	PROC permet (à METIER) de tester FONCTION
115	ORG est éliminé par FONCTION	ORG est tué par FONCTION
116	PROC associe PCHIMIQUES	PROC combine PCHIMIQUES
117	PCHIMIQUE est associé à PROC	PCHIMIQUE est responsable de FONCTION
118	PROC est évaluée	(METIER) contrôle PROC
119	MAL relève de FONCTION	MAL être liée à FONCTION
120	FONCTION est abaissée (avant PROC)	(METIER) baisse FONCTION (avant PROC)
121	PROC s'accompagne de PROC	PROC est suivie de PROC
122	FONCTION s'accompagne de FONCTION	FONCTION entraîne FONCTION
123	FONCTION se poursuit	FONCTION suit son cours
124	PROC expose à MAL	PROC donne accès à MAL
125	METIER dépiste MAL	METIER découvre MAL
126	FONCTION est observée chez STAS	(METIER) enregistre FONCTION chez STAS
127	METIER observe MAL dans MAL	METIER rencontre MAL dans MAL
128	STAS subit MAL	STAS fait MAL
129	STAS subit PROC	(METIER) fait PROC à STAS
130	PCHIMIQUE s'associe à FONCTION	PCHIMIQUE est lié à FONCTION
131	STAS relève de MAL	STAS souffre de MAL
132	PROC relève de PROC	PROC fait partie de PROC
133	METIER relève MAL	METIER note MAL
134	STAS relève de PROC	STAS a besoin de PROC
135	FONCTION relève de PROC	FONCTION provient de PROC
136	MAL est relevée chez ORG	MAL est enregistrée chez ORG

TAB. D.14 – Liste des PSS alignés (4).

137	PCHIMIQUE est éliminé par ANA	PCHIMIQUE est évacué (du corps) par ANA
138	MAL est relevée (chez STAS)	(METIER) constate MAL (chez STAS)
139	PROC s'observe chez STAS	PROC concerne STAS
140	FONCTION induite par PROC	FONCTION causée par PROC
141	ORG induit MAL	ORG provoque MAL
142	AGENT induit PROC	AGENT entraîne PROC
143	METIER observe ANA	METIER constate (association de) ANA
144	MAL s'observe au cours de MAL	MAL se manifeste au cours de MAL
145	PROC est analysée	(METIER) vérifie PROC
146	AGENT est activé	(METIER) met en marche AGENT
147	PCHIMIQUE active FONCTION	PCHIMIQUE rend actif FONCTION
148	MAL se manifeste	MAL apparaît
149	ORG développer FONCTION	ORG manifeste FONCTION
150	AGENT est développé (à partir de ORG)	(METIER) créé AGENT (à partir de ORG)
151	MAL se développe	MAL évolue
152	METIER développe PROC	METIER fait PROC
153	PROC abaisse FONCTION	PROC fait baisser FONCTION
154	STAS présente FONCTION	STAS manifester FONCTION
155	ORG développe ANA	ORG créé ANA
156	FONCTION est réalisée par ANA	FONCTION est assurée par ANA
157	METIER relève AGENT	METIER constater AGENT
158	ORG affecte ANA	ORG compromet ANA
159	PROC s'observe chez STAS	PROC concerne STAS
160	PROC est évoquée	(METIER) suggère/pense à PROC
161	MAL évoque PROC	MAL fait penser à PROC
162	PROC évoque MAL	PROC fait soupçonner MAL
163	PROC est envisagée	(METIER) considère PROC
164	PROC est envisagée chez STAS	(METIER) entreprend PROC chez STAS
165	FONCTION subit FONCTION	FONCTION sont modifiées
166	FONCTION est réalisée par PCHIMIQUE	FONCTION est faite par PCHIMIQUE
167	METIER relève STAS	METIER note STAS
168	PCHIMIQUE est développé	(METIER) créé PCHIMIQUE
169	FONCTION réalise FONCTION/MAL	FONCTION provoque FONCTION/MAL
170	STAS analyse MAL	STAS reconnaît MAL
171	AGENT analyse FONCTION	AGENT examine FONCTION
172	STAS est évalué	(METIER) examine STAS
173	STAS est détecté par PROC	STAS est repéré par PROC
174	ORG produit FONCTION	ORG présente FONCTION
175	MAL affecte STAS	MAL touche STAS
176	MAL est traitée chez STAS	(METIER) soigne MAL chez STAS
177	MAL est traitée par PROC	(METIER) soigne MAL par PROC
178	METIER signale MAL chez STAS	METIER découvre MAL chez STAS
179	MAL est signalée chez STAS	(METIER) détecte MAL chez STAS

TAB. D.15 – Liste des PSS alignés (5).

180	MAL se présente	MAL se manifeste
181	MAL se présente sur PROC	MAL se montre grâce à PROC
182	PROC est recommandée chez STAS	(METIER) conseille PROC chez STAS
183	MAL relève de FONCTION	MAL est associée à FONCTION
184	PROC associe PROCEDURES	PROC combine PROCEDURES
185	FONCTION est associée à FONCTION	FONCTION est liée à FONCTION
186	FONCTION s'observe	(METIER) constate FONCTION
187	STAS relève de PROC	STAS bénéficie de PROC
188	ANA est activée	ANA est rendue active
189	PCHIMIQUE active FONCTION	PCHIMIQUE rend FONCTION active
190	STAS développe ANA	STAS produit ANA
191	METIER diagnostique STAS	METIER examine STAS
192	MAL évolue en MAL	MAL se transformer en MAL
194	ANA contrôle FONCTION	ANA assure FONCTION
195	FONCTION contrôle FONCTION	FONCTION régule FONCTION
196	PROC survient chez STAS	(METIER) effectue PROC sur STAS
197	FONCTION est synthétisée par ANA	FONCTION est produite par ANA
198	MAL colonise ANA	MAL envahit ANA
199	FONCTION est synthétisée par AGENT	FONCTION est produite par le canal de AGENT
200	ANA synthétise FONCTION	ANA produit FONCTION
201	PROC est relevée	(METIER) enregistre PROC
202	MAL affecte FONCTION	MAL touche FONCTION
203	QQCHSE affecte PCHIMIQUE	QQCHSE concerne PCHIMIQUE
204	PROC est évoquée chez STAS	(METIER) fait PROC chez STAS
205	STAS inhale PCHIMIQUE	STAS aspire PCHIMIQUE
206	MAL est associée à PROC	MAL est liée à PROC
207	AGENT affecte ANA	AGENT influence ANA
208	STAS traité à l'aide de PCHIMIQUE	(METIER) soigne STAS à l'aide de PCHIMIQUE
209	PROC se traduit par PROC	PROC entraîne PROC
210	PROC s'accompagne de FONCTION	PROC entraîne FONCTION
211	MAL traduit MAL	MAL reflète MAL
212	MAL traduit FONCTION	MAL aboutit à FONCTION
213	METIER pratique PROC	METIER fait PROC
214	PROC est associée à PROC	PROC entraîne PROC
215	PROC est pratiquée (par ANA)	(METIER) fait PROC (par ANA)
216	FONCTION s'accompagne de AGENT	FONCTION entraîne AGENT
217	FONCTION isolée	FONCTION se manifeste sans aucun autre symptôme
218	ANA transmet ANA	ANA conduit ANA
219	FONCTION est transmise	FONCTION est communiquée
221	ANA sécrète FONCTION	ANA produit FONCTION

TAB. D.16 – Liste des PSS alignés (6).

222	FONCTION exprime MAL	FONCTION décrit MAL
223	MAL s'exprime par FONCTION	MAL se manifeste par FONCTION
224	FONCTION est altérée	FONCTION est modifiée
225	FONCTION altère FONCTION	FONCTION modifie FONCTION
226	PROC dépiste FONCTION	PROC permet d'exposer FONCTION
227	FONCTION est observée	(METIER) remarque FONCTION
228	PROC induire MAL	PROC entraîne MAL
229	METIER élimine ANA	METIER évacue ANA
230	ANA élimine ANA	ANA détruit ANA
231	MAL se dissémine	MAL se répand
232	ANA se dissémine dans ANA	ANA se répand dans ANA
233	MAL associe MALADIES	MAL combine MALADIES
234	PCHIMIQUE inhibe FONCTION	PCHIMIQUE bloque FONCTION
235	ORG est excrété	ORG est évacuée
236	ANA est éliminée	ANA est évacuée
237	FONCTION relève de FONCTION	FONCTION dérive de FONCTION
239	PCHIMIQUE est évalué	(METIER) contrôle PCHIMIQUE
240	ANA élimine PCHIMIQUE	ANA évacue PCHIMIQUE
241	MAL est éliminée par FONCTION	MAL est détruite par FONCTION
242	PROC relève de STAS	PROC dépend de METIER
243	STAS présente MAL	STAS souffre de MAL

Table des figures

1.1 Représentation graphique de la stemma (tirée de Tesnière 1953, p. 4). 21
1.2 Dicovalence : exemple de cadre de valence 38

2.1 Page d'accueil du CISMeF : formulaire de requêtes 49
2.2 Quelques sujets abordés sur les forums doctissimo. 51
2.3 Exemple d'annotation Cordial. 60

3.1 Schéma de la méthode. 66
3.2 Exemple de ligne de commande Wget. 67
3.3 Processus d'annotation des corpus et acquisition des PSS. 68
3.4 Exemple d'annotation Cordial. 69
3.5 Étapes du processus de sélection des PSS. 92
3.6 Format de présentation des PSS pour la validation. 99
3.7 Étapes du processus d'alignement des PSS à simplifier. 102
3.8 Modèle d'analyse de la distribution des arguments du verbe dans un corpus donné. 108
3.9 Méthode FN vs. notre méthode. 111

4.1 Schéma de la méthode : l'annotation syntaxique. 114
4.2 Comparaison de la fréquence des verbes dans les corpus. 117
4.3 Distribution de 30 verbes dans les différents corpus. 128
4.4 Schéma de la méthode : sélection des verbes et PSS. 158
4.5 Schéma général de la méthode : validation des PSS. 163
4.6 Processus de validation des PSS. 163
4.7 Les combinaisons de réponses. 166
4.8 Schéma de la méthode : la simplification des PSS. 175

5.1 Exemple d'évaluation d'une phrase. 199
5.2 Repérage du verbe pivot par Cordial. 199
5.3 Exemple de faux COI. 202
5.4 Problème de raccordement des arguments. 204
5.5 Analyse syntaxique d'une phrase avec forme verbale complexe. 207

Liste des tableaux

2.1 Taille et contenu des 4 corpus. 52

3.1 Exemple de n-gramme. 72
3.2 Exemple de bigramme. 73
3.3 Exemple de tête multicatégorielle : *risque*. 77
3.4 Exemple de tête multicatégorielle : *manœuvre*. 77
3.5 Cas d'ambiguïté des termes portant les catégories D et F. 81
3.6 Exemple d'application de la méthode fréquentielle d'annotation avec la tête *implantation*. 83
3.7 Répartition des experts et des PSS. 100

4.1 La fréquence verbale dans les 4 types de corpus. 116
4.2 Les verbes et leurs fréquences dans les 4 sous-corpus. 119
4.3 Verbes avec fréquence supérieure à la moyenne dans les corpus experts (PRO et/ou ETU). 121
4.4 Verbes avec fréquence supérieure à la moyenne dans FOR et/ou VUL. 123
4.5 Les verbes ayant un schéma fréquentiel similaire dans PRO et FOR. 125
4.6 Proximité entre les corpus PRO, FOR et VUL, en termes de nombre de verbes (1). 129
4.7 Proximité entre les corpus ETU, FOR et VUL, en termes de nombre de verbes (2). 129
4.8 Résultat de l'annotation des termes-arguments. 131
4.9 Types et nombres de PSS selon les corpus. 133
4.10 Quelques verbes avec leurs nombres de PSS. 134
4.11 Quelques PSS fréquents (1). 139
4.12 Quelques PSS fréquents (2). 140
4.13 Distribution syntaxique des PSS dans les corpus. 142
4.14 Variation sémantique autour d'une structure argumentale de base. 148
4.15 Répartition syntaxique des PSS de *relever* et *accompagner* dans les corpus PRO et FOR. 150
4.16 *Relever* dans les constructions transitives directe et indirecte. 150
4.17 *Accompagner* dans les constructions transitives directe et indirecte. 151
4.18 La variation lexicale au sein des PSS. 152

4.19 Récapitulatif du processus de sélection des verbes. 159
4.20 Liste des verbes sélectionnés pour la simplification 159
4.21 Résultats obtenus classés par catégories. 164
4.22 Récapitulatif de la validation. 165
4.23 Les 19 patrons validés partiellement (1/3 experts). 168
4.24 Quelques patrons sélectionnés à l'unanimité mais ayant moins de 10 occurrences dans le corpus. 170
4.25 Les PSS ne nécessitant pas de simplification. 173
4.26 Exemples de groupes de verbes candidats équivalents pour l'alignement. 176
4.27 Quelques PSS alignés avec leurs équivalents. 177
4.28 Récapitulatif du résultat de la simplification. 179
4.29 Exemples de résultats de l'analyse des collocations verbe-terme : cas du verbe *accompagner*. 183
4.30 Quelques verbes autour des catégories fréquentes du corpus : P et D. 183
4.31 Quelques collocations verbe-terme dans le corpus. 185
4.32 Illustration de la préférence lexicale des verbes dans le corpus PRO. 186
4.33 Illustration de la préférence lexicale des verbes dans le corpus FOR. 187

5.1 Liste des participants à la campagne EASY. 193
5.2 Performances de Cordial dans EASY : mesures en précision (p) et f-mesure (f) par type de corpus pour tous les constituants et toutes les relations (Paroubek *et al.*, 2007). 195
5.3 Performances globales de Cordial dans PASSAGE 1 (Laurent *et al.*, 2009). . . . 196
5.4 Performances de Cordial sur les textes médicaux et courriers électroniques (*mail*) (Laurent *et al.*, 2009). 196
5.5 Résultats de l'évaluation de l'annotation syntaxique (PRO). 200
5.6 Résultats de l'évaluation de l'annotation syntaxique (ETU). 200
5.7 Résultats de l'évaluation de l'annotation syntaxique (VUL). 200
5.8 Résultats de l'évaluation de l'annotation syntaxique (FOR). 200
5.9 Nombre de phrases contenant des faux COI (N = 50). 203
5.10 Résultats de l'évaluation de l'annotation sémantique des corpus. 210
5.11 Les différentes sources du bruit. 211
5.12 Les différentes sources du silence. 212
5.13 Résultats de l'évaluation par les linguistes. 216
5.14 Les phrases portant l'étiquette *non*. 217
5.15 Résultats de l'évaluation par les non-linguistes. 218
5.16 Les phrases portant l'étiquette *non*. 219

5.17 Les phrases portant l'étiquette *non* chez les linguistes et les non-linguistes collectivement. 220
5.18 Quelques phrases ayant des groupes nominaux compréhensibles en position d'arguments. 221

A.1 Liste des étiquettes utilisées par Cordial pour l'annotation des fonctions grammaticales. 227

B.1 Quelques exemples de termes pluriels avec leurs catégories. 229
B.2 Les termes mal orthographiés du corpus des forums (1) 230
B.3 Les termes mal orthographiés du corpus des forums (2) 231
B.4 Les mots mal orthographiés du corpus des forums (3) 232
B.5 Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (1). . 233
B.6 Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (2). . 234
B.7 Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (3). . 235
B.8 Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (4). . 236
B.9 Liste des têtes multicatégorielles Snomed avec leurs différentes catégories (5). . 237
B.10 Têtes dont la catégorie la plus fréquente enregistre un pourcentage ≥ 90. . . . 237
B.11 Têtes dont la catégorie la plus fréquente enregistre un pourcentage ≥ 90. . . . 238
B.12 Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (1). . 239
B.13 Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (2). . 240
B.14 Têtes dont la catégorie la plus fréquente enregistre un pourcentage < 90 (3). . 241

C.1 Liste des mots outils utilisés (1). 243
C.2 Liste des mots outils utilisés (2). 244
C.3 Liste des déterminants complexes utilisés (1). 245
C.5 Liste des 147 verbes de réalisation. 246
C.4 Liste des déterminants complexes utilisés (2). 247

D.1 Les 50 phrases (originales) évaluées (1). 249
D.2 Les 50 phrases (originales) évaluées (2). 250
D.3 Les 50 phrases (originales) évaluées (3). 251
D.4 Les 50 phrases (originales) évaluées (4). 252
D.5 Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (1). 253
D.6 Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (2). 254
D.7 Liste des phrases après remplacement des verbes par des équivalents compréhenvibles et après simplification des PSS (3). 255

D.8 Liste des 243 PSS sélectionnés pour validation par les experts (1). 256
D.9 Liste des 243 PSS sélectionnés pour validation par les experts (2). 257
D.10 Liste des 243 PSS sélectionnés pour validation par les experts (3). 258
D.11 Liste des PSS alignés (1). 259
D.12 Liste des PSS alignés (2). 260
D.13 Liste des PSS alignés (3). 261
D.14 Liste des PSS alignés (4). 262
D.15 Liste des PSS alignés (5). 263
D.16 Liste des PSS alignés (6). 264

Résumé en francais

Mots clés : langue de spécialité, langage médical, communication médecin/patient, verbe spécialisé, structure argumentale, simplification de textes

Grâce à l'évolution de la technologie à travers le Web, la documentation relative à la santé est de plus en plus abondante et accessible à tous, plus particulièrement aux patients, qui ont ainsi accès à une panoplie d'informations sanitaires. Malheureusement, la grande disponibilité de l'information médicale ne garantit pas sa bonne compréhension par le public visé, en l'occurrence les non-experts. Notre projet de thèse a pour objectif la création d'une ressource de simplification de textes médicaux, à partir d'une analyse syntaxico-sémantique des verbes dans quatre corpus médicaux en français qui se distinguent de par le degré d'expertise de leurs auteurs et celui des publics cibles. La ressource conçue contient 230 patrons syntaxico-sémantiques des verbes (appelés PSS), alignés avec leurs équivalents non spécialisés. La méthode semi-automatique d'analyse des verbes appliquée pour atteindre notre objectif est basée sur quatre tâches fondamentales : l'annotation syntaxique des corpus, réalisée grâce à l'analyseur syntaxique Cordial (Laurent, Dominique et al, 2009) ; l'annotation sémantique des arguments des verbes, à partir des catégories sémantiques de la version française de la terminologie médicale Snomed Internationale (Côté, 1996) ; l'acquisition des patrons syntactico-sémantiques et l'analyse contrastive du fonctionnement des verbes dans les différents corpus. Les patrons syntaxico-sémantiques des verbes acquis au terme de ce processus subissent une évaluation (par trois équipes d'experts en médecine) qui débouche sur la sélection des candidats constituant la nomenclature de la ressource de simplification. Les PSS sont ensuite alignés avec leurs correspondants non spécialisés, cet alignement débouche sur le création de la ressource de simplification, qui représente le résultat principal de notre travail de thèse. Une évaluation du rendement du contenu de la ressource a été effectuée avec deux groupes d'évaluateurs : des linguistes et des non-linguistes. Les résultats montrent que la simplification des PSS permet de faciliter la compréhension du sens du verbe en emploi spécialisé, surtout lorsque un certains paramètres sont réunis.

English Abstract

Keywords : language for specific purposes, medical language, doctor/patient communication, specialized verb readings, argument structure, text simplification

With the evolution of Web technology, healthcare documentation is becoming increasingly abundant and accessible to all, especially to patients, who have access to a large amount of health information. Unfortunately, the ease of access to medical information does not guarantee its correct understanding by the intended audience, in this case non-experts. Our PhD work aims at creating a resource for the simplification of medical texts, based on a syntactico-semantic analysis of verbs in four French medical corpora, that are distinguished according to the level of expertise of their authors and that of the target audiences. The resource created in the present thesis contains 230 syntactico-semantic patterns of verbs (called PSS), aligned with their non-specialized equivalents. The semi-automatic method applied, for the analysis of verbs, in order to achieve our goal is based on four fundamental tasks : the syntactic annotation of the corpora, carried out thanks to the Cordial parser (Laurent *et al.*, 2009) ; the semantic annotation of verb arguments, based on semantic categories of the French version of a medical terminology known as Snomed International (Côté, 1996) ; the acquisition of syntactico-semantic patterns of verbs and the contrastive analysis of the verbs behaviors in the different corpora. The PSS, acquired at the end of this process, undergo an evaluation (by three teams of medical experts) which leads to the selection of candidates constituting the nomenclature of our text simplification resource. These PSS are then aligned with their non-specialized equivalents, this alignment leads to the creation of the simplification resource, which is the main result of our PhD study. The content of the resource was evaluated by two groups of people : linguists and non-linguists. The results show that the simplification of PSS makes it easier for non-experts to understand the meaning of verbs used in a specialized way, especially when a certain set of parameters is collected.

Bibliographie

Abdaoui, A., Azé, J., Bringay, S., Grabar, N. & Poncelet, P. (2014). Analysis of forum posts written by patients and health professionals. In *e-Health - For Continuity of Care - Proceedings of MIE2014, the 25th European Medical Informatics Conference, Istanbul, Turkey, August 31 - September 3, 2014*, p. 1185.

Abeillé, A. & Barrier, N. (2004). Enriching a french treebank. In *LREC*.

Abney, S. (1997). Part-of-speech tagging and partial parsing. In *Corpus-based methods in language and speech processing*, p. 118–136. Springer.

Abrahamsson, E., Forni, T., Skeppstedt, M. & Kvist, M. (2014). Medical text simplification using synonym replacement : Adapting assessment of word difficulty to a compounding language. In *In Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, Gothenburg, Sweden.

Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M. & Wächter, T. (2009). Biomedical word sense disambiguation with ontologies and metadata : automation meets accuracy. *BMC bioinformatics*, **10**(1), 28.

AMA (1999). Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, **281**(6), 552–7.

Androutsopoulos, I. & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, **38**, 135–187.

Angrosh, M., Nomoto, T. & Siddharthan, A. (2014). Lexico-syntactic text simplification and compression with typed dependencies. In *COLING*, p. 1996–2006.

Anspach, R. R. (1988). Notes on the sociology of medical discourse : The language of case presentation. *Journal of Health and Social Behavior*, p. 357–375.

Ashby, W. J. (1988). Français du canada/français de france : divergence et convergence. *French Review*, p. 693–702.

Augustyn, M., Hamou, S. B., Bloquet, G., Goossens, V., LOISEAU, M. & RINCK, F. (2008). Constitution de ressources pédagogiques numériques : le lexique des affects. *Autour des langues et du langage*, p. 407.

Bagola, B., Niederehe, H. & Wolf, L. (2007). Français du canada/français de france : actes du viii e colloque international.

Baker, M. C. (1997). Thematic roles and syntactic structure. In *Elements of grammar*, p. 73–137. Springer.

Balahur, A. (2013). Sentiment analysis in social media texts. In *4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, p. 120–128.

Barque, L., Marin, R., Jugnet, A. & Huyghe, R. (2009). Two types of deverbal activity nouns in french. *Invited Talks.*

Barzilay, R. & Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, p. 25–32 : Association for Computational Linguistics.

Beckman, H., Kaplan, S. H. & Frankel, R. (1989). Outcome based research on doctor-patient communication : A review. *Communicating With Medical Patients. Newbury Park, Calif : Sage Publications*, p. 223–227.

Belder, D. & Moens, M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, p. 19–26 : ACM.

Bensing, J. (1991). Doctor-patient communication and the quality of care. *Social science & medicine*, **32**(11), 1301–1310.

Berkenkotter, C., Hanganu-Bresch, C. & Dreher, K. (2015). Descriptive psychopathology in asylum case histories : The case of john horatio baldwin. *ASp. la revue du GERAS*, (68).

Bessières, P., Nazarenko, A. & Nédellec, C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. In *CIDE 2001 : colloque international sur le document électronique*, p. 165–183.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biran, O., Brody, S. & Elhadad, N. (2011). Putting it simply : a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers, Vol 2*, p. 496–501 : Association for Computational Linguistics.

Blanchard, C. G., Ruckdeschel, J. C., Blanchard, E. B., Arena, J. G., Saunders, N. L. & Malloy, E. D. (1983). Interactions between oncologists and patients during rounds. *Annals of Internal Medicine*, **99**(5), 694–699.

Borin, L., Grabar, N., Hallett, C., Toporowska Gronostaj, M., Kokkinakis, D., Williams, S., Willis, A. *et al.* (2007a). Empowering the patient with language technology.

Borin, L., Gronostaj, M. T. & Kokkinakis, D. (2007b). Medical frames as target and tool. *Frame 2007 : Building frame semantics resources for Scandinavian and Baltic languages*, p.1̃1.

Bouhaddou, O. & Warner, H. (1994). An interactive patient information and education system (medical housecall) based on a physician expert system (iliad). *Medinfo. MEDINFO*, **8**, 1181–1185.

Brouwers, L., Bernhard, D., Ligozat, A.-L. & François, T. (2012). Simplification syntaxique de phrases pour le français. In *TALN*, p. 211–224.

Brouwers, L., Bernhard, D., Ligozat, A.-L. & François, T. (2014). Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, p. 47–56.

Brown, S., Dligach, D. & Palmer, M. (2011). Verbnet class assignment as a wsd task. In *9th International Conference on Computational Semantics*, Oxford, UK.

Brunt, R. (2008). Medical english since the mid-nineteenth century. *In : Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*.

Buchholz, S. & Marsi, E. (2006). Conll-xshared task on multilingual dependency parsing. In *In Proc. of CoNLL*, p. 149–164.

Buller, M. K. & Buller, D. B. (1987). Physicians' communication style and patient satisfaction. *Journal of Health and Social Behavior*, p. 375–388.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S. & Pinkal, M. (2006). The salsa corpus : a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, p. 969–974.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S. & Pinkal, M. (2009). Using framenet for the semantic analysis of german : Annotation, representation, and automation. *Multilingual FrameNets in Computational Lexicography : methods and applications*, p. 209–244.

Bélisle, L. (1974). *Dictionnaire général de la langue française au Canada*. Bélisle, Montréal Sondec.

Cabot, C., Soualmia, L. F., Dahamna, B. & Darmoni, S. J. (2016). Ecmt : Indexation multi-terminologique de documents biomédicaux. In *1er Forum Franco-Québécois d'Innovation en Santé, Polytechnique Montréal*.

Cajolet-Laganière, H. (2009). Marques et indicateurs géographiques dans le dictionnaire général du français de l'équipe franqus. In *Français du Canada-Français de France VIII : Actes du huitième Colloque international de Trèves, du 12 au 15 avril 2007*, volume 23, p. 121 : Walter de Gruyter.

Cajolet-Laganière, H. & Maillet, N. (1995). Caractérisation des textes techniques québécois. *Présence francophone*, (47), 113–137.

Callan, J. & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, p. 460–467.

Canada, S. (1998). *Outils de communication I-une meilleure communication médecin-patient pour de meilleurs résultats auprès des patients*. Technical report, Ottawa : Santé Canada.

Canada, S. (2001). *La communication efficace à votre service. Outils de communication II. Guide de ressources*. Technical report, Ottawa : Santé Canada.

Candel, D. (1984). Une approche de la langue des physiciens. *Langue française*, (64), 93–108.

Candido, A. J., Maziero, E., Gasperin, C., Pardo, T. A. S., Specia, L. & Aluisio, S. M. (2009). Supporting the adaptation of texts for poor literacy readers : a text simplification editor for Brazilian Portuguese. In *EdAppsNLP '09 Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 34–42.

Candito, M., Nivre, J., Denis, P. & Anguiano, E. (2010). Benchmarking of statistical dependency parsers for french. In *International Conference on Computational Linguistics*, p. 108–116.

Candito, M. & Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

Carroll, J., Minnen, G., Canning, Y., Devlin, S. & Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, p. 7–10 : Citeseer.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S. & Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of EACL*, volume 99, p. 269–270.

Castellví, M. T. C. (2002). Textos especializados y unidades de conocimiento : metodología y tipologización. In *Texto, terminología y traducción*, p. 15–36.

Cavalla, C. (2010). Les écrits universitaires des étudiants étrangers : quelles normes présenter ?

Chall, J. S. & Dale, E. (1995). *Readability Revisited : The New Dale-Chall Readability Formula*. Brookeline Books.

Chandrasekar, R. & Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge Based Systems*, **10**(3), 183–190.

Chapman, K., Abraham, C., Jenkins, V. & Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology*, **12**(6), 557–566.

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P. & Vandenbussche, P.-Y. (2012). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In *23es journées francophones d'Ingénierie des connaissances*, p. 33–48.

Charpy, J.-P. (2015). Maurizio gotti, stefania maci, michele sala (eds.), insights into medical communication. bern, berlin, brussels, frankfurt am main, new york, oxford, vienna : Peter lang, 2015. *ASp. la revue du GERAS*, (68), 121–131.

Chebil, W., Soualmia, L. F., Omri, M. N. & Darmoni, S. J. (2014). Extraction possibiliste de concepts mesh à partir de documents biomédicaux. *Revue d'Intelligence Artificielle*, **28**(6), 729–752.

Chmielik, J. & Grabar, N. (2011). Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques. *TAL*, **51**(2), 151–179.

Cho, N. (2003). Linguistic features of electronic mail. *S. Herring (ed.), Computer-Mediated Conversation*.

Christie, F. (2002). The development of abstraction in adolescence in subject english. *Developing advanced literacy in first and second languages : Meaning with power*, p. 45–66.

Christy, N. P. (1979). English is our second language.

Chute, C. G., Cohn, S., Campbell, K., Oliver, D. & Campbell, J. (1996). The content coverage of clinical classifications. for the computer-based patient record institute's work group on codes & structures. *J Am Med Inform Assoc*, **3**(3), 224–33.

Chy, Y., Parsons, J., Mamdani, M., Lebovic, G., Shah, B., Bhattacharyya, O., Laupacis, A. & Straus, S. (2012). Designing and evaluating a web-based selfmanagement site for patients with type 2 diabetes - systematic website development and study protocol. *BMC Medical Informatics and Decision Making 2012, 12 :57*, **12**, 57.

Coch, J. (1996). Overview of alethgen. In *Demonstrations and Posters of the Eighth International Natural Language Generation Workshop (INLG'96)*, p. 25–28.

Coch, J. (1998). Interactive generation and knowledge administration in multimeteo. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, p. 300–303.

Collet, T. (1997). La réduction des unités terminologiques complexes de type syntagmatique. *Meta : Journal des traducteursMeta :/Translators' Journal*, **42**(1), 193–206.

Condamines, A. (1992). Aide à l'acquisition de connaisances par la spécification de la terminologie d'un domaine de spécialité. *Proc. of the 3 rd Journées d'Acquisition des Connaissances, Dourdan, France*.

Condamines, A. (1993). Un exemple d'utilisation de connaissances de sémantique lexicale : acquisition semi-automatique d'un vocabulaire de spécialité. *Cahiers de lexicologie*, **62**, 25–65.

Condamines, A. & Bourigault, D. (1999). Alternance nom/verbe : explorations en corpus spécialisés. In *Cahiers de l'Elsap*, p. 41–48, Caen, France.

Côté, R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Coulon, R. (1972). French as it is written by french sociologists. *Bulletin pédagogique des IUT (18)*, p. 11–25.

Culotta, A. & Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, p. 423 : Association for Computational Linguistics.

Da Cunha, I., Cabré, M. T., SanJuan, E., Sierra, G., Torres-Moreno, J. M. & Vivaldi, J. (2011). Automatic specialized vs. non-specialized sentence differentiation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 266–276 : Springer.

Daelemans, W., Höthker, A. & Sang, E. F. T. K. (2004). Automatic sentence simplification for subtitling in dutch and english. In *In proceedings of the 4th international conference on Language resources and Evaluation (LREC)*.

Darmoni, S. J., Leroy, J.-P., Baudic, F., Douyère, M., Piot, J. & Thirion, B. (1999). Cismef : catalogue et index des sites médicaux francophones. *Cahiers d'études et de recherches francophones/Santé*, **9**(2), 123–128.

De La Clergerie, E. V., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P. & Vilnat, A. (2008). Passage : from french parser evaluation to large sized treebank. In *Proceedings of the 6th Language Resource and Evaluation Conference (LREC'08)*, volume 100, p.P̃2.

Debusmann, R. & Kuhlmann, M. (2010). Dependency grammar : Classification and exploration. In *Resource-adaptive cognitive processes*, p. 365–388. Springer.

Deléger, L. & Zweigenbaum, P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, p. 146–50.

Deléger, L. & Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, p. 2–10 : Association for Computational Linguistics.

Delpech, E. (2011). Un protocole d'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée. *Revue des Nouvelles Technologies de l'information*, (spécial Qualité des Données et des Connaissances/Évaluation des méthodes d'Extraction de Co), 23–48.

Diana, M. & Roberto, N. (2007). Semeval-2007 task 10 : English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, p. 48–53 : Association for Computational Linguistics.

Djemaa, M., Candito, M., Muller, P. & Vieu, L. (2016). Corpus annotation within the french framenet : methodology and results. In *Proceedings of LREC 2016*, Portoroz, Slovenia.

Dolbey, A. (2009). *Bioframenet : a framenet extension to the domain of molecular biology*. Ph.d. thesis, Department of Linguistics, UC Berkeley.

Dolbey, A., Ellsworth, M. & Scheffczyk, J. (2006). Bioframenet : A domain-specific framenet extension with links to biomedical ontologies. In *KR-MED*, volume 222.

Dowty, D. (1991). Thematic proto-roles and argument selection. *language*, p. 547–619.

Drouin, P. (2003). Termostat web 3.0.

DuBay, W. H. (2007). The classic readability studies. *Online Submission*.

Dubois, J. & Dubois-Charlier, F. (1997). *Les verbes français*. Larousse.

Elhadad, N. (2006). Comprehending technical texts : predicting and defining unfamiliar terms. In *AMIA*, p. 239–243.

Elhadad, N. & Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, p. 49–56 : Association for Computational Linguistics.

Eynde, K. & Mertens, P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, **13**(1), 63–104.

Faber, P. (2012). *A cognitive linguistics view of terminology and specialized language*, volume 20. Walter de Gruyter.

Fábregas, A. & Marín, R. (2012). The role of aktionsart in deverbal nouns : State nominalizations across languages. *Journal of Linguistics*, **48**(01), 35–70.

Fábregas, A., Marín, R. & McNally, L. (2012). From psych verbs to nouns. *Demonte & McNally (eds.)*.

Falkenjack, J., Mühlenbock, K. H. & Jönsson, A. (2013). Features indicating readability in swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) ; May 22-24 ; 2013 ; Oslo University ; Norway. NEALT Proceedings Series 16*, number 085, p. 27–40 : Linköping University Electronic Press.

Fang, Z. (2005). Scientific literacy : A systemic functional linguistics perspective. *Science education*, **89**(2), 335–347.

Faure, P. (2010). Des discours de la médecine multiples et variés à la langue médicale unique et universelle. *In : ASp la revue du GERAS*, (58), 73–86.

Feyrer, C. (2016). 6 textes et discours en médecine. *Manuel des langues de spécialité*, **12**, 147.

Fillmore, C. (1968). *The case for case*, In Universals, Ed., *Linguistic Theory*, p. 1–88.

Fillmore, C. (1976). *Topics in lexical semantics*, In I. U. Press, Ed., *Current Issues in Linguistic Theory*, p. 76–138.

Fillmore, C. (1982). *Frame Semantics*, In H. P. Co, Ed., *Linguistics in the morning calm*, p. 111–137.

Fillmore, C. J. & Baker, C. (2010). A frames approach to semantic analysis.

Fleischman, S. (2003). Language and medicine. *The handbook of discourse analysis*, **18**, 470–502.

Foley, W. A. & Van Valin Jr, R. D. (1984). Functional syntax and universal grammar. *Cambridge Studies in Linguistics London*, (38).

Fontenelle, T. (2009). sémantique des cadres et lexicographie. *Lexique*, (19), 162–177.

Fournier, C. & Kerzanet, S. (2007). Communication médecin-malade et éducation du patient, des notions à rapprocher : apports croisés de la littérature. *Santé Publique*, **19**(5), 413–425.

François, T. & Watrin, P. (2011). On the contribution of mwe-based features to a readability formula for french as a foreign language. In *RANLP*, p. 441–447 : Citeseer.

François, J., Le Pesant, D. & Leeman, D. (2007). Présentation de la classification des verbes français de jean dubois et françoise dubois-charlier. *Langue Française*, **153**(1), 3–19.

Gardiner, R. (2008). The transition from 'informed patient'care to 'patient informed'care. *Stud Health Technol Inform*, **137**, 241–56.

Gauducheau, N. (2008). La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives. *Bulletin de psychologie*, (4), 389–404.

Gautier, L. (2014). Des langues de spécialité à la communication spécialisée : un nouveau paradigme de recherche à l'intersection entre sciences du langage, info-com et sciences cognitives ? *Etudes Interdisciplinaires en Sciences humaines*, **1**, 225–245.

Gendron, J. (1966). *Tendances phonétiques du française parlé au Canada*, volume 2. C. Klincksieck.

Giacomini, L. (2015). Context-dependent variation of lsp collocations : A corpus-based analysis. *Procedia-Social and Behavioral Sciences*, **198**, 140–148.

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, **28**(3), 245–288.

Gladkij, A. V. & Mel'čuk, I. A. (1975). Tree grammars. i. a formalism for syntactic transformations in natural languages. *Linguistics*, **13**(150), 47–82.

Gläser, R. (1990). *Fachtextsorten im Englischen*, volume 13. Gunter Narr Verlag.

Göpferich, S. (1995). *Textsorten in Naturwissenschaften und Technik : pragmatische Typologie-Kontrastierung-Translation*. Narr.

Grabar, N. (2004). *Terminologie m'edicale et morphologie. Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique*. Thèse de doctorat, Universit'e Paris 6, Paris.

Grabar, N. & Dumonet, L. (2015). Automatic computing of global emotional polarity in french health forum messages. In *Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings*, p. 243–248.

Grabar, N. & Hamon, T. (2004). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle (RIA)*, **18**(1), 57–85.

Grabar, N. & Hamon, T. (2014). Automatic extraction of layman names for technical medical terms. In *2014 IEEE International Conference on Healthcare Informatics, ICHI 2014, Verona, Italy, September 15-17, 2014*, p. 310–319.

Grabar, N., Krivine, S. & Jaulent, M.-C. (2007). Classification of health webpages as expert and non expert with a reduced set of cross-language features. In *AMIA Annual Symposium Proceedings*, volume 2007, p. 284 : American Medical Informatics Association.

Grabar, N., Zweigenbaum, P., Soualmia, L. & Darmoni, S. (2002). Les utilisateurs de Doc'CISMEF peuvent-ils trouver ce qu'ils cherchent ? Une étude de l'adéquation du vocabulaire des requêtes au MeSH. In *Journées Francophones d'Informatique Médicale (JFIM)*, Québec, Canada.

Grabar, N., Zweigenbaum, P., Soualmia, L. & Darmoni, S. J. (2003). Matching controlled vocabulary words. *Studies in Health Technology and Informatics*, **95**, 445–450.

Graesser, A. C., McNamara, D. S. & Kulikowich, J. M. (2011). Coh-metrix providing multilevel analyses of text characteristics. *Educational researcher*, **40**(5), 223–234.

Grimshaw, J. (1990). *Argument structure*. Cambridge : MIT Press.

Grivel, L. (2011). *La recherche d'information en contexte : Outils et usages applicatifs*. Lavoisier.

Gross, G. & Kiefer, F. (1995). La structure événementielle des substantifs. *Folia linguistica*, **29**(1-2), 43–66.

Guilbert, L. (1973). La spécificité du terme scientifique et technique. *Langue française*, (17), 5–17.

Haas, P. & Huyghe, R. (2010). Les propriétés aspectuelles des noms d'activités. *Cahiers Chronos*, **21**, 103–118.

Haas, P., Huyghe, R. & Marín, R. (2008). Du verbe au nom : calques et décalages aspectuels. In *Congrès Mondial de Linguistique Française 2008*, p. 2051–2065.

Hall, J. A., Roter, D. L. & Katz, N. R. (1988). Meta-analysis of correlates of provider behavior in medical encounters. *Medical care*, **26**(7), 657–675.

Halliday, M. (1998). Things and relations. *Reading science : Critical and functional perspectives on discourses of science*, p. 185–235.

Halliday, M. A. & Hasan, R. (1989). Language, context, and text : Aspects of language in a social-semiotic perspective.

Hamon, T., Nazarenko, A. & Gros, C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of the 36th Annual Meeting of the*

Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, p. 498–504 : Association for Computational Linguistics.

Hancke, J., Vajjala, S. & Meurers, D. (2012). Readability classification for german using lexical, syntactic, and morphological features. In *COLING*, p. 1063–1080.

Haralambous, Y. & Lavagnino, E. (2011). La réduction de termes complexes dans les langues de spćialité. *TAL*, **52**(1), 37–68.

Harris, Z. S. (1971). *Structures mathématiques du langage*, volume 3. Dunod.

Heid, U. (1994). On ways words work together-topics in lexical combinatorics. *Martin, W. et al*, p. 226–257.

Heid, U. (2001). 8.4. 4 collocations in sublanguage texts : Extraction from corpora ulrich heid. *Handbook of Terminology Management : Application-oriented terminology management*, **2**, 788.

Heid, U. (2009). Aspects of lexical description for electronic dictionaries. *eLEX2009*, p.1̃.

Heid, U. & Freibott, G. (1991). Collocations dans une base de données terminologique et lexicale. *Meta*, **36**(1), 77–91.

Heilman, M. & Smith, N. A. (2010). Extracting simplified statements for factual question generation. In *Proceedings of QG2010 : The Third Workshop on Question Generation*, p. 11–20.

Heribert, P. (1983). Hvad gør et almensproget verbum til et fagsprogligt verbum ?–verbernes terminologisering. *AScLA-Symposiet'Oversættelse og Tolkning'4.-6. oktober 1982*, p. 189–201.

Heribert, P. (1985). Termer og deres fagsproglige omgivelser-fagsproglig fraseologi. *Nordisk terminologikursus*, p. 296–336.

Heribert, P. (1987). Terms and their lsp environment-lsp phraseology. *Meta : Journal des traducteursMeta :/Translators' Journal*, **32**(2), 149–155.

Herring, S. (2003). Winning and losing : Abbreviations and routines as community register markers on a social mud. *S. Herring (ed.), Computer-Mediated Conversation*.

Herring, S. & Androutsopoulos, J. (2015). *Computer-mediated Discourse*, In D. S. Deborah Tannen & S. E. Heidi Hamilton (Eds.), Eds., *The Handbook of Discourse Analysis*, p. 127–151.

Herring, S. C. (1998). Le style du courrier électronique : variabilité et changement. *Terminogramme, 84*, **85**, 9–16.

Herring, S. C. & Zelenkauskaite, A. (2008). Gendered typography : Abbreviation and insertion in italian itv sms. *IULC Working Papers*, **8**(3).

Heslot, J. (1983). Récit et commentaire dans un article scientifique. *DRLAV*, **29**, 133–154.

Hibbard, J. H. & Peters, E. (2003). Supporting informed consumer health care decisions : data presentation approaches that facilitate the use of information in choice. *Annual review of public health*, **24**(1), 413–433.

Hoard, J. E., Wojcik, R. & Holzhauser, K. (1992). An automated grammar and style checker for writers of simplified english. In *Computers and Writing*, p. 278–296. Springer.

Hoffmann, L. (1983). Fachtextlinguistik. *Fachsprache*, **5**(2), 57–68.

Hoffmann, L. (1985). Kommunikationsmittel fachsprache. eine einführung. zweite völlig neu bearbeitete auflage. In *Tübingen : Gunter Narr Verlag.(= Forum für Fachsprachenforschung, Band 1)*.

Huenerfauth, M., Feng, L. & Elhadad, N. (2009). Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, p. 3–10 : ACM.

Hutchins, W. J. & Somers, H. L. (1992). *An introduction to machine translation*, volume 362. Academic Press London.

Huyghe, R. & Marín, R. (2007). L'héritage aspectuel des noms déverbaux en français et en espagnol. *Faits de langues*, (30), 265–274.

Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Compututational Linguistic*, **24**(1), 2–40.

Inui, K., Fujita, A., Takahashi, T., Iida, R. & Iwakura, T. (2003). Text simplification for reading assistance : a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, p. 9–16 : Association for Computational Linguistics.

Ischreyt, H. (1965). *Studien zum Verhältnis von Sprache und Technik*. Verlag Schwann.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic inquiry*, **18**(3), 369–411.

Jackendoff, R. (1990). *Semantic structures*. Cambridge : MIT Press.

Jackendoff, R. S. (1972). Semantic interpretation in generative grammar.

Jacobi, D. (1989). Reformulation et socialisation des connaissances dans des discours de vulgarisation scientifique. *Etudes de lettres*, **4**, 23–44.

Jacobi, D. (1993). Les terminologies et leur devenir dans les textes de vulgarisation scientifique.

Jacobi, D. (1994). Lexique et reformulation intradiscursive dans les documents de vulgarisation scientifique. *Français scientifique et thechnique et dictionaire de langue. Paris : Didier Érudition*, p. 77–91.

Jacquemin, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique, Université de Nantes.

Jacquemin, C. & Tzoukermann, E. (1999). Nlp for term variant extraction : A synergy of mor- phology, lexicon, and syntax. *Strzalkowski T, ed, Natural language information retrieval of Text, Speech and Language Technology*, **7**, 25–74.

Jacques, M. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. PhD thesis, Atelier national de reproduction des thèses.

Jonnalagadda, S. & Gonzalez, G. (2011). Biosimplify : an open source sentence simplification engine to improve recall in automatic biomedical information extraction. *CoRR*, **abs/1107.5744**.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C. & Gonzalez, G. (2010). Towards effective sentence simplification for automatic processing of biomedical text. *CoRR*, **abs/1001.4277**.

Kahane, S. (2000). Grammaires de dépendance, t.a.l. *Hermès*, **41**(1).

Kahane, S. & Polguère, A. (1998). *Workshop on Dependency-Based Grammars, ACL/COLING'98*. Technical report, Montréal.

Kandula, S., Curtis, D. & Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. In *AMIA Annu Symp Proc*, p. 366–70.

Karel van den, E. & Piet, M. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, **13**, 63–104.

Kasper, S. (2008). A comparison of thematic role theories.

KHARRAZI, H. (2009). Improving healthy behaviors in type 1 diabetic patients by interactive frameworks. In *AMIA*, p. 322–326.

KIDA, T. (2001). Multimodalité dans l'interaction natif/non natif : cas de négociation du sens. In *Communication présentée à la conférence : The multimodality of Human Communication : Theories, problems, and Applications*, p. 3–6.

KILGARRIFF, A., BAISA, V., BUŠTA, J., JAKUBÍČEK, M., KOVÁŘ, V., MICHELFEIT, J., RYCHLÝ, P. & SUCHOMEL, V. (2014). The sketch engine : ten years on. *Lexicography*, **1**(1), 7–36.

KILGARRIFF, A., RYCHLY, P., SMRZ, P. & TUGWELL, D. (2004). The sketch engine. In *Proc. Euralex.*, p. 105–116, Lorient, France.

KINCAID, J. P., FISHBURNE JR, R. P., ROGERS, R. L. & CHISSOM, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical report, DTIC Document.

KIPPER, K., DANG, H. & PALMER, M. (2000). Class-based construction of a verb lexicon. In *The Seventh National Conference on Artificial Intelligence AAAI/IAAI*, p. 691–696.

KIPPER-SCHULER, K. (2005). *VerbNet : A broad-coverage comprehensive verb lexicon*. Thèse de doctorat, niversity of Pennsylvania, Philadelphia, PA.

KOKKINAKIS, D. & TOPOROWSKA GRONOSTAJ, M. (2006). Comparing lay and professional language in cardiovascular disorders corpora. In J. C. U. PHAM T., Ed., *WSEAS Transactions on Biology and Biomedicine*, p. 429–437.

KORHONEN, A. & BRISCOE, T. (2004). Extended lexical-semantic classification of english verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, Boston, MA.

LAURENT, D. & NÈGRE, S. (2006). *Cordial, le TAL et les aides à la rédaction*. Technical report, Journées de l'ATALA, Paris.

LAURENT, D., NÈGRE, S. & SÉGUÉLA, P. (2009). L' analyseur syntaxique Cordial dans Passage. *Actes de TALN*, **9**.

LERAT, P. (1995). *Les langues spécialisées*. Presses universitaires de France.

LERAT, P. (2002). Qu'est-ce que le verbe spécialisé ? le cas du droit. *Cahiers de Lexicologie*, **80**, 201–211.

Lerner, E. B., Jehle, D. V., Janicke, D. M. & Moscati, R. M. (2000). Medical communication : do our patients understand ? *The American journal of emergency medicine*, **18**(7), 764–766.

Leroy, G., Endicott, J. E., Mouradi, O., Kauchak, D. & Just, M. (2012). Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, **707**(10).

Levin, B. (1993). *English Verb Classes and Alternation : A Preliminary Investigation*. The University of Chicago : Press.

L'Homme, M. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, **73**(2), 61–84.

L'Homme, M. (2003). Capturing the lexical structure in special subject fields with verbs and verbal derivatives a model for specialized lexicography. *IJL*, **16**(4), 403–422.

L'Homme, M. (2009). *Le DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet*. Technical report, Observatoire de linguistique Sens-Texte (OLST).

L'Homme, M. (2012a). Adding syntactico-semantic information to specialized dictionaries : an application of the FrameNet methodology. *Lexicographica*, **28**, 233–252.

L'Homme, M. (2012b). Le verbe terminologique un portrait de travaux récent. In *Congrès Mondial de Linguistique Française-CMLF*, p. 93–107.

L'Homme, M. & Bodson, C. (1997). Modèle de description des verbes specialises combinant base de connaissances et hypertexte. In *Congres international de terminologie*, p. 381–398, San Sebastian, Espagne.

L'Homme, M.-C. (1992). *Contribution à l'analyse grammaticale de la langue de spécialité : le mode, le temps et la personne du verbe dans quelques textes scientifiques écrits à vocation pédagogique*. PhD thesis, Université Laval.

L'Homme, M.-C. (1996). Formes verbales de temps et texte scientifique. *Le Langage et l'homme*, **31**(2-3), 107–123.

Lin, J. & Wilbur, W. J. (2007). Syntactic sentence compression in the biomedical domain : facilitating access to related articles. *Information Retrieval*, **10**(4-5), 393–414.

Lorente, M. (2002). Verbos y discurso especializado. *Estudios de lingüística española*.

Lothar, H., Hartwig, K. & Herbert, E. (1998). Fachsprachen/languages for special purposes. ein internationales handbuch zur fachsprachenforschung und terminologiewissenschaft/an international handbook of special-language and terminology research. 2 halbband./vols. (handbücher zur sprach-und kommunikationswissenschaft 14).

Lovis, C. & Baud, R. H. (2000). Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language. *Journal of the American Medical Informatics Association*, **7**(4), 378–391.

Lovis, C., Michel, P.-A., Baud, R. & Scherrer, J.-R. (1995). Word segmentation processing : a way to exponentially extend medical dictionaries. *Medinfo*, **8**(pt 1), 28–32.

Lucia, S. (2010). Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, p. 30–39 : Springer.

Maglie, R. (2015). "'can you read this leaflet ?' : User-friendliness of patient information leaflets in the uk and in italy". *ASp. la revue du GERAS*, (68).

Magri-Mourgues, V. (2015). Les noms déverbaux en-ment et le corpus poétique. *Le Français Moderne-Revue de linguistique Française*, (1), 146–164.

Marcoccia, M. (2000). La représentation du nonverbal dans la communication écrite médiatisée par ordinateur. *Communication et organisation*, (18).

Marcoccia, M. (2004). La communication écrite médiatisée par ordinateur : faire du face à face avec de l'écrit. *Journée d'étude sur le traitement automatique des nouvelles formes de communication écrite*, **5**.

Marneffe, M., Maccartney, B. & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, p. 449–454.

Màrquez, L., Carreras, X., Litkowski, K. C. & Stevenson, S. (2008). Semantic role labeling : an introduction to the special issue.

Maynor, N. (1994). The language of electronic mail : Written speech ? *Publication of the American Dialect Society*, **78**(1), 48–54.

McCray, A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.

McCray, A. T., Srinivasan, S. & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 235 : American Medical Informatics Association.

Medero, J. & Ostendorf, M. (2011). Identifying targets for syntactic simplification. In *SLaTE*, p. 69–72.

Meeuwesen, L., Schaap, C. & van der Staak, C. (1991). Verbal analysis of doctor-patient communication. *Social Science Medicine*, **32**(10), 1143 – 1150.

Meinschaefer, J. (2003). Nominalizations of french psychological verbs. *Josep Quer, Jan Schroten, Mauro Scorretti, Petra Sleeman & Els Verheugt (éds.), Selected Papers from Going Romance*, p. 231–246.

Meinschaefer, J. (2005). Deverbal nouns in spanish. *Lingue e linguaggio*, **4**(2), 215–228.

Melčuk, I. A. (1988). *Dependency syntax : theory and practice*. SUNY press.

Mel'cuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot / Aupelf-UREF.

Mertens, P. (2010). Restrictions de sélection et réalisations syntagmatiques dans dicovalence : conversion vers un format utilisable en tal.

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.

Minard, A., Ligozat, A., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P. & Grouin, C. (2011). Hybrid methods for improving information access in clinical documents : concept, assertion, and relation identification. *J Am Med Inform Assoc*, **18**(5), 588–93.

Möhn, D. & Pelka, R. (1984). *Fachsprachen*, volume 30. Niemeyer.

Mortureux, M. (1991). Impersonnel et indéfini dans un discours scientifique. *L'impersonnel ; mécanismes linguistiques et fonctionnement littéraires*, p. 199–206.

Mortureux, M.-F. (1985). Linguistique et vulgarisation scientifique. *Information (International Social Science Council)*, **24**(4), 825–845.

Mortureux, M.-F. (1988). La vulgarisation scientifique : parole médiane ou dédoublée. *Vulgariser la science. Le procès de l'ignorance, Seyssel : Champ Vallon*, p. 118–148.

Mougeon, F. (1995). *Quel français parler : initiation au français parlé au Canada et en France*. Number 3. Éditions du GREF.

Mühlenbock, K. & Johansson Kokkinakis, S. (2009). Lix 68 revisited-an extended readability measure. *Proceedings of Corpus Linguistics 2009*.

Murray, D. E. (1990). Cmc. *English Today*, **6**(03), 42–46.

Nciri, M. (2009). La communication dans la relation médecin-malade. *Espérance Médicale*, **16**(164).

Nelson, J., Perfetti, C., Liben, D. & Liben, M. (2012). Measures of text difficulty : Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.

Nickel, G. (1999). Fachsprachen/languages for special purposes, ein internationales handbuch zur fachsprachenforschung und terminologiewissenschaft/an international handbook of special-language and terminology research 1. *IRAL, International Review of Applied Linguistics in Language Teaching*, **37**(4), 334.

Nilsson, J., Riedel, S. & Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, p. 915–932 : sn.

Nivre, J. (2005). Dependency grammar and dependency parsing. *MSI report*, **5133**(1959), 1–32.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. *www.nlm.nih.gov/mesh/meshhome.html*.

Norman, C. D. & Skinner, H. A. (2006). ehealth literacy : essential skills for consumer health in a networked world. *Journal of medical Internet research*, **8**(2).

Névéol, A., Grosjean, J., Darmoni, S. & Zweigenbaum, P. (2014). Language resources for french in the biomedical domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

Ong, L., de Haes, J., Hoos, A. & Lammes, F. (1995). Doctor-patient communication : A review of the literature. *Social Science Medicine*, **40**(7), 903–918.

Palmer, M., Gildea, D. & Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, **3**(1), 1–103.

Panevová, J. (1974). On verbal frames in functional generative description i. *Prague Bulletin of Mathematical Linguistics*, **22**(6–3), 3–40.

Panevová, J. (1975). On verbal frames in functional generative description ii. *Prague Bulletin of Mathematical Linguistics*, **23**, 17–52.

Parker, R. M. (2006). What an informed patient means for the future of healthcare. *Pharmacoeconomics*, **24**(2), 29–33.

Paroubek, P. (2009). *Rapport sur la seconde campagne d'evaluation, Livrable D16, Projet PASSAGE*. Technical report, LIMSI-CNRS.

Paroubek, P., Vilnat, A., Robba, I. & Ayache, C. (2007). Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. *Actes de TALN*, **2007**.

Patel, V. L., Branch, T. & Arocha, J. F. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, **65**(3), 193–211.

Pearson, J. (1998). *Terms in Context*. Amsterdam/Philadelphia : John Benjamins.

Pecman, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique : thèse...* PhD thesis, Nice.

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

Pendleton, D., Schofield, T., Tate, P. & Havelock, P. (2003). *The new consultation : developing doctor-patient communication*. OUP Oxford.

Petersen, S. E. & Ostendorf, M. (2007). Text simplification for language learners : a corpus analysis. In *SLaTE*, p. 69–72 : Citeseer.

Petkevic, V. (1995). A new formal specification of underlying structures. *Theoretical linguistics*, **21**(1), 1–61.

Peykevič, V. (1988). New dependency based specification of underlying representations of sentences. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, p. 512–514 : Association for Computational Linguistics.

Pimentel, J. (2011). Description de verbes juridiques au moyen de la sémantique des cadres. In *TOTH*.

Poibeau, T. (2001). Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états. In *Actes de la Conférence Française de Traitement Automatique de la Langue,(TALN'2001)*.

Portelance, C. (1991). Fondements linguistiques de la terminologie. *Meta : Journal des traducteursMeta :/Translators' Journal*, **36**(1), 64–70.

Pouliquen, B. (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations*. PhD thesis, Université Rennes 1.

Putz, M. (2008). Approaching linguistic complexity in medical care. *International Journal of Anthropology*, **23**(3-4), 275–284.

Qing, Z.-T., Sergey, G., Hyeoneui, K., Alla, K. & Douglas, R. (2007). Making texts in electronic health records comprehensible to consumers : a prototype translator. In *AMIA*, p. 846–50.

Quirk, C., Menezes, A. & Cherry, C. (2005). Dependency treelet translation : Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 271–279 : Association for Computational Linguistics.

Raman, C., Christine, D. & Bangalore, S. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 1041–1044 : Association for Computational Linguistics.

Rey, A. (1979). *La terminologie : noms et notions*, In P. universitaires de France, Ed., *"Que sais-je ?"*.

Richard, J.-F., Barcenilla, J., Brie, B., Charmet, E., Clement, E. & Reynard, P. (1993). Le traitement de documents administratifs par des populations de bas niveau de formation. *Le Travail Humain*, p. 345–367.

Roelcke, T. (2010). Fachsprachen. 3., neu bearb. *Aufl. Berlin : Erich Schmidt*.

Roelcke, T. (2014). Zur gliederung von fachsprache und fachkommunikation. *Fachsprache*, **37**(3-4), 154–178.

Roger, A. C. & Robboy, S. (1980). Progress in medical information management. systematized nomenclature of medicine (snomed). *JAMA [Internet]*, **8**(243), 756–62. disponible sur : http ://jama.ama-assn.org/cgi/reprint/243/8/756.pdf, consulté le 07/03/2017.

Roter, D. & Hall, J. A. (2006). *Doctors talking with patients/patients talking with doctors : improving communication in medical visits*. Greenwood Publishing Group.

Roter, D. L., Hall, J. A. & Katz, N. R. (1987). Relations between physicians 'behaviors and analogue patients' satisfaction, recall, and impressions. *Medical care*, p. 437–451.

Ruppenhofer, J., Boas, H. C. & Baker, C. F. (2013). The framenet approach to relating syntax and semantics. *Dictionaries. An international encyclopedia of lexicography, volume Supplementary volume : Recent developments with special focus on computational lexicography*, p. 1320–1329.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. & Scheffczyk, J. (2006). *FrameNet II Extended Theory and Practice*. Berkeley, California : International Computer Science Institute. Distributed with the FrameNet data.

Schmidt, T. (2008). *The Kicktionary revisited.*, In A. Rothkegel & J. L. (eds), Eds., *Text Resources and Lexical Knowledge*, p. 239–251.

Schmidt, T. (2009). *The Kictionary-A Multilingual Lexical Resource of Football Language*, In H. Boas, Ed., *Multilingual FrameNets in Computational Lexicography*, p. 101–134.

Schuemie, M. J., Kors, J. A. & Mons, B. (2005). Word sense disambiguation in the biomedical domain : an overview. *Journal of Computational Biology*, **12**(5), 554–565.

Schulte im Walde, S. & Brew, C. (2002). Inducing german semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 223–230 : Association for Computational Linguistics.

Segen, J. C. (1992). *The dictionary of modern medicine*. CRC Press.

Sgall, P., Hajicová, E. & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Sgall, P., Nebeskỳ, L., Goralciková, A. & Hajicová, E. (1969). *A functional approach to syntax : in generative description of language*. JSTOR.

Siddharthan, A. (2002). An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings*, p. 64–71 : IEEE.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language & Computation*, **4**(1), 77–109.

Siddharthan, A. (2014a). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *EACL*, p. 722–731.

Siddharthan, A. (2014b). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

Simoni Aurembou, M. (2000). *Français du Canada-Français de France : Actes du cinquième Colloque international de Bellême du 5 au 7 juin 1997*, volume 13. Walter de Gruyter.

Smith, C. & Wicks, P. (2008). PatientsLikeMe : Consumer health vocabulary as a folksonomy. In *Proceedings of the AMIA 2008 Symposium*, p. 682–686.

Specia, L., Jauhar, S. & Mihalcea, R. (2012). Semeval-2012 task 1 : English lexical simplification. In **SEM 2012*, p. 347–355.

Stein, A., Kocourek, R. & Rey, A. (1992). La langue française de la technique et de la science. vers une linguistique de la langue savante. deuxième édition augmentée, refondue et mise à jour avec une nouvelle bibliographie.

Swales, J. (1990). *Genre analysis : English in academic and research settings*. Cambridge University Press.

Tapas, K. & Orr, D. (2009). Predicting the readability of short web summaries. In *WSDM*, p. 202–211, Barcelona, Spain.

Tartier, A. (2006). Variation terminologique et analyse diachronique. In *TALN*.

Tateisi, Y., Ohta, T. & Tsujii, J. (2004). Annotation of predicate-argument structure on molecular biology text. In Springer, Ed., *In Proceedings of the Workshop on the 1st International Joint Conference on Natural Language Process (IJCNLP*, Hainan Island, China.

Tellier, C. (2008). *Verbes spécialisés en corpus médicale : une méthode de description pour la rédaction d'articles terminologiques*. Thèse de doctorat, Université de Montréal.

Tesnière, L. (1953). *l'Esquisse d'une syntaxe structurale*. Paris : Klincksieck.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris : Klincksieck.

Todirascu, A., Pado, S., Krisch, J., Kisselew, M. & Heid, U. (2012). French and german corpora for audience-based text type classification. In *LREC*, volume 2012, p. 1591–1597.

Tokuhisa, R., Inui, K. & Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, p. 881–888 : Association for Computational Linguistics.

Tran, T. M., Chekroud, H., Thiery, P. & Julienne, A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient ? *Ethica Clinica*, **53**, 34–43.

Vajjala, S. & Meurers, D. (2014). Readability assessment for text simplification : From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, **165**(2), 194–222.

Van Den Eynde, K. & Mertens, P. (2006). Le dictionnaire de valence dicovalence : manuel d'utilisation. *Manuscript, Leuven*.

Veel, R. (1997). Learning how to mean—scientifically speaking : Apprenticeship into scientific discourse in the secondary school. *Genre and institutions : Social processes in the workplace and school*, p. 161–195.

Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca.

Vergne, J. (1999). Etude et modélisation de la syntaxe des langues à l'aide de l'ordinateur, analyse syntaxique automatique non combinatoire. *Habilitation à diriger les recherches*.

Villoing, F. & Namer, F. (2008). «interpréter les noms déverbaux : quelle relation avec la structure argumentale du verbe de base ? le cas des noms en–oir (e) du français. *Actes du Congrès Mondial de Linguistique Française, Paris, 9-12 juillet 2008*, p. 1539–1557.

Vilnat, A., Monceaux, L., Paroubek, P., Robba, I., Gendner, V., Illouz, G. & Jardino, M. (2004). Annoter en constituants pour évaluer des analyseurs syntaxiques. *actes de TALN-04*.

Vor der Brück, T., Hartrumpf, S. & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica*, **32**(4).

Wagner, W., Schmid, H. & Schulte Im Walde, S. (2009). Verb sense disambiguation using a predicate-argument-clustering model. In *In Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts*, p. 23–28.

Waisman, Y., Siegal, N., Chemo, M., Siegal, G., Amir, L., Blachar, Y. & Mimouni, M. (2003). Do parents understand emergency department discharge instructions ? a survey analysis. *IMAJ-RAMAT GAN-*, **5**(8), 567–570.

Wandji, T. O. (2014). Les modèles de description du verbe dans les travaux de linguistique. *Terminologie et TAL, 21ème TAL, Marseilles*, p. 37–48.

Wandji Tchami, O. (2016). Acquiring verb frames for a text simplification lexicon in the medical domain. In *In Proceedings of the 12th Terminology and Knowledge Engineering Conference (TKE)*, Copenhague, Denmark.

Wandji Tchami, O. & Grabar, N. (2014). Towards automatic distinction between specialized and non-specialized occurrences of verbs in medical corpora. In *Proceedings of Computerm*, p. 114–124, Dublin, Ireland.

Wandji Tchami, O., Grabar, N. & Heid, U. (2015). Syntagmatic behaviors of verbs in medical texts : Expert communication vs. forums of patients. In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*, p. 99–106, Universidad de Granada, Granada, Spain.

Wandji Tchami, O., Heid, U. & Grabar, N. (2016). French specialised medical constructions : Lexicographic treatment and corpus coverage in general and specialized dictionaries. In *In proceedings of the XVIIth EURALEX conference*, Tbilissi, Georgia.

Wandji Tchami, O., L'Homme, M. & Grabar, N. (2013). Discovering semantic frames for a contrastive study of verbs in medical corpora. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, Villetaneuse.

Wandji Tchami, O., L'Homme, M. & Grabar, N. (2014). Frame semantics-based study of verbs across medical genres. In *e-Health - For Continuity of Care - Proceedings of MIE2014, the 25th European Medical Informatics Conference, Istanbul, Turkey, August 31 - September 3, 2014*, p. 1075–1079.

White, P., Singleton, A. & Jones, R. (2004). Copying referral letters to patients : the views of patients, patient representatives and doctors. *Patient education and counseling*, **55**(1), 94–98.

Williams, N. & Ogden, J. (2004). The impact of matching the patient's vocabulary : a randomized control trial. *Family Practice*, **21**(6), 630–635.

Williams, S. & Reiter, E. (2005). Generating readable texts for readers with low basic skills.

Wüster, E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In G. R. et H. Felber, Ed., *Textes choisis de terminologie*, volume I. Fondements théoriques de la terminologie, p. 55–114. Université de Laval, Québec : GISTERM. sous la direction de V.I. Siforov.

Wüster, E. (1985). *Introduction to the General Theory of Terminology and Therminological Lexicography*.

Ye, P. & Baldwin, T. (2006). Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler. In *Australasian Language Technology Workshop*, p. 141–148, Sydney, Australia.

Yusuke, S. & Satoshi, S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, p. 65–71 : Association for Computational Linguistics.

Zeng, X. & Parmanto, B. (2003). Evaluation of web accessibility of consumer health information websites. In *AMIA 2003*, p. 743–7.

Zeng-Treiler, Q., Kim, H., Goryachev, S., Keselman, A., Slaugther, L. & Smith, C. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, p. 1117–1121, Brisbane, Australia.

Zeng-Treiler, Q. & Tse, T. (2006). Exploring and developing consumer health vocabularies. *J of Am Med Infor Ass*, **13**, 24–29.

Zeng-Treiler, Q., Tse, T., Divita, G., Keselman, A., Crowell, J. & Browne, A. C. (2006). Exploring lexical forms : first-generation consumer health vocabularies. In *AMIA 2006*, p. 1155–1155.

Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *Journal of biomedical informatics*, **36**(4), 326–333.

Zribi-Hertz, A. (1982). La construction" se-moyen" du français et son statut dans le triangle moyen-passif-réfléchi. *Lingvisticae Investigationes*, **6**(2), 345–401.