Thèse de doctorat de l'Université Paris 6

Spécialité : Informatique médicale

Présentée par : Natalia Grabar

Pour obtenir le grade de Docteur de l'Université Paris 6

Sujet de la thèse :

Terminologie médicale et morphologie.
Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique

Soutenue le 8 décembre 2004

Devant le jury composé de :

Olivier Bodenreider		Examinateur
Bernard Fradin	Directeur de Recherche CNRS	Rapporteur
Benoît Habert	Professeur, Paris X	Rapporteur
Stéfan Darmoni	Professeur, Rouen	Examinateur
Jean-Gabriel Ganascia	Professeur, Paris 6	Examinateur
Monique Slodzian	Professeur, INaLCO, Paris	Examinatrice
Pierre Zweigenbaum	Chercheur, DSI–AP-HP, INaLCO, INSERM	Directeur de thèse

Souvent le terme varie
Fol est qui s'y fie

Remerciements

Je tiens tout d'abord à remercier les chercheurs et les personnes qui ont travaillé et travaillent sur les sujets abordés dans ma thèse. Cette thèse est un *intertexte*, né et enrichi grâce à ces travaux.

Je remercie Monique Slodzian de m'avoir mise sur la 'bonne voie' du Traitement Automatique des Langues et des corpus, et de m'avoir fait confiance dans les travaux sur des projets nationaux et européens.

Je remercie Pierre Zweigenbaum qui a assumé la direction de mon travail de thèse, m'a transféré sa passion, (un peu de) ses connaissances et surtout la rigourosité dans le travail. Avec Pierre, je remercie toute l'équipe du Service d'Informatique Médicale du CHU de la Pitié-Salpêtrière de m'avoir accueillie pendant la durée de ma thèse.

Je remercie l'action 5 du GdR 2220 du CNRS de m'avoir transmis de la possion pour la morphologie.

Pour avoir pu utiliser les ressources terminologiques existantes du domaine médical, j'adresse mes remerciements à :

- le Dr. R.A. Côté pour la version précommerciale du Répertoire d'Anatomopathologie de la SNOMED en français et I. Emelin pour la version en russe ;
- l'INSERM pour la traduction partielle du thesaurus MeSH ;
- et la NLM d'avoir mis à disposition des chercheurs, à travers l'UMLS, les ressources terminologiques et lexicaux du domaine médical.

Je remercie Stefan Darmoni et son équipe pour le cadre de travail qu'il m'a offert avec le portail médical francophone CISMeF et pour l'accès aux documents indexés dans ce portail.

Je remercie les chercheurs pour les outils et les ressources linguistiques mis à ma disposition :

- Didier Bourigault pour l'outil `Lexter` d'acquisition de candidats termes dans les corpus,
- Nathalie Aussenac et Patrick Séguéla pour les patrons lexico-syntaxiques de repérage de relations sémantiques entre les termes,
- Yannick Toussaint pour l'outil d'annotation morphosyntaxique et Pierre Zweigenbaum pour sa transformation en mode `emacs`,
- INaLF pour l'entraînement de l'étiqueteur `Brill` et sa mise à disposition des chercheurs,
- Fiammetta Namer pour le lemmatiseur `Flemm`,
- Thierry Hamon et Didier Bourigault pour les ressources synonymiques de Le Robert,
- les services d'hématologie et de cardiologie du CHU de la Pitié-Salpêtrière pour l'accès aux documents hospitaliers.

Je remercie mes collègues et amis pour les discussions et collaborations ; Thierry Hamon et Pierre Zweigenbaum pour l'aide dans la maîtrise de l'encodage des caractères cyrilliques sous LaTeX ; mes proches pour leur soutien permanent.

Je remercie le service d'hématologie du CHU Henri Mondor pour les conditions presqu'idéales de travail et de rédaction.

Je remercie mes relecteurs, Thierry Hamon, Fiammetta Namer et Pierre Zweigenbaum, pour leurs remarques pertinentes et toujours pas suffisantes.

Mais par-dessus tout, je remercie mes rapporteurs, Bernard Fradin et Benoît Habert, d'avoir accepté cette lourde tâche ; et bien sûr les membres du jury.

Résumé

Notre travail s'articule autour de la variation terminologique et en particulier autour de la variation morphologique des termes.

En premier temps, nous effectuons l'acquisition de données morphologiques à partir des données terminologiques structurées. Notre méthode donne des résultats d'une précision élevée (au-dessus des 90 %) et d'un rappel assez important.

En deuxième temps, nous effectuons deux expériences afin d'évaluer les ressources morphologiques acquises : une application proche de la recherche d'information et la structuration des termes. À côté d'autres normalisations, les ressources morphologiques (paires de mots reliés morphologiquement et les règles de désuffixation de traitement des pluriels réguliers) apportent un gain faible mais non négligeable. De manière générale, différents types de normalisations donnent des résultats complémentaires et les ressources spécifiques du domaine sont plus utiles que les ressources de la langue générale.

Mots clés. Terminologie, variation terminologique, morphologie, médecine, acquisition de ressources morphologiques, productivité morphologique, appariement de termes, structuration de terminologies.

Abstract

In this work we study the terminological variations and particularly variations existing on the morphological level.

We first aim at acquiring morphological data from structured terminologies. Our method gives a very good precision (more than 90%) and an important recall.

We then realize two experiences to evaluate acquired morphological resources : information retrieval and terminology structuring. Beside the other normalisations, the morphology (words pairs and stemming rules for regular plurals) gives a minor but important gain. In a general way, different normalisations give complementary results, and domain specific resources are more usefull that the general language resources.

Key-words. Terminology, terminological variation, morphology, medecine, acquisition of morphological resources, morphological productivity, term matching, terminology structuring.

Table des matières

1 Introduction **1**
1.1 Objectifs du travail 4
1.2 Plan du travail 4

I. État de l'art **6**

2 Variation terminologique **7**
2.1 Types de variation terminologique 7
2.1.1 Techniques d'appariement lexical 8
2.1.2 Données terminologiques ou lexicales structurées 13
2.2 La variation terminologique en contexte 13
2.2.1 Fusion de terminologies 14
2.2.2 Recherche d'information, indexation et encodage 24
2.2.3 Acquisition terminologique 29
2.3 Discussion et conclusion 32
2.4 Précision des objectifs du travail 33

3 Morphologie **35**
3.1 Morphologie, une sous-discipline de la linguistique 36
3.1.1 Objet de la morphologie 36
3.1.2 Matériau de la morphologie 36
3.1.3 Deux sous-domaines de la morphologie 39
3.1.4 Familles morphologiques de mots 47
3.1.5 Pour les cas discutables, définissons quelques frontières... 48
3.2 Morphologie et TAL 54
3.2.1 Étiquetage morphosyntaxique 54
3.2.2 Production de lexiques morphologiques 56
3.2.3 Systèmes d'analyse morphosémantique 57
3.3 Conclusion 58

II. Explorations morphologiques des données textuelles 59

4 Induction de connaissances morphologiques à partir de terminologies structurées 59

4.1 Introduction . . . 60
4.2 Travaux en acquisition automatique de connaissances morphologiques . . . 61
4.2.1 Exploitation de dictionnaires électroniques . . . 61
4.2.2 Exploitation de corpus . . . 62
4.2.3 Exploitation d'un corpus et d'une liste de termes . . . 63
4.2.4 Exploitation des couples de suffixes et de leurs fréquences . . . 64
4.2.5 Exploitation d'approches distributionnelles . . . 64
4.2.6 Application d'algorithmes d'apprentissage . . . 65
4.2.7 Exploitation des règles de formation des lexèmes . . . 65
4.2.8 Bilan . . . 65
4.3 Apprentissage à partir de séries de synonymes . . . 66
4.3.1 Méthode . . . 66
4.3.2 Évaluation des résultats : précision et rappel . . . 68
4.3.3 Matériel . . . 68
4.3.4 Analyse et discussion des ressources morphologiques induites avec les séries de synonymes en trois langues . . . 69
4.3.5 Conclusion et perspectives . . . 76
4.4 Apprentissage à partir de séries de synonymes étiquetées et lemmatisées . . . 77
4.4.1 Adaptation de la méthode . . . 78
4.4.2 Préparation de données étiquetées et lemmatisées . . . 78
4.4.3 Analyse et discussion des ressources morphologiques induites avec des données étiquetées et lemmatisées . . . 79
4.4.4 Conclusion et perspectives . . . 83
4.5 Apprentissage à partir d'autres relations sémantiques encodées dans les terminologies . . . 84
4.5.1 Méthode . . . 84
4.5.2 Matériel . . . 85
4.5.3 Analyse et discussion des ressources morphologiques induites avec diverses relations sémantiques . . . 85
4.5.4 Conclusion et perspectives . . . 88
4.6 Ajustement et maximisation des suffixes . . . 89
4.6.1 Adaptation de la méthode . . . 89
4.6.2 Matériel . . . 89
4.6.3 Analyse et discussion des ressources morphologiques induites avec la maximisation des suffixes . . . 90
4.6.4 Conclusion et perspectives . . . 91
4.7 Bilan de l'acquisition automatique de ressources morphologiques à partir de terminologies structurées et quelques perspectives . . . 92

5 Productivité quantitative des procédés morphologiques de l'adjectivation dénominale 95
5.1 Introduction . . . 95
5.2 Travaux en quantification de la productivité morphologique . . . 96
5.3 Méthode pour l'étude de la productivité de l'adjectivation dénominale en corpus . . . 99
5.4 Collecte et préparation du corpus . . . 99
5.4.1 Le corpus journalistique . . . 100
5.4.2 Le corpus hospitalier . . . 101
5.4.3 Le corpus Web . . . 101
5.4.4 Corpus et occurrences : les corpus à comparer . . . 103
5.5 Lexiques nominal et adjectival . . . 105
5.6 Association semi-automatique des adjectifs dénominaux avec leurs bases . . 105
5.6.1 Collecte semi-automatique des règles d'association . . . 106
5.6.2 Recensement des adjectifs dénominaux . . . 107
5.6.3 Fusion de suffixes 'allomorphiques' en familles . . . 107
5.7 Analyse et discussion du comportement des adjectifs dénominaux en corpus 107
5.7.1 Différences de productivité des procédés d'adjectivation dénominale à l'intérieur des groupes de comparaison . . . 109
5.7.2 Étude des adjectivations 'concurrentes' . . . 115
5.8 Conclusion et perspectives . . . 120

III. Application des ressources morphologiques 122

6 Appariement des requêtes des utilisateurs avec le vocabulaire contrôlé. Étude des logs d'un portail médical français 123
6.1 Introduction . . . 123
6.2 Travaux en analyse des logs . . . 124
6.3 Méthode d'appariement des requêtes des utilisateurs et des termes d'indexation 126
6.3.1 Normalisations au niveau des caractères . . . 127
6.3.2 Normalisations au niveau des mots : application de connaissances morphologiques . . . 127
6.3.3 Correction orthographique : comparaison approximative de mots . . . 128
6.3.4 Enchaînement des normalisations . . . 128
6.3.5 Occurrences et types . . . 128
6.4 Description du matériel . . . 129
6.4.1 Log des requêtes à CISMeF . . . 129
6.4.2 Vocabulaire d'indexation : le MeSH . . . 129
6.4.3 Normalisations morphologiques . . . 130
6.4.4 Mots vides . . . 130
6.5 Analyse et discussion des appariements des vocabulaires . . . 131
6.6 Conclusion et perspectives . . . 134

7 Application de connaissances morphologiques en structuration de terminologies 135
7.1 Introduction . . . 135
7.2 Relations dans les terminologies structurées . . . 136
7.2.1 Relations hiérarchiques . . . 136
7.2.2 Relations synonymiques et antonymiques . . . 137
7.2.3 Relations transversales . . . 138
7.3 Travaux en structuration de terminologies . . . 139
7.3.1 Étude de la structure des termes . . . 139
7.3.2 Étude du contexte des termes . . . 141
7.4 Méthodes d'induction et d'évaluation de relations hiérarchiques . . . 143
7.4.1 Détection d'inclusions lexicales . . . 143
7.4.2 Evaluation par rapport au référentiel existant . . . 144
7.5 Préparation et description du matériel . . . 146
7.5.1 Liste de termes . . . 146
7.5.2 Structuration de référence des termes : structure originale du MeSH 146
7.5.3 Connaissances morphologiques . . . 146
7.5.4 Synonymes . . . 147
7.5.5 Mots vides . . . 149
7.6 Analyse et discussion des relations induites et des termes placés . . . 149
7.6.1 Évolution des relations induites avec les inclusions lexicales . . . 150
7.6.2 Évaluation des inclusions lexicales par rapport au MeSH . . . 152
7.6.3 Analyse manuelle de 'nouvelles' relations . . . 155
7.6.4 Structuration des termes provenant du corpus de la cogénération . . . 159
7.7 Conclusion et perspectives . . . 161

IV. Discussion, conclusion et perspectives 163

8 De l'utilité des ressources morphologiques pour le traitement de la variation terminologique 163
8.1 Complémentarité des différents niveaux de traitement . . . 163
8.2 À chaque variation morphologique ses ressources . . . 164
8.3 À chaque ressource ses heuristiques . . . 166
8.4 Acquisition à partir de terminologies structurées ou de corpus ? . . . 167
8.5 Ressources linguistiques ou règles de désuffixation ? . . . 169
8.6 Ressources spécifiques du domaine ou de la langue générale ? . . . 170

9 Perspectives 171

Références bibliographiques 172

Annexes **191**

A Acquisition de ressources morphologiques à partir de terminologies structurées **191**

A.1 Apprentissage à partir de séries de synonymes 191
A.1.1 Extrait des séries de synonymes du Répertoire d'anatomopathologie (196 séries sur 2 344) 191
A.1.2 Règles (fréquence ≥ 5) 194
A.1.3 Suffixes (fréquence ≥ 5) 197
A.1.4 Préfixes (fréquence ≥ 5) 199
A.1.5 Familles 203
A.2 Apprentissage à partir de séries de synonymes étiquetées et lemmatisées . . 208
A.2.1 Jeu d'étiquettes morphosyntaxiques `Brill` INaLF 208
A.2.2 Règles (fréquence ≥ 5) 209
A.2.3 Suffixes (fréquence ≥ 5) 210
A.2.4 Préfixes (fréquence ≥ 5) 212
A.2.5 Familles 213
A.3 Apprentissage à partir d'autres relations sémantiques encodées dans les terminologies 218
A.3.1 Règles (fréquence ≥ 5) 218
A.3.2 Suffixes (fréquence ≥ 5) 219
A.3.3 Préfixes (fréquence ≥ 5) 220
A.3.4 Familles 222
A.4 Ajustement et maximisation des suffixes 227
A.4.1 Règles (fréquence ≥ 5 : 117 règles sur 501) 227
A.4.2 Suffixes (fréquence ≥ 5 : 94 suffixes sur 582) 227
A.4.3 Préfixes (fréquence ≥ 5 : 54 préfixes sur 1 251) 228

Index **229**

Glossaire **231**

Liste des tableaux

1.1 Exemples de la variabilité des termes dans le domaine médical. 2

2.1 Les termes préférés et les termes synonymes dans la SNOMED. 18
2.2 Les relations hiérarchiques de la SNOMED. 19

3.1 Contraintes catégorielles pour l'affixation. 41
3.2 Contraintes catégorielles avec la conversion. 42
3.3 Contraintes catégorielles des composés populaires. 43
3.4 Contraintes catégorielles des composés savants. 46
3.5 Quelques exemples d'expressions d'origine syntaxique. 49

4.1 Les termes préférés et les termes synonymes dans la SNOMED. 66
4.2 Le rappel pour la flexion et l'affixation en anglais. λ=4 avec notre méthode. 72
4.3 Nombre de séries des termes à l'amorçage : relations sémantiques de la SNOMED ([STD], [HIE], [TRN]) et relations hiérarchiques de la CIM10 [CIM10]. 85
4.4 Quelques exemples de règles avec des suffixes maximisés. 90

5.1 Procédés morphologiques d'adjectivation dénominale retenus. 96
5.2 Taille des corpus initiaux après segmentation, étiquetage et lemmatisation par `TreeTagger` et `Flemm` : types et occurrences des lemmes. 104
5.3 Groupes de comparaison de corpus d'étude. Les corpus dont la taille a été réduite sont suivis du facteur de réduction appliqué ($/N$). 105
5.4 Allomorphes regroupés automatiquement puis complétés manuellement. . . 108
5.5 Nombre de couples {*nom*, *adjectif dérivé*} repérés par les règles de dérivation adjectivale avec la liste de noms de référence dans les différents groupes de comparaison. 109
5.6 Fréquence et spécificité des suffixes concurrents. 116

7.1 Exemples de relations induites avec les inclusions lexicales et se trouvant dans la structure originale du MeSH. 'Indirect' signifie que la distance entre les deux termes est supérieur à 1. 151
7.2 Relations analysées : la tailles des échantillons. 155
7.3 Induction des relations avec les inclusions lexicales et leur validation. . . . 160
7.4 Distribution des relations typées par sous-domaines. 160

Table des figures

4.1 Acquisition de connaissances morphologiques en trois langues : français, anglais, russe. λ=4. 70
4.2 Acquisition de connaissances morphologiques avec des données étiquetées et lemmatisées. λ=4. 80
4.3 Acquisition de connaissances morphologiques avec diverses relations sémantiques encodées dans les terminologies. λ=4. 86

5.1 Positionnement des suffixes d'adjectivation dénominale dans un espace bidimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre langue de spécialité (*web-signes*) et langue générale (*Le Monde*). Échelle logarithmique. 111
5.2 Positionnement des suffixes d'adjectivation dénominale dans un espace bidimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre trois sous-domaines médicaux des documents du Web : hématologie, néphrologie et neurologie. Échelle logarithmique. 112
5.3 Positionnement des suffixes d'adjectivation dénominale dans un espace bidimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre deux sous-domaines médicaux des documents hospitaliers : hématologie et cardiologie. Échelle logarithmique. 113
5.4 Positionnement des suffixes d'adjectivation dénominale dans un espace bidimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre deux genres médicaux du sous-domaine d'hématologie : comptes rendus hospitaliers et documents du Web. Échelle logarithmique. 114

6.1 Évolution des appariements à chaque étape de la normalisation. 131
6.2 Évolution mensuelle des occurrences et des types restant non appariés. . . 133

7.1 Relations taxinomiques dans une terminologie structurée 137
7.2 Relations synonymiques et antonymiques dans une terminologie structurée 138
7.3 Relations transversales dans une terminologie structurée 139
7.4 Illustration des deux méthodes d'évaluation des résultats. À gauche une relation correcte est trouvée dans la structure originale du MeSH. À droite un terme est correctement placé dans la structure originale du MeSH. . . . 145
7.5 Quantification des relations induites et des termes placés. 150

7.6 Utilisation réelle des ressources linguistiques à chaque étape des normalisations. 152
7.7 Évaluation des relations et placements induits : rappel et précision. 153

Chapitre 1

Introduction

La langue médicale dispose d'un vocabulaire très riche. Ses nombreuses terminologies totalisent un nombre impressionnant de termes[1] et saisissent ainsi plusieurs de ses facettes. Mais elles ne peuvent cependant pas prétendre à décrire le vocabulaire médical dans son intégralité. Cela serait, d'ailleurs, impossible à réaliser faute d'une source exhaustive (experts, corpus, etc.). Par ailleurs, d'un corpus ou expert à un autre, il peut exister des variations notables dans le recensement et la dénomination des objets et notions du domaine. Par exemple, pour parler d'une atteinte morphologique (formation de kystes) localisée dans le rein, nous pouvons employer indifféremment les termes *rein kystique* ou *kyste du rein* (exemple 5 du tab. 1.1). La difficulté provient également de l'évolution permanente du vocabulaire. Les terminologies au vocabulaire incomplet constituent une pierre d'achoppement potentielle pour les applications du traitement automatique des langues (TAL). Par exemple, en recherche d'information, la non reconnaissance d'un terme d'indexation dans les documents analysés mène à la production de résultats lacunaires. L'impossibilité d'une couverture intégrale du vocabulaire médical demande de disposer de moyens linguistiques et informatiques qui permettraient d'apparier un terme d'indexation avec les expressions qui apparaissent dans les documents, même si leurs formes ne coïncident pas... À condition, bien sûr, que les moyens mis en œuvre soient adaptés à la variation en question. Et il existe en effet différents types de variation terminologique.

Dans le tab. 1.1 nous en donnons quelques exemples puisés dans différents projets menés dans l'équipe : encodage et indexation des dossiers patient (Zweigenbaum *et al.*, 1995; Blanquet & Zweigenbaum, 1999), indexation et recherche d'information (Darmoni *et al.*, 2001; Zweigenbaum *et al.*, 2001). Dans les deux cas, il s'agit de mettre en correspondance deux ensembles de termes : vocabulaire contrôlé (*terme cible*) et vocabulaire libre (*expression source*). Les expressions sources correspondent aux expressions telles qu'elles apparaissent dans les documents ou bien dans les requêtes des utilisateurs. Les termes cibles sont les termes 'officiels', car recensés par une terminologie. Dans la dernière colonne du tableau nous indiquons les terminologies en question :

[1] L'UMLS version 2004, un produit terminologique qui vise la fusion de terminologies médicales existantes (sec. 2.2.1 page 21), recense ainsi plus d'un million de concepts décrits avec plus de 3 millions de termes anglais et 'seulement' 59 385 termes français.

	Expression source	Terme cible	Terminologie
1	Aspect de leucémie aiguë de type LAM2	2904HM : LAM0, LAM1, LAM2 (leucémie aiguë myéloblastique)	Thesaurus d'hématologie
2	Elle met en évidence une double sténose serrée de l'interventriculaire antérieure distale modérément calcifiée	Sténose de l'artère interventriculaire antérieure	Fiche Menelas
3	interaction entre médicaments et alimentation	INTERACTION ALIMENT-MEDICAMENT	MeSH
4	sténose carotidienne gauche	4331CC : Sténose, obstruction, athérome ou embolie de l'artère carotide, sans infarctus cérébral	Thesaurus de chirurgie cardiaque
5	kyste du rein	REIN KYSTIQUE	MeSH
6	sténose valve aorte	STENOSE AORTIQUE VALVULAIRE	MeSH
7	Il s'agit d'un patient dyslipidémique qui présente une hérédité coronaire et qui est un ancien tabagique	Facteur de risque cardiovasculaire	Fiche Menelas

TAB. 1.1 – Exemples de la variabilité des termes dans le domaine médical.

- Les deux thesaurus, d'hématologie et de chirurgie cardiaque, sont des extensions de la CIM[2] et visent la couverture des spécialités médicales correspondantes.
- Les fiches Menelas ont été élaborées à partir de la CIM dans le cadre du projet Menelas (Zweigenbaum *et al.*, 1995). Elles composent l'ontologie des maladies coronariennes alors constituée.
- Et enfin le MeSH est un thesaurus général du domaine medical utilisé en particulier pour la recherche d'information.

Dans l'exemple 5 déjà cité, la variante *kyste du rein* correspond au terme officiel *rein kystique* enregistré dans la terminologie médicale MeSH. Voyons maintenant d'autres exemples du tableau 1.1 et les différentes variations terminologiques qu'ils présentent :

- Dans les exemples 1 et 2, l'appariement entre les expressions sources et les termes

[2]La CIM, une des terminologies médicales internationales, est surtout utilisée pour l'encodage des dossiers patient (sec. 2.2.1 page 19).

cibles est possible à travers les mots qu'ils ont en commun (*leucémie aiguë*, *LAM2* et *sténose*, *interventriculaire*, *antérieure*), bien que ces mots n'apparaissent pas dans le même ordre, ni dans le même contexte.
- Dans l'exemple 3, l'ordre des mots est également différent. Nous notons de plus la présence de la coordination (*et*), de l'adverbe (*entre*), de la ponctuation (-), mais aussi des formes de mots reliées morphologiquement : variation flexionnelle du nombre {*médicament*, *médicaments*} et suffixation {*aliment*, *alimentation*}.
- Les exemples 4, 5 et 6 présentent également les différences d'ordre des mots, l'apparition des mots grammaticaux (prépositions, articles, adverbes,...) et des mots reliés morphologiquement à travers une suffixation ({*carotide*, *carotidienne*}, {*kyste*, *kystique*}, {*valve*, *valvulaire*}, {*aorte*, *aortique*}).
- Dans le dernier exemple 7, le terme cible *facteur de risque cardiovasculaire* peut être reconnu dans l'expression source à travers *dyslipidémique*, *hérédité coronaire* et *ancien tabagique*, qui sont ses hyponymes (termes plus spécifiques). Mais, pour mener à bien cet appariement, l'information sémantique de ce type doit déjà être encodée, par exemple dans une terminologie structurée.

Certaines des variations du tab. 1.1 peuvent être neutralisées assez facilement : suppression de différences dues à l'ordre des mots, à la ponctuation, aux mots grammaticaux, à la casse ou aux accents. Mais pour d'autres, dues par exemple à la présence de mots reliés morphologiquement, nous avons besoin de ressources linguistiques adéquates. Tandis que la reconnaissance de termes reliés avec des relations hiérarchiques demande de disposer d'une terminologie structurée qui encode l'information nécessaire. Remarquons que les termes présentent rarement un seul type de variation : différentes combinaisons apparaissent dans les exemples donnés.

Le but de notre travail consiste à proposer et à tester des appariements de termes qui présentent des différences formelles. Parmi les variations terminologiques montrées, nous nous intéressons particulièrement au traitement de la variation morphologique. Trois principaux types de relations morphologiques sont habituellement distingués :

- la flexion qui gère les variations de genre, de nombre, de temps, de mode, etc. : {valvule, valvule*s*} ;
- l'affixation qui met en relation les lexèmes appartiennent souvent, mais pas toujours, à des catégories syntaxiques différentes : {valvul(e)*/NOM*, valvul*aire/ADJ*} ;
- la composition qui combine au moins deux éléments lexicaux : {valvul(e), valvulo*pathie*}.

Avec la flexion la sémantique des termes ne varie pas beaucoup. Avec l'affixation elle est également sauvegardée, sinon reste très proche. Tandis qu'avec la composition, la différence sémantique entre le composé et ses éléments peut être importante. Dans des applications automatiques, où, lors du regroupement des variantes de termes, la préservation de leur équivalence sémantique est primordiale, les ressources flexionnelles et dérivationnelles trouvent leur utilisation plus 'naturellement'. La composition doit être utilisée sous contrainte.

1.1 Objectifs du travail

Le traitement de la variation morphologique des termes demande de disposer de ressources morphologiques adéquates. Or, si de telles ressources existent dans certaines langues (anglais, allemand, néerlandais), elles ne sont pas actuellement disponibles en français. La situation est encore plus déficitaire pour les langues de spécialité, la langue médicale dans notre cas. Le premier objectif de notre travail consiste donc à proposer des méthodes pour l'acquisition de ressources morphologiques pour le français médical. Le deuxième objectif consiste à vérifier l'utilité de ces ressources morphologiques dans le traitement de la variation terminologique et de les évaluer. Nous le faisons à travers deux applications :

- recherche d'information : nous cherchons à effectuer l'appariement du vocabulaire 'contrôlé' d'indexation avec les expressions des utilisateurs ;
- structuration des termes avec des relations hiérarchiques : nous cherchons à réduire la variation parmi les termes à structurer.

Dans les deux cas, il s'avère que les ressources morphologiques sont d'une aide importante dans la neutralisation de la variation terminologique. Nos expériences, de même que les travaux d'autres chercheurs, montrent que lors du traitement des variations des termes au niveau morphologique, il est pertinent de prendre en compte des variations plus élémentaires : ordre des mots, ponctuation, mots grammaticaux, différences de casse et d'accent, etc. D'autant plus que les termes combinent souvent ces différentes variations. Il apparaît également que, si les connaissances morphologiques apportent des solutions dans l'appariement des termes, elles ne permettent pas de résoudre tous les problèmes qui se posent. D'autres moyens doivent être mis en œuvre.

1.2 Plan du travail

Notre travail se présente sous forme de chapitres autonomes, bien qu'il existe une dépendance et des renvois entre eux. Chaque chapitre est centré sur une problématique et, souvent, l'expérience qui lui correspond.

La première partie de notre travail est consacrée à l'état de l'art. Nous commençons par présenter plus en détail la variation terminologique (chap. 2), que nous situons dans différents contextes : mise en correspondance des termes recensés dans une terminologie avec des expressions des utilisateurs, mise en correspondance de termes provenant de différentes terminologies d'un domaine, mise en correspondance des termes proposés par différents outils automatiques de constitution de terminologies. Comme nous nous proposons de traiter la variabilité terminologique au niveau morphologique, nous présentons la discipline qu'est la morphologie (chap. 3) : d'abord ses assises théoriques et les objets qu'elle manipule et crée (sec. 3.1), ensuite les travaux en morphologie qui se situent du côté du TAL (sec. 3.2).

La deuxième partie de notre travail est consacrée aux explorations morphologiques de données textuelles. Elles vise à satisfaire notre premier objectif : acquisition de ressources morphologiques. Nous proposons ainsi une méthode pour l'acquisition automatique de

connaissances morphologiques à partir des terminologies structurées (chap.4). Dans le chapitre suivant (chap. 5) nous effectuons une étude sur la productivité des procédés suffixaux d'adjectivation dénominale (par exemple {*kyst(e)*, *kystique*}) dans différents corpus du domaine médical. Une des idées sous-jacentes à cette étude concerne la spécificité des procédés affixationnels aux sous-domaines et genres médicaux, mais également à la langue médicale en général. Nous cherchons alors à contraster la productivité des suffixes dans des corpus qui se différencient par :

- leur niveau de spécialisation (documents hospitaliers destinés aux spécialistes *vs* document provenant du Web et destinés au public non averti) ;
- leurs domaines de spécialité (par exemple, cardiologie, neurologie) ;
- leurs genres (documents hospitaliers *vs* documents du Web).

La troisième partie de notre travail, liée à notre deuxième objectif, vise à montrer l'efficacité des ressources morphologiques acquises pour le traitement de la variation terminologique. Nous présentons ainsi deux expériences dans lesquelles nous utilisons nos ressources. La première expérience met à l'épreuve les connaissances morphologiques dans un contexte proche de la recherche d'information (chap. 6) : nous effectuons une analyse des `logs` du portail médical CISMeF et étudions dans quelle mesure les connaissances morphologiques améliorent l'appariement des mots des requêtes soumises par les utilisateurs et des termes contrôlés utilisés pour l'indexation des documents. La deuxième expérience concerne la structuration de terminologies (chap. 7) : nous implémentons l'hypothèse d'inclusion lexicale pour le repérage de relations hiérarchiques entre les termes. Cette hypothèse stipule qu'il existe une relation hiérarchique entre deux termes si l'un de ces termes est inclus dans l'autre. Le terme inclus est considéré comme hyperonyme, le terme incluant comme hyponyme. Nous utilisons des connaissances linguistiques, dont les ressources morphologiques, dans l'appariement des termes et la réduction de la variation terminologique entre eux.

Dans chaque chapitre nous commençons par poser la problématique qui lui est propre et présentons les travaux antérieurs. Nous décrivons ensuite le matériel utilisé et les méthodes proposées et appliquées. Et enfin, nous analysons les résultats obtenus, dressons les perspectives et concluons.

Dans la dernière partie de notre travail, nous faisons une discussion générale (chap. 8) et donnons quelques perspectives à nos expériences (chap. 9).

Chapitre 2

Variation terminologique

Nous consacrons ce chapitre à la présentation de la variation terminologique. Nous voulons montrer en quoi elle consiste et les moyens mis en œuvre pour la neutraliser. Nous présentons d'abord différents types de la variation terminologique (sec. 2.1), que nous regroupons de manière suivante :

1. ceux qui sont accessibles aux ressources générales de la langue et peuvent être traités avec des *techniques d'appariement lexical* : traitements au niveau des caractères, au niveau morphologique, abstraction de l'ordre des mots, suppression des mots vides, transformations morphosyntaxiques, etc. ;
2. ceux qui exigent des connaissances terminologiques du domaine et doivent ainsi s'appuyer sur les *ressources terminologiques structurées*.

Nous situons ensuite la variation terminologique dans trois contextes applicatifs (sec. 2.2) : (1) fusion de terminologies existantes, (2) appariement de requêtes d'utilisateurs avec des termes contrôlés ou des documents, et (3) appariement de candidats termes proposés par les outils automatiques de constitution de terminologies. Nous présentons alors les traitements automatiques utilisés d'habitude pour le regroupement des variantes des termes.

À la fin du chapitre, nous faisons une discussion des travaux présentés (sec. 2.3) et précisons les objectifs de notre travail (sec. 2.4).

2.1 Types de variation terminologique

Les raisons qui se trouvent à l'origine de la variation terminologique sont, entre autres, les suivantes (Grabar & Hamon, 2004a) :

- Variation régionale. D'un pays à un autre, les dénominations peuvent varier :
  - dans une même langue : *ligne directrice* au Canada *vs recommandations* en France[1],
  - dans des langues différentes : par exemple, la variation sur les noms propres dans la dénomination des maladies : *maladie de Weber-Christian* en français *vs maladie de Pfeifer-Weber-Christian* en allemand (Bodenreider & Zweigenbaum, 2000).

[1]Nous remercions Stéfan Darmoni pour cet exemple.

- Locuteur. D'un locuteur à un autre, les dénominations peuvent également varier :
  - différence d'idiolectes entre des locuteurs spécialistes d'un domaine :
    sténose de l'aorte vs aorte sténotique,
  - différence de spécialisation et de niveaux de langue entre des locuteurs de cultures médicales différentes :
    infarctus du myocarde vs crise cardiaque, *rhagade vs crevasse*.
- Diachronie. La dénomination des objets et notions peut varier d'une époque à l'autre :
  oculiste vs ophtalmologue
  oculiste, formé sur la base latine *oculus* (*œil*), est apparu vers 1503 (Robert, 1993), *ophtalmologue*, formé sur la base grecque *ophthalmos* (*œil*), est apparu vers 1840 (Robert, 1993).
  Notons également qu'à côté de *ophtalmologue*, enregistré par (Robert, 1993), il existe *ophtalmologiste*, enregistré par (Manuila *et al.*, 2001). Les procédés morphologiques, ici les suffixes, ont donc aussi tendance à varier selon les époques.
- Stratégie commerciale. La dénomination peut varier afin de marquer la différence entre des produits similaires mais proposés par des industriels différents :
  airbag vs coussin de sécurité vs coussin gonflable.

Mais ce qui nous intéresse c'est de disposer d'une typologie dictée par les besoins du TAL. Nous cherchons donc à regrouper les variantes terminologiques en fonction des phénomènes linguistiques qu'elles présentent. Nous supposons en même temps que cette typologie sera profitable pour définir les traitements nécessaires à chaque niveau de la variation. Nous en distinguons deux grands types : les variations accessibles à la langue générale et pouvant être traitées avec des techniques d'appariement lexical (sec. 2.1.1) et les variations qui nécessitent l'application de données terminologiques structurées (sec. 2.1.2). Nous présentons notre typologie dans un ordre qui pourrait correspondre à l'ordre croissant de difficulté de traitement. Mais n'importe lequel de ces traitements peut devenir vite complexe en fonction des techniques et ressources utilisées.

2.1.1 Techniques d'appariement lexical

Dans cette section, nous décrivons les variantes qui apparaissent au niveau des caractères (casse, accent, orthographe), de l'ordre des mots[2] dans les termes, de mots vides, de mots morphologiquement proches mais formellement différents, de modifications morphosyntaxiques et d'insertion ou de suppression d'éléments dans la structure des termes. Les exemples proviennent des terminologies qui seront présentées plus loin, des données obtenues lors de nos expériences et des travaux cités.

Niveau des caractères. Les *variations au niveau des caractères* englobent :

- la variation de casse : *anévrisme* - *Anévrisme*,

[2]Lorsque nous utilisons le terme *mot* il l'est dans le sens informatique : il s'agit d'une chaîne de caractères obtenue suite à la segmentation. Quant à la segmentation, elle peut être faite sur les espaces et/ou la ponctuation et/ou les symboles ($, *, etc) et/ou les nombres.

– l'omission d'accents : *anévrisme* - *anevrisme*,
– les variantes orthographiques : *anévrisme* - *anévrysme*,
– les accents erronés : *anévrisme* - *anèvrisme*,
– les fautes de frappe : *anévrisme* - *anévrise*.

Les termes qui présentent des variantes de casse sont les plus faciles à traiter et à apparier : la mise en minuscules ou en majuscules des caractères est aisée. De plus, ce traitement n'apporte que très peu d'ambiguïté et de bruit[3] éventuel. La confusion peut cependant apparaître entre les noms propres et les noms communs :

pompe vs Pompe, *pierre vs Pierre*.

Les variantes d'accentuation ou les accents erronés sont aussi faciles à traiter si l'on cherche à supprimer les accents. Par contre, si le but est leur restauration ou leur rectification, la tâche devient vite difficile. Il s'agit d'une part de trouver les règles (contextuelles) d'apparition des caractères accentués (Zweigenbaum & Grabar, 2002b) et d'autre part de désambiguïser en contexte les mots qui présentent différentes accentuations possibles (Spriet & El-Bèze, 1997; Simard, 1998).

Quant aux variantes orthographiques, elles peuvent être enregistrées a priori, sinon traitées avec des techniques phonémiques ou bien en calculant la distance des chaînes (Levenshtein, 1966). Les fautes de frappe sont trop imprévisibles pour pouvoir être enregistrées à l'avance. On peut les traiter avec les mêmes techniques.

Ordre des mots. L'*ordre des mots*, qui reflète l'organisation syntaxique des termes, constitue une autre source de variation. Nous verrons plus loin que la syntaxe des termes a tendance à varier d'une terminologie à une autre, en fonction de l'application à laquelle une terminologie est destinée. Dans les exemples :

Binders, Abdominal - *Abdominal Binders*
Patches, Eye - *Eye Patches*

l'ordre des mots n'est pas significatif. Il peut être ignoré et il l'est généralement dans les applications comme la recherche d'information ou l'indexation de documents. Les termes sont alors traités comme des *sacs de mots* : leur structure originale n'est plus considéré et les mots sont triés et traités dans l'ordre alphabétique. La ponctuation étant supprimée en même temps, nous obtenons la représentation suivante des exemples donnés plus haut :

Abdominal Binders - *Abdominal Binders*
Eye Patches - *Eye Patches*

Ces sacs de mots sont appariables sans encombre.

Mais « *les mots diversement rangés font un divers sens et les sens diversement rangés font différents effets* » (Pascal, 1662, pensée 66). (McCray *et al.*, 1994) appellent cet effet *venetian blind* - *blind venetian*. Dans ces cas, l'ordre des mots devient significatif et change le sens des termes en question. L'abstraction de l'ordre des mots dans les termes peut donc être cause d'erreurs.

[3]Le *bruit* correspond aux données indésirables et/ou fausses proposées par un système automatique.

'Mots vides'. Les mots dits *mots vides* peuvent également être source de variation des termes. Dans les exemples qui suivent, nous opposons des termes qui omettent les mots vides (articles, prépositions, etc.) à leurs formes 'normales' :

abandon traitement	*abandon du traitement*
carie racine dentaire	*carie de la racine dentaire*
implantation prothese vasculaire	*implantation d'une prothèse vasculaire*

Les articles et les prépositions sont certainement considérés ici comme dénués de sens et mis de côté. Comme pour l'ordre des mots, l'utilisation des mots vides dans les termes semble également être guidée par les applications.

De manière générale, l'ensemble des mots vides regroupe :

- les mots grammaticaux (articles, prépositions, pronoms, adverbes, etc.) qui peuvent être ignorés lors des traitements et ceci pour deux raisons : ils sont polysémiques et très fréquents dans les documents,
- les mots centraux d'un domaine qui sont, de ce fait, également très fréquents dans les documents.

Il va de soi qu'il n'existe pas de liste unique de mots vides : ils sont dépendants des domaines et des applications[4].

Variation morphologique. Les variations de termes se réalisent également au *niveau morphologique*, comme en témoigne l'exemple suivant :

hématome anévrismal vs anévrisme de l'hématome.

Dans cet exemple, *anévrisme* et *anévrismal* sont en relation morphologique (suffixation).

La morphologie permet de prendre en compte les variations des termes dont les mots impliquent des procédés de formation (morphologie constructionnelle) ou de variation (morphologie flexionnelle). Nous en faisons ici une présentation minimale, qui est reprise et développée dans le chapitre suivant.

Les principaux procédés de la morphologie constructionnelle sont l'affixation, la conversion et la composition. Nous mentionnons également la supplétion, car elle correspond à des données différentes et demande des traitements (acquisition, stockage) particuliers du point de vue du TAL. La morphologie constructionnelle manipule des *lexèmes*, unités linguistiques abstraites dans le sens où elles sont dépourvues de marques flexionnelles (Lyons, 1968, p. 198) cité dans (Fradin, 2003). C'est lorsque ces lexèmes sont récupérés par la syntaxe, et éventuellement après avoir subi des variations flexionnelles, qu'ils deviennent des mots ou formes, ou des *mots-formes* (dans la terminologie de (Melčuk, 1993)). Au sein de la morphologie constructionnelle, nous distinguons donc les quatre procédés suivants :

- La conversion traite des lexèmes qui ont la même forme graphique, mais dont les catégories syntaxiques et le sens sont différents :

  {muqueuse/*Adj*, muqueuse/*Nom*}
  - *muqueux/Adj* (*muqueuse/Adj*) : *Qui se rapporte au mucus ou aux muco-sités, qui en contient ou en produit* (Manuila *et al.*, 2001),

[4]Quelques listes de mots vides sont disponibles sur l'Internet, par exemple `http://www.unine.ch/info/clef/`, `http://www.up.univ-mrs.fr/~veronis/data/antidico.txt`.

- *muqueuse/Nom* : *Membrane de revêtement des cavités naturelles de l'organisme, à surface toujours légèrement humide* (Manuila *et al.*, 2001).

Lorsque les données ne sont pas étiquetées morphosyntaxiquement, la conversion est invisible aux outils du TAL.

– L'affixation traite la formation des lexèmes avec des affixes (préfixes et suffixes). Les lexèmes affixés ont souvent une catégorie syntaxique différente de la catégorie de leur base. L'affixe applique une instruction sémantique sur la base. L'exemple qui suit présente la formation d'un adjectif à partir d'un nom (adjectivation dénominale) :

{anévrism(e)*/Nom*, anévrism*al/Adj*}

- *anévrisme* : « *Dilatation au niveau de la paroi d'une artère, apparaissant là où la résistance est diminuée par une lésion, une malformation, un traumatisme...* » (Manuila *et al.*, 2001),

- *anévrismal* : « *Qui se rapporte à un anévrisme, qui en a les caractères.* » (Manuila *et al.*, 2001).

Mais l'affixation permet également de former des lexèmes de la même catégorie syntaxique que leur base :

{*hématomètre/Nom*, *hématométrie/Nom*}.

– La supplétion prend en charge la formation de lexèmes qui présentent des bases sémantiquement équivalentes mais dont les langues d'origine sont différentes. Leurs formes graphiques sont donc également différentes. Les exemples qui suivent correspondent à la formation d'adjectifs dénominaux :

{foie*/Nom*, hépat*ique/Adj*} (*hépatique* : « *Qui se rapporte au foie.* » (Manuila *et al.*, 2001)),

{estomac*/Nom*, gastr*ique/Adj*} (*gastrique* : « *Qui se rapporte à l'estomac.* » (Manuila *et al.*, 2001)).

Nous voyons que, dans ces exemples, l'instruction sémantique des suffixes d'adjectivation dénominale est exactement la même que dans le cas d'affixation sur une base non supplétive *anévrisme*. Les bases supplétives proviennent essentiellement du latin et du grec. Un des moyens des plus sûrs dans le traitement de la supplétion consiste à enregistrer les bases correspondantes *a priori*.

– La composition traite également la formation des lexèmes, mais au moyen de la combinaison d'au moins deux composants. La catégorie syntaxique peut changer et le sens du composé résulte de la combinaison des sens de ses composants :

{anévrisme, anévrismo*rraphie*}

- *anévrisme* : « *Dilatation au niveau de la paroi d'une artère, apparaissant là où la résistance est diminuée par une lésion, une malformation, un traumatisme...* » (Manuila *et al.*, 2001),

- *-rraphie* : « *Suffixe*[5] *d'origine grecque signifiant suture* » (Manuila *et al.*, 2001),

→ *anévrismorraphie* : « *Cure chirurgicale d'un anévrisme, qui consiste*

[5] *-rraphie* est considéré comme un suffixe dans (Manuila *et al.*, 2001), mais il s'agit plutôt d'un élément de composition supplétif.

à ouvrir la poche anévrismale, puis à suturer par l'intérieur les orifices artériels qui y débouchent. » (Manuila *et al.*, 2001).

Le composé peut donc manifester une modification sémantique importante par rapport à ses composants.

À côté de la morphologie constructionnelle, la morphologie flexionnelle s'occupe de l'insertion des lexèmes dans les structures syntaxiques. Elle traite la variation des lexèmes quant à leur genre, nombre, cas, etc. Les exemples qui suivent présentent une variation du nombre et du genre :

{anévrisme, anévrisme*s*}
{anévrism*al*, anévrism*ale*}.

Lors de l'appariement des termes, la préservation de l'équivalence sémantique est primordiale. Les applications automatiques doivent donc privilégier la conversion, l'affixation, la flexion et la supplétion ; la composition devant être soumise à une utilisation contrôlée.

La variation morphologique peut être neutralisée avec des techniques non linguistiques, comme par exemple l'application des règles de désuffixation (Lovins, 1968; Porter, 1980), ou au contraire avec l'utilisation de ressources linguistiques, qui se présentent souvent sous forme de paires de mots reliés morphologiquement, par exemple :

{muqueuse*/Adj*, muqueuse*/Nom*}
{anévrism(e)*/Nom*, anévrism*al/Adj*}
{foie*/Nom*, hépat*ique/Nom*}
{estomac*/Nom*, gastr*ique/Nom*}
{anévrisme, anévrisme*s*}
{anévrism*al*, anévrism*ale*}

On cherche alors à remplacer la flexion par son lemme et le lexème construit par sa base.

Variations morphosyntaxiques. Les *variations morphosyntaxiques* regroupent plusieurs des phénomènes déjà vus. En laissant de côté les variations provenant du niveau des caractères, les variations morphosyntaxiques gèrent l'ordre des mots, leurs formes morphologiques et, en plus, leurs dépendances syntaxiques. Les exemples qui suivent peuvent être pris en charge à ce niveau :

sténose de l'aorte - *aorte sténosée*
kyste du rein - *rein kystique*

Le traitement des variations morpholo-syntaxiques demande des connaissances issues d'une étude linguistique et, plus particulièrement, d'une étude syntaxique des termes (Jacquemin & Tzoukermann, 1999).

Insertion et suppression des éléments. *L'insertion ou la suppression des éléments* dans les termes peut également être une source de variation. En voilà quelques exemples :

Insertion :	*diffraction des rayons X*	*diffraction des neutrons ou des rayons X*
	cellule du sang	*cellule mononucléaire du sang*
Suppression :	*usagers en trafic local*	*usagers*
	usagers en trafic local	*trafic local*

Pour neutraliser cette différence, il faut également disposer de règles qui gèrent les dépendances syntaxiques. Il faut aussi avoir des informations sur les types d'éléments dont l'insertion ou la suppression sauvegarde la sémantique du terme (Jacquemin, 1999), voire recourir au contexte pour le vérifier (Jacques, 2003).

2.1.2 Données terminologiques ou lexicales structurées

Dans cette section, nous présentons les variations qui peuvent être neutralisées grâce à l'utilisation de ressources terminologiques ou lexicales structurées avec des relations synonymiques, hiérarchiques et transversales (pour plus de détails sur la structuration voir la sec. 7.2 page 136).

La synonymie relie des termes ou des lexèmes sémantiquement substituables et constitue une piste dans l'appariement des termes. De telles ressources peuvent provenir :

- de ressources générales, par exemple le Petit Robert : *anévrisme - varice*,
- de ressources spécifiques au domaine traité, par exemple la terminologie médicale SNOMED : *hématome anévrismal - hématome pulsatile*.

La pertinence et l'efficacité de ces deux types de synonymes ne sont pas égales lorsqu'ils sont appliqués aux documents des domaines de spécialité (sec. 7.6.2 page 152).

Les terminologies structurées hiérarchiquement offrent des relations de spécialisation entre les termes. Par exemple, le terme *facteur de risque cardiovasculaire*, de l'exemple 7 du tab. 1.1 page 2, peut être apparié avec ses hyponymes :

dyslipidémie	`est-un`	*facteur de risque cardiovasculaire*
hérédité coronaire	`est-un`	*facteur de risque cardiovasculaire*
ancien tabagique	`est-un`	*facteur de risque cardiovasculaire*

Mais il faut que cette information soit encodée dans la terminologie qui est utilisée. Plus loin dans ce chapitre, nous montrons que le recensement et la structuration des termes peut être dépendante des documents, des domaines et des applications. Il faudrait donc disposer de ressources adaptées au contexte de travail.

Les relations autres que hiérarchiques et synonymiques, que nous appelons relations transversales, peuvent également être utilisées dans l'appariement des termes. L'exemple suivant recourt à une relation de ce type (relation d'association) :

accident vs prevention accident.

Ainsi, en recherche d'information, lorsque les utilisateurs recherchent les documents relatifs à *accident*, ils peuvent trouver également les documents sur la *prévention des accidents*.

2.2 La variation terminologique en contexte

Nous présentons maintenant trois contextes applicatifs d'apparition de la variation terminologique. Mais notons tout de suite qu'elle a beaucoup de potentiel et ne reste pas cantonnée à ces contextes :

1. Fusion des termes 'contrôlés' provenant de différentes terminologies existantes (sec. 2.2.1). L'objectif généralement visé est l'*interopérabilité sémantique* entre ces terminologies.

Dans le domaine médical, où il existe un large éventail de terminologies, le besoin d'apparier leurs termes est assez présent. Or le passage des termes provenant d'une terminologie à ceux d'une autre n'est pas évident.

2. Appariement des requêtes des utilisateurs avec les termes d'indexation ou des documents (indexation, recherche d'information) ou des termes d'indexation avec les documents (encodage des dossiers patient). Nous présentons ces contextes dans la sec. 2.2.2.
3. Dans le domaine de l'acquisition terminologique, appariement de candidats termes fournis par différents outils automatiques (sec. 2.2.3).

Pour chaque contexte, nous présentons les techniques de regroupement des variantes.

2.2.1 Fusion de terminologies

Une terminologie a pour vocation de recenser et d'organiser le vocabulaire d'un domaine de spécialité et donc la connaissance qui y est manipulée. Dans de nombreux secteurs d'activité, les terminologies visent à combler un vide (Zweigenbaum, 2004). Notons par exemple les secteurs où les ressources terminologiques électroniques ne sont pas disponibles et doivent être constituées : télécommunications (Maedche & Staab, 2000), tourisme (Giraldo & Reynaud, 2002), aéronautique (Jeannin & Monceaux, 2003), cogénération (Grabar & Jeannin, 2002). Mais la situation est tout autre dans le domaine médical, où il existe de nombreux produits terminologiques, voir par exemple (Chute, 1995), qui reflètent la multiplicité de besoins existant dans ce domaine (Ingenerf & Giere, 1998; Zweigenbaum, 1999). Nous présentons trois types de terminologies qui correspondent aux trois besoins majeurs du domaine médical. Les besoins étant différents, les produits terminologiques correspondants sont également différents. Leurs appellations, mais surtout les contenus et les structures, varient :

- Un *thesaurus*, par exemple le MeSH, est utilisé pour l'indexation des connaissances médicales et la recherche d'information dans les bases de données et les portails médicaux. Un thesaurus assure l'accès à la littérature scientifique et technique du domaine. Les termes d'un thesaurus sont, le plus souvent, des créations artificielles dont le but est de cerner aussi précisément que possible le sens d'un concept[6].
- Les *nomenclatures*, par exemple la SNOMED, sont utiles dans les traitements informatiques des dossiers patient. Elles recensent les termes 'réels' et permettent ainsi d'accéder aux informations contenues par exemple dans les documents hospitaliers.
- Les *classifications*, par exemple la CIM, sont utilisées pour l'encodage des dossiers patients dans des buts statistiques. À la différence des nomenclatures, les termes

[6]Nous utilisons le terme *concept* pour signifier des notions et des objets d'une manière abstraite. Le terme *terme* se rapporte aux expressions linguistiques qui sont utilisées pour dénommer le concept en question. Sans entrer dans des discussions philosophiques, qui sortiraient de nos compétences, les termes se retrouvent dans les textes, les concepts dans nos cerveaux. Par extension, les produits terminologiques recensent les termes, expressions linguistiques, tandis que les ontologies visent une description plus abstraite d'un domaine à travers le recensement des primitives sémantiques et des règles logiques qui les gèrent.

contenus dans les classifications ne sont pas des créations naturelles, mais correspondent à un métalangage artificiel. Ils servent à spécifier des classes, par exemple de causes de maladies et de mortalité.

Les trois terminologies mentionnées (le MeSH, la SNOMED et la CIM) sont des terminologies 'génériques' : elles visent à couvrir le plus exhaustivement possible le domaine médical et sont utilisées dans de nombreux pays à travers le monde. Mais il existe également des terminologies plus locales, consacrées à un des sous-domaines de la médecine, ou ayant une portée nationale (Zweigenbaum, 2004), ou encore destinées à un corps de métier, par exemple les infirmiers (Bakken, 1999). Signalons aussi que l'existence et l'accessibité des outils automatiques pour la constitution de terminologies stimulent la production d'autres terminologies (Le Moigno *et al.*, 2002). Leur objectif consiste souvent à satisfaire les besoins là où les terminologies existantes, n'étant pas adaptées, échouent. Le choix alors se trouve entre l'ajustement des terminologies existantes et le développement de nouvelles.

L'interopérabilité entre tous ces produits terminologiques n'est pas toujours évidente. Car étant conçus avec des objectifs différents, ils contiennent des informations différentes, et ceci à plusieurs niveaux : concepts biomédicaux recensés, termes ou expressions linguistiques de ces concepts, relations sémantiques entre eux. Il existe pourtant de nombreux cas où l'interopérabilité serait souhaitable (Zweigenbaum, 2004) :

- les données enregistrées pour un patient à l'aide d'une terminologie comme la SNOMED ne sont pas utilisables pour rechercher dans la littérature les articles scientifiques qui traitent des mêmes symptômes mais sont indexés avec le thesaurus MeSH ;
- les données diagnostiques enregistrées pour les études statistiques avec la CIM ne sont pas réutilisables pour noter des effets secondaires de médicaments décrits avec les termes d'autres terminologies.

À travers la présentation d'UMLS, nous verrons les techniques utilisées pour neutraliser la variation et permettre l'interaction entre ces différentes terminologies.

Nous commençons par présenter trois produits terminologiques clés du domaine médical : le MeSH pour l'indexation et la recherche d'information, la SNOMED pour l'informatisation des dossiers patient, et la CIM pour l'encodage et étude statistique des dossiers patient. Pour chacune de ces terminologies, nous présentons d'abord les objectifs qui ont été à leur origine, ensuite la nature des concepts et des termes, et la nature des relations. Nous faisons une présentation assez détaillée de ces produits terminologiques, car nous les utilisons dans nos différentes expériences. Cette présentation devrait également permettre de mieux voir que le contenu de ces terminologies varie en fonction des applications pour lesquelles elles ont été conçues, et de préparer l'introduction d'un dernier produit terminologique : l'UMLS. L'objectif de l'UMLS est justement la fusion et l'intégration de plusieurs terminologies médicales existantes. Les travaux autour d'UMLS illustrent les difficultés qui se posent alors et les techniques d'appariement des terminologies.

Thesaurus MeSH pour la recherche d'information

Contexte et objectifs. Le MeSH (Medical Subject Headings) (NLM, 2001)[7] est un thesaurus conçu dans les années 60 par la NLM (U. S. National Library of Medicine) pour la recherche d'information dans sa base de données biomédicales Medline[8]. Il s'agit d'aider l'indexation des documents biomédicaux (articles et ouvrages scientifiques et techniques). L'indexation est effectuée manuellement par des indexeurs professionnels. Elle est contrôlée et consiste à détecter dans les articles les thèmes principaux abordés et à les décrire avec les termes consignés dans le thesaurus MeSH. Grâce à l'indexation, l'utilisateur peut ensuite identifier les documents qui répondent aux mots clés de sa requête.

Le MeSH peut être considéré comme un thesaurus à grain fin (Zweigenbaum, 1999) : pour indexer et représenter un grand nombre de documents biomédicaux les concepts MeSH doivent assurer une couverture maximale du domaine.

Termes. Étant créés dans le cadre d'une application des sciences de l'information, les termes du MeSH sont des expressions artificielles qui ne correspondent pas toujours aux expressions réellement utilisées dans les documents biomédicaux. Les termes du MeSH sont créés avec le souci de refléter le plus fidèlement possible leur signification et de faciliter ainsi leur utilisation par les indexeurs et les utilisateurs humains. Dans les exemples du MeSH qui vont suivre, les termes anglais sont en caractères minuscules, les termes français, lorsque la traduction existe, sont en caractères majuscules non accentués. Nous remarquons tout d'abord les termes avec des virgules dont la syntaxe ne respecte pas celle des groupes nominaux :

ACTINOMYCETALES, INFECTION
ADDISON, MALADIE
CANAL ARTERIEL, PERSISTANCE
FER, COMPOSES
ABERRATION CHROMOSOMIQUE, ANOMALIES

Il y a également des énumérations qui permettent d'englober des termes d'un niveau hiérarchique inférieur :

DELIRE, DEMENCE, TROUBLES MNESIQUES ET COGNITIFS
HORMONES, SUBSTITUTS HORMONES, ET ANTAGONISTES HORMONAUX

D'autres termes témoignent des omissions d'articles et de prépositions :

ABANDON TRAITEMENT
CARIE RACINE DENTAIRE
IMPLANTATION PROTHESE VASCULAIRE

Une révision importante des termes du MeSH français est en cours à l'occasion du projet VUMeF par l'équipe de l'INSERM qui est responsable de sa maintenance (Darmoni *et al.*, 2003). Elle devrait mener à des termes plus naturels : accentués, casse mixte, avec le respect de la syntaxe, etc.

[7] `http://www.nlm.nih.gov/mesh/meshhome.html`
[8] `www.ncbi.nlm.nih.gov/pubmed`

Relations. Les concepts du MeSH sont structurés. Comme pour les termes, la structure du MeSH vise à organiser les concepts d'une manière aussi claire et intuitive que possible (Nelson *et al.*, 2001). Les termes principaux ou préférés (*main headings*), utilisés en indexation, sont reliés à leurs variantes ou termes équivalents (*entry terms*). Les termes équivalents, de leur côté, pointent sur les *main headings* et en constituent autant de points d'accès. La relation d'équivalence ou de substitution regroupe la synonymie et la quasi-synonymie (Nelson *et al.*, 2001). Par exemple, *Laser Scalpel* et *Laser Knife* sont synonymes et constituent des pointeurs vers *Laser Surgery* (*CHIRURGIE LASER*) (Nelson *et al.*, 2001). Un ensemble de termes équivalents constitue un concept. Chaque concept est représenté par son terme préféré, qui semble correspondre au *main heading*. Les concepts, à leur tour, sont regroupés en une classe de descripteurs (ou un descripteur). Par exemple les concepts suivants forment un descripteur :

Coronary Disease (*VAISSEAUX CORONAIRES, MALADIES*)
Coronary Occlusion
Coronary Stenosis

où le nom du descripteur correspond au concept préféré *Coronary Disease*.

Les relations hiérarchiques sont situées au niveau macroscopique, celui des descripteurs. Chaque descripteur reçoit un identifiant unique qui indique sa spécificité. Le MeSH propose actuellement jusqu'à 9 niveaux hiérarchiques. Chaque niveau est exprimé avec un code alphanumérique ou numérique. Les niveaux sont séparés avec des points. Les relations hiérarchiques incluent indifféremment :

- la relation qui relie un terme générique à ses termes spécifiques (l'*hyponymie*) : A02.633.565 (*FIBRE MUSCULAIRE*) ⇒ A02.633.565.600 (*FIBRE MUSCULAIRE CONTRACTION RAPIDE*),
- la relation qui relie la partie à un tout (la *méronymie*) : A01.456.505 (*FACE*) ⇒ A01.456.505.733 (*NEZ*),
- et la relation considérée comme souhaitable en recherche d'information, car reliant des termes proches (l'*aboutness*) : G03.850.110 (*ACCIDENT*) et G03.850.110.060 (*PREVENTION ACCIDENT*).

Les relations hiérarchiques étendent un terme générique à ses termes spécifiques effectuant une 'explosion' de la requête. La recherche est alors dite basée sur les concepts.

Le MeSH est structuré en 15 axes hiérarchiques : c'est une terminologie multiaxiale. Parmi les axes du MeSH, on trouve : Anatomie (A), Organismes (B), Maladies (C), Sciences biologiques (G), etc. Dans les identifiants hiérarchiques des termes, leur axe est signifié par le premier caractère alphabétique.

Un terme MeSH, et un descripteur, peuvent appartenir à plus d'un axe hiérarchique. Ainsi, le terme *NEZ* est situé à deux endroits dans la structure du MeSH (*ORGANES SENS* et *REGION CORPS*). Dans chaque position, il représente un sens différent car les identifiants hiérarchiques qu'il reçoit et ses termes génériques et spécifiques sont différents (Nelson *et al.*, 2001) :

A09 (*ORGANES SENS*) ⇒ A09.531 (*NEZ*) ⇒ A09.531.940 (*ORGANE VOMERONASAL*)

A01 (*REGION CORPS*) ⇒ A01.456 (*TETE*) ⇒ A01.456.505 (*FACE*) ⇒ A01.456.505.733 (*NEZ*)

Le thesaurus indique également des relations d'association, qui pointent sur des termes avec une valeur définitoire ou bien représentent toute autre relation entre les termes des différents axes. Une relation associative spécifique indique que certaines combinaisons de descripteurs et/ou de termes forment un autre descripteur (Nelson *et al.*, 2001).

Taille. Au début de 2003, le MeSH comportait 21 973 descripteurs et 23 512 *entry terms*. La traduction française du MeSH, assurée par l'INSERM (INSERM, 2000), est une traduction partielle. Dans la version de la fin de 2001, elle comporte environ 19 000 termes et 9 000 synonymes.

Nomenclature SNOMED pour l'informatisation du dossier patient

Contexte. La nomenclature SNOMED (Nomenclature systématique des médecines humaine et vétérinaire) (Côté *et al.*, 1997) a été conçue par le CAP (College of Americain Pathologists à Northfield) en 1965. Sa première version visait la description des lésions anatomo-pathologiques et radiologiques. Elle a ensuite été étendue à toute la médecine.

Objectifs et termes. La SNOMED comporte des termes 'naturels' tels qu'ils apparaissent dans les documents médicaux. Et, malgré l'utilisation de l'abréviation métalangagière *SAI* (« *Sans autre indication* »), la SNOMED est particulièrement adaptée à l'informatisation et aux traitements informatiques du dossier médical.

Relations. Dans la SNOMED, comme dans le MeSH, un concept est formé de termes synonymes. Un code alphanumérique (*Code* dans le tab. 2.1) sert à identifier un concept et ses termes (ils ont tous le même code). La classe des termes indique le terme préféré d'un concept ('01') et ses synonymes ('02', '03' et '05'). Les classes '02' et '03' sont réservées aux substantifs et aux groupes nominaux, la classe '05' aux synonymes adjectivaux ou éléments de composition comme *cardio-*.

Code	Classe	Termes
F-00470	01	*symbiose*
F-00470	02	*commensalisme*
F-00470	05	*symbiotique*
F-00470	05	*commensal*

TAB. 2.1 – Les termes préférés et les termes synonymes dans la SNOMED.

Les concepts sont organisés hiérarchiquement en onze axes sémantiques (Morphologie (M), Topographie (T), Fonction (F), Organismes vivants (L), Diagnostics (D), etc.). C'est une terminologie multiaxiale (Lussier *et al.*, 1998). Les relations hiérarchiques comprennent

l'hyponymie et la méronymie (tab. 2.2). Le premier caractère dans le code alphanumérique indique l'axe sémantique. Les 5 chiffres qui suivent représentent 5 niveaux hiérarchiques. Plus le niveau est haut, plus il y a de zéro en fin de code. Dans les exemples du tableau 2.2 le code alphanumérique reflète la subsomption hiérarchique des concepts :

M-12000 (*fracture, SAI*) `est-une` M-10000 (*blessure, SAI*)
M-12010 (*fracture transverse*) `est-une` M-12000 (*fracture, SAI*)
M-12020 (*fracture oblique*) `est-une` M-12000 (*fracture, SAI*)
M-12200 (*fracture ouverte, SAI*) `est-une` M-12000 (*fracture, SAI*)

Code	Classe	Termes
hyponymie		
M-10000	01	*blessure, SAI*
M-12000	01	*fracture, SAI*
M-12010	01	*fracture transverse*
M-12020	01	*fracture oblique*
M-12200	01	*fracture ouverte, SAI*
méronymie		
T-20000	01	*appareil respiratoire, SAI*
T-21000	01	*nez, SAI*
T-21010	01	*muqueuse nasale*
T-21100	01	*partie externe du nez*

TAB. 2.2 – Les relations hiérarchiques de la SNOMED.

Certains concepts de l'axe Diagnostics de la SNOMED pointent, avec des relations de référence ou transversales, vers des concepts 'plus élémentaires'. Ces concepts se trouvent dans les axes autres que Diagnostics, par exemple :

D3-40220 (*thrombose pulmonaire*) → T-44000 (*artère pulmonaire, SAI*)
↘ M-35100 (*thrombus, SAI*)

Dans la SNOMED les relations transversales sont définitoires et compositionnelles (Spackman & Campbell, 1998; Lussier *et al.*, 1998; Zweigenbaum, 1999). Nous détaillons ce point dans la sec. 7.2.3 page 138.

Taille. La SNOMED comporte actuellement 109 023 concepts (164 180 libellés). Dans nos travaux, nous utilisons la version précommerciale du Répertoire d'anatomopathologie en français, qui nous a été aimablement donnée par le Dr. R. A. Côté. Elle contient 9 098 concepts (12 554 libellés), ce qui représente environ 10 % de l'ensemble.

Classification CIM pour l'encodage des dossiers patient et leur étude statistique

Contexte et objectifs. La CIM (Classification internationale des maladies et des problèmes de santé connexes) (OMS, 1995) apparaît au XIX^{e} siècle, faisant suite aux travaux

d'établissement des causes de mortalité et de morbidité qui débutent à partir du 16e siècle. En 1946, l'OMS (Organisation Mondiale de la Santé) a repris le maintien de la CIM en vue de son utilisation à l'échelle mondiale. Actuellement c'est la version 10 de la CIM (CIM-10) qui est utilisée. La CIM est étroitement liée aux systèmes de santé publique des pays du monde. Elle permet l'analyse, l'interprétation et la comparaison des données de mortalité et de morbidité (OMS, 1995, p. 2). Grâce à ces études statistiques, la CIM permet d'assister la lutte contre les maladies, de rompre l'enchaînement des phénomènes morbides ou de mettre en route un traitement efficace (OMS, 1995, p. 33).

En France, la CIM est utilisée pour l'encodage des dossiers patient dans le cadre des études statistiques réalisées pour le PMSI (Programme de médicalisation du système d'information) : enregistrement et comparaison des activités hospitalières des différents hôpitaux, allocation budgétaire de différents hôpitaux et services.

Termes. Les termes, appelés aussi rubriques, sont souvent des expressions d'un métalangage plutôt que des expressions que l'on trouve réellement dans les documents hospitaliers. La CIM regroupe les termes en concepts ou classes. Ils constituent des instructions qui guident le choix d'une classe à laquelle affecter un dossier patient donné (Zweigenbaum, 1999). Chaque classe est identifiée avec un code alphanumérique.

Les termes comportent plusieurs conventions de lecture dont (OMS, 1995, p. 24–27) :

- entre les parenthèses, sont indiqués des qualifieurs facultatifs, qui ne modifient pas le sens : K510 (*Entéro-colite ulcéreuse (chronique)*) ;
- entre les crochets, sont indiqués les synonymes :
  K51 (*Recto-colite hémorragique [colite ulcéreuse]*) ;
- *SAI* signifie « *Sans autre indication* » ou « *Non précisé* » ;
- *NCA* signifie « *Non classé ailleurs* » ;
- *Autre* signifie que la maladie ne peut pas être classée dans les classes qui précèdent :
  A80 (*Poliomyélite aiguë*)
  A800 (*Poliomyélite paralytique aiguë, associée au virus vaccinal*)
  A803 (*Poliomyélites paralytiques aiguës, autres et sans précision*)

Relations. La CIM est une terminologie monoaxiale. Les concepts sont classés suivant le siège anatomique des maladies (*Maladies cardio-vasculaires*, *Troubles mentaux et du comportement*) ou bien suivant leur étiologie ou cause (*Maladies infectieuses*, *Tumeurs*). Le principe de classification par siège anatomique a été adopté au 19e siecle. La CIM est divisée en 21 chapitres. Le premier caractère du code est une lettre associée, sauf quelques exceptions, à un chapitre donné (OMS, 1995, p. 15), par exemple :

A00-B99 (*Maladies infectueuses et parasitaires*)
E50-E90 (*Troubles mentaux et du comportement*)
G00-G99 (*Maladies du système nerveux*)

La spécification des concepts est signifiée par la spécification des codes alphanumériques : plus le code est long plus le concept est spécifique :

A80 (*Poliomyélite aiguë*)
A800 (*Poliomyélite paralytique aiguë, associée au virus vaccinal*)
A801 (*Poliomyélite paralytique aiguë, virus sauvage importé*)
ou bien par l'adjonction d'autres chiffres adoptées pour identifier différents sièges ou variétés supplémentaires (OMS, 1995, p. 17).

Taille. La version anglaise du volume analytique (structuré) de la CIM comporte 12 318 classes (13 505 libellés), sa traduction française 10 800 classes (9 412 libellés). Il existe également un index alphabétique de la CIM qui recense des entrées supplémentaires.

UMLS ou Unification de terminologies

Contexte et objectifs. L'UMLS (Unified Medical Language System) est un projet lancé par la NLM dans le but de faciliter la recherche et l'intégration des informations provenant de différentes terminologies biomédicales (NLM, 2003). Ce besoin a été ressenti face au volume grandissant d'informations et à la difficulté d'intégration de ces informations du fait de la différence du vocabulaire et de la structuration des différentes terminologies (NLM, 2003, p. 10). La première version de l'UMLS, parue en 1988, s'est constamment enrichie de données provenant de nouvelles terminologies. En 2003, elles sont au nombre de 100 en 15 langues. L'UMLS est composé de quatre parties :

1. Le *Metathesaurus* enregistre les termes des concepts recensés,
2. Le *Réseau sémantique* établit un ensemble hiérarchisé de 'types sémantiques' généraux pour les concepts,
3. Le *Specialist Lexicon* présente les mots des termes et fournit des programmes pour les traitements informatiques,
4. Les *Sources* indiquent et décrivent les terminologies source.

Termes. Les termes de l'UMLS proviennent des terminologies source, dont le MeSH, la SNOMED et la CIM. Ils sont recensés dans le Metathesaurus et reçoivent un identifiant unique. L'UMLS enregistre toutes les informations (relations, définitions, etc.) sur ces termes. Il suffit, par exemple, que deux termes soient reliés avec une relation quelconque dans une de ces terminologies pour que cette information apparaisse dans l'UMLS. Pour chaque terme, ses terminologies source sont également mémorisées. Lors de l'utilisation d'UMLS, en sélectionnant une source donnée, il devient possible d'extraire le vocabulaire provenant de cette source et toutes les informations qui s'y rapportent.

Relations. Les termes équivalents provenant de terminologies source sont regroupés en un même concept (NLM, 2003, p. 14). Chaque concept reçoit un identifiant unique. Selon la présence et sa signification dans les terminologies source, un terme peut être lié un ou plusieurs concepts.

Lorsque les termes ont des graphies identiques, la fusion des terminologies s'amorce sans encombre. Mais lorsque les graphies ne sont pas identiques, leur appariement devient

possible grâce à l'utilisation des ressources du Specialist Lexicon, qui permettent de traiter les variations suivantes (McCray *et al.*, 1994) :

- Pour la normalisation de la casse les majuscules sont mises en minuscules :
  ANOMALIES CONGENITAL MULTIPLE vs Anomalies congenital multiple
- La ponctuation (tiret, espace, virgule, etc.) est supprimée :
  Nerve, Abducens vs Nerve (Abducens)
- Certaines des variantes orthographiques sont enregistrées dans les tables de ressources lexicales du Specialist Lexicon :
  Abdominal pain vs Abd. pain
- La morphologie flexionnelle est gérée avec des règles régulières de flexion et des tables d'exceptions. Par exemple, la mise au pluriel régulier d'une forme en *-y* :
  Batteries vs battery
- L'affixation est gérée avec des tables de paires de mots du Specialist Lexicon :
  ABDOMEN PAIN vs Abdominal Pains
- La supplétion est également gérée avec des tables du Specialist Lexicon :
  Renal disease vs Kidney diseases
- L'ordre des mots n'est pas pris en compte :
  Binders, Abdominal vs Abdominal Binders
  Patches, Eye vs Eye patches
- Et enfin, les mots vides sont supprimés :
  splenic artery aneurysm vs Aneurysm of splenic artery

Les appariements lexicaux (des termes anglais) de ce type sont donc possibles à travers les ressources lexicales du Specialist Lexicon d'UMLS, ses programmes de normalisation (`norm`), ses index normalisés et ses programmes de traitements morphologiques (`lvg`).

En plus des appariements lexicaux des termes, les relations sémantiques des terminologies source sont également exploitées : la signification d'un terme est alors donnée par ses synonymes, ses termes génériques ou spécifiques, mais également par tous les termes avec lesquels il est relié (NLM, 2003, p. 55). Les méthodes de cet appariement sont centrées en particulier sur la recherche de synonymes et de *co-hyponymes* (termes qui ont le même hyperonyme et qui peuvent, dans certains cas, être substituables) (Bodenreider *et al.*, 1998). Mentionnons également des travaux qui exploitent les terminologies structurées pour le calcul de la distance sémantique entre termes (Degoulet *et al.*, 1998; Bousquet *et al.*, 2001).

Les relations sémantiques de chaque terminologie source s'enrichissent donc mutuellement : de nouvelles relations s'ajoutent et d'autres, déjà existantes mais sous-spécifiées, sont typées de manière plus précise. Par exemple, il existe dans le MeSH une relation associative sous-spécifiée entre les termes *Atrial Fibrillation* et *Arrhythmia*. La nature de cette relation a pu être précisée pendant la constitution de l'UMLS (NLM, 2003, p. 17) :

Atrial Fibrillation `est-un` *Arrhythmia*

Mais l'accumulation d'informations sémantiques provenant de différentes terminologies peut apporter des ambiguïtés et des inexactitudes. (Bodenreider, 2001) étudie par exemple les relations hiérarchiques circulaires dans l'UMLS : lorsqu'un concept peut être son propre fils ou un descendant quelconque de lui-même. Parmi les causes principales de cette cir-

cularité l'auteur indique les différences de granularité des terminologies, de spécificité des termes, de choix des relations qui sont considérées comme hiérarchiques. (Cimino, 2001) étudie les cas d'ambiguïté et de redondance : lorsqu'un concept est assigné à des types sémantiques considérés comme exclusifs, il est ambigu ; lorsque les termes équivalents n'ont pas pu être appariés et les concepts fusionnés, ces concepts sont redondants.

Les concepts du Metathesaurus sont liés à un réseau sémantique, dont le but est de catégoriser les concepts et de fournir des relations qui existent entre eux. Le réseau sémantique se compose de types sémantiques et de relations. Dans la version de 2003, le réseau contient 135 types et 54 relations sémantiques autres que la synonymie (NLM, 2003, p. 55). Chaque concept est assigné à au moins un type sémantique. Les types et les relations sémantiques sont eux-mêmes structurés en une hiérarchie (NLM, 2003, p. 66–68). Ces couches superposées de structures rapprochent l'UMLS des ontologies ou du moins le situent entre les terminologies et les ontologies (Zweigenbaum, 2004).

Taille. Dans la version de 2003, le Metathesaurus comporte 875 255 concepts (2 140 000 termes) en anglais, dont 23 966 concepts (34 630 termes) sont en français. L'UMLS est disponible sous licence gratuite auprès de la NLM. Certaines terminologies source imposent des contraintes supplémentaires pour leur utilisation.

Conclusion et Discussion

Parmi les nombreuses terminologies existant dans le domaine médical, nous en avons présenté quatre. Chacune d'elles a été conçue avec des objectifs et dans des cadres applicatifs différents. Nous avons vu que les cadres applicatifs conditionnent directement le contenu de ces terminologies.

Les concepts recrutés ne sont pas les mêmes. La CIM recense les causes de maladies et de mortalité. L'exhaustivité de ces causes est primordiale. En 2003, par exemple, la CIM a été enrichie du terme *Pneumonie atypique*. Notons néanmoins que sa couverture laisse à désirer et que de nombreux produits terminologiques spécifiques à différentes spécialités médicales sont créés pour compléter la CIM. Le MeSH, par contre, cherche à indexer les articles et les ouvrages scientifiques. Avec le MeSH, le champ d'exploration est large : il s'agit de la description des spécialités médicales mais également de la formation et de la recherche scientifique. La SNOMED recense également les diagnostics, et tout particulièrement les signes et symptômes. Mais d'une part, elle se concentre sur la terminologie des dossiers patient. Et d'autre part, elle recense aussi des procédures chirurgicales, la description anatomique du corps humain, etc. On ne peut donc pas espérer avoir un recouvrement sémantique complet entre ces terminologies. En effet, (Wang *et al.*, 2001) rapportent que le recouvrement de la SNOMED RT (un dérivé de la SNOMED) et de Clinical Terms V3 (une norme britannique), à travers les relations de synonymie et d'hyperonymie, est d'environ 28 %.

Les termes varient aussi : le MeSH et la CIM les créent artificiellement afin de les rendre les plus explicites possible ; la SNOMED les recense dans les documents médicaux.

Les principes d'établissement des relations sémantiques sont également différents. La synonymie peut englober des expressions du même type syntaxique ou non (la SNOMED enregistre la synonymie entre les noms et les adjectifs). La relation hiérarchique correspond dans certains cas à une relation `est-un`, dans d'autres elle peut englober la relation partitive (comme dans la SNOMED), des relations d'association (comme dans le MeSH) ou encore des relations sous-spécifiées.

Malgré ces différences, il existe un besoin réel de mettre en correspondance les termes de ces différentes terminologies. Nous avons alors décrit, à travers la présentation de l'UMLS, les techniques d'appariement des informations contenues dans les différentes terminologies. L'appariement devient partiellement possible avec les ressources et programmes lexicaux fournis par l'UMLS et la structure des terminologies source.

Les variations au niveau morphologique sont gérées à travers les programmes lexicaux `lvg` du Specialist Lexicon. Pour les variations flexionnelles, ces programmes utilisent des règles flexionnelles régulières et des tables d'exceptions. Pour l'affixation, ils utilisent des tables de paires de mots, y compris l'affixation sur des bases supplétives. La composition n'est pas traitée par `lvg`.

2.2.2 Recherche d'information, indexation et encodage

Nous nous situons maintenant dans un contexte proche de la recherche d'information ou de l'indexation et de l'encodage des documents. Ce contexte n'est pas moins important que le précédent, puisqu'il concerne l'usage effectif qui est fait des termes dans un domaine et implique les usagers et les acteurs de ce domaine. L'appariement des termes dans ce contexte suit un modèle semblable à celui de la section précédente : l'ordre des mots, la casse, la ponctuation, les accents sont d'habitude ignorés. Les variations du niveau morphologique sont généralement prises en compte à travers l'application de règles de désuffixation ou l'utilisation de ressources motivées linguistiquement. La proximité sémantique des termes (synonymie, hyperonymie, etc.) est détectée à travers des données structurées[9]. Nous consacrons cette section aux traitements morphologiques.

Pour commencer, nous présentons les expériences effectuées avec des formes non traitées ('brutes') des termes. Nous rapportons alors les travaux qui appliquent des traitements morphologiques. Nous montrons ensuite qu'il existe une controverse autour de l'utilité des traitements morphologiques dans ce contexte.

Comparaison des formes 'brutes'

Quel est le succès d'appariement des termes d'indexation et des requêtes des utilisateurs lorsque aucun traitement lexical n'est effectué ? Il n'est pas élevé, comme le rapportent (Furnas *et al.*, 1987) à travers des expériences dans une base documentaire. Les auteurs demandent aux utilisateurs :

– de deviner des noms de commandes d'un éditeur de texte ;

[9]Voir par exemple les communications du workshop *Using semantics for information retrieval and filtering* de LREC 2002.

– d'indexer des recettes de cuisine ;
– de donner un hyperonyme de mots de la langue courante ;
– et de donner des synonymes pour des mots de la langue courante.

Les propositions des utilisateurs sont ensuite comparées avec les termes d'indexation 'officiels'. Leur correspondance est évaluée selon quatre protocoles :

– seulement le premier mot clé proposé par l'utilisateur est accepté ;
– le mot clé le plus fréquent proposé par les utilisateurs est accepté ;
– les trois mots clés les plus fréquents proposés par les utilisateurs sont acceptés ;
– les 15 mots clés les plus fréquents proposés par les utilisateurs sont acceptés.

Ce travail montre qu'avec un seul mot clé l'appariement est possible dans 7 à 18 % des cas, avec le mot clé le plus fréquent il varie entre 16 à 36 %, avec les 3 mots clés les plus fréquents entre 38 et 67 %, et enfin avec 15 mots clés les plus fréquents entre 60 et 80 %. Donc, un plus grand choix laissé aux utilisateurs augmente leur possibilité de trouver le bon mot clé et les documents recherchés. L'étendue de cette réussite est grand : 7 à 18 %, 16 à 36 %, etc. Les valeurs les plus élevées sont obtenues avec les recettes culinaires, et les moins élevées avec les commandes informatiques.

Une expérience similaire, mais dans l'indexation et la recherche d'images d'actions, montre un appariement plus réussi qui va de 76 à 90 % (Turner, 1995).

Appariements à travers des traitements lexicaux

Deux modèles sont utilisés pour le traitement de la variation morphologique des termes : la réduction des variantes morphologiques avec des règles de désuffixation ou, plus rarement, avec des ressources morphologiques.

Application de règles de désuffixation. L'application de règles de désuffixation est effectuée à travers des outils appelés désuffixeur ou *stemmer*. Les *désuffixeurs* permettent de réduire les variantes morphologiques des mots à leur *stem*, qui correspond, dans l'idéal, à une base ou bien à un lemme, selon que l'affixation et/ou la flexion sont traitées. Les règles sont établies manuellement et couvrent un ensemble d'affixes à supprimer ou à remplacer. Les premiers désuffixeurs, devenus standards, sont ceux de Lovins (Lovins, 1968) et Porter (Porter, 1980). Tous les deux traitent la flexion et l'affixation des mots anglais.

Le désuffixeur de (Lovins, 1968) utilise une liste d'environ 250 suffixes. Après la suppression du suffixe le plus long d'un mot, l'algorithme vérifie si la chaîne de caractères restante contient au moins 3 caractères et, si oui, recompose une finale du mot[10] [11] :

magnesia	$=(ia\|\epsilon)\Rightarrow$	*magnes*	$=(s\|s)\Rightarrow$	*magnes*
magnesite	$=(ite\|\epsilon)\Rightarrow$	*magnes*	$=(s\|s)\Rightarrow$	*magnes*
magnetize	$=(ize\|\epsilon)\Rightarrow$	*magnet*	$=(t\|t)\Rightarrow$	*magnet*
magnetometric	$=(ic\|\epsilon)\Rightarrow$	*magnetometr*	$=(tr\|ter)\Rightarrow$	*magnetometer*
magnetometry	$=(y\|\epsilon)\Rightarrow$	*magnetometr*	$=(tr\|ter)\Rightarrow$	*magnetometer*

[10] Ces exemples ont été reconstitués selon le cours *Indexation et recherche d'information* assuré par Ch. Jacquemin à Paris 11 (disponible sur `www.limsi.fr/individu/jacquemin/`).

[11] Dans les exemples qui suivent, ϵ représente une chaîne vide.

L'algorithme de (Porter, 1980) utilise une liste d'environ 1 200 suffixes 'minimaux' gérés avec des règles contextuelles. Il a un fonctionnement itératif et prévoit jusqu'à 5 suppressions de suffixes dans un mot :

formalize	$=(alize\|al)\Rightarrow$	*formal*	$=(al\|\epsilon)\Rightarrow$	*form*
operator	$=(ator\|ate)\Rightarrow$	*operate*	$=(ate\|\epsilon)\Rightarrow$	*oper*
feudalism	$=(alism\|al)\Rightarrow$	*feudal*	$=(al\|\epsilon)\Rightarrow$	*feud*

Cet algorithme, initialement appliqué à l'anglais, a donné lieu au développement d'une série de désuffixeurs *Snowball*[12] pour d'autres langues (Porter, 2001), y compris pour le français. C'est sans doute le désuffixeur le plus utilisé.

Discussion. Comme le montrent les exemples de désuffixation avec Lovins et Porter, les chaînes de caractères obtenues après la désuffixation ne correspondent pas toujours à des mots de la langue. Ce qui peut générer du bruit lorsque les stems ne regroupent pas que des mots reliés sémantiquement. Par exemple (Porter, 1980) regroupe ensemble les mots *general*, *generous*, *generation* et *generic*, puisqu'ils sont tous réduits à *gener* (Hull, 1996). Par ailleurs, le non regroupement de mots reliés sémantiquement crée des silences[13]. De ce point de vue, les algorithmes de désuffixation ne sont que relativement fiables.

L'application des désuffixeurs a donné des résultats contradictoires. (Harman, 1991), cité dans (Hull, 1996; Kraaij & Pohlmann, 1996), compare trois désuffixeurs en anglais (suppression de *-s* final, Lovins et Porter) avec la *baseline* (la non-application des traitements, ici de la désuffixation). La *baseline* s'avère alors meilleure que n'importe lequel des désuffixeurs. Alors que dans d'autres travaux, des versions adaptées des désuffixeurs s'avèrent être meilleures que la *baseline* (Krovetz, 1993; Gaussier *et al.*, 2000).

Face à ces résultats contradictoires, (Church, 1995) effectue une étude probabiliste de la corrélation entre les formes de mots passibles d'être traitées par les désuffixeurs. Selon son hypothèse, deux mots sont d'autant plus corrélés et donc substituables, qu'ils apparaissent plus souvent dans les mêmes documents. Par conséquence, plus ils sont substituables, plus l'application de la désuffixation est justifiée. Selon cette hypothèse :

- lorsque la corrélation entre deux mots est égale à 1 il s'agit bien d'un même mot, ce qui justifie la désuffixation ;
- lorsque la corrélation est égale à 0 les deux mots sont bien distincts et l'affixation n'a pas lieu d'être appliquée : c'est la *baseline*.

L'expérience est faite avec 999 paires des pluriels réguliers en *-s* : {*hostage*, *hostages*}. Elle montre que la corrélation est toujours supérieure à 0 (il ne s'agit pas de deux mots différents), mais reste toujours inférieure à 1 (il ne s'agit donc non plus d'un seul mot). En réalité, la corrélation ne dépasse pas le seuil de 0,7. Cette corrélation suggère qu'il ne s'agit pas d'un mot ni de deux mais d'environ 1,5. L'utilisation des désuffixeurs n'est donc justifiée qu'à moitié. Parmi les meilleures corrélations se trouvent : 0,52 pour {*hostage*, *hostages*}, 0,43 pour {*drug*, *drugs*}. La même expérience, appliquée aux variantes de casse (premier caractère en majuscule), montre des résultats similaires (meilleures corrélations : 0,65 pour

[12] `http://snowball.tartarus.org/`

[13] Le *silence* correspond aux résultats désirables mais non fournis par un système.

{*Hurricane*, *hurricane*}, 0,44 pour {*Museum*, *museum*}). Les mots affixés (adverbes en *-ly* à partir des adjectifs) montrent des corrélations plus faibles (0,24 pour {*sexual*, *sexually*}, 0,15 pour {*illegal*, *illegally*}). L'auteur signale également que, parmi les paires de mots étudiées et en fonction des corpus, certaines paires sont mieux corrélées que d'autres. Ainsi selon les corpus, un même mot devrait ou ne devrait pas être désuffixé. En conséquence, lorsque la désuffixation est utilisée, il est nécessaire de l'adapter aux corpus.

La notion de substituabilité de cette expérience est sujette aux questionnements : la distribution de deux mots dans les mêmes documents est-elle suffisante pour qu'ils soient considérés comme substituables ? Toutefois, cette expérience montre des résultats plus optimistes que ceux de (Harman, 1991) : la désuffixation devient fondée, surtout lorsqu'elle est adaptée aux corpus (et donc aux domaines) traités.

Utilisation de connaissances motivées linguistiquement. Toujours face aux résultats contradictoires fournis par les algorithmes de désuffixation, d'autres chercheurs ont choisi de constituer des ressources morphologiques adaptées aux corpus et de les utiliser ensuite pour réduire les variations morphologiques. Ces ressources consistent souvent en un ensemble de paires de mots reliés morphologiquement :

{*journal*, *journals*}
{*journal*, *journalistic*}
{*journalistic*, *journalistically*}

Elles sont comparées avec les désuffixeurs et la *baseline*. Deux études (Hull, 1996; Kraaij & Pohlmann, 1996) font une telle comparaison pour l'anglais et le néerlandais respectivement.

Dans (Hull, 1996), la *baseline* est opposée aux trois désuffixeurs (suppression de *-s* final, Lovins étendu et Porter) et à environ un million de paires de mots en relation morphologique. Ces paires sont générées avec des règles de flexion et d'affixation (préfixation et suffixation) à partir de quelques 77 000 formes de base. Il s'agit de ressources linguistiques de Xerox. Les désuffixeurs et les ressources morphologiques sont utilisés à travers l'outil de recherche SMART[14] sur le corpus et les requêtes de TREC (Text REtrieval Conference). Le calcul du rappel et de la précision[15] montre que la *baseline* produit les résultats les moins bons et que les désuffixeurs sont légèrement meilleurs que les ressources linguistiques. Plusieurs modifications sont effectuées par l'auteur pour mettre en valeur les ressources morphologiques :

- Le rappel et la précision sont observés sur les 5 à 15 premiers documents (recherche superficielle) et les 50 à 150 premiers documents (recherche fouillée) ;
- Les règles de préfixation sont éliminées car elles génèrent trop de bruit ;

[14]Disponible à l'adresse `ftp.cs.cornell.edu/pub/smart/`

[15]Le rappel et la précision sont deux mesures d'évaluation des résultats adoptées à l'origine dans le domaine de la recherche d'information. La *précision* indique le pourcentage de documents corrects parmi ceux qui ont été trouvés : il s'agit de l'exactidude des résultats. Le *rappel* indique le pourcentage de documents trouvés par rapport à ce qui est attendu : il s'agit de la complétude des résultats. Les notions de silence et de bruit sont intimement liées avec la précision et le rappel. Le silence est le complémentaire du rappel, le bruit de la précision.

- Les requêtes courtes et longues sont distinguées car elles ne montrent pas les mêmes résultats ;
- La priorité dans l'analyse des résultats est donnée aux requêtes les plus sensibles à la différence des traitements morphologiques appliqués.

Suite à ces modifications, les ressources morphologiques et surtout l'affixation deviennent les meilleures. L'auteur constate aussi que les traitements morphologiques sont surtout utiles avec des requêtes et des documents courts. Cette observation est également faite par (Krovetz, 1993).

(Kraaij & Pohlmann, 1996) présentent des expériences semblables sur le néerlandais. Les auteurs comparent la *baseline* avec les traitements morphologiques : Porter adapté au néerlandais (98 règles) et ressources linguistiques compilées à partir de la base Celex (Burnage, 1990)[16]. Le désuffixeur et les paires de mots peuvent traiter la flexion, l'affixation et partiellement la composition. Plusieurs versions du désuffixeur et surtout plusieurs combinaisons de ressources morphologiques sont testées (flexion seule, flexion et affixation, flexion et composition, etc.). Cette expérience montre :

- la supériorité des traitements morphologiques par rapport à la *baseline* ;
- qu'il n'existe pas de grande différence entre le désuffixeur Porter et les ressources morphologiques. Notons que la même conclusion est faite par (Hull, 1996) avant que l'auteur ne modifie les principes d'évaluation des résultats et n'ajuste les ressources morphologiques ;
- que la flexion seule donne de meilleurs résultats uniquement lorsqu'elle est combinée avec l'affixation ;
- que la prise en compte de la composition est utile dans une langue concaténatoire comme le néerlandais ;
- et que, de manière générale, le rappel est amélioré au détriment de la précision.

Du fait de cette dernière constatation, les auteurs concluent que les traitements de la variation morphologique sont surtout utiles pour l'amélioration du rappel.

Conclusion et discussion

Nous avons présenté des traitements morphologiques utilisés dans les domaines de l'indexation et de la recherche d'information, qu'elles soient contrôlées ou libres. L'appariement de mots bruts laisse une faible possibilité de retrouver le bon terme (Furnas *et al.*, 1987), et la possibilité d'utiliser une syntaxe plus évoluée peut être une cause d'erreurs (Jansen *et al.*, 2000; Jones *et al.*, 1998). Les traitements lexicaux peuvent donc s'avérer utiles. Le traitement de la variation morphologique, qui peut être prise en charge par des désuffixeurs ou par des ressources morphologiques, donne des résultats contradictoires. Ils semblent dépendre de plusieurs facteurs :

- longueur des requêtes et des documents (Krovetz, 1993; Hull, 1996) ;

[16]Les auteurs n'indiquent pas la taille des ressources morphologiques, mais la version 3.1 de Celex (1990) couvrait 124 000 lemmes en néerlandais. Le nombre de paires de mots est au moins tout aussi important que dans l'expérience de (Hull, 1996).

- type flexionnel des langues ((Popovic & Willett, 1992) cité dans (Kraaij & Pohlmann, 1996)) :
  - les langues avec un système morphologique relativement simple peuvent être prises en charge par les désuffixeurs,
  - les langues avec un système morphologique complexe sont difficilement descriptibles avec les règles de désuffixation, elles sont plus favorables à l'utilisation de ressources morphologiques ;
- principes et mesures d'évaluation des résultats (Hull, 1996) ;
- contenu des ressources morphologiques (Hull, 1996; Kraaij & Pohlmann, 1996).

Il semble néanmoins que l'utilisation de connaissances linguistiques améliore les résultats (Krovetz, 1993; Hull, 1996; Gaussier *et al.*, 2000), et ce d'autant plus si elles sont adaptées au domaine étudié (Church, 1995; Jacquemin, 1997a; Xu & Croft, 1998). On peut aussi se demander si l'évaluation des résultats à travers la précision et le rappel reflète réellement les performances des appariements. Si les normalisations ne sont pas appliquées, le système reste aveugle face à de nombreux phénomènes linguistiques. Les normalisations indésirables pourraient être filtrées *a priori* à travers les ressources ou *a posteriori* avec les résultats.

2.2.3 Acquisition terminologique

Dans cette section, nous nous situons dans le contexte de la constitution (semi)automatique de terminologies. Les outils d'acquisition terminologique mettent en œuvre des stratégies différentes, dont voici une liste non exhaustive :

- recherche de segments répétés dans une fenêtre de n mots (Lafon, 1984; Frath *et al.*, 2000; Dias *et al.*, 2003),
- recherche de patrons syntaxiques correspondant aux groupes nominaux (boîte à outils `Xelda` de Xerox, module GN de l'outil `AlethIP` d'EDF),
- recherche de patrons syntaxiques correspondant aux syntagmes nominaux et adjectivaux, mais cette fois à l'intérieur des frontières syntaxiques (catégories syntaxiques qui ne peuvent pas faire partie des termes : coordination, pronoms, verbes conjugués, adverbes, etc.). Les groupes terminologiques sont ensuite décomposés en syntagmes minimaux (`Lexter` (Bourigault, 1993)),
- repérage de syntagmes répétés autour de connecteurs grammaticaux (*de*, *de l'*, *du*, etc.) et/ou d'ancres lexicales déjà connues (`Ana` (Enguehard *et al.*, 1992)),
- application de patrons syntaxiques des bitermes et ensuite de filtres statistiques et de règles de variation (`Acabit` (Daille, 1995)).

Qui plus est, les patrons syntaxiques utilisés par les différents outils ne sont pas les mêmes. Parmi les outils cités certains sont destinés spécifiquement à l'acquisition terminologique (comme `Lexter`, `Ana` ou `Acabit`), d'autres au traitement de la langue naturelle en général (`Xelda` et `AlethIP`). Avec ces derniers, les groupes nominaux extraits ne sont pas spécifiques des structures terminologiques, mais décrivent plutôt la décomposition de la phrase en syntagmes. Tandis que les outils dédiés à l'acquisition terminologique recherchent spécifiquement des structures des termes.

Parmi les outils terminologiques, il existe en plus des différences applicatives : `Lexter` visait à l'origine le traitement des documents électroniques en vue de leur indexation, `Ana` et `Acabit` la recherche d'information. Les termes obtenus avec les deux derniers sont donc plus concis, en particulier avec `Acabit` qui se concentre sur le repérage des bitermes.

La plupart de ces outils analysent des documents déjà prétraités avec des étiqueteurs morphosyntaxiques. La performance variable des étiqueteurs utilisés, de même que les différences de stratégies et de patrons syntaxiques, sont autant de facteurs qui conditionnent la variabilité des candidats termes proposés par les outils d'acquisition terminologique. Cette variabilité est potentiellement importante et ne manque pas de se manifester lors des comparaisons de candidats termes provenant de différents outils ou lors de l'intégration d'outils dans une même plateforme.

Dans l'évaluation d'outils terminologiques, il existe des travaux qui comparent les interfaces, les ressources externes utilisées, la convivialité, etc. (Béguin *et al.*, 1997). D'autres effectuent une comparaison plus détaillée à travers une analyse linguistique des sorties (Habert *et al.*, 1997; Spackman & Hersh, 1996) et leur intégrabilité (Hamon, 2000). Ce sont ces derniers qui nous intéressent.

(Habert *et al.*, 1997) comparent ainsi les sorties de deux outils : `AlethIP/GN` conçu pour les traitements généraux de la langue naturelle et `Lexter` (Bourigault, 1993) destiné à l'acquisition terminologique. L'analyse est effectuée à travers une comparaison des arbres syntaxiques des candidats termes. Elle montre que les arbres identiques proposés par ces deux outils ne représentent que la moitié des résultats et que les groupes nominaux uniques correspondent à environ 10 % pour chaque outil. Le reste correspond essentiellement à la subsomption, ou inclusion, d'arbres syntaxiques. De manière générale, `AlethIP/GN` extrait des arbres moins nombreux mais plus profonds et plus complexes que ceux fournis par `Lexter`. Les auteurs considèrent que ces différences sont dues aux facteurs suivants :

- La finalité des outils : `Lexter` est conçu pour l'acquisition terminologique, `AlethIP` pour le traitement de la langue en général. Le module de l'extraction de groupes nominaux d'`AlethIP` ne constitue qu'un des modules et ne concerne pas spécifiquement l'acquisition terminologique.
- Le fonctionnement : `AlethIP` utilise des dictionnaires externes, `Lexter` exploite des règles de découpage, des patrons syntaxiques et des règles de décomposition des groupes nominaux.
  - L'utilisation de dictionnaires permet à `AlethIP` de remettre en cause l'étiquetage morphosyntaxique, tandis que `Lexter` conserve l'étiquetage d'origine.
  - Par contre, `Lexter` décompose les groupes nominaux, tandis que `AlethIP` les garde intacts.

  Ces facteurs causent donc des différences dans les résultats quant à la catégorisation syntaxique des mots et la taille des groupes nominaux.
- La portée d'analyse : `AlethIP` effectue l'analyse au niveau de la phrase, `Lexter` effectue une analyse superficielle locale. `AlethIP` cherche à décomposer la phrase en syntagmes, tandis que `Lexter` effectue la recherche de syntagmes qui correspondent aux patrons terminologiques entre les frontières syntaxiques.

De ces deux outils, c'est bien sûr `Lexter` qui est mieux adapté à l'acquisition terminologique. Et ses stratégies de recherche des groupes terminologiques et de leur découpage le confirment.

(Spackman & Hersh, 1996) effectuent une comparaison de termes proposés par deux autres outils : `Xelda` et `Clarit` (Evans & Lefferts, 1993), un module de repérage de syntagmes du système du même nom destiné à la recherche d'information. Les auteurs constatent également des subsomptions d'arbres terminologiques dans les deux sens. Ils comparent ces résultats avec une liste de termes de référence. Le rappel total des deux outils est de 85,8 %. Parmi lesquels, 59,8 % de termes sont identifiés par les deux outils à la fois, 8,9 % sont uniques à `Xelda` et 17,1 % à `Clarit`. Malheureusement, les auteurs n'indiquent pas les structures syntaxiques qui manquent pour chaque outil.

La différence qui existe entre les candidats termes peut constituer une pierre d'achoppement lors de l'intégration des outils dans une même plateforme. Un exemple d'une telle intégration est présenté dans (Hamon, 2000, p. 133–134). L'auteur utilise les sorties de `Lexter` et d'un autre outil terminologique, `Faster` (Jacquemin, 1997b), et constate que la mise en correspondance des deux ensembles de candidats termes n'est pas aisée. Nous avons vu ci-dessus que `Lexter` élimine les conjonctions de coordination et les adverbes, considérés comme frontières entre les termes. `Faster`, par contre, étant destiné au repérage des variantes morphosyntaxiques des termes, conserve les variantes avec les conjonctions de coordination et les adverbes. Sans des traitements appropriés, il est impossible d'apparier les termes fournis par ces deux outils.

Différents outils d'acquisition terminologique, mais également de structuration, ne fournissent donc pas les mêmes ensembles de termes. Et même parmi les termes proposés par un même outil il existe des variantes terminologiques. Afin de faciliter le travail de validation et d'intégration de données, il est souhaitable d'appliquer des normalisations pour la réduction de ces variations.

Discussion et conclusion

La constitution de ressources terminologiques s'appuie, dans les travaux cités, sur l'exploration des données textuelles et la détection des expressions qui peuvent correspondre à des termes. Ces expressions sont repérées avec des stratégies de détection et de sélection fort différentes. Les termes proposés par les outils automatiques sont donc également potentiellement différents. La comparaison des sorties des outils d'acquisition terminologique (Spackman & Hersh, 1996; Habert *et al.*, 1997; Hamon, 2000) montre que les différences proviennent : (1) de l'objectif qui se trouve à l'origine de la conception des outils (analyse générale de la langue, acquisition terminologique pour l'indexation, pour la recherche d'information), (2) des stratégies de fonctionnement des outils (analyse et découpage de la phrase, décomposition en termes) et les ressources linguistiques utilisées (dictionnaires, règles) et (3) de la portée de l'analyse syntaxique (globale à une phrase ou locale). Par ailleurs, (Bourigault & Habert, 1998) indiquent que l'influence de l'application visée dans la constitution de ressources terminologiques joue un rôle primordial : pour un même domaine et à partir du même corpus, les termes et les liens retenus sont fonction de l'application pour

laquelle cette terminologie est créée. Ainsi, plutôt que de parler de la terminologie d'un domaine, il convient de parler de différents produits terminologiques qui correspondent à des usages précis (Bourigault & Habert, 1998; Bourigault & Slodzian, 1999). La constitution de terminologies devrait donc être guidée par l'application.

Il apparaît par ailleurs que les outils conçus spécifiquement pour l'acquisition terminologique fournissent de meilleurs résultats que les outils génériques de TAL (Habert *et al.*, 1997). Notons néanmoins que, dans un contexte multilingue, les outils non spécifiques à l'acquisition terminologique, mais qui appliquent des stratégies semblables dans différentes langues, peuvent fournir des résultats plus comparables et alignables que des outils développés séparément dans chaque langue (Grabar & Haag, 2003).

2.3 Discussion et conclusion

Nous avons vu que les contenus des terminologies génériques du domaine médical varient, de même que les stratégies des outils d'acquisition terminologique. Nous pouvons donc nous demander quels critères adopter, lors de la constitution de ressources terminologiques, dans le choix des termes 'officiels' parmi toutes les variantes pouvant exister ? Une des causes de cette variabilité, qui est souvent revenue durant notre exposé, provient du cadre applicatif où ces terminologies s'inscrivent. Le contenu des terminologies montre ainsi une tendance à varier selon leurs objectifs et les applications où elles sont utilisées. Les terminologies médicales présentées (le MeSH, la CIM, la SNOMED) sont effectivement 'fléchées'. Le choix des termes et des relations entre eux dépend des applications visées.

Face aux terminologies, de toute manière lacunaires (Hersh *et al.*, 1996; Spackman & Hersh, 1996), ce qui devient de plus en plus saillant 'grâce' aux traitements automatiques des documents, les techniques d'appariement de termes deviennent donc une aide importante dans la reconnaissance des variantes. Il s'avère ainsi plus crucial de développer des ressources et des outils qui permettent de neutraliser la variation terminologique que de poser des restrictions et des conventions dans le choix et l'usage des expressions terminologiques 'contrôlées' (Bodenreider *et al.*, 2002).

Nous avons présenté un certain nombre de types de variations terminologiques qui relèvent de différents niveaux (caractère, morphologie, mots vides, syntaxe, hyponymes, etc.) et demandent la mise en place de différentes techniques et ressources. Nous nous sommes concentrée en particulier sur les variations morphologiques et leur traitement. Ces variations peuvent être neutralisées avec des approches non linguistiques (application de règles de désuffixation, qui tronquent un mot dès qu'un 'suffixe' y a été reconnu et éventuellement recomposent sa finale) et des approches utilisant des ressources linguistiquement motivées (paires de mots reliés morphologiquement, règles morphologiques de variations flexionnelles régulières).

L'application même de ces traitements est sujette à discussion. Les résultats sont plutôt contradictoires et semblent dépendre de plusieurs facteurs : longueur des requêtes et des documents ; type flexionnel des langues ; principes et mesures d'évaluation des résultats ; contenu des ressources linguistiques. Il semble néanmoins que l'utilisation de connais-

sances linguistiques améliore les résultats (Krovetz, 1993; Hull, 1996; Gaussier *et al.*, 2000; de Loupy, 2000), d'autant plus que ces connaissances sont adaptées au domaine étudié (Church, 1995; Jacquemin, 1997a; Xu & Croft, 1998).

2.4 Précision des objectifs du travail

Dans la suite de notre travail, nous continuons de traiter la variation des termes au niveau morphologique. Face au manque de fiabilité des règles de désuffixation, nous misons en particulier sur l'utilisation de ressources motivées linguistiquement.

Des expériences présentées jusqu'ici, nous retenons que ces ressources sont utilisées avec succès dans la reconnaissance de variantes de termes relatives à :

- la flexion :
  {anévrisme, anévrisme*s*}, {anévrismal, anévrismal*e*} ;
- la suffixation (la préfixation semble bruiter les résultats (Hull, 1996)) :
  {anévrism(e)*/Nom*, anévrism*al/Adj*}, {biologie*/Nom*, biologiste*/Nom*} ;
- la supplétion :
  {foie*/Nom*, hépat*ique/Nom*}, {estomac*/Nom*, gastr*ique/Nom*} ;
- la conversion, qui devient visible aux outils du TAL uniquement dans les documents étiquetés morphosyntaxiquement :
  {muqueuse*/Adj*, muqueuse*/Nom*} ;
- la composition, surtout pour les langues concaténatoires et à condition d'être utilisée avec précaution :
  {anévrisme, anévrismo*rraphie*}.

Lorsque ces variations sont régulières, ce qui est le cas par exemple de la majorité des flexions, elles peuvent être traitées avec les règles correspondantes. Dans d'autres cas, il faut disposer de paires de mots qui mettent en relation les mots reliés morphologiquement.

Ces expériences nous font également retenir que la pertinence des ressources morphologiques augmente lorsque ces dernières sont adaptées aux corpus et domaines traités. Et que, pour fournir plus d'efficacité, le traitement des variantes morphologiques doit être combiné avec le traitement d'autres variations des termes.

Après avoir présenté plus en détail les procédés morphologiques de formation de lexèmes et de leurs variations (chap. 3), nous présentons les explorations automatiques des données textuelles. D'abord à travers l'acquisition de ressources morphologiques, ce qui constitue le premier objectif de notre travail (chap. 4), et avec une étude de la spécificité de certaines suffixations selon la spécialisation des documents, les domaines et les genres médicaux (chap. 5).

Ensuite nous passons à notre deuxième objectif qui consiste à observer par nous-même la contribution réelle apportée par les ressources morphologiques dans le traitement de la variation terminologique. Deux expériences sont alors présentées. La première correspond à une application proche de la recherche d'information (chap. 6). Nous effectuons une analyse des `log` du portail médical CISMeF pour comparer le lexique des termes d'indexation avec celui des utilisateurs. Nous appliquons les ressources morphologiques et les règles de

désuffixation de variations régulières. La deuxième se situe du côté de la structuration de terminologies (chap. 7). Nous appliquons l'hypothèse d'inclusion lexicale pour détecter des relations hiérarchiques entre deux termes. Dans les deux expériences, les traitements morphologiques sont couplés à d'autres traitements de normalisation (casse, ordre des mots, ponctuation, nombres, mots vides, synonymes).

Chapitre 3

Morphologie

Dans le chapitre précédent, nous avons décrit la variabilité des termes, les contextes d'apparition de cette variabilité et les moyens informatiques utilisés pour la neutraliser. Nous nous sommes concentrée sur les traitements de la variation des termes au niveau morphologique. Ce chapitre est consacré à la morphologie, une sous-discipline de la linguistique[1]. Pour décrire le potentiel de la morphologie dans la neutralisation de la variation terminologique, nous présentons d'abord ses assises théoriques (sec. 3.1) : les mécanismes mis en œuvre dans la formation et variation des lexèmes et leurs tenants sémantiques. Les faits morphologiques que nous présentons sont alors dictés par les besoins de description des données, en français, anglais et russe, induites et analysées dans le chapitre 4. Travaillant avec des données écrites, nous nous attachons surtout au traitement des 'graphèmes', bien que dans l'exposé proposé ici, les lexèmes soient souvent considérés selon leurs plans sémantique, morphologique, catégoriel et phonologique. Mais c'est surtout le traitement de la graphie des lexèmes qui nous intéresse. Ensuite, dans la section sec. 3.2, nous présentons les travaux en morphologie qui se situent du côté du TAL : il s'agit de l'utilisation des ressources morphologiques pour l'étiquetage morphosyntaxique, l'analyse sémantique, etc. Nous terminons avec un bilan, en particulier pour sélectionner les procédés de morphologies dérivationnelle et flexionnelle utiles pour le traitement de la variation terminologique (sec. 3.3).

[1]Nous sommes reconnaissante aux chercheurs de l'action 5 du GdR 2220 du CNRS 'Bases de données constructionnelles et productivité des procédés morphologiques en corpus' (Georgette Dal, Bernard Fradin, Nabil Hathout, Fiammetta Namer) d'avoir partagé avec nous et avec autant de passion leurs connaissances sur le domaine. Nous remercions particulièrement Fiammetta Namer et notre rapporteur Bernard Fradin d'avoir effectué une relecture attentive et critique de ce chapitre et de nous avoir signalé plusieurs références bibliographiques en rapport avec les sujets abordés. L'ensemble de leurs remarques nous a beaucoup aidée à améliorer cette présentation. Les imprécisions et incohérences qui persistent ne sont inputables qu'à nous-mêmes.

3.1 Morphologie, une sous-discipline de la linguistique

La morphologie est une sous-discipline de la linguistique qui met en œuvre des procédés et des règles spécifiques de formation et de variation des lexèmes. Certains de ces procédés, lorsqu'ils assurent une continuité sémantique entre les lexèmes, peuvent être utilisés dans le repérage des variantes des termes. Nous précisons d'abord l'objet de la morphologie (sec. 3.1.1), nous décrivons ensuite le matériau qu'elle manipule (sec. 3.1.2), ses deux branches : formation et variation des lexèmes (sec. 3.1.3), et la notion de famille morphologique (sec. 3.1.4). À la fin, nous discutons quelques cas qui n'ont pas d'explication unanime en morphologie (sec. 3.1.5).

3.1.1 Objet de la morphologie

La morphologie assure une double tâche. D'une part, la morphologie dérivationnelle ou constructionnelle s'occupe de la construction du lexique, à travers l'étude des lexèmes[2] complexes construits par les moyens de la grammaire (Fradin & Kerleroux, 2003) : affixation, conversion et composition parmi les procédés centraux. D'autre part, la morphologie flexionnelle propose une description des variations que les lexèmes subissent lorsqu'ils participent à la formation des énoncés. Avant de présenter ces deux morphologies, nous introduisons le matériau qu'elles manipulent.

3.1.2 Matériau de la morphologie

Le matériau morphologique est manipulé par les règles de formation de lexèmes et les règles de leur réalisation syntaxique. On peut en distinguer trois types : les lexèmes, les bases non autonomes et les affixes (Fradin & Kerleroux, 2003). Les lexèmes et les bases non autonomes peuvent servir de base dans les procédés de formation de lexèmes. Après la description de ce matériau, nous présentons l'allomorphie qui peut l'altérer formellement.

Lexèmes

Les lexèmes sont des instances fondamentales du signe linguistique. Ils sont multidimensionnels, étant constitués de plusieurs plans de représentations, dont les plans phonologique, morphologique, catégoriel et sémantique (Fradin & Kerleroux, 2003)[3]. Ils correspondent à un élément obligatoire lors de la formation de lexèmes complexes, construits ou fléchis. Les lexèmes utilisés de manière autonome par la syntaxe et qui servent également de base aux opérations de construction sont appelés *bases autonomes*. Les lexèmes peuvent être simples ou bien complexes. Des lexèmes construits ont déjà subi des opérations de construction (affixation ou composition, entre autres). Par exemple, la base non construite autonome *cellule*

[2]Rappelons qu'un *lexème* est une unité linguistique abstraite dans le sens où elle n'est pas pourvue de marques flexionnelles, contrairement aux *mots-formes* qui en sont pourvus (Lyons, 1968, p. 198).

[3]Ces plans peuvent aussi être attachés au niveau des règles de formation, comme dans (Corbin, 1987).

permet de construire l'adjectif *cellulaire*. Cet adjectif peut servir de base à la construction du nom *cellularité*. On peut avoir ainsi plusieurs niveaux de dérivations morphologiques.

Bases non autonomes

Les *bases non autonomes*, également appelées *bases supplétives*, *archéoconstituants*, *composants néolatins*, *bases savantes* ou *primitifs supplétifs* (Namer, 2003a), n'apparaissent pas à l'état autonome et ne constituent pas un matériau direct pour la syntaxe. Étant des *bases supplétives*, elles peuvent se substituer aux bases autonomes dans la formation des lexèmes :

{foie*/Nom*, hépat*ique/Adj*}
{queue*/Nom*, caud*al/Adj*}
{estomac*/Nom*, gastr*ique/Adj*}

Dans ces exemples, les bases qui servent à la formation des adjectifs sont formellement différentes des noms auxquels ces adjectifs correspondent. Ces bases supplétives, bien qu'elles ne soient pas autonomes actuellement, ont existé, dans la plupart des cas, à l'état autonome dans certaines langues. La langue médicale recourt largement à l'utilisation de bases supplétives, essentiellement d'origine grecque et latine.

La catégorie grammaticale des bases supplétives est définie en fonction des règles de construction qui s'appliquent à elles et en fonction des bases autonomes auxquelles elles peuvent se substituer. Dans les exemples ci-dessus, *hépat-*, *caud-* et *gastr-* sont des bases supplétives qui correspondent à des noms, car (1) les adjectifs en question, dits relationnels ou dénominaux, sont formés à partir des noms et (2) ces bases supplétives sont des équivalents sémantiques et référentiels des noms correspondants : *foie/Nom*, *queue/Nom* et *estomac/Nom*. La sélection d'une des bases dépend entre autre des contraintes phonologiques (Lignon, 1999; Plénat, 2001; Plénat & Roché, 2001b), de sa disponibilité,... Peut-être aussi des moyens traditionnels de formation de lexèmes dans une langue de spécialité.

Affixes

Les *affixes* sont des unités infralexématiques sans statut syntaxique. Ce ne sont pas des signes linguistiques mais ils entrent dans la construction de tels signes (Fradin & Kerleroux, 2003). Les affixes dérivationnels participent à la formation des lexèmes, tandis que les affixes flexionnels marquent leurs variations.

Affixes dérivationnels. Les *affixes dérivationnels* participent à la formation de lexèmes. Selon la place qu'ils occupent dans un lexème par rapport à la base, on distingue les *préfixes*, qui apparaissent avant la base, et les *suffixes*, qui apparaissent après la base. Les affixes *-aire*, *-al* et *-ique* sont des suffixes car ils apparaissent après la base à laquelle ils s'appliquent :

cellul*aire/Adj*, péricard*ique/Adj*, hépat*ique/Adj*, caud*al/Adj*, gastr*ique/Adj*.

Dans les exemples suivants, les affixes *anti-* et *avant-* sont des préfixes :

*anti*grippe/Adj, *avant*-bras/Nom.

En fonction des règles d'affixation, les affixes forment des lexèmes avec une sémantique et une catégorie syntaxique données. La règle de formation des adjectifs relationnels sélectionne ainsi le suffixe *-aire* pour construire des adjectifs sur des bases nominales avec un sens général « *Relatif à N* » :

{cellul*e/Nom*, cellul*aire/Adj*}.

Affixes flexionnels. Les *affixes flexionnels* marquent la variation flexionnelle des lexèmes. En français, en fonction de leur catégorie syntaxique, les lexèmes ont les dimensions de variations flexionnelles suivantes (Huot, 2001, p. 105) :

- substantifs :
  - nombre : {cellule, cellule*s*},
- adjectifs, pronoms :
  - nombre : {anévrism*al*, anévrism*aux*} ;
  - genre : {muqu*eux*, muqu*euse*} ;
- verbes :
  - temps : {piqu*e*, piqu*ait*},
  - mode : {piqu*e*, piquer*ait*},
  - personne : {piqu*ons*, piqu*ent*},
  - nombre : {piqu*es*, piqu*ez*},
  - genre avec les formes verbales au participe passé : {pris, pris*e*},
  - infinitif : piqu*er*, prend*re*.

Ces traits flexionnels des catégories syntaxiques sont sujets à questionnement. Il existe par exemple une discussion sur la nature, flexionnelle ou dérivationnelle, du genre et du nombre, comme le montrent les ouvrages de G. Corbett (Corbett, 1991; Corbett, 2000)[4].

Allomorphie

Les *allomorphies* sont des variations phonologiques et donc formelles des éléments morphologiques, déclenchées par les contextes morphologiques (Corbin, 1987, p. 289). Le plus souvent, les allomorphies sont déclenchées dans les bases par les affixes :

- postériorisations phoniques déclenchées par certains suffixes :

  {abdom*e*n*/Nom*, abdom*i*n*al/Adj*}
  {aborig*è*ne*/Nom*, aborig*i*n*al/Adj*}
  {m*e*r*/Nom*, m*a*r*in/Adj*}
  {h*eu*re*/Nom*, h*o*r*aire/Adj*}

- transformation de */ai/* en */a/* déclenché par *-ien* :

  {ov*ai*re*/Nom*, ov*a*r*ien/Adj*}
  {coron*ai*re*/Nom*, coron*a*r*ien/Adj*}
  {urtic*ai*re*/Nom*, urtic*a*r*ien/Adj*}

[4]Nous remercions Fiammetta Namer de nous avoir signalé ces ouvrages de Greville G. Corbett. Toutefois, nous n'avons pas pu les consulter au moment de la rédaction de la thèse.

Mais les bases peuvent aussi déclencher les variations allomorphiques des affixes. Ainsi, les préfixes de négation *dé-* et *a-* se réalisent différemment selon que la base commence par une voyelle ou un *h* non aspiré (*dés-*, *an-*) ou bien une consonne (*dé-*, *a-*) :

préfixe	base commençant par une voyelle ou un *h*	base commençant par une consonne
dé-	*dés*infecter, *dés*équilibrer, *dés*hydrater	*dé*liter, *dé*générer, *dé*saturer
a-	*an*aérobie, *an*algésique, *an*hydride	*a*typique, *a*symptomatique

En outre, l'allomorphie peut apparaître dans un affixe sous l'influence d'un autre affixe. Ainsi, le suffixe *-able* se transforme en *-ible* si la base comporte, par exemple, la séquence *-sc-* (Corbin, 1987, p. 315–316) :

fermentesc*ible*, irasc*ible*, revivisc*ible*.

Dans (Corbin, 1987, p. 293–294), les allomorphies et les supplétions sont considérées comme deux extrémités d'un même continuum. Ainsi pour qu'une alternance phonologique soit considérée comme allomorphique, il faut qu'elle soit reproductible et donc explicable par des propriétés phonologiques. Si ce n'est pas le cas, comme dans les exemples qui suivent, il s'agit de bases supplétives :

{*vierge*, *virginité*}
{*nager*, *natation*}
{*roche*, *rupestre*}.

Notons que dans (Bonami *et al.*, 2005), les bases allomorphiques peuvent entrer dans la collection de thèmes d'un lexème et être ensuite sélectionnées par les opérations de construction ou par la flexion.

3.1.3 Deux sous-domaines de la morphologie

Comme nous l'avons déjà noté, nous distinguons deux morphologies : la morphologie dérivationnelle, qui décrit les procédés de construction de lexèmes, et la morphologie flexionnelle, qui décrit la variation de ces lexèmes lors de leur insertion dans les structures syntaxiques.

Morphologie dérivationnelle ou constructionnelle

Le but de la *morphologie dérivationnelle ou constructionnelle* consiste à proposer une grammaire explicite de la construction des lexèmes[5]. Les règles de construction réunissent entre autre les contraintes catégorielles, morphologiques, sémantiques et phonologiques. Soit la règle de formation des adjectifs relationnels (ou dénominaux) sur les bases nominales[6] :

$$[[X]_N(\text{aff})]_A$$

[5]Notons que (Corbin, 1987; Corbin, 1991), dont les travaux sont à l'origine de la morphologie constructionnelle, ne parle pas de règles de formation de lexèmes, mais de mots.

[6]Nous utilisons ici la notation linéaire proposée dans (Corbin, 1987). Notons qu'il existe d'autres notations possibles, linéaires ou non, par exemple celle utilisée dans (Fradin, 2003), qui rend compte de différents plans de représentations du lexème base et du lexème construit.

Selon (Corbin, 1987), la contrainte catégorielle de cette règle prévoit l'application d'un paradigme d'affixes (aff) à une base nominale $[X]_N$ pour la formation d'un adjectif A. La contrainte morphologique prévoit que le paradigme d'affixes pouvant intervenir dans l'opération comporte, par exemple, les suffixes *-aire*, *-al*, *-el*, *-eux* et *-ique*. La contrainte sémantique prévoit que le sens prédictible et général des adjectifs formés par cette règle est « *Relatif à N* » ou bien « *Qui est en rapport avec N* ». Notons que la contrainte sémantique prime sur la contrainte catégorielle (Dal, 1997), et que la contrainte phonologique impose une tendance à la dissimilation, ce qui évite la jonction d'une base et d'un suffixe trop proches phonologiquement (Lignon, 1999; Plénat & Roché, 2001a), comme ce serait le cas de *cellulal* ou *cellulel*. Quelques exemples d'adjectifs formés par cette règle sur les bases *cellule*, *anévrysme*, *clone* et *péricarde* et leur sens tel que donné par les dictionnaires :

- cellul*aire/Adj* : « *Qui se rapporte aux cellules ; Qui est constitué par des cellules* » (Manuila *et al.*, 1970)
- anévrysm*al/Adj* : « *Qui se rapporte à un anévrysme ; Qui en a les caractères* » (Manuila *et al.*, 2001)
- clon*al/Adj* : « *Relatif aux clones* » (TLFi, 2001)
- péricard*ique/Adj* : « *Qui se rapporte ou qui appartient au péricarde* » (Manuila *et al.*, 1970)

La morphologie dérivationnelle met en œvre plusieurs procédés de formation des lexèmes, parmi lesquels l'affixation, la conversion et la composition sont considérés comme centraux (Fradin, 2003, sec. 5.3). Parmi les procédés de formation de lexèmes qui se rapprochent de la morphologie mais lui restent périphériques, (Fradin, 2000) distingue entre autres la *composition cachée* (*concealed compounding*) et les *mots valises*. La série suivante de lexèmes correspond à la composition cachée, où la forme tronquée du lexème *handicap* est combinée avec d'autres lexèmes tronqués ou non. Les deux gardent leur contenu sémantique (Fradin, 2003, p. 213) :

handi-contact, *handiplace*, *handivie*, *handimag*, *handitel*, *handiscol*.

Quant aux mots valises, leur formation est soumise à des patrons basés sur des segments phonologiquement identiques ou similaires (Σ) dans les lexèmes combinés :

$A\Sigma\ \Sigma B \Rightarrow A\Sigma B$ *métropole politique* $\Rightarrow$ *métropolitique*
$A\Sigma\ B\Sigma \Rightarrow AB\Sigma$ *potiron maron* $\Rightarrow$ *potimaron*
$A\Sigma\ B\Sigma B' \Rightarrow A\Sigma B'$ *hippie épidémie* $\Rightarrow$ *hippidémie*.

Bien que la composition cachée, les mots valises et d'autres procédés périphériques (accourcissements, affixation secrétive, siglaisons, etc.) soient de plus en plus productifs en langue (voir par exemple sec. 8.4 page 167), nous ne les considérons pas ici. Essentiellement à cause de leur caractère irrégulier et peu prévisible. Par contre, l'affixation, la conversion et la composition peuvent être traitées avec les règles et contraintes proposées par la morphologie constructionnelle (voir par exemple (Dal *et al.*, 1999; Hathout *et al.*, 2001; Namer & Zweigenbaum, 2004)).

Affixation. L'*affixation* consiste en la combinaison d'une base et d'un affixe. Selon le type d'affixe qui intervient dans les opérations morphologiques en trois langues étudiées

(français, anglais et russe), une affixation peut être préfixale ou suffixale. On appelle les lexèmes qui résultent d'une affixation *lexèmes construits*, *lexèmes affixés*, *lexèmes dérivés*, *dérivations* ou bien *affixations*. Le passage du lexème affixé vers sa base s'appelle la *racinisation* ou la *désuffixation*[7].

RC	Préfixation	Suffixation
Adj → Adj	{mortel, *im*mortel}	{propre, propr*et*}
Adj → Nom		{mortel, mortal*ité*}
Adj → V	{niais, *dé*niais(er)}	{banal, banal*is(er)*}
Adj → Adv		{banal, banal*ement*}
Nom → Adj	{grippe, *anti*grippe}	{mort, mort*el*}
Nom → Nom	{pesanteur, *a*pesanteur}	{maison, maisonn*ette*}
Nom → V	{moustique, *dé*moustiqu(er)}	{pacte, pact*is(er)*}
V → Adj		{pli(er), pli*able*}
V → Nom		{balay(er), balay*age*}
V → V	{coud(re), *dé*coud(re)}	{tap(er), tap*ot(er)*}

TAB. 3.1 – Contraintes catégorielles pour l'affixation.

Le tab. 3.1, compilé à partir de (Corbin, 1987, p. 479) et (Corbin, 1993), présente les contraintes ou les rapports catégoriels (RC) possibles entre une base et un lexème formé par préfixation et suffixation. L'affixation prend en entrée deux éléments : un lexème, autonome ou non, et un affixe. Elle produit en sortie toujours un lexème construit de catégorie majeure : noms, adjectifs, verbes. Outre les rapports catégoriels indiqués dans le tableau, les numéraux cardinaux et les pronoms peuvent être suffixées pour former, respectivement, des noms et des adjectifs ordinaux et des verbes (Corbin, 1993) :

{deux, douz*aine*}, {deux, deux*ième*},
{tu, tu*toy(er)*}, {vous, vou*voy(er)*}.

Les rapports catégoriels sont gérés par des contraintes sémantiques (Dal, 1997). Par exemple, les structures suffixées A → A, N → N et V → V correspondent à la construction d'un sens évaluatif : la diminution (*maisonnette*, *propret*, *tapoter*) et la péjoration (*vinasse*). La structure suffixée N → A correspond à la construction des adjectifs dénominaux ou relationnels (*cellulaire*, *bronchique*).

Les affixations peuvent être iso- et transcatégorielles. Dans les *affixations transcatégorielles*, le lexème affixé n'a pas la même catégorie syntaxique que sa base. Les suffixes comme les préfixes peuvent former des affixations transcatégorielles (Corbin, 1992, p. 129). Par exemple, la formation des adjectifs dénominaux avec les suffixes *-aire/Adj* et *-ique/Adj* et de préfixation avec *anti-* et *a-* :

[7]Chez certains auteurs (Dal & Jacquemin, 1999), la *racinisation* s'oppose à la *désuffixation* (*stemming*) par le fait que cette dernière correspond à un procédé linéaire de suppression des affixes ou des chaînes finales en général. Tandis que la racinisation sous-entend une analyse linguistique et morphologique. C'est également l'acception que nous en faisons. Mais notons que les deux termes s'emploient généralement avec le même sens de *stemming*.

{cellul*e/Nom*, cellul*aire/Adj*}
{bronch*e/Nom*, bronch*ique/Adj*}
{*pétale/Nom*, *a*pétale/Adj}
{*grippe/Nom*, *anti*grippe/Adj}.

Dans les *affixations isocatégorielles*, la base et le lexème affixé appartiennent à la même catégorie syntaxique :

{cellule*/Nom*, cellul*ette/Nom*}
{*ven(ir)/Ver*, re*ven(ir)/Ver*}
{*aérobie/Nom*, an*aérobie/Nom*}.

Les affixations peuvent se former sur des bases autonomes (exemples 1 à 3) ou bien supplétives (exemples 4 à 6) :

(1) {cellul*e/Nom*, cellul*aire/Adj*}
(2) {anévrism*e/Nom*, anévrism*al/Adj*}
(3) {péricard*e/Nom*, péricard*ique/Adj*}
(4) {foie*/Nom*, hépat*ique/Adj*}
(5) {queue*/Nom*, caud*al/Adj*}
(6) {estomac*/Nom*, gastr*ique/Adj*}.

Conversion. La *conversion* est un autre procédé de formation de lexèmes. La conversion peut prendre en entrée un seul élément de n'importe quelle catégorie. Elle produit en sortie un seul lexème de catégorie majeure. Le tab. 3.2, compilé à partir de (Corbin, 1987, p. 479) et (Corbin, 1993), présente quelques rapports catégoriels possibles entre un lexème convert et sa base. Mais les catégories syntaxiques mineures et les éléments morphologiques peuvent également être touchées par la conversion :

Peser le pour et le contre avant d'agir.
Il y a l'hibernateur, [...], le synthéticien, le physicien, [...], l'endocrinologue, l'hématologue, et quelques autres ciens et logues, les meilleurs du continent.

RC	Conversion
Adj → Nom	{alcoolique, alcoolique}
Adj → V	{blanc, blanch(ir)}
Nom → Adj	{orange, orange}
Nom → V	{singe, sing(er)}
V → Adj	
V → Nom	{vol(er), vol}

TAB. 3.2 – Contraintes catégorielles avec la conversion.

La forme graphique et phonologique des lexèmes converts peut changer par rapport à leur base : {*blanc/A*, *blanch(ir)/V*}. Mais ce changement n'est pas systématique. Rappelons par ailleurs, que la désinence de l'infinitif des verbes, qui est un affixe flexionnel, ne concerne pas la morphologie constructionnelle (Corbin, 1987, p. 124–129), (Fradin, 2003, p. 137–138).

Notons que les exemples du tab. 3.2 concernent le type le plus courant de conversion, la *conversion catégorielle* (Fradin, 2003, p. 157–161). (Melčuk, 1996) et (Fradin, 2003, p. 161–165) distinguent en outre la *conversion valencielle*, la *conversion de classe morphologique* et la *conversion flexionnelle* que nous ne présentons pas ici.

Composition. La *composition* est un procédé de formation des lexèmes qui permet de combiner au moins deux lexèmes, autonomes ou non. Les lexèmes qui résultent de la composition sont dits *lexèmes composés* ou *compositions*. Le sens d'un lexème composé est défini par (1) le sens produit par la règle qui impose une structure et un schéma interprétatif de combinaison des composants et (2) le sens de chacun des composants (Corbin, 1992). Dans un lexème composé, un des composants tient généralement place de l'élément sémantiquement recteur. Selon la place occupée par cet élément, mais également selon la nature et l'origine des composants, on distingue la composition populaire et la composition savante (Corbin, 1993).

Composition populaire. Dans une *composition populaire* française, l'élément recteur se trouve en première position. Ce type de composition combine très souvent des bases autonomes de catégories majeures pour former des noms ou des adjectifs. Le tab. 3.3, compilé à partir des informations rencontrées dans (Corbin, 1993; Namer, 2003a; Villoing, 2002), en présente quelques exemples. Le composant souligné correspond au composant recteur, il se trouve toujours en première position.

	$\rightarrow N$	$\rightarrow A$
$\underline{N} + N$	*bébé-éprouvette*, *cas-témoin*, *médecin-généraliste*	
$\underline{A} + A$		*gris-bleu*, *sourd-muet*
$\underline{V} + N$	*porte-drapeau*, *garde-barrière*	*casse-pieds*

TAB. 3.3 – Contraintes catégorielles des composés populaires.

Les composés de type $[\underline{N_1}N_2]_{N_3}$ reçoivent une interprétation selon laquelle le lexème construit N_3 appartient à une sous-classe sémantique de N_1, conformément à l'hypothèse de l'inclusion lexicale énoncée dans (Kleiber & Tamba, 1990). Le deuxième composant N_2 apparaît ainsi comme spécifieur du premier N_1 (Corbin, 1994). La transmission des propriétés de N_2 à N_3 varie selon un continuum : l'interprétation de N_3 peut aller d'un véritable hyponyme de N_1 (généralement par projection d'une seule propriété de N_2, comme dans *bébé-éprouvette*, *cas-témoin*) à une combinaison étroite des propriétés de N_1 et N_2 dans N_3 (phénomène appelé 'hybridation', comme dans *médecin-généraliste*, *porte-fenêtre*, *libraire-éditeur*) (Lesselingue, 2003). Par contre, dans ce type de structure, rien n'indique lesquelles des propriétés de N_2 sont transmises à N_3 (Habert & Jacquemin, 1993). L'élement recteur N_1 du lexème composé $[\underline{N_1}N_2]_{N_3}$ joue le rôle sémantique dominant et impose ses traits syntaxiques à N_3. Si N_1 et N_2 ne sont pas du même genre, c'est N_1 qui détermine le genre

du composé N_3 : *bébé-éprouvette*, *hôpital-ville* sont tous les deux masculins conformément au genre de leurs premiers composants (Lesselingue, 2003).

Les composés de type $[\underline{A_1}A_2]_{A_3}$ reçoivent une interprétation similaire à celle de $[\underline{N_1}N_2]_{N_3}$ (Corbin, 1993). Le premier constituant est interprété comme l'hyperonyme du lexème composé : *gris-bleu* est un type de *gris*, *sourd-muet* est un type de *sourd*.

Enfin, les constructions de type $[\underline{V}N_1]_{N_2/A}$ combinent un composant verbal et un composant nominal pour former des noms ou des adjectifs :

rince-bouche, *lave-vaisselle*, *pèse-bébé*, *chauffe-plat*.

Le fait qu'un des composants soit d'origine verbale soulève la question de l'origine morphologique ou syntaxique des composés... déjà à partir du XIX^{e} siècle. Les débats portent sur le statut de la forme verbale et de la construction. Selon les analyses du XIX^{e} siècle, la forme verbale peut correspondre :

- à la forme fléchie du verbe à la 2^{e} personne de singulier de l'impératif (Darmesteter, 1875) cité dans (Villoing, 2002),
- à la forme fléchie du verbe à la 3^{e} personne de singulier du présent de l'indicatif (Meunier, 1872) cité dans (Villoing, 2002),
- au *thème verbal* (forme verbale non fléchie) (Boucherie, 1876) cité dans (Villoing, 2002).

Le dernier cas seulement, ne présentant pas une forme fléchie, correspond à une construction morphologique. L'analyse de ces composés dans le cadre de la grammaire générative du XX^{e} siècle penche également vers la syntaxe parce que :

- la morphologie ne constitue un domaine autonome de la grammaire (Lieber, 1992) cité dans (Villoing, 2002),
- et si elle en constitue un, elle ne peut pas former des lexèmes composés en général (Barbaud, 1991) cité dans (Villoing, 2002)
- ou des composés de type $[VN_1]_{N_2}$ (Di Sciullo & Williams, 1987; Zwanenburg, 1991) cités dans (Villoing, 2002).

(Villoing, 2002, p. 172–201) montre que le composant verbal correspond au thème et que, par conséquence, ces composés sont formés par la morphologie :

- Les formes verbales ne sont pas toujours homographes des formes flexionnelles à l'impératif présent de singulier :

  guérit-tout vs guéris, *abat-jour vs abats*, *cuit-œuf vs cuis* ;

  ni à l'indicatif de présent à la 3^{e} personne du singulier :

  boitout vs boit, *faitout vs fait*.

  On ne peut donc pas étendre les conclusions faites sur la base des homographes (*porte-plume*, *ouvre-boîte*, *cueille-fruits*), même s'ils sont fréquents, à toutes les formations.
- Ces formes verbales ne sont marquées ni pour un mode, ni pour un temps, ni pour une personne et correspondent, de ce côté aussi, au thème verbal.
- Si l'on prend en compte tous les homographes possibles du thème verbal (3^{e} personne de singulier de présent ou 2^{e} personne de singulier de l'impératif), les constructions ne sont pas homogènes. Alors que le fait de considérer la forme verbale comme un thème conduit à un traitement unique.

L'analyse du même type de composés en italien conduit (Scalise, 1994) à la même conclusion : l'élément verbal dans un composé $[VN_1]_{N_2}$ n'est pas une forme fléchie, mais un thème[8]. L'analyse sémantique de ces composés montre qu'il existe différentes relations non syntaxiques, de type prédicat-patient et prédicat-agent entre le V et le N : changement d'état (*chauffe-plat* (le plat a été chauffé)), incrémentalité (*épluche-légumes* (les légumes ont été (à moitié ou complètement) épluchés)), affectitude (*casse-noix* (la noix a été cassée)), etc. (Villoing, 2002, p. 202–252).

Composition savante. Dans une *composition savante* les lexèmes sont souvent des bases non autonomes d'origine latine ou grecque. En outre, le lexème recteur se trouve en dernière position, sauf de rares exceptions héritées souvent de la langue source (*misogyne*, *misanthrope*, *philosophe*) (Corbin, 1993). Graphiquement, les lexèmes sont concaténés directement ou bien au moyen d'un tiret ou d'un *interfixe* (une voyelle, le plus souvent *-o-* ou *-i-*, qui s'intercale entre deux composants pour améliorer la phonologie du lexème) (Corbin, 1993; Namer, 2003a).

(Cottez, 1985) (cité dans (Fradin, 2000)) distingue trois types de composés savants :

- *Emprunt* : le lexème existe en latin ou en grec et il est utilisé avec le même sens en français. C'est le cas du lexème *hémorragie* (« *Écoulement de sang hors d'un vaisseau sanguin lésé, à la surface du corps ou à l'intérieur d'un organe ou d'un tissu* » (Manuila *et al.*, 2001)), créé en latin avec des éléments grecs (Bloch & von Wartburg, 2002).
- *Adaptation* : le lexème existe en latin ou en grec mais il est utilisé en français avec un sens différent. Par exemple, *physiologie* signifie en grec « *Sciences naturelles* » mais prend en français le sens de « *Science qui traite des fonctions normales d'un organisme animal, végétal ou humain* » (Manuila *et al.*, 2001).
- *Création* : les composants existent en latin ou en grec, mais le lexème est formé actuellement. Il s'agit par exemple de *chronologie*, *chimiothérapie*, *taxidermie*.

Dans ce qui suit, nous ne faisons pas la différence entre ces trois types de composition, nous les analysons comme un seul procédé. Mais cette distinction, en particulier entre l'emprunt et l'adaptation *vs* la création, peut être importante quant à la place de l'élément recteur dans le composé : le grec et le latin ont formé, dans de très rares cas, des lexèmes où le composant recteur se trouve en première position, comme dans les exemples cités ci-dessus (*misogyne*, *misanthrope*, *philosophe*). Signalons aussi que (Fradin, 2003, p. 197) considère que les procédés de construction de la morphologie savante peuvent différer de ceux existant dans la langue ordinaire. Il n'est donc pas sûr qu'on doive les considérer comme ayant forcément appartenu à la grammaire de la langue.

Les composants savants appartiennent à l'une des catégories grammaticales majeures, le résultat de cette composition est un nom ou un adjectif. Le tab. 3.4, compilé à partir des informations rencontrées dans (Corbin, 1993; Namer, 2003a; Fradin, 2003), en présente quelques exemples. Le composant souligné correspond au composant recteur, il se trouve généralement en dernière position.

[8]Nous remercions Fiammetta Namer de nous avoir signalé cette référence.

	$\rightarrow N$	$\rightarrow A$
$N + \underline{N}$	*gastroentérite*	
$\underline{N} + A$	*éocène*	
$A + \underline{N}$	*leucémie*, *mégalithe*	*macrocéphale*, *anglophone*
$A + \underline{A}$	—	
$\underline{V} + N$	*philosophe*	*misogyne*
$N + \underline{V}$	*homicide*, *sexologue*	*crucifère*, *sexophobe*, *anthropophage*

TAB. 3.4 – Contraintes catégorielles des composés savants.

Selon (Corbin, 1993), dans les composés de type $[N_1 \underline{N_2}]_{N_3}$ le composant recteur N_2 se trouve en relation d'hyperonymie par rapport au composé N_3. Ainsi, *gastroentérite* est un type de *entérite*. *Entérite*, à son tour, est un type d'*inflammation* (représentée par le suffixe *-ite*[9]). Les structures $[A\underline{N}]$ peuvent former des noms (*similicuir*, *mégalithe*) et des adjectifs (*macrocéphale*, *anglophone*). Dans les deux cas, on observe également la relation d'hyperonymie entre le composant recteur et le composé (Corbin, 1993). Les structures $[\underline{V} N_1]_{N_2}$ et $[N_1 \underline{V}]_{N_2}$, qui impliquent les composants verbaux, peuvent recevoir une interprétation du procès (*homicide*), sinon d'un nom d'agent ou d'instrument (*anthropophage*, *crucifère*) (Corbin, 1993).

Le bon usage voudrait que les composés savants comportent des éléments d'une même origine (grecque, latine, française,...). Mais il existe énormément de composés qui mélangent des composants d'origines variées. On les appelle *composés hybrides* (Riegel *et al.*, 1998, p. 550) ou *chimères* (Bouché, 1994, p. 241) :

> adénocancer : du grec *adên* (glande) et du latin *cancer* (crabe),
> scanographie : du l'anglais *to scan* (scruter) et du grec *graphein* (écrire).

Morphologie flexionnelle

La variation des lexèmes lors de leur insertion dans les énoncés est du ressort de la *morphologie flexionnelle*. Elle consiste en l'application de marques flexionnelles au moyen d'*affixes flexionnels* (de *désinences* ou de *flexions*) à un lexème ou un lemme. Le *lemme*, qui couvre une notion proche de celle de *lexème*, correspond aux formes 'citationnelles' : infinitif pour les formes verbales, singulier pour les noms, singulier masculin pour les adjectifs et les pronoms. Les lemmes ou les lexèmes, qui ont subi une variation flexionnelle, sont dits *formes fléchies*, *flexions*, *mots-formes*, *formes*, *words*, etc.

Les variations flexionnelles pertinentes pour le français sont celles du nombre (pour les noms, les adjectifs, les verbes et les pronoms), du genre (pour les adjectifs, les pronoms et quelques formes verbales) et de la personne, du temps et du mode (pour les verbes).

[9]Les composants de type *-ite* montrent la fragilité des frontières entre l'affixation et la composition. Nous expliquons nos positions sur ce point dans la sec. 3.1.5 page 51.

3.1.4 Familles morphologiques de mots

Pour la morphologie constructionnelle, la notion de *famille morphologique* couvre l'ensemble des lexèmes reliés par des relations formelles et sémantiques et gérés par des règles de formation (Corbin, 1987, p. 89). À l'origine d'une famille se trouve l'élément morphologiquement le plus élémentaire, qui donne naissance aux autres membres de cette famille.

(Corbin, 1987, p. 119) nous donne un exemple didactique : dans le Robert Méthodique, les mots associés sémantiquement et formellement à *labeur* et dérivés de la racine *-lab-* sont regroupés selon leur degré de dérivation de la manière suivante :

collaborer, *collaborateur*, *collaboration* ;
élaborer, *élaboration* ;
labeur, *laborieux*, *laborieusement* ;
laboratoire, *laborantin*.

D'après les critiques de (Corbin, 1987, p. 119), cette liste mêle les dérivés de premier, deuxième et troisième degrés. Une chaîne dérivationnelle plus correcte serait (Corbin, 1987, p. 89)[10] :

labeur → °*laborer* → *collaborer* → *collaborateur*
↘ *collaboration*
↘ *élaborer* → *élaboration*
↘ *laboratoire*
↘ *laborant* → *laborantin*
↘ *laborieux* → *laborieusement*

Dans cette chaîne, le lexème *labeur* se trouve à la racine de la famille, chaque flèche représente un degré de dérivation, et pour chaque couple des lexèmes reliés morphologiquement, il existe une règle de formation.

La notion de la famille morphologique se trouve à la base des travaux de (Gruaz, 1998), dont l'objectif est l'organisation synchronique du lexique. Nous verrons que les bases théoriques et les résultats sont très différents de ce qui se fait dans le courant de la morphologie constructionnelle. Ainsi, bien que (Gruaz, 1998) parle des chaînes dérivatives et de la morphologie, il considère que la forme des lexèmes doit être écartée et que l'étymologie des mots doit être modulée par leur sens. Ce sont donc les sèmes communs des lexèmes qui définissent si ces lexèmes appartiennent à la même famille. Les sèmes sont établis selon les articles dictionnairiques proposés par le Lexis et, ponctuellement, par le Petit Robert. La famille `GÉNÉRER` comporte ainsi les entrées suivantes selon les degrés de dérivation :

1er degré de dérivation : *générer*, *genèse*, *génie*, *engendrer* ;
2e degré de dérivation : *progéniture* ;
3e degré de dérivation : *germain*.

[10] Le signe ° devant un mot est utilisé pour signifier un lexème lacunaire. Dans la chaîne dérivationnelle de l'exemple, le lexème °*laborer* est lacunaire : il est possible et prédictible morphologiquement mais n'est pas attesté par l'usage ni dans les dictionnaires. Notons que les bases de ce type peuvent faire partie de l'espace thématique des lexèmes (Bonami *et al.*, 2005)

Nous remercions Fiammetta Namer de nous avoir signalé que les lexèmes *génération* et *génératif* sont absents de la famille. Il faudrait certainement en chercher l'explication dans les articles dictionnairiques correspondants.

Selon (Gruaz, 1998), les liens entre les mots dans une famille peuvent être motivés morphologiquement, comme dans {*renard*, *renardeau*}, au moyen de l'affixation, ou bien avec un 'morphème zéro', comme dans la famille JOUR :

jour$_0$ *Espace de temps correspondant à une rotation complète de la Terre sur elle-même* ;
jour$_1$ *Espace de temps qui se situe entre le lever et le coucher du soleil* ;
jour$_2$ *Lumière que le soleil répand sur la Terre* ;
jour$_3$ *Ouverture qui donne de la lumière.*

Remarquons qu'il s'agit dans ces derniers exemples d'une dérivation sémantique plutôt que morphologique. Voir également les critiques à propos du morphème zéro faites dans (Corbin, 1987, p. 134–139).

3.1.5 Pour les cas discutables, définissons quelques frontières...

La morphologie garde encore de nombreuses zones d'ombre. Nous nous limitons ici à discuter celles relatives à notre travail : acquisition et utilisation de ressources morphologiques pour le traitement de la variation terminologique. Tout d'abord, en sachant que la morphologie et la syntaxe enrichissent toutes les deux le lexique, nous confrontons et tâchons de distinguer les constructions que chacune propose, ceci afin d'en délimiter la portée. Ensuite, nous tâchons de délimiter la morphologie flexionnelle et la morphologie dérivationnelle, ce qui vise à établir une hiérarchie dans l'application des affixes mais également à délimiter les effets sémantiques des procédés morphologiques. Et enfin, nous distinguons l'affixation et la composition, en sachant que certains composants actuels sont proches des affixes dérivationnels. Là aussi, l'enjeu porte sur les effets sémantiques. Pour chaque cas, nous tâchons de positionner les débats existants et de distinguer les traits discriminants.

...entre les constructions morphologiques et syntaxiques

Si la morphologie et la syntaxe permettent toutes les deux d'enrichir le lexique, elles n'ont pas accès au même matériau et la nature des constructions qui en résultent n'est pas la même (Corbin, 1994). De manière générale, (Corbin, 1992, p. 50) avance le critère suivant : « *Ne pas compter au nombre des unités lexicales relevant d'une opération lexicale de composition les unités qui peuvent être générées sans dommage par d'autres composantes de la grammaire.* »

Ainsi, contrairement à la morphologie, la syntaxe a accès aux 'mots grammaticaux', à la négation et aux mots-formes. Les expressions issues de la syntaxe correspondent souvent à des syntagmes nominaux (SN), verbaux (SV) ou autres bien formés. Les expressions du tab. 3.5 sont d'origine syntaxique. Pour plus d'exemples de ce type, voir (Fradin, 2003,

Expression	Structure syntaxique, contenu
rein kystique	SN
sténose de l'aorte	SN, préposition, déterminant
dermatite diffuse et nodulaire	SN, conjonction
pneumonie non résolue	SN, négation
accident relié à une activité professionnelle	SN, flexion, préposition, déterminant

TAB. 3.5 – Quelques exemples d'expressions d'origine syntaxique.

p. 199–206). De son côté, la morphologie seule a accès aux éléments supplétifs et aux affixes dérivationnels. Ainsi, les constructions qui suivent sont d'origine morphologique :

kystique, *gastrique*, *biochimie*, *cellulaire*, *construction*, *dermatite*.

Dans ces exemples, *-ique*, *-aire* et *-tion* sont des affixes dérivationnels et *gastr-* une base supplétive, *bio-* et *-ite* pouvant également être considérés comme des bases supplétives.

Les cas difficiles à classer relèvent de la composition populaire : il s'agit de constructions $[VN_1]_{N_2}$ et $[N_1N_2]_{N_3}$.

Analyse des composés $[VN_1]_{N_2}$. La difficulté d'analyse des composés $[VN_1]_{N_2}$ est due au fait qu'ils mettent en jeu une forme verbale. Nous avons discuté ce point plus haut : les travaux de (Scalise, 1994; Villoing, 2002) montrent que la forme verbale de ces composés correspond à la forme non fléchie du verbe, le thème verbal, et qu'il s'agit par conséquent de constructions morphologiques.

Analyse des formations $[N_1N_2]_{N_3}$. L'analyse de formations de type NN est complexe car elles peuvent être formées par la syntaxe et par la morphologie. En outre, la variété de ces formations est grande. (Fradin, 2003, p. 195–206) en donne les exemples suivants :

(a) le *président Mitterrand*, l'*ami Pierre*, le *camarade Staline*

(b) *code Morse*, *projet Delors*, *plan Marshall*, *voitures Peugeot*

(c) *secteur éducation*, *impôt sécheresse*, *disquette programme*, *langage auteur*

(d) *case départ*, *impression laser*, *stylo-bille*, *sortie piétons*, *cigarette filtre*

(e) *auteur compositeur*, *présentateur vedette*, *moissonneuse-batteuse*, *boulanger-pâtissier*

(f) *cité dortoir*, *bateau-lavoir*, *voiture-balai*, *problème clé*

(Fradin, 2003, p. 195–206) recense dans la littérature différents tests linguistiques et définit ainsi les relations sémantiques entre N_1 et N_2. Il apparaît alors que :

- Les composés de type (f) présentant des propriétés distinctes des unités d'autres types et des schémas interprétatifs plus contraints sont engendrés par la morphologie. Du point de vue sémantique, le composé N_3 est un sous-type de N_1. Le N_2 dénote un caractère (fonctionnel ou aspectal) de N_1 pris comme propriété discriminante.
- D'autres unités sont d'origine syntaxique :

- Dans les expressions (a), le N_2, qui est un nom propre, fonctionne comme un appellatif et effectue une identification du N_1.
- Les expressions (b) forment des appellations propres. Le N_2, qui est également un nom propre, désigne celui qui est à l'origine du référent de N_1.
- Dans les expressions (c) et (d), le N_3 désigne également un sous-type de N_1. La relation entre N_1 et N_2 se fonde sur le N_1 en ce sens qu'elle explicite la relation typiquement associée à la dimension de classification dont dépend ce dernier :

  fichier image : *objet de type fichier qui contient une image*,
  sortie piétons : *partie d'un édifice par où sortent les piétons*.

  D'autre part, la multiplicité des schémas interprétatifs de (c) et (d) et l'insertion possible d'une préposition dans (d) font également pencher ces expressions du côté de la syntaxe.
- L'insertion de la coordination dans les expressions (e) indique qu'elles sont également d'origine syntaxique.

De manière générale, les composés morphologiques présentent des schémas interprétatifs plus contraints et ne semblent pas supporter l'insertion des mots grammaticaux, prépositions et coordinations dans les tests présentés ici.

...entre la morphologie flexionnelle et la morphologie constructionnelle

Nous avons vu que la morphologie se compose de deux branches : (1) la morphologie constructionnelle, qui traite de la formation des lexèmes, et (2) la morphologie flexionnelle, qui traite de leur variation, relative au genre, nombre, personne, temps, etc., lors de leur insertion dans les énoncés. En français, la morphologie flexionnelle semble être gouvernée par un modèle différent de celui de la morphologie constructionnelle, même si les deux branches mettent en jeu des affixes. Nous tâchons donc de présenter les principales différences telles que signalées dans (Corbin, 1987, p. 5–6), (Riegel *et al.*, 1998) et (Huot, 2001, p. 105) :

1. Il ne semble pas exister de contraintes sémantiques lors de l'application des règles de flexion, car tous les substantifs ou adjectifs peuvent être mis au pluriel (et au féminin pour les adjectifs). Notons toutefois les *noms massifs ou massiques* (*le blé*, *l'argent*, *le sucre*) qui ne peuvent pas être mis au pluriel. Par contre l'application d'affixes dérivationnels est contrainte par la sémantique des bases et des affixes.
2. Les affixes flexionnels ont une valeur grammaticale : ils comportent des indications pour que les lexèmes soient assemblés de façon à former des énoncés grammaticalement corrects. Les affixes dérivationnels véhiculent une instruction sémantique lors de la formation des lexèmes.
3. La flexion ne détermine pas, en principe, la catégorie grammaticale des mots issus de son adjonction. Par contre, les opérations de la morphologie constructionnelle sont dépendantes de la contrainte catégorielle autant au niveau des bases qu'au niveau des lexèmes construits.
4. Contrairement aux constructions morphologiques, le sens des formes fléchies est toujours prédictible à partir du sens des lemmes, la lexicalisation étant très rare

dans le domaine flexionnel. Il existe néanmoins des cas de lexicalisation des pluriels (définitions de (Robert, 1993)) :

- *noce* : « *L'ensemble des personnes qui assistent à un mariage, qui forment le cortège du mariage.* »
- *noces* : « *Mariage, ensemble des réjouissances qui accompagnent un mariage.* »
- *oreillon* : « *Partie mobile de l'armure de tête, qui protège l'oreille et la joue.* »
- *oreillons* : « *Maladie infectieuse, épidémique et contagieuse d'origine virale, caractérisée par une inflammation des glandes parotides et des douleurs dans l'oreille.* »

5. Les règles flexionnelles s'appliquent après les règles de formation des lexèmes. Les affixes flexionnels apparaissent donc toujours en fin de lexème. Il existe néanmoins des composés où un des composants est fléchi : *casse-pieds*, *épluche-légumes* ; ou des formations où la morphologie constructionnelle intervient après la morphologie flexionnelle : *je-m'en-foutisme*, *fleurdeliser*, *grand-priser*, *TOCqué*.
6. Les affixes flexionnels ne sont pas exclusifs les uns des autres : on peut avoir un affixe de féminin suivi d'un affixe de pluriel (dans le mot *anévrismales* la flexion de pluriel (*-s*) se trouve après la flexion de féminin (*-e*)). Mais leurs possibilités d'apparition en fin de mot sont contraintes, ainsi que leur ordre. Les règles dérivationnelles, par contre, gèrent des contraintes phonologiques, sémantiques et catégorielles.

Rappelons néanmoins qu'il existe des travaux qui montrent que la frontière entre les morphologies flexionnelle et constructionnelle n'est pas étanche. Les ouvrages de G. Corbett lèvent la question sur la nature, flexionnelle ou dérivationnelle, du genre et du nombre (Corbett, 1991; Corbett, 2000). (Aliquot-Suengas, 1996) montre que l'affixe flexionnel *-é* (*fermé*, *mangé*) est étymologiquement lié à trois affixes dérivationnels : (1) *-é(e)* de formation des adjectifs dénominaux (*âgé(e)*, *vertébré(e)*), (2) *-ée* de formation des noms dénominaux (*année*, *cuillerée*) et (3) *-é* ou *-ée* de formation des noms d'état dénominaux (*chanoinée*, *comté*).

...entre l'affixation et la composition

Si certains lexèmes se trouvent à la frontière entre l'affixation et la composition, comme *entérite* page 45, c'est essentiellement à cause du statut ambigu des éléments qui interviennent dans leur formation (*-ite* dans le cas d'*entérite*). Il faut donc définir s'il s'agit d'un élément de composition ou bien d'un affixe. Du point de vue sémantique, le rôle des composants et des affixes est différent (Corbin, 1992). Les composants ont un pouvoir référentiel, et le sens du composé correspond souvent à la combinaison des sens de ses composants en fonction de la règle de formation impliquée. Tandis qu'un affixe exécute une instruction sémantique sur la base. Dans cette opération trois facteurs interviennent : la base, l'affixe et la règle. D'autres principes, mettant en jeu des critères catégoriels et formels, aident à faire la distinction entre les composants et les affixes (Iacobini, 1997; Namer, 2003a) :

- Les composants n'ont généralement pas de position fixe dans un composé, et, selon leur environnement, ils peuvent subir des variations allomorphiques. La base nominale *-gyn-* (du grec *femme*) peut être réalisée de différentes manières, en fonction des contextes morphologiques où elle apparaît :

  gynogenèse, *mysogyne*, *androgyne*, *gynandre*.

  Les composants ont un sens référentiel et peuvent servir de base lors de l'affixation (la base nominale *hépat-* permet de former l'adjectif : *hépatique*) ou être combinés avec d'autres composants pour former des composés (*hépatomégalie*). Ils ont une catégorie stable, inhérente ou bien héritée des bases autonomes auxquelles ils se substituent.
- Les affixes ont une place fixe dans les lexèmes affixés (les suffixes suivent la base et les préfixes la précèdent). Ils n'ont pas de sens référentiel et ne peuvent pas servir de base dans une construction affixale. Ils n'appartiennent à aucune partie du discours, mais ils définissent la catégorie des lexèmes qu'ils forment.
- Dans les cas difficiles, il faut analyser le sens morphologiquement construit du composé pour statuer sur la catégorie, le sens et donc le type de ses éléments. L'exploitation des informations historico-étymologiques proposées dans les dictionnaires peut également donner des pistes d'analyse. Dans ce dernier cas, il faut cependant être prudent (Corbin, 1987), voire contraster les informations de plusieurs dictionnaires.

Nous analysons maintenant des composants d'origine grecque ou latine, ancêtres de composants nominaux, verbaux et les préfixes.

Ancêtres de composants nominaux. Avec la plupart des bases supplétives, ancêtres de noms, l'ambiguïté n'apparaît pas. Par exemple la base supplétive *-gyn-* d'origine grecque a, en français, un sens référentiel (« *femme* »), une catégorie syntaxique (nom) ; elle peut apparaître dans plusieurs endroits dans un composé et sa forme peut alors varier : *gynogenèse*, *mysogyne*, *androgyne*, *gynandre*.

Par contre, les éléments suffixés comme *-ose* et *-ite* ont un statut ambigu entre celui d'un composant et d'un affixe. Ces éléments sont proches des affixes : (1) ils peuvent apparaître dans un lexème uniquement en position finale ; (2) ils ne peuvent pas former de lexèmes par seule combinaison avec un affixe. Mais ils sont également proches des bases : (1) ils ont un sens référentiel : *-ite* signifie « *maladie de nature inflammatoire* » (Manuila *et al.*, 2001), *-ose* peut signifier, selon les contextes, « *maladie non inflammatoire, état, dégénération, prolifération, condition ou excès* » (Dujols *et al.*, 1991; Manuila *et al.*, 2001) ; (2) étant des primitifs supplétifs, ils se substituent aux bases autonomes correspondantes dont ils héritent la catégorie grammaticale (noms).

L'évolution diachronique subie par les unités morphologiques grecques comme *-ite* et *-ose* fournit des éléments de réponse. *-ite* correspondait à l'origine à un suffixe d'adjectivation dénominale (comme *-eux* ou *-ique*) et était employé dans des expressions de type *nosos arthritis*, qui signifiait « *maladie des articulations* ». Avec le temps le nom *nosos*, qui signifie « *maladie* » (Manuila *et al.*, 2001), est devenu facultatif et l'adjectif *arthritis* porteur du sens de toute l'expression (Pacak *et al.*, 1980). Depuis le suffixe *-ite* s'est spécifié sémantiquement : il signifie actuellement « *maladie de nature inflammatoire* » (Manuila *et al.*, 2001). Tandis que *arthrite* signifie « *inflammation des articulations* » (Manuila

et al., 2001). *-ose* et *-ase*, de leur côté, correspondaient en grec à un suffixe de nominalisation déverbale (comme *-ment* ou *-tion*) (Wolff, 1987). Leur évolution est comparable avec celle de *-ite*.

Nous considérons ces éléments comme composants non autonomes supplétifs de noms, essentiellement à cause de leur charge sémantique importante et du fait qu'ils ont des correspondants autonomes dans la langue. Mais il s'agit effectivement d'un cas limite entre l'affixation et la composition : les affixes en question ont acquis au cours de leur histoire des propriétés sémantiques de composants. D'ailleurs (Corbin & Paul, 1999, p. 62–65) ne classent ces unités morphologiques ni parmi les affixes ni parmi les composants. Ils les appellent *classifieurs suffixoïdes*. Car ces unités permettent d'identifier, dans la forme des lexèmes construits, la classe superordonnée à laquelle appartiennent les référents de ces lexèmes : *-ite* réfère aux maladies inflammatoires, *-ase* à la classe des enzymes, etc.

Descendants de composants verbaux. Le statut morphologique des éléments descendant d'un verbe (*-phage*, *-cide*, *-fère*, *-gène*) est plus proche de celui des composants : (1) ils ont un sens référentiel (*-phage* signifie « *manger* », *-cide* signifie « *abattre, tuer* ») ; (2) étant des primitifs supplétifs, ils se substituent aux bases autonomes correspondantes dont ils héritent la catégorie grammaticale (verbes) ; (3) ils peuvent apparaître dans des positions différentes dans un composé. La plupart de ces éléments n'apparaissent que dans les composés et ne forment pas de lexèmes par affixation (Namer, 2003a). Là encore, nous les considérons comme composants non autonomes supplétifs.

Les préfixes et leurs ancêtres. Les préfixes ne constituent pas une classe homogène : si certains d'entre eux ont toujours connu des emplois non autonomes (*re-*, *dé-*, *dys-*), d'autres proviennent d'éléments de composition (*archéo-*, *pseudo-*) (Iacobini, 1997) ou de prépositions et d'adverbes (*sur-*, *entre-*, *sans-*) (Amiot, 2001). Leur emploi en tant que préfixes témoigne de leur grammaticalisation, dont le degré varie selon les unités morphologiques.

(Iacobini, 1997) propose une série de critères pour distinguer un composant d'un affixe, dont ceux déjà cités ci-dessus : position fixe ou non dans les lexèmes, possibilité de combinaison avec un autre affixe lors de la formation de lexèmes, sémantique et catégorie grammaticale. Par ailleurs, la distinction entre un adjectif et un préfixe peut être faite grâce aux tests sémantiques (Namer, 2003a) :

1. si le premier élément possède un sens référentiel, et donc si son sens est combiné avec celui de la base nominale, il s'agit d'un adjectif, et le sens construit est une composition des sens des parties ;
2. si le premier élément exerce une prédication sur la base nominale et sert à former un nouveau sens à partir de celui du N, il s'agit d'un préfixe.

On distingue ainsi deux types pour *micro-* (Corbin, 1993) :

microfaune (« *faune de petite taille* ») : *micro-* signifie « *petit* » ⇒ composant adjectival
microseconde (« *fraction de seconde* ») : *micro-* signifie « *fraction de* » ⇒ préfixe

Cet exemple montre qu'il existe deux entrées de *micro-* ou bien que ces deux constructions ont été formées à des moments différents de la grammaticalisation du composant *micro-*. Le vacillement d'unités morphologiques de ce type se trouve à l'intérieur de la morphologie constructionnelle.

Lorsque les ancêtres de préfixes sont les prépositions et les adverbes, leur niveau de grammaticalisation peut également être différent (Amiot, 2001). Si cette grammaticalisation est achevée, comme pour *sur-*, *sous-* ou *contre-*, il s'agit d'affixation. Sinon, les constructions sont sémantiquement hétérogènes (comme avec *sans-* : *sans-papier*, *sans-abri*, *sans-dos*). Dans ce dernier cas, (Amiot, 2001) propose de considérer les formations comme des composés. Sachant que la morphologie constructionnelle, et donc la composition, n'ont pas accès aux mots grammaticaux les éléments en cours de grammaticalisation doivent avoir un statut intermédiaire (Amiot, 2004).

Notons également la discussion proposée par (Fradin, 2003, 196-197) notamment face aux constructions comme :

surpasser, *surestimer*, *surprendre*
sous-payer, *sous-entendre*, *sous-louer*
contre-attaquer, *contredire*, *contrefaire*

où les prépositions *sur*, *sous*, etc. ne se construisent pas avec les verbes. Par contre, les prépositions qui se construisent syntaxiquement avec les verbes (*à*, *de*) n'apparaissent jamais dans des constructions semblables. Cette constatation renforce l'analyse sémantique de (Amiot, 2001) et éloigne ces constructions des analyses syntaxiques.

3.2 Morphologie et TAL

Nous présentons ici les travaux du TAL qui investissent le domaine de la morphologie. Il s'agit en particulier de l'étiquetage morphosyntaxique (sec. 3.2.1), de la constitution de lexiques morphologiques (sec. 3.2.2) et de l'analyse morphosémantique des lexèmes (sec. 3.2.3). Nous présentons également l'utilisation de ressources morphologiques pour le traitement de la variation terminologique (chap. 2) et l'acquisition automatique de ressources morphologiques (chap. 4).

3.2.1 Étiquetage morphosyntaxique

Les *étiqueteurs morphosyntaxiques* proposent d'assigner une étiquette grammaticale, munie parfois de traits morphologiques, aux mots[11] d'une phrase. Il en existe pour de nombreuses langues. Les étiqueteurs sont souvent couplés avec la fonction de *lemmatisation*, qui consiste à définir, en même temps, les lemmes des mots. Ils fonctionnent souvent avec un lexique de mots déjà connus. Ce lexique comporte une liste de mots munis d'étiquettes (et de traits) morphosyntaxiques. En cas d'*ambiguïté*, lorsque plus d'une étiquette et/ou

[11]Nous employons le terme *mot* au sens informatique : il d'agit des unités lexicales obtenues après la segmentation d'une phrase sur les espaces blancs et éventuellement d'autres signes de ponctuation.

de traits sont disponibles pour un mot, le système doit résoudre cette ambiguïté et assigner l'étiquette et les traits les plus probables dans un contexte donné. Le système doit également assigner une étiquette aux mots inconnus. Différentes techniques peuvent alors être employées. (de Loupy, 1994) en distingue trois, selon le mode d'acquisition de règles : techniques à base de règles manuelles, techniques statistiques et techniques mixtes. Nous nous focalisons ici sur deux types de règles qui peuvent être utilisées pour la résolution d'ambiguïtés et l'étiquetage des mots inconnus : les *règles contextuelles*, qui s'appuient sur les mots proches du mot courant, et les *règles lexicales*, qui s'appuient sur la forme du mot courant. L'étiqueteur conçu par (Brill, 1993), fait usage des règles des deux types, mais ce sont les règles lexicales qui nous intéressent particulièrement car elles établissent un lien avec le système morphologique d'une langue.

Les règles probabilistes lexicales de `Brill` se repèrent par rapport à la finale des mots, qui est de quatre caractères au maximum dans les expériences de la sec. 4.4 page 77. Les chaînes finales correspondent souvent aux affixes flexionnels et parfois dérivationnels :

```
ique   hassuf      4  ADJ :sg  51     suffixe d'adjectivation
es     hassuf      2  SBC :pl  23.5   flexion de pluriel
es     deletesuf   2  ADJ :pl  6      flexion de pluriel
al     hassuf      2  ADJ :sg  13     suffixe d'adjectivation
ale    hassuf      3  ADJ :sg  17     suffixe d'adjectivation + flexion de féminin
```

Quelques explications sur le format : la première colonne contient la chaîne finale qui indique la marque flexionnelle et/ou dérivationnelle par rapport à laquelle se repère le système ; la deuxième colonne indique l'opération effectuée par le système lors de l'assignation des étiquettes (`hassuf` : le mot contient la chaîne finale donnée, `deletesuf` : la chaîne finale peut ne pas apparaître dans ce mot ailleurs dans le texte analysé) ; la troisième colonne indique la longueur de la chaîne de caractères qui sert de repère ; la quatrième la catégorie morphosyntaxique que cette règle va associer au mot ; la cinquième colonne indique sans doute le rang de fiabilité de la règle lexicale. Dans la dernière, nous avons ajouté les commentaires sur la nature des chaînes finales : correspondent-elles aux affixes flexionnels ou dérivationnels. Notons qu'il n'est pas toujours aussi aisé de faire cette caractérisation. Par exemple dans la règle :

```
ires hassuf 4 ADJ :pl 6
```

la chaîne qui sert de repère correspond sans doute à une sous-chaîne du suffixe dérivationnel *-aire* suivi de l'affixe de pluriel *-s*.

Dans la même optique d'étiquetage morphosyntaxique, un autre système conçu par (Spyns, 1994) pour l'analyse des mots médicaux néerlandais inconnus du lexique est également basé sur des marqueurs morphologiques. Dans ce système, la partie droite du mot, qui correspond le plus souvent à son affixe flexionnel, indique sa catégorie et ses traits morphologiques. Remarquons que les affixes dérivationnels ont également un pouvoir de catégorisation.

Le lemmatiseur `Flemm` (Namer, 2000)[12], également basé sur l'application de règles, se propose de corriger et d'affiner l'étiquetage réalisé par d'autres outils actuellement : `Brill`

[12] `www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm`

et `TreeTagger` (Schmid, 1994). Mais les règles de `Flemm` gèrent la variation des mots et non leur agencement dans une phrase : la correction des erreurs est plus locale, elle est faite en fonction de la graphie des formes fléchies. C'est donc la graphie qui conduit à la sélection d'une règle par défaut pour la correction de l'étiquetage et le traitement des mots inconnus. Un autre point spécifique du système consiste en l'utilisation d'une liste d'exceptions dans l'application de règles morphologiques.

3.2.2 Production de lexiques morphologiques

Un *lexique morphologique* se présente souvent sous forme de couples {*lemme*, *forme*} pour les flexions et sous forme de couples {*base*, *dérivé*} pour les dérivations. Signalons que l'établissement de tels lexiques peut être basé sur les données morphologiques acquises automatiquement (sec. 4.2 page 61).

L'établissement de lexiques flexionnels est bien avancé dans plusieurs langues, dont le français. Par exemple, le système `Intex` (Silberztein, 1993) propose de tels lexiques pour le français et une dizaine d'autres langues. La base `Celex` (Burnage, 1990) couvre l'anglais, l'allemand et le néerlandais.

La morphologie dérivationnelle par contre n'est pas aussi bien investie et très peu de langues ont des ressources dérivationnelles disponibles. On peut ainsi citer de nouveau la base `Celex` (Burnage, 1990) qui propose de riches ressources pour le néerlandais, l'anglais et l'allemand : plus de 124 000 lemmes couverts en néerlandais (version 3.1 de 1990), et presque 52 000 lemmes en allemand (version 2.5 de 1995). En français, il existe des travaux isolés. Par exemple, l'établissement d'une liste de noms déverbaux (Berche *et al.*, 1997), où les données acquises automatiquement avec l'outil de désuffixage `findaffix` d'`Unix` sont validées manuellement. Une investigation plus fondamentale du domaine dérivationnel français par le TAL a commencé avec le projet `Franlex` (Dal & Jacquemin, 1999) et continue avec l'analyseur DeriF (Namer, 1999) du projet `MorTAL` (Dal *et al.*, 1999; Hathout *et al.*, 2001). L'approche qui se trouve à la base de ces travaux est inspirée par la morphologie constructionnelle. Dans cette approche, parallèlement à la construction de la forme des mots, la contrepartie sémantique est également traitée : le sens d'un mot construit peut ainsi être calculé à partir de sa base, de l'affixe et de la règle de construction des mots (Corbin, 1992). L'opération inverse, calcul du sens de la base à partir d'une forme dérivée, est également disponible dans DeriF (Namer, 2002). Actuellement, `MorTAL` permet de traiter la préfixation avec *re-*, *é-*, *a-*, *dé-* et *in-* ; la suffixation avec *-able*, *-ité*, *-et(te)*, *-is(er)*, *-ifi(er)*, *-aie*, *-aille*, *-age*, *-oir*, *-tion*, *-ment* et *-eur* ; et la conversion (Namer, 2002). L'analyseur DeriF s'attaque actuellement à la composition (Namer & Zweigenbaum, 2004).

Pour la langue médicale, la seule base de données morphologiques publiquement disponible se trouve dans le Specialist Lexicon d'UMLS (McCray *et al.*, 1994), elle permet de traiter la flexion et l'affixation de la langue médicale anglaise. Cette base se compose d'un lexique, de règles de formation et de variation des mots réguliers. Les allomorphies, les bases supplétives et les variantes orthographiques sont listées explicitement dans des tables spécifiques. Les programmes pour la génération de formes fléchies et affixées et pour la réduction de ces formes à leur lemme ou base sont également disponibles. Dans la

continuité des travaux d'UMLS, il existe des initiatives de constitution de lexiques morphologiques (flexion et affixation) du domaine médical en allemand (Schulz *et al.*, 1999) et en français (Zweigenbaum *et al.*, 2003a).

3.2.3 Systèmes d'analyse morphosémantique

Il existe dans le domaine médical une tradition d'analyse morphosémantique de lexèmes composés. Les travaux débutent avec (Pacak *et al.*, 1980), où les auteurs analysent les lexèmes en *-ite* et en proposent différents patrons interprétatifs en fonction des classes sémantiques auxquelles appartiennent les composants du lexème analysé. Le travail recense environ 750 composants d'origine grecque ou latine. Chacun d'eux est classé sous un des axes sémantiques proposés par la Systematized Nomenclature of Pathology (SNOP, 1965), ancêtre de la SNOMED : topologie, morphologie, étiologie ou fonction. Les composants supplétifs sont 'lexicalement interprétés' et remplacés par les composants autonomes correspondants. *-ite* est ainsi remplacé par *inflammation*. Les lexèmes sont analysés de la gauche vers la droite. À chaque segmentation le composant le plus long est recherché. En fonction des types sémantiques des composants, un ou plusieurs patrons sont appliqués. Les auteurs ont établi au total 14 patrons sémantiques. Cette étude, étendue de l'anglais à d'autres langues (français, italien, allemand et tchèque), a montré que le comportement des composants analysés est assez similaire à travers ces langues. D'autres travaux investissent les lexèmes avec d'autres finales : *-tomie*, *-stomie*, *-ectomie* (Wolff, 1987) ; *-ose* (Dujols *et al.*, 1991). Ou bien ils visent une couverture maximale de composés médicaux (Dujols *et al.*, 1991; Lovis *et al.*, 1995; Schulz *et al.*, 1999; Namer & Zweigenbaum, 2004).

La majorité de ces travaux font du *paraphrasage* : un lexème est décomposé et paraphrasé selon le patron sémantique reconnu. Quelques exemples de (Pacak *et al.*, 1980) :

iridochoroiditis	*inflammation of iris and choroid*
keratoconjunctivitis	*inflammation of conjunctiva with cornification*
laminitis	*inflammation of lamina*

Par contre le travail de (Schulz *et al.*, 1999) reste dans la lignée des travaux en recherche d'information : les termes médicaux sont décomposés en bases et composants qui sont alors traités alphabétiquement comme des 'sacs de composants'. La structure des termes telle que gérée par les patrons interprétatifs (Pacak *et al.*, 1980; Wolff, 1987; Dujols *et al.*, 1991; Lovis *et al.*, 1995) ou les règles de formation de lexèmes (Namer & Zweigenbaum, 2004) n'est pas prise en compte. Dans cette approche, il suffit que deux termes comportent des bases et composants équivalents pour être considérés équivalents eux-mêmes :

Hyperemesis gravidarum :
⇒ (1) *hyper*, *emet*, *gravit*
⇒ (2) *hyper/high/augmented*, *emet/vomit*, *gravid/pregnan/gest*
Excessive vomiting in pregnancy :
⇒ (1) *excessiv*, *vomit*, *pregnan*
⇒ (2) *excessiv/high/augmented*, *vomit/emet*, *pregnan/gravid/gest*

3.3 Conclusion

Notre objectif étant l'utilisation de ressources morphologiques pour le traitement de la variation des termes, il faut définir quels procédés des morphologies flexionnelle et constructionnelle il convient de prendre en compte.

L'apport sémantique de la flexion est faible : le genre, le nombre, la personne, etc. ne modifient généralement pas la sémantique des lexèmes et la lexicalisation des formes fléchies est très rare. Avec les trois procédés principaux de la morphologie dérivationnelle (affixation, conversion et composition), l'apport sémantique dans les lexèmes construits est plus important. Lors de la suffixation, le suffixe opère une instruction sémantique sur la base mais la sémantique du lexème construit reste proche de celle de la base, sauf pour les cas de glissement de sens qui ne sont pas du ressort de la morphologie mais de l'usage. Par contre les préfixes ne constituent pas une classe homogène : elle est nourrie par les adverbes, les prépositions et les éléments de composition. Très peu d'éléments sont d'origine préfixale. Selon le degré de grammaticalisation des nouvelles unités, les constructions doivent être considérées comme des affixations ou des compositions. La conversion semble présenter des schémas sémantiques proches de la suffixation. Enfin, avec la composition l'équivalence semantique est rompue : il existe, selon la position des composants, une relation hyperonymique ou bien des relations spécifiques, entre eux.

Pour le traitement de la variation terminologique la flexion, la suffixation et la conversion semblent donc être les plus pertinentes. Par contre, la préfixation et la composition, qui apportent des ruptures sémantiques, ne sont pas souhaitables, et si les ressources correspondantes sont à utiliser elles doivent l'être de manière contrôlée.

Mais, comme nous l'avons noté, les ressources morphologiques sont plutôt lacunaires en français, et encore plus en français médical. Le chapitre suivant est donc consacré à l'acquisition de ressources morphologiques.

Chapitre 4

Induction de connaissances morphologiques à partir de terminologies structurées

Dans le chapitre 2, nous avons montré que la variation terminologique apparaît dans différents contextes et nous nous sommes alors proposé de la traiter en nous concentrant surtout sur la variation morphologique. Dans le chapitre 3, nous avons décrit les moyens mis en œuvre par la morphologie lors de la formation des lexèmes et de leurs variations. Nous avons également noté les procédés morphologiques potentiellement utiles pour le traitement des variantes de termes : la flexion, la suffixation et la conversion. La préfixation et la composition, présentant souvent des ruptures sémantiques entre les lexèmes base et les lexèmes construits, doivent être utilisées sous contraites sémantiques. Dans ce chapitre, nous présentons nos travaux sur l'acquisition de ressources morphologiques. De telles ressources ne sont pas disponibles en français, sinon de manière fragmentaire. La situation est encore plus lacunaire lorsqu'il s'agit d'une langue de spécialité comme la langue médicale. Ces travaux correspondent ainsi au premier objectif que nous nous sommes fixé : l'acquisition de ressources morphologiques. Nous avons noté auraparavant qu'il existe deux approches pour le traitement de la variation des termes : une approche de type formel qui consiste en suppression systématique de suffixes en fin des mots et une approche qui implique des connaissances morphologiques motivées linguistiques. Nous misons sur l'application de l'approche linguistique et donc sur l'utilisation de ressources motivées morphologiquement. De telles ressources ont l'avantage de refléter le système morphologique d'une langue (Gaussier, 1999). Nous soutenons également que ces ressources doivent être adaptées au domaine scientifique ou technique traité pour fournir de meilleurs résultats.

Dans la suite de ce chapitre, nous présentons nos motivations pour l'acquisition automatique de ressources morphologiques (sec. 4.1) et les travaux existants (sec. 4.2). Ensuite nous décrivons les méthodes automatiques que nous proposons (sections 4.3 à 4.6) en précisant à chaque fois le matériel nécessaire et en analysant les résultats obtenus. Notons qu'un échantillon de ces résultats peut être consulté à l'annexe A. Nous terminons le chapitre en dressant un bilan et en marquant quelques perspectives (sec. 4.7).

4.1 Introduction

Le traitement de la variation morphologique des termes demande de disposer de ressources morphologiques adaptées de préférence à la langue de spécialité étudiée. Mais une analyse de telles ressources montre qu'elles sont plutôt lacunaires. Les ressources flexionnelles sont les plus étudiées dans plusieurs langues à travers, par exemple, de nombreux outils d'étiquetage et de lemmatisation (Brill, 1993; Schmid, 1994; Namer, 2000; Silberztein, 1993; Mertens, 1999). Par ailleurs, la base `Celex` (Burnage, 1990) comporte une description des systèmes flexionnels de l'anglais et de l'allemand. Quant à l'affixation de la langue générale[1], elle est bien représentée en anglais et en allemand, toujours grâce à la base `Celex`. En français par contre les ressources pour le traitement de l'affixation est actuellement en cours de développement dans le cadre du projet `MorTal` (Dal *et al.*, 1999; Hathout *et al.*, 2001). Quant à la langue médicale, la seule base de données morphologiques publiquement disponible se trouve dans le Specialist Lexicon d'UMLS. Cette base permet de traiter la flexion et l'affixation de l'anglais. La composition semble être absente dans les applications de la langue générale. Dans la langue médicale, les ressources sur la composition sont constituées essentiellement dans le cadre de travaux d'analyse morphosémantique des termes médicaux (sec. 3.2.3 page 57), elles ne sont pas publiques et restent réservées à l'utilisation des équipes qui les ont conçues. Notons que sur l'exemple d'UMLS il existe des travaux de constitution de ressources morphologiques pour la langue médicale autre que l'anglais : l'allemand (Schulz *et al.*, 1999) et le français (Zweigenbaum *et al.*, 2003a). Ces ressources en cours d'élaboration ont pour but de couvrir la flexion, l'affixation et la composition. À l'heure actuelle, il n'existe donc pas de ressources morphologiques disponibles en français, que ce soit pour la langue générale ou médicale.

Afin de pouvoir traiter les variations morphologiques des termes médicaux en français, nous avons à satisfaire un premier objectif de notre travail : proposer une méthode pour l'acquisition automatique de ressources morphologiques et constituer de telles ressources. Parmi les procédés morphologiques notés comme intéressants pour notre tâche, nous avons la suffixation et la conversion (en cas de données étiquetées) pour la formation des lexèmes, et la flexion pour leurs variations. La préfixation et la composition, impliquant des ruptures sémantiques, doivent être utilisées sous contraintes sémantiques.

Dans la suite de ce chapitre, nous verrons que les travaux en TAL et en acquisition emploient des acceptions des termes de la morphologie différentes de celles que nous avons vues dans le chapitre 3. C'est le cas des termes comme *préfixe* et *suffixe*. L'hypothèse souvent partagée dans les travaux d'acquisition de ressources morphologiques suppose que les

[1]Dans notre travail, nous faisons une opposition entre la langue générale et la langue de spécialité, ici, la langue médicale. Une *langue de spécialité* couvre un domaine scientifique ou technique. Elle comporte le lexique spécifique à ce domaine, sa terminologie, mais elle peut également présenter des particularités syntaxiques, morphologiques ou autres. La *langue générale*, par contre, n'est pas consacrée à un domaine de spécialité précis. Elle reflète souvent l'usage quotidien de la langue, tel que fait par exemple dans la presse grand public. Les articles des journaux ne sont évidemment pas complètement neutres de ce point de vue : ils concernent différents domaines de spécialités (agriculture, cuisine, politique, etc.), mais on considère généralement que le niveau de leur spécialisation n'est pas élevé et le mélange de différents domaines les neutralise les uns par rapport aux autres.

mots reliés morphologiquement contiennent une *racine commune*. Du point de vue de l'acquisition automatique, une racine commune est représentée par une *chaîne de caractères commune* souvent en position initiale dans les mots comparés. Cette chaîne initiale commune est souvent appelée 'préfixe' et les chaînes finales non communes 'suffixes'. Il va de soi que les segments de mots trouvés avec les techniques automatiques et appelés 'préfixe' ou 'suffixe' ne correspondent pas toujours au découpage morphologique réel de la langue. Et il serait donc plus correct de parler de *chaînes de caractères* initiales ou finales. De la même manière, différentes segmentations d'un même mot conduisent vers l''allomorphie'. Il ne s'agit bien sûr pas de l'allomorphie morphologique.

4.2 Travaux en acquisition automatique de connaissances morphologiques

Les traitements 'standards' de la variation morphologique, faits au moyens des désuffixeurs comme ceux de Lovins (Lovins, 1968) et Porter (Porter, 1980), ne permettent pas de traiter toutes les variations nécessaires et surtout ils ne sont pas adaptés aux différents domaines techniques. Des travaux plus récents cherchent à proposer des outils et des ressources plus performants quant à la couverture des phénomènes linguistiques traités et à leur spécificité à un domaine technique ou scientifique. La plupart de ces travaux visent la constitution de ressources morphologiques à travers la segmentation des mots et la détection de chaînes de caractères communes. Mais le fait que les mots partagent une chaîne initiale commune ne garantit pas en soi qu'ils soient reliés sémantiquement. Par exemple, un outil `Unix` comme `findaffix` permet de repérer les mots qui partagent une chaîne initiale commune d'une longueur donnée. Parmi ceux qui partagent une chaîne initiale commune de trois caractères, nous avons, dans une liste de 8 874 formes du domaine médical, 17 formes qui commencent par *tro*. `findaffix` les regroupe en une seule famille :

> *trocart*, *trochléaire*, *trois*, *troisième*, *troisièmement*, *trompe*, *trompes*, *tronc*, *troncs*, *trop*, *trophique*, *trophoblaste*, *trophoblastique*, *tropicale*, *trou*, *trouble*, *trouvé*.

Mais on peut y distinguer en réalité au moins 11 familles morphologiques (Grabar & Zweigenbaum, 1999a). Le souci principal des travaux en acquisition automatique de ressources morphologiques consiste donc à trouver un contexte sémantique qui garantisse que les mots reliés formellement soient également reliés sémantiquement. De nombreuses pistes sont exploitées.

4.2.1 Exploitation de dictionnaires électroniques

Les dictionnaires peuvent fournir un contexte sémantique très contraint. Les résultats dépendent alors de la richesse et de la complétude de ces dictionnaires.

(Krovetz, 1993) utilise la version électronique du *Longman Dictionary of Contemporary English*. Ayant pour but de proposer un désuffixeur plus performant pour le traitement de la flexion et de l'affixation, l'auteur exploite ce dictionnaire de deux manières :

- sa structure est utilisée pour le repérage de mots reliés formellement et sémantiquement : si une entrée est proche formellement d'un des mots de sa définition, ils sont considérés être proches sémantiquement aussi. Par exemple, la définition de *cylindrical* contient le mot *cylinder* ;
- un système de désuffixation à la Porter est appliqué itérativement, mais avant chaque itération le système vérifie si le mot désuffixé correspond à une des entrées du dictionnaire. Si c'est le cas, la désuffixation s'arrête. Le but est alors de faciliter l'interprétation sémantique en réduisant les formes traitées plutôt à un mot existant qu'à une racine.

Le même dictionnaire est utilisé dans (Pentheroudakis & Vanderwende, 1993). Tout d'abord, les auteurs établissent un lexique flexionnel en utilisant les tables de conjugaison et les listes d'exceptions proposées par le dictionnaire. Pour la constitution du lexique dérivationnel, plusieurs informations sont exploitées :

- comme dans (Krovetz, 1993), le système vérifie si dans la définition d'une entrée il y a des mots qui lui sont reliés formellement : comme la définition de *journalism* comporte le mot *journal*, ces deux mots sont supposés être reliés sémantiquement ;
- par ailleurs, si une entrée dictionnairique comporte une liste de suffixes le système est capable de former les mots suffixés. Par exemple l'article *journalism* indique les suffixes *-istic/Adj* et *-istically/Adv* et le système peut alors construire *journalistic/Adj* et *journalistically/Adv* ;
- et enfin, les entrées sont considérées proches sémantiquement si leurs définitions comportent des mots proches formellement. Comme c'est le cas de *geographer* et *geography* et de *cartographer* et *cartography*.

(Hathout, 2001) utilise un dictionnaire de synonymes. Les mots simples du dictionnaire sont comparés deux par deux et retenus s'ils présentent une alternance de suffixes vérifiée sur au moins une autre paire de mots. Par exemple, l'alternance *er|ation* est retenue car elle est vérifiée avec trois bases verbales : *adorer*, *vénérer* et *permuter*. Les mots mis en relation au moyen des alternances, ou des règles, doivent en outre partager un préfixe d'au moins 3 caractères : les mots du couple {ador*er*, ador*ation*} partagent un préfixe de 4 caractères *ador-*. Les alternances de suffixes permettent de construire un graphe dans lequel l'auteur cherche les analogies constructionnelles à travers les quadruplets de mots comme : {*fouiller*, *fouilleur*} / {*fureter*, *fureteur*}. L'existence de tels quadruplets confirme la relation morphologique. L'extraction des quadruplets peut être faite de manière stricte ou lâche. L'exemple ci-dessus représente un quadruplet strict : les deux alternances de suffixes sont identiques. Par contre, les quadruplets {*changer*, *changement*} / {*permuter*, *permutation*} et {*réunir*, *réunion*} / {*mélanger*, *mélange*} sont lâches, car les suffixes en alternance ne sont pas identiques, mais correspondent à des opérations morphologiques équivalentes.

4.2.2 Exploitation de corpus

Les corpus de textes permettent d'effectuer une acquisition de ressources morphologiques spécifiques à un domaine.

Dans (Xu & Croft, 1998), l'hypothèse de proximité sémantique pose que des mots reliés morphologiquement apparaissent souvent dans une fenêtre de mots d'une certaine taille (50 à 100). Les mots reliés morphologiquement sont détectés avec un désuffixeur 'agressif' basé sur le repérage des n premiers caractères dans les mots ($n = 3$). Ensuite un calcul d'associations statistiques entre les mots est appliqué pour filtrer les cooccurrents. Les paires de mots filtrés peuvent être regroupées en familles avec des relations dérivationnelles et flexionnelles :

uniformity, *uniformily*, *uniformed*, *uniforms*, *uniform*.

Le mot le plus court de la famille (ici *uniform*) est considéré comme sa racine. Selon les expériences, l'exploitation d'un corpus spécialisé permet effectivement d'obtenir les données spécifiques au corpus et au domaine. Par exemple, dans un corpus de finances, les mots *stock* et *stocks* se trouvent reliés, mais ce n'est pas le cas dans un corpus d'histoire médiévale où *stocks* signifie un outil de torture. La portabilité des données obtenues à partir d'un corpus sur un autre dépend de la similarité entre ces deux corpus. Cette approche fournit des données d'autant plus fiables que le corpus est représentatif du domaine ou de l'application, suffisamment volumineux et répétitif. Les mots rares, par contre, reçoivent des valeurs d'association plutôt faibles et ne sont pas associés.

Une expérience similaire, appliquée sur un corpus médical français, a été effectuée dans (Zweigenbaum *et al.*, 2003b). La chaîne initiale commune est de 4 caractères et la fenêtre de mots varie entre 50 et 200 mots. Pour la segmentation des mots et la détection des préfixes et des suffixes, les auteurs utilisent la méthode que nous décrivons dans la sec. 4.3.

4.2.3 Exploitation d'un corpus et d'une liste de termes

(Jacquemin, 1997a) travaille dans un contexte plus restreint : il utilise une liste de termes contrôlés, un corpus et un ensemble de règles de réécriture morphosyntaxique des termes. L'auteur exploite le fait que dans les termes contrôlés et ceux des corpus les mots peuvent varier morphologiquement tout en sauvegardant les dépendances syntaxiques. Ainsi, les paires de termes :

{*gene expression*, *genic expression*}
{*gene expression*, *genes expression*}

donnent les paires de mots {*gene*, *genic*} et {*gene*, *genes*}.

Les meilleurs résultats sont obtenus avec une distance de 2 mots entre les mots d'un terme, sans permutation et avec une longueur de suffixe pouvant aller de 2 à 5 caractères. En plus des relations morphologiques fréquentes dans les variantes de termes (flexion, affixation), cette approche permet de repérer :

- des relations morphologiques rares : {*North Africa*, *Northern Africa*}, {*pulsating flow*, *pulsatile flow*}, {*biliary cirrhosis*, *biliary cirrhosia*},
- des variantes et erreurs orthographiques : {*coccidioides immitis*, *coccidiodes immitis*},
- des variations extra-morphologiques : {*alpha lipoprotein*, *alpha1 lipoprotein*}.

Les paires de mots reliés sont ensuite regroupées en familles morphologiques. Étant basée sur un corpus et un lexique spécialisés, cette approche propose également des paires de mots spécifiques au domaine.

4.2.4 Exploitation des couples de suffixes et de leurs fréquences

Dans (Gaussier, 1999), lors de la comparaison des mots, c'est la fréquence du couple des suffixes qui assure la fiabilité du contexte sémantique. Si cette fréquence est suffisamment élevée (au moins 2), les mots mis en relation sont considérés comme étant reliés formellement et sémantiquement. Par exemple, les mots *deplorable* et *deploringly* fournissent à la segmentation le couple de suffixes *able|ingly*. Comme ce couple se retrouve dans d'autres paires de mots, il est retenu comme marqueur de la relation morphologique entre les mots. Les paires de mots sont par ailleurs groupées en famille. Là encore, ce sont les suffixes qui assurent la cohérence sémantique des familles. L'auteur suppose qu'il existe une compatibilité entre les suffixes : l'apparition d'un suffixe coïncide avec l'apparition d'un autre suffixe et les suffixes de différentes familles ne se mélangent pas.

4.2.5 Exploitation d'approches distributionnelles

Les résultats fournis par les approches distributionnelles sont filtrés avec des pondérations statistiques ou autres informations.

Pour découvrir les affixes les plus fréquents d'une langue, (Déjean, 1998) applique l'approche distributionnelle. Ainsi, si après une séquence de caractères se trouve un nombre important de caractères différents, le système détecte une frontière entre les morphèmes :

- *direc-* peut être suivi uniquement de *t*, donc il n'y a pas de frontière ;
- *direct-* peut être suivi de *i*, *l*, *o* et *e* (comme dans direct*ion*, direct*ly*, direct*or* ou direct*ed*), donc il y a une frontière.

Pour enrichir la liste d'affixes, l'auteur regarde si les affixes déjà repérés peuvent alterner avec d'autres. Ainsi, à partir des mots où le système reconnaît les affixés déjà répertoriés (light, light*s*, light*ed*, light*ing*, light*ly*, light*er*), il est possible d'en apprendre d'autres (light*ness*, light*est*, light*en*) et d'élargir ainsi la famille.

(Schone & Jurafsky, 2001) entame l'acquisition de suffixes sur un corpus avec une approche semblable à celle de (Déjean, 1998), mais les préfixes les plus fréquents sont identifiés en même temps. Le système est capable de traiter non seulement la préfixation ou la suffixation, mais également la *circumfixation*, propre aux langues comme l'allemand et le néerlandais, qui consiste en un ajout simultané d'un préfixe et d'un suffixe. Un premier filtrage des paires de mots est fait avec Latent Semantic Analysis (LSA). Le résultat de ce filtrage est ensuite confirmé avec une mesure de similarité et pondéré par la fréquence des affixes et le contexte syntaxique des mots. Finalement, le système effectue le regroupement des paires de mots en familles sur la présence de mots communs.

(Urrea, 2000) effectue également une segmentation des mots avec l'approche distributionnelle proche de celle des deux travaux précédents. Les segmentations obtenues sont évaluées avec plusieurs calculs, dont celui utilisé dans (Pirrelli & Yvon, 1999) (voir le paragraphe suivant). La moyenne des différents scores donne un score global aux candidats suffixes et préfixes. Les affixes flexionnels, qui sont plus redondants en langue, se laissent découvrir plus facilement que les affixes dérivationnels.

4.2.6 Application d'algorithmes d'apprentissage

Les algorithmes d'apprentissage sont également appliqués en acquisition de ressources morphologiques. Ces algorithmes sont entraînés sur une base d'exemples qu'ils généralisent ensuite aux mots inconnus de la base initiale. Leurs performances dépendent des régularités et des analogies dans la langue, et de la complétude de la base d'apprentissage.

(van den Bosch *et al.*, 1996) trouvent la base d'exemples nécessaires dans le lexique néerlandais `Celex` (Burnage, 1990).

(Pirrelli & Yvon, 1999) proposent un algorithme d'apprentissage 'paresseux', qui n'est pas sans rappeler l'expérience de (Hathout, 2001). Dans un 4-tuple de type *eat:eater = cheat:X*, ils peuvent calculer la forme manquante *X* (*cheater*) en supposant qu'il existe une analogie dans la formation de cette forme en référence aux formes déjà connues (ici, par analogie avec *eat:eater*, il s'agit de l'ajout du suffixe *-er* à la forme *cheat*). Les informations à calculer peuvent être plus évoluées et à chaque forme on peut attacher sa prononciation, son étiquette morphosyntaxique ou autre.

(Theron & Cloete, 1997) se basent sur un ensemble de paires de mots reliés morphologiquement pour déduire les règles de leur formation. Ces règles sont exprimées dans le modèle de la morphologie à deux niveaux (Koskenniemi, 1983).

4.2.7 Exploitation des règles de formation des lexèmes

Les règles de formation de lexèmes issues d'études linguistiques peuvent également servir de base fiable à l'acquisition de ressources morphologiques. Ces règles sont souvent surgénératrices et le résultat doit être filtré.

Les règles proposées par la morphologie constructionnelle (Corbin, 1987; Corbin, 1991) sont appliquées dans le projet `MorTAL` (Dal *et al.*, 1999; Hathout *et al.*, 2001). Les formes générées automatiquement sont validées avec une liste de référence compilée à partir du TLFi (Trésor de la Langue Française) et sur l'Internet (Namer, 2002).

Dans le cadre du projet `Verbaction`, (Berche *et al.*, 1997) utilisent une liste de verbes également extraits du TLF. À l'aide de l'outil `Unix findaffix` ils appliquent les règles de formation des noms déverbaux d'action et gardent ceux qui existent effectivement dans le TLF. Une validation manuelle complète ce filtrage. Une extension du projet `Verbaction` est proposée dans (Tanguy & Hathout, 2002). La recherche de nouveaux lexèmes affixés est faite sur l'Internet, de même que leur filtrage : les lexèmes doivent se trouver dans un même document pour entretenir une relation sémantique.

(Viegas *et al.*, 1996) valident les lexèmes générés avec un dictionnaire et un corpus.

4.2.8 Bilan

L'acquisition de ressources morphologiques est basée essentiellement sur la recherche d'une chaîne initiale commune lors de la comparaison de deux mots. Cette chaîne commune, ou préfixe, garantit la parenté formelle des mots comparés. Pour s'assurer de leur parenté sémantique, les chercheurs contraignent le contexte sémantique d'apparition de ces

mots. L'exploitation de la structure des dictionnaires et de différentes informations qu'ils contiennent, des corpus, des règles morphologiques acquises sur les données ou établies par les morphologues, des calculs statistiques, etc. servent ainsi pour établir la proximité sémantique entre les mots qui sont proches formellement. Dans la suite de ce chapitre, nous décrivons nos expériences en acquisition de ressources morphologiques. Nous trouvons le contexte, qui garantit la parenté sémantique des mots, grâce à l'exploitation de ressources terminologiques structurées. Ainsi, lorsque les termes se trouvent en relation sémantique dans une terminologie, nous supposons que les mots de ces termes, qui sont proches formellement, sont également proches sémantiquement. La proximité sémantique des termes assure ainsi la proximité sémantique des mots qui les composent.

Les différentes expériences en acquisition de ressources morphologiques que nous présentons maintenant ont été effectuées en étroite collaboration avec Pierre Zweigenbaum. Nous présentons chacune de ces expériences : acquisition à partir de séries de synonymes (sec. 4.3), de données étiquetées et lemmatisées (sec. 4.4), à partir d'autres relations sémantiques entre termes (sec. 4.5) et avec une normalisation des segmentations (sec. 4.6).

4.3 Apprentissage à partir de séries de synonymes

Comme le montre le tab. 4.1, les termes synonymes peuvent comporter des mots morphologiquement proches, par exemple {pneumo*nie*, pneumo*pathie*} du concept D2-50140 de la SNOMED. Rappelons que le *Code* correspond à l'identifiant de concepts de la SNOMED et regroupe les termes synonymes. La *Classe* 01 indique le terme préféré, 02 son synonyme nominal (nom ou groupe nominal), 05 son synonyme adjectival. *SAI*, qui apparaît dans les libellés des termes, veut dire « *Sans autres indications* ».

Code	Classe	Termes
D2-50140	01	pneumonie, SAI
D2-50140	02	pneumopathie inflammatoire, SAI
F-00470	01	symbiose
F-00470	02	commensalisme
F-00470	05	symbiotique
F-00470	05	commensal
T-51110	01	palais dur, SAI
T-51110	02	voûte palatine

TAB. 4.1 – Les termes préférés et les termes synonymes dans la SNOMED.

4.3.1 Méthode

La méthode d'acquisition de connaissances morphologiques que nous proposons effectue d'abord un amorçage à travers une acquisition de règles morphologiques sur les séries de

synonymes. Ces règles sont ensuite généralisées sur une liste de mots de référence. Pour que deux mots soient reliés morphologiquement, il faut : (1) qu'ils partagent une chaîne de caractères initiale commune, ou préfixe (au moins trois ou quatre caractères selon les expériences) ; (2) que l'alternance de chaînes finales, qui forment une règle morphologique, se vérifie sur d'autres paires de mots comparés ; (3) et que les mots en question ne soient pas identiques. Nous détaillons maintenant les étapes principales de la méthode.

Constitution des séries de synonymes. Tout d'abord, en nous basant sur les codes alphanumériques des termes de la SNOMED, nous constituons les séries de synonymes. Si une série comporte au moins deux termes elle est utilisée dans les étapes qui suivent. À partir des données du tab. 4.1 nous obtenons les séries de synonymes suivantes :

D2-50140 *pneumonie, SAI* ; *pneumopathie inflammatoire, SAI*
F-00470 *symbiose* ; *commensalisme* ; *symbiotique* ; *commensal*
T-51110 *palais dur, SAI* ; *voûte palatine*

Alignement de paires de mots reliés morphologiquement et leur segmentation. À cette étape, pour chaque paire de termes, nous considérons chaque paire de mots $\{PS_1, PS_2\}$ qui partagent un préfixe P, ou une chaîne initiale commune, 'suffisamment long'. Nous travaillons avec une chaîne initiale de trois et de quatre caractères. Ce seuil reste un paramètre de la méthode. Les paires de mots suivantes sont ainsi alignées et segmentées :

D2-50140 {*pneumonie*, *pneumopathie*} $\Longrightarrow$ {pneumo*nie*, pneumo*pathie*}
F-00470 {*symbiose*, *symbiotique*} $\Longrightarrow$ {symbio*se*, symbio*tique*}
F-00470 {*commensal*, *commensalisme*} $\Longrightarrow$ {commensal, commensal*isme*}
T-51110 {*palais*, *palatine*} $\Longrightarrow$ {pala*is*, pala*tine*}

Induction de préfixes, de suffixes et de règles morphologiques. À partir des paires de mots alignés et segmentés, nous induisons des 'règles morphologiques' qui correspondent aux alternances de suffixes, ou de chaînes finales, des mots segmentés. Chaque paire $\{PS_1, PS_2\}$, où P est la chaîne initiale commune maximale aux deux mots, est généralisée en règle $S_1|S_2$. Celle-ci peut être interprétée de la manière suivante : étant donné un mot qui finit avec une chaîne de caractères S_1, nous pouvons former un mot où S_1 est remplacé par S_2. Les règles sont symétriques : S_2 peut également être remplacé par S_1. Pour qu'une règle soit retenue, il faut qu'elle se vérifie sur au moins deux paires de mots et que le préfixe soit suffisamment long. Sur chaque paire de mots non identiques, nous induisons ainsi une règle $S_1|S_2$, mais également les préfixes communs P et les suffixes $\{S_1, S_2\}$ (ϵ marque une chaîne vide) :

Code	Préfixe	Suffixes	Règle
D2-50140	*pneumo-*	*-nie*, *-pathie*	*nie*|*pathie*
F-00470	*symbio-*	*-se*, *-tique*	*se*|*tique*
F-00470	*commensal-*	*-*ϵ, *-isme*	ϵ|*isme*
T-51110	*pala-*	*-is*, *-tine*	*is*|*tine*

Généralisation des règles et induction de nouvelles paires de mots. Nous considérons que les règles induites peuvent être généralisées à d'autres mots de la liste de référence. Ceci nous permet d'identifier les paires de mots qui se trouvent en même relation morphologique et d'étendre ainsi les connaissances morphologiques d'amorçage. La règle *se*|*tique*, induite sur la paire {symbio*se*, symbio*tique*} permet ainsi de mettre en relation d'autres mots et de générer les nouvelles paires :

{acantholy*se*, acantholy*tique*}, {artériosclér*ose*, artériosclér*otique*}, {cholesta*se*, cholesta*tique*}, {hémoly*se*, hémoly*tique*}, {hémosta*se*, hémosta*tique*}, {hyposta*se*, hyposta*tique*}, {mélan*ose*, mélan*otique*}, {nécrobi*ose*, nécrobi*otique*}, {néphr*ose*, néphr*otique*}, {psych*ose*, psych*otique*}, {scléro*se*, scléro*tique*}, {spongio*se*, spongio*tique*}, {sténo*se*, sténo*tique*}, {synap*se*, synap*tique*}, {thrombo*se*, thrombo*tique*}.

Constitution de familles morphologiques. Les mots mis en relation morphologique dans une paire sont fusionnés en familles morphologiques lorsque :

- ils ont été alignés pendant l'étape d'amorçage, ils ont alors un même préfixe P ;
- paires de mots qui ont des mots communs.

Toujours par rapport aux termes du tab. 4.1, nous obtenons à la généralisation les familles morphologiques suivantes :

- *pneumocoque*, *pneumocoques*, *pneumoniae*, *pneumonie*, *pneumopathie*, *pneumopathies*
- *symbiose*, *symbiotique*
- *commensal*, *commensalisme*
- *palais*, *palatine*, *palatines*

4.3.2 Évaluation des résultats : précision et rappel

Pour estimer la qualité des résultats obtenus, nous cherchons à les évaluer en termes de précision et de rappel. La précision des paires de mots et des familles morphologiques est évaluée manuellement, avec la consultation des dictionnaires, des terminologies et des spécialistes du domaine. Elle correspond à la proportion des données correctes par rapport à la totalité des données induites.

L'évaluation du rappel demande par contre de disposer d'une ressource de référence. Comme nous l'avons montré, de telles ressources ne sont actuellement disponibles que pour l'anglais grâce à Specialist Lexicon de l'UMLS. En utilisant le programme `lvg`, nous pouvons générer les paires de mots flexionnelles (`lvg -m -fi`) et dérivationnelles (`lvg -m -fRf`). Ces paires de mots nous servent de référence pour l'évaluation du rappel obtenu sur des sources en anglais. Le rappel correspond à la proportion des données induites avec nos méthodes par rapport aux données proposées par le Specialist Lexicon.

4.3.3 Matériel

Le matériel nécessaire pour l'application de notre méthode consiste en une terminologie structurée avec des relations de synonymie et en une liste de mots de référence. Nous avons

travaillé sur trois langues : français, anglais et russe.

Terminologie structurée avec des relations de synonymie. Nous utilisons les versions suivantes de la SNOMED qui comporte des relations de synonymie :

- En français nous utilisons la version précommerciale du Répertoire d'anatomopathologie (Côté, 1996), qui nous a été gracieusement donnée par le Dr. Roger Côté. Le Répertoire comporte 12 555 termes.
- En anglais nous utilisons toute la SNOMED Internationale version 3.5, incluse dans le Metathesaurus de l'UMLS de 1999. Les termes de la SNOMED sont reconnaissables à leur identifiant de source (SAB=SNMI98 : Source Abbreviation = Snomed International, 1998). Nous obtenons 128 855 termes.
- En russe nous utilisons une version du Répertoire en cours de traduction (Emeline *et al.*, 1995), qui nous a été gracieusement donnée par Yvan Emelin. Le Répertoire en russe comportait à l'époque des expériences (1999-2000) 13 462 termes.

Liste de mots de référence. La liste de mots de référence est établie à partir des termes de la SNOMED et de la CIM :

- En français nous utilisons la version française du Répertoire d'anatomopathologie de la SNOMED et les termes principaux de la version française de la CIM-10 (OMS, 1995). Nous obtenons un total de 8 874 formes de mots.
- En anglais nous utilisons la SNOMED Internationale version 3.5 dans son intégralité et les termes de la CIM-9-CM. Les deux terminologies se trouvent dans le Metathesaurus de l'UMLS (pour la CIM-9-CM, SAB=ICD-9-CM). Nous obtenons une liste de 49 627 formes de mots.
- Pour le russe nous utilisons le Répertoire avec un total de 9 871 formes de mots.

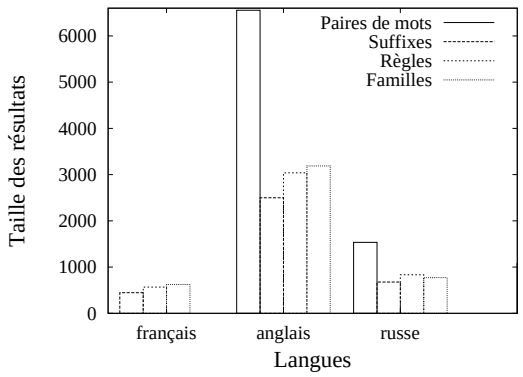
Lors de la segmentation des termes en mots, les espaces et la ponctuation, dont le tiret, sont considérés comme séparateurs. Les formes avec des nombres ne sont pas prises en compte. Tous les mots sont convertis en minuscules.

4.3.4 Analyse et discussion des ressources morphologiques induites avec les séries de synonymes en trois langues

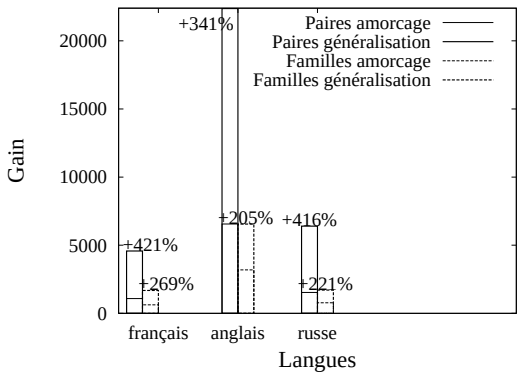
La méthode proposée est appliquée aux trois langues : français, anglais et russe (Grabar, 1999b; Grabar & Zweigenbaum, 1999a; Grabar, 1999a; Grabar & Zweigenbaum, 1999b; Grabar & Zweigenbaum, 2000a). Les données de départ ne sont pas de la même taille pour chacune des langues :

- les données en anglais sont les plus complètes (environ 10 fois plus volumineuses que dans les deux autres langues) : 128 855 termes qui donnent 26 295 séries de synonymes ;
- vient ensuite le russe avec 13 462 termes et 2 636 séries de synonymes ;
- et le français avec 12 555 termes et 2 344 séries de synonymes.

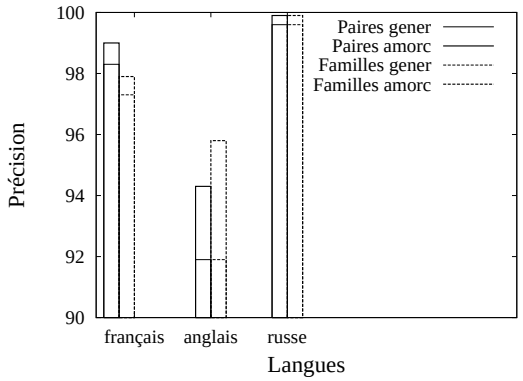
Les résultats induits sont globalement proportionnels à ces volumes de départ.



(a) Étape d'amorçage



(b) Étape de généralisation



(c) Précision

FIG. 4.1 – Acquisition de connaissances morphologiques en trois langues : français, anglais, russe. λ=4.

Étape d'amorçage. Le graphique 4.1(a) montre les résultats obtenus à l'étape de l'amorçage dans les trois langues, avec une longueur du préfixe λ fixée à quatre caractères. La figure indique les nombres de paires de mots, de suffixes, de règles et de familles morphologiques. Les résultats sont proportionnellement comparables dans les trois langues, sauf pour les familles morphologiques qui sont proportionnellement moins nombreuses en russe : il s'agit sans doute de formes flexionnelles, très présentes dans cette langue, et qui, une fois regroupées, forment un nombre de familles moins important que dans les deux autres langues. Par ailleurs, nous pouvons calculer le rendement des paires de mots quant à la production de suffixes et de règles. Nous avons 6 556 paires de mots en anglais, 1 086 en français et 1 535 en russe. Le rendement moyen des paires de mots est le plus élevé en russe, où il faut 1,8 paires de mots pour trouver une règle et 2,2 paires de mots pour

un suffixe. Il est le moins élevé en anglais, avec 2,1 paires de mots pour une règle et 2,6 pour un suffixe. En français, ce rendement est moyen avec 1,9 paires pour une règle et 2,4 pour un suffixe. Ces calculs reflètent, à notre avis, la richesse morphologique des langues étudiées, qui est élevée en russe et faible en anglais. Le français occupe, de ce point de vue, une place intermédiaire.

Les données d'amorçage en français et en russe constituent environ 10 % des données en anglais, où les ressources de la SNOMED sont les plus complètes. Ce qui explique la taille des données générées. Par contre, si on nivelle les données en prélevant un échantillon de 10 % sur les séries de synonymes en anglais, la comparaison des résultats offre un tableau plus réaliste des systèmes morphologiques de ces trois langues. Avec ce matériel d'amorçage, le nombre de règles en anglais devient 445, donc moins important qu'en français (567) ou en russe (834). Ce qui confirme la gradation de la richesse morphologique entre ces trois langues.

Étape de généralisation. À l'étape de généralisation, les règles morphologiques sont appliquées aux mots de la liste de référence, ce qui permet d'induire de nouvelles paires de mots reliés morphologiquement. Ces paires viennent enrichir les familles morphologiques existantes et en former de nouvelles. Le graphique 4.1(b) présente le gain en paires de mots et en familles morphologiques effectué lors de cette étape. La partie basse de chaque colonne présente la taille des données à l'amorçage et la partie haute à la généralisation. Nous pouvons ainsi voir que le nombre des paires à la généralisation augmente dans toutes les langues : un gain de 421 % de paires en français, de 416 % en russe et de 341 % en anglais. Il en va de même pour les familles morphologiques avec un gain de 269 % en français, de 221 % en russe et de 205 % en anglais. La taille moyenne des familles, non représentée sur les graphiques, augmente également. Elle passe de 2,9 à 3,48 mots par famille en anglais, de 2,74 à 3,38 en russe et de 2,53 à 3,08 en français. La richesse et la complétude du vocabulaire en anglais, par rapport aux deux autres langues, permet d'obtenir des familles plus peuplées. Le russe va au-devant du français avec son système flexionnel plus développé : en plus des catégories flexionnelles existant en français, le russe présente le cas qui s'applique aux noms, adjectifs, pronoms et participes.

Précision. L'évaluation de la précision des paires de mots et des familles est effectuée manuellement. En anglais, en raison de la taille des résultats, la précision est évaluée sur un échantillon (l'intervalle de confiance est de 2,2 et 1,3 pour les paires de mots et de 1,6 et 1,5 pour les familles morphologiques, lors de l'amorçage et de la généralisation respectivement). La fig. 4.1(c) montre que la précision est excellente dans les trois langues. Même si elle est un peu moins bonne pour l'anglais, elle ne descend pas au-dessous de 90 %. La généralisation (barre fine) la détériore légèrement par rapport à l'étape d'amorçage (barre épaisse). Les règles morphologiques induites dans un contexte sémantique contraint gardent donc leur exactitude lorsqu'elles sont appliquées à une liste plus grande de mots du domaine. Le fait que cette liste comporte les mots des terminologies, donc contrôlés et plutôt centraux dans le domaine médical, doit contribuer à l'obtention d'aussi bons résultats.

Rappel. Le rappel est évalué uniquement pour les données en anglais, la seule des langues étudiées qui dispose de ressources morphologiques du domaine médical que nous pouvons prendre comme référence. Le programme `lvg` du Specialist Lexicon de l'UMLS nous permet donc de générer les paires de mots en relation flexionnelle (`lvg -m -fi`) et dérivationnelle (`lvg -m -fRf`). Nous comparons nos résultats à ces paires de mots et évaluons ainsi leur rappel. Le tab. 4.2 présente les résultats de cette évaluation. La flexion de l'UMLS est couverte à 91,2 % dans nos résultats, l'affixation ne l'est qu'à 79,2 %. Les silences sont dus essentiellement au matériel linguistique de départ : une terminologie comme la SNOMED, avec une bonne couverture du domaine médical (Elkin *et al.*, 2001), reste néanmoins loin de recenser tous les termes et toutes leurs variantes (Hersh *et al.*, 1997). Mais en même temps, notre méthode permet d'obtenir un volume de paires de mots en anglais bien plus important que ce qui est proposé par `lvg`. Plusieurs raisons peuvent être mentionnées pour expliquer cette différence : environ 10 % des paires des mots générées avec notre méthode sont erronées ; `lvg` ne génère pas les mots composés, alors que notre méthode le fait ; les règles induites avec notre méthode peuvent combiner un ou plusieurs 'suffixes élémentaires' (par exemple, $\epsilon|es$, $\epsilon|le$, $aire|es$).

Éléments de connaissances	Méthode	`lvg`	Rappel
Paires de mots	25 740		
Flexion		2 697	91,2 %
Affixation		2 973	79,2 %

TAB. 4.2 – Le rappel pour la flexion et l'affixation en anglais. λ=4 avec notre méthode.

Un échantillon des résultats obtenus à la généralisation pour chaque langue traitée (règles, suffixes, préfixes et familles morphologiques) se trouvent dans l'annexe A.1 page 191. Ici, nous discutons ces résultats et en présentons quelques exemples. Les fréquences, qui servent de base au classement, sont calculées à partir du nombre de paires de mots qui fournissent une règle, un suffixe ou un préfixe. Les familles sont classées dans l'ordre alphabétique.

Règles. Parmi les règles en français, celles de la variation flexionnelle (nombre et genre) viennent en premier : $\epsilon|s$, $\epsilon|e$, $\epsilon|es$. La première dérivation, avec la règle $e|ique$, apparaît au 4e rang, elle est suivie par $aire|e$ et $ation|é$. La première composition[2] apparaît au 14e rang avec la règle $e|ose$. Elle est immédiatement suivie par $ite|o$ et $e|ome$. Notons que certaines règles combinent plus d'une opération morphologique. Par exemple, la variation simultanée du genre et du nombre ($\epsilon|es$, $\epsilon|le$) ou le passage d'une forme au pluriel à un adjectif en *-aire* (règle $aire|es$).

En anglais la situation est similaire : les règles flexionnelles arrivent en tête avec $\epsilon|s$. La règle $a|us$ (2e rang) reflète la particularité de la terminologie anglaise qui fait un emploi fréquent des mots latins et préserve leurs flexions d'origine. Les flexions des formes verbales

[2]Sur notre acception de la composition, voir la discussion de la sec. 3.1.5 page 51.

arrivent au 8e($\epsilon|ing$) et 10e($\epsilon|ed$) rangs. La première dérivation ($\epsilon|al$) se trouve au 11e rang ; et la première composition ($a|osis$) est au 12e rang.

La particularité des règles en russe est qu'elles couvrent surtout la flexion : genre (*ая*|*ый*, *ая*|*ое*), nombre (*ое*|*ый*, *ая*|*ые*), de très nombreux cas (*ая*|*ого*, *ого*|*ый*, *а*|*ы*, *й*|*х*) parmi les règles les plus fréquentes. Comme nous l'avons noté, le système flexionnel du russe est plus riche que celui du français ou de l'anglais. Les règles dérivationnelles apparaissent à partir du 22e rang ($\epsilon|$*ный*, *ие*|*ный*, $\epsilon|$*ная*, *а*|*ная*) ; elles impliquent les suffixes de formation d'adjectifs. Les règles de composition pointent à partir du 56e rang (*ма*|*саркома*).

Dans les trois langues étudiées la flexion vient donc en premier. L'anglais a la particularité d'utiliser des mots latins. L'abondance des règles flexionnelles en russe laisse entrevoir la richesse de son système flexionnel qui, en plus des catégories existant en français, présente le cas. L'affixation apparaît en deuxième, après la flexion. La composition arrive en dernier. Certaines règles combinent plus d'une opération morphologique : variation simultanée du genre et du nombre ($\epsilon|es$, $\epsilon|le$), passage d'une forme au pluriel à un adjectif (règle $aire|es$).

Suffixes. Les remarques sur les suffixes sont similaires à celles sur les règles. En français, les marques flexionnelles viennent en premier (*-s*, *-e*, *-es*). Les suffixes dérivationnels commencent avec *-é* au 5e rang et *-aire* au 6e. Tout de suite après, au 7e rang, vient le premier composant *-me* (forme tronquée de *-ome*). Ce composant réapparaît en entier au 10e rang. Quant au suffixe *-se* (8e rang), il est ambigu et peut être obtenu suite à différentes segmentations :

Mot 1	Mot 2	Préfixe	Suffixes	Règle
muqueu*x*	muqueu*se*	*muqueu-*	*-x*, *-se*	$x\|se$
fibro*se*	fibro*kystique*	*fibro-*	*-se*, *-kystique*	$se\|kystique$
psycho*se*	psycho*tique*	*psycho-*	*-se*, *-tique*	$se\|tique$
hémosta*se*	hémosta*tique*	*hémosta-*	*-se*, *-tique*	$se\|tique$

Nos techniques de recherche de la chaîne initiale commune la plus longue et la nature des mots alignés et comparés font qu'à la segmentation nous obtenons un suffixe *-se*, qui peut correspondre en réalité à l'affixe flexionnel *-euse* et aux composants *-ose* et *-stase*.

Notons que dans d'autres cas, toujours en fonction de notre méthode et des mots alignés, des mots tronqués de différentes manières peuvent donner lieu à des suffixes 'allomorphes' :

Mot 1	Mot 2	Préfixe	Suffixes	Règle
oblitéra*tion*	oblitéra*nte*	*oblitéra-*	*-tion*, *-nte*	$tion\|nte$
oblitér*ation*	oblitér*é*	*oblitér-*	*-ation*, *-é*	$ation\|é$
obstructi*on*	obstructi*f*	*obstructi-*	*-on*, *-f*	$on\|f$
obstru*ction*	obstru*é*	*obstru-*	*-ction*, *-é*	$ction\|é$
progress*ion*	progress*ant*	*progress-*	*-ion*, *-ant*	$ion\|ant$
progressi*on*	progressi*f*	*progressi-*	*-on*, *-f*	$on\|f$

Ici un même mot peut fournir des suffixes allomorphes : *-tion* et *-ation* pour le premier exemple, *-on* et *-ction* pour le deuxième, et *-ion* et *-on* pour le troisième. D'autres suffixes sont dans ce cas :

-ose (rang 11) *vs* *-atose* (rang 35),
-ation (rang 19) *vs* *-ion* (rang 38) *vs* *-tion* (rang 40),

-ique (rang 9) *vs* *-que* (rang 14) *vs* *-tique* (rang 37).

En anglais aussi, les suffixes flexionnels, dont les suffixes latins, viennent en premier. L'affixation (*-ic*) est en 7e position, et la composition (*-osis*) en 8e.

Comme pour les règles, les suffixes en russe concernent essentiellement la flexion. Le premier suffixe dérivationnel apparaît au 14e rang (*-ный*), le second au 16e rang (*-ная*). Le premier composant *-ма* (forme tronquée de *-ома* (*-ome*)) apparaît au 20e rang.

Parmi les suffixes induits, ceux de la flexion sont les plus fréquents. Comme pour les règles, l'affixation vient en deuxième, et les composants en troisième. Nous avons par ailleurs signalé les cas d'ambiguïté et d'allomorphie des suffixes repérés, ceci dans les trois langues étudiées. Ils sont dus à notre méthode de segmentation des mots en préfixes et en suffixes.

Préfixes. La méthode de segmentation a le même effet sur les 'préfixes' : ils peuvent également être segmentés de différentes manières selon les paires de mots d'où ils sont induits et former ainsi des allomorphies dues à la segmentation :

Mot 1	Mot 2	Préfixe	Suffixes	Règle
fibro*se*	fibro*kystique*	*fibro-*	*-se*, *-kystique*	*se*|*kystique*
fibro*blastique*	fibro*sarcome*	*fibro-*	*-blastique*, *-sarcome*	*blastique*|*sarcome*
fibr*ose*	fibr*euse*	*fibr-*	*-ose*, *-euse*	*ose*|*euse*
fibr*illaire*	fibr*eux*	*fibr-*	*-illaire*, *-eux*	*illaire*|*eux*

À chaque alignement la segmentation est donc apprise localement.

En anglais, nous trouvons rapidement des préfixes comme *hyper-* (rang 1), *hypo-* (rang 18), *anti-* (rang 36) ou *trans-* (rang 45) qui n'appartiennent pas spécifiquement au domaine médical, mais y sont très utilisés. Mais la plupart des préfixes, surtout parmi les plus fréquents, sont bien caractéristiques du domaine. Nous avons également quelques allomorphes dus à la segmentation : *neuro-* et *neur-*, *phosph-* et *phospho-* ; et des variantes orthographiques : *oesophag-* et *esophag-*.

Si les préfixes en russe sont majoritairement spécifiques du domaine, dès les premiers rangs nous trouvons quelques préfixes neutres : *тяжел-*, rang 22 (*difficile*, *lourd*) et *открыт-* rang 28 (*ouvert*) qui sont très usités dans le domaine médical. À part ce mélange de langages spécialisé et non spécialisé, le russe présente également un mélange de racines slaves (*почечн-*, *легочн-*, *костн-*) et de racines grecques et latines (*фибро-* = *fibro-*). Parfois une même notion existe dans les deux versions : *костн-* et *остео-* (*osteo-*), *нерв-* et *нейро-* (*neuro-*). La segmentation des mots en russe fournit également des allomorphes (*фибро-*, *фиброз-* et *фиброзн-*).

Les préfixes présentent un mélange qu'on peut situer sur l'axe de leur spécialisation. Bien que les préfixes du domaine soient dominants, ceux de la langue générale apparaissent également. Ce qui montre qu'il n'existe pas de frontière étanche entre les langages spécialisés et non spécialisés mais que les lexèmes peuvent acquérir une signification spécifique dans un registre et un contexte donnés. Le russe en plus présente une terminologie 'double' qui consiste à employer des racines slaves et des racines d'origine grecque ou latine. Parfois une même notion peut exister en deux versions. Par ailleurs, comme nous l'avons remarqué

pour les suffixes, les préfixes induits peuvent se présenter sous formes allomorphiques suite aux différents alignements et segmentations.

Familles morphologiques. Parallèlement aux remarques faites au sujet des préfixes et suffixes, les familles morphologiques présentent beaucoup de flexions, surtout en russe et en français. La flexion anglaise est complétée par la flexion latine. En fonction des familles, les dérivations et compositions peuvent être fréquentes. En anglais et en français, nous remarquons aussi des familles constituées de variantes orthographiques. Dans les trois langues étudiées les familles sont formées autour des mots spécifiques du domaine médical mais des mots 'neutres' ou provenant d'autres domaines sont aussi fréquents.

Analyse des erreurs. Nous avons signalé plus haut les cas d''allomorphie' dus à la segmentation : la segmentation d'un mot en préfixe et en suffixe peut varier selon les alignements où il apparaît. Mais les allomorphies de ce type peuvent être traitées avec des méthodes de recherche de similarité entre deux chaînes de caractères, par exemple (Jacquemin, 1997a), et si besoin regroupées.

La principale source d'erreurs concerne les cas où les chaînes initiales communes n'ont pas de parenté sémantique, ceci malgré les contraintes concernant le contexte sémantique et la longueur minimale du préfixe. Les préfixes peuvent correspondre alors à des bases dérivationnelles différentes :

{*trait*, *traitement*}, {*version*, *versant*}, {*classique*, *classe*}, {*chronique*, *chrome*},
{*clinoïde*, *clinique*}, {*glande*, *gland*}, {*malaise*, *malais*}, {*porteur*, *porte*}.

Le couple erroné {*chronique*, *chrome*} par exemple est généré avec la règle *me*|*nique* induite sur un couple correct {*polyembryome*, *polyembryonique*}.

Dans d'autres couples la chaîne initiale commune peut correspondre à un préfixe dérivationnel. La base, qui pourrait assurer la parenté sémantique des mots du couple, se trouve alors après le préfixe :

{*microgliome*, *microcytaire*}, {*hyperplasie*, *hyperkératose*}, {*hétérogreffe*, *hétérologue*}.

Certaines erreurs proviennent directement du matériel d'amorçage, comme le couple {*antérieur*, *antéflexion*} :

M-31030 : *déplacement antérieur*, *antéflexion* $\Longrightarrow$ {ant*érieur*, *antéflexion*},

où le préfixe *anté-* correspond à un segment du lexème dans ant*érieur* et au préfixe dans *anté*flexion. Notons néanmoins que la règle correspondante (*rieur*|*flexion*) ne se vérifie que sur ce couple de mots.

La longueur de la chaîne initiale commune peut également jouer un rôle dans la génération et le filtrage des erreurs. Ainsi, lorsque le préfixe est de trois caractères, la précision diminue. Et l'augmentation de la longueur du préfixe à quatre caractères permet d'éviter les erreurs comme :

{*mal*, *malaire*}, {*dur*, *dural*}, {*scie*, *sciatique*}, {*vis*, *vision*}, {*sel*, *sellaire*} ;

mais elle passe sous silence les paires de mots correctes où le préfixe commun ne dépasse pas trois caractères ou bien qui présentent des allomorphies avant le 4^{e} caractère :

{*oedème*, *oedémateux*}, {*mycétome*, *mycosique*}, {*osseuse*, *ossification*},

Les erreurs qui apparaissent dans les couples de mots sont ensuite propagées sur les familles morphologiques. Par exemple la famille formée autour du préfixe *hyper-* :

hypercellulaire, *hypercellularité*, *hyperchylomicronémie*, *hyperlipoprotéinémie*, *hyperostose*, *hyperplasie*, *hyperplasique*, *hypertrophie*, *hypertrophique*, *hypertrophié*.

Acquisition sur des mots inversés. La même méthode appliquée aux mots inversés, lus de droite à gauche, par exemple *erialucsumartni erialucsum* au lieu de *intramusculaire* et *musculaire*, permet de les regrouper par rapport à leurs chaînes finales. Dans ce cas, ce sont les préfixes qui composent les règles. Par exemple, l'ajout de préfixes :

$\epsilon|intra$, $\epsilon|sub$, $\epsilon|hémi$, $\epsilon|épi$, $\epsilon|péri$, $\epsilon|multi$, $\epsilon|mal$;

ou leur substitution :

- *inflamma*|*réac* et *dépigmenta*|*malnutri* avec la finale *-tion* ;
- *bili*|*vésicul*, *lymph*|*réticul*, *réticulolymph*|*réticul* et *intermédi*|*nodul* avec la finale *-aire*.

La règle *esc*|*esch* indique par contre une variante orthographique dans {*escarre*, *escharre*}.

Cette application de la méthode est intéressante pour la détection des préfixes, par exemple à travers les règles comme : $\epsilon|intra$, $\epsilon|sub$, $\epsilon|hémi$, $\epsilon|épi$, etc. Mais, comme nous l'avons noté, la proximité sémantique des paires de mots associés avec de telles règles n'est pas assurée : l'apport sémantique de ces préfixes est souvent trop important pour sauvegarder la continuïté morphologique de la paire de mots. Les résultats générés avec cette application ne semblent donc pas être intéressants pour le traitement de la variation terminologique. Par contre, ils offrent une vue complémentaire dans la segmentation des mots et sont potentiellement intéressants pour le recensement de préfixes d'une langue.

4.3.5 Conclusion et perspectives

Nous avons présenté les résultats d'acquisition de ressources morphologiques obtenus à partir de séries de synonymes extraites d'une terminologie structurée. Ce contexte correspond à un 'reverse engineering', dans le sens où nous ne travaillons pas sur des données brutes mais exploitons des connaissances encodées auparavant. Notre méthode est donc dépendante de l'existence de telles ressources. Pour la rendre plus générale, nous allons l'appliquer à d'autres relations sémantiques pouvant exister dans une terminologie (sec. 4.5).

Ce contexte sémantique très contraint permet d'obtenir une excellente précision pour les paires de mots et pour les familles morphologiques en français et russe (entre 97,3 et 99,9%). En anglais la précision se trouve entre 91,9±1,5% et 95,8±1,6%. La méthode, appliquée sur des langues de trois types différents (familles romane, germanique et slave), donne ainsi des résultats très performants. Le rappel, évalué uniquement pour les données en anglais, est de 91,2% pour la flexion et de 79,2% pour l'affixation.

Avec notre méthode nous obtenons des paires de mots qui, suite à une segmentation permettent d'induire des 'préfixes', des 'suffixes' et des 'règles morphologiques'. Par ailleurs les paires de mots sont regroupées en familles morphologiques.

Les erreurs identifiées lors de la révision humaine des paires de mots et des familles sont enregistrées. Elles peuvent constituer une liste d'exceptions pour les expériences futures (McCray *et al.*, 1994; Namer, 2000). Nous les appliquons ainsi pour le filtrage dans les expériences présentées dans le chapitre 5.

Notre méthode n'effectue pas de segmentation itérative en préfixes et suffixes. Cependant, sans parler de la flexion, beaucoup de mots médicaux ont plus de deux éléments morphologiques :

actino-myc-os-ique, *adéno-myom-ateux*, *an-algési-que*, *apo-névr-ot-ique*.

L'ordre des opérations, leurs dépendances et les apports sémantiques de chaque élément doivent être étudiés dans un modèle linguistique plus élaboré, par exemple dans une application de la morphologie constructionnelle (Namer, 1999; Namer, 2003a). Cette perspective va bien au-delà de nos objectifs qui consistent plutôt à appréhender et amorcer la constitution de ressources morphologiques pour une langue et un domaine qui n'en disposent pas. Elle se rapproche des travaux en analyse morphosémantique (sec. 3.2.3 page 57) et présente une perspective intéressante.

Chaque règle morphologique induite a une valeur associative : étant donné deux suffixes de cette règle, elle propose de relier deux mots qui comportent ces suffixes et un préfixe d'une longueur minimale donnée. Mais souvent elle ne représente pas une opération morphologique précise, comme la formation des adjectifs dénominaux, des noms déverbaux, etc. De ce côté aussi, une étude linguistique devra compléter et affiner les résultats.

Les suffixes les plus fréquents, comme les règles, concernent la flexion. En anglais la flexion couvre également la flexion latine. L'affixation vient ensuite et la composition suit. Nous repérons également quelques variantes orthographiques. La flexion, très présente dans les règles et les suffixes, est souvent combinée avec les procédés de formation des lexèmes : affixation et composition.

Si des données étiquetées syntaxiquement sont disponibles, les règles induites deviendront plus précises et pourront bloquer la génération de certaines paires de mots erronées. Et si en plus nous disposons de données lemmatisées, les règles et les suffixes seront 'nettoyés' des flexions. Des affixes dérivationnels et des composants pourront ainsi émerger à la surface. C'est ce que nous présentons dans la section suivante.

4.4 Apprentissage à partir de séries de synonymes étiquetées et lemmatisées

Dans la section précédente, la méthode a été appliquée à des données 'brutes', sans utilisation de connaissances linguistiques *a priori*. Dans cette section, notre intérêt porte sur l'influence des connaissances linguistiques, en particulier de l'étiquetage morphosyntaxique et de la lemmatisation (Grabar, 1999a; Zweigenbaum & Grabar, 1999). Nous nous attendons à ce que les données étiquetées syntaxiquement permettent de repérer les cas de conversion, qui ne sont pas détectés sinon. Elles rendent également les règles plus contraintes en précisant la catégorie grammaticale des suffixes obtenus. Tandis que les

données lemmatisées permettent de faire l'abstraction de la flexion et de faire ainsi émerger l'affixation et la composition.

4.4.1 Adaptation de la méthode

Dans les données étiquetées, l'étiquette morphoyntaxique assignée à chaque mot est séparée de lui par une barre oblique « / »[3] :

tissu/SBC:sg égaré/ADJ:sg pendant/PREP la/DTN:sg manipulation/SBC:sg technique/ADJ:sg

La même méthode est alors appliquée pour l'acquisition de connaissances morphologiques sauf que l'étiquette et le séparateur « / » doivent faire partie des suffixes. L'étiquette est ainsi considérée comme une prolongation du suffixe et permet de le spécifier.

4.4.2 Préparation de données étiquetées et lemmatisées

Lorsqu'on travaille sur une version étiquetée et lemmatisée des séries de synonymes, le reste du matériel (liste de mots de référence) doit également être étiqueté et lemmatisé. Nous présentons ici le travail de préparation de ces données effectué pour le français. Notons que le travail d'étiquetage a été également effectué pour le russe (Grabar, 1999a), à travers une adaptation et un entraînement de l'étiqueteur morphosyntaxique de (Brill, 1995) (désormais `Brill`).

Étiquetage des données. Nous avons étiqueté les termes du Répertoire d'anatomopathologie en français avec `Brill`. Séparément, nous avons étiqueté à la main les mots des termes de la CIM-10 qui n'apparaissent pas dans le Répertoire. `Brill` est un étiqueteur générique que l'on peut entraîner sur un corpus préétiqueté. Il est fourni par l'auteur avec les données résultant de l'entraînement sur un corpus journalistique anglais. Nous utilisons la version `Brill` INaLF, adaptée au français par l'INaLF à travers un entraînement sur des textes littéraires (Lecomte, 1998). Un jeu d'environ 50 étiquettes est alors utilisé.

Bien que la SNOMED recense les expressions 'naturelles' de la langue, ces expressions, qui se présentent souvent comme des groupes nominaux, ont une syntaxe différente de celle de phrases. Notons essentiellement :

- l'omission d'article devant un nom : *radical libre*, *bras artificiel*
- un nom initial suivi d'au moins un adjectif :
  prothèse valvulaire cardiaque artificielle, *vis métallique orthopédique*.

Pour obtenir de meilleurs résultats, nous avons adapté `Brill` INaLF aux données spécifiques que sont les termes : nous l'avons d'abord lancé sur les termes de la SNOMED ; ensuite, sur un échantillon de 10 % (1 301 termes), nous avons corrigé les erreurs à la main[4]. Lors

[3] La signification des étiquettes morphosyntaxiques est indiquée dans l'annexe A.2.1 p. 208.

[4] La rectification des erreurs a été effectuée sous `emacs`. Nous remercions Yannick Toussaint qui nous a fourni gracieusement un programme d'assignation et de correction manuelle des étiquettes morphosyntaxiques.

de la correction, nous avons ajouté une nouvelle étiquette *PFX* (préfixe). Elle s'applique aux préfixes et aux composants, qui se terminent souvent en *-o* et sont séparés par un trait d'union du reste du mot. Et comme le trait d'union est un des caractères utilisés lors de la segmentation des termes en mots, ces éléments sont automatiquement isolés. Il est donc assez logique de leur assigner une étiquette correspondante :

D0-10440 *dermatose/SBC:sg pustuleuse/ADJ:sg sous/PFX -/- cornée/SBC:sg*
D1-50460 *fibromatose/SBC:sg pseudo/PFX -/- sarcomateuse/ADJ:sg*
C-10072 *radio/PFX -/- isotope/SBC:sg ,/, SAI/ABR*
D1-21000 *ostéo/PFX -/- arthrose/SBC:sg ,/, SAI/ABR*
D3-20030 *fibro/PFX -/- élastose/NN endocardique/ADJ:sg*

L'étape suivante a consisté en un ré-entraînement de l'étiqueteur sur l'échantillon corrigé manuellement. Ce qui a permis d'une part d'enrichir le dictionnaire de mots médicaux inconnus jusqu'ici, et d'autre part d'acquérir des règles contextuelles et lexicales spécifiques aux constructions terminologiques et à la langue médicale. La nouvelle version de `Brill` a été ensuite appliquée à tous les termes du Répertoire.

Si la version `Brill` INaLF a montré une précision de 76,4 %, après le ré-entraînement nous obtenons une précision de 91,4 %. Notons néanmoins que 13 % des mots restent encore inconnus du système et ont été étiquetés à la main.

Lemmatisation des données. Nous avons appliqué `Flemm` (Namer, 2000) pour lemmatiser les termes et mots munis d'étiquettes `Brill`. `Flemm` permet également d'affiner les étiquettes en ajoutant des traits morphologiques (genre, nombre) et de les rectifier lorsqu'une distorsion entre la forme des mots et l'étiquette assignée est détectée.

4.4.3 Analyse et discussion des ressources morphologiques induites avec des données étiquetées et lemmatisées

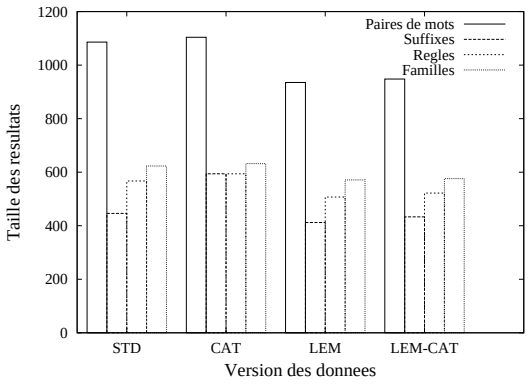
Nous avons appliqué la méthode aux données en français sur trois versions du matériel enrichi linguistiquement : étiquetage [CAT], lemmatisation [LEM], lemmatisation et étiquetage [LEM-CAT]. Nous comparons les nouveaux résultats avec la version brute des données [STD] telle que décrite dans la sec. 4.3.4.

Étape d'amorçage. La fig. 4.2(a) présente la taille des données induites à l'étape d'amorçage : paires de mots, règles, suffixes et familles morphologiques. Avec les données [CAT], le volume de toutes les données augmente par rapport à [STD]. Ceci est dû aux mots polycatégoriels et aux conversions, qui ne sont pas repérés sur les données [STD] :

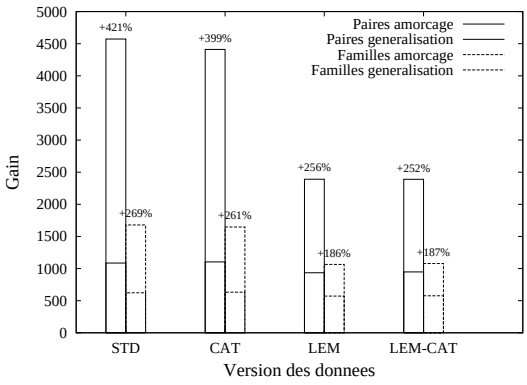
rompu/ADJ:sg - rompu/ADJ2PAR:sg
muqueuse/SBC:sg - muqueuse/ADJ:sg
résistant/ADJ:sg - résistant/VNCNT.

Le même fait influence le contenu des suffixes et des règles. Il s'agit en particulier de :

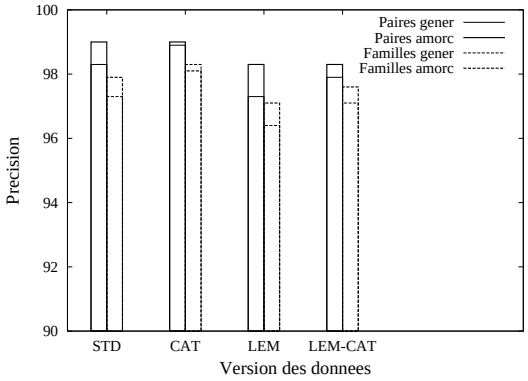
– l'apparition de règles de conversion :
ϵ*/ADJ:sg*|ϵ*/SBC:sg*, ϵ*/ADJ:sg*|ϵ*/VNCNT*, ϵ*/ADJ2PAR:sg*|ϵ*/ADJ:sg* ;



(a) Étape d'amorçage



(b) Étape de généralisation



(c) Précision

FIG. 4.2 – Acquisition de connaissances morphologiques avec des données étiquetées et lemmatisées. λ=4.

- la spécialisation de règles existantes :
  - la règle *aire*|*é* est spécialisée de deux manières :
    aire/ADJ:sg|*é/ADJ:sg* $\Longleftarrow$ {fasciculaire/ADJ:sg, fasciculé/ADJ:sg}
    aire/ADJ:sg|*é/SBC:sg* $\Longleftarrow$ {cavitaire/ADJ:sg, cavité/SBC:sg} ;
  - la règle $\epsilon|e$ ne s'applique plus qu'aux adjectifs :
    ϵ/*ADJ:sg*|*e/ADJ:sg* $\Longleftarrow$ {nasalϵ/*ADJ:sg*, nasal*e/ADJ:sg*}.
    Cette règle génère 27 paires en moins par rapport à [STD], dont les erreurs :

    gland - *glande*, *gain* - *gaine*, *vaccin* - *vaccine*, *coup* - *coupe*.

Avec les données [LEM], la taille des données générées diminue, elle est inférieur à [CAT] et [STD] : la lemmatisation permet de s'abstraire des formes fléchies en les réduisant à leurs lemmes. Le nombre de paires de mots et de règles est réduit en proportion : respecti-

vement 14 % et 10 % en moins. Les règles qui n'apparaissent plus concernent les flexions (mise au féminin $e|\epsilon$ ou au pluriel $s|\epsilon$) ou bien les règles qui combinent la flexion avec l'affixation ou la composition ($e|al$ *vs* $e|ale$). Les résultats de l'expérience [LEM-CAT] sont un croisement des deux expériences précédentes. La lemmatisation des données réduit la flexion, l'étiquetage fait ressortir les mots polycatégoriels et permet de spécifier les suffixes morphosyntaxiquement.

Étape de généralisation. La fig. 4.2(a) montre le gain réalisé à la généralisation pour les paires de mots et les familles morphologiques. L'introduction de contraintes syntaxiques fait diminuer le gain en paires de mots : il passe de 421 % avec [STD] à 399 % avec [CAT]. En effet, certaines règles étant spécifiées syntaxiquement, comme ϵ*/ADJ:sg*|*e/ADJ:sg*, ne s'appliquent plus que dans ce contexte syntaxique précis. Le nombre et la taille des familles morphologiques diminuent également. Avec les données lemmatisées [LEM], le gain diminue drastiquement à 256 % : la réduction des suffixes flexionnels qui apparaissaient séparément ou en combinaison avec l'affixation et la composition se fait sentir. Les familles morphologiques diminuent encore plus : leur taille, car presque toutes elles comportaient la flexion, et leur nombre, car certaines familles ne comportaient que des variantes flexionnelles. Le gain avec la version [LEM-CAT] est de 252 %, il est légèrement inférieur à [LEM] : le filtrage syntaxique compense la présence des conversions.

Précision. La précision des familles morphologiques est toujours un peu inférieure à celle des paires (fig. 4.2(c)). De manière générale, la généralisation des données fait diminuer la précision, qui reste pourtant très élevée. La précision des paires de mots avec [CAT] passe à 98,3 % (augmentation d'un % par rapport à [STD]) : plusieurs paires incorrectes ont pu être filtrées avec les contraintes syntaxiques. La précision des paires [LEM] est moins bonne que celle obtenue avec [STD] : avec un volume des paires beaucoup moins élevé certaines erreurs persistent et prennent un pourcentage plus important. Mais ici aussi la différence est faible, moins d'un %. La combinaison [CAT-LEM] réduit le nombre d'erreurs par rapport à [LEM] et, avec 97,1 % de précision, atteint quasiment le niveau [STD]. La précision [CAT-LEM] *vs* [LEM] a presque les mêmes rapports que [CAT] *vs* [STD].

À côté des erreurs filtrées par l'étiquetage syntaxique (*gland* - *glande*, *gain* - *gaine*, etc.) certaines paires de mots correctes ne sont plus associées. Cela vient en partie de l'existence des mots polycatégoriels : le matériel d'amorçage ni la liste de référence ne répertorient toutes leurs étiquettes possibles. Par exemple la paire {*gris*, *grise*} où *gris* est ambigu entre le singulier et le pluriel. Et comme dans le lexique de référence seule la forme plurielle est attestée *gris/ADJ:pl* :

T-A3040 *noyaux/SBC:pl gris/ADJ:pl centraux/ADJ:pl*,

la règle de mise au féminin ϵ*/ADJ:sg*|*e/ADJ:sg* ne peut donc pas être appliquée. Il en est de même pour la règle de formation des pluriels *l*|*ux* spécifiée en *l/ADJ:sg*|*ux/ADJ:pl* : elle ne peut être appliquée qu'aux adjectifs et ne couvre plus les noms comme {*canal/SBC:sg*, *canaux/SBC:pl*}. Et comme la règle similaire pour la mise au pluriel des noms *l/SBC:sg* |*ux/SBC:pl* n'a pas été induite, ces deux mots ne sont pas associés. Ou une autre règle

$\epsilon|ose$ induite sur le couple {*hyalin*, *hyalinose*}. Dans la version étiquetée, cette règle devient ϵ*/ADJ:sg*|*ose/SBC:sg* et ne peut plus associer deux noms comme {*cholestérol/SBC:sg*, *cholestérolose/SBC:sg*}.

L'étiquetage syntaxique améliore la précision, mais le rappel a tendance à diminuer face à l'incomplétude du matériel d'amorçage et aux mots polycatégoriels non répertoriés. Il faudrait assigner toutes les étiquettes possibles aux mots polycatégoriels, comme le suggère (Krovetz, 1993).

Nous analysons maintenant les résultats obtenus avec les données étiquetées et lemmatisées à l'étape de généralisation. L'échantillon de ces résultats se trouve dans l'annexe A.2 page 208. Nous discutons les règles morphologiques et les suffixes parce que c'est là qu'apparaissent les principales différences par rapport aux données [STD].

Règles. L'injection de connaissances syntaxiques a principalement deux effets : spécialisation syntaxique des règles et émergence de la conversion. Ces différences sont observables par rapport aux règles [STD]. Ainsi, la règle de mise au pluriel régulier ($\epsilon|s$), qui apparaît au 1er rang avec [STD], occupe maintenant les deux premiers rangs sous ses formes spécifiées applicables aux adjectifs (ϵ*/ADJ:sg*|*s/ADJ:pl*) et aux noms (ϵ*/SBC:sg*|*s/SBC:pl*). La règle $\epsilon|e$, 3e rang avec [STD], est spécifiée en ϵ*/ADJ:sg*|*e/ADJ:sg*, 3e rang avec [CAT], et ne s'applique plus que pour former les féminins d'adjectifs. La première affixation, formation d'adjectifs dénominaux, se trouve au 5e rang (*e/SBC:sg*|*ique/ADJ:sg*). La première composition (*ite/SBC:sg*|*o/PFX*) au 16e. Au 8e rang apparaît une règle de conversion (ϵ*/ADJ:sg*|ϵ*/SBC:sg*). Par ailleurs, certaines règles mélangent toujours la flexion avec l'affixation et la composition, mais d'une manière plus explicite : la règle *aire/ADJ:sg*|*es/SBC:pl* associe un adjectif au singulier avec un nom au pluriel.

Comme la lemmatisation réduit les formes flexionnelles, les règles de flexion n'apparaissent plus dans [LEM]. L'affixation occupe les premiers rangs (*e*|*ique*, *aire*|*e*, *e*|*que*, *ation*|*é*, *al*|*o*). La première règle de composition (*e*|*ose*) apparaît au 6e rang. Si la plupart des règles s'appliquent à des noms et des adjectifs, les règles comme *é*|*er*, 23e rang, et *é*|*ant*, 31er rang, impliquent les formes verbales.

Dans les données [LEM-CAT], le premier rang est occupé par une règle d'affixation *e/SBC*|*ique/ADJ*. Au deuxième rang se trouve une règle de conversion ϵ*/ADJ*|ϵ*/SBC*. La première composition (*e/SBC*|*ose/SBC*) apparaît au 7e rang.

Suffixes. Par rapport aux suffixes [STD] les suffixes [CAT] sont spécifiés syntaxiquement. Nous pouvons ainsi distinguer deux *-s* et deux *-e*, dont l'un s'applique aux adjectifs (*-s/ADJ:pl*, *-e/ADJ:sg*) et un autre aux noms (*-s/SBC:pl*, *-e/SBC:sg*). Par ailleurs, les suffixes pluriels et singuliers peuvent être distingués au moyen de leur trait morphologique du nombre (*-al/ADJ:sg vs -ux/ADJ:pl*). La première affixation (*-aire/ADJ:sg*) apparaît au 6e rang, le premier composant (*-me/ADJ:sg*) au 8e. Avec les données [STD], nous avons vu que le suffixe *-se* est ambigu hors contexte : il peut provenir de l'affixe flexionnel *-euse* ou des composants *-ose* et *-stase*. L'étiquetage syntaxique permet de le désambiguïser en

partie :

Mot 1	Mot 2		Suffixes
muqueu*x/ADJ:sg*	muqueu*se/ADJ:sg*	$\Longrightarrow$	*-x/ADJ:sg*, *-se/ADJ:sg*
caséeu*x/ADJ:sg*	caséeu*se/ADJ:sg*	$\Longrightarrow$	*-x/ADJ:sg*, *-se/ADJ:sg*
fibro*se/SBC:sg*	fibro*kystique/ADJ:sg*	$\Longrightarrow$	*-se/SBC:sg*, *-kystique/ADJ:sg*
psycho*se/SBC:sg*	psycho*tique/ADJ:sg*	$\Longrightarrow$	*-se/SBC:sg*, *-tique/ADJ:sg*
hémosta*se/SBC:sg*	hémosta*tique/ADJ:sg*	$\Longrightarrow$	*-se/SBC:sg*, *-tique/ADJ:sg*

L'étiquette syntaxique précise que le *-se* adjectival provient d'une flexion et *-se* nominal des composants. Par contre, la confusion persiste face aux conversions comme *muqueuse/SBC:sg*.

Avec les données [CAT-LEM], les suffixes flexionnels n'apparaissent pas et l'affixation occupe les premiers rangs. Le premier composant *-me* (*-ome* tronqué) apparaît au 6e rang.

Notons par ailleurs que le lemmatiseur, même performant pour le français général, ne dispose pas toujours de règles nécessaires pour traiter convenablement le langage médical. La particularité linguistique, que nous discutons ici, concerne l'emploi des mots spécifiques du domaine avec des flexions irrégulières et les mots latins. La difficulté apparaît surtout avec des mots latins, noms ou adjectifs, terminés par un *-s*. Lorsque ce sont des noms pluriels ou singuliers, la suppression de *-s* final n'est pas toujours pertinente :

singulier			pluriel		
alcalis/SBC:pl	$\Longrightarrow$	*alcali*	*adipositas/SBC:sg*	$\Longrightarrow$	*adiposita*
fascias/SBC:pl	$\Longrightarrow$	*fascia*	*cutis/SBC:sg*	$\Longrightarrow$	*cuti*
fèces/SBC:pl	$\Longrightarrow$	*fèce*	*hydrops/SBC:sg*	$\Longrightarrow$	*hydrop*

Mais la plupart de noms singuliers de ce type sont néanmoins correctement traités. Par contre, les cas d'erreurs de lemmatisation des adjectifs concernent toujours des formes au singulier :

abducens/ADJ:sg	$\Longrightarrow$	*abducen*
biceps/ADJ:sg	$\Longrightarrow$	*bicep*
chronicus/ADJ:sg	$\Longrightarrow$	*chronicu*
rectus/ADJ:sg	$\Longrightarrow$	*rectu*

Le lemmatiseur, comme l'étiqueteur, devrait être adapté au vocabulaire du domaine de spécialité, ici à la langue médicale.

4.4.4 Conclusion et perspectives

Avec les données étiquetées syntaxiquement, les règles induites deviennent plus spécifiques et peuvent bloquer certaines erreurs. Elles font également émerger la conversion. La précision des paires de mots à la généralisation augmente légèrement : elle est de 98,9 %. Par contre, nous constatons en même temps des silences dus aux lacunes concernant les mots polycatégoriels. Dans notre matériel, ces mots ne reçoivent pas toutes les étiquettes possibles et ne peuvent donc plus être traités par les règles.

Avec la lemmatisation la flexion est réduite et les règles et suffixes font émerger deux autres procédés de formation de lexèmes : affixation et composition. La précision des paires de mots diminue d'un % par rapport aux données [STD], elle est alors de 97,3 %. Comme

certaines erreurs peristent et le volume des données générées est moins important, ces erreurs prennent des proportions plus importantes.

Les données lemmatisées et étiquetées font abstraction de la flexion et bénéficient des contraintes syntaxiques. Nous sommes en présence de trois procédés de formation des lexèmes : conversion, affixation et composition. La précision des paires de mots est presqu'aussi élevée qu'avec les données [STD] : 97,9 %.

Malgré les corrections manuelles des données certaines erreurs d'étiquetage persistent. Une correction progressive devrait être faite et on pourra alors refaire cette expérience avec des données plus correctes.

Dans les deux expériences présentées jusqu'alors nous nous basons sur l'existence de relations de synonymie. Mais ces relations ne sont pas présentes dans toutes les terminologies structurées. Il serait donc intéressant de généraliser la méthode tout en profitant du contexte sémantiquement contraint des terminologies. C'est ce que nous faisons dans la section suivante en exploitant d'autres relations sémantiques du Répertoire d'anatomopathologie et de la CIM-10.

4.5 Apprentissage à partir d'autres relations sémantiques encodées dans les terminologies

Nous distinguons trois types de relations dans les terminologies : les relations lexicales, dont la synonymie, les relations hiérarchiques et les relations transversales (sec. 7.2 page 136). La relation hiérarchique, de nature conceptuelle et souvent la plus répandue, présente un niveau obligatoire dans la structuration des termes. Les relations lexicales font leur apparition dans les produits terminologiques. Quant aux relations transversales, elles sont très rares bien que, lorsqu'elles sont étiquetées, elles permettent d'effectuer une description plus fine de la connaissance du domaine. Afin de rendre notre méthode d'induction de ressources morphologiques plus générique, nous l'appliquons à toutes ces relations, en exploitant le Répertoire d'anatomopathologie (relations hiérarchiques et transversales) et la CIM (relations hiérarchiques) (Grabar & Zweigenbaum, 2000b). Nous supposons donc que le contexte sémantiquement contraint reste sauvegardé avec ces autres types de relations.

4.5.1 Méthode

Nous utilisons la même méthode que pour les termes synonymes (sec. 4.3.1), mais l'amorçage est fait sur des séries de termes en relation hiérarchique ou transversale.

Les séries de termes reliés hiérarchiquement sont constituées sur la base de leurs codes alphanumériques, leur niveau hiérarchique étant souvent marqué à travers ce code. Dans la SNOMED, le code à 5 chiffres représente 5 niveaux hiérarchiques. Plus le niveau est haut, plus il y a de zéros dans le code. Avec la spécialisation, les zéros sont précisés par d'autres chiffres :

D2-53000 *pneumoconiose, SAI*
D2-53400 *pneumopathie liée à l'inhalation de poussières, SAI*
D2-53900 *pneumonie par irradiation*

Dans la CIM, la spécialisation des concepts est signifiée par un principe similaire (ajout et spécialisation des chiffres du code alphanumérique) :

A04 *Autres infections intestinales bactériennes*
A040 *Infection entéropathogène à Escherichia coli*
A041 *Infection entérotoxigène à Escherichia coli*
A045 *Entérite à Campylobacter*
A047 *Entérocolite à Clostridium difficile*
A049 *Infection intestinale bactérienne, sans précision*

Pour la constitution des séries de termes reliés avec les relations transversales, nous avons utilisé l'information encodée dans la SNOMED :

D2-50140 (*pneumonie, SAI*) → T-28000 (*poumon, SAI*)
P1-57450 (*appendicectomie, SAI*) → T-59200 (*appendice vermiculaire, SAI*)

4.5.2 Matériel

La liste de mots de référence est la même que dans les expériences précédentes : 8 874 formes de la SNOMED et de la CIM10. Les séries de termes de la SNOMED et de la CIM10 sont reliées par des relations hiérarchiques et transversales.

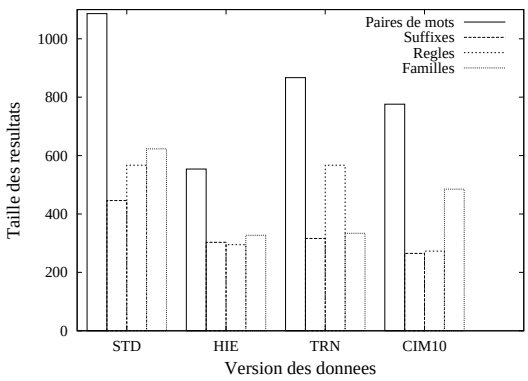
4.5.3 Analyse et discussion des ressources morphologiques induites avec diverses relations sémantiques

Nous décrivons ici les expériences effectuées avec les données en français : relations hiérarchiques [HIE] et transversales [TRN] de la SNOMED, et relations hiérarchiques de la CIM10 [CIM10] que nous comparons avec les relations synonymiques de la SNOMED [STD]. La taille des données de départ est inégale selon les terminologies et les relations. La CIM10 propose ainsi moins de relations hiérarchiques que la SNOMED (tab. 4.3).

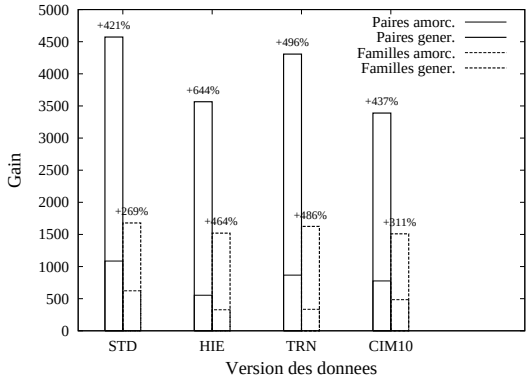
STD	HIE	TRN	CIM10
2 344	1 949	2 082	1 554

TAB. 4.3 – Nombre de séries des termes à l'amorçage : relations sémantiques de la SNOMED ([STD], [HIE], [TRN]) et relations hiérarchiques de la CIM10 [CIM10].

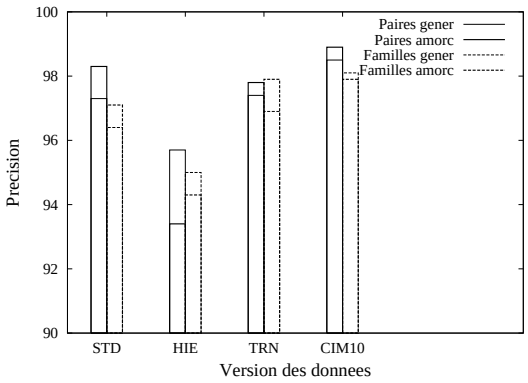
Étape d'amorçage. La comparaison des volumes de données induites à l'amorçage (fig. 4.3(a)) montre également une différence notable. Ce sont les données [STD] qui permettent d'induire les données les plus volumineuses. Les relations hiérarchiques de la CIM-10 [CIM10], comparées aux relations hiérarchiques de la SNOMED [HIE], montrent une productivité supérieure pour les paires de mots et les familles morphologiques. Mais le nombre



(a) Étape d'amorçage



(b) Étape de généralisation



(c) Précision

FIG. 4.3 – Acquisition de connaissances morphologiques avec diverses relations sémantiques encodées dans les terminologies. λ=4.

de suffixes et de règles est supérieur avec [HIE] de la SNOMED. Les relations transversales de la SNOMED [TRN], avec une différence entre le nombre de suffixes et de règles plus importante qu'ailleurs, montrent une grande capacité combinatoire de ces suffixes.

Étape de généralisation. La figure 4.3(b) montre le gain réalisé par les règles à l'étape de généralisation. La version [HIE] a le gain le plus élevé en paires de mots (644 %) et en familles morphologiques (464 %), tandis que [STD] atteint respectivement 421 % et 269 %. Mais du point de vue du volume des données [STD] en propose le plus à la généralisation. Notons que, malgré le gain réalisé par la CIM (437 % pour les paires de mots et 311 % pour les familles), les volumes des données qu'elle permet de générer restent inférieurs à ceux de la SNOMED.

Précision. La précision pour les paires de mots et les familles à la généralisation est excellente pour toutes les méthodes : entre 97 % et 98 % (fig. 4.3(c)). La précision des familles morphologiques est également excellente, à peine inférieure d'un % à la précision des paires de mots. Notons qu'avec [HIE] la précision augmente à la généralisation.

Un échantillon des résultats se trouve dans l'annexe A.3 page 218. Ici, nous montrons les différences qui existent entre les suffixes et règles induits avec les expériences [STD], [HIE], [TRN] et [CIM10].

Suffixes. Sur 891 suffixes au total, il n'y en a que 66 qui sont proposés par toutes les 4 expériences ([STD], [HIE], [TRN] et [CIM10]). Il s'agit bien sûr des flexions (*-e*, *-s*, *-x*, *-ë*), y compris les flexions latines (*-a*, *-um*, *-us*). Mais dans la majorité des cas, ce sont des suffixes dérivationnels (*-aire*, *-al*, *-ant*, *-ique*, *-ation*) et leurs variantes flexionnelles et de segmentation : *-aires* et *-ire* pour *-aire* ; *-aque*, *-cique*, *-que*, *-ques* et *-tique* pour *-ique*. Nous avons aussi quelques composants très fréquents du domaine : *-ite*, *-me*, *-opathie*, *-ose*, *-pathie*.

108 suffixes sont partagés par trois expériences et 177 par deux expériences. Plus de la moitié des suffixes (540) sont uniques à l'une des expériences : 143 avec [STD], 171 avec [HIE], 77 avec [TRN] et 149 avec [CIM10]. La principale différence de ces suffixes uniques provient des composants, qui peuvent être élémentaires (*-glie*) ou bien décomposables à leur tour (*-(o)sarcomatose*) :

STD : *-dermie*, *-blastomatose*, *-leucémie*, *-flexion*, *-glie*, *-graphie*, *-greffe*, *-logue*, *-micrographie*, *-monoblastique*, *-(o)moteur*, *-(o)phage*, *-philie*, *-(o)rraphie*, *-plastique*, *-synaptique*

HIE : *-adénocarcinome*, *-alanine*, *-angéite*, *-arthrite*, *-basal*, *-blaste*, *-calice*, *-cervical*, *-corps*, *-cytose*, *-globine*, *-lipidémie*, *-logie*, *-nucléaire*, *-activité*, *-(o)rrhexis*, *-pathologie*, *-plasmique*

TRN : *-allergique*, *-carde*, *-dynie*, *-ectomie*, *-cervicite*, *-myocardique*, *-dysplasie*, *-gliome*, *-génèse*, *-myélocèle*, *-salpinx*, *-thorax*

CIM10 : *-biotiques*, *-cérébrale*, *-fongiques*, *-musculaire*, *-mycose*, *-globuline*, *-rhumatismal*, *-sexuel*, *-stimulants*, *-toniques*, *-pulmonaires*

Comme dans les expériences précédentes, nous avons les variantes flexionnelles :

-logique (STD, HIE, TRN) *vs -logiques* (CIM10)
-rachidien (HIE) *vs -rachidiens* (CIM10)
-crânien (HIE) *vs -crânienne* (CIM10) *vs -crâniennes* (CIM10) *vs -crâniens* (CIM10)
-apical (STD, TRN) *vs -apicaux* (TRN)

et les variantes de segmentation des composants :

-chrome (STD) *vs -hrome* (HIE) *vs -ochrome* (HIE)
-nécrose (CIM10) *vs -onécrose* (TRN)
-oéthylène (STD) *vs -éthylène* (HIE)

Il est intéressant de remarquer que la plupart des suffixes uniques de la CIM10 sont au pluriel : *-atoses*, *-bles*, *-ites*, *-thies*, *-logiques*, *-rachidiens*, *-crâniennes*, *-crâniens*, etc. Il

s'agit certainement d'une particularité du métalangage de la CIM où les formes au pluriel acquièrent une signification plus générique.

Règles. Sur la totalité des 1299 règles, 47 sont proposées à la fois par les quatre expériences. Il s'agit des règles de flexion (*es*|ϵ, *e*|ϵ, *s*|ϵ, *s*|*e*), mais surtout des règles d'affixation, avec ou sans variation flexionnelle (*e*|*aire*, *e*|*al*, *ique*|*e*, *nel*|ϵ, *o*|*aire*, *o*|*al*, *que*|*e*, *s*|*aire*, *s*|*dien*, *tique*|*se*, *x*|*cique*). Et de quelques règles de composition : *opathie*|*aque*, *ose*|*euse*, *o*|*ite*, *us*|*ose*.

Deux tiers des règles (921) sont uniques à l'une des expériences : 252 règles à STD, 304 à HIE, 190 à TRN, et 175 à CIM10 :

STD : *me*|*dermie*, *logue*|*greffe*, *se*|*philie*, *plastique*|*greffe*

HIE : *blaste*|ϵ, *métriales*|*cervicales*, *peroxidase*|*cytochimique*, *nucléaire*|*corps*, *orrhexis*|*aire*

TRN : *myocardique*|*carde*, *ique*|*ectomie*, *thorax*|*rragie*, *ogliome*|*e*

CIM10 : *mycose*|*blastomycose*, *ventriculaire*|*cérébrale*, *fongiques*|*biotiques*

Les différentes terminologies et relations examinées peuvent identifier différentes relations morphologiques. Les suffixes et règles de flexion et d'affixation composent un noyau commun assez stable, même s'il en existe de nombreuses variantes de segmentation. La principale différence dans les expériences présentées provient des composants et des règles correspondantes.

Les relations hiérarchiques et transversales mettent en jeu des termes qui sont sémantiquement plus éloignés que des synonymes. Nous nous attendions à ce que ces relations fournissent le plus grand nombre de composants. Mais il apparaît que toutes les relations exploitées, y compris la synonymie, sont productives en composants et en règles de composition. Les résultats de chacune des relations exploitées sont plutôt complémentaires. Les expériences [HIE] et [TRN] permettent par exemple d'induire une règle de composition assez générale *ite*|ϵ, non fournie par [STD]. Cette règle génère des paires comme :

{*tympan*, *tympanite*}, {*vagin*, *vaginite*}, {*sinus*, *sinusite*}, {*méat*, *méatite*}.

4.5.4 Conclusion et perspectives

L'exploitation des différentes relations encodées dans une terminologie permet de généraliser la méthode et la rend applicable à n'importe quelle terminologie structurée. Nous avons exploité trois types de relations : lexicales, hiérarchiques et transversales. L'apport de chacune est important et leur combinaison améliore le rappel. La précision à l'étape de généralisation reste très bonne, entre 97,3 % et 98,5 %.

Même si la prise en compte de différentes relations dans les terminologies permet d'enrichir les connaissances morphologiques, le lexique des terminologies n'est pas exhaustif : des lexèmes sont absents ou bien ne reçoivent pas toutes les formes flexionnelles, ni les affixations ou compositions. Un travail sur des corpus (Xu & Croft, 1998; Zweigenbaum *et al.*, 2003b) doit permettre d'acquérir une plus grande variété de règles et de traiter un plus grand nombre de lexèmes[5].

[5]Ces attentes ne pas tout à fait justifiées, comme le montre la comparaison de la sec. 8.4 page 167.

La principale différence entre les différentes relations et terminologies explorées concerne les suffixes et règles de composition. Les suffixes de flexion et d'affixation semblent former un noyau commun assez stable, bien qu'ils présentent de nombreuses variantes flexionnelles et de segmentation. Il serait intéressant d'homogénéiser les variantes de segmentation et éventuellement de rapprocher les suffixes et préfixes induits des éléments morphologiques réels. C'est ce que nous essayons de faire dans la section suivante en effectuant un apprentissage endogène sur l'ensemble des formes de la liste de référence.

4.6 Ajustement et maximisation des suffixes

Notre objectif ici est d'éviter des troncations aléatoires des mots et d'induire des suffixes et préfixes linguistiquement plus pertinents. Pour cela, nous mettons en place une 'maximisation des suffixes' en effectuant un apprentissage endogène sur la totalité des formes de la liste de référence.

4.6.1 Adaptation de la méthode

La méthode standard est focalisée sur la détection des chaînes initiales communes maximales lors de la comparaison d'une paire de mots. Les préfixes sont alors étendus au maximum vers la droite et le premier caractère non commun marque la frontière entre le préfixe et les suffixes. Avec la maximisation des suffixes, un calcul complémentaire est effectué pour étendre en même temps ceux-ci au maximum vers la gauche. Ainsi lors de l'application d'une règle, les paires de mots qui la vérifient sont recherchées. Comme dans la méthode standard, une paire est retenue si le préfixe commun est suffisamment long. En un deuxième temps, nous effectuons un calcul supplémentaire pour rechercher, dans la paire de mots en question, un segment commun complémentaire qui étend les suffixes vers la gauche. Si ce segment est présent dans toutes les paires de mots d'une règle, c'est la version étendue des suffixes de cette règle qui est retenue. Les préfixes sont diminués en proportion.

La maximisation de suffixes est une étape facultative dans la méthode, laissée au choix de l'utilisateur. La maximisation ne modifie pas les paires de mots ni les familles induites. Elle touche simplement les segments de mots : préfixes et suffixes.

4.6.2 Matériel

La méthode ainsi modifiée peut être appliquée à n'importe quelle version du matériel. Nous l'appliquons ici aux séries de synonymes lemmatisées.

4.6.3 Analyse et discussion des ressources morphologiques induites avec la maximisation des suffixes

Le nombre des paires de mots mis en relations morphologiques et des familles est le même que dans l'expérience [LEM]. Il en va de même pour l'évaluation de la précision. Ce sont les segments de mots, préfixes et suffixes, qui sont modifiés par cette version de la méthode. Le nombre de règles et leurs rangs restent les mêmes mais les suffixes qui les composent peuvent changer formellement, comme dans les exemples du tab. 4.4. Dans ce tableau, nous indiquons la règle induite avec la méthode standard, le segment complémentaire calculé avec la méthode [MAX] et la règle qui résulte de cette maximisation.

Règle	Ajout		Règle étendue
$que\|e$	(i)	$\Longrightarrow$	$ique\|ie$
$x\|se$	(eu)	$\Longrightarrow$	$eux\|euse$
$ux\|l$	(a)	$\Longrightarrow$	$aux\|al$
$ve\|f$	(i)	$\Longrightarrow$	$ive\|if$
$me\|sarcome$	(o)	$\Longrightarrow$	$ome\|osarcome$
$me\|se$	(o)	$\Longrightarrow$	$ome\|ose$

TAB. 4.4 – Quelques exemples de règles avec des suffixes maximisés.

Avec la méthode [MAX] nous obtenons 583 suffixes et 1 252 préfixes, alors que dans la version [LEM] il y en avait 409 et 1 431 respectivement. Ainsi, le nombre de suffixes augmente et celui des préfixes diminue. Nous discutons maintenant ces différences. Un échantillon des résultats [MAX] se trouve dans l'annexe A.4 page 227.

Suffixes. L'examen des suffixes les plus fréquents montre que leurs rangs et fréquences d'apparition dans les paires de mots ne sont pas les mêmes que dans la version [LEM]. Il en va de même pour leur nombre.

Le premier suffixe dans les deux listes, -*e*, voit sa fréquence diminuer de 639 avec [LEM] à 556 avec [MAX] : il est maximisé selon les contextes et apparaît alors sous ses formes étendues.

Le suffixe qui occupe le 2^{e} rang dans [MAX] apparaît 288 fois dans les paires de mots. Le même suffixe est au 10^{e} rang dans les résultats [LEM], avec une fréquence de 102. En même temps, le suffixe [LEM] -*me*, 6^{e} rang 17 occurrences, n'apparaît pas dans les résultats [MAX]. Il est fusionné avec -*ome*, comme dans les paires suivantes :

{*lipome*, *lipoblastique*}, {*hémato*, *hématome*}, {*chondrome*, *chondroblastome*}.

Par ailleurs, le suffixe -*e* peut également être maximisé jusqu'à -*ome* à travers par exemple la règle étendue $ome|omateux$, qui est vérifiée sur les paires de mots comme :

{*lipome*, *lipomateux*}, {*athérome*, *athéromateux*}, {*fibrome*, *fibromateux*}, {*adénofibrome*, *adénofibromateux*}, {*cylindrome*, *cylindromateux*}.

Ainsi, selon les contextes, les suffixes [LEM] *-me* et *-e* sont maximisés jusqu'à *-ome*. Ce qui augmente la fréquence de ce dernier dans les paires de mots et le situe au 2e rang.

Le suffixe *-o* occupe le 3e rang dans les deux cas, mais il a une fréquence un peu plus importante dans [MAX] : il est détecté, grâce à la maximisation, dans plus de paires de mots.

Le suffixe *-ique* remonte aussi à la surface : 254 occurrences avec [MAX] *vs* 176 avec [LEM]. Comme dans le cas de *-ome*, il existe plus de paires de mots où il est détectable grâce à la maximisation.

L'extension des suffixes vers la gauche, lorsque c'est possible, conduit vers leur réarrangement. Certains, comme *-me*, disparaissent en nourrissant leurs formes étendues (*-ome*). D'autres, comme *-e*, sont étendus dans certains contextes seulement et selon ces contextes. À coté de cette tendance réunificatrice, les suffixes qui apparaissent dans les règles qui se vérifient que sur une seule paire de mots sont étendus au maximum vers la gauche... jusqu'au début de chaque mot de la paire. Les 'suffixes' correspondront alors aux mots entiers. Ceci peut être considéré comme un inconvénient de la méthode face aux règles hapax :

variante|*variation*, *utéro*|*utérin*, *tricholeucémie*|*tricholymphocyte*, *thrombus*|*thrombosé*, *thorax*|*thoracique*, *thrombopénie*|*thrombocytopénie*, etc.

Ce cas apparaît avec des emprunts (*thorax*, *thrombus*, etc.), avec des composés (*tricholeucémie*, *thrombopénie*, etc.) et les paires de lexèmes en relations morphologiques rares (*variante*|*variation*, *tache*|*tacheté*).

Dans certains cas, selon le contexte fourni par les paires de mots, la maximisation des suffixes est bloquée car elle ne se vérifie pas sur toutes les paires de mots. Par exemple la règle *se*|*tique*, qui génère entre autres la paire {*sténose*, *sténotique*}, n'est pas étendue jusqu'à *ose*|*otique*. Car d'autres paires, également générées par cette règle :

{*synapse*, *synaptique*}, {*cholestase*, *cholestatique*}, {*ostéolyse*, *ostéolytique*}, etc.

bloquent une telle maximisation.

Préfixes. Les préfixes, dont le nombre diminue, sont réajustés en même temps que les suffixes. Parmi les préfixes les plus fréquents, c'est le *-o-* final qui disparaît :

fibr-, *myél-*, *lymph-*, *osté-*, *adén-*.

Ce *-o-* est alors joint aux suffixes ou bien apparaît isolément. Notons que les données [LEM] peuvent comporter les deux versions de tels préfixes, voire plus :

fibro- et *fibr-* ; *myélo-*, *myél-* et *myélom-*.

4.6.4 Conclusion et perspectives

Lors de la maximisation des suffixes, ni les paires de mots ni les familles ne sont touchées. Par contre la décomposition des lexèmes en préfixes et suffixes n'est pas la même. Si l'ensemble des paires de lexèmes générées par une règle permet d'effectuer une extension des suffixes vers la gauche, le segment commun complémentaire est alors ajouté aux suffixes de la règle. Mais si une des paires ne permet pas la maximisation, celle-ci est bloquée et

les suffixes sont alors inchangés. Dans les cas de règles hapaxiques, suite à une extension maximale vers la gauche, ce sont les mots entiers qui sont retenus comme 'suffixes'.

Dans la majorité des cas, les lexèmes ainsi segmentés fournissent des préfixes et suffixes qui se rapprochent plus d'éléments morphologiques linguistiquement pertinents. Si nécessaire, un regroupement des éléments qui résultent alors pourrait être effectué en un deuxième temps avec la méthode de (Jacquemin, 1997a), comme nous le faisons dans la sec. 5.6.3 page 107.

4.7 Bilan de l'acquisition automatique de ressources morphologiques à partir de terminologies structurées et quelques perspectives

Comme dans d'autres travaux en acquisition de ressources morphologiques, notre méthode est basée sur la recherche de mots possédant une chaîne initiale commune suffisamment longue. Nous trouvons un contexte sémantiquement contraint à travers les relations sémantiques encodées dans une terminologie structurée. La méthode proposée demande donc un type particulier de matériel d'amorçage, les terminologies structurées. Elle est donc dépendante de leur existence et de leur disponibilité. Si la synonymie est une denrée terminologique plutôt rare, l'application de la méthode aux différentes relations sémantiques (lexicales, hiérarchiques et transversales) permet de la rendre plus générique et plus facilement applicable. La contrainte qui garantit la proximité sémantique des mots formellement proches reste forte avec tous ces types de relations et donne une précision très élevée. Nous avons appliqué la méthode à des langues de trois familles différentes : romane (français), germanique (anglais) et slave (russe). La précision se montre excellente (supérieure à 90 %) dans ces expériences. Malgré le faible taux d'erreur, les résultats induits doivent néanmoins être soumis à une validation humaine. Le rappel des paires de mots induites en anglais a été évalué par rapport aux résultats founis par `lvg` d'UMLS, la seule base de données morphologiques du domaine médical disponible publiquement. Le rappel pour les paires flexionnelles est de plus de 90 %. Il est moins bon pour l'affixation : un peu moins de 80 %. Il serait intéressant d'appliquer la méthode à d'autres langues indo-européennes, mais également à des langues non indo-européennes, par exemple les langues idéographiques (japonais, chinois) ou les langues sémitiques.

Lors de l'induction des connaissances morphologiques nous effectuons une segmentation des mots, ce qui nous permet d'obtenir (1) les chaînes initiales maximales communes, appelées 'préfixes', et (2) les chaînes finales, appelées 'suffixes'. Dans la méthode standard la segmentation est effectuée localement sur chaque paire de mots. Un même mot, en fonction des mots avec lesquels il est aligné, peut donc être segmenté de différentes manières et donner lieu à des 'allomorphies' dues à la segmentation. La maximisation des suffixes conduit vers une extension des suffixes vers la gauche et permet de réduire en partie de telles segmentations allomorphiques. Les méthodes de recherche de similarité entre deux chaînes de caractères, par exemple (Jacquemin, 1997a), peuvent également conduire vers

un regroupement de telles 'allomorphies'.

Les suffixes appris sur une paire de mots forment une 'règle'. Ces règles ont une valeur associative : étant donné deux suffixes, elles proposent de relier deux mots qui comportent ces suffixes et un préfixe d'une longueur minimale donnée. Mais souvent elles ne représentent pas une opération morphologique précise, comme la formation des adjectifs dénominaux, des noms déverbaux, etc. En même temps, la méthode n'est pas itérative : les mots sont soumis à une seule segmentation, alors que la plupart des mots du domaine se composent de plus de deux éléments morphologiques. Une segmentation plus fine, avec l'établissement d'un ordre dans les opérations morphologiques et de dépendances sémantiques, demande un modèle morphologique plus élaboré, comme celui implémenté dans l'analyseur `DeriF` (Namer, 1999). Une telle analyse rejoint les travaux en analyse morphosémantique des mots médicaux (sec. 3.2.3 page 57) et constitue une perspective de ce travail. Plus immédiatement, des connaissances sémantiques pourraient être introduites à travers un typage sémantique des suffixes. Ce typage peut être obtenu par exemple avec les axes sémantiques de la SNOMED où ces suffixes apparaissent le plus souvent.

Les données induites avec notre méthode permettent donc d'effectuer un premier abord du système morphologique d'une langue et en fournissent une description sommaire. Elles sont directement utilisables pour le traitement de la variation terminologique à travers, par exemple, les paires de mots, les familles ou mêmes les règles.

L'application des règles aux mots des terminologies médicales conduit à des résultats d'une précision remarquable. Mais le bruit alors généré constitue quant même une pierre d'achoppement à l'utilisation de ces règles. Leur précision risque par ailleurs de diminuer si elle sont appliquées aux mots provenant de corpus textuels dont la langue n'est pas aussi spécialisée et contrôlée que dans les terminologies. La fiabilité des règles peut alors être contrôlée à travers un seuil de confiance, basé par exemple sur le nombre d'exemples où ces règles sont vérifiées. Ce type de mesure de pertinence est souvent utilisé dans les travaux similaires (Gaussier, 1999; Pirrelli & Yvon, 1999). Il pourrait servir à réduire la taille de la chaîne initiale commune et à aider le traitement des mots courts sans perdre de précision. Des listes d'exceptions aux règles peuvent également être utilisées. Notons que leur existence sous-entend un travail déjà important dans l'analyse des données générées.

Si nous comparons le rendement des paires de mots en suffixes et en règles, il apparaît que le russe a le système morphologique le plus riche ou du moins qui se laisse découvrir le plus vite et avec peu de matériel. Il est suivi par le français et par l'anglais. Notons que la découverte des systèmes morphologiques n'est pas directement proportionnelle à la taille du matériel d'amorçage. Comme dans d'autres applications du TAL, nous atteignons assez rapidement un niveau minimal. Pour aller au-delà, il faut disposer de données d'amorçage bien plus volumineuses et variées. Effectivement les terminologies, même avec une bonne couverture du domaine médical comme la SNOMED (Elkin *et al.*, 2001), restent néanmoins loin de recenser tous les termes du domaine (Hersh *et al.*, 1997). Pour mieux couvrir le domaine médical, il faudrait compléter le matériel d'amorçage avec des corpus ou bien compléter l'acquisition avec des méthodes qui travaillent à partir de corpus (Xu & Croft, 1998; Zweigenbaum *et al.*, 2003b).

Nous avons appliqué la méthode à différentes relations sémantiques dans différentes ter-

minologies médicales. La flexion et l'affixation constituent un noyau commun assez stable. La différence provient de la composition et des 'allomorphies' de segmentation. Les résultats induits à travers ces différentes relations et terminologies s'avèrent complémentaires. Nous avons également appliqué la méthode au domaine de l'agriculture à travers le thesaurus AgroVoc (Zweigenbaum & Grabar, 2000). Il serait intéressant de l'appliquer à d'autres domaines de spécialité et de faire une comparaison des règles et suffixes induits. Nous nous attendons à ce que ce soit la composition qui constitue la principale différence dans les données générées.

Le fait de varier la taille de la chaîne initiale commune de 3 à 4 caractères améliore la précision, mais passe sous silence les paires de mots où le préfixe commun ne dépasse pas trois caractères ou bien qui présentent des allomorphies morphologiques avant le 4^{e} caractère. Les allomorphies, généralement non détectées avec notre méthode, de même que les supplétions, pourraient être enregistrées à l'avance (McCray *et al.*, 1994; Namer, 2000), au même titre que les exceptions aux règles morphologiques.

Lorque le travail est effectué sur des données étiquetées syntaxiquement, les règles deviennent syntaxiquement plus spécifiques et bloquent ainsi la génération de certaines paires de mots incorrectes. De manière générale, l'étiquetage fait augmententer la précision. Et il fait ressortir les conversions. Mais les mots polycatégoriels, qui ne reçoivent pas toutes les étiquettes possibles dans le matériel dont nous disposons, font diminuer le rappel : certaines paires de mots ne peuvent plus être générées. Il faudrait pouvoir assigner toutes les étiquettes possibles aux mots polycatégoriels, comme le suggère (Krovetz, 1993). L'utilisation de données lemmatisées permet de faire abstraction de la flexion et de se concentrer sur les procédés de formation des lexèmes, l'affixation et la composition. Avec la lemmatisation, la précision diminue : certaines erreurs persistent et, face au volume plus petit des données induites, prennent des proportions plus grandes.

La principale source d'erreurs générées avec notre méthode concerne les cas où les chaînes initiales communes n'ont pas de parenté sémantique, ceci malgré les contraintes concernant le contexte sémantique et la longueur minimale du préfixe. Les préfixes peuvent alors correspondre à des bases différentes ou à la présence de préfixes dérivationnels.

Nous avons travaillé indépendamment sur différentes langues. Mais il peut être intéressant de mettre en parallèle les terminologies internationales pour collecter le lexique et les ressources morphologiques multilingues (Baud *et al.*, 1997; Schulz *et al.*, 1999). Le travail en multilingue pourrait permettre d'améliorer les ressources morphologiques dans chacune des langues traitées : les chaînons manquants dans une langue peuvent être complétés sur la base des données d'autres langues. Par ailleurs, l'apprentissage des *cognats* (mots graphiquement et sémantiquement proches dans différentes langues : *sténose* en français *vs stenosis* en anglais) est une autre perspective que présente le travail sur les données multilingues alignées, comme les terminologies internationales.

Par rapport aux procédés décrits dans le chapitre 3, notre méthode permet de découvrir la flexion, la conversion, l'affixation (suffixation et préfixation) et la composition savante. La composition populaire ne semble pas apparaître dans les terminologies explorées. Nous disposons donc de ressources utiles pour le traitement de la variation morphologique des termes : flexion, suffixation et conversion.

Chapitre 5

Productivité quantitative des procédés morphologiques de l'adjectivation dénominale

La productivité des procédés morphologiques est globalement reflétée par la fréquence de leurs emplois. Notons cependant que cette notion est souvent utilisée sans être définie et peut recouvrir de ce fait différentes réalités morphologiques. Nous nous intéressons ici à la productivité quantitative des procédés morphologiques relatifs à la formation des adjectifs dénominaux. Une des idées sous-jacentes concerne la spécificité de ces procédés dans le domaine médical, selon ses sous-domaines et genres.

5.1 Introduction

Parmi les différents emplois de *productivité morphologique*, deux approches se dégagent : qualitative et quantitative (Dal, 2003). Elles peuvent être rapprochées des notions de disponibilité et de rentabilité (Corbin, 1987, p. 176–178) :

- La *disponibilité* d'un affixe concerne sa possibilité de construire des lexèmes. Elle prend en compte la nature de l'affixe et les contraintes de son application. Cette notion correspond aux approches qualitatives, qui montrent une image binaire de la productivité des affixes : ils sont productifs ou non productifs.
- La *rentabilité* d'un affixe concerne la possibilité qu'il a de s'appliquer à un grand nombre de bases et de former un grand nombre de lexèmes construits. Cette notion correspond aux approches quantitatives et consiste souvent en un recensement des formations d'un affixe. Dans ce cas, la productivité d'un procédé est une variable qui se situe sur un continuum qui va du non productif au très productif.

Dans ce chapitre, nous nous intéressons à la productivité quantitative et mesurons la productivité des procédés morphologiques de formation d'adjectifs dénominaux dans

Suffixe	Exemples
-aire	{*corpuscule/SBC*, *corpusculaire/ADJ*}
-ais	{*écosse/SBC*, *écossais/ADJ*}
-al	{*hémorroïde/SBC*, *hémorroïdal/ADJ*}
-é	{*sténose/SBC*, *sténosé/ADJ*}
-el	{*catégorie/SBC*, *catégoriel/ADJ*}
-eux	{*nausée/SBC*, *nauséeux/ADJ*}
-ien	{*microbe/SBC*, *microbien/ADJ*}
-ier	{*nourrice/SBC*, *nourricier/ADJ*}
-if	{*supression/SBC*, *supressif/ADJ*}
-in	{*utérus/SBC*, *utérin/ADJ*}
-ique	{*péricarde/SBC*, *péricardique/ADJ*}
-oire	{*respiration/SBC*, *respiratoire/ADJ*}
-ois	{*chine/SBC*, *chinois/ADJ*}

TAB. 5.1 – Procédés morphologiques d'adjectivation dénominale retenus.

la langue médicale française[1]. Ces adjectifs sont formés à partir de noms (d'où le nom d''adjectifs dénominaux') avec un ensemble de suffixes et reçoivent l'instruction sémantique générale « *Relatif à N* ». Ces suffixes, recensés en majorité dans les expériences antérieures (chap. 4), sont présentés dans le tab. 5.1. Notons tout de suite que *-é* est ambigu : dans les traitements automatiques, il peut être confondu avec *-é* de formation des participes.

Dans la suite de ce chapitre, nous présentons des travaux autour de la quantification de la productivité morphologique (sec. 5.2), décrivons notre méthode pour contraster la productivité des suffixes d'adjectivation dénominale dans différents corpus (sec. 5.3), le matériel (sec. 5.4 à 5.6) et discutons les résultats obtenus (sec. 5.7). Nous terminons avec une conclusion et des perspectives (sec. 5.8).

5.2 Travaux en quantification de la productivité morphologique

Les premiers travaux qui ont su formaliser la notion de productivité morphologique et la mesurer quantitativement ont consisté essentiellement en un recensement du nombre de types[2] qui comportent un affixe donné : dans les dictionnaires et parmi les néologismes

[1]Ce travail a été effectué dans le cadre de l'action 5 du GdR 2220 du CNRS 'Bases de données constructionnelles et productivité des procédés morphologiques en corpus'. Il a fait objet d'une publication (Grabar & Zweigenbaum, 2003) et a été présenté lors de la journée d'étude sur la productivité morphologique organisée par l'action 5 du GdR Morphologie le 12 mai 2003 à Paris.

[2]En linguistique de corpus, la notion de *type* correspond à chaque forme (ou vocable) différente qui apparaît dans ce corpus. Le type est souvent opposé à l'*occurrence*, qui correspond elle au nombre total d'apparitions d'un type (forme) dans le corpus.

de 'colloquial speech' (Schultink, 1962) ; uniquement dans les dictionnaires (Aronoff, 1976; Bauer, 1998) ; ou parmi les néologismes provenant du discours spontané (Clark & Clark, 1979) (travaux cités dans (Baayen & Renouf, 1996)). D'autres chercheurs appliquent des calculs probabilistes (écart réduit de la loi binomiale, analyse vectorielle, loi de Zipf, etc.) sur des corpus diachroniques (Brunet, 1981; Brunet, 1999) ou sur les données des dictionnaires (Wimmer & Altmann, 1995).

Nous nous intéressons aux calculs de productivité pouvant être effectués sur des corpus synchroniques et ayant un lien avec les travaux des morphologues. Deux définitions :
«

> *...by productivity as a morphological phenomenon we understand the possibility for language users to coin, unintentionally, a number of formations which are in principle uncountable...* » (Schultink, 1961)
>
> « *...the statistically determinated readiness with which an element enters into new combinations.* » (Bolinger, 1948)

auraient ainsi inspiré les statisticiens qui les formalisent et proposent la mesure de l'indice de productivité $\mathcal{P}$ (Baayen, 1991; Baayen & Lieber, 1991) : $\mathcal{P} = \frac{n_1}{N}$, où n_1 est le nombre d'*hapax* (types possédant une seule occurrence dans un corpus) d'un procédé morphologique et N le nombre total de ses occurrences rencontrées dans ce corpus.

La notion d'hapax semble être intimement liée aux néologismes (Aronoff, 1976; Bauer, 1998; Baayen & Renouf, 1996). Dans le dernier travail par exemple les auteurs montrent que ce sont les hapax qui fournissent le plus de néologismes (mots absents du dictionnaire *Webster's third new international dictionary of the English language*). Quant à la notion d'occurrence, elle est très exploitée dans la communauté d'origine des auteurs, la psycholinguistique. Dans ce domaine, de nombreux travaux et expériences ont montré que les mots les plus fréquents sont activés par les usagers dans leur stock lexical (ou lexique mental) plus facilement et plus vite que les mots moins fréquents. Il en va ainsi pour la reconnaissance orale de mots dans un flux bruité, pour les erreurs d'orthographe, pour la reconnaissance du lexique spécialisé par les spécialistes d'un domaine (travaux cités dans (Baayen, 1991)). Ainsi, en psycholinguistique (voir la synthèse de (Meunier, 2003)), on avance l'hypothèse selon laquelle (1) les mots fréquents sont stockés dans la mémoire tels quels et l'accès à ces mots est direct ; (2) les mots avec des fréquences faibles ne sont pas stockés en mémoire et l'accès à ces mots est basé sur des règles morphologiques. Faisant le lien avec la productivité, H. Baayen suppose que les mots fréquents, qui ont une forte activation, correspondent à des procédés non productifs : ils sont stockés en mémoire directement. Les mots peu fréquents par contre, qui ont un niveau d'activation bas, ne sont pas stockés en mémoire. Ils sont formés avec des affixes productifs et l'accès à ces mots se fait à travers des règles de formation (Baayen, 1991, p. 126), leurs bases et affixes.

Appliqué à un corpus, $\mathcal{P}$ permet d'exprimer la probabilité de rencontrer de nouveaux types avec un affixe donné lorsque N de ses occurrences ont déjà été rencontrées dans ce corpus (Baayen & Lieber, 1991, p. 809–810). La valeur de $\mathcal{P}$ peut donc être vue comme la potentialité qu'a un affixe à former des lexèmes étant donné un état de langue observé. La valeur de $\mathcal{P}$ est dépendante du corpus exploré. Il existe également une dépendance directe entre n_1 et N : plus n_1 est important plus $\mathcal{P}$ est élevé, mais avec l'augmentation de la taille

du corpus, et donc de N, $\mathcal{P}$ tend vers zéro (Baayen & Lieber, 1991, p. 836–838). Ce fait peut être vu comme une limite de la mesure lorsqu'elle est appliquée à de très gros corpus.

La mesure de productivité $\mathcal{P}$ a été critiquée par (Van Marle, 1991). Les critiques concernent (1) le caractère incertain de la notion d'hapax, (2) l'utilisation du nombre d'occurrences et (3) les résultats obtenus avec l'application de la mesure $\mathcal{P}$. En ce qui concerne les hapax, d'une part les lexèmes qui sont hapax dans un corpus ne le sont pas forcément dans un autre ; et d'autre part avec l'augmentation de la taille du corpus le nombre d'hapax diminue. Ce dernier point est également signalé dans (Baayen & Lieber, 1991). Notons également la discussion menée dans (Dal, 2003) sur le rapport entre les hapax et les néologismes. Quant au nombre d'occurrences, il aurait tendance à refléter l'usage des lexèmes, tandis que c'est le nombre de types qui reflète le système linguistique d'une langue ((Van Marle, 1991, p. 151), (Brunet, 1981, p. 427)). Le nombre de types serait ainsi plus pertinent pour évaluer la productivité d'un procédé morphologique. Mais (Van Marle, 1991) note qu'il ne suggère aucunement de modifier la formule de $\mathcal{P}$ en remplaçant le nombre d'occurrences N par le nombre de types, mais que la productivité semble être plus difficile à cerner que ne le laisse entendre la mesure $\mathcal{P}$. Et enfin la dernière critique de (Van Marle, 1991), qui concerne les résultats obtenus avec l'application de $\mathcal{P}$ et leur interprétation. Dans (Baayen, 1991) l'auteur compare la productivité des suffixes néerlandais : *-ster* ($\mathcal{P} =$ 0,231) de formation de noms personnels féminins apparaît ainsi être plus productif que *-er* de formation de noms personnels masculins ($\mathcal{P} = 0{,}076$). Ces résultats correspondent aux intuitions de (Baayen, 1991). Mais d'après (Van Marle, 1991, p. 154–155), *-er*, appartenant à une catégorie non marquée sémantiquement, est plus flexible et donc potentiellement plus productif que *-ster*. (Van Marle, 1991) considère ainsi contre-intuitif que les affixes marqués sémantiquement reçoivent un degré de productivité plus élevé.

Face à ces critiques, (Baayen, 1992) actualise la mesure $\mathcal{P}$, formalise la mesure de productivité globale $\mathcal{P}*$, et propose deux autres mesures I et $\mathcal{A}$ dont nous ne parlerons pas ici. Ainsi, (Baayen, 1992, p. 186) rappelle que l'observation de la productivité $\mathcal{P}$ doit se faire dans un espace bi-dimensionnel, où l'axe y indique l'indice de $\mathcal{P}$ et l'axe x la taille de V (nombre de types). Une telle projection offre une vue plus complète et globale de la productivité d'un affixe. L'auteur va plus loin dans la formalisation de cette relation et propose une nouvelle mesure $\mathcal{P}*$, ou l'indice de productivité globale : $\mathcal{P}* = \frac{n_1}{h_1}$ où n_1 est le nombre d'hapax d'un procédé et h_1 est le nombre total d'hapax dans un corpus. $\mathcal{P}*$ peut être vu comme une mesure de la contribution relative d'un procédé morphologique à accroître le vocabulaire d'un corpus à un moment t (Baayen, 1992, p. 190–194). Appliquée aux affixes néerlandais, les valeurs obtenues avec $\mathcal{P}*$ indiquent que *-er* est effectivement plus productif que *-ster*, ce qui correspond finalement aux intuitions de (Van Marle, 1991). Parmi ces deux mesures $\mathcal{P}$ et $\mathcal{P}*$, la première est dite conditionnée par l'affixe et la deuxième par les hapax (Baayen, 1992, p. 192). La première convient pour distinguer entre les affixes productifs et les affixes non productifs. Tandis que la deuxième est susceptible d'établir une échelle de productivité des affixes productifs (Baayen, 1992, p. 194). Un autre point distinctif est que $\mathcal{P}$ est dépendant de la taille des corpus et devient inexploitable sur de très gros corpus (plus de 100 000 000 occurrences), $\mathcal{P}*$ n'étant pas sensible à ce facteur (Baayen, 1992, p. 205).

5.3 Méthode pour l'étude de la productivité de l'adjectivation dénominale en corpus

Pour l'étude de la productivité des suffixes d'adjectivation dénominale nous appliquons la mesure $\mathcal{P}$ (Baayen & Lieber, 1991). Bien qu'étant critiquée par les morphologues, cette mesure reste l'une des plus utilisées dans le domaine. Elle a été conçue pour distinguer les procédés morphologiques productifs et non productifs en corpus synchronique. Elle correspond au rapport qui existe entre le nombre d'hapax n_1 d'un procédé morphologique et le nombre total de ses occurrences N rencontrées dans un corpus : $\mathcal{P} = \frac{n_1}{N}$. Rappelons que $\mathcal{P}$ reflète la potentialité d'un affixe à former des lexèmes dans un corpus donné. Les valeurs de $\mathcal{P}$ sont donc toujours associées au corpus étudié.

La collecte et préparation du matériel pour le calcul de $\mathcal{P}$ occupent une place prépondérante dans cette expérience. Nous utilisons des corpus textuels : ils représentent l'état actuel de la langue, contiennent des données langagières attestées, qui peuvent de plus être appuyées par leurs fréquences d'occurrence. Les données quantitatives collectées dans les corpus peuvent donc être directement utilisées pour le calcul de $\mathcal{P}$. Pour l'étude de la productivité des suffixes d'adjectivation dénominale, nous cherchons à les contraster dans différents sous-corpus qui varient selon leurs degré de spécialisation, genre et sous-domaine. Nous nous attendons à ce que ces paramètres influencent le contenu des corpus (Baayen, 2001, p. 208) et constituent une différence dans l'emploi d'adjectifs dénominaux. Les corpus étudiés doivent répondre à une série d'exigences, dont la représentativité du domaine médical. Nous décrivons les principes de la collecte et de la préparation des corpus dans la sec. 5.4. Nous décrivons ensuite le recensement des lexiques adjectival et nominal (sec. 5.5). Pour associer les adjectifs avec leurs bases nominales, nous utilisons les règles d'association. Les étapes de filtrage et de préparation de ces règles sont décrites dans la sec. 5.6. Et enfin les résultats de l'application de la méthode sont discutés dans la sec. 5.7.

5.4 Collecte et préparation du corpus

Un des soucis principaux lors de la constitution de corpus concerne leur représentativité du domaine étudié. Le corpus exploité doit donc couvrir suffisamment la langue médicale, formée de nombreuses spécialités et représentée par différents genres textuels. Mais même ainsi, les corpus fournissent une vue partielle de la langue. L'avantage est que les textes sont produits par les locuteurs réels du domaine et qu'ils reflètent l'état actuel de la langue, le délai entre leur production et leur utilisation étant potientiellement court.

Parmi les différents paramètres qui influent sur le contenu et la forme des documents, nous avons distingué, dans des travaux antérieurs de constitution du corpus *CLEF* du français contemporain (Habert *et al.*, 2001), les dimensions externes (facteurs liés au contexte de production du texte) et internes (facteurs que l'on retrouve dans le texte lui-même) des documents. Dans cette démarche, nous avons repris, fusionné et complété les propositions de (Sinclair, 1994) et de (Biber, 1994). Parmi les paramètres externes, nous avons distingué les informations de nature bibliographique (origine, localisation,

échantillonnage) de celles liées au contexte de production et de réception (mode de production et de transmission, créateur, destinataire, objectifs). Les paramètres internes comprennent la langue, la taille du texte, le niveau de style, l'interaction avec le public, la personnalisation, le genre, le sous-domaine et la technicité et spécialisation des textes.

Trois des paramètres internes sont pris en compte dans l'étude présente : technicité et spécialisation des documents, leurs sous-domaines et leurs genres. La *spécialisation* reflète le degré de spécialisation technique et scientifique d'un corpus. Dans notre expérience, nous contrastons un corpus de langue générale destiné à un large public de lecteurs et produit à partir de matériel journalistique et un corpus de la langue de spécialité représentée par la langue médicale. Les *sous-domaines* d'une langue de spécialité correspondent à ses différentes spécialités qui peuvent se recouvrir partiellement. Dans (Habert *et al.*, 2001), nous avons distingué environ 70 sous-domaines médicaux (cardiologie, néphrologie, neurologie, hématologie, rhumatologie, pédiatrie, etc.). Le sous-domaine d'un texte peut être décrit à l'aide de mots clés pris par exemple dans un thesaurus de référence, comme le MeSH. Et enfin, à l'intérieur d'un même domaine, différents *genres* (ou *registres*) peuvent être rencontrés. Dans (Habert *et al.*, 2001), nous avons recensé une trentaine de genres de textes médicaux, correspondant à des contextes caractéristiques de production ou de réception des documents : soin médical (rapports médicaux, comptes rendus d'hospitalisation, lettres), enseignement (polycopiés, questions de test, notes d'étudiants), diffusion de connaissances (articles de revues et de conférences, thèses de doctorat, articles de vulgarisation), sources de connaissances de référence (dictionnaires, encyclopédies, monographies, ouvrages, guides de bonnes pratiques, documents officiels, systèmes de codage). Dans la suite de cette section, nous précisons les sous-domaines et genres contrastés.

Le sous-domaine influe sur le lexique des documents, qui varie en fonction des notions qui y sont décrites. Le niveau de spécialisation et le genre exercent également une influence sur le contenu. Par exemple, les termes employés entre spécialistes dans un compte rendu d'hospitalisation peuvent être différents de ceux que l'on trouvera, à domaine constant, dans un manuel de cours, dans un article scientifique ou encore dans les documents à destination des patients : *grippe vs influenza*, *fièvre jaune vs tiphus amaril*, *crise cardiaque vs infarctus du myocarde*, *rhume vs rhinopharyngite*. Nous nous attendons à ce que ces différences se répercutent sur d'autres traits linguistiques, comme la morphologie et en particulier sur la formation d'adjectifs dénominaux. Dans la suite de cette section, nous décrivons la constitution et le contenu du corpus journalistique (sec. 5.4.1), du corpus de comptes rendus hospitaliers (sec. 5.4.2) et du corpus Web (sec. 5.4.3). C'est à l'intérieur de ces corpus que nous trouvons les paramètres retenus pour l'étude (sec. 5.4.4).

5.4.1 Le corpus journalistique

Le corpus journalistique (*presse*) que nous utilisons a été constitué par (Namer, 2003b)[3] pour une expérience similaire. Il a été collecté sur le Web, sur les sites de cinq périodiques : *Le Monde*, *Libération*, *Sciences et Avenir*, *Pour la Science* et *Challenges*. Ce corpus au

[3]Nous remercions Fiammetta Namer d'avoir mis à notre disposition ce corpus.

format texte totalise plus de 27,5 millions d'occurrences. Comme nous l'indiquons plus loin, c'est le corpus *Le Monde* que nous utilisons dans notre expérience.

5.4.2 Le corpus hospitalier

Les comptes rendus hospitaliers, qui composent majoritairement notre corpus médical, correspondent à un des genres du domaine médical. Ils contiennent l'information relative à la prise en charge, au diagnostic et au traitement d'un patient. Ce type de document est accessible uniquement à travers une collecte directe auprès des services hospitaliers[4]. Les documents hospitaliers sont anonymisés pour masquer les données personnelles (noms des patients et des médecins, dates, coordonnées postales et téléphoniques, références de dossiers, etc.) conformément à la loi *Informatique et Liberté*.

Dans notre expérience, nous utilisons deux corpus de comptes rendus provenant de deux sous-domaines médicaux :

Cardiologie : le corpus *cr-cardio*, constitué au cours du projet européen Menelas (Zweigenbaum *et al.*, 1995), est composé de comptes rendus d'hospitalisation et de lettres de sortie concernant des patients affectés de maladies coronariennes. Il compte 282 documents pour un total de 96 487 occurrences.

Hématologie : le corpus *cr-hémato*, rassemblé pour des tests d'aide au codage de diagnostics (Blanquet & Zweigenbaum, 1999), compte 100 comptes rendus d'hospitalisation totalisant 58 264 occurrences.

5.4.3 Le corpus Web

Pour constituer le corpus Web, nous exploitons le portail médical français de référence CISMeF (Catalogue et Index des Sites Médicaux Francophones) (Darmoni *et al.*, 2001)[5]. Il recensait, en septembre 2002, plus de 11 000 documents. Chaque document est indexé par un ou plusieurs mots clés du thesaurus médical MeSH, ce qui permet en particulier de noter la spécialité médicale concernée. Une équipe de documentalistes et d'informaticiens médicaux assure la mise à jour hebdomadaire des documents indexés dans CISMeF, ce qui garantit la qualité et la fiabilité de ces documents du point de vue médical. Les auteurs du catalogue CISMeF ont ainsi constitué et documenté ce que l'on peut considérer comme un corpus médical diversifié. Nous avons utilisé ce catalogue pour constituer les corpus Web médicaux[6]. Nous considérons que les documents Web, étant destinés dans leur majorité à un large public, s'inscrivent dans le genre de vulgarisation. Pour étudier les oppositions à l'intérieur de ce genre, nous avons constitué trois corpus Web médicaux s'opposant par leurs sous-domaines : hématologie (*web-hémato*), neurologie (*web-neuro*) et néphrologie (*web-néphro*). Pour étudier l'opposition sur la spécialisation des corpus, nous avons constitué

[4]Nous remercions les services de cardiologie et d'hématologie de la Pitié-Salpêtrière de nous avoir permis d'accéder aux documents hospitaliers.

[5]`www.chu-rouen.fr/cismef`

[6]Nous remercions l'équipe CISMeF, et en particulier Stéfan Darmoni, de nous avoir permis d'utiliser les documents recensés dans ce portail médical.

un corpus (*web-signes*) qui correspond à un extrait transversal des ressources de CISMeF : tous les documents ayant pour thème la catégorie MeSH *C23* (*signes et symptômes*), tous domaines et genres confondus.

Nous présentons maintenant les étapes qui permettent de passer d'une spécification de ressources CISMeF à un corpus de travail : (1) détermination des adresses URL des documents indexés par CISMeF et répondant à une spécification donnée, dans notre cas, à un des axes sémantiques du MeSH ; (2) téléchargement des documents correspondants ; (3) filtrage des documents et conversion en texte brut.

Détermination des adresses URL des pages à télécharger. Le catalogue CISMeF s'accompagne d'un moteur de recherche interne (Darmoni *et al.*, 2001). Une requête spécifiant les mots clés appropriés (par exemple, *néphrologie*) renvoie une page listant les documents indexés par ces mots clés avec pour chacune son adresse URL (Uniform Resource Locator). Étant des utilisateurs réguliers de CISMeF, nous avons téléchargé une copie de cette page.

Téléchargement des documents. Les URL recensées dans CISMeF pointent la plupart des cas sur la page Web contenant l'information indexée. Cependant, il arrive que les indexeurs de l'équipe CISMeF indiquent une page se trouvant un peu plus haut dans la structure hiérarchique du site concerné. C'est notamment le cas lorsque le document visé est fractionné en plusieurs pages : l'URL indique alors la page d'accueil du document, qui prend souvent la forme d'un sommaire. C'est également le cas lorsque le site change fréquemment le nom de la page visée : les indexeurs préfèrent alors pointer vers une page de nom plus stable. Pour prendre en compte ces cas, nous avons décidé de télécharger non seulement les URL indiquées par CISMeF, mais aussi tous les documents qui se trouvent hiérarchiquement un niveau plus bas sur le même site.

À partir d'une liste d'URL, nous téléchargeons automatiquement les documents à l'aide d'un 'aspirateur de Web' `wget` (outil GNU fonctionnant sous GNU/Linux comme sous d'autres systèmes d'exploitation), paramétré pour télécharger les URL de la liste et les URL filles :

```
wget --mirror --no-parent --level=1 --input-file=fichier-urls.txt
     --force-directories
```

Filtrage et conversion des documents. Les documents ainsi téléchargés peuvent être de types divers. Les plus courants sont des documents au format HTML, mais on trouve également des documents aux formats PDF, RTF, images, etc. Nous avons conservé uniquement les documents HTML, que nous pouvons convertir facilement en texte. Pour déterminer le type de document, le nom du fichier n'est pas suffisant : en effet, certains documents, produits par des programmes sur les serveurs Web (scripts CGI, ASP, etc.), ont un nom qui ne se termine pas par `.html` ou une autre extension spécifique. Nous avons utilisé la commande Unix `file`, qui détermine le type d'un fichier en analysant ses premières

lignes. Nous nous sommes fondés sur le résultat de `file` pour sélectionner les documents de type HTML.

Les traitements appliqués par la suite travaillent sur du texte brut. Nous avons donc converti dans ce format les documents téléchargés. Il était important lors de cette conversion de préserver les frontières de paragraphes. La conversion des documents HTML a été effectuée par un programme Perl fondé sur le module `Format::FormatText` en cherchant à respecter cette contrainte.

Le fait de télécharger davantage de documents que ceux explicitement listés par CISMeF, apporte des pages qui donnent une traduction de la page d'origine (en anglais, allemand, espagnol, etc.). Pour détecter et éliminer ces pages, nous avons mis en place un identificateur de langue suivant le principe décrit par (Grefenstette & Nioche, 2000). Il se fonde sur des listes de mots grammaticaux fréquents et spécifiques à chaque langue. Pour le français, par exemple, nous avons :

au aux avec ces cette dans est leur nous ont par pas plus pour qui ses sont sur une vous.

Nous avons repris telles quelles les listes publiées par (Grefenstette & Nioche, 2000).

5.4.4 Corpus et occurrences : les corpus à comparer

Tous les corpus constitués ont été segmentés, étiquetés et lemmatisés avec l'étiqueteur `TreeTagger` (Schmid, 1994) [7], et ensuite corrigés avec le lemmatiseur `Flemm`. Ce même traitement a été réalisé par (Namer, 2003b) sur le corpus *presse*. L'étiquetage nous a permis de recenser les formes étiquetées comme noms et adjectifs, dont nous retenons les lemmes. Mais certain mots, étiquetés comme noms ou adjectifs, présentent du bruit. Il s'agit par exemple de mots contenant des chiffres ou des caractères non alphabétiques, qui peuvent apparaître dans les documents d'origine ou bien résulter de la conversion en texte ou de la segmentation automatique en mots :

).Vancomycine
...pour
.chirurgie
1.-infection
10ans

Les données de la colonne *Toutes catégories* du tab. 5.2 incluent ce bruit de segmentation. Avant d'extraire les noms et les adjectifs, nous filtrons tous les mots contenant des chiffres ou des caractères non alphabétiques, ou encore des lettres majuscules. Ce dernier filtrage permet essentiellement d'éliminer les noms propres, souvent non reconnus comme tels et donc mal traités par les outils automatiques. Les colonnes *Adjectifs* et *Noms* décrivent les données épurées. À titre indicatif, après le filtrage du corpus *presse*, sa taille descend de 27 531 441 à 26 531 554 occurrences et de 201 571 à 88 892 types. Et dans le corpus *websignes*, la liste d'adjectifs non filtrés comprendrait 43 785 types (21 659 une fois filtrés), et la liste de noms 40 514 (38 025). Pour chaque corpus, nous avons comptabilisé le nombre de types et d'occurrences de ses lemmes nominaux et adjectivaux. La première constatation

[7] `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

	Toutes catégories		Adjectifs		Noms	
Corpus	types	occurrences	types	occurrences	types	occurrences
web-signes	142 545	5 204 901	21 659	507 162	38 025	1 188 574
web-hémato	42 230	1 032 216	7 893	115 658	14 101	241 939
web-neuro	106 153	3 794 737	17 612	369 130	31 402	871 155
web-néphro	30 524	726 545	6 053	95 176	10 699	171 421
cr-cardio	5 233	96 487	1 016	12 407	1 635	19 067
cr-hémato	5 306	58 264	1 008	6 304	1 838	13 980
presse	201 571	27 531 441	23 098	4 199 427	58 808	10 413 246

TAB. 5.2 – Taille des corpus initiaux après segmentation, étiquetage et lemmatisation par `TreeTagger` et `Flemm` : types et occurrences des lemmes.

à l'examen du nombre d'occurrences dans les corpus entiers (*Toutes catégories*) est qu'ils sont très inégaux en taille.

Or (Baayen, 2001) insiste sur le fait que les statistiques lexicales sont systématiquement affectées par la taille de l'échantillon de texte sur lequel elles sont mesurées. Nous ne pouvons donc pas travailler directement sur des corpus aussi variés en taille. Nous avons donc effectué un nivellement de la taille des corpus à l'intérieur de chaque groupe à contraster (spécialisation, sous-domaines et genres médicaux). Ce nivellement est fait par échantillonnage, le souci principal étant de veiller à la représentativité du corpus obtenu par rapport au corpus complet initial. L'échantillonnage a été effectué en tirant au hasard des textes entiers dans chaque corpus à réduire. L'intégrité de chaque texte est ainsi préservée, ce qui préserve son homogénéité interne. La mesure comparée lors du nivellement est le nombre d'occurrences d'un corpus. Nous supposons que ses autres caractéristiques (nombres de types, de noms, d'adjectifs, d'adjectifs dénominaux, d'hapax, etc.) sont inhérentes à ses dimensions stylistiques et seront conservées dans la réduction. Pour la langue générale, comme le corpus *Le Monde* possède une taille très proche de *web-signes*, nous avons préféré prendre ce corpus en entier plutôt que d'échantillonner les différents corpus de presse dont nous disposions.

Le tab. 5.3 présente les groupes de comparaison et les corpus nivelés que nous utilisons. Les tailles des corpus médical et général sont ainsi proches de 5 200 000 occurrences, les tailles des corpus venant du Web se situent entre 700 000 et 1 000 000 occurrences, et les corpus de comptes rendus médicaux avoisinent les 58 000 occurrences. Les corpus du sous-domaine d'hématologie comptent en moyenne 59 000 occurrences. Nous pouvons voir qu'avec des tailles de corpus comparables, le nombre d'adjectifs qu'ils contiennent peut varier jusqu'à une vingtaine de pourcents (comptes rendus médicaux, hématologie *vs* cardiologie ; *Le Monde vs* langue médicale) et le nombre de noms encore plus.

Corpus	Toutes catégories types	Toutes catégories occurrences	Adjectifs types	Adjectifs occurrences	Noms types	Noms occurrences
Langue générale (presse) *vs* langue médicale (Web)						
web-signes	142 545	5 204 901	21 659	507 162	38 025	1 188 574
Le Monde	88 052	5 233 503	20 453	415 337	42 513	1 116 548
Sous-domaines — documents Web : hématologie *vs* néphrologie *vs* neurologie						
web-hémato	42 230	1 032 216	7 893	115 658	14 101	241 939
web-neuro/4	45 077	950 114	7 749	91 260	14 698	218 379
web-néphro	30 524	726 545	6 053	95 176	10 699	171 421
Sous-domaines — comptes rendus médicaux : hématologie *vs* cardiologie						
cr-hémato	5 306	58 264	1 008	6 304	1 838	13 980
cr-cardio/2	4 004	58 213	807	7 786	1 299	11 652
Genres — hématologie : comptes rendus hospitaliers *vs* documents Web						
cr-hémato	5 306	58 264	1 008	6 304	1 838	13 980
web-hémato/15	7 588	60 658	1 594	7 355	3 037	14 445

TAB. 5.3 – Groupes de comparaison de corpus d'étude. Les corpus dont la taille a été réduite sont suivis du facteur de réduction appliqué ($/N$).

5.5 Lexiques nominal et adjectival

Les lexiques d'adjectifs sont compilés à partir de chaque corpus analysé. La taille de ces listes correspond aux nombres de types indiqués dans le tab. 5.3.

Dans le but de complétude lors de la reconnaissance d'adjectifs dénominaux, le lexique de référence de noms joint une composante médicale et une composante de langue générale. La composante médicale provient de deux terminologies médicales : le Répertoire d'anatomopathologie et la CIM (4 290 noms) et des corpus médicaux (55 517 noms). La composante générale provient du dictionnaire de l'ABU[8] (36 445 noms) et du corpus *presse* (58 808 noms). La fusion de toutes ces listes donne un lexique de référence de 105 417 noms différents.

5.6 Association semi-automatique des adjectifs dénominaux avec leurs bases

Nous avons noté que certains des lemmes étiquetés comme adjectifs constituent des erreurs. Nous avons pu en éliminer une partie grâce au filtrage de lemmes avec des chiffres, des caractères non alphabétiques ou des lettres majuscules. Mais d'autres erreurs restent. Pour assurer une meilleure qualité lors du recensement des adjectifs dénominaux, nous

[8]Association des bibliophiles universels, `abu.cnam.fr/DICO`.

utilisons des règles d'association pour effectuer un deuxième filtrage. Ainsi, nous recensons un adjectif comme dénominal s'il peut être associé à un nom de la liste de référence à travers une des règles de formation d'adjectifs dénominaux. Nous décrivons maintenant plus en détail les étapes de ce recensement : collecte des règles d'association (sec. 5.6.1), recensement des adjectifs dénominaux (sec. 5.6.2) et leur fusion en familles (sec. 5.6.3).

5.6.1 Collecte semi-automatique des règles d'association

Nous nous basons sur 513 règles induites à partir des séries de synonymes étiquetées et lemmatisées (sec. 4.4 page 77) avec un préfixe minimal commun de quatre caractères et une maximisation des suffixes. Nous avons vu que les relations sémantiques des terminologies constituent une contrainte favorable pour obtenir des données d'une précision élevée. Et la taille fixée à quatre caractères renforce cette contrainte. La maximisation des suffixes améliore la segmentation des mots et augmente ainsi la qualité des règles d'association. Quant à l'étiquetage et à la lemmatisation, ils permettent de réduire la flexion et d'appliquer ensuite des heuristiques catégorielles pour le repérage des règles de formation d'adjectifs dénominaux. Effectivement, parmi les 513 règles nous trouvons les types suivants :

- règles d'affixation :
  e/SBC|ique/ADJ {épigastr*e/SBC*, épigastr*ique/ADJ*},
  ϵ/ADJ|ment/ADV {clinique/*ADJ*, clinique*ment/ADV*},
- règles de conversion :
  ϵ/ADJ|ϵ/SBC {liquide/*ADJ*, liquide/*SBC*},
- règles de composition :
  e/SBC|ose/SBC {polype/*SBC*, polyp*ose/SBC*},
- ou bien des règles non élémentaires, qui associent un dérivé avec un composé sans passer par la base :
  aire/ADJ|ose/SBC {parasit*ose/SBC*, parasit*aire/ADJ*},
  inique/ADJ|ome/SBC {myél*ome/SBC*, myél*inique/ADJ*}.

Le premier filtrage que nous effectuons consiste à ne garder que les règles transcatégorielles, qui peuvent être rapprochées de l'affixation et de la conversion et qui peuvent filtrer la composition. Une règle comme *e/SBC|ose/SBC* n'apparaît donc plus. Suite à ce filtrage, nous obtenons 288 règles. Le filtrage suivant est effectué manuellement, il consiste à supprimer les règles transcatégorielles non élémentaires comme *aire/ADJ|ose/SBC* ou *inique/ADJ|ome/SBC*. Il nous reste alors 119 règles transcatégorielles. Un dernier filtrage, semi-automatique, consiste à garder les règles qui permettent d'associer un nom à son adjectif dénominal. Nous éliminons ainsi les règles dont le suffixe adjectival est vide, en supposant que l'affixation se fait par ajout d'affixes et non par leur suppression. Une analyse manuelle supplémentaire permet de filtrer des règles de formation de noms à partir d'adjectifs (*le/ADJ|ilité/SBC*), à partir d'autres catégories syntaxiques ou de formations latines. Il nous reste alors 73 règles. Nous complétons cette base manuellement en ajoutant des règles pour trois suffixes adjectivaux (*-in*, *-ais*, *-ois*) qui ne figurent pas dans les données générées dans les expériences antérieures. Nous disposons alors de 89 règles.

5.6.2 Recensement des adjectifs dénominaux

Pour recenser les adjectivations dénominales {*nom*, *adjectif dénominal*}, nous appliquons l'étape de généralisation de la méthode d'acquisition de ressources morphologiques (sec. 4.3 page 66) en appliquant ces 89 règles aux listes de noms et d'adjectifs. Ce recensement est semi-automatique :

- la partie automatique consiste à appliquer la méthode et à associer, à travers les règles, les bases nominales potentielles avec leurs adjectifs potentiels ;
- la partie humaine est la validation des couples candidats {*nom*, *adjectif dénominal*} ainsi générés.

La partie automatique génère généralement du bruit car les règles ne considèrent que la forme et la catégorie syntaxique des mots et pas leurs propriétés morphologiques, phonologiques ou sémantiques. L'objectif de la validation humaine est de limiter ce bruit. La partie automatique conduit également à un risque de silence : les règles établies, bien que complétées manuellement pour les formations manquantes, ne couvrent sans doute pas tout type de formations. Ceci en particulier dans le corpus de la langue générale.

Pour chaque corpus analysé, les adjectifs proviennent uniquement de ce corpus. Quant à la liste de noms, dans un souci de complétude des résultats, elle a été compilée à partir de plusieurs sources.

5.6.3 Fusion de suffixes 'allomorphiques' en familles

Lors de la description des expériences dans le chapitre 4, nous avons vu que la segmentation des mots en préfixes et suffixes est fonction des mots alignés et peut varier d'un alignement à un autre. C'est ce que nous avons appelé l'"allomorphie' due à la segmentation. La maximisation des suffixes réduit en partie cette allomorphie mais ne la supprime pas. Par exemple, le suffixe adjectival *-aire* se réalise ainsi dans nos règles : *-naire*, *-ulaire* et *-culaire*. Il faut donc regrouper ces allomorphes par familles. Nous avons cherché à le faire automatiquement en appliquant un algorithme de calcul de similarité entre deux chaînes de caractères (Jacquemin, 1997a). Les propositions générées, validées manuellement et complétées par la suite, nous ont fourni 10 familles de suffixes avec leurs formes allomorphiques (tab. 5.4). C'est donc à l'intérieur de ces familles que nous effectuons les décomptes nécessaires au calcul de l'indice $\mathcal{P}$ et à son interprétation : nombres d'hapax, d'occurrences et de types d'un suffixe.

5.7 Analyse et discussion du comportement des adjectifs dénominaux en corpus

Dans notre expérience, les groupes de comparaison sont constitués de corpus comparables en nombre d'occurrences. Nous supposons ainsi que les procédés morphologiques étudiés sont fonction des dimensions stylistiques de spécialisation, sous-domaine et genre.

Suffixe	Allomorphes regroupés
-aire	*-aire*, *-naire*, *-ulaire*, *-culaire*
-ais	*-ais*, *-çais*, *-elais*, *-nnais*, *-nais*, *-gais*, *-lais*
-al	*-al*, *-ial*, *-inal*, *-péritonéal*, *-arial*
-é	*-é*, *-né*, *-gé*
-el	*-el*, *-iel*, *-ionnel*, *-tiel*
-eux	*-eux*, *-ineux*
-ien	*-en*, *-ien*, *-gien*, *-arien*, *-idien*, *-éen*
-in	*-in*, *-guin*, *-atin*, *-llin*
-ique	*-ique*, *-tique*, *-atique*
-ois	*-ois*, *-cois*, *-çois*, *-geois*

TAB. 5.4 – Allomorphes regroupés automatiquement puis complétés manuellement.

Notons que dans des expériences similaires, (Namer, 2003b) retient l'ensemble des documents disponibles et (Fradin *et al.*, 2003; Gaeta & Ricca, 2003) effectuent un échantillonnage en nivelant les données par le nombre d'occurrences des affixations étudiées.

La mesure $\mathcal{P}$ permet de calculer la probabilité de rencontrer un nouvel adjectif avec un suffixe donné dans le corpus étudié. $\mathcal{P}$ manifeste une tendance générale à diminuer avec l'augmentation de la taille des corpus : le nombre d'occurrences N tend alors vers l'infini et le nombre d'hapax n_1 vers zéro. $\mathcal{P}$ devient carrément inexploitable dans les très gros corpus (plus de 100 000 000 occurrences (Baayen, 1992, p. 205)), ce qui ne constitue pas une limite dans notre étude. À cause de cette même dépendance à la taille de corpus, $\mathcal{P}$ sera plus élevé dans un corpus plus petit. Ce fait ne devrait pas être perceptible non plus : nos groupes de comparaison comportent des corpus de tailles comparables. Mais nous verrons que les petites différences des données de départ influencent légèrement les valeurs de $\mathcal{P}$.

Dans le tab. 5.5, nous présentons le nombre de règles qui ont pu être appliquées dans chaque corpus étudié et le nombre de paires {*nom*, *adjectif dénominal*} alors associées. Nous pouvons voir que le nombre de paires reste relativement comparable à l'intérieur de chaque groupe de comparaison, sauf dans le groupe de corpus contrastés par leur niveau de spécialisation : dans le corpus *Le Monde*, avec un nombre de règles utilisées assez élevé, le nombre de paires l'est beaucoup moins que dans son correspondant médical *web-signes*. La liste de noms de référence comporte pourtant les noms provenant des sources spécialisées et non spécialisées. Ce fait pourrait être rapproché avec la constatation de (L'Homme, 2004) : les adjectifs dénominaux sont très fréquents dans le corpus médical analysé par l'auteur ; mais aussi du fait que nos règles proviennent en majorité du domaine médical et ne sont sans doute pas très spécifiques du corpus général. La partie automatique de notre méthode, qui consiste en l'association des noms avec leurs adjectifs dénominaux potentiels génère du bruit : les règles ne considèrent que la forme des mots et pas leurs propriétés morphologiques ou sémantiques. L'objectif de la validation humaine est de limiter ce bruit. D'un autre côté, la base de règles, même complétée manuellement, n'est certainement pas parfaite, ce qui peut à l'inverse être cause de silence : des dérivés effectifs ne sont pas

Corpus	Règles	Couples
web-signes	79	3053
Le Monde	72	1871
web-hémato	75	1841
web-neuro/4	75	1774
web-néphro	74	1497
cr-hémato	62	414
cr-cardio/2	52	301
cr-hémato	62	414
web-hémato/15	65	628

TAB. 5.5 – Nombre de couples {*nom*, *adjectif dérivé*} repérés par les règles de dérivation adjectivale avec la liste de noms de référence dans les différents groupes de comparaison.

relevés, car aucune des règles fournies ne sait les mettre en relation avec leur base. Les règles apprises sur corpus (Zweigenbaum *et al.*, 2003b) devraient permettre de réduire le silence. Le lexique de noms de référence, qui fournit les bases candidates, peut lui aussi être source de silence. Il pourrait également être complété.

Dans la suite de cette section, nous contrastons d'abord l'indice $\mathcal{P}$ des suffixes d'adjectivation dénominale à l'intérieur de chaque groupe de comparaison (sec. 5.7.1). L'expérience menée a montré par ailleurs qu'il existe des adjectivations 'concurrentes' : lorsqu'une base nominale autorise l'application de plus d'un suffixe et permet de former plus d'un adjectif dénominal. Nous présentons une ébauche de leur analyse dans la sec. 5.7.2.

5.7.1 Différences de productivité des procédés d'adjectivation dénominale à l'intérieur des groupes de comparaison

Les corpus à l'intérieur de chaque groupe de comparaison sont différenciés selon les paramètres suivants :

1 spécialisation des documents : un corpus du domaine médical (*web-signes*) est contrasté avec un corpus de langue générale (*Le Monde*) ;

2, 3 sous-domaines médicaux différents à l'intérieur d'un genre : hématologie (*web-hémato*), neurologie (*web-neuro/4*) et néphrologie (*web-néphro*) dans les corpus de documents vulgarisés du Web, et hématologie (*cr-hémato*) et cardiologie (*cr-cardio/2*) dans les corpus de comptes rendus hospitaliers ;

4 genres différents à l'intérieur du domaine de l'hématologie : comptes rendus (*cr-hémato*) et documents du Web (*web-hémato/15*).

Dans les corpus examinés, le nombre d'hapax pour un procédé morphologique donné est souvent très faible, de l'ordre de quelques unités à quelques dizaines. Nous avons effectué une vérification manuelle de tous les hapax pour supprimer les erreurs. Parmi les dérivés hapaxiques erronés, nous avons rencontré des fautes d'orthographe (*ganglionaire*,

similiaire, *microsomial*, *differenciel*, *gestationel*), des variantes orthographiques (*macro-économique vs macroéconomique*, *micro-circulatoire vs microcirculatoire*), des absences ou erreurs d'accentuation (*ulceratif*, *endothelial*, *operationnel*, *couteux*), et des problèmes de césure (*franco-bri- tannique* qui donne l'adjectif *tannique*). Les données présentées dans ce qui suit tiennent compte de cette correction.

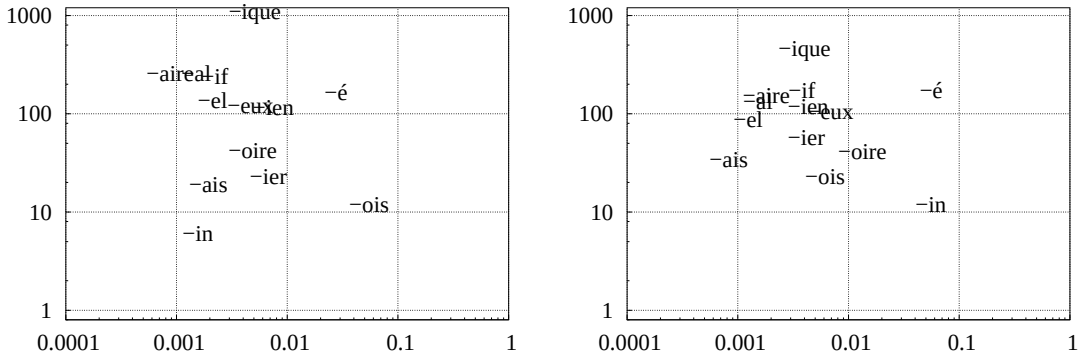
Nous faisons l'analyse en projetant l'indice $\mathcal{P}$ sur l'axe x dans un espace bi-dimensionnel, l'axe y étant réservé pour le nombre de types V d'un suffixe. Dans cette projection, un indice $\mathcal{P}$ élevé décale les suffixes vers la droite du graphe, un nombre de types V élevé les décale vers le haut. De manière générale, plus un suffixe est productif, plus il tend à se positionner dans le coin supérieur droit du graphe. Deux suffixes S_1 et S_2 étant donnés, ou un suffixe dans deux corpus, nous distinguons quatre cas de figures :

1. $\mathcal{P}_{S_1} > \mathcal{P}_{S_2}$ et $V_{S_1} > V_{S_2}$: S_1, avec un indice $\mathcal{P}$ et un nombre de V plus élevés, est plus productif que S_2. Il se rapproche alors du point 'idéal' (coin supérieur droit) ;
2. $\mathcal{P}_{S_1} > \mathcal{P}_{S_2}$ et $V_{S_1} = V_{S_2}$: S_1, avec un $\mathcal{P}$ plus élevé mais un nombre V égal, tend à être plus productif que S_2. Par son $\mathcal{P}$, il se rapproche encore du point 'idéal' ;
3. $\mathcal{P}_{S_1} = \mathcal{P}_{S_2}$ et $V_{S_1} > V_{S_2}$: S_1, avec un nombre de V supérieur mais un indice $\mathcal{P}$ égal, tend encore à être plus productif que S_2. Il permet de former plus d'adjectifs et son V élevé le rapproche alors du point 'idéal' ;
4. $\mathcal{P}_{S_1} > \mathcal{P}_{S_2}$ et $V_{S_1} < V_{S_2}$. Ce cas de figure ne permet pas de trancher facilement en faveur de la productivité ou non productivité d'un suffixe : S_2, ayant un vocabulaire V plus grand, montre logiquement un $\mathcal{P}$ plus petit, et vice versa. Pour pouvoir en dire davantage, on pourrait par exemple prélever des échantillons de corpus de tailles croissantes et comparer les courbes d'évolution de $\mathcal{P}$. Sinon faire appel à d'autres modèles (Baayen, 2001). L'approfondissement de ce cas reste une perspective[9].

1. Langue générale et langue de spécialité. La première comparaison contraste un corpus médical (*web-signes*) et un corpus de langue générale (*Le Monde*). La fig. 5.1 présente les résultats. Le suffixe *-é*[10] montre la productivité la plus élevée dans les deux corpus, mais c'est dans le corpus de presse que son $\mathcal{P}$ est plus important. Le suffixe *-in* a également un $\mathcal{P}$ plus important dans le corpus de presse, mais le faible nombre d'occurrences qu'il y montre (non représenté sur les graphiques, mais pris en compte dans le calcul de $\mathcal{P}$) peut être considéré comme insuffisant pour examiner sa productivité. Les suffixes *-eux* et *-oire*, avec des nombres de types V à peu près égaux dans les deux corpus, montrent un $\mathcal{P}$ plus élevé dans le corpus de presse. Quant au corpus médical, le suffixe *-el* y a un $\mathcal{P}$ deux fois plus important que dans le corpus de presse, et son nombre de types V y est également deux fois plus élevé. Avec des nombres de types à peu près égaux dans les deux corpus, *-ien* a un $\mathcal{P}$ plus élevé dans le corpus médical. *-ique* et *-al* y sont également plus productifs, avec des V plus élevés et des indices $\mathcal{P}$ pas très différents. Notons que *-ique* est

[9]Nous remercions Pierre Zweigenbaum de nous avoir aidée dans la compréhension et l'interprétation de ce cas de figure.

[10]Rappelons que *-é* reste ambigu entre la formation des adjectifs dénominaux et des participes si la totalité de ses formations n'a pas été revue à la main.



(a) Langue médicale (*web-signes*) (b) Langue générale (*Le Monde*)

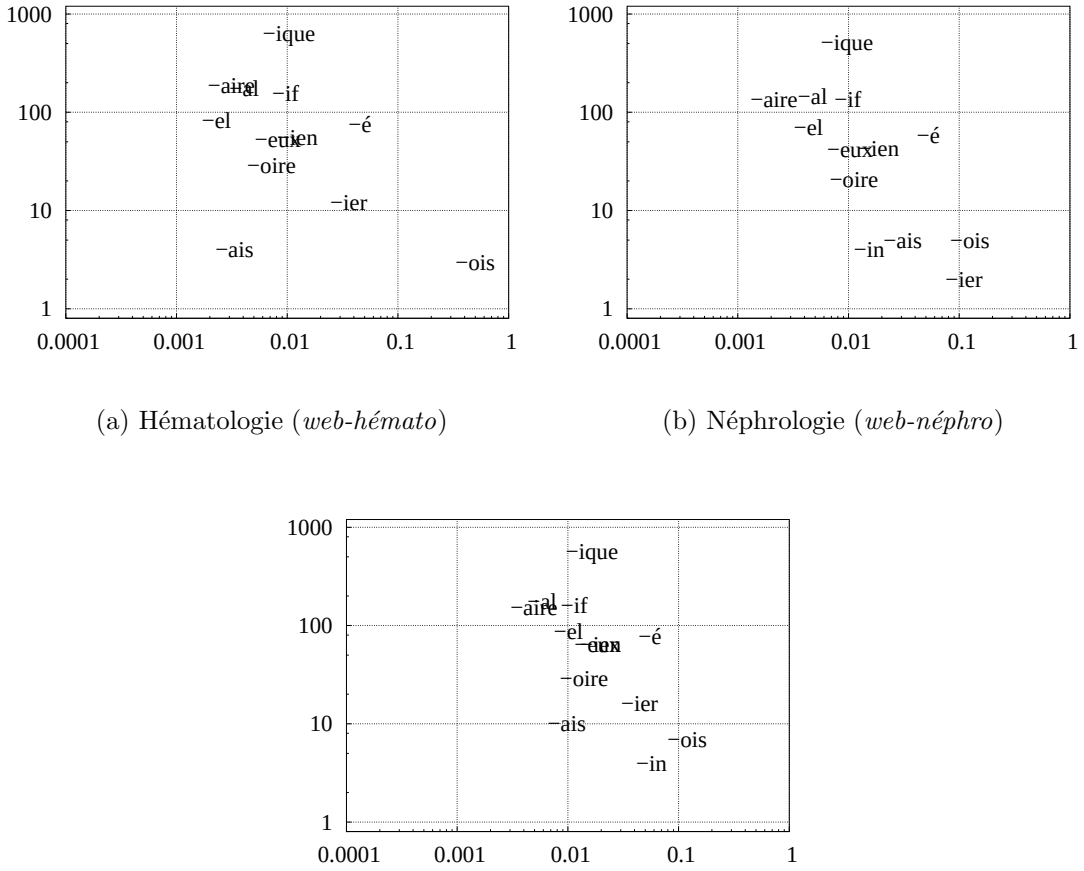
FIG. 5.1 – Positionnement des suffixes d'adjectivation dénominale dans un espace bi-dimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre langue de spécialité (*web-signes*) et langue générale (*Le Monde*). Échelle logarithmique.

de loin le suffixe le plus fréquent dans les deux corpus. Les cinq suffixes qui restent (*-ais*, *-ier*, *-ois*, *-aire* et *-if*) se trouvent dans le quatrième cas de figure, difficile à interpréter avec les projections que nous analysons. En résumé, le corpus *Le Monde* se fait remarquer par les suffixes *-é*, *-eux* et *-oire* ; le corpus médical *web-signes* par *-el*, *-ien*, *-ique* et *-al*. Remarquons néanmoins que la plupart des différences ne sont pas flagrantes.

2-3. Comparaison des sous-domaines à l'intérieur d'un genre. Les sous-domaines médicaux correspondent à des communautés scientifiques et sociologiques distinctes. Nous faisons deux oppositions : dans le genre des documents vulgarisés du Web, nous contrastons les spécialités d'hématologie (*web-hémato*), de néphrologie (*web-néphro*) et de neurologie (*web-neuro/4*) ; dans le genre des comptes rendus hospitaliers, nous opposons les spécialités d'hématologie (*cr-hémato*) et de cardiologie (*cr-cardio/2*).

La fig. 5.2 présente les résultats de la première opposition, au sein des documents du Web. Remarquons tout de suite que les procédés morphologiques du corpus *web-neuro* ont globalement l'indice $\mathcal{P}$ le plus élevé. Comme dans le groupe de comparaison précédent, le suffixe *-é* est encore le plus productif et *-ique* le plus fréquent. C'est le corpus de neurologie qui se distingue par la productivité de nombreux suffixes, par rapport à un ou même les deux corpus du groupe. *-eux* et *-if* y ont ainsi une productivité globale plus élevée, même si les différences sont faibles par rapport aux deux autres corpus. *-al*, *-el*, *-ique* et *-oire*, avec un V comparable, y montrent un indice $\mathcal{P}$ plus élevé. *-aire*, avec un $\mathcal{P}$ plus élevé, a un V plus élevé que dans le corpus de néphrologie mais moins élevé que le corpus d'hématologie. *-ien* a également un $\mathcal{P}$ plus élevé, mais un nombre de types comparable par rapport au corpus d'hématologie. Il a plus de types et un $\mathcal{P}$ comparable dans le corpus de néphrologie.

Si *-é* est globalement productif dans les trois corpus, il se trouve en meilleure position dans *web-néphro*. Les suffixes *-ais*, *-ier*, *-in* et *-ois* ont des occurrences faibles dans au moins un des corpus et manifestent en plus le rapport délicat du 4e cas de figure. Au total, on note peu de différences importantes dans ces trois corpus entre les procédés morphologiques les plus fréquents. Mais de manière générale, c'est dans le corpus de neurologie que les suffixes sont les plus productifs (*-eux*, *-if*, *-al*, *-el*, *-ique*, etc.) et dans le corpus d'hématologie qu'ils le sont le moins.



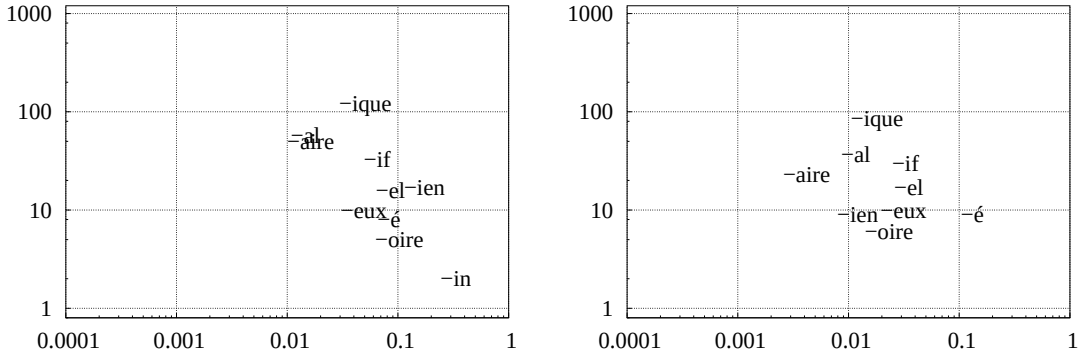
(a) Hématologie (*web-hémato*)

(b) Néphrologie (*web-néphro*)

(c) Neurologie (*web-neuro/4*)

FIG. 5.2 – Positionnement des suffixes d'adjectivation dénominale dans un espace bi-dimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre trois sous-domaines médicaux des documents du Web : hématologie, néphrologie et neurologie. Échelle logarithmique.

La fig. 5.3 oppose l'hématologie et la cardiologie dans les comptes rendus d'hospitalisation. Globalement les suffixes ont un nombre de types et d'occurrences plus bas mais



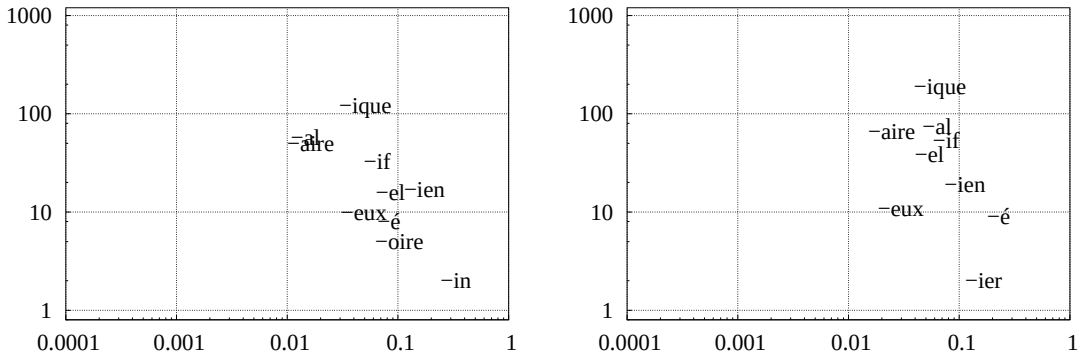
(a) Hématologie (*cr-hémato*)

(b) Cardiologie (*cr-cardio/2*)

FIG. 5.3 – Positionnement des suffixes d'adjectivation dénominale dans un espace bi-dimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre deux sous-domaines médicaux des documents hospitaliers : hématologie et cardiologie. Échelle logarithmique.

un $\mathcal{P}$ plus élevé par rapport aux deux groupes de comparaison précédents : la taille des corpus dans le groupe de comptes rendus est plus petite. *-é* est le seul suffixe qui montre une productivité supérieure dans le corpus *cr-cardio/2*, mais il a une fréquence trop basse pour être pris en compte. *-ais*, *-ier*, *-ois*, *-in*, *-oire* et *-ien* sont d'autres suffixes avec des fréquences non significatives. Parmi les suffixes qui restent, *-aire*, *-al* et *-ique* ont un indice $\mathcal{P}$ et un nombre V supérieurs dans le corpus *cr-hémato*. *-el*, *-eux* et *-if*, avec un nombre de types comparable, y montrent un $\mathcal{P}$ supérieur. En résumé, la productivité dans *cr-hémato* est plus importante pour plusieurs procédés morphologiques : *-ique*, *-if*, *-el*, *-aire*, etc. Le corpus *cr-cardio/2* est plus répétitif, avec davantage d'occurrences pour les mêmes types : 1 008 types pour 6 304 occurrences pour l'hématologie contre 807 types et 7 786 occurrences pour la cardiologie.

4. Entre les genres médicaux à l'intérieur d'un sous-domaine. Les genres médicaux peuvent également être corrélés à des communautés sociologiques différentes (par exemple : étudiants, patients, praticiens, scientifiques, etc.), et à des emplois lexicaux et morphologiques différenciés. Deux genres sont constrastés dans le sous-domaine de l'hématologie : des documents du Web (*web-hémato/15*) et des comptes rendus médicaux (*cr-hémato*). La fig. 5.4 présente les données sur la productivité des procédés morphologiques dans ces corpus. De nombreux suffixes (*-ais*, *-el*, *-ier*, *-if*, *-in*, *-oire*, *-ois* et *-é*) ont des fréquences basses dans ce groupe de comparaison, ce qui les rend statistiquement insignifiants. Les suffixes *-aire*, *-al* et *-ique* sont plus productifs dans le corpus Web autant par rapport à l'indice $\mathcal{P}$ que par le nombre de types V. *-eux* et *-ien* sont par contre plus



(a) Documents hospitaliers (*cr-hémato*) (b) Documents du Web (*web-hémato/15*)

FIG. 5.4 – Positionnement des suffixes d'adjectivation dénominale dans un espace bi-dimensionnel : productivité $\mathcal{P}$ et nombre de types V. Comparaison entre deux genres médicaux du sous-domaine d'hématologie : comptes rendus hospitaliers et documents du Web. Échelle logarithmique.

productifs dans le corpus des documents hospitaliers.

Bilan. Les observations restent attachées au corpus spécifique et au groupe de comparaison dans lesquels elles ont été effectuées. Par exemple, un même corpus *cr-hémato* fait émerger les suffixes *-eux* et *-ien* par rapport à *web-hémato* (fig. 5.4(a)), mais *-aire*, *-al*, *-ique*, *-el*, *-eux* et *-if* par rapport à *cr-cardio/2* (fig. 5.3(a)), avec un suffixe commun (*-eux*) entre les deux groupes. La généralisation des constatations à d'autres textes du même genre ou du même domaine n'est pas acquise — et ce d'autant plus que les corpus sont de petite taille. L'indice $\mathcal{P}$ est ainsi très sensible à la taille de corpus. D'une part, dans le corpus *web-hémato* échantillonné (fig. 5.4(b)) les suffixes sont globalement plus productifs, tandis que dans ce même corpus pris en entier (fig. 5.2(a)) l'indice $\mathcal{P}$ est moins élevé avec un nombre de types globalement plus élevé. D'autre part, dans le premier groupe de comparaison, le $\mathcal{P}$ de *-é* légèrement supérieur dans le corpus de presse peut être considéré comme neutralisé par la taille des données de départ (tab. 5.3) : le nombre d'adjectifs et de noms étant plus élevé dans le corpus médical, l'indice $\mathcal{P}$ y est automatiquement moins élevé. De manière générale les constrastes observés ne sont pas très flagrants au fil des comparaisons, sauf peut-être : (1) le suffixe *-el* dans le 1er groupe, (2) le corpus *web-neuro/4* dans le 2e, (3) le corpus *cr-hémato* dans le 3e. Les graphiques montrent par ailleurs que les couples de suffixes *-aire* et *-al*, *-eux* et *-ien* se trouvent souvent à proximité. *-ique* est de loin le plus fréquent et *-é* le plus productif dans les corpus de taille plus importante. Les procédés de formation d'adjectifs dénominaux présentent donc globalement peu de différences. Ils semblent constituer une sorte de trame morphologique. Pour confirmer ces

résultats il faudrait effectuer des expériences sur d'autres corpus, voire appliquer d'autres calculs. Il serait ainsi intéressant d'effectuer une analyse dynamique des corpus en prélevant des échantillons de taille croissante (Baayen, 2001). Ce qui permettrait peut-être d'avoir un jugement sur les suffixes avec des fréquences faibles (quelques dizaines d'occurrences), considérés actuellement comme statistiquement non pertinents.

Notons aussi que nous avons trouvé une certaine difficulté à interpréter les données statistiques : appliquée au matériau linguistique, l'interprétation devrait être appuyée par des analyses linguistiques parallèles.

5.7.2 Étude des adjectivations 'concurrentes'

Ce que nous appelons les adjectivations 'concurrentes' correspond aux cas où une base nominale sélectionne plus d'un suffixe et produit ainsi plus d'un adjectif[11] :

grippe $\Longrightarrow$ *grippal* *vs* *grippé*
sulfure $\Longrightarrow$ *sulfurique* *vs* *sulfureux*
muscle $\Longrightarrow$ *musculaire* *vs* *musculeux*

Si les adjectifs concurrents avaient exactement le même sens construit, l'économie linguistique les empêcherait de coexister. Mais comme ces adjectifs restent dans la langue, nous supposons qu'il y a une différence sémantique entre eux. Effectivement, le sens d'un lexème construit résulte de la combinaison de l'instruction sémantique de la règle, du sens hérité de la base et du sens de l'affixe (Corbin, 1991). Selon l'auteur, la formation de tous les adjectifs dénominaux est soumise à une même opération morphologique qui transmet aux adjectifs l'instruction sémantique générale « *Relatif à N* ». Il faudrait sans doute chercher la différence entre les adjectifs concurrents dans le sens des suffixes et les traits sémantiques que ces suffixes sélectionnent dans les bases. En nous basant sur les écrits des morphologues, nous tâchons de dégager un ensemble de facteurs statistiques et sémantiques capables d'expliquer l'existence d'adjectifs concurrents. Nous faisons pour terminer quelques synthèses sémantiques de ces procédés dans le domaine médical. La phonologie, également importante lors de la formation de lexèmes (Lignon, 1999; Plénat & Roché, 2001a), n'est pas considérée.

Préparation et nettoyage des données. Les données analysées proviennent du corpus de presse (Namer, 2003b) (ensemble de documents provenant de *Le Monde*, *Science et Avenir*, *le Nouvel Observateur*, *Challenges* et *Pour la Science*) et de l'ensemble de corpus médicaux (*web-signes*, *web-neuro*, *web-néphro*, *web-hémato*, *cr-cardio*, *cr-hémato*). Lors du nettoyage des adjectifs générés automatiquement, nous avons repéré des cas d'adjectifs dénominaux multiples pour une même base qui ne doivent pas être considérés comme des adjectivations concurrentes au sens où nous l'entendons ici (entre parenthèses sont indiquées les fréquences de chaque forme) :

[11] Cette étude a été effectuée avec Fiammetta Namer et Pierre Zweigenbaum et présentée à l'occasion de la journée d'étude sur la productivité morphologique organisée par l'action 5 du GdR Morphologie le 12 mai 2003 à Paris.

- Réalisation de /s/ en 't' ou 'c' devant 'i' :
  adventice $\Longrightarrow$ *adventiciel* (5) *vs* *adventitiel* (3)
  interstice $\Longrightarrow$ *intersticiel* (22) *vs* *interstitiel* (1087)
- Blocage de l'allomorphie 'ai' → 'a' devant *-ien* :
  urticaire $\Longrightarrow$ *urticairien* (4) *vs* *urticarien* (17)
- Association de noms français avec des adjectifs d'origine étrangère (anglais, etc.) :
  autosome $\Longrightarrow$ *autosomal* (47) *vs* *autosomique* (383)
  myocarde $\Longrightarrow$ *myocardial* (23) *vs* *myocardique* (545)
- Spécialisation du sens des adjectifs :
  catégorie $\Longrightarrow$ *catégoriel* (2) (*Relatif à catégorie*)
  catégorie $\Longrightarrow$ *catégorique* (6) (non *Relatif à catégorie*)
- Bases qui, bien que formellement très semblables, ont des origines différentes :
  klôn (*pousse*) → *clone* $\Longrightarrow$ *clonal* (328)
  klonos (*agitation*) → *clone* $\Longrightarrow$ *clonique* (140)

Fréquences des suffixes concurrents. D'après (Aronoff & Anshen, 1998), parmi les affixes concurrents il y en a toujours un qui s'applique par défaut. C'est celui-là, présentant le moins de contraintes, qui va montrer les fréquences les plus hautes. Le tab. 5.6 indique les fréquences des suffixes qui s'appliquent aux mêmes bases nominales dans les corpus *presse* et *web-signes*. Nous avons également recensé le nombre de types correspondants. Mais comme leur évolution est comparable avec les occurrences, nous indiquons seulement ces dernières. Dans le couple *-al*/*-é*, le suffixe *-al* apparaît 472 fois dans le corpus de presse, *-é* 96 fois. Selon l'hypothèse énoncée ci-dessus, *-é* est plus contraint que *-al* dans le corpus de presse. Il en est de même dans le corpus médical. Les paires de suffixes *-el* et *-aire*, *-ique* et *-é* sont dans le même cas. La spécificité des suffixes des paires *-aire* et *-al*, *-eux* et *-ique* est inversée. Les paires de suffixes qui ont des fréquences trop insignifiantes dans un des corpus ne sont pas représentées dans le tableau.

	presse		*web-signes*	
Suffixes	Freq	Spécificité	Freq	Spécificité
-al/*-é*	472/96	*-al* > *-é*	263/12	*-al* > *-é*
-aire/*-al*	29/87	*-aire* < *-al*	1720/289	*-aire* > *-al*
-aire/*-el*	49/145	*-aire* < *-el*	40/616	*-aire* < *-el*
-eux/*-ique*	67/134	*-eux* < *-ique*	284/32	*-eux* > *-ique*
-ique/*-é*	176/25	*-ique* > *-é*	1209/280	*-ique* > *-é*

TAB. 5.6 – Fréquence et spécificité des suffixes concurrents.

Interaction entre les suffixes et les bases nominales. Les suffixes ont la propriété de sélectionner un type sémantique de la base nominale à laquelle ils peuvent s'appliquer et parfois d'y sélectionner des propriétés sémantiques différentes (Corbin, 1992) :

-ique semble être propre au domaine technique : c'est avec le développement des écrits techniques qu'il montre des fréquences remarquables (Brunet, 1981). Appliqué aux mêmes bases que *-eux*, *-ique* est plus spécifique dans le corpus médical, moins spécifique dans le corpus journalistique. Appliqué aux mêmes bases que *-é*, *-ique* est moins spécifique dans les deux corpus.

-eux s'applique à des noms concrets perceptibles sensoriellement, il indique une relation de consubstentialité et une cause interne (Corbin, 1992) :
peau boutonneuse est couverte de boutons, est susceptible de secréter des boutons. Le tableau 5.6 le montre plus spécifique que *-ique* dans le corpus de presse, moins dans le corpus médical.

-é indique une cause externe d'un état (*patient grippé*). Mais, selon (Aliquot-Suengas, 1996, p. 228–233), il marque surtout une idée de typicité et de saturation : *artère méningée* se trouve dans les méninges et en est typique. Il est plus spécifique que *-al* ou *-ique* dans les deux corpus.

-ien, *-aire*, *-al*, *-el*, etc. semblent pouvoir s'appliquer à des bases très variées. Par exemple *-ien* (Corbin, 1992) : zone territoriale (paris*ien*) ; objet concret, partie du corps (crân*ien*) ; nom propre (parkinson*ien*).

Interaction entre les bases nominales et les noms recteurs. Les contraintes portent non seulement sur la base nominale N_2 sélectionnée par les suffixes mais également sur le nom recteur N_1, avec lequel l'adjectif dérivé apparaît dans un groupe nominal N_1 $[[N_2]\text{aff}]_A$. Il existe deux types de relations entre N_1 et N_2 (Mélis-Puchulu, 1991) : l'appartenance et la possession. Avec la relation d'*appartenance*, le nom recteur N_1 représente le tout et la base nominale N_2 une de ses parties. Cette relation concerne *-é*, *-aire*, *-eux*, *-in* et *-ique* :

couche musculeuse : « *couche composée de tissu musculaire, de muscles* »,
viscère herniée : « *viscère qui comporte une hernie (orifice)* »,
thrombose hémorroïdaire : « *thrombose avec des hémorroïdes* »,
acide phosphorique : « *acide dans la composition duquel entre le phosphore* ».

Avec la relation de *possession*, le lien entre N_1 et N_2 est inverse : N_2 représente le tout, N_1 une de ses parties. Cette relation concerne *-al*, *-aire*, *-el*, *-ien*, *-in*, *-ique* et *-ois* :

fracture vertébrale : « *fracture des vertèbres, qui se trouve dans la colonne vertébrale* »,
collapsus pharyngé : « *collapsus (affaissement) qui se trouve dans le pharynx* »,
cellule kystique : « *cellule qui se trouve dans un kyste, qui compose un kyste* ».

Notons que les suffixes *-aire*, *-in* et *-ique* sont ambigus et peuvent apparaître dans les constructions des deux types :

- Un *disque lombaire* ou un *abcès lombaire* se trouvent dans la région lombaire, la partie basse du dos. Tandis que la *colonne lombaire* est une partie de la colonne vertébrale.
- Une *hémorragie kystique* ou une *infection kystique* sont localisées dans un kyste, tandis qu'un *ovaire kystique* ou un *rein kystique* ont tendance à développer des kystes.

Mais surtout les relations sémantiques pouvant exister entre le nom recteur N_1 et la base nominale N_2 dans un domaine de spécialité comme la médecine ne peuvent pas toujours

être exprimées avec les deux relations génériques proposées par (Mélis-Puchulu, 1991). La caractérisation est difficile et incomplète dans les groupes nominaux comme : *complication hémorroïdaire* (complication développée suite à l'apparition d'un hémorroïde), *pathologie hémorroïdaire* (pathologie caractérisée par l'apparition d'hémorroïdes), *douleur herniaire* (douleur causée par des hernies), *chirurgie herniaire* (chirurgie qui a pour vocation de traiter les hernies). Dans le paragraphe suivant, nous essayons de dégager des relations plus spécifiques. Pour ceci, nous effectuons d'abord un typage sémantique des noms N_1 et N_2, en nous basant sur les axes sémantiques de la SNOMED, et ensuite un typage de leurs relations.

Typage sémantique des noms et de leurs relations sémantiques. Nous avons analysé les syntagmes nominaux N_1 $[[N_2]\text{aff}]_A$ dont les adjectifs sont formés sur des bases nominales N_2 appartenant aux axes sémantiques (1) de produits chimiques et de médicaments (*azote* (*nitre*), *phosphore*, *sulfone*, *soufre*, *vaccin*), (2) de pathologies ou de maladies (*grippe*, *hernie*, *kyste*, *macule*, *tuberculose*), et (3) de parties du corps (*lobule*, *muscle*, *pharynx*, *vertèbre*, *vésicule*).

Les syntagmes où N_2 désigne un produit chimique sont assez aisés à analyser :

- Si N_2 est suffixé en *-al* et N_1 correspond également à un produit chimique, le sens de l'expression correspond à la relation générale d'appartenance : *agent calmant sulfonal*, *souche vaccinale*. Si N_1 est une pathologie, le sens de l'expression reflète une causalité (N_1 *qui est dû à* N_2) : *encéphalomyélite post-vaccinale*, *papule vaccinale*, *réaction vaccinale*. Sinon, l'expression donne un sens général « N_1 *relatif à* N_2 » : *calendrier vaccinal*, *échec vaccinal*, *rappel vaccinal*.
- Si N_2 est suffixé en *-é* et N_1 désigne également un produit chimique le sens global du syntagme est « N_1 *riche en* N_2 », avec une idée de saturation (Aliquot-Suengas, 1996) : *composé nitré*, *sel nitré*, *dérivé sulfoné*. Lorsque N_1 désigne une pathologie le sens de l'expression pourrait être « N_1 *causé par (la présence/absence de)* N_2 » : *nécrose phosphorée*, *carence phosphorée*.
- Si N_2 est suffixé en *-ique* ou en *-eux* le sens est « N_1 N_2*ique contient plus d'atomes de* N_2 *que* N_1 N_2*eux* » : *oxyde nitrique vs oxyde nitreux*, *acide sulfurique vs acide sulfureux*, mais aussi *vapeur nitreuse*, *vapeur sulfureuse*, *ester phosphorique*. Dans la langue générale *sulfureux* est surtout employé au sens figuré (« *Qui rappelle le démon, l'enfer, qui sent le soufre* » (Robert, 1993)) : *caractère sulfureux*, *contexte sulfureux*, *idée sulfureuse*, *le Polonais sulfureux*, etc.

Lorsque N_2 désigne une pathologie N_1 correspond souvent à une partie du corps ou une autre pathologie :

- Si N_2 est suffixé en *-é* et N_1 est un être animé ou une de ses parties du corps, le sens de l'expression est « N_1 *subit un état* N_2 » ou une localisation (« N_2 *localisé dans* N_1 »): *patient grippé*, *anse herniée*, *tumeur enkystée*.
- Si N_2 est suffixé en *-eux* et N_1 est un être animé, le sens de l'expression est « N_1 *a tendance à* N_2 », où N_2 constitue une propriété intrinsèque de N_1 : *malade tuberculeux*. Si N_1 est une pathologie, nous avons une relation de causalité ou de conséquence

(« N_1 *a pour cause* N_2 ») : *éruption maculeuse*, *angine tuberculeuse*. Dans d'autres cas, le sens générique « N_1 *relatif à* N_2 » et la relation d'appartenance semblent être suffisants : *antécédent tuberculeux*, *foyer tuberculeux*.

- Si N_2 est suffixé en *-ique* et N_1 est une pathologie, nous avons une relation de causalité (« N_1 *causé par* N_2 », « N_1 *avec le développement de* N_2 ») : *adénomatose kystique*, *cancer kystique*, *hémorragie kystique*. Si N_1 est une partie du corps nous avons une relation de localisation à travers l'appartenance « N_2 *localisé dans* N_1 » (*ovaire kystique*, *rein kystique*) ou la possession « N_1 *localisé dans* N_2 » (*cellule kystique*, *liquide kystique*). Sinon, il s'agit du sens général « N_1 *relatif à* N_2 » : *aspect kystique*, *cartographie kystique*.
- Lorsque N_2 est suffixé en *-al* et N_1 est une autre pathologie, le sens de l'expression est « N_1 *relatif au cours de* N_2 », « N_1 *causé par* N_2 » : *complication grippale*, *otite grippale*. Lorsque N_1 est un produit chimique, essentiellement un médicament dans les exemples étudiés, le sens est « N_1 *utilisé contre* N_2 » : *médicament antigrippal*, *vaccin grippal*. Dans d'autres cas, le sens général « N_1 *relatif à* N_2 » est suffisant : *saison grippale*, *variant grippal*.
- Et enfin, lorsque N_2 est suffixé en *-aire* et N_1 est une autre pathologie, nous avons une relation de causalité (« N_1 *dû à* N_2 », « N_1 *lié à* N_2 ») : *douleur herniaire*, *gastrite herniaire*, *hémorragie maculaire*, *œdème maculaire*. Lorsque N_1 est une partie du corps, il s'agit de la localisation à travers l'appartenance (« N_1 *est formé de* N_2 », « N_1 *contient* N_2 ») : *orifice herniaire*, *sac herniaire*, *aire maculaire*. Dans d'autres cas, la relation générale est suffisante (« N_1 *relatif à* N_2 ») : *chirurgie herniaire*, *laser maculaire*.

Lorsque N_2 est une partie du corps N_1 peut appartenir à de nombreux axes sémantiques. Dans les différentes configurations, la relation privilégiée est la localisation : la partie du corps sert à localiser une pathologie, une autre partie du corps, un acte médical, etc.

- Lorsque N_2 est suffixé en *-al*, la relation reflète une localisation. N_1 peut alors être une pathologie (« N_1 *se trouve dans* N_2 ») : *arthrose vertébrale*, *fracture vertébrale* ; un acte médical ou les soins (« N_1 *est localisé dans* N_2 », « N_1 *est pratiqué sur* N_2 ») : *angiographie vertébrale*, *biopsie vertébrale* ; ou une autre partie du corps. Dans ce dernier cas, nous trouvons la possession (« N_1 *se trouve dans* N_2 ») : *artère vertébrale*, *disque vertébral* ; ou l'appartenance (« N_2 *se trouve dans* N_1 ») : *colonne vertébrale*. Dans d'autres cas, la relation générale est suffisante : *jet urétéral*, *flux vertébral*.
- Lorsque N_2 est suffixé en *-é* et N_1 est un être animé ou une de ses parties du corps, les relations sont « N_1 *pourvu de* N_2 » (*animal vertébré*, *hôte vertébré*), « N_1 *localisé dans* N_2 » (*amygdale pharyngée*, *muqueuse pharyngée*) ou « N_1 *en forme de* N_2 » (*masse lobulée*), avec une idée de saturation et de localisation. Si N_1 est une pathologie, les relations peuvent également être diverses : « N_1 *en forme de* N_2 » (*tumeur lobulée*) ou « N_1 *situé dans* N_2 » (*douleur pharyngée*, *œdème pharyngé*). Si N_1 représente un acte médical ou les soins, nous avons une relation de localisation (« N_1 *est pratiqué sur* N_2 ») : *anesthésie pharyngée*, *désinfection pharyngée*. Dans d'autres cas, nous avons encore la relation générale : *aspiration pharyngée*, *lumière pharyngée*.

- Lorsque N_2 est suffixé en *-aire* il sert à localiser N_1. Celui-ci peut alors être une partie du corps : « N_1 *se trouve dans* N_2 » (*artère lobulaire*, *paroi musculaire*), « N_2 *est composé de* N_1 » (*masse musculaire*) ; une pathologie (*lésion lobulaire*, *dystrophie musculaire*, *éruption vésiculaire*) ; un acte médical ou des soins (*biopsie musculaire*). Dans le reste des cas, l'adjectif indique une relation générale : *apparence lobulaire*, *tonus musculaire*, *origine vésiculaire*.
- Lorsque N_2 est suffixé en *-eux*, la relation spécifique privilégiée est encore la localisation. N_1 peut à son tour désigner une partie du corps (*bandelette musculeuse*, *couche musculeuse*) ; ou une pathologie (*infiltration musculeuse*, *angine vésiculeuse*, *pharyngite vésiculeuse*). Sinon, la relation est de type général (*poussée vésiculeuse*).
- Et enfin, lorsque N_2 est suffixé en *-ien*, nous avons toujours la relation de localisation (« N_1 *localisé dans* N_2 »). N_1 peut désigner alors une pathologie : *sténose pharyngienne*, *zona pharyngien* ; ou une partie du corps : *membrane pharyngienne*.

Bilan. Nous avons présenté une analyse d'adjectivations concurrentes observées sur des corpus médicaux spécialisés et de langue générale. Les observations sont basées sur les écrits de morphologues, mais sont également inspirées par les travaux en analyse morphosémantique des composés savants (sec. 3.2.3 page 57). Une comparaison plus poussée des composés savants et des constructions syntaxiques reste à faire. Il reste également à établir un protocole plus strict et complet pour l'analyse des adjectivations concurrentes.

Nous avons analysé les suffixes d'adjectivation dans les contextes d'occurrence des adjectifs qu'ils forment : leur apport sémantique est alors un peu plus évident à détecter. Mais l'instruction de l'affixe est intimement liée avec le sens hérité de la règle et le sens de la base nominale. De manière générale, les gloses sémantiques générales donnent une indication quant au sens de l'affixe et de l'expression mais peuvent être plus précises dans un domaine de spécialité comme la médecine. L'analyse est alors appuyée par le typage sémantique du nom de base de l'adjectif et du nom recteur du syntagme nominal. Nous avons ainsi noté quelques constantes : les adjectifs en *-é* traduisent souvent une idée de saturation (Aliquot-Suengas, 1996) ; les adjectifs formés sur des bases nominales désignant une partie du corps servent principalement à localiser les pathologies, les actes médicaux, etc. ; les adjectifs formés sur des bases nominales désignant des pathologies induisent une relation de causalité. Cette information sémantique à des intérêts applicatifs, par exemple en structuration de termes avec des relations transversales (Grabar & Hamon, 2004b).

5.8 Conclusion et perspectives

Cette étude de la productivité des suffixes d'adjectivation dénominale doit être considérée comme une ébauche de comparaison à travers les domaines et genres. Les résultats qu'elle fait ressortir sont partiels et demandent à être assis plus solidement par d'autres études statistiques et par une étude morphologique parallèle. Elle peut également être étendue à d'autres procédés morphologiques et d'autres corpus. Sur de petits corpus, une estimation d'intervalle de confiance pourrait être associée aux mesures appliquées. Des

résultats plus complets pourront également être obtenus si nous complétons la base de règles et de noms par exemple avec des méthodes basées sur corpus (Zweigenbaum *et al.*, 2003b). Le cas de figure où les valeurs $\mathcal{P}$ et V sont croisées ($\mathcal{P}_{S_1} > \mathcal{P}_{S_2}$ et $V_{S_1} < V_{S_2}$) devrait aussi être étudié un peu mieux, d'autant plus qu'il concerne un certain nombre de suffixes. Cette étude apporterait une vue supplémentaire sur la totalité des suffixes étudiés. Des expériences complémentaires devraient donc mieux cerner la variation intra-domaniale et intra-genre.

L'étude des adjectivations concurrentes montre que l'apport sémantique des affixes va souvent au-delà des gloses sémantiques générales. Ces gloses peuvent être spécifiées dans un domaine de spécialité comme la médecine. Dans les analyses effectuées, elles dépendent des suffixes mais aussi des types sémantiques des noms N_1 et N_2 : les adjectifs en *-é* traduisent souvent une idée de saturation (Aliquot-Suengas, 1996) ; les adjectifs formés sur des bases nominales désignant une partie du corps servent principalement à localiser les pathologies, les actes médicaux, etc. ; les adjectifs formés sur des bases nominales désignant des pathologies induisent une relation de causalité. De manière générale, les calculs statistiques appliqués au matériau linguistique devraient être appuyés par une analyse linguistique parallèle : il est difficile de retrouver toute la palette de critères qui influent sur les formations morphologiques à travers une formule mathématique.

Nous avons étudié la productivité des opérateurs d'adjectivation dénominale en analysant des corpus différenciés par des critères socio-linguistiques comme le sous-domaine ou le genre. Une démarche inverse peut également être entreprise : se baser sur les procédés morphologiques pour catégoriser les documents en catégories homogènes répondant aux notions de genre, sous-domaine ou autre. Des études montrent en effet que la morphologie (flexionnelle et dérivationnelle), à côté du lexique, de la ponctuation, de l'information syntaxique ou autre, constitue une source de critères pour le typage des documents (Baayen, 1994; Rillof, 1995; Habert *et al.*, 2000; Arppe, 2002; Valette & Grabar, 2004). Ainsi, contrairement à de nombreuses applications où l'on cherche à neutraliser la variation terminologique, le profilage des documents semble tirer profit de la morphologie, en offrant un 'gain d'information' supplémentaire.

Chapitre 6

Appariement des requêtes des utilisateurs avec le vocabulaire contrôlé. Étude des logs d'un portail médical français

L'expérience que nous présentons dans ce chapitre se situe du côté de la recherche d'information. Elle consiste en un appariement des expressions réelles des utilisateurs, qui proviennent dans nos expériences des logs du portail médical CISMeF, et d'un vocabulaire contrôlé, qui provient lui du thesaurus médical MeSH utilisé dans CISMeF pour l'indexation des documents[1].

6.1 Introduction

Une des vocations premières du TAL et de la linguistique de corpus consiste à faire face à une masse d'informations sans cesse croissante, à laquelle nous avons accès grâce à de nombreuses bases de données et à l'Internet. Dans le domaine médical, qu'il s'agisse des professionnels de la santé ou du grand public, les informations sont tout aussi abondantes : (1) les bases de données bibliographiques comme Medline[2] offrent un accès à la littérature scientifique ; (2) les sites thématiques ou les portails médicaux généralistes (CISMeF[3] ou HON[4]) offrent des documents de diverses natures. L'approche classique en recherche d'informations consiste à indexer les documents avec des mots clés qui leur sont spécifiques. L'utilisateur doit alors les 'deviner' pour trouver les documents en question. Pour aider les utilisateurs dans cette recherche, les documents peuvent être indexés à l'aide d'un *voca-*

[1] Dans la présentation de cette expérience, nous reprenons nos publications antérieures, par exemple (Grabar *et al.*, 2002) et (Grabar *et al.*, 2003).

[2] `www.ncbi.nlm.nih.gov/pubmed`

[3] `www.chu-rouen.fr/cismef`.

[4] Health on the Net : `www.hon.ch`.

bulaire contrôlé et donc connu. Ainsi, Medline, CISMeF et HON, utilisent des mots clés du thesaurus MeSH. Nous pouvons noter plusieurs avantages de l'emploi d'un thesaurus comme le MeSH. (1) Avec sa structure hiérarchique, il permet d'effectuer l'*explosion de concept*. Ainsi, en interrogeant la base avec une notion générique (par exemple, *maladies cardiaques*), l'utilisateur peut retrouver les documents indexés par ses termes plus spécifiques (par exemple, *ischémies myocardiques*) et par là même augmenter le rappel. (2) Il est régulièrement maintenu (par la NLM) et de nouveaux concepts biomédicaux y sont ajoutés. (3) Étant un thesaurus 'officiel', il est connu des utilisateurs mais surtout des documentalistes. Cependant, on ne peut pas attendre de tous les utilisateurs qu'ils maîtrisent les termes d'un vocabulaire contrôlé. Le MeSH compte environ 20 000 descripteurs et, malgré l'ajout de 8 000 synonymes, un utilisateur non spécialiste a peu de chances d'employer d'emblée le terme canonique désignant ce qu'il cherche.

À travers les portails médicaux comme Medline ou CISMeF, les utilisateurs interrogent les bases sous forme de requêtes libres. Ces requêtes sont ensuite mises en correspondance avec les termes d'indexation. La variation terminologique apparaît donc comme un obstacle à l'appariement des requêtes des utilisateurs avec les termes de l'indexation. Dans l'expérience que nous présentons, nous nous proposons d'étudier l'apport de connaissances morphologiques dans l'appariement lexical des variantes terminologiques. Ce travail ne concerne pas directement le fonctionnement en ligne de l'outil d'interrogation : il est basé sur l'étude des logs du portail médical et définit ainsi les conditions de faisabilité de la mise en correspondance des termes et des requêtes. Nous cherchons ainsi à savoir dans quelle mesure les requêtes libres employées par les utilisateurs comportent un vocabulaire comparable à celui des termes cible. Et si les écarts observés peuvent être réduits par l'injection de connaissances linguistiques.

Dans la suite de ce chapitre, nous présentons des travaux en analyse des logs (sec. 6.2), mais notons tout de suite que notre expérience se rapproche surtout des travaux en traitement de la variation terminologique par l'appariement lexical. Nous décrivons ensuite notre matériel : les requêtes, la terminologie cible et les ressources morphologiques (sec. 6.4). Nous expliquons la méthode suivie pour comparer et normaliser le vocabulaire des requêtes et le vocabulaire cible (sec. 6.3). Nous détaillons et discutons ensuite les résultats de ces comparaisons (sec. 6.5). Et enfin, nous concluons (sec. 6.6).

6.2 Travaux en analyse des logs

Un des principaux soucis des systèmes de recherche d'information concerne l'évaluation de la recherche et de l'interface et leur amélioration en fonction des attentes des utilisateurs (Jones *et al.*, 2000). Mais comme la recherche d'information sur le Web est anonyme, les systèmes de recherche savent peu de choses sur leurs utilisateurs. Pour combler ce vide, un des moyens d'observer les utilisateurs et leur interaction avec le système passe par l'étude des logs. Un *log* est un fichier qui garde les traces des connexions au serveur : adresse Internet (IP) des machines, mots clés des utilisateurs, durée des connexions, etc. L'analyse des logs, appelée aussi *transaction log analysis*, permet ainsi de collecter des informations

sur le comportement des utilisateurs et de tirer les conclusions utiles à l'amélioration de l'interface, du fonctionnement du système, etc.

Nous présentons les études réalisées autour de deux moteurs de recherche généraux, `Excite` (Jansen *et al.*, 2000) et `AltaVista` (Silverstein *et al.*, 1998), et un moteur de recherche spécialisé de la librairie électronique `NZDL` (New Zealand Digital Library) consacrée à l'informatique et au multimédia (Jones *et al.*, 1998). Le premier travail de ce type a été réalisé par (Jansen *et al.*, 1998) sur les logs d'`Excite`. Bien que les travaux qui ont suivi s'en soient inspirés, il existe peu de critères en commun.

La taille des logs analysés est très variable : elle est d'environ 30 000 requêtes sur le portail spécialisé `NZDL`, d'environ 51 000 avec le moteur général `Excite`, et pas loin d'un milliard avec `AltaVista`. Le nombre d'utilisateurs avoisine les 18 000 avec `Excite` et 285 millions avec `AltaVista`. L'analyse des logs est concentrée autour de trois pôles :

- les *sessions* : temps de connexion, nombre moyen de requête par connexion, IP de la machine, nombre moyen de pages consultées, etc.,
- les *requêtes* : nombre moyen de mots par requête, utilisation d'opérateurs booléens et de modifieurs, etc.,
- les *termes* : casse, présence de nombres et symboles, typage sémantique, etc.

Le nombre moyen de requêtes par session correspond au temps passé avec le moteur. Les utilisateurs d'`Excite` en soumettent en moyenne trois, et ceux d'`AltaVista` et de `NZDL` deux. Une fois la requête soumise, les utilisateurs des moteurs généraux vont rarement au-delà de la 2^{e} page de résultats, et 58 % des utilisateurs d'`Excite` ne regardent que la 1ère page. Par contre sur le moteur thématique `NZDL` l'assiduité des utilisateurs est plus grande, avec plus de 87 % qui vont au-delà de la 1ère page : ils sont prêts à creuser considérant sans doute qu'ils peuvent trouver l'information recherchée. Le fait que les utilisateurs se bornent à consulter uniquement les premières pages de résultats demande de proposer des systèmes avec une précision élevée (Jansen *et al.*, 2000).

Lors de la composition des requêtes, les utilisateurs font aussi peu d'effort. L'usage des opérateurs booléens (`ET`, `OU`, `NON`, etc.) et des modifieurs (`NEAR`) est très faible. `Excite` constate en plus que lorsqu'ils sont utilisés, 20 % d'erreurs de syntaxe sont vite atteints, ce qui détériore les réponses du système. Sur `NZDL`, 73 % des utilisateurs se contentent du paramétrage par défaut : les informations sur ce paramétrage doivent donc être visibles et explicitement signalées en cas de modification. En ce qui concerne le nombre moyen de mots par requête, le moteur thématique `NZDL` présente les valeurs les plus élevées : 2,43 contre 2,3 avec les moteurs généraux. Les utilisateurs d'un moteur thématique semblent ainsi pouvoir mieux cibler leurs demandes. Mais la différence avec les bases documentaires classiques, où les requêtes comprennent entre 7 et 15 mots, reste grande.

En analysant les termes des logs d'`Excite`, (Jansen *et al.*, 2000) ont typé sémantiquement les mots des requêtes les plus fréquentes, qui représentent pourtant rarement plus d'un % de la totalité des requêtes. Ces mots concernent le sexe, la géographie et les voyages, l'économie et la recherche de travail, les images, etc. Par ailleurs, l'application de la loi de Zipf montre la particularité du lexique utilisé dans les documents du Web (Jansen *et al.*, 2000) : la courbe est allongée aux extrémités où se trouvent les mots très fréquents et les hapax. (Jones *et al.*, 1998) considèrent que les mots les plus fréquents sont exploitables lors

de la constitution des index, ce qui permettrait de fournir plus rapidement les réponses aux mots clés les plus fréquents. L'analyse des termes de NZDL montre en plus que parmi les requêtes soumises, 5,61 % des mots ne trouvent pas de réponses. Ces mots correspondent aux erreurs d'orthographe (*repesentation*, *sofhware*, *algorythm*), aux noms propres et noms de marque (*geobase*, *gillingham*) et aux termes techniques non répertoriés (*jsort*, *octrie*, *echocardiology*).

En résumé, il existe, dans ces expériences, une différence nette dans le comportement des utilisateurs des moteurs de recherche généraux et spécialisés. Les moteurs de recherche spécialisés reçoivent des requêtes plus développées et présentent une consultation des résultats plus fouillée. La thématique ciblée et peut-être l'accoutumance des utilisateurs et leur connaissance de la base sont la cause de cette différence. Mais (Jones *et al.*, 1998) constatent néanmoins une fidélité modérée des utilisateurs de NZDL, qui correspond à la 'règle 80/20' : 80 % des utilisateurs ne font qu'une seule visite sur le site et 20 % restent fidèles. De manière générale, les auteurs concluent que la recherche d'information sur le Web est bien différente de la recherche documentaire classique. Le peu d'effort manifesté par les utilisateurs doit être compensé par l'adaptation de l'interface et la précision des systèmes de recherche.

6.3 Méthode d'appariement des requêtes des utilisateurs et des termes d'indexation

Les travaux en analyse des logs que nous venons de présenter sont concentrés sur une étude statistique des informations (nombre moyen de requêtes par session, de mots dans une requête, etc.). L'étude lexicale des requêtes est rarement effectuée et reste au niveau des constatations. L'expérience que nous proposons consiste également en une analyse des logs. Mais nous appliquons aux requêtes des traitements de TAL dans le but d'observer l'apport des connaissances linguistiques, et surtout morphologiques, lors de l'appariement du vocabulaire des utilisateurs avec le vocabulaire contrôlé.

La méthode mise en œuvre consiste à segmenter en mots les termes source (requêtes des utilisateurs) et cible (thesaurus d'indexation, le MeSH), et ensuite à comparer les vocabulaires résultants après des normalisations successives. Tous les traitements sont appliqués parallèlement au vocabulaire source et au vocabulaire cible.

La segmentation en mots est faite sur les caractères non alphanumériques (espaces, ponctuations et autres caractères). Le résultat de cette étape est une liste de mots source, et une liste de mots cible. Ces listes permettent de faire une première évaluation du recouvrement des vocabulaires : on recense ainsi l'ensemble des mots source qui se trouvent dans la cible (les mots communs aux deux listes) et son complémentaire, les mots source 'inconnus' des mots cible. La suite des traitements est effectuée à chaque fois sur les mots source restant inconnus.

Les deux vocabulaires (mots sources restant inconnus et mots cible) sont soumis à une série de normalisations réalisées au niveau des caractères (sec. 6.3.1) et à des normalisations

qui demandent un apport de connaissances linguistiques (sec. 6.3.2). Les mots 'vides' (mots considérés comme non pertinents dans les vocabulaires comparés) sont supprimés des deux vocabulaires. Les normalisations appliquées permettent d'observer l'apport en termes de rappel du moteur de recherche qui reçoit ces requêtes, à travers le nombre d'appariements effectués. Par contre, nous n'avons aucun contrôle sur l'évolution de la précision.

6.3.1 Normalisations au niveau des caractères

Les normalisations au niveau des caractères ont été conditionnées par la nature du MeSH (mots en majuscules non accentuées, hormis les termes en minuscules ajoutés par l'équipe CISMeF), et par la nature non prévisible des mots employés dans les requêtes des utilisateurs de CISMeF (mots en minuscules ou en majuscules, accentués ou non). Nous effectuons deux types de traitement à ce niveau : la minusculisation et la désaccentuation. Ce type de normalisation est élémentaire, mais important du fait des différences signalées.

La suppression des accents (par exemple, chaque occurrence des lettres *éèêë* est convertie en *e*) est nécessaire pour homogénéiser l'utilisation des caractères accentués, mais également pour éliminer les mots accentués incorrectement (par exemple, *athlètisme*). Cette transformation peut effacer les différences de signification entre certains mots. Par exemple, les mots *sténose* et *sténosé* sont réduits à une seule forme : *stenose*. Vu l'ambiguïté que ce traitement apporte, nous avons initialement testé son application en fin de chaîne des traitements pour ne pas déteriorer la précision de l'appariement. Mais nous avons alors remarqué que l'absence ou bien la différence d'accentuation bloquaient dans beaucoup de cas l'application des connaissances linguistiques. Nous avons donc finalement testé son application en début de chaîne des traitements (après la minusculisation). Les connaissances linguistiques, qui sont appliquées par la suite, sont également désaccentuées.

6.3.2 Normalisations au niveau des mots : application de connaissances morphologiques

Les traitements au niveau des mots consistent en des normalisations morphologiques, effectuées avec des connaissances linguistiques et des règles de désuffixation des pluriels réguliers. Les ressources morphologiques couvrent la flexion et l'affixation, qui correspondent chacune à une étape séparée dans les traitements. Les règles de désuffixation prennent en compte uniquement certaines variations flexionnelles. Ces traitements sont appliqués séparément ou bien combinés, ce qui donne trois modules :

- Application des ressources morphologiques ;
- Application des règles de désuffixation des pluriels réguliers ;
- Combinaison des ressources morphologiques et des règles de désuffixation.

Application de ressources morphologiques. L'application des ressources morphologiques (sec. 6.4.3) est faite en deux étapes : lemmatisation et racinisation. La lemmatisation réduit les formes fléchies d'un mot à leur forme canonique. La racinisation se place du côté

de la morphologie dérivationnelle et permet par exemple de réduire les adjectifs aux noms à partir desquels ils ont été formés. Il s'agit de réunir les mots qui ont le même noyau morphologique. La racinisation est appliquée aux mots du vocabulaire cible et sur les listes de mots qui restent inconnus à la sortie de la lemmatisation.

Application des règles de désuffixation des pluriels réguliers. Ces règles sont basées sur une heuristique de suppression de la marque du pluriel utilisée en recherche d'information, où elles montrent de bons résultats (Savoy, 2002). Elles consistent à supprimer les finales en *-s*, à réduire les *-aux* en *-al* et à supprimer les finales en *-x*.

Combinaison de ressources morphologiques et de règles. La combinaison de connaissances flexionnelles et de règles de désuffixation permet de viser la précision (par l'application de connaissances flexionnelles) et d'augmenter le rappel (avec les règles de réduction des pluriels). Les connaissances linguistiques sont appliquées d'abord, ensuite, sur les mots non traités par ces connaissances, nous appliquons les règles.

6.3.3 Correction orthographique : comparaison approximative de mots

Les mots qui restent inconnus après les normalisations précédentes sont peut-être des mots du MeSH mal orthographiés par les utilisateurs, situation fréquente dans les travaux similaires. (Jones *et al.*, 1998; Hawking *et al.*, 2000) notent ainsi jusqu'à 13 % de requêtes mal orthographiées. Nous avons donc tenté une *correction orthographique* sur ces mots avec l'outil `ispell` d'Unix et le vocabulaire cible comme dictionnaire de référence. Cependant, comme une correction orthographique est trop aléatoire sur les mots courts, nous avons exclu de cette correction les mots de longueur inférieure à cinq lettres. Dans le but d'une corretion entièrement automatique, nous avons également exclu les propositions de correction multiples pour un même mot.

6.3.4 Enchaînement des normalisations

L'enchaînement des normalisations est effectué dans l'ordre suivant : d'abord les normalisations au niveau des caractères (minusculisation et désaccentuation) et la suppression des mots vides, ensuite les trois modules de normalisations morphologiques appliqués en parallèle et, enfin, la correction orthographique. Dans cette chaîne, seule la désaccentuation avait une place flottante (car étant potentiellement une cause de perte de précision). Elle est finalement effectuée après la minusculisation.

6.3.5 Occurrences et types

A chaque étape des traitements, nous calculons les appariements et les mots restant à traiter. Nous le faisons en terme d'occurrences et de types. Si le nombre d'occurrences

reste le même tout au long des traitements (sauf lorsque les mots vides sont supprimés), le nombre de types diminue au fil des normalisations (par exemple, plusieurs formes fléchies pouvent être réduits à un seul lemme). D'autre part, comme nous le verrons lors de la présentation des résultats d'appariement (fig. 6.1), ils ne sont pas parallèles pour les occurrences et les types.

6.4 Description du matériel

Le matériel utilisé comprend (1) les requêtes reçues par le moteur de recherche CISMeF de septembre 2000 à janvier 2001, (2) le thesaurus d'indexation MeSH utilisé dans CISMeF pour l'indexation des documents Web, (3) des connaissances morphologiques et règles de désuffixation et (4) une liste de mots vides.

6.4.1 Log des requêtes à CISMeF

Les requêtes analysées du portail de sites médicaux francophones CISMeF ont été reçues par son moteur de recherche peu après son ouverture, entre septembre 2000 et janvier 2001. Ce sont des requêtes simples, sans utilisation d'opérateurs logiques ni de modifieurs, d'où nous avons éliminé les requêtes effectuées par l'équipe CISMeF (sur la base de l'adresse de la machine source) et les requêtes vides. Nous obtenons ainsi 108 660 requêtes (29 092 requêtes uniques). Nous avons cherché à tenir compte du fait que certaines requêtes sont plus fréquentes que d'autres. Pour cela, nous avons compté pour chaque requête le nombre de machines différentes d'où elle a été envoyée. Cela donne 76 341 occurrences de *requêtes.machines* (requêtes différentes émises à partir de machines différentes). C'est sur cet ensemble de requêtes que nous travaillons. Il présente 131 570 occurrences et 21 112 types de mots. Les requêtes reçues par CISMeF sont composées en caractères minuscules ou bien majuscules, accentués ou bien non accentués.

6.4.2 Vocabulaire d'indexation : le MeSH

À l'époque de l'expérience, CISMeF utilise la traduction française du MeSH 2001 pour l'indexation des documents. Elle comporte 19 971 termes, 9 151 synonymes et 83 qualificatifs. Sur les 19 971 termes français du MeSH, 5 336 étaient employés dans CISMeF au moment de l'expérience, soit 23 % de la totalité disponible en français. Ces termes sont enrichis par l'équipe de CISMeF de 38 'métatermes' et 101 'types de ressources' (Darmoni *et al.*, 2001)[5]. Les métatermes correspondent aux spécialités médicales (dermatologie, néphrologie, etc.) ; les types de ressources donnent une indication sur le genre des documents recensés (support de cours, article scientifique, etc.). L'ensemble de termes que nous avons pris pour cette expérience comporte au total 29 035 termes différents (termes +

[5] À titre d'information, en octobre 2004, CISMeF utilise presque la moitié des descripteurs MeSH traduits en français : 10 638 sur 22 568, mais aussi 84 qualificatifs, 193 types de ressources et 105 métatermes. Nous remercions Magaly Douyère de l'équipe CISMeF pour ces informations.

synonymes - doublons). Ce qui constitue le vocabulaire cible : 58 912 occurrences et 21 475 types de mots. Les termes français du MeSH sont écrits en majuscules non accentuées. Les termes employés dans CISMeF ont été mis en casse mixte (minuscules, emploi 'normal' des majuscules) et pour certains réaccentués.

6.4.3 Normalisations morphologiques

Ressources morphologiques linguistiquement motivées. Les connaissances morphologiques que nous avons induites à partir de terminologies médicales (chap. 4) ont été mises en application ici pour lemmatiser et raciniser les mots. Elles ont été validées manuellement puis complétées à partir de dictionnaires généraux[6] et de différents corpus médicaux étiquetés. Ces ressources comportent 308 847 couples {*forme*, *lemme*} (par exemple, {*abdominaux*, *abdominal*}) pour la lemmatisation et 1 041 couples {*base*, *dérivé*} (par exemple, {*abdominal*, *abdomen*}) pour la racinisation. La flexion verbale occupe une grande place dans ces couples flexionnelles, mais est en principe peu utile ici.

Règles de désuffixation des pluriels réguliers. Nous avons implémenté des règles de traitement des pluriels réguliers. De nombreux travaux en recherche d'information (par exemple (Savoy, 2002)) ont montré l'efficacité de telles règles pour l'augmentation du rappel. Ce sont des règles très générales de réduction des pluriels :

- passage d'un pluriel en *-s* à un singulier : *sténoses* → *sténose*;
- réduction des *-aux* en *-al* : *abdominaux* → *abdominal*;
- suppression des *-x* finals qui restent : *genoux* → *genou*.

Par contre, comme nous n'utilisons pas de liste d'exceptions, les règles seront également appliquées aux formes non plurielles : *AIDS*, *anis*, *anticorps*, *Bordeaux*, *pharynx*. Dans la mesure où ces 'normalisations' sont appliquées également au vocabulaire source, elles ne sont pas très gênantes.

6.4.4 Mots vides

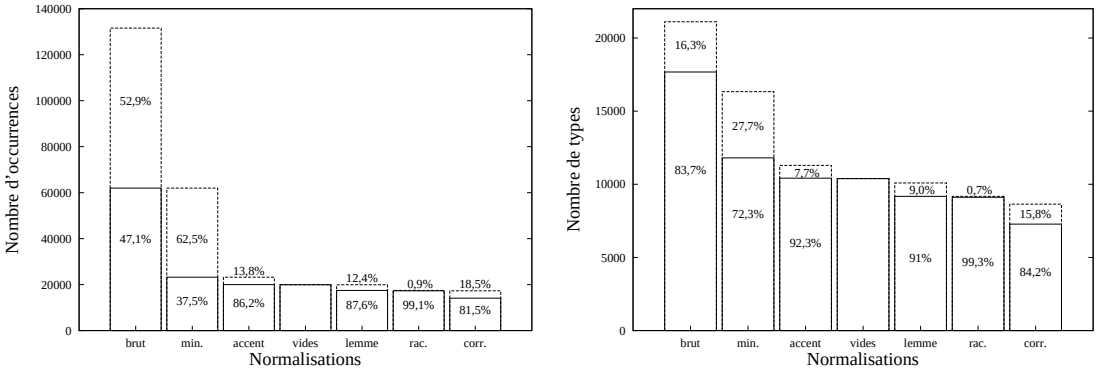
Nous avons établi une liste de mots vides dans une optique de recherche d'information dans le domaine médical (Zweigenbaum *et al.*, 2002). Elle comprend des mots grammaticaux (articles, prépositions, conjonctions, pronoms), des numéraux, certains adverbes et adjectifs et quelques formes verbales, pour un total de 199 formes. Nous avons également utilisé une liste de 217 mots vides mise au point pour la recherche d'information (Savoy, 2002)[7]. La fusion de ces deux listes a donné au total 344 mots vides.

[6] Lexique de l'ABU : `abu.cnam.fr/DICO`

[7] Cette liste est disponible à l'adresse `www.unine.ch/info/clef`.

6.5 Analyse et discussion des appariements des vocabulaires

L'étude de la faisabilité d'appariement entre les mots des requêtes des logs et ceux du vocabulaire contrôlé correspond à une étude et un traitement de la variation terminologique, ce qui différencie notre travail de la 'transaction log analysis' traditionnelle. Notre expérience consiste donc en des normalisations de différents niveaux appliquées progressivement et en parallèle aux vocabulaires source et cible.



(a) Occurrences

(b) Types

FIG. 6.1 – Évolution des appariements à chaque étape de la normalisation.

La figure 6.1 présente les résultats de ces appariements pour l'ensemble de la période allant de septembre 2000 à janvier 2001 en termes d'occurrences (fig. 6.1(a)) et de types (fig. 6.1(b)). Chaque colonne correspond à une étape : *brut* appariement direct, *min.* minusculisation, *accent.* désaccentuation, *vides* suppression des mots vides, *lemme* lemmatisation, *rac.* racinisation et *corr.* correction orthographique. Pour l'ensemble de mots source restant à traiter à chaque étape, nous indiquons le pourcentage de mots appariés (partie haute) et le pourcentage de mots restant à traiter (partie basse). Les mots qui restent non appariés à l'étape n correspondent à l'ensemble traité à l'étape $n + 1$.

Le vocabulaire cible (MeSH) comporte 58 912 occurrences et 21 475 types. Le vocabulaire source (requêtes) comporte 21 112 types et 131 570 occurrences (colonne *brut*). Parmi ces mots, 3 438 types et 69 602 occurrences sont appariés directement avec les mots du MeSH. 17 674 types et 61 968 occurrences restent inconnus. L'appariement des mots originaux, sans aucun traitement, permet donc de reconnaître 52,9 % d'occurrences, mais seulement 16,3 % de types. La minusculisation apparie en plus 62,5 % d'occurrences (27,7 % de types) dans l'ensemble du vocabulaire source qui reste. 23 217 occurrences (11 806 types) restent inconnues. La désaccentuation permet d'apparier encore 13,8 % d'occurrences (7,7% de types). À cette étape, 20 004 occurrences (10 420 types) restent

inconnues. L'élimination de 33 mots vides correspond à la suppression de 85 occurrences. Ce traitement ne produit pas de nouveaux appariements. Les meilleurs résultats de normalisation flexionnelle sont obtenus avec la combinaison de connaissances linguistiques et de règles de désuffixation. Ce sont donc ces résultats que nous présentons ici. La lemmatisation apparie 12,4 % d'occurrences (9% de types) et la racinisation 0,9% d'occurrences (0,7% de types) en plus.

La comparaison directe des vocabulaires permet donc de traiter 52,9 % des occurrences, ce qui est plutôt élevé. Par contre seulement 16,3 % des types sont reconnus. Ce qui signifie que sans traitements supplémentaires plus de 85 % de types restent non reconnus et que les utilisateurs obtiendraient donc des résultats incomplets et d'une qualité moyenne. Mais sachant que la majorité des termes français du MeSH sont écrits en majuscules non accentuées et que les utilisateurs écrivent avec une casse et une accentuation variables, les résultats de ce premier appariement peuvent être considérés comme élevés. Cela peut provenir de plusieurs origines : (1) un grand nombre d'utilisateurs de CISMeF sont des documentalistes, qui connaissent le MeSH (...en anglais) ; (2) un grand nombre de pathologies courantes sont des mots clés MeSH, utilisés par CISMeF et parfois écrits en minuscules accentuées. Par exemple, nous trouvons parmi les vingt requêtes les plus fréquentes en janvier 2002 les mots : *euthanasie, anorexie, alcoolisme, hepatite/hépatite, pneumonie, sida, varicelle, paludisme* ; (3) enfin, certains sites proposent des liens consistant en des requêtes toutes prêtes à CISMeF : ces requêtes pointent systématiquement sur les bons mots clés.

L'utilisation des normalisations au niveau des caractères et des normalisations morphologiques amène à un taux final de reconnaissance de l'ordre de 87% d'occurrences et 66 % de types de mots. Moins d'un septième des occurrences du vocabulaire des utilisateurs n'est alors pas interprétable. La correction orthographique propose de corriger environ 15,8 % de types (18,5 % d'occurrences). Finalement, dans l'ensemble initial de mots source, 65,5 % de types et 89,3% d'occurrences sont reconnus avec les normalisations appliquées. Notons néanmoins que dans les cas où les mots clés des utilisateurs ne sont pas appariés avec les termes du MeSH, le moteur de CISMeF dispose d'un recours : il génère une recherche sur le contenu intégral des notices puis des documents.

Parmi les normalisations appliquées, la normalisation de la casse apporte le plus gros gain (29,5 % d'occurrences et 21,4 % de types par rapport au vocabulaire source original), la désaccentuation ajoute 2,4 % d'occurrences et 4,1 % de types aux mots originaux reconnus. Au total, 84,8 % d'occurrences et 41,8 % de types du vocabulaire source sont reconnus. Étant donné la simplicité de ces méthodes, les normalisations de caractères devraient être considérées comme la base en recherche d'information dans CISMeF. La contribution totale des connaissances morphologiques est beaucoup plus faible : 2 % pour les occurrences et 4,6 % pour les types (ce qui donne 86,8 % pour les occurrences et 46,4 % pour les types par rapport au vocabulaire d'origine). Alors que le recensement de ces ressources demande davantage d'effort. Nous entrons ici dans un schéma classique dans la résolution de problèmes : au-delà d'un certain point, les efforts nécessaires pour améliorer les résultats sont de plus en plus importants pour un gain qui va en diminuant. Remarquons néanmoins que l'apport des connaissances morphologiques est plus élevé pour les types (4,6 %) que pour les occurrences (2 %). Notons également que la couverture de nos ressources morpho-

logiques pour la lemmatisation et surtout la racinisation n'est pas parfaite. Nous l'avons noté : ces ressources font défaut de façon générale pour le français et encore plus pour les domaines de spécialité comme la médicine.

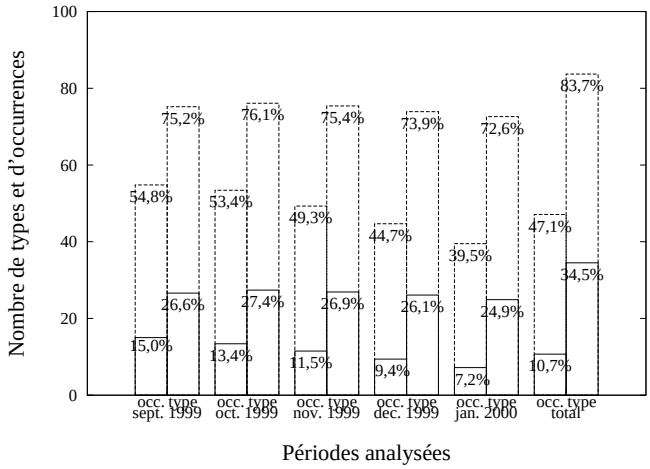


FIG. 6.2 – Évolution mensuelle des occurrences et des types restant non appariés.

La figure 6.2 montre l'évolution mensuelle de mots qui restent inconnus à chaque mois (*sept. 1999*, *oct. 1999*, etc.) de la période analysée. Les décomptes sont également donnés en occurrences (*occ.*) et en types (*type*). Le pourcentage d'occurrences et de types restant à traiter sont indiqués pour les appariements initiaux sans aucune normalisation (partie haute) et avec toute la chaîne des normalisations (partie basse). Nous pouvons voir que le pourcentage des occurrences des mots inconnus diminue d'une manière assez importante d'un mois à l'autre. Par contre cette diminution est très faible pour les types. La dernière colonne (*Total*) correspond aux résultats obtenus lorsque les requêtes de chaque mois sont traitées ensemble. Si le nombre total d'occurrences non appariées correspond à la moyenne des occurrences de tous les mois, le nombre de types non appariés est supérieur à cette moyenne : de nouveaux types inconnus viennent s'ajouter chaque mois.

Les normalisations appliquées, tout en augmentant le rappel, peuvent mener à des ambiguïtés et donc à la diminution de la précision. C'est rarement le cas pour la mise en minuscules, un peu plus pour la désaccentuation (*sténose* et *sténosé* → *stenose*). Quant aux ressources de lemmatisation et de racinisation, elles ont été validées lors de leur génération et sont, en principe, correctes.

Nous avons constaté par ailleurs que les ressources morphologiques motivées linguistiquement et les règles de désuffixation sont complémentaires. En effet, les règles de désuffixation, appliquées aux pluriels réguliers là où nous ne pouvons pas recenser toutes les formes au singulier et au pluriel, permettent d'augmenter le rappel. Il serait intéressant d'appliquer le même type de règles pour le traitement des féminins, voire des affixations. Rappelons néanmoins que le traitement des féminins avec les règles n'est pas évident, surtout sur des données non étiquetées comme dans notre expérience ; la racinisation est également sujette

à la génération de bruit (voir l'analyse des erreurs dans la sec. 4.3.4 page 75). Les règles de désuffixation de telles formations pourraient alors être accompagnées d'exceptions.

Enfin, rappelons que les mesures effectuées concernent le recouvrement des mots individuels. La plupart des requêtes et des termes MeSH comprennent plusieurs mots et leur appariement demande des conditions et des traitements supplémentaires (voir par exemple (Jacquemin, 1996; Zweigenbaum *et al.*, 2002)). De manière générale, les normalisations que nous appliquons permettent d'évaluer l'évolution du rappel, mais pas de la précision. Tandis que les évaluations habituelles des systèmes de recherche visent à mesurer les deux (par exemple, (Gaudinat & Boyer, 2002) pour un autre portail médical).

6.6 Conclusion et perspectives

Les mesures effectuées dans ce travail montrent l'importance de traitements de 'bas niveau' dans l'appariement des requêtes en texte libre avec les termes d'un vocabulaire contrôlé. Elles indiquent aussi l'intérêt mais également les limites des connaissances morphologiques pour aider cette mise en correspondance. Si l'apport de la lemmatisation est net, celui de la racinisation reste extrêmement faible dans cette étude. Il faut toutefois noter la différence quantitative des connaissances mises en jeu : trois cents fois plus importantes pour la lemmatisation que pour la racinisation. Des connaissances plus complètes pour cette dernière pourraient donc modifier ces résultats. L'utilisation combinée des connaissances morphologiques avec les règles de désuffixation des pluriels réguliers améliore le rappel. Du fait qu'il s'agit de la flexion régulière, la précision reste potentiellement élevée. L'établissement de règles pour les féminins, les variations flexionnelles non régulières ou la morphologie constructionnelle demanderait d'appliquer en parallèle des listes d'exceptions.

Une correction orthographique est en revanche un facteur d'amélioration important. Cependant, ses résultats sont moins fiables, et ses conditions d'utilisation restent à préciser.

La proportion de mots non traités par les méthodes étudiées, qui reste importante, témoigne de la nécessité de prendre en compte d'autres méthodes d'appariement et d'accès. Parmi ces autres méthodes, rappelons l'emploi de synonymes (Pouliquen *et al.*, 2002) ou de similarités distributionnelles. Le repli sur un accès classique en texte intégral est une autre option.

Une autre perspective consiste à s'approcher des travaux classiques en *transaction log analysis* (Jones *et al.*, 2000; Jansen *et al.*, 2000; Silverstein *et al.*, 1998), qui permettraient éventuellement d'apporter des amélioration au système de recherche et à l'interface d'interrogation.

Les résultats de cette étude fournissent une mesure utile de l'adéquation entre vocabulaire des utilisateurs et vocabulaire d'indexation ; ils peuvent devenir l'une des métriques servant au suivi régulier de CISMeF. Notons également que les normalisations et les ressources testées ont été depuis intégrées dans le traitement des requêtes en ligne dans CISMeF et ont donné lieu à d'autres travaux (Soualmia & Darmoni, 2004; Névéol, 2004; Névéol *et al.*, 2004).

Chapitre 7

Application de connaissances morphologiques en structuration de terminologies

Dans ce chapitre, nous présentons un autre contexte applicatif où les termes doivent être mis en relation : la structuration des terminologies. Les termes à structurer peuvent être contrôlés ou non contrôlés, selon qu'il s'agit de l'enrichissement d'une terminologie existante ou de la création d'une nouvelle terminologie. Les ressources morphologiques peuvent alors être utilisées, à côté d'autres ressources, pour la normalisation des termes lors de leur structuration.

7.1 Introduction

La structuration d'une terminologie repose sur la détection de *relations* entre les termes d'un domaine. L'ensemble de ces termes et relations forme un *système* dans le but de représenter la connaissance du domaine étudié. À la différence de listes de termes 'à plat', les terminologies structurées permettent à partir d'un terme d'accéder à d'autres termes du domaine qui lui sont liés. Par exemple, en recherche d'information, l'utilisation de la relation de synonymie permet d'enrichir la requête avec des termes équivalents, et l'*explosion de concept*, qui correspond à la prise en compte de termes plus spécifiques que le terme en question, active les relations hiérarchiques. Les terminologies structurées sont relativement rares mais le domaine médical fait figure d'exception, car ses principales terminologies sont structurées. Le besoin de construction de nouvelles terminologies ou bien d'enrichissement ou d'adaptation de terminologies existantes y existe néanmoins. L'approche manuelle de construction et de structuration de terminologies ne peut plus suivre ces besoins. Elle n'est pas non plus adaptée au traitement de la masse d'information existante. Des approches (semi)automatiques sont proposées pour prendre la relève. Les outils informatiques sont alors considérés comme une assistance à la constitution de terminologies structurées ; et les données générées automatiquement, avant d'être utilisées, sont soumises à des 'valideurs'

humains qui prennent la décision finale quant aux termes et relations entre eux. Cette intervention manuelle représente la partie coûteuse dans le processus de construction et de structuration de terminologies. Pour l'alléger, certains chercheurs travaillent avec des ressources 'riches', qui s'opposent alors aux ressources 'pauvres'.

L'acquisition à partir de ressources pauvres est basée sur l'étude des données textuels brutes, dont on fait émerger les termes et les relations entre eux. Cette approche est préconisée par le courant de la terminologie textuelle (Bourigault & Slodzian, 1999), où les documents produits dans un domaine de spécialité sont supposés contenir l'information nécessaire à la création de terminologies de ce domaine. Les terminologies ainsi constituées n'ont pas une vocation 'universelle', mais sont conçues dans le cadre d'une application donnée du domaine exploré. Les travaux qui explorent les ressoures riches sont basés sur les informations contenues dans les dictionnaires généraux ou encyclopédies (par exemple (Calzolari, 1991; Montemagni & Vanderwende, 1992; Agirre *et al.*, 1994)), ou des `DTD` (Document Type Definition) de documents `XML` (Giraldo & Reynaud, 2002) dans les travaux plus récents. Le contenu de ces sources influence directement le contenu des terminologies obtenues. (Grefenstette, 1994) signale quelques biais de ces approches : (1) les mises à jour des dictionnaires sont rares, (2) les termes spécialisés en sont absents et, (3) lorsqu'ils sont répertoriés, leurs sens et relations restent ambigus et souvent non acceptables pour un domaine de spécialité. Quant à l'utilisation des `DTD`, les arbres et les noms d'éléments sous-entendent déjà une structuration du domaine spécifique à un site Web et varient d'un site à l'autre (Giraldo & Reynaud, 2002). Dans d'autres travaux, toujours dans le but d'accélerer le processus, la constitution de terminologies structurées est amorcée sur une ressource sémantiquement riche et ensuite complétée avec des données provenant de sources pauvres (Maedche & Staab, 2000). Dans nos travaux, nous exploitons les ressources pauvres : corpus de textes et listes de termes 'à plat'.

Dans la suite du chapitre, nous précisons les types de relations distingués dans une terminologie structurée (sec. 7.2) et établissons un état de l'art en structuration de terminologies à partir de ressources pauvres (sec. 7.3). Nous présentons ensuite notre méthode de structuration (sec. 7.4) et le matériel utilisé (sec. 7.5). Nous analysons et discutons alors les relations induites (sec. 7.6), et concluons (sec. 7.7).

7.2 Relations dans les terminologies structurées

Nous distinguons trois types de relations dans les terminologies structurées : les relations hiérarchiques, synonymiques et antonymiques, et transversales[1].

7.2.1 Relations hiérarchiques

Les *relations hiérarchiques* relient deux concepts dont l'un recouvre une notion 'plus large' que l'autre. Elles permettent d'organiser les termes en un arbre hiérarchique et

[1]La présentation des types des relations dans une terminologie structurée est fortement inspirée de celles faites dans (Zweigenbaum & Grabar, 2000; Grabar & Hamon, 2004a).

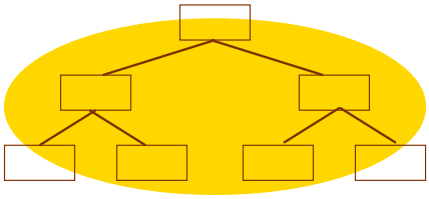


FIG. 7.1 – Relations taxinomiques dans une terminologie structurée

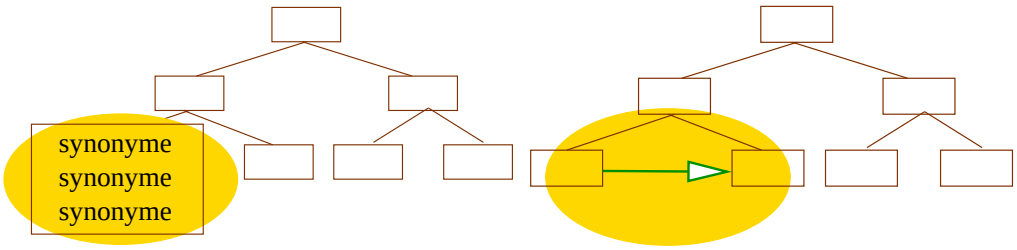
d'établir ainsi une arborescence du domaine (figure 7.1). Ces relations sont aussi appelées génériques, hyperonymie ou relation `est-un`. Le terme générique est appelé *hyperonyme*, le terme spécifique *hyponyme*. Par exemple dans la SNOMED, l'*appendicite* est une des *maladies de l'appendice*.

Les *relations partitives*, appelées aussi relation `partie-de` ou méronymie, identifient les parties d'un objet et servent à décrire les objets manufacturés ou artefacts. Elles sont parfois regroupées avec la relation hiérarchique et participent alors aussi à la constitution de l'arborescence du domaine. Le terme qui désigne le tout est appelé *holonyme*, le terme qui désigne la partie *méronyme*. Mais ces relations peuvent être distinguées de la relation hiérarchique et apparaître comme une relation à part entière ou encore tomber dans l'anonymat des relations transversales. Dans la SNOMED, l'*appendice vermiculaire* fait partie du *gros intestin*, qui fait à son tour partie de l'*appareil digestif*.

Ces relations structurantes sont les plus étudiées car elles fournissent la colonne vertébrale des terminologies structurées. Mettre en place une hiérarchie de concepts selon la relation `est-un` est un acte clé dans la construction d'une terminologie ou d'une ontologie. En recherche d'information en particulier mais aussi dans d'autres applications, ces relations permettent d'adapter la granularité de la recherche selon les besoins de l'utilisateur. L'explosion de concept prend en compte, pour un concept générique, ses concepts spécifiques. Ce qui permet d'améliorer le rappel et de proposer des réponses plus ciblées. Et si un terme ne figure pas dans une terminologie, la détection et l'utilisation d'un hyperonyme permet de proposer néanmoins une réponse. La relation de taxinomie peut revêtir un aspect plus pragmatique. Par exemple, dans le MeSH, elle couvre également la relation d'*aboutness* (`voir-aussi`), considérée comme souhaitable en recherche d'information, car reliant des termes proches. Ainsi, les termes *accident* et *prevention accident* font partie de la relation hiérarchique.

7.2.2 Relations synonymiques et antonymiques

La *synonymie* (figure 7.2(a)) met en relation des expressions qui ont, dans un contexte donné, des sens équivalents. Sortie de ce contexte, la relation de synonymie peut s'atténuer ou bien perdre de la pertinence. Dans les terminologies, les synonymes sont attachés à un même concept. Par exemple le concept `T-50100` de la SNOMED regroupe les termes *voies digestives*, *tractus gastro-intestinal* et *tube digestif*. La relation de synonymie ne constitue



(a) Synonymie (b) Antonymie

FIG. 7.2 – Relations synonymiques et antonymiques dans une terminologie structurée

pas une information obligatoire dans les terminologies. Mais lorsqu'elle est encodée, la synonymie peut être utilisée de différentes manières : en indexation et analyse de contenu, les synonymes aident dans l'appariement d'expressions formellement différentes ; en recherche d'information, les synonymes viennent enrichir la requête et permettent ainsi d'augmenter le rappel ; en rédaction technique, ils permettent d'éviter les répétitions ; dans les travaux de fusion de terminologies, la relation de synonymie sert à apparier les concepts équivalents provenant de différentes sources ; la terminologie normative en fait une utilisation discriminative et les synonymes doivent alors être strictement préférés à leurs équivalents. Notons également qu'en construction d'ontologie, le terme préféré peut recevoir une vocation formelle et servir d'étiquette symbolique à son concept.

L'*antonymie* (figure 7.2(b)) relie des notions contraires ou opposées. Formellement, elle met en relation des termes co-hyponymes qui ne sont pas synonymes (Amsili, 2003). L'antonymie apparaît rarement dans les terminologies. Elle peut pourtant servir pour cerner une requête et éliminer ainsi les documents qui comportent un terme indésirable en recherche d'information ou dans d'autres applications.

7.2.3 Relations transversales

Les *relations transversales* correspondent aux relations entre des termes situés dans les différentes branches de la hiérarchie (figure 7.3). Elles permettent d'identifier des rapports entre termes que les relations synonymiques, antonymiques ou hiérarchiques ne peuvent pas refléter. Il s'agit par exemple des rôles thématiques (objet, instrument, agent, patient, etc.), des relations causales, temporelles, spatiales, mais surtout des relations spécifiques à un domaine.

Dans une terminologie traditionnelle, la relation transversale est exprimée le plus souvent avec un lien `voir-aussi`, et la relation d'*aboutness* s'en rapproche aussi. `voir-aussi` sert de pointeur général vers des notions apparentées, sans préciser la nature de ce rapport. Il joue alors un rôle d'association entre les termes. L'explicitation et le typage de telles relations est un apport très important d'information. Ces relations peuvent correspondre à

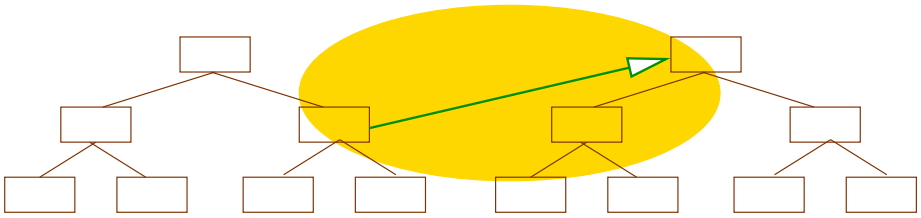


FIG. 7.3 – Relations transversales dans une terminologie structurée

des liens définitoires. Dans la SNOMED par exemple, qui parle de relations 'de référence', elles sont utilisées pour décomposer un concept en d'autres concepts plus élémentaires. Le terme *appendicite* possède ainsi une relation de localisation anatomique avec *appendice vermiculaire* et une relation d'atteinte morphologique avec *inflammation* :

D2-50140 *appendicite* → `localisé-dans` → T-28000 *appendice vermiculaire*
D2-50140 *appendicite* → `a-pour-atteinte` → M-40000 *inflammation*

Cette dernière utilisation des relations transversales est propre aux terminologies dites *post-coordonnées*, qui permettent de composer un terme complexe étant donnée une série de termes simples (Spackman & Campbell, 1998; Zweigenbaum, 1999). Les relations transversales peuvent servir en recherche d'information, où elles peuvent aider l'utilisateur à élargir la requête ou bien, lorsqu'ils ne sont pas sûrs des mots clés, à la formuler en sélectionnant un des termes proposés.

7.3 Travaux en structuration de terminologies

Les travaux en structuration de terminologies que nous présentons ici effectuent une acquisition à partir de ressources pauvres : corpus de textes ou liste de termes. L'acquisition est basée sur l'étude de la structure interne des termes (sec. 7.3.1) ou bien sur leur contexte (sec. 7.3.2). Dans les deux cas, les approches peuvent être endogènes, c'est-à-dire que seules les informations présentes dans le corpus sont utilisées, ou bien basées sur l'exploitation de connaissances extérieures, sous forme de ressources lexicales, de marqueurs, etc.[2].

7.3.1 Étude de la structure des termes

L'étude de la structure interne des termes permet d'acquérir des relations hiérarchiques et des relations synonymiques ou bien des variantes morphosyntaxiques.

Inclusions lexicales des termes. L'hypothèse d'inclusion lexicale décrite notamment par (Kleiber & Tamba, 1990) consiste à identifier des relations d'hyperonymie lexicalement

[2]Cette présentation est fortement inspirée de l'exposé fait dans (Grabar & Hamon, 2004a).

fondées selon le principe suivant : « *...pour passer d'un hyponyme à son hyperonyme, il suffit d'enlever le second élément spécificateur du composé.* »
Par exemple, à partir du terme *sténose sévère*, par suppression du modifieur *sévère*, nous obtenons son hyperonyme *sténose*. Ce type de formation de termes est naturellement utilisé dans les langues de spécialité.

(Bodenreider *et al.*, 2001) appliquent cette hypothèse pour acquérir des relations d'hyperonymie et compléter ainsi la structure d'UMLS, en se basant sur :

- une liste de modifieurs (ou spécificateurs) adjectivaux (*autoimmune*, *hemolytic*, etc.)
- les termes complexes dont la structure est conforme à la syntaxe. Ne sont pas traités les termes artificiels comme *Patient transfer, in-hospital, unit-to-unit*.

Les termes hyperonymes sont déduits par suppression des modifieurs dans les termes traités. Par exemple, les modifieurs *autoimmune* et *hemolytic* sont supprimés du terme *autoimmune hemolytic anemia* pour produire trois nouveaux termes *hemolytic anemia*, *autoimmune anemia* et *anemia*, ainsi que les relations d'hyperonymie entre ces termes. Le traitement d'environ 63 000 termes et 5 400 modifieurs adjectivaux produit environ 60 000 relations d'hyperonymie parmi lesquelles 50 % sont déjà présentes dans l'UMLS et 25 % mettent en relation des termes co-hyponymes d'UMLS. Les 25 % de relations qui restent sont absentes de l'UMLS et doivent être analysées manuellement.

L'analyse syntaxique des termes peut également conduire à l'identification d'inclusions lexicales. L'utilisation des dépendances syntaxiques entre les composants permet de décomposer un terme en tête et expansion (Bourigault, 1993). Cette technique, qui permettait à l'origine de détecter des termes plus élémentaires, s'est avérée utile en structuration de termes en un réseau. Ainsi, l'analyse syntaxique indique que le terme *sténose sévère* a pour tête *sténose*.

Règles de transformation pour le repérage de synonymes et de variantes morphosyntaxiques. L'objectif des travaux qui exploitent des règles de transformation consiste à décrire et à regrouper les variantes morphosyntaxiques ou les synonymes. Ainsi, *sténose de l'aorte* et *sténose aortique* sont des variantes morphosyntaxiques désignant le même terme.

La détection de variantes morphosyntaxiques peut être utilisée pour identifier des variantes de termes contrôlés fournis par un thesaurus ou bien pour repérer les variantes à l'intérieur d'un même corpus (Jacquemin, 1996). L'acquisition de ces variantes repose sur l'hypothèse que la relation de dépendance entre les composants du terme contrôlé est préservée dans ses variantes. Trois variations morphosyntaxiques sont autorisées :

- Coordination de termes :
  X ray diffraction / X ray or neutron diffraction
  (*diffraction des rayons X / diffraction des neutrons ou des rayons X*) ;
- Insertion et juxtaposition de modifieurs :
  blood cell / blood mononuclear cell
  (*cellule du sang /cellule mononucléaire du sang*) ;
- Permutation :

data transmission / transmission of data
(*transmission de données / transmission de données*).

Pour acquérir des relations de synonymie entre termes complexes, (Hamon & Nazarenko, 2001a) font l'hypothèse de la propagation compositionnelle des relations de synonymie. Ainsi, deux termes complexes sont considérés comme synonymes si leurs composants sont identiques ou synonymes. Trois règles sont définies pour inférer des relations de synonymie entre les termes complexes (Hamon & Nazarenko, 2001b) :

- Variation sur l'expansion : les têtes sont identiques et les expansions sont synonymes *action de protection / action de sauvegarde* ;
- Variation sur la tête : les têtes sont synonymes et les expansions sont identiques *capacité faible / puissance faible* ;
- Variation sur la tête et l'expansion : les têtes et les expansions sont synonymes *classement d'équipement / classification de matériel*.

Afin d'éviter la surgénération, les termes mis en relation doivent être attestés en corpus.

7.3.2 Étude du contexte des termes

L'étude du contexte des termes consiste à repérer les régularités d'apparition de ces termes dans les textes afin de détecter les relations qui existent entre eux. La nature des relations peut être définie à l'avance, comme dans le cas des marqueurs et des patrons lexico-syntaxiques ou bien typée *a posteriori* à l'intérieur des classes de termes.

Les marqueurs. Les *marqueurs* sont des éléments textuels caractéristiques d'une relation. Dans les exemples qui suivent *is a* et *including* sont des marqueurs de l'hyperonymie, et *entraîne* est un marqueur de la causalité. La détection d'un marqueur dans un segment de texte déclenche une analyse détaillée de ce contexte.

Les marqueurs permettent de détecter différentes relations : hyperonymie, méronymie, synonymie et relations transversales (Kavanagh, 1995), causalité (Garcia, 1998). Pour identifier les phrases contenant les relations de synonymie, les définitions, l'introduction de nouveaux termes et l'expansion d'abréviations dans des corpus techniques anglais, (Pearson, 1998) utilise des marqueurs suivants : *i.e.*, *e.g.*, *( )*, *known as*, *called*, qui lui permettent de détecter les phrases comme :

Cells or organs which can detect stimuli (e.g. smell, temperature, touch,...)
Amœba. It swims using short hairs called cilia. The cilia also "P 19"...
A signal of limited duration known as a "measuring signal"...

L'application de ces marqueurs sur trois corpus portant sur des domaines différents montre des résultats variables selon les corpus.

La coordination peut également relier des termes sémantiquement proches. C'est la piste suivie par (Lame, 2002, sec.6.1) dans son travail avec les textes juridiques. Les conjonctions de coordination *et* et *ou* permettent ainsi de détecter les relations entre :

aérodromes et *installations à usage aéronautique*
artisants et *commerçants*
déductions fiscales et *réductions d'impôt*.

Patrons lexico-syntaxiques. L'utilisation de *patrons lexico-syntaxiques* repose également sur la présence d'indices linguistiques (*such as*, *including*, *tel que*, *certain*) dans les textes. Mais à la différence des marqueurs, les patrons identifient la relation recherchée plus précisément en définissant à la base les contraintes syntaxiques ou typographiques du contexte, la position des termes en relation, etc. L'utilisation des patrons lexico-syntaxiques dans le traitement de corpus spécialisés débute avec le travail de (Hearst, 1992), qui les exploite pour l'acquisition de relations d'hyperonymie.

Cette approche a été reprise et implémentée par (Morin, 1999) pour assister le repérage de patrons lexico-syntaxiques. Les patrons spécifiques du corpus étudié sont définis à l'aide d'un thesaurus : (1) les termes du thesaurus, qui ont une relation d'hyperonymie entre eux, sont projetés sur un corpus ; (2) les segments de texte pertinents sont utilisés pour construire les patrons candidats ; (3) les patrons pertinents sont sélectionnés manuellement et (4) ensuite projetés sur le corpus pour détecter de nouveaux termes en relation d'hyperonymie. Ainsi, la projection de la relation d'hyperonymie *glycérol* `est-un` *polyol* permet d'extraire le patron « `SN tel que LISTE_SN` » de la phrase :

> *L'hydrolyse des substrats est activée par le glucose et les polyols tels que le sorbitol et le glycérol.*

Et l'application de ce patron sur le corpus permet de détecter à partir de la phrase :

> *Des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par des méthodes de routine.*

de nouvelles relations d'hyperonymie :

> *sodium* `est-un` *cation*,
> *potassium* `est-un` *cation*,
> *calcium* `est-un` *cation* et
> *magnésium* `est-un` *cation*.

La définition de patrons lexico-syntaxiques pour la détection de différentes relations (hyperonymie, méronymie, synonymie, relations transversales) peut être basée sur les marqueurs caractéristiques proposés dans la littérature (Séguéla & Aussenac-Gilles, 1999). Le souci principal est alors leur adaptation aux nouveaux corpus à traiter. Les patrons peuvent aussi provenir d'une étude linguistique du corpus traité (Condamines & Rebeyrolle, 1998).

Constitution de classes de termes. La constitution de classes sémantiques de termes est basée sur l'étude de la similarité de la distribution (occurrences et voisinages) des termes en corpus. Le principe distributionnel (Harris, 1971), qui se trouve à la base de ces travaux, stipule que les termes qui apparaissent dans les contextes similaires ont un sens proche.

La représentation du contexte par l'analyse distributionnelle peut être une 'fenêtre de mots' ou une relation de dépendance syntaxique. Pour obtenir cette dernière, il faut effectuer une analyse syntaxique. Deux mots ou termes sont alors considérés comme sémantiquement proches s'ils partagent les mêmes dépendances. Ces dépendances peuvent être recherchées au sein de :

- groupes nominaux, adjectivaux, prépositionnels, etc. (Grefenstette, 1994),
- groupes nominaux (Habert *et al.*, 1996; Assadi, 1998),

– groupes verbaux (Bourigault & Fabre, 2000).

Un calcul d'association entre les mots ou termes permet ensuite de forger les classes.

Les groupes de similarité ou les classes peuvent avoir une forte proximité sémantique interne. Par exemple (Grefenstette, 1994) obtient les groupes suivants :

large : *small*, *important*, *major*, *great*, *various*, *main*, *different*, *field*, *new*

patient : *case*, *group*, *child*, *day*, *treatment*, *woman*

Un typage manuel des relations conduit à la détection de : antonymie (*large*, *small*) ; synonymie (*large*, *important*, *great*) ; méronymie (*patient*, *group*) ; hyperonymie (*patient*, *child*, *woman*) ou bien des relations plus spécifiques et dépendantes du corpus et du domaine (*patient*, *treatment*).

Le domaine de la fouille de textes propose des méthodes qui exploitent le modèle des règles d'association. Ainsi, (Toussaint & Simon, 2000) prennent en compte l'apparition de termes dans un document et en déduisent des règles d'association du type :

$Terme_1, Terme_2 \Rightarrow Terme_A, Terme_B, ...$

Une telle règle associe les termes $Terme_1$ et $Terme_2$ aux termes $Terme_A$ et $Terme_B$ et suppose qu'il existe une relation sémantique entre ces deux groupes de termes. Comme dans les travaux précédents, le typage des relations entre les termes de chaque règle est effectué manuellement. L'application de l'approche sur un corpus de résumés d'articles scientifiques dans le domaine de l'agriculture permet de calculer des règles d'association qui infèrent des relations d'hyperonymie (*histamine* `est-un` *biogenic amine*), de cohyponymie (*spermidine* `est-un-frère` *putrescine*) et des relations spécifiques au domaine (*acids* `se-transforment-en` *esters*, *silica* `utilisé-pour` *chromatography*).

La difficulté principale des approches de classification réside dans l'interprétation indispensable des classes construites. L'intervention de l'utilisateur, qu'il soit terminologue ou expert, est primordiale et coûteuse.

7.4 Méthodes d'induction et d'évaluation de relations hiérarchiques

L'inclusion lexicale est naturellement utilisée dans la formation de termes. Nous voulons donc exploiter ce fait pour la détection de relations hiérarchiques. De plus, une telle approche se prête facilement au traitement de la variation terminologique et à l'injection de connaissances supplémentaires, y compris au niveau morphologique. Aux termes du MeSH mis à plat nous appliquons l'hypothèse d'inclusion lexicale pour induire des relations hiérarchiques entre eux (sec. 7.4.1). Nous utilisons ensuite la structure originale du MeSH comme le référentiel pour évaluer le rappel et la précision des relations induites avec notre méthode (sec. 7.4.2).

7.4.1 Détection d'inclusions lexicales

La méthode que nous proposons pour la détection de relations hiérarchiques est basée sur l'étude d'inclusions lexicales. Elle consiste à vérifier si un terme P (*parent*) est inclus

dans un autre terme F (*fils*). Nous supposons que ce type d'inclusion constitue une clé de la relation hiérarchique entre les termes, comme dans l'exemple :

acides gras / *acides gras indispensables*.

Notre méthode s'apparente donc aux travaux, décrits dans la sec. 7.3.1, qui appliquent la même hypothèse. Mais elle se distingue par le fait qu'elle n'est pas restreinte à l'analyse d'une liste prédéfinie de modifieurs (Bodenreider *et al.*, 2001), elle n'exploite pas les dépendances syntaxiques dans les groupes nominaux (*têtes* et *expansions* syntaxiques) (Bourigault, 1993).

Pour détecter la relation d'inclusion entre deux termes, nous vérifions si tous les mots du terme P apparaissent dans le terme F. Nous effectuons les tests sur les termes segmentés, qui se présentent alors chacun sous forme de *sac de mots* : l'ordre des mots dans les termes n'est plus respecté et les mots sont ordonnés alphabétiquement. Les tests sont faits d'abord sur les formes brutes des mots et ensuite avec une série de normalisations (les ressources utilisées pour chaque normalisation sont décrites dans la section 7.5) :

- normalisation de base : conversion en caractères minuscules, suppression d'accents, de la ponctuation, de nombres et de mots vides ;
- normalisations avec des ressources morphologiques flexionnelles (de la langue médicale et générale), dérivationnelles et allomorphiques ;
- normalisations avec des synonymes de la langue médicale, de la langue générale et les synonymes de la langue générale filtrés sur les corpus médicaux.

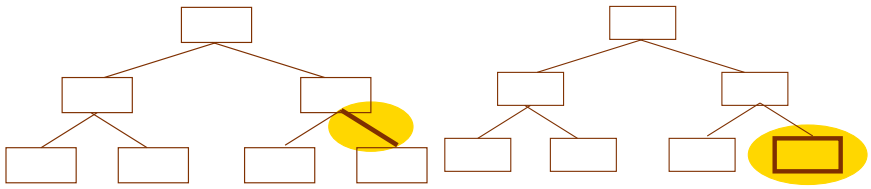
Pour accélérer le calcul des inclusions, les termes sont indexés par les mots qui les composent. Lorsque les normalisations sont appliquées, les termes sont indexés avec les mots normalisés. Nous considérons alors que le terme P est lexicalement inclus dans F si tous les mots normalisés de P apparaissent dans F. Les ressources linguistiques utilisées sont également soumises aux normalisations. Comme les termes du MeSH sont écrits en majuscules non accentuées et les ressources morphologiques et synonymiques en minuscules accentuées, les normalisations de base permettent de ramener à une base commune les deux ensembles. Les normalisations morphologiques et synonymiques permettent d'effectuer des tests d'appariement plus complets. De manière générale, les normalisations lexicales appliquées à la liste de termes couvrent un nombre croissant de variations terminologiques et permettent d'induire un nombre croissant de relations entre les termes.

L'objectif de cette expérience est donc de vérifier la validité de l'hypothèse de l'inclusion lexicale sur une liste de termes médicaux. Mais aussi d'examiner l'influence des normalisations appliquées et des connaissances linguistiques injectées sur les résultats induits.

7.4.2 Evaluation par rapport au référentiel existant

Nous évaluons les relations induites et les termes placés en les comparant avec la structure originale du MeSH. Nous évaluons les résultats selon deux méthodes (figure 7.4) :

1. la première méthode prend en compte le nombre de relations induites et les compare avec les relations qui existent dans la structure originale du MeSH. Nous cherchons à savoir si nous induisons toutes les relations présentes dans le MeSH ;



(a) Induction d'une relation (b) Placement d'un terme

FIG. 7.4 – Illustration des deux méthodes d'évaluation des résultats. À gauche une relation correcte est trouvée dans la structure originale du MeSH. À droite un terme est correctement placé dans la structure originale du MeSH.

2. la deuxième méthode prend en compte le placement des termes individuels dans la structure hiérarchique du MeSH. Nous cherchons alors à savoir si nous arrivons à placer les termes dans au moins une position correcte dans la hiérarchie du MeSH.

En sachant que les termes du MeSH peuvent avoir plusieurs positions hiérarchiques et donc plusieurs 'pères', il est plus facile de placer un terme (2) que de le placer dans toutes ses positions et sous tous ses pères (1).

Dans chaque cas, le rappel et la précision sont calculés selon deux versions :

- stricte (ou directe) : seules les relations directes entre un père et un fils sont considérées comme correctes,
- tolérante (ou indirecte) : toute relation vers un ancêtre est considérée comme correcte.

Le rappel direct ou strict R_s est mesuré avec le nombre de relations directes d induites divisé par le nombre de relations directes D existant dans le MeSH (26 094). Le rappel tolérant R_t est mesuré par le nombre de relations directes et indirectes $d+i$ induites, divisé par le nombre total (directes et indirectes) de relations du MeSH $D + I$ (95 815) :

$$R_s = \frac{d}{D}; R_t = \frac{d+i}{D+I}$$

L'évaluation de la précision est faite selon les mêmes principes : si d est le nombre de relations directes induites, i le nombre de relations indirectes induites, et n le nombre de relations extra-MeSH induites, les précisions stricte P_s et tolérante P_t correspondent à :

$$P_s = \frac{d}{d+i+n}; P_t = \frac{d+i}{d+i+n}$$

Nous avons également testé un schéma mixte : le poids donné à chaque relation dépendait de la distance entre les deux termes reliés dans la structure originale du MeSH. Plus la distance est grande plus le poids de la relation induite est faible. Mais nous ne présentons pas ses résultats, car ce schéma donnait des évaluations très proches du schéma tolérant.

7.5 Préparation et description du matériel

Dans ces expériences nous utilisons une liste de termes à structurer (sec. 7.5.1), le référentiel de leur structuration (sec. 7.5.2), des connaissances morphologiques (sec. 7.5.3), des connaissances synonymiques (sec. 7.5.4) et une liste de 'mots vides' (sec. 7.5.5).

7.5.1 Liste de termes

La liste de 19 638 termes à structurer est extraite du MeSH (voir ci-dessous).

7.5.2 Structuration de référence des termes : structure originale du MeSH

Le thesaurus international MeSH comporte des termes qui ne sont pas des expressions naturelles, mais ont été composés par ses créateurs dans le but de décrire le mieux possible les concepts du domaine. La version française du MeSH 2001 que nous utilisons comporte 19 638 termes, 26 094 relations directes père-fils et 95 815 relations directes ou bien indirectes père-fils. Les termes du MeSH français sont écrits en majuscules non accentuées.

7.5.3 Connaissances morphologiques

Connaissances flexionnelles. Pour normaliser les variations flexionnelles des mots, nous utilisons en parallèle deux ressources : ressources spécifiques au domaine médical et ressources provenant du lexique général. Les ressources flexionnelles spécifiques au domaine médical (*lem-med*) ont été induites à partir de la SNOMED et de la CIM avec la méthode appliquée aux terminologies structurées (chap. 4). Elles comportent 3 470 couples {*lemme*, *forme*}. Les ressources flexionnelles de la langue générale (*lem-gen*) proviennent du lexique ABU (Association des bibliophiles universels)[3] et de corpus médicaux (cardiologie, hématologie, soins médicaux, monographies de médicaments) étiquetés et lemmatisés. Elles contienent 308 812 couples {*lemme*, *forme*}.

Connaissances dérivationnelles. Pour les dérivations (*rac-med*), nous utilisons 2 418 couples {*base*, *dérivé*} induits à partir de terminologies structurées (chap. 4).

Connaissances allomorphiques. Pour compléter les dérivations avec des formations supplétives et allomorphiques, nous avons extrait de la SNOMED les synonymes simples qui appartiennent à des 'classes' différentes (01, 02 et 03 pour les noms et les syntagmes nominaux, 05 pour les adjectifs et les éléments de composition). Nous supposons ainsi que parmi ces synonymes nous trouverons des mots reliés morphologiquement. Après un filtrage manuel, nous obtenons 462 couples de type {*estomac*, *gastrique*}. Nous disposons alors de 2 716 couples dérivationnels (*allom*).

[3] `http://abu.cnam.fr/DICO/`

7.5.4 Synonymes

Comme pour les flexions, nous utilisons deux ensembles de synonymes entre mots : les synonymes de la langue médicale (dictionnaire Masson[4]) et de la langue générale (le Petit Robert). Dans ce dernier, les renvois, simples et doubles, qui semblent avoir été utilisés pour établir les relations de synonymie, peuvent correspondre en réalité à plusieurs relations (synonymes, analogies, thèmes d'expressions, variantes orthographique, sous-entrées, hyperonymes, etc.) mettant ainsi ensemble des données qui sont fondamentalement hétérogènes (Marcus, 2003) déjà dans la langue générale. Nous avons donc décidé de les filtrer afin de les adapter au domaine médical.

Un autre de nos soucis concernait la directionnalité de cette relation. La synonymie, reliant des mots ou expressions contextuellement interchangeables, peut sans doute être considérée comme une relation bidirectionnelle, où tous les synonymes d'une famille sont 'égaux' quant à leurs emplois. Nos ressources synonymiques par contre sont ordonnées : {*entrée dictionnairique*, *famille de synonymes*} ou {*entrée dictionnairique*, *synonyme*}. Nous avons préféré de garder cet ordre pour nos expériences. Ce qui semble être raisonnable, surtout face aux ressources de la langue générale qui regroupent des lexiques de différents registres :

> {*boulimie*, *cynorexie*}, {*boulimie*, *hyperorexie*}, {*boulimie*, *hyperphagie*}, {*boulimie*, *sitiomanie*}, {*boulimie*, *faimcalle*}, {*boulimie*, *appétit*}, {*boulimie*, *avidité*}.

Dans les ressources spécialisées la stabilité sémantique des synonymes est plus grande. *Boulimie* y reçoit quatre synonymes :

> {*boulimie*, *cynorexie*}, {*boulimie*, *hyperoxie*}, {*boulimie*, *hyperphagie*}, {*boulimie*, *polyorexie*}.

À titre d'exemple, si nous considérons la synonymie comme une relation bidirectionnelle, la famille *boulimie* est 'enrichie', dans les ressources de la langue générale, de *faim*, *fringale* et *frénésie*. Et les mots qui ont un sens spécialisé dans le domaine médical mais qui sont également employés hors du domaine ont une expansion sémantique encore plus importante : *ventre* passe de 11 synonymes à 19, *trouble* de 22 à 64.

Synonymes spécifiques au domaine médical. L'ensemble de synonymes spécifiques au domaine médical (*syno-med*) provient du dictionnaire médical Masson. Il comporte 831 paires de synonymes simples.

Synonymes de la langue générale. L'ensemble de synonymes de la langue générale (*syno-gen*) provient du dictionnaire Le Petit Robert[5]. Il comporte 140 141 paires de synonymes simples. Pour les raisons qui nous venons d'exposer, nous avons décidé de les filtrer sur des corpus médicaux afin de les adapter au domaine.

[4]En ligne à l'époque de ce travail sur `www.AtMedica.com`

[5]Nous remercions l'INaLF et Didier Bourigault d'avoir rendu disponibles ces ressources et Thierry Hamon de nous les avoir fournies nettoyées et formatées.

Synonymes de la langue générale filtrés sur les textes médicaux. Les synonymes de la langue générale sont filtrés à travers un corpus médical d'environ 8,5 millions d'occurrences (voir sec. 5.4 page 99). Le corpus contient des documents hospitaliers (lettres, comptes rendus hospitaliers) et des documents grand public collectés sur le Web à travers le portail médical CISMeF. Nous appliquons trois approches pour effectuer le filtrage : (1) projection de patrons lexico-syntaxiques, (2) projection de marqueurs de coordination et (3) calculs d'association entre les mots.

Les patrons lexico-syntaxiques de synonymie proviennent du travail de (Séguéla & Aussenac-Gilles, 1999). Nous avons par exemple les patrons suivants (`X` et `Y` correspondent aux termes synonymes) :

«

`X appelé Y »`
« `X est défini comme 1-MOT Y »`
« `X est confondu avec Y »`
« `X n'est autre que 1-MOT Y »`

Nous trouvons alors des contextes qui valident les couples suivants :

- {*gonflement*, *œdème*} : *L'œdème* `est défini comme` *un gonflement palpable produit par l'expansion du volume interstitiel liquidien.*
- {*rhinopharynx*, *cavum*} : *Le rhinopharynx* `appelé` *cavum est situé sous la base du crâne, en arrière des fosses nasales, au-dessus de l'oropharynx et en avant des 2 premières vertébres cervicales.*
- {*repos*, *sommeil*} : *Trop souvent, repos* `est confondu avec` *sommeil et activité avec éveil.*

Notons que nous n'avons pas détecté de patrons dans les documents hospitaliers. Cela semble raisonnable : ces documents s'adressent à des spécialistes avec, comme but principal, la transmission d'informations sur les patients. Par contre dans les documents du Web, destinés à un public plutôt non averti, les reformulations et le recours à la synonymie sont fréquents. Notons aussi que nous avons projeté les synonymes sans effectuer une analyse syntaxique. Une telle projection touche vite à ses limites. Dans le premier exemple qui suit, l'un des synonymes (*court*) fait partie d'un terme plus grand (*texte court*) et reste donc sous-spécifié dans cette relation de synonymie. Le deuxième exemple présente le même cas : il s'agit d'un *syndrome hépato-rénal* qui `est-une` sorte d'*infection* :

- {*résumé*, *court*} : *A chaque maladie est associé un texte court* `appelé` *"résumé" (moins de 200 mots) présentant les grandes caractéristiques dans un langage relativement simple.*
- {*syndrome*, *affection*} : *Cette affection est* `appelée` *le syndrome hépato-rénal qui est défini comme une augmentation progressive de la créatinine plasmatique, sans cause évidente autre chez un patient atteint de maladie hépatique avancée.*

Les patrons lexico-syntaxiques permettent de retenir 46 paires de synonymes.

Notre deuxième piste vient du travail de (Lame, 2002, sec.6.1), où l'auteur note que dans les documents juridiques les marqueurs de coordination (*et*, *ou*) relient des termes possédant des sens proches. Partant de cette constatation, nous supposons que si ces marqueurs relient

dans notre corpus les mots d'une paire de synonymes, il s'agit effectivement de synonymes du domaine médical. Nous avons ajouté d'autres marqueurs qui s'avèrent fréquents dans les documents médicaux : la coordination négative *ni* et la particule de la négation *pas*. À l'analyse des résultats, il apparaît que ces marqueurs relient plutôt des co-hyponymes, comme l'a déjà remarqué (Pearson, 1998) :

- {*bruit*, *souffle*} : *Examen cardiaque : bruits bien frappés aux 4 foyers sans souffle* `ni` *bruit surajoutés.*
- {*orthopnée*, *dyspnée*} : *Examen cardio-vasculaire : pas de dyspnée,* `pas` *d'orthopnée, présence d'œdèmes des membres inférieurs avec un godet positif.*

Les marqueurs de coordination utilisés permettent de retenir 1 736 paires de synonymes.

Une dernière approche de filtrage dérive de la méthode décrite dans (Zweigenbaum *et al.*, 2003b), où les auteurs appliquent une mesure statistique (*log likelihood ratio* (Manning & Schütze, 1999)) pour détecter l'association entre les mots reliés morphologiquement. Nous l'utilisons pour le filtrage des paires de synonymes, dans une fenêtre de 2*150 *mots pleins*[6]. Nous obtenons, par exemple, les associations suivantes :

{*abcès*, *phlegmon*}, {*biopsie*, *ponction*}, {*dernier*, *culot*}, {*signal*, *appel*}.

Ce filtrage propose aussi des erreurs, comme {*dernier*, *culot*} dans :

Dernier culot reçu le 24/09/1888.

Où *culot* signifie « *amas d'érythrocytes* » ou « *pochette de globules rouges* » (il s'agit d'une transfusion de globules rouges) et *dernier* est un qualificatif. Le sens « *fond, dépôt, résidu* » du couple {*dernier*, *culot*} est assez éloigné de cette acception. En gardant 60 % meilleurs couples, tels que classés par ordre décroissant de mesure d'association, nous obtenons 15 589 paires.

Le filtrage complet nous donne un ensemble de 16 154 paires de synonymes (*syno-gen-f*), ce qui correspond à une réduction de presque 90 % par rapport à l'ensemble synonymique de départ.

7.5.5 Mots vides

Nous utilisons une liste de mots vides composée uniquement de mots grammaticaux qui apparaissent dans les termes du MeSH et qui ne sont pas ambigus. Elle comprend au total 15 mots, essentiellement des articles et prépositions : *au*, *aux*, *d'*, *de*, *des*, *du*, *en*, *et*, *l'*, *la*, *le*, *les*, *ses*, *un*, *une*.

7.6 Analyse et discussion des relations induites et des termes placés

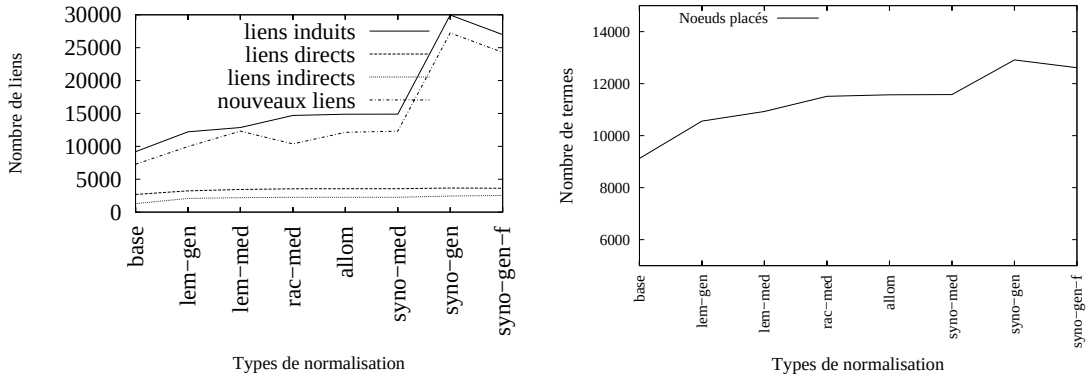
Dans la présentation des résultats, nous discutons les points suivants : l'évolution du nombre de relations induites avec les inclusions lexicales (sec. 7.6.1), leur évaluation par

[6]Les mots pleins s'opposent aux mots vides. Il s'agit donc des catégories majeures : noms, adjectifs, verbes et adverbes.

rapport à la structure originale du MeSH (sec. 7.6.2), l'analyse des relations hors-MeSH (sec. 7.6.3) et l'application de la méthode à d'autres termes (sec. 7.6.4)[7].

7.6.1 Évolution des relations induites avec les inclusions lexicales

Nous avons appliqué notre méthode de détection des inclusions lexicales à une liste à plat de 19 638 termes du MeSH. La figure 7.5 montre les résultats quantitatifs obtenus à chaque étape des normalisations : normalisation des caractères et suppression des mots vides (*base*), application des ressources flexionnelles de la langue générale (*lem-gen*) et de la langue médicale (*lem-med*), application des ressources dérivationnelles (*rac-med*) et complétées avec les supplétions et allomorphies (*allom*), application des synonymes de la langue médicale (*syno-med*), de la langue générale (*syno-gen*) et des synonymes de la langue générale filtrés (*syno-gen-f*). La figure 7.5(a) montre le nombre de relations induites. Nous indiquons les valeurs pour la totalité des relations induites, parmi lesquelles les liens directs et indirects du MeSH et les nouveaux liens. La figure 7.5(b) indique le nombre de termes placés sous un terme plus général avec notre méthode.



(a) Liens induits

(b) Termes placés

FIG. 7.5 – Quantification des relations induites et des termes placés.

Comme attendu, le nombre de relations augmente avec l'application des normalisations successives. Les ressources flexionnelles de la langue médicale (*lem-med*) permettent de relier plus de termes que les ressources flexionnelles de la langue générale (*lem-gen*) : 12 857 *vs* 12 210 relations. La même observation est vraie pour le placement des termes : la couverture est également meilleure avec l'utilisation des ressources flexionnelles spécifiques du domaine (10 929 *vs* 10 560 termes). Comme les ressources *lem-med* sont plus performantes, la suite des normalisations est enchaînée à cette branche. Les paires dérivationnelles de

[7]Dans cette présentation nous reprenons les discussions faites dans nos publications antérieures (Grabar & Zweigenbaum, 2002a; Grabar & Zweigenbaum, 2002b; Grabar & Jeannin, 2002).

Normalisation	Père P	Fils F
base $_{direct}$	accouchement	accouchement provoque
base $_{indirect}$	acides gras	acides gras indispensables
lem-gen $_{direct}$	intervention chirurgicale	interventions chirurgicales obstetricales
lem-gen $_{indirect}$	intervention chirurgicale	interventions chirurgicales voies biliaires
lem-med $_{direct}$	agents adrenergiques	inhibiteurs captage agent adrenergique
lem-med $_{indirect}$	chromosomes humains	chromosome humain 21
rac-med $_{direct}$	aberration chromosomique, anomalies	aberrations chromosomes sexuels, anomalies
rac-med $_{indirect}$	eosinophilie	poumon eosinophile
allom $_{direct}$	poumon	eau extravasculaire pulmonaire
allom $_{indirect}$	estomac	cellule parietale gastrique
syno-med $_{direct}$	saccharose	sucrose alimentaire
syno-med $_{indirect}$	—	—
syno-gen $_{direct}$	fracture machoire	fracture maxillaire inferieur
syno-gen $_{indirect}$	therapeutique	traitement par art

TAB. 7.1 – Exemples de relations induites avec les inclusions lexicales et se trouvant dans la structure originale du MeSH. 'Indirect' signifie que la distance entre les deux termes est supérieur à 1.

mots augmentent encore le nombre de relations induites (14 695 avec *rac-med vs* 12 857 avec *lem-med*) et de termes placés (11 511 *vs* 10 929). L'utilisation des allomorphies et supplétions (*allom*) permet d'induire 189 relations et de placer 61 termes en plus qu'avec *rac-med*. Toutes les ressources synonymiques sont appliquées en parallèle après la totalité des normalisations morphologiques (*allom*). Les synonymes médicaux induisent 8 nouvelles relations et placent en plus 7 termes. L'application de synonymes de la langue générale montre un gain impressionnant : 15 085 relations induites et 1 314 termes placés de plus qu'avec *allom*. Avec les synonymes généraux filtrés le gain est légèrement moins important : 12 102 relations et 1 039 termes placés de plus qu'avec *allom*. Dans le tableau 7.1 nous présentons des exemples des termes reliés à chaque étape de normalisation : nous indiquons les termes père P et fils F correspondant aux relations directes et indirectes dans la structure originale du MeSH.

À travers la figure 7.6 nous voulons montrer l'utilité des différentes ressources. Pour chaque étape de normalisation, nous indiquons le nombre de mots dans la ressource linguistique en question, le nombre de mots dans les termes du MeSH (15 446 après la suppression des mots vides et des nombres), et le nombre de mots qui sont réellement traités lors du calcul des inclusions lexicales. Nous pouvons ainsi voir que la totalité des ressources est rarement utilisée, surtout avec les ressources de la langue générale (*lem-gen* et *syno-gen*). À côté de la contribution à l'induction des relations, c'est un autre facteur qui milite pour l'adaptation de ressources linguistiques générales au domaine étudié.

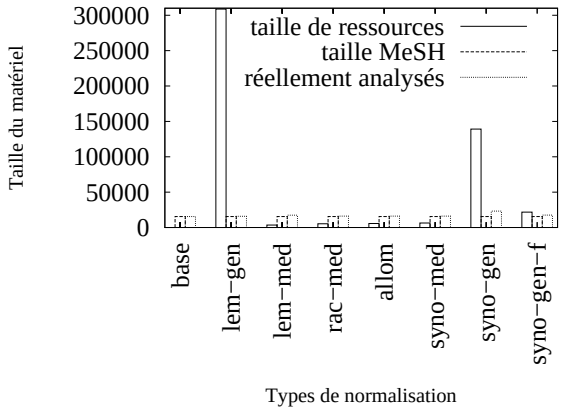


FIG. 7.6 – Utilisation réelle des ressources linguistiques à chaque étape des normalisations.

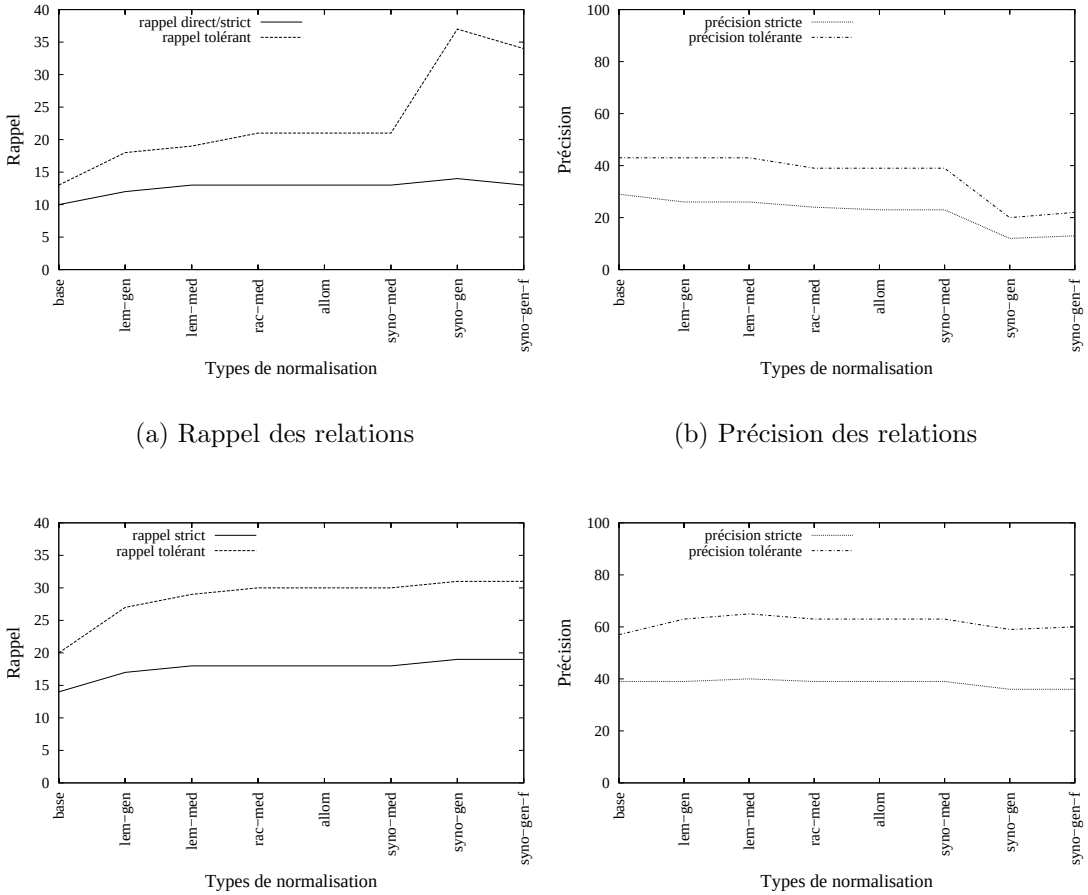
Bilan

L'injection successive de connaissances linguistiques permet d'augmenter le nombre de relations induites et de termes placés. Notons que les relations sont plus sensibles à l'injection de telles connaissances. Alors que la normalisation lexicale de *base* permet de relier les termes grâce à l'appariement des mots identiques, les connaissances linguistiques (ressources morphologiques et synonymiques) introduisent un aspect sémantique dans l'appariement des termes. La lemmatisation ajoute une flexibilité face aux variantes flexionnelles des mots. Et la racinisation permet de relier les termes dont les mots, bien que formellement différents, sont en relation morphologique et sémantique très proche. Et enfin, les paires de synonymes aident à induire des relations entre les termes en appariant les mots qui sont en relation sémantique proche, au moins dans certains contextes. Nous avons noté également que les ressources spécifiques au domaine s'avèrent plus utiles lors des traitements.

7.6.2 Évaluation des inclusions lexicales par rapport au MeSH

L'évaluation des données induites par rapport à la structure originale du MeSH nous permet de calculer leurs rappel et précision. Dans tous les cas, nous prenons en compte la nature des relations induites (directe ou indirecte par rapport à la structure originale du MeSH) et calculons les variantes stricte et tolérante. Rappelons que le *rappel strict* et la *précision stricte* prennent en compte uniquement les relations directes ; le *rappel tolérant* et la *précision tolérante* prennent en compte les relations directes et indirectes. La figure 7.7 montre les résultats de cette évaluation.

Lorsque nous considérons les relations, le rappel augmente avec l'injection de connaissances linguistiques supplémentaires. Les connaissances morphologiques (flexion, dérivation et ensuite les allomorphies et supplétions) participent à l'amélioration du rappel, qui va de 10,3 % avec la normalisation de *base* jusqu'à 13,6 avec *allom* pour le rappel strict, et de 13,7 % à 21,6 % pour le rappel tolérant. Notons que le rappel des relations obtenues



(a) Rappel des relations

(b) Précision des relations

(c) Rappel des termes placés

(d) Précision des termes placés

FIG. 7.7 – Évaluation des relations et placements induits : rappel et précision.

avec les connaissances morphologiques spécifiques du domaine (*lem-med*, *rac-med* ou *allom*) est meilleur : 19,3, 21,4 et 21,6 % contre 18,3 % obtenus avec *lem-gen*. L'application de synonymes du domaine médical augmente très légèrement le rappel, par contre avec les synonymes de la langue générale il atteint un sommet. Le rappel strict augmente ainsi jusqu'à 14 % et le tolérant jusqu'à 37,6 %. Avec les synonymes filtrés, le rappel strict est de 13,9 % et le tolérant de 34,6 %. Comme c'est souvent le cas, l'évolution de la précision est opposée : l'injection de connaissances morphologiques (dérivation *vs* flexion) et ensuite synonymiques apporte plus de 'risque' dans la génération de relations incorrectes entre les termes. Par rapport à la configuration de *base*, la précision stricte diminue de 29,3 % à 23,9 % avec *allom* et *syno-med*, et jusqu'à 12,2 % avec *syno-gen*. Les synonymes filtrés la remontent à 13,5 %. Lorsque nous acceptons les relations quelle que soit leur nature (directe ou indirecte), la précision est plus élevée que lorsque seules les relations directes

sont considérées. Par exemple, avec la normalisation de base, l'approche tolérante montre une précision de 43,3 % alors que la précision stricte est de 29,3 %. Et avec *syno-gen* et *syno-med*, la précision tolérante est respectivement de 20,4 % et 39 % et la stricte de 12,2 % et 23,9 %. Selon la chaîne de normalisation et le schéma d'évaluation, jusqu'à 43,8 % de relations induites sont correctes (approche tolérante avec *lem-med*), et jusqu'à 14 % de relations directes par rapport au MeSH (approche stricte avec *syno-gen*).

Quant aux placements des termes, les courbes du rappel et de la précision sont moins sensibles aux normalisations. Le rappel augmente avec l'utilisation de ressouces flexionnelles, mais ensuite il évolue beaucoup plus doucement. La précision augmente avec l'injection des connaissances morphologiques (*lem-med*) et ensuite diminue, surtout à l'étape d'injection des synonymes de la langue générale non filtrés (*syno-gen*). Jusqu'à 31 % des termes MeSH sont correctement placés sous leurs ancêtres (approche tolérante avec *syno-gen* ou *syno-gen-f*) ; et les placements proposés sont corrects jusqu'à 65 % (approche tolérante avec *lem-med*).

De manière générale, les connaissances linguistiques spécifiques au domaine médical sont plus profitables dans nos traitements que les connaissances qui proviennent de la langue générale. La précision élevée obtenue à ces étapes en est le témoin.

L'induction de nouvelles relations semble être plus réactive à l'injection de connaissances linguistiques, et surtout aux ressources synonymiques. La précision des relations chute avec les synonymes de la langue générale, elle est légèrement améliorée avec les synonymes généraux filtrés. L'évolution du rappel est alors inverse : il augmente fortement. Tandis que le placement des termes, tout en ayant des valeurs de rappel et de précision plus élevées, montre une évolution très douce. Comme attendu, il est plus facile de placer un terme dans un seul endroit de la hiérarchie que de trouver tous ses parents existants dans le MeSH.

Bilan

La comparaison avec la structure originale du MeSH montre qu'avec l'injection de ressources linguistiques et selon le schéma de normalisation, jusqu'à 43,8 % de relations induites sont correctes (approche tolérante avec *lem-med*), et jusqu'à 14 % de relations directes par rapport au MeSH (approche stricte avec *syno-gen*). En ce qui concerne les placements de termes, jusqu'à 31 % des termes MeSH sont correctement placés sous leurs ancêtres (approche tolérante avec *syno-gen* ou *syno-gen-f*) et ces placements sont corrects jusqu'à 65 % (approche tolérante avec *lem-med*). Ici encore, nous notons que l'apport des ressources linguistiques, morphologiques ou synonymiques, est meilleur lorsqu'elles sont spécifiques au domaine traité. Nous observons que leur contribution est similaire à celle signalée par (Hamon *et al.*, 1998) : les ressources de la langue générale permettent d'augmenter le rappel, tandis que les ressources spécifiques au domaine améliorent la précision. L'adaptation des ressources de la langue générale au domaine traité permet de réduire la différence entre les deux. Les valeurs de rappel et de précision plus élevées pour l'induction des relations montrent qu'il est plus facile de placer un terme dans un endroit que de trouver tous ses parents.

Normalisations	Tête	Expan.	Autres
base	22	31	14
rac-med	37	57	33

TAB. 7.2 – Relations analysées : la tailles des échantillons.

7.6.3 Analyse manuelle de 'nouvelles' relations

Jusqu'ici nous avons analysé et évalué, parmi les relations induites, celles qui existent dans la structure originale du MeSH. Mais notre méthode génère jusqu'à 70 % de relations considérées comme 'erronées' ou bien 'nouvelles', c'est-à-dire de relations qui n'existent pas dans le MeSH. Nous analysons maintenant la nature de ces relations et cherchons à savoir pourquoi elles ne sont pas encodées dans le MeSH. Selon la position du terme père dans le terme fils, nous distinguons trois ensembles de relations nouvelles :

- le terme père se trouve en position de *tête* (syntaxique et sémantique) dans le terme fils : *absorption / absorption intestinale*;
- le terme père se trouve en position d'*expansion* dans le terme fils : *abdomen / tumeur abdomen*;
- le terme père se trouve dans d'*autres* types de positions dans le terme fils.

Pour l'analyse, nous avons prélevé un échantillon de 20 % des relations induites avec la normalisation de *base* et avec la chaîne de normalisations allant jusqu'à *rac-med*. Le tableau 7.2 indique le nombre de relations analysées de chaque type, au total 194 relations.

Ensemble tête

Nous avons adopté un critère heuristique pour déterminer la tête des termes : elle doit apparaîte en début de termes. Cette heuristique marche bien en général (*penicillines / penicilline g*), sauf quelques exceptions où la tête ne se trouve pas au début des termes.

Termes 'inversés'. Parmi les termes du MeSH, créés pour la plupart artificiellement, il s'agit des termes qui ont une structure 'inversée' et comportent généralement une virgule :

filoviridae / filoviridae, infections
quinones / quinone reductases
neurone / neurone moteur, maladie

Nous avons traité ces termes tels quels. Mais ils devraient être réorganisés à la main pour être analysés dans les catégories correspondantes ou bien ignorés (Bodenreider *et al.*, 2001).

Énumérations. À côté des termes inversés, nous avons d'autres cas où le terme incluant n'est pas le fils du terme inclus qui se trouve en position tête. Ceci pour des raisons syntaxiques. Dans les termes :

acides amines / acides amines, peptides et proteines

le terme incluant est une énumération et ses constituants sont reliés entre eux avec un OU logique. Le terme inclus (*acides amines*), au même niveau que d'autres constituants (*peptides* et *proteines*), est un terme plus spécifique que le terme *acides amines, peptides et proteines*. Le sens de la relation hiérarchique est ainsi inversé.

Rupture sémantique. La sémantique peut être rompue si les traits définitoires ne sont pas transmis du terme père au terme fils. Par exemple, un *chat moulé* n'est pas un *chat* (un mammifère, qui...). Dans les données générées, nous avons trouvé plusieurs 'chats moulés' :

personnalite/personnalite compulsive
voix/voix œsophagienne

Personnalite décrit le comportement caractéristique d'un individu, tandis que *personnalite compulsive* correspond aux troubles psychiques. Les troubles (ou maladies) et le comportement sont des objets différents dans le MeSH. De la même manière, *voix* se rapporte à la physiologie humaine, tandis que *voix œsophagienne* aux traitements et à la thérapeutique.

La tête est ambiguë. Ce cas est proche du précédent, mais il est dû au choix des libellés des termes dans le MeSH. Les termes comme *absorption*, *investissement*, etc. sont polysémiques. Pour déterminer le sens précis, ils doivent être spécifiés par leurs contextes :

investissement/investissement (psychanalyse),
absorption/absorption cutanee,
goitre/goitre ovarien.

Investissement tout seul a un sens financier, tandis que dans *investissement (psychanalyse)*, ce mot reçoit un autre sens spécifique. De la même manière, *absorption* a un sens spécifique en chimie, et *goitre* employé tout seul correspond au désordre de la glande thyroïdale. Il faut noter que ces cas sont souvent non-ambigus dans la version originale du MeSH, en anglais. Par exemple, *investissement (psychanalyse)* en français est la traduction de *cathexis* en anglais.

Des situations similaires apparaissent lorsque le libellé du terme père est sous-spécifié :

acides/acides pentanoiques,
acne/acne rosacee.

Dans ces exemples, *acides* correspond à *acides inorganiques* et *acne* à *acne vulgaris*, mais la convention adoptée veut que le terme simple représente le concept correspondant.

Engagement ontologique. Les cas qui viennent sont intéressants car, bien qu'étant des relations hiérarchiques potentielles, ils sont absents dans le MeSH : les concepteurs du MeSH ont décidé de ne pas les encoder ou bien de les encoder différemment. Par exemple :

amyotrophies/amyotrophies spinales enfance,
hyperplasie/hyperplasie epitheliale focale,
centre public sante/centre public sante mentale,
rectocolite/rectocolite hemorragique,
penicillines/penicilline g.

Les terminologies médicales (par exemple MeSH, SNOMED), différencient les *signes et symptômes* et les *maladies* (état pathologique plus caractérisé). C'est le cas de *amyotrophies* et *hyperplasie* (*signes et symptômes*) *vs amyotrophies spinales enfance* et *hyperplasie epitheliale focale* (*maladies* du système nerveux ou de la bouche).

Pour des raisons que nous ignorons, le terme *centre public sante mentale* ne partage pas tous les attributs de *centre public sante*, ce qui les empêche d'être en relation hiérarchique dans le MeSH : ils sont frères. Il en est de même pour *rectocolite hemorragique* et *rectocolite*, qui sont tous les deux fils de *colite*.

Penicillines a été choisi dans le MeSH pour désigner une classe thérapeutique de médicaments (sous *antibiotiques*, sous *actions chimiques*), tandis que *penicilline g* est considéré comme une *substance chimique*.

La structuration impliquée dans ces exemples reflète l'engagement ontologique des concepteurs du MeSH, qui ne peut pas être induit avec les approches automatiques et donc les inclusions lexicales.

Erreur de traduction. La relation induite entre *acide linoleique* et *acide linoleique alpha* se trouvait parmi les erreurs. L'analyse de la structure du MeSH en anglais et les recherches effectuées sur la page Web du MeSH ont alors montré que le terme anglais *alpha linolenic acide* traduit par *acide linoleique alpha* aurait dû être traduit par *acide linolenique alpha*.

Ensemble expansion

Dans la majorité des cas, lorsque le terme père se trouve en position d'expansion (position finale) dans le terme fils, nous supposons que la tête sémantique du terme fils est modifiée. La relation induite n'est pas donc une relation hiérarchique.

Relation partitive. Plusieurs relations induites sont des relations partitives (partie du corps, continent et son peuple, substances chimiques) :

abdomen/muscle droit abdomen,
amerique centrale/indien amerique centrale,
argent/nitrate argent.

Relations spécifiques. Sinon, ce sont des relations plus spécifiques, que nous avons essayé de caractériser :

virus myxome `causé-par` *myxome*,
comportement alimentaire `atteint-par` *troubles comportement alimentaire*,
bovin `localise` *pneumonie interstitielle atypique bovin*,
hopital `a-pour-caractéristique` *capacite lits hopital*,
services sante `a-pour-état` *fermeture service sante*,
macrophage `a-pour-état` *activation macrophage*.

Notons que certaines de ces relations impliquent des noms et les dérivations adjectivales dénominales correspondantes :

cubitus `contient` *nerf cubital*,
genes `subit-analyse` *epreuve complementation genetique*.

Énumérations. Dans cet ensemble, nous avons également des cas d'énumération (voir plus haut, page 155) :

immunodepresseurs / antineoplasiques et immunodepresseurs.

Ambiguïté syntaxique. Dans le cas d'ambiguïté syntaxique relevé, deux termes (*oncogene* et *antigene viral oncogene*) comportent le mot *oncogene*, qui est un nom dans le premier terme et un adjectif dans le second.

Ensemble 'autre'

Dans cet ensemble, le terme père semble se trouver dans des positions autres que tête ou expansion. Il peut alors être non-contigu dans le terme fils et contenir des modifieurs ou autres éléments insérés :

bacterie aerobie/bacterie gram-negatif aerobie,
sommeil/troubles sommeil extrinseques,
dent/implantation dentaire sous-periostee,
diphosphate/uridine diphosphate acide glucuronique.

Les termes 'pères' de ces exemples sont en réalité en position tête ou expansion. Mais comme nous n'effectuons pas d'analyse syntaxique des termes, ces cas restent sous-spécifiés dans l'ensemble 'autre'.

Homographe. Nous avons détecté un cas d'homographie due à l'absence de caractères accentués dans le MeSH français. Dans *cilie/cellule ciliee externe* les mots *cilie* (nom, organisme invertébré) et *ciliee* (forme fléchie de l'adjectif *cilie*, qui caractérise un type de cellule) sont regroupés avec la lemmatisation. Dans la version accentuée du MeSH l'adjectif serait noté *cilié*, ce qui le rendrait non ambigu.

Erreurs

Et enfin, nous avons détecté des cas d'erreurs, dont certaines sont dues aux dérivations. Deux mots peuvent être correctement reliés par une relation dérivationnelle dans l'absolu, mais lorsqu'ils apparaissent dans les termes cette relation peut perdre (une partie de) sa pertinence :

contracture/contraction musculaire,
biologie/testament biologique,

ici, *contracture* et *contraction* ont été correctement reliés par une relation dérivationnelle, mais dans ces exemples *contracture* signifie une maladie du muscle et *contraction* le fonctionnement normal du muscle. Dans le deuxième exemple, l'adjectif *biologique* est dérivé du nom *biologie*, mais son sens s'avère être plus spécifique que le sens de *biologie*.

Bilan

L'analyse des relations hors-MeSH montre, entre autre, que les relations hiérarchiques ne sont pas les seules à être proposées à travers notre application de l'hypothèse des inclusions lexicales. Des relations plus spécifiques du domaine émergent dans les ensembles *expansion* et 'autre'. D'autre part, toutes les relations induites, même potentiellement correctes, ne satisfont pas toujours les choix de modélisation du domaine ni l'engagement ontologique des concepteurs du MeSH. Ces choix ontologiques ne peuvent pas être induits avec des méthodes automatiques et correspondent à une part humaine lors de la constitution des terminologies.

Parmi les améliorations possibles de la méthode, le recours au réseau sémantique d'UMLS pourrait être utile pour faciliter le typage des relations et leurs sélection. L'étiquetage morphosyntaxique (des termes à structurer et des ressources linguistiques) devrait également être utilisé pour limiter la génération des relations erronées (Jacquemin & Tzoukermann, 1999). La réaccentuation des termes du MeSH constitue une autre amélioration possible des résultats. Notons que la réaccentuation a été entreprise (Zweigenbaum & Grabar, 2002a), et que, de manière indépendate, l'INSERM propose, depuis 2004, le MeSH sous forme accentuée.

7.6.4 Structuration des termes provenant du corpus de la cogénération

Nous avons également appliqué notre méthode de structuration, à côté d'autres méthodes, aux termes provenant d'un corpus sur la cogénération (Grabar & Jeannin, 2002)[8]. L'évaluation des relations induites a été effectuée manuellement, car il n'existe pas de référentiel dans ce domaine.

La méthode est appliquée à quatre sous-corpus distingués dans le domaine de la cogénération : écologie, économie, règlementation et technique. Nous avons testé les inclusions lexicales avec une normalisation de *base* : (1) nous ne disposons pas de ressources morphologiques complètes et validées du domaine et (2) les termes traités n'ont pas été validés, ce qui augmente le risque de génération de bruit. Il est effectivement très élevé (tab. 7.3). Ce tableau montre également que le nombre de relations induites avec les inclusions lexicales est très variable selon les corpus et que le taux de rejet des relations est important. Le domaine technique présente ainsi la meilleure précision avec seulement 88,3 % de bruit. Pour faciliter les validations, il serait intéressant de séparer les relations induites en des

[8]La cogénération est une technique de production combinée d'électricité et de chaleur (eau chaude, vapeur). Ce travail a été réalisé dans le cadre du projet national SAFIR visant à concevoir un métamoteur de recherche sémantique.

ensembles tête, expansion et autre. Nous nous attendons à ce que l'ensemble tête fournisse le plus de relations pertinentes.

Domaine	Candidats termes	Relations induites	Relations sélectionnées	% de bruit
Écologie	18 865	17 702	940	94,7 %
Économie	24 747	1 309	73	94,4 %
Règlementation	18 595	33 374	172	99,5 %
Technique	16 553	17 353	2 135	88,3 %

TAB. 7.3 – Induction des relations avec les inclusions lexicales et leur validation.

	anto	caract.	conduit	hypero	méro	produit	syno	utilise	induit	punit	règle	consomme
Écologie	7	137	3	547	39	20	4	161	-	1	10	3
Économie	3	13	-	49	-	-	-	-	-	-	8	-
Règlementation	-	53	-	95	-	-	-	-	2	5	24	-
Technique	4	33	9	1 852	70	35	30	102	-	-	-	-

TAB. 7.4 – Distribution des relations typées par sous-domaines.

Nous avons en même temps typé manuellement les relations (Grabar & Jeannin, 2002). Le résultat de ce typage est présenté dans le tableau 7.4. Pour chaque domaine nous indiquons les types de liens et le nombre de relations correspondantes. Les relations hyperonymiques (`hypero`) correspondent généralement à la majorité des relations sélectionnées. Si elles sont présentes dans chaque corpus, elles ne sont pas les seules à avoir été distinguées : il existe aussi la relation partitive (`méro`), les relations lexicales (`anto` et `syno`) et un ensemble de relations transversales (`produit`, `utilise`, `conduit`, etc.). La relation `caractérise` regroupe des relations sous-spécifiées : elle est proche du traditionnel `voir aussi`.

Bilan

L'application de la méthode à des termes non validés montre une précision très faible. Il serait donc intéressant d'effectuer une validation des termes avant de chercher à les structurer. Les inclusions lexicales ont permis de détecter un ensemble de 12 relations parmi lesquelles des relations taxinomiques, lexicales et transversales. Cette expérience a montré également l'intérêt de l'utilisation de différentes approches de structuration des termes : elles s'avèrent être complémentaires (Kavanagh, 1995). Mais n'oublions pas que chaque approche de structuration demande un travail de validation et d'analyse conséquent.

7.7 Conclusion et perspectives

Nous avons présenté une application de l'hypothèse des inclusions lexicales pour la détection des relations hiérarchiques entre les termes. Nous supposons que le terme P est un père éventuel du terme fils F si tous les mots de P apparaissent dans F. Les tests sont effectués avec des normalisations linguistiques successives (morphologiques et synonymiques). Nous appliquons notre méthode aux termes d'une terminologie existante, le MeSH, et comparons les relations induites avec celles encodées dans la structure originale du MeSH. Cette évaluation montre ainsi qu'avec l'injection de ressources linguistiques et selon le schéma de normalisation, jusqu'à 43,8 % de toutes les relations induites et jusqu'à 14 % de relations directes se trouvent dans le MeSH. Et jusqu'à 65 % de termes du MeSH sont placés correctement sous leur ancêtre. Ces chiffres font apparaître qu'il est plus facile de placer un terme que de le placer dans tous les bons endroits de la hiérarchie. Nos expériences ont montré que la précision des ressources linguistiques, morphologiques ou synonymiques, est meilleure lorsqu'elles sont spécifiques au domaine traité. Mais les ressources de la langue générale permettent d'augmenter le rappel (Hamon *et al.*, 1998).

À côté des relations hiérarchiques, la méthode propose d'autres types de relations : partitive, lexicales mais surtout transversales. Toutes les relations induites, même potentiellement correctes, n'apparaissent pas dans les terminologies : l'engagement ontologique, qui définit les connaissances à encoder, sous-entend une décision humaine que les approches automatiques ne peuvent pas induire.

La précision que nous obtenons n'est pas très élevée. Une amélioration possible pourrait être obtenue grâce à l'utilisation des données étiquetées (termes et ressources linguistiques), ce qui permettrait de spécifier les mots syntaxiquement et d'effectuer une analyse des dépendances syntaxiques dans les termes. Ce dernier point conduirait par ailleurs à un véritable traitement des termes complexes, contrairement au traitement des sacs de mots que nous faisons actuellement. L'utilisation du MeSH accentué est une autre source d'amélioration. Notons qu'une initiative de son accentuation a été entreprise dans les travaux de (Zweigenbaum & Grabar, 2002b). Par ailleurs, l'INSERM, le traducteur officiel du MeSH en français, en fournit maintenant une version accentuée. Si nous disposons de ressources morphologiques plus complètes, les résultats présenteront un meilleur rappel. Pour compléter ces ressources, les approches à partir de corpus (Zweigenbaum *et al.*, 2003b) constituent une des pistes, de même que l'application parallèle des ressources morphologiques et des règles de désuffixation pour le traitement des variations flexionnelles régulières.

Les termes du MeSH sont des expressions artificielles créées par les concepteurs du MeSH afin de représenter le mieux possible les concepts médicaux. Le traitement de telles expressions avec des moyens automatiques présente parfois des difficultés. Nous avons également appliqué la méthode aux termes de la SNOMED, qui eux sont plus proches des expressions trouvées dans les documents textuels. Par contre, nous n'avons pas pu évaluer complètement les relations induites : la structure hiérarchique de la SNOMED est plus difficile à décoder à travers ses identifiants. Une des difficultés, il nous semble, vient du fait que, pour le nombre de concepts recensés dans la SNOMED, sa hiérarchie n'est

pas très profonde. Ainsi, les concepts *oreillette* (T-32100), *oreillette droite* (T-32200) et *oreillette gauche* (T-32300) sont tous fils de *cœur* (T-32000). Alors qu'avec notre méthode, nous induisons les relations hiérarchiques *oreillette / oreillette droite* et *oreillette / oreillette gauche*. Mais nous pouvons nous trouver ici aussi face à un engagement ontologique des développeurs. Dans tous les cas, il serait intéressant de pousser plus en avant l'évaluation des relations obtenues avec les termes de la SNOMED. D'autant plus que la SNOMED propose des relations transversales.

Pour l'adaptation de ressources synonymiques générales au corpus, il serait aussi intéressant d'appliquer l'approche distributionnelle (Nazarenko *et al.*, 2001), ce qui permettrait de les filtrer à travers leur substituabilité paradigmatique. Alors que les méthodes appliquées jusqu'ici (patrons lexico-syntaxiques, marqueurs, associations des mots dans une fenêtre donnée) exploitent les relations syntagmatiques entre les mots.

Dans nos expériences, nous avons aussi calculé un poids de mise en relation de deux termes pour évaluer la 'distance sémantique' entre eux. Nous nous sommes inspirés alors de (Aronson, 2001) : les normalisations de base ont un coût nul, la lemmatisation 0,1, la racinisation 0,3 et les synonymes 0,5. Nous disposons donc d'une 'distance' entre deux termes mis en relation. Pour le moment, nous n'avons pas exploité cette information. Elle pourrait fournir une limite au-delà de laquelle les relations seraient analysées avec plus de précaution. Mais il est possible que ce genre de limite soit surtout utile dans les applications automatiques comme la recherche d'information. C'est d'ailleurs dans ce contexte que (Aronson, 2001) la met en place.

Notre méthode, moyennant de faibles modifications, a été adaptée à la détection des termes équivalents et contraires (Grabar & Hamon, 2004b). Elle génère alors des relations similaires aux variations morphosyntaxiques (Jacquemin, 1996) et aux termes contraires (Daille, 2003). Par rapport à d'autres applications de la même hypothèse, notre méthode n'est pas liée par une liste déterminée de modifieurs comme dans (Bodenreider *et al.*, 2001), et, contrairement à (Bourigault, 1993), elle vise à proposer des relations sémantiques entre termes. Toujours contrairement à ce dernier travail, notre méthode n'effectue pas actuellement une analyse syntaxique des termes. Par rapport à ces deux travaux, nous générons plus de propositions, au risque d'une diminution de la précision. Mais le point fort de notre méthode est la possibilité d'application des normalisations linguistiques.

Les inclusions lexicales, comme d'autres méthodes, sont sensibles à certains types de relations apparaissant dans certains contextes. Pour obtenir une structure des termes plus complète, il est plus intéressant d'appliquer cette hypothèse en combinaison avec d'autres méthodes (Kavanagh, 1995; Grabar & Jeannin, 2002).

Chapitre 8

De l'utilité des ressources morphologiques pour le traitement de la variation terminologique

À travers nos travaux où nous utilisons et évaluons des ressources morphologiques acquises à partir de termiologies médicales (mise en relation des vocabulaires de l'indexation et des utilisateurs, structuration des termes), nous montrons que ces ressources sont bénéfiques pour le traitement de la variation terminologique. Cette constatation est assortie toutefois de plusieurs remarques relatives à l'acquisition de ressources morphologiques et à leur utilisation, que nous reprenons ici. Elles peuvent être vues comme des recommandations vis-à-vis des traitements automatiques dans lesquels un appariement des termes et/ou expressions est visé.

8.1 Complémentarité des différents niveaux de traitement

Nous avons distingué plusieurs types de variations terminologiques qui se réalisent à différents niveaux linguistiques, dont la casse, l'orthographe, la morphologie, la syntaxe, la synonymie, etc. Chacun de ces types de variation demande des méthodes de traitement et des ressources spécifiques, chaque variation n'étant accessible que lorsque la connaissance requise est disponible. Par conséquent, pour obtenir des résultats plus complets dans l'appariement des termes et expressions, différentes méthodes et ressources doivent être utilisées de manière complémentaire.

Certaines méthodes apportent des gains très importants dans les appariements. C'est le cas de certaines normalisations au niveau des caractères (casse, accentuations) ou au niveau des mots (ordre ds mots). En plus, les efforts fournis pour leur traitement sont minimes. Ce sont des traitements peu coûteux, peu ambigus et efficaces. Vu la simplicité de ces méthodes, elles devraient être considérées comme la base des traitements de la variation

terminologique, surtout dans les contextes que l'on sait instables et non soignés de ce point de vue (requêtes des utilisateurs, sources non accentuées, certains types de documents du Web (messages, forums, sites personnels), etc.).

Les traitements se compliquent lorsque les variations dépassent le niveau formel des termes et touchent à la sémantique. Là, des ressources externes sont nécessaires. Il en est ainsi par exemple, pour les variations morphologiques des termes et la synonymie. Deux questions se posent alors. D'une part, si l'on veut traiter ce genre de variations il faut disposer des ressources nécessaires. Et si elles ne sont pas disponibles, ce qui est souvent le cas, il faut disposer de moyens et de sources pour les constituer. Notons aussi que parfois les ressources en question sont disponibles mais ne sont pas forcément adaptées aux données et/ou à la tâche. Dans ce cas, la raison voudrait que leur utilisation efficace soit conditionnée par leur adaptation. D'autre part, comme il n'existe pas d'unanimité sur l'utilité des ressources qui impliquent la sémantique, l'on peut se poser justement la question sur leur utilisation, leur efficacité et surtout sur leur apport. Face aux efforts de constitution ou d'adaptation de ces ressources, elles risquent de (re)devenir ambiguës une fois reprojetées sur les textes. Nos expériences ont aussi montré qu'en arrivant aux normalisations morphologiques, nous entrons dans un schéma classique de la résolution de problèmes : les efforts nécessaires pour améliorer les résultats sont de plus en plus importants pour un gain qui va en diminuant. La contribution des connaissances morphologiques est très faible comparée à celle des normalisations au niveau des caractères, alors que le recensement des ressources morphologiques demande justement davantage d'effort. Le questionnement sur l'évaluation de l'apport de chaque normalisation pousse aussi à se demander si les mesures de rappel et de précision et les conditions de leur application y sont adaptées. Car, sans les normalisations, les systèmes automatiques resteraient aveugles face à plusieurs types de variation de termes. Mais en même temps, l'injection des connaissances requises augmente le 'risque' d'effectuer de mauvais appariements, ce qui est reflété de manière générale par la diminution de la précision. Nous pensons donc qu'il n'est pas très judicieux d'évaluer de la même manière des appariements bruts et des appariements obtenus avec des normalisations. Une étape de filtrage pourrait ainsi être prévue avant l'évaluation du rappel et de la précision des données normalisées.

Chaque connaissance supplémentaire injectée donne ainsi accès à des appariements nouveaux. La morphologie permet d'effectuer des appariements là où d'autres méthodes et ressources ne peuvent pas agir. Il faut ainsi disposer de ressources morphologiques pour mettre en relation les termes dont les mots présentent des variations de nombre, de genre, d'adjectivation, etc. La même remarque doit néanmoins être prise dans l'autre sens : si la morphologie conduit aux appariements, elle n'a abolument pas accès à toutes les variations (Light, 1996) et doit donc être combinée avec d'autres informations.

8.2 À chaque variation morphologique ses ressources

De la même manière que différents types de variation sont accessibles à différentes méthodes et ressources, différentes ressources morphologiques donnent accès chacune à un

type donné de variation.

Nous l'avons noté : la morphologie met en œuvre plusieurs procédés pour la formation des lexèmes et leurs variations. La morphologie constructionnelle recourt ainsi à l'affixation (ou dérivation), la conversion et la composition, parmi les procédés centraux. La supplétion et l'allomorphie peuvent également intervenir dans ces formations. La morphologie flexionnelle décrit les variations de nombre, genre, personne, temps, mode, etc. L'utilisation de ressources couvrant ces procédés conduit à la découverte des variantes de termes correspondantes. Il n'existe donc pas une seule ressource morphologique mais des ressources morphologiques requises pour une variation. Une raison de les distinguer est qu'elles n'assurent pas toutes le même apport sémantique dans les termes.

La flexion n'a ainsi qu'un faible apport et les lexicalisations de pluriels ou de féminins, même si elles existent, sont très rares dans ce domaine. La suffixation et la conversion ont un apport sémantique plus important. Elles appliquent de nombreuses instructions aux bases permettant entre autre de construire des adjectifs dénominaux (sens général étant « *Relatif à N* »), des noms déverbaux d'agent (« *Celui qui V* ») ou d'action (« *Action de V* »), des adjectifs déverbaux (« *Qui peut être V* »). Dans ces formations, les règles de construction et les affixes appliquent une instruction à la base, et comme seule la base véhicule le sens référentiel, la sémantique du lexème construit lui reste proche.

L'apport sémantique est beaucoup plus important en cas de préfixation et de composition, on peut même parler d'une rupture sémantique. Ce qui conduit vers la rupture tient à la nature des éléments combinés. Dans la composition, nous avons ainsi au moins deux éléments de composition qui sont porteurs chacun d'un sens référentiel. Lors de leur combinaison, qui est gérée par la règle morphologique, les deux sens référentiels sont articulés pour en former un nouveau. Selon la place des composants dans le lexème construit, nous pouvons avoir des relations d'hyperonymie ou des relations transversales : *gastroentérite* est ainsi `une sorte d'`*entérite*, *entérite* est `une sorte d'`*inflammation* `localisée dans` la *muqueuse de l'intestin grêle*. Quant aux préfixes actuels, ils sont nourris par des éléments aux antécédents différents : certains d'entre eux ont toujours connu des emplois préfixaux (*re-*, *dé-*, *dys-*), d'autres proviennent d'éléments de composition (*archéo-*, *pseudo-*) ou de prépositions et d'adverbes (*sur-*, *entre-*, *sans-*). Leur emploi en tant que préfixes témoigne donc de leur grammaticalisation, dont le degré varie selon les éléments morphologiques. Lorsque la grammaticalisation est avancée, la charge sémantique diminue, devient plus générale et donc proche d'une instruction. En attendant (que la grammaticalisation s'achève), la préfixation et la composition sont assez loin des relations construites avec l'affixation, la flexion ou la conversion, que l'on peut caractériser de proches. La préfixation et la composition doivent donc être utilisées avec précaution dans les applications automatiques où l'on doit assurer une continuité sémantique, en particulier dans le traitement de la variation terminologique.

Différentes ressources morphologiques seront donc utilisées selon que les applications visent à détecter des variations de termes très proches ou non. Un maximum d'information morphologique peut être utile là où le rappel est privilégié. Dans ces cas, la préfixation et la composition peuvent intervenir. Elles peuvent alors être comparées avec la relation d'*aboutness* (ou *voir aussi*) qui s'avère utile en recherche d'information. Par contre, lorsque

seules les relations d'équivalence sont recherchées, l'on devrait se satisfaire avec la flexion, (la majeure partie de) la suffixation et la conversion.

8.3 À chaque ressource ses heuristiques

La découverte de ressources morphologiques peut se faire de différentes manières. L'étude des grammaires et l'expertise des morphologues constituent l'une des pistes. Mais les applications informatiques ont souvent besoin de données linguistiques plus exhaustives que celles utilisées par les humains et il est alors nécessaire d'indiquer des relations morphologiques entre mots tout à fait banales et évidentes. C'est là que des méthodes d'acquisition automatique s'avèrent utiles car elles permettent d'effectuer un dépouillage morphologique plus systématique des sources. Les ressources découvertes sont alors à l'image des données de départ et de la sensibilité des outils d'acquisition.

La piste principale pour la découverte de mots reliés morphologiquement est basée le plus souvent sur la présence d'une sous-chaîne commune suffisamment longue et se trouvant souvent en première position. C'est la piste suivie dans nos expériences et dans d'autres travaux présentés. Le contexte sémantique de l'acquisition garantit la qualité des résultats. Si la découverte de mots aux sous-chaînes initiales communes permet de mettre au jour des relations morphologiques entre eux, cela ne permet pas de typer automatiquement ces relations. Des heuristiques supplémentaires sont alors nécessaires. Dans nos expériences, elles sont basées sur l'étiquetage morphosyntaxique et la lemmatisation, deux traitements linguistiques des données textuelles qui peuvent être considérés actuellement comme basiques.

La découverte et/ou le typage de la flexion relève ainsi directement de l'étiquetage et de la lemmatisation. Si le travail est fait à partir de données étiquetées, il suffit d'enregistrer les formes des mots qui apparaissent dans les textes et les lemmes correspondants. Notons que les erreurs d'étiquetage alors générées seront répercutées (et amplifiées) lors de l'application des ressources flexionnelles. Et comme les erreurs portent rarement sur les catégories mineures ou la ponctuation, mais sur les catégories majeures (noms, adjectifs, etc.), la correction de l'étiquetage peut alors être utile. Nous le faisons dans nos expériences avec les terminologies structurées (12 555 termes de la SNOMED, qui comportent 8 874 types). Mais lorsque l'on travaille sur des données plus importantes la correction de l'intégralité de ces données est inimaginable. Elle peut être concentrée alors sur un sous-ensemble (hapax, étiquetages multiples, noms, etc.). La flexion peut être aussi typée à travers les règles flexionnelles qui, reflétant la grammaire d'une langue, sont en nombre fini. On devra ensuite éliminer les erreurs, du point de vue synchronique, comme {*rat*, *rate*} ou {*chapeau*, *chapelle*}.

La conversion catégorielle peut être détectée uniquement sur des données étiquetées et lemmatisées. Mais elle est aisée à détecter, ceci d'autant plus si l'étiquetage et la lemmatisation sont corrects : les mots mis en relation doivent avoir la même forme, les affixes flexionnels, dont la finale des verbes à l'infinitif, mis de côté, et des catégories syntaxiques différentes.

Le typage des procédés d'affixation est plus inégal : il y en a plusieurs types et certains d'entre eux sont formellement proches de la composition. L'affixation transcatégorielle est ainsi repérable heuristiquement à travers le fait qu'elle permet de former des lexèmes d'une autre catégorie syntaxique que celle de la base. Quant à la composition, on peut considérer qu'elle met en relation des mots de la même catégorie grammaticale : principalement des noms dans les données médicales. Pour différencier les compositions des affixations iso-catégorielles, il est possible de se baser sur la longueur du 'suffixe'. Dans les compositions, où il correspond aux éléments lexicaux, il est souvent plus long. Rappelons néanmoins que *-ite*, *-ase* et autres sont courts mais composants quand même. Il reste aussi des cas où nous avons des relations morphologiques enchaînées entre les mots : adjectif ou forme au pluriel reliés à un composé sans que cette relation mentionne la forme de base des deux formations ({*arthropathies*, *arthropathique*}, {*carcinomateuse*, *carcinoïde*}).

De manière générale, l'exploration morphologique des données textuelles est souvent basée sur des heuristiques grossières et approximatives. La précision de l'étiquetage et de la lemmatisation laisse aussi à désirer. Tous ces facteurs peuvent causer des inductions erronées des systèmes d'acquisition automatiques. Une telle exploration morphologique des données textuelles propose donc un premier débroussaillage. Les résultats de ce débroussaillage peuvent servir tels quels, une fois validés, ou bien constituer le matériel d'une étude linguistique plus poussée, et être ensuite incorporés dans un modèle morphologique plus évolué.

8.4 Acquisition à partir de terminologies structurées ou de corpus ?

Les ressources morphologiques utilisées dans nos expériences ont été acquises à partir de terminologies structurées. Ce contexte sémantique permet d'induire des données d'une précision très élevée (toujours supérieure à 90 %) et d'un rappel intéressant (entre 79 et 91 % par rapport aux données en anglais encodées dans l'UMLS). Nous avons montré par ailleurs que cette méthode peut être appliquée à tout type de relations sémantiques entre termes. Les résultats obtenus avec ces différentes relations s'avèrent complémentaires, surtout dans la découverte des éléments de la composition. En plus d'une cohésion sémantique forte entre termes, les terminologies génériques, comme la SNOMED, le MeSH ou autre, offrent une couverture importante du domaine (Elkin *et al.*, 2001). Nous pouvons donc prétendre disposer des termes principaux, tout en sachant que les terminologies ne recensent pas tous les termes ni tout le lexique médical (Hersh *et al.*, 1997). Mais le frein principal de la méthode reste qu'elle est dépendante de la disponibilité des terminologies structurées qui constituent notre matériel d'amorçage.

Nous avons donc décidé de comparer les données induites avec notre méthode et les données induites avec une approche à partir de corpus textuels (Zweigenbaum *et al.*, 2003b). Les expériences décrites dans (Zweigenbaum *et al.*, 2003b) travaillent sur des données étiquetées et lemmatisées. Nous avons donc choisi l'expérience correspondante (sec. 4.4

page 77) à partir des terminologies structurées. Nous avons noté que la segmentation des mots en 'préfixes' et 'suffixes' peut varier en fonction des mots qui sont alignés et qu'un même mot peut être segmenté de différentes manières. Ces segmentations multiples, ou les 'allomorphies', constituent la principale difficulté lors des comparaisons. La première étape dans la préparation des données consiste donc à fusionner les éléments morphologiques identiques ou proches. Comme auparavant, pour fusionner les 'suffixes', nous utilisons l'algorithme de distance entre chaînes de caractères proposé dans (Jacquemin, 1997a) avec un seuil de proximité fixé à 0,5 (0 correspond aux chaînes identiques et au-dessus de 0,5, il nous semble, l'algorithme génère trop de bruit). Pour la fusion des 'préfixes' nous utilisons une des étapes de notre méthode (sec. 4.3.1 page 68) qui effectue la fusion des paires de mots en familles morphologiques. Le but des familles étant donc de regrouper des 'allomorphies' de 'préfixes' et les lexèmes formés avec. Le travail est fait avec les parties du discours accessibles à la morphologie : noms, adjectifs, verbes et adverbes.

Les résultats de la comparaison montrent que nous découvrons plus de 'préfixes' à partir de corpus : presque 4 000 familles contre 1 078 avec les terminologies structurées. Ce qui veut dire qu'en travaillant sur corpus nous découvrons, à travers la variété des bases, un plus grand vocabulaire. Quant aux 'suffixes', les dérivations semblent encore former le noyau commun. Les corpus présentent toutefois de très nombreuses variations dans leurs segmentations, de même que de nombreuses flexions verbales quasiment absentes des terminologies. La différence principale provient, comme dans la comparaison des données morphologiques générées à partir de différentes relations sémantiques entre termes, des éléments de composition. Ceux-ci sont plus variés dans les terminologies. Les corpus ont par ailleurs l'avantage de refléter plus immédiatement l'état actuel de la langue : nous trouvons des compositions cachées (*handiplace*) ou des mots-valises (*sidaction*)[1] qui sont absents des terminologies officielles. Pour résumer, les principales différences viennent des éléments morphologiques à sens référentiel (bases dans les affixations et élements de composition), des flexions verbales et des procédés périphériques de formation de lexèmes, comme la composition cachée et les mots-valises.

Avec les familles induites à partir des terminologies structurées, nous avons une précision de 97,1 % (31 familles erronées) ; avec les corpus, elle est de 76,8 % (913 familles incorrectes). En ce qui concerne la précision des couples de mots, elle avoisine 98 % avec les terminologies et se trouve entre 98 et 30 %, en fonction du rang des association des mots, dans les données générées à partir de corpus. Parmi les erreurs, nous avons les mêmes que celles signalées auparavant : chaînes initiales communes n'ayant pas de parenté sémantique, préfixes dérivationnels se trouvant en première position. Mais nous trouvons également des mots étrangers, des mots collés à d'autres mots (qui se trouvaient déjà dans les documents d'origine ou bien sont apparus lors de la conversion des documents `html` en texte), la césure, et finalement les erreurs qui proviennent des documents sources (accentuations, orthographe, etc.). L'étiquetage et la lemmatisation ajoutent une autre couche d'erreurs.

En résumé, les corpus présentent une plus grande diversité de bases, de formations périphériques comme les mots-valises ou la composition cachée, mais aussi plus de bruit

[1] Voir (Fradin, 2000; Fradin, 2003) pour une description de ces procédés.

(erreurs de génération, mots étrangers, erreurs d'orthographe, etc.). Ils demandent donc un travail plus important lors de la validation. Les corpus offrent par ailleurs des ressources morphologiques qui y sont plus fidèlement adaptées. Quant aux terminologies, offrant un contexte sémantique plus contraint, elles montrent une précision élevée et une assez grande couverture en fonction des domaines médicaux couverts. Elles sont plus généreuses en éléments de composition (mais il s'agit peut-être d'une particularité de la SNOMED). Les deux types d'approches conduisent vers la constitution de ressources morphologiques pour une langue et, plus spécifiquement, pour un domaine de spécialité, tout en en offrant des vues supplémentaires.

8.5 Ressources linguistiques ou règles de désuffixation ?

En revenant sur le débat sur l'efficacité des normalisations morphologiques par rapport à la *baseline*, nous pensons avoir montré à travers nos expériences que ces normalisations sont nécessaires si l'on veut découvrir les variantes de termes qui se produisent au niveau morphologique. Nous discutons ici les contributions relatives de ressources morphologiques motivées linguistiquement, qui se présentent souvent sous forme de paires de mots, et de règles de désuffixation. Les différentes expériences citées auparavant montrent que les résultats dépendent de plusieurs facteurs.

Tout d'abord, la longueur des requêtes et des documents exerce une influence : les documents et requêtes courts, n'offrant pas beaucoup d'autres amorces dans l'appariement, sont plus sensibles à l'injection de connaissances morphologiques.

L'apport relatif des connaissances morphologiques varie aussi en fonction du type flexionnel des langues : les langues avec un système morphologique relativement simple peuvent être prises en charge par les désuffixeurs, tandis que les langues avec un système morphologique complexe sont difficilement descriptibles avec les règles de désuffixation, elles sont alors plus favorables à l'utilisation de ressources morphologiques. Pour une langue comme le français, nos expériences montrent qu'il existe une complémentarité des règles et des ressources linguistiques. Les règles sont efficaces dans le traitement des formations flexionnelles régulières, comme les pluriels. Par contre, en affixation où il existe des irrégularités morphologiques, sémantiques ou autres, des allomorphies et supplétions, il est préférable de recourir aux ressources motivées linguistiquement. Lors de l'acquisition de ressources morphologiques à partir de corpus (Zweigenbaum *et al.*, 2003b) nous nous sommes aperçus ainsi que la précision des données varie en fonction des règles : certaines d'entre elles ne génèrent pas d'erreurs, ou très peu ; tandis que d'autres présentent une très faible précision. Notons à ce sujet l'expérience d'établissement de relations morphologiques entre les *synsets* dans `WordNet` (Fellbaum & Miller, 2003), donc sur des lexèmes anglais (langue considérée comme peu riche morphologiquement). Dans cette expérience, la détection des affixations est effectuée en lançant un `grep` sur les affixes traités. Les affixations sont ensuite validées et reliées manuellement à leurs bases. Les auteurs justifient cette démarche par le fait que les programmes informatiques ne peuvent pas traduire les règles de la langue et encore moins la connaissance lexico-sémantique beaucoup trop

irrégulière. Il s'agit sans doute d'une position un peu extrêmiste vis-à-vis des traitements automatiques et du TALN. Il peut être plus raisonnable de les considérer comme une aide dans l'exploration systématique des données textuelles. Cette aide s'avère d'autant plus efficace que les programmes informatiques sont contraints quant aux objectifs 'demandés'.

L'évaluation de l'apport des normalisations morphologiques dépend aussi des principes et mesures utilisés. Nous nous sommes demandée ainsi si les mesures de précision et de rappel et les conditions de leur application reflètent fidèlement cet apport. (Hull, 1996) va plus loin et effectue plusieurs modifications lors de l'évaluation : (1) Le rappel et la précision sont observés sur les 5 à 15 premiers documents, qui représentent une recherche superficielle, et les 50 à 150 premiers documents qui correspondent à une recherche fouillée ; (2) La priorité dans l'analyse des résultats est donnée aux requêtes les plus sensibles à la différence des traitements morphologiques appliqués. Suite à ces modifications, il s'avère que les ressources morphologiques, surtout celles concernant la suffixation, sont plus utiles que la désuffixation.

Notons aussi que les règles de désuffixation sont plus exhaustives et s'avèrent plus utiles lorsque le rappel est privilégié. Mais il est vrai qu'elles effectuent un traitement 'agressif' des données. Les listes d'exceptions aux règles générales constituent donc une précaution utile. L'existence de ces listes sous-entend déjà une analyse et une validation de ressources morphologiques. Les exceptions devront par ailleurs être adaptées aux nouvelles données traitées. Quant aux ressources morphologiques, elles demandent plus d'effort lors de leur constitution. Mais elles effectuent des normalisations plus fines.

8.6 Ressources spécifiques du domaine ou de la langue générale ?

Nos expériences de structuration de termes montrent que les ressources spécifiques au domaine (morphologiques et synonymiques) permettent de relier les termes avec une meilleure précision. Les ressources morphologiques spécifiques au domaine peuvent donner également un rappel plus élevé (comme dans notre expérience de structuration de termes). Quant au rappel avec les synonymes, il est plus élevé avec les ressources génériques. Lorsque des ressources spécifiques au domaine n'existent pas ou ne sont pas accessibles, les ressources génériques peuvent être adaptées à ce domaine, comme nous le faisons avec les synonymes. Cet effort supplémentaire est alors recompensé par un rapport précision / rappel plus intéressant. Les expériences de (Burgun & Bodenreider, 2001; Bodenreider & Burgun, 2002; Bodenreider *et al.*, 2003) montrent que lors de l'adaptation de ressources générales, elles peuvent également être enrichies de notions, de relations et surtout de termes spécifiques au domaine.

Chapitre 9

Perspectives

Les perspectives principales concernent les améliorations pouvant être apportées à l'acquisition de ressources morphologiques : leur qualité et leur couverture. Nous reprenons ici quelques unes de celles déjà signalées. Il serait ainsi intéressant de pouvoir segmenter les mots en sous-chaînes plus proches des éléments morphologiques et donc plus directement utilisables dans différentes applications. Sans oublier qu'un mot (typographique ou informatique) peut être composé de plus de deux élements. L'application de la méthode aux langues d'autres familles (langues sémitiques par exemple) ou bien dans un contexte multilingue (acquisition de cognats) est un autre pôle d'investigation. La combinaison de notre méthode avec les approches à partir de corpus constitue également un facteur potentiel d'amélioration. Le recours à des méthodes de pondération des résultats est dans tous les cas souhaitable et permettra d'aborder avec plus d'aisance les validations. Par ailleurs, les corpus textuels offrent toujours une source d'information pour des études ponctuelles des procédés morphologiques, qu'elles aient une visée linguistique ou applicative.

Mais les perspectives vont bien sûr au-delà des explorations morphologiques des données textuelles. Elles concernent aussi les domaines applicatifs abordés dans notre travail : indexation et recherche des documents, catégorisation, structuration des terminologies. La principale perspective est alors la prise en compte des dépendances syntaxiques dans les termes complexes, ce qui nous permettrait de dépasser le niveau de traitement du mot. Vu tout ce que nous avons rapporté, une comparaison plus poussée et complète des règles de désuffixation et des ressources morphologiques s'impose. Les ressources morphologiques devront alors être complétées. Et finalement, un de nos intérêts particuliers concerne la constitution et la structuration des terminologies. D'une part en ce qui concerne le repérage des relations entre termes et leur typage. D'autre part l'aide aux validations qui fait, de manière générale, défaut dans ce domaine.

Bibliographie

Agirre, E., Arregi, X., Artola, X., de Llarraza, A. D., & Sarasola, K. (1994). Intelligent dictionary help systems. In M. Brekke, I. Andersen, T. Dahl & J. Myking, Eds., *Applications and implementations of current LSP research*, Norway.

Aliquot-Suengas, S. (1996). *Référence collective/Sens collectif. La notion du collectif à travers les noms suffixés du lexique français.* Thèse de doctorat en linguistique, Université de Lille III, Lille, France.

Amiot, D. (2001). Préfixes ou prépositions ? Les cas de *sur-*, *sans-*, *contre-* et les autres. *Lexique*, **16**.

Amiot, D. (2004). *Des préfixes aux prépositions. Théorie morphologique et grammaticalisation.* Habilitation à diriger la recherche, Université d'Artois, Arras.

Amsili, P. (2003). L'antonymie en terminologie : quelques remarques. In *Terminologie et Intelligence Artificielle (TIA)*, pp. 31–40, Strasbourg.

Aronoff, M. (1976). *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Aronoff, M. & Anshen, F. (1998). *Morphology and the lexicon: lexicalisation and productivity*, In *The handbook of morphology*, pp. 237–247. A. Spencer and A.M. Zwicky.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

Arppe, A. (2002). The usage patterns and selectional preferences of synonyms in a morphologically rich language. In *Journées de traitement automatique des données textuelles (JADT)*, Saint Malo.

Assadi, H. (1998). *Construction d'ontologies à partir de textes techniques – Application aux systèmes documentaires.* Thèse de doctorat en informatique, Université de Paris 6, Paris, France.

Baayen, H. (1991). Quantitative aspects of morphological productivity. *Yearbook of Morphology*, pp. 109–149.

Baayen, H. (1992). On frequency, transparency and productivity. *Yearbook of Morphology*, pp. 181–208.

Baayen, H. (1994). Derivational productivity and text typology. *Journal of quantitative linguistics*, **1**(1), 16–34.

BAAYEN, H. (2001). *Word frequency distributions*, volume 18 of *Text, Speech and Language Technology*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

BAAYEN, H. & LIEBER, R. (1991). Productivity and English derivation: a corpus-based study. *Linguistics*, **29**(5), 801–843.

BAAYEN, H. & RENOUF, A. (1996). Productive lexical innovations in an English newspaper. *Language. Journal of the Linguistics Society of America*, pp. 69–98.

BAKKEN, S. (1999). On a path to a reference terminology for nursing concepts: the science and synergy. In C. G. CHUTE, Ed., *International Medical Informatics Assiciation Working Group 6 (IMIA WG6)*, Phoenix, Arizona.

BARBAUD, P. (1991). Fondements grammaticaux de l'acquisition des mots composés. *Canadian Journal of Linguistics*, **36**(3), 215–253.

BAUD, R. H., LOVIS, C., RASSINOUX, A.-M., MICHEL, P.-A. & SCHERRER, J.-R. (1997). Extracting linguistic knowledge from an international classification. In *Medical Informatics in Europe (MIE)*.

BAUER, L. (1998). Is there a class of neoclassical compounds, and if so is it productive? *Linguistics*, **36**(3), 403–422.

BERCHE, A., MOUGIN, F., HATHOUT, N. & LECOMTE, J. (1997). *Verbaction : constitution d'un lexique déverbal du français*. Rapport technique, INIST.

BIBER, D. (1994). Representativeness in corpus design. *Linguistica Computazionale*, **IX-X**, 377–408. Current Issues in Computational Linguistics: in honor of Don Walker.

BLANQUET, A. & ZWEIGENBAUM, P. (1999). A lexical method for assisted extraction and coding of ICD-10 diagnoses from free text patient discharge summaries. *Supplementary for the Journal of American Medical Informatics Association*.

BLOCH, O. & VON WARTBURG, W. (2002). *Dictionnaire étymologique de la langue française*. Paris: PUF.

BODENREIDER, O. (2001). Circular hierarchical relationships in the UMLS: etiology, disgnosis, treatment, complications and prevention. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

BODENREIDER, O. & BURGUN, A. (2002). Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In *Proceedings of the First Global WordNet Conference*, pp. 223–230.

BODENREIDER, O., BURGUN, A. & MITCHELL, J. A. (2003). Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In *Medical Informatics in Europe (MIE)*, pp. 379–384.

BODENREIDER, O., BURGUN, A. & RINDFLESCH, T. C. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In URI INIST CNRS, Ed., *Terminologie et Intelligence artificielle (TIA)*, pp. 11–21, Nancy.

BODENREIDER, O., MITCHELL, J. A. & MCCRAY, A. T. (2002). Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, pp. 61–65.

BODENREIDER, O., NELSON, S. J., HOLE, W. T. & CHANG, F. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In C. G. CHUTE, Ed., *A paradigm shift in health care information systems: clinical infrastructures for the 21st certury. Annual Symposium of the American Medical Informatics Association (AMIA)*, pp. 815–819, Orlando, Philadelphia. Disponible à `http://www.nlm.nih.gov/mesh/beyond.html`.

BODENREIDER, O. & ZWEIGENBAUM, P. (2000). Stratégies d'identification des noms propres à partir de nomenclatures médicales parallèles. *Traitement automatique des langues (TAL)*, **41**(3), 727–758.

BOLINGER, D. L. (1948). *On defining the morpheme*, In *Forms of English, Accent, Morpheme, Order*, pp. 183–189. Harvard University Press: Cambridge, MA.

BONAMI, O., BOYÉ, G. & KERLEROUX, F. (2005). *L'allomorphie radicale et la relation flexion-construction.*

BOUCHERIE, A. (1876). Bibliographie. *Revue des langues romanes*, **II**, 264–275.

BOUCHÉ, P. (1994). *Les mots de la médecine.* Paris: Belin.

BOURIGAULT, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *Traitement Automatique des Langues (TAL)*, pp. 105–117.

BOURIGAULT, D. & FABRE, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, **25**, 131–152.

BOURIGAULT, D. & HABERT, B. (1998). Evaluation of terminology extractors: Principles and experimentation. In *Language Resources and Evaluation (LREC)*, pp. 299–305, Grenade.

BOURIGAULT, D. & SLODZIAN, M. (1999). Pour une terminologie textuelle. In *Terminologie et Intelligence Artificielle (TIA)*, Nantes.

BOUSQUET, C., JAULENT, M.-C., CHATELLIER, G. & DEGOULET, P. (2001). Réutilisation de la nomenclature multi-axiale SNOMED pour mesurer une distance sémantique entre termes médicaux. In URI INIST CNRS, Ed., *Terminologie et Intelligence artificielle (TIA)*, pp. 139–148, Nancy.

BRILL, E. (1993). *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, Philadelphia. Disponible à `http://www.cs.jhu.edu/~brill/acadpubs.html`. Visité le 23/08/99.

BRILL, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, **21**(4), 543–565.

BRUNET, E. (1981). *Les suffixes*, In *Le vocabulaire français de 1789 à nos jours. D'après les données du Trésor de la langue française*, pp. 415–493. Librairie Slatkine.

BRUNET, E. (1999). *Aperçu statistique sur l'évolution du vocabulaire français*, In *Nouvelle histoire de la langue française*, pp. 675–627. Éditions du Seuil.

BURGUN, A. & BODENREIDER, O. (2001). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *Proceedings of NAACL Workshop WordNet and Other lexical Ressources: Applications, Extensions and Customizations*, pp. 77–82.

BURNAGE, G. (1990). *CELEX - A Guide for Users*. University of Nijmegen: Centre for Lexical Information.

BÉGUIN, A., JOUIS, C. & MUSTAFA, W. (1997). Évaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus. In AUPELF-UREF & FMA, Eds., *Actes des Premières Journées Scientifiques et Techniques (JST)*, pp. 419–126, Avignon.

CALZOLARI, N. (1991). Acquiring and representing semantic information in a lexical knowledge base. In J. PUSTEJOVSKY, Ed., *Proceedings of the ACL workshop on lexical semantics and knowledge representation (SIGLEX)*, pp. 188–197.

CHURCH, K. W. (1995). One term or two? In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pp. 310–318, Seattle, US.

CHUTE, C. G. (1995). Medical concept representation: from classification to understanding. In *Annual Synposium on Computer Applications in Medical Care (SCAMC)*, Washington. Tutoriel T33.

CIMINO, J. (2001). Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

CLARK, H. H. & CLARK, E. V. (1979). When nouns surface as verbs. *Language*, **55**, 767–811.

CONDAMINES, A. & REBEYROLLE, J. (1998). CTKB : A corpus-based approach to a terminological knowledge base. In *Proceedings of Computerm'98 (First Workshop on Computational Terminology)*, pp. 29–35, Coling-ACL'98, Université de Montréal, Montréal, Quebec, Canada.

CORBETT, G. G. (1991). *Gender*. Cambridge: Cambridge University Press.

CORBETT, G. G. (2000). *Number*. Cambridge: Cambridge University Press.

CORBIN, D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 1. Lille: Presse universitaire de Lille.

CORBIN, D. (1991). La formation des mots : structures et interprétations. *Lexique*, **10**, 7–30.

CORBIN, D. (1992). Hypothèse sur les frontières de la composition nominale. *Cahiers de grammaire*, **17**, 25–55.

CORBIN, D. (1993). French (indo-european : Romance). *Encyclopédie internationale de morphologie*, **121**.

Corbin, D. (1994). Locutions, composés, unités polylexématiques : lexicalisation et mode de construction. In *La locution*, pp. 53–101, ENS Fontenay/Saint-Cloud.

Corbin, D. & Paul, J. (1999). Aperçu sur la créativité morphologique dans la terminologie de la chimie. *La banque des mots*, **60**, 51–68.

Cottez, H. (1985). *Dictionnaire des structures du vocabulaire savant*. Paris: Le Robert.

Côté, R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Côté, R. A., Brochu, L. & Cabana, L. (1997). *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.

Daille, B. (1995). Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *Traitement Automatique des Langues (T.A.L.)*, **36**(1-2), 101–118.

Daille, B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pp. 9–16.

Dal, G. (1997). Du principe d'unicité catégorielle au principe d'unicité sémantique: incidence sur la formalisation du lexique construit morphologiquement. *Bulag (numéro spécial)*, pp. 105–115.

Dal, G. (2003). Productivité morphologique : définitions et notions connexes. *Langue française*, **140**, 3–23.

Dal, G. & Jacquemin, C. (1999). *FRANLEX. Base de données lexicales sur la morphologie dérivationnelle en français contemporain*. Description du projet, SILEX - LIMSI-CNRS. Disponible à `http://m17.limsi.fr/Individu/jacquemi/FRANLEX/index.html`. Visité le 17/08/99.

Dal, G., Namer, F. & Hathout, N. (1999). Construire un lexique dérivationnel : théorie et réalisations. In P. Amsili, Ed., *Traitement Automatique des Langues Naturelles (TALN)*, pp. 115–124, Cargèse.

Darmesteter, A. (1875). *Traité de la formation des mots composés dans la langue française comparée aux autres langues romanes et au latin*. Paris: Librairie Honoré Champion. Version revue et refondue en 1894.

Darmoni, S. J., Thirion, B., Leroy, J.-P., Douyère, M., Lacoste, B., Godard, G., Rigolle, I., Brisou, M., Videau, S., Goupy, E., Piot, J., Quéré, M., Ouazir, S. & Abdulrab, H. (2001). A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *MIIM*, **26**(3), 165–178.

Darmoni, S. J., Éric Jarousse, Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., Joubert, M., Vallée, H., Côté, R., Buemi, A., Bourigault, D., Recourcé, G., Jeanneau, S. & Rodrigues, J.-M. (2003). VUMeF : Extending the french involvement in the UMLS Metathesaurus. In *Annual Symposium of the American Medical Informatics Association (AMIA)*.

DE LOUPY, C. (1994). La méthode d'étiquetage d'Eric Brill. *Traitement automatique des langues (TAL)*, **36**(1-2), 37–46.

DE LOUPY, C. (2000). *Évaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche d'information*. Thèse de doctorat en informatique, Université d'Avignon et des Pays de Vaucluse, Aix-Marseille, France.

DEGOULET, P., SAUQUET, D., JAULENT, M.-C., ZAPLETAL, E. & LAVRIL, M. (1998). Rationale and design considerations for a semantic mediator in health information systems. *Methods in Informatics and Medicin (MIM)*, **37**, 518–526.

DI SCIULLO, A.-M. & WILLIAMS, E. (1987). *On the definition of Word*. Cambridge: MIT Press.

DIAS, G., CARAPINHA, L., TRINDADE, R., MOTA, S. & RIBEIRO, M. (2003). Construire et accéder à une base de données d'expressions figées à partir des ressources de la toile. In *Terminologie et intelligence artificielle (TIA)*, pp. 92–101, Strasbourg.

DUJOLS, P., AUBAS, P., BAYLON, C. & GRÉMY, F. (1991). Morphosemantic analysis and translation of medical compound terms. *Methods in Informatics and Medicin (MIM)*, **30**, 30–35.

DÉJEAN, H. (1998). Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, pp. 295–299, Adelaide.

ELKIN, P., RUGGIERI, A., BROWN, S., BUNTROCK, J., BAUER, B., WAHNER-ROEDLER, D., LITIN, S., BEINBORN, J., BAILEY, K. & BERGSTROM, L. (2001). A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

EMELINE, I. V., LEVENSON, R., PEROV, Y. L. & RYKIV, V. V. (1995). A Russian version of SNOMED-International. In R. A. GREENES, Ed., *Medical Informatics in Europe (MIE)*.

ENGUEHARD, C., MALVACHE, P. & TRIGANO, P. (1992). Indextion de textes : l'apprentissage des concepts. In *International Conference on Computational Linguistics (COLING)*, Nantes, 23-28 août.

EVANS, D. A. & LEFFERTS, R. G. (1993). Design and evaluation of the CLARIT-TREC-2 system. In *Text REtrieval Conference*, pp. 137–150.

FELLBAUM, C. & MILLER, G. A. (2003). Morphosemantic links in WordNet. *Traitement Automatique des Langues (TAL)*, **44**(2), 69–80.

FRADIN, B. (2000). Combining forms, blends and related phenomena. *Extragrammatical and marginal morphology*, pp. 11–59.

FRADIN, B. (2003). *Nouvelles approches en morphologie*. Paris: Presses universitaires de France (PUF).

FRADIN, B., HATHOUT, N. & MEUNIER, F. (2003). La suffixation en *-et* et la question de productivité. *Langue française*, **140**, 56–78.

FRADIN, B. & KERLEROUX, F. (2003). Introduction. *Langages : Quoi de neuf en morphologie ?*, **152**, 3–11.

FRATH, P., OUESLATI, R. & ROUSSELOT, F. (2000). Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des Connaissances "Evolutions récentes et nouveaux défis"*, chapter 18. Eyrolles.

FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M. & BELL, S. T. D. (1987). The vocabulary problem in human-system communication: an analysis and a solution. In *Communications of the ACM*, pp. 964–971. Disponible à `http://citeseer.nj.nec.com/furnas87vocabulary.html`. Visité le 10/09/2002.

GAETA, L. & RICCA, D. (2003). Frequency and productivity in Italian derivation: a comparison between corpus-based and lexico-graphical data. *Italian Journal of Linguistics (Rivista di Linguistica)*, **15**(1).

GARCIA, D. (1998). *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique* COATIS. Thèse de doctorat nouveau régime en informatique, Université de Paris-Sorbone (Paris IV), Paris, France.

GAUDINAT, A. & BOYER, C. (2002). Automatic extraction of MeSH terms from Medline abstracts. In R. BAUD & P. RUCH, Eds., *Workshop Natural Language Processing in Biomedical Applications of Medical Informatics in Europe (MIE)*, pp. 63–68, Cyprus.

GAUSSIER, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In A. KEHLER & A. STOLCKE, Eds., *ACL workshop on Unsupervised Methods in Natural Language Learning*, College Park, Md.

GAUSSIER, E., GREFENSTETTE, G., HULL, D. & ROUX, C. (2000). Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues (TAL)*, **41**(2), 473–493.

GIRALDO, G. & REYNAUD, C. (2002). Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine. In *Actes de la conférence Journées Francophones de l'Ingénierie des Connaissances (IC)*, Rouen.

GRABAR, N. (1999a). *Acquisition automatique de connaissances morphologiques à partir d'un thésaurus: Prise en compte de contraintes syntaxiques en français et en russe.* Mémoire de DEA, INaLCO.

GRABAR, N. (1999b). *Préparation de ressources lexicales pour l'aide au codage automatique par extension de la requête.* Mémoire de DESS, INaLCO.

GRABAR, N. & HAAG, K. (2003). Des textes parallèles vers une terminologie trilingue. In *Terminologie et intelligence artificielle (TIA)*, pp. 102–111, Strasbourg.

GRABAR, N. & HAMON, T. (2004a). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle (RIA)*, **18**(1).

GRABAR, N. & HAMON, T. (2004b). Repérage de relations sémantiques entre termes : sur la piste de la morphologie dérivationnelle. In *Ingénierie des connaissances (IC)*, pp. 237–248, Lyon.

GRABAR, N. & JEANNIN, B. (2002). Contribution de différents outils à la construction d'une terminologie pour la recherche d'information. In C. GRÉBOVAL, Ed., *Ingénierie des connaissances (IC)*, Rouen. Poster.

GRABAR, N. & ZWEIGENBAUM, P. (1999a). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Traitement Automatique de Langues Naturelles (TALN)*, pp. 175–184.

GRABAR, N. & ZWEIGENBAUM, P. (1999b). Language-independent automatic acquisition of morphological knowledge from synonym pairs. *JAMIASUP*, pp. 77–81.

GRABAR, N. & ZWEIGENBAUM, P. (2000a). Automatic acquisition of domain-specific morphological resources from thesauri. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pp. 765–784.

GRABAR, N. & ZWEIGENBAUM, P. (2000b). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, pp. 310–314.

GRABAR, N. & ZWEIGENBAUM, P. (2002a). Lexically-based terminology structuring: a feasibility study. In *Workshop terminologie de Language Resources and Evaluation (LREC)*, Canaries.

GRABAR, N. & ZWEIGENBAUM, P. (2002b). Lexically-based terminology structuring: Some inherent limits. In L.-F. CHIEN, B. DAILLE, K. KAGEURA & H. NAKAGAWA, Eds., *Proceedings of Second International Workshop on Computational Terminology (COMPUTERM 2002)*, pp. 36–42, Taipei, Taiwan: ACLCLP.

GRABAR, N. & ZWEIGENBAUM, P. (2003). Productivité à travers domaines et genres : dérivés adjectivaux et langue médicale. *Langue française*, **140**, 102–125.

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. (2003). Matching controlled vocabulary. In *Medical Informatics in Europe (MIE)*, Saint-Malo, France.

GREFENSTETTE, G. (1994). *Explorations in automatic thesaurus discovery.* Kluwer Academic Publishers.

GREFENSTETTE, G. & NIOCHE, J. (2000). Estimation of English and non-English language use on the WWW. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pp. 237–246, Paris.

GRUAZ, C. (1998). La hiérarchie dérivationnelle dans les familles synchroniques. Prototypie, polysémie et homonymie. In *Congrès Linguistique et Philologie romanes*, Bruxelles.

HABERT, B., GRABAR, N., JACQUEMART, P. & ZWEIGENBAUM, P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus Linguistics*, Lancaster.

HABERT, B., HERVIOU-PICARD, M.-L., BOURIGAULT, D., QUATRAIN, R. & ROUMENS, M. (1997). Un outil et une méthode pour comparer deux extracteurs de groupes nominaux. In *Journées Scientifiques et Techniques FRANCIL*, Avignon.

HABERT, B., ILLOUZ, G., LAFON, P., FLEURY, S., FOLCH, H., HEIDEN, S. & PRÉVOST, S. (2000). Profilage de textes : cadre de travail et expérience. In M. RAJMAN, Ed., *5èmes Journées d'Analyse des Données Textuelles (JADT)*, Lausanne.

HABERT, B. & JACQUEMIN, C. (1993). Noms composés, termes, dénominations complexes : problématiques et traitements automatiques. *Traitement Automatique des Langues (TAL)*, **34**(2), 5–41.

HABERT, B., NAULLEAU, E. & NAZARENKO, A. (1996). Symbolic word clustering for medium-size corpora. In *International Conference on Computational Linguistics (COLING'96)*, volume 1, pp. 490–495, Copenhagen, Danmark.

HAMON, T. (2000). *Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales*. Thèse de doctorat en informatique, Université Paris Nord, Paris.

HAMON, T. & NAZARENKO, A. (2001a). Detection of synonymy links between terms: experiment and results. In *Recent Advances in Computational Terminology*, pp. 185–208. John Benjamins.

HAMON, T. & NAZARENKO, A. (2001b). Exploitation de l'expertise humaine dans un processus de constitution de terminologie. In *Traitement Automatique des Langues Naturelles (TALN)*, pp. 213–222, Tours.

HAMON, T., NAZARENKO, A. & GROS, C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, pp. 498–504, Université de Montréal, Montréal, Quebec, Canada.

HARMAN, D. (1991). How effective is suffixing? *Journal of the American Society of Information Science*, **42**(1), 7–15.

HARRIS, Z. S. (1971). *Structures mathématiques du langage*. Paris: Monographies de linguistique mathématique. Dunod. Traduit par C. Fuchs.

HATHOUT, N. (2001). Analogies morpho-syntaxiques. In *Traitement Automatique des Langues Naturelles (TALN)*, Tours.

HATHOUT, N., NAMER, F. & DAL, G. (2001). An experimental constructional database: the MorTAL project. In P. BOUCHER, Ed., *Morphology book*. Cambridge, MA: Cascadilla Press.

HAWKING, D., BAILEY, P. & CRASWELL, N. (2000). *An Intranet Reality Check For TREC Ad Hoc*. Technical report, CSIRO Mathematical and Information Sciences, Canberra, Australia. Disponible à `http://pigfish.vic.cmis.csiro.au/~nickc/pubs/`.

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France. Disponible à `http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Gery1/index.html`. Visité le 26/08/99.

HERSH, W. R., CAMPBELL, E. H., EVANS, D. A. & BROWNLOW, N. D. (1996). Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. In *Annual Symposium of American Medical Informatics Association (AMIA)*, pp. 159–163.

HERSH, W. R., CAMPBELL, E. M. & MALVEAU, S. E. (1997). Assessing the feasibility of large-scale NLP in a corpus of ordinary medical records: a lexical analysis. *Journal of American Medical Informatics Association (JAMIA)*, **4 suppl.**, 580–584.

HULL, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science*, **47**(1), 70–84.

HUOT, H. (2001). *Morphologie, forme et sens des mots du français*. Paris: Armand Collin.

IACOBINI, C. (1997). Distinguishing derivational prefixes from initial combining forms. In *First mediterranean conference of morphology*, Mytilene, Island of Lesbos, Greece.

INGENERF, J. & GIERE, W. (1998). Concept-oriented standartization and statistics-oriented classification: continuing the classification versus nomenclature controversy. *Methods in Informatics and Medicin (MIM)*, **37**, 527–539.

INSERM (2000). *Thésaurus Biomédical Français/Anglais*. Institut National de la Santé et de la Recherche Médicale, Paris.

JACQUEMIN, C. (1996). A symbolic and surgical acquisition of terms through variation. In S. WERMTER, E. RILOFF & G. SCHELER, Eds., *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 425–438, Springer.

JACQUEMIN, C. (1997a). Guessing morphology from terms and corpora. In *ACM SIGIR*.

JACQUEMIN, C. (1997b). *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. (1999). Syntagmatic and paradigmatic representations of term variation. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 341–348, University of Maryland.

JACQUEMIN, C. & TZOUKERMANN, E. (1999). NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In T. STRZALKOWSKI, Ed., *Natural Language Processing and Information Retrieval*, chapter 2, pp. 25–74. Boston, MA: Kluwer.

JACQUES, M.-P. (2003). Repérage de termes réduits : intérêt et limites de l'analyse distributionnelle. In *Terminologie et intelligence artificielle (TIA)*, pp. 135–144, Strasbourg.

JANSEN, B. J., SPINK, A., BATEMAN, J. & SARACEVIC, T. (1998). Real life information retrieval: A study of user queries on the web. In *SIGIR Forum*, pp. 5–17. Disponible à `http://jimjansen.tripod.com/academic/pubs/ipm98/ipm98.html`. Visité le 08/02/2002.

JANSEN, B. J., SPINK, A. & SARACEVIC, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, **36**(2), 207–227.

JEANNIN, B. & MONCEAUX, A. (2003). Constitution de dictionnaires bilingues à partir de données terminologiques trilingues. In F. ROUSSELOT, Ed., *Terminologie et Intelligence Artificielle (TIA)*, Strasbourg. Poster.

JONES, S., CUNNINGHAM, S. J. & MCNAB, R. J. (1998). Usage analysis of a digital library. In *ACM DL*, pp. 293–294. Disponible à `http://citeseer.nj.nec.com/steve98usage.html`. Visité le 08/02/2002.

JONES, S., CUNNINGHAM, S. J., MCNAB, R. J. & BODDIE, S. J. (2000). A transaction log analysis of a digital library. *Internal Journal on Digital Libraries*, **3**(2), 152–169. Disponible à `http://citeseer.nj.nec.com/330107.html`. Visité le 08/02/2002.

KAVANAGH, J. (1995). *The Text Analyser: A Tool for Extracting Knowledge From Text*. Master of computer science thesis, University of Ottawa, Ottawa, Canada.

KLEIBER, G. & TAMBA, I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, **98**, 7–32. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

KOSKENNIEMI, K. (1983). *Two-level morphology: a general computational model for word-form recognition and production*. PhD thesis, University of Helsinki Department of General Linguistics, Helsinki.

KRAAIJ, W. & POHLMANN, R. (1996). Viewing stemming as recall enhancement. In *SIGIR*.

KROVETZ, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202.

LAFON, P. (1984). *Dépouillements et statistiques en lexicométrie*. Paris-Genève: Champion-Slatkine.

LAME, G. (2002). *Construction d'ontologies à partir de textes. Une ontologie du droit dédié à la recherche d'information sur le Web*. Thèse de doctorat en informatique temps réel, robotique et automatique, École des Mines de Paris, Paris.

LE MOIGNO, S., CHARLET, J., BOURIGAULT, D., DEGOULET, P. & JAULENT, M.-C. (2002). Terminology extraction from text to build an ontology in surgical intensive care. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, pp. 430–434.

LECOMTE, J. (1998). Le catégoriseur Brill14-JL5 / WinBrill-0.3. INaLF. Catégoriseur d'Eric Brill, entraîné pour le français à l'INaLF.

LESSELINGUE, C. (2003). Les noms composés $[NN]_N$ 'holonymiques' : illustration de la spécificité sémantique des unités construites morphologiquement. *Silexicales*, **3**, 100–107. 3^e rencontres du Forum de Morphologie. Actes du colloque de Villeneuve d'Ascq. 19-21 septembre 2002.

LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, **707**(10).

L'Homme, M.-C. (2004). Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies. In *Journées d'étude Terminologie, Ontologie et représentation des connaissances*, Lyon.

Lieber, R. (1992). *Deconstructing morphology : word formation in syntactic theory*. Chicago/London: The University of Chicago Press.

Light, M. (1996). Morphological cues for lexical semantics. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL)*, University of California, Santa Cruz, CA, USA.

Lignon, S. (1999). Suffixasser ou suffixailler ? *Silexicales*, **2**, 117–126. 2^e rencontres du Forum de Morphologie. Actes du colloque de Toulouse. 29-30 avril 1999.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical translation and computational linguistics*, **11**, 22–31.

Lovis, C., Michel, P.-A., Baud, R. & Scherrer, J.-R. (1995). Word segmentation processing: a way to exponentially extend medical dictionaries. In *Medical Informatics in Europe (MIE)*, pp. 28–32.

Lussier, Y. A., Rothwell, D. J. & Côté, R. A. (1998). The SNOMED model : a knowledge source for the controlled terminology of the computerized patient record. *Methods in Informatics and Medicin (MIM)*, **37**, 161–164.

Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press. 2 vol.

Maedche, A. & Staab, S. (2000). Mining ontologies from text. In R. Dieng & O. Corby, Eds., *EKAW 2000*, Juan les Pins.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Manuila, A., Manuila, L., Nicoulin, M. & Lambert, H. (1970). *Dictionnaire français de médicine et de biologie*. Paris: Masson et Cie. 3 volumes.

Manuila, L., Manuila, A., Lewalle, P. & Nicoulin, M. (2001). *Dictionnaire médical*. Paris: Masson. 9^e édition.

Marcus, A. (2003). *Dictionnaires électroniques et hypertextualité. Analyse critique des renvois doubles du Grand Robert*. Mémoire de DESS, CRIM/INaLCO. Sous la direction de David Piotrovsky.

McCray, A. T., Srinivasan, S. & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual SCAMC*, pp. 235–239.

Melčuk, I. (1993). *Cours de morphologie générale. Première partie : Le mot*. Montréal: Presses Universitaires de Montréal - CNRS Éditions.

Melčuk, I. (1996). *Cours de morphologie générale. Troisième partie : Moyens morphologiques. Quatrième partie : Syntactiques morphologiques*. Montréal: Presses Universitaires de Montréal - CNRS Éditions.

MERTENS, P. (1999). Un algorithme pour la génération de l'intonation dans la parole de synthèse. In P. AMSILI, Ed., *Traitement Automatique des Langues Naturelles (TALN)*, pp. 233–242, Cargèse.

MEUNIER, F. (2003). La notion de productivité morphologique : modèles psycholinguistiques et données expérimentales. *Langue française*, **140**, 24–37.

MEUNIER, L. F. (1872). *Étude sur les composés syntactiques en grec*. Paris: A. Durand and Pedone Lauriel.

MONTEMAGNI, S. & VANDERWENDE, L. (1992). Structural patterns vs. string patterns for extracting semantic information from dictionnaries. In *International Conference on Computational Linguistics (COLING)*, pp. 546–552, Nantes.

MORIN, E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL)*, **40**(1), 143–166.

MÉLIS-PUCHULU, A. (1991). Les adjectifs dénominaux : les adjectifs de 'relation'. *Lexique*, **10**, 33–60.

NAMER, F. (1999). Traitement automatique de la dérivation : le cas des noms et adjectifs en *-et(te)*. *Silexicales*, **2**.

NAMER, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, **41**(2), 523–547.

NAMER, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Traitement Automatique de la Langue Naturelle (TALN)*, pp. 235–244, Nancy.

NAMER, F. (2003a). *Les mots composés morphologiquement*. Note technique faite dans le cadre du projet UMLF, LANDISCO, Université Nancy 2.

NAMER, F. (2003b). Productivité morphologique, représentativité et complexité de la base : le système MoQuête. *Langue française*, **140**, 79–101.

NAMER, F. & ZWEIGENBAUM, P. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, San-Francisco.

NAZARENKO, A., ZWEIGENBAUM, P., HABERT, B. & BOUAUD, J. (2001). Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, pp. 327–351. John Benjamins.

NELSON, S. J., JOHNSTON, D. & HUMPHREYS, B. L. (2001). Relationships in Medical Subject Headings. In C. A. BEAN & R. GREEN, Eds., *Relationships in the organization of knowledge*, pp. 171–184, New York: Kluwer Academic Publishers.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. `http://www.nlm.nih.gov/mesh/meshhome.html`.

NLM (2003). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. `www.nlm.nih.gov/research/umls/`.

Névéol, A. (2004). Indexation automatique de ressources de santé à l?aide d?un vocabulaire contrôlé. In *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, Fès, Maroc.

Névéol, A., Douyère, M., Rogozan, A. & Darmoni, S. J. (2004). Construction de ressources terminologiques en santé pour un système d?indexation automatique. In *Journées INTEX/NOOJ 2004*, Tours, France.

OMS (1995). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

Pacak, M. G., Norton, L. M. & Dunham, G. S. (1980). Morphosemantic analysis of -itis forms in medical language. *Methods in Medical Informatics (MIM)*, **19**(2), 99–105.

Pascal, B. (1662). *Pensées*. Livre de poche de la Librairie générale française. Édition consultée : 1962.

Pearson, J. (1998). *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Pentheroudakis, J. & Vanderwende, L. (1993). Automatically identifying morphological relations in machine-readable dictionaries. In *Ninth annual confenrence of the UW Center for the New OED an d Text Research*, pp. 114–131.

Pirrelli, V. & Yvon, F. (1999). The hidden dimension: a paradigmatic view of data-driven NLP. *JETAI*, **11**, 391–408.

Plénat, M. (2001). Jean-Louis Fossat : fossatissime. Note sur la morphophonologie des dérivés en *-issime*. *Cahier d'études romanes*, **10**. Hommage à Jean-Louis Fossat.

Plénat, M. & Roché, M. (2001a). Entre morphologie et phonologie : la suffixation décalée. *Lexique*, **16**.

Plénat, M. & Roché, M. (2001b). Prosodic constraints on suffixation in French. In J. DeCesaris, Ed., *Selected papers from the Third Mediterranean Morphological Meeting*, Barcelona, Espagne.

Popovic, M. & Willett, P. (1992). The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society of Information Science (JASIS)*, **43**(5), 384–390.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

Porter, M. (2001). *Snowball: a language for stemming algorithms*. Technical report. Disponible à `http://snowball.tartarus.org/texts/introduction.html`. Visité le 5/09/2003.

Pouliquen, B., Delamarre, D. & Le Beux, P. (2002). Indexation de textes médicaux par extraction de concepts, et ses utilisations. In *Journées de traitement automatique des données textuelles (JADT)*, pp. 617–628.

Riegel, M., Pellat, J.-C. & Rioul, P. (1998). *Grammaire méthodique du français*. Paris: PUF.

RILLOF, E. (1995). Little words can make a big difference for text classification. In *SIGIR*, pp. 130–136, Seattle, Washingthon.

ROBERT (1993). *Le nouveau petit Robert*. Paris: Dictionnaires Le Robert.

SAVOY, J. (2002). *Morphologie et recherche d'information*. Cahier de recherche en informatique CR-I-2002-01, Université de Neuchatel, Division économique et sociale, Faculté de Droit et des Sciences Économiques.

SCALISE, S. (1994). *Morfologia*. Bologna: Il Mulino.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49, Manchester, UK.

SCHONE, P. & JURAFSKY, D. (2001). Knowledge-free induction of inflectional morphologies. In *Workshop NA de ACL*.

SCHULTINK, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, **2**, 110–125.

SCHULTINK, H. (1962). *De morfologische valentie van het ongelede adjectief in modern Nederlands [The morphological valency of the simplex adjective in modern Dutch]*. The Hague: Van Goor and Zonen.

SCHULZ, S., ROMACKER, M., FRANZ, P., ZAISS, A., KLAR, R. & HAHN, U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*.

SILBERZTEIN, M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris: Masson.

SILVERSTEIN, C., HENZINGER, M., MARAIS, H. & MORICZ, M. (1998). *Analysis of a Very Large AltaVista Query Log*. Technical report, Digital systems research center. Disponible à `http://citeseer.nj.nec.com/silverstein98analysis.html`. Visité le 08/02/2002.

SIMARD, M. (1998). Automatic insertion of accents in French text. In *3rd Conference on Empirical Methods in Natural Language processing*, Grenade.

SINCLAIR, J. (1994). *EAGLES. Corpus typology*. Technical report, EAG-CWG-IR-2. Disponible à `http://www.ilc.pi.cnr.it/EAGLES96/`. Visité le 02/03/2003.

SNOP (1965). *Systematized Nomenclature of Pathology*. College of American Pathologists, University of Chicago, États-Unis, Illinois.

SOUALMIA, L. F. & DARMONI, S. J. (2004). Combining different standards and different approches for health information retrieval in a quality-controlled gateway. *International Journal of Medical Informatics (IJMI)*.

SPACKMAN, K. & CAMPBELL, K. (1998). Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. In *Journal of American Medical Informatics Association (JAMIA)*, pp. 740–744.

SPACKMAN, K. A. & HERSH, W. R. (1996). Recognizing noun phrases in medical discharge summaries: an evaluation of two natural language parsers. In *Annual Symposium of American Medical Informatics Association (AMIA)*, pp. 155–158.

SPRIET, T. & EL-BÈZE, M. (1997). Réaccentuation automatique de textes. In *FRACTAL*, Besançon.

SPYNS, P. (1994). A robust category guesser for Dutch medical language. In *Proceedings of ANLP 94 (ACL)*, pp. 150–155.

SÉGUÉLA, P. & AUSSENAC-GILLES, N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes d'Ingénierie des Connaissances (IC)*, pp. 79–88, Palaiseau, France.

TANGUY, L. & HATHOUT, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Traitement Automatique des Langues Naturelles (TALN)*, pp. 245–254, Nancy.

THERON, P. & CLOETE, I. (1997). Automatic acquisition of two-level morphological rules. In *ANLP*, pp. 103–110.

TLFI (2001). *Trésor de la Langue Française - I.* INaLF/ATILF. Disponible à l'adresse `www.tlfi.fr`.

TOUSSAINT, Y. & SIMON, A. (2000). Building and interpreting term dependencies using association rules extracted from Galois lattices. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pp. 1686–1692, Paris.

TURNER, J. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images : research results. In *Meeting of American Society of Indexers*, pp. 2–10.

URREA, A. M. (2000). Automatic discovery of affixes by means of a corpus : a catalog of Spanish affixes. *Journal of quantitative linguistics*, **7**(2), 97–114.

VALETTE, M. & GRABAR, N. (2004). Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP. In *Journées de traitement automatique des données textuelles (JADT)*, Liège, Belgique.

VAN DEN BOSCH, A., DAELEMANS, W. & WEIJTERS, T. (1996). Morphological analysis as classification: an inductive-learning approach. In *International Conference on Computational Linguistics (COLING)*.

VAN MARLE, J. (1991). The relationship bertween morphological productivity and frequency: a comment on Baayen's performance-oriented conception of morphological productivity. In G. BOOIJ & J. VAN MARLE, Eds., *Yearbook of Morphology*, pp. 151–163, Netherlands: Kluwer Academic Publishers.

VIEGAS, E., GONZALES, M. & LONGWELL, J. (1996). Morpho-semantics and constructive derivational morphology: A transcategorical approach. In *Memoranda in Computer and Cognitive Science*. Disponible à `http://crl.nmsu.edu/Research/Projects/mikro/htmls/misc-htmls/mikro-pub.%html`. Visité le 17/08/99.

VILLOING, F. (2002). *Les mots composés $[VN]_{N/A}$ du français : réflexions épistémologiques et propositions d'analyse*. Thèse de doctorat, Université de Paris X - Nanterre, Paris. Discipline : Sciences du langage.

WANG, A. Y., BARRETT, J. W., BENTLEY, T., MARKWELL, D., PRICE, C., SPACKMAN, K. A. & STEARNS, M. Q. (2001). Mapping between SNOMED RT and Clinical Terms version 3: a key component of the SNOMED CT development process. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

WIMMER, G. & ALTMANN, G. (1995). A model of morphological productivity. *Journal of quantitative linguistics*, **2**(3), 212–216.

WOLFF, S. (1987). Automatic coding of medical vocabulary. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., *Medical Language Processing. Computer Management of Narrative Data*, chapter 7, pp. 145–162. New-York: Addison-Wesley.

XU, J. & CROFT, B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, **16**(1), 61–81.

ZWANENBURG, W. (1991). Dépendances et hiérarchies en morphologie lexicale. *L'Information grammaticale*, **50**, 38–42.

ZWEIGENBAUM, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Information de Santé Innovation Stratégie (ISIS)*, **2**(3), 27–47.

ZWEIGENBAUM, P. (2004). L'UMLS entre langue et ontologie : une approche pragmatique. *Revue d'intelligence artificielle (RIA)*, **18**.

ZWEIGENBAUM, P., BACHIMONT, B., BOUAUD, J., CHARLET, J., BEN SAÏD, M., BOISVIEUX, J.-F., DUJOLS, P., FIESCHI, M., VOLOT, F., BEUX, P. L., DELAMARRE, D., BURGUN, A., KEANE, M. T., NANGLE, B., FLINTER, S., MCALLISTER, B., MC NAMARA, S., JOHNSTON, R., MCDERMOTT, M., O'SULLIVAN, K., WILLEMS, J. L., SPYNS, P., OGONOWSKI, A., GUILLOTIN, T., FARGUES, J., LANDAU, M.-C., BRADEA, A., MOENS, M., WHITTEMORE, G., GROVER, C. & MIKHEEV, A. (1995). MENELAS *Final Report*. Deliverable report AIM-MENELAS 17, DIAM-SIM/INSERM U.194.

ZWEIGENBAUM, P., BAUD, R., BURGUN, A., NAMER, F., ÉRIC JARROUSSE, GRABAR, N., RUCH, P., DUFF, F. L., THIRION, B. & DARMONI, S. (2003a). Towards a Unified Medical Lexicon for French. In *Medical Informatics in Europe (MIE)*.

ZWEIGENBAUM, P. & GRABAR, N. (1999). A contribution of medical terminology to medical language processing resources: Experiments in morphological knowledge acquisition from thesauri. In *Work Group 6 of IMIA*, pp. 155–167.

ZWEIGENBAUM, P. & GRABAR, N. (2000). Liens morphologiques et structuration de terminologie. In *Ingénierie des connaissances (IC)*, pp. 325–334.

ZWEIGENBAUM, P. & GRABAR, N. (2002a). Accentuation de mots inconnus : application au thesaurus biomédical MeSH. In *Ttraitement Automatique des Langues Naturelles (TALN)*, pp. 53–62.

ZWEIGENBAUM, P. & GRABAR, N. (2002b). Restoring accents in unknown biomedical words: application to the French MeSH thesaurus. *International Journal of Medical Informatics (IJMI)*, pp. 113–126.

ZWEIGENBAUM, P., GRABAR, N. & DARMONI, S. J. (2001). L'apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée. In *Traitement Automatique des Langues Naturelles (TALN)*, pp. 403–408.

ZWEIGENBAUM, P., GRABAR, N. & DARMONI, S. J. (2002). Projection de requêtes en langue naturelle sur les termes du MeSH : l'apport de connaissances morphologiques. In *INFSANT 2002*, pp. 55–62.

ZWEIGENBAUM, P., HADOUCHE, F. & GRABAR, N. (2003b). Apprentissage de relations morphologiques en corpus. In *Traitement Automatique des Langues Naturelles (TALN)*.

Annexe A

Acquisition de ressources morphologiques à partir de terminologies structurées

A.1 Apprentissage à partir de séries de synonymes

A.1.1 Extrait des séries de synonymes du Répertoire d'anatomopathologie (196 séries sur 2 344)

Français

A-00100 : *dispositif, SAI*; *appareil, SAI*
A-38100 : *dispositif de protection, SAI*; *appareil de protection, SAI*
C-10092 : *produit chimique organique, SAI*; *composé organique, SAI*
C-20070 : *émanations chimiques, SAI*; *vapeurs chimiques, SAI*
C-20821 : *chloroforme*; *trichlorméthane*
D0-10430 : *pemphigoïde, SAI*; *pemphigus bénin, SAI*
D0-90150 : *panniculite fébrile nodulaire récidivante non suppurée*; *maladie de Weber-Christian*
D2-02400 : *polype nasal, SAI*; *polype de la cavité nasale, SAI*
D2-53140 : *amiantose*; *asbestose*
D2-60100 : *syndrome de détresse respiratoire de l'adulte, SAI*; *SDRA*
D3-81670 : *angéite allergique*; *vasculite leucocytoclastique*
D4-85150 : *absence congénitale de l'épithélium germinal des testicules*; *aplasie germinale des testicules*
D5-34300 : *ulcère peptique, SAI*; *ulcère gastro-duodénal, SAI*
D5-80700 : *cirrhose biliaire*; *cirrhose cholestatique*; *angiocholite chronique non suppurée*
D6-11110 : *obésité constitutionnelle*; *obésité familiale*
D6-34750 : *pseudo-goutte, SAI*; *chondrocalcinose articulaire idiopathique*; *chondrocalcinose articulaire due aux cristaux de pyrophosphate*
D6-50430 : *glycogénose de type VII*; *maladie de Tarui*
D6-60210 : *abêtalipoprotéinémie*; *maladie de Bassen-Kornzweig*
D6-70232 : *mucopolysaccharidose, MPS-III-B*; *maladie de Sanfilippo de type B*; *déficit en alpha-N-acétylglucosaminidase*
D6-72760 : *déficit en triacylglycérol lipase pancréatique*; *absence congénitale de la lipase pancréatique*
D6-76130 : *maladie de Tay-Sachs*; *déficit sévère en hexosaminidase A*
D6-98300 : *diabète insipide néphrogénique héréditaire, SAI*; *diabète insipide pitresso-résistant*
D7-11600 : *hypoplasie rénale, SAI*; *hypoplasie rénale d'origine inconnue*
D7-12434 : *néphrite tubulo-interstitielle fongique aiguë*; *n.t.-i. fongique aiguë*
D7-23570 : *prolapsus de l'urètre*; *urétrocèle*
D7-71460 : *pyomètre*; *pyométrie*
D8-00110 : *grossesse molaire, SAI*; *môle de grossesse, SAI*
D9-13100 : *anorexie mentale*; *anorexie psychogène*
DB-02110 : *adénomatose pluri-endocrinienne, type 1 : syndrome de Wermer*
DC-11440 : *aplasie chronique constitutionnelle isolée de la lignée rouge*; *anémie de Blackfan-Diamond*
DC-38000 : *polyglobulie, SAI*; *polyglobulie secondaire, SAI*
DE-12100 : *granulome inguinal*; *granulome vénérien*
DE-40000 : *mycose, SAI*; *maladie d'origine mycosique, SAI*
DE-69240 : *infection à Filaria ozzardi*; *infection à Mansonella ozzardi*
F-06100 : *état post-vaccination*; *état post-inoculation*
F-32320 : *shunt, SAI*; *dérivation, SAI*

F-58605 : *vésicule biliaire non fonctionnelle* ; *absence de fonction de la vésicule biliaire* ; *vésicule biliaire exclue*
F-65830 : *acide uridylique* ; *uridine-monophosphate*
F-81170 : *ménorragie* ; *hyperménorrhée*
F-C0B00 : *chimiotactisme, SAI* ; *chimiotaxie, SAI*
F-D0105 : *arrêt de la maturation érythrocytaire* ; *arrêt de la maturation érythroïde*
G-0008 : *pas d'antécédents de* ; *aucun antécédent de*
G-1010 : *diagnostic clinique* ; *dx clinique*
G-4004 : *état post* ; *après*
G-7170 : *période définie* ; *période limitée*
G-A213 : *inégal* ; *inégalité de*
G-A358 : *fasciculaire* ; *fasciculé*
G-A417 : *local* ; *localisé* ; *topique* ; *région limitée*
G-A506 : *à répéter* ; *répété*
G-A617 : *épais* ; *épaissi*
G-C001 : *dû à* ; *causé par*
G-D820 : *isogreffe* ; *greffe isogénique* ; *greffe syngénique*
J-00000 : *profession, SAI* ; *métier, SAI* ; *travail, SAI*
L-25116 : *Streptococcus pneumoniae* ; *Diplococcus pneumoniae* ; *pneumocoque*
L-36212 : *virus de l'herpès simplex, type 2 : HSV-2*
L-55841 : *Necator americanus* ; *ankylostome des Amériques*
L-58050 : *trématode du foie, SAI* ; *douve du foie, SAI*
M-01000 : *anomalie morphologique, SAI* ; *modification morphologique, SAI* ; *changement morphologique, SAI* ; *aspect tissulaire anormal, SAI*
M-01530 : *fusion, SAI* ; *fusionné*
M-02501 : *variante normale de la taille* ; *variante normale du volume*
M-03000 : *masse, SAI* ; *bosse, SAI* ; *masse localisée, SAI*
M-05150 : *dédoublement* ; *duplication* ; *dédoublé* ; *dupliqué*
M-09030 : *artéfact, SAI* ; *artéfact morphologique, SAI*
M-11100 : *brûlure, SAI* ; *lésion par brûlure, SAI* ; *brûlure thermique* ; *brûlé*
M-12590 : *pseudarthrose, SAI* ; *néarthrose* ; *néoarthrose*
M-14100 : *plaie pénétrante* ; *plaie perforante* ; *perforation traumatique*
M-17010 : *amputation par traumatisme* ; *formation anatomique amputée par traumatisme*
M-20700 : *communication anormale congénitale, SAI* ; *lien anormal congénital, SAI* ; *insertion anormale congénitale, SAI*
M-22170 : *prolapsus congénital, SAI* ; *ptose congénitale, SAI*
M-26360 : *vestige du canal de Müller* ; *vestige du canal paramésonéphrique* ; *hydatide sessile de Morgagni*
M-30040 : *calcul urique* ; *calcul uratique*
M-31050 : *prolapsus* ; *déplacement inférieur* ; *déplacement caudal* ; *ptose* ; *procidence* ; *déplacement vers le bas*
M-32360 : *anévrisme cylindrique* ; *anévrisme tubulaire*
M-33152 : *comédon à follicule fermé* ; *point blanc*
M-33780 : *kyste chyleux* ; *kyste lymphatique* ; *chylocèle*
M-35120 : *thrombus non occlusif* ; *thrombus annulaire* ; *thrombose partielle*
M-36410 : *spongiose* ; *oedème péricellulaire de l'épiderme*
M-37004 : *hémorragie massive* ; *saignement massif*
M-38350 : *ulcère superficiel* ; *érosion* ; *ulcération superficielle*
M-40200 : *inflammation séreuse, SAI* ; *transsudat inflammatoire, SAI*
M-41300 : *inflammation fibrineuse aiguë, SAI* ; *inflammation séro-fibrineuse aiguë, SAI*
M-42000 : *inflammation subaiguë, SAI* ; *inflammation chronique et aiguë, SAI* ; *inflammation chronique active, SAI* ; *infiltration inflammatoire subaiguë, SAI*
M-43610 : *abcès chronique* ; *abcès froid*
M-44710 : *granulome fibro-caséeux* ; *nodule fibro-caséeux*
M-50062 : *corps hyalins* ; *drusen*
M-50560 : *dégénérescence spongieuse* ; *changement spongieux* ; *status spongiosus*
M-52200 : *artériolosclérose, SAI* ; *artérioloscléreux*
M-54370 : *carnification* ; *carnisation*
M-54810 : *infarctus limitrophe* ; *infarctus de la zone vascularisée limitrophe* ; *infarctus de la zone artérielle limitrophe*
M-55880 : *dépôt d'or* ; *chrysopexie*
M-57560 : *neurones minéralisés* ; *neurones incrustés*
M-58130 : *atrophie neuropathique* ; *atrophie neurale* ; *atrophie névritique*
M-61110 : *aneuploïdie, SAI* ; *modification du nombre de chromosomes, SAI*
M-63300 : *macrocyte* ; *mégalocyte* ; *macrocytose*
M-67000 : *atypie cellulaire, SAI* ; *présence de cellules atypiques, SAI*
M-67210 : *macronucléus* ; *caryomégalie*
M-69060 : *réaction osmiophile* ; *osmiophilie* ; *osmiophile*
M-71090 : *pseudo-hypertrophie* ; *fausse hypertrophie*
M-72100 : *hyperplasie lobulaire* ; *hyperplasie multilobulaire*
M-72920 : *kératose inversée* ; *kératose folliculaire inversée*
M-74320 : *maladie fibrokystique, SAI* ; *maladie fibrokystique non proliférative*
M-75560 : *naevus vasculaire* ; *hamartome vasculaire, SAI*
M-76810 : *polype fibro-épithélial* ; *papillome fibro-épithélial* ; *polype fibreux* ; *acrochordon*
M-78172 : *nodules de la microglie* ; *nodules névrogliques*
M-79370 : *endomètre atrophique* ; *endomètre post-ménopausique* ; *endomètre sénile* ; *endomètre ménopausique*
M-80102 : *carcinome in situ, SAI* ; *carcinome intra-épithélial, SAI* ; *carcinome non invasif, SAI*
M-80702 : *carcinome épidermoïde in situ, SAI* ; *carcinome épidermoïde non-invasif, SAI* ; *carcinome intra-épidermique, SAI* ; *carcinome épidermoïde intra-épithélial* ; *carcinome spinocellulaire intra-épithélial* ; *carcinome spinocellulaire in situ, SAI*
M-80823 : *carcinome lympho-épithélial* ; *lympho-épithéliome* ; *carcinome nasopharyngé* ; *tumeur de Regaud* ; *tumeur de Schminke*
M-81201 : *papillome excréto-urinaire* ; *papillome de la vessie* ; *papillome urothélial*
M-81453 : *carcinome de type diffus* ; *adénocarcinome de type diffus*
M-81700 : *hépatome bénin* ; *adénome solitaire du foie* ; *adénome hépatocellulaire* ; *adénome à cellules hépatiques*
M-82210 : *polypose multiple* ; *polypes adénomateux multiples*
M-82630 : *adénome tubulovilleux, SAI* ; *adénome tubulopapillaire*
M-83130 : *adénofibrome à cellules claires* ; *cystadénofibrome à cellules claires*
M-83601 : *adénomatose endocrinienne multiple* ; *polyadénomatose endocrine*
M-83900 : *adénome des annexes cutanées* ; *tumeur des annexes cutanées* ; *tumeur annexielle*
M-84503 : *cystadénocarcinome papillaire, SAI* ; *adénocarcinome papillokystique*
M-84813 : *adénocarcinome mucipare* ; *carcinome mucipare* ; *adénocarcinome mucosécrétant* ; *carcinome mucosécrétant*
M-85050 : *papillomatose intracanalaire diffuse* ; *papillomatose intracanalaire, SAI*
M-85603 : *carcinome glandulaire et malpighien* ; *carcinome mixte glandulaire et spinocellulaire* ; *carcinome mixte glandulaire et épidermoïde*
M-86203 : *tumeur maligne de la granulosa* ; *carcinome de la*

granulosa ; *carcinome à cellules de la granulosa* ; *folliculome malin*
M-86600 : *tumeur à cellules du hile* ; *tumeur des cellules hilaires* ; *tumeur à cellules de Berger*
M-87203 : *mélanome malin, SAI* ; *mélanocarcinome* ; *naevocarcinome* ; *mélanosarcome, SAI* ; *mélanome, SAI*
M-87453 : *mélanome malin desmoplasique* ; *mélanome malin neurotropique*
M-88003 : *sarcome, SAI* ; *tumeur maligne des parties molles* ; *tumeur mésenchymateuse maligne* ; *sarcome des tissus mous*
M-88211 : *fibrome envahissant* ; *tumeur desmoïde extra-abdominale* ; *tumeur desmoïde, SAI* ; *fibrome desmoïde*
M-88523 : *liposarcome myxoïde* ; *myxoliposarcome*
M-89311 : *endométriose cytogène* ; *myose stromale endolymphatique* ; *endométriose stromale* ; *myose stromale, SAI*
M-90000 : *tumeur de Brenner solide* ; *tumeur de Brenner, SAI* ; *tumeur de Brenner kystique*
M-90503 : *mésothéliome malin* ; *mésothéliome, SAI*
M-90800 : *tératome bénin* ; *tératome adulte, SAI* ; *tératome adulte kystique, SAI* ; *dysembryome adulte*
M-91200 : *hémangiome, SAI* ; *angiome, SAI* ; *angiochoriome*
M-91803 : *ostéosarcome, SAI* ; *sarcome ostéogénique* ; *ostéochondrosarcome* ; *sarcome squelettogène* ; *sarcome ostéoblastique*
M-92511 : *tumeur à cellules géantes des parties molles, SAI* ; *tumeur à cellules géantes des tissus mous, SAI*
M-93103 : *améloblastome malin* ; *adamantinome malin (à l'exception du tibia et des os longs M-92613)* ; *épithélioma adamantin*
M-93841 : *astrocytome sous-épendymaire à cellules géantes* ; *astrocytome sous-épendymaire à grosses cellules fusiformes*
M-95003 : *neuroblastome, SAI* ; *sympathoblastome*
M-95350 : *méningiome angioblastique* ; *méningiome hémangioblastique*
M-95903 : *lymphome malin, SAI* ; *lymphome, SAI*
M-96823 : *lymphome malin diffus à grandes cellules à noyaux non encochés* ; *lymphome malin à grandes cellules à noyaux non encochés*
M-97013 : *maladie de Sézary* ; *syndrome de Sézary*
M-97661 : *lésion immunoprolifératives angiocentrique* ; *granulomatose lymphoïde*
M-98503 : *leucémie à cellules lymphosarcomateuses* ; *leucosarcomatose*
M-99403 : *leucémie à tricholymphocytes* ; *leucémie à tricholeucocytes* ; *tricholeucémie* ; *leucémie à cellules chevelues*
P1-03050 : *résection radicale* ; *résection étendue*
P1-0D000 : *transplantation, SAI (acte chirurgical)* ; *greffe, SAI (acte chirurgical)*
P3-45060 : *cytopathologie, révision des lames et rapport par le médecin d'un spécimen d'origine non génitale* ; *cytopathologie, révision de spécimen obtenu par lavage bronchioalvéolaire*
T-01120 : *couche claire de l'épiderme* ; *stratum lucidum*
T-02120 : *peau de la face, SAI* ; *peau du visage, SAI*
T-04280 : *prolongement axillaire du sein* ; *lobe de Spence du sein*
T-11085 : *os des extrémités, SAI* ; *os d'un membre, SAI*
T-1151D : *anneau fibreux périphérique du disque intervertébral, SAI* ; *anneau fibreux du disque intervertébral, SAI*
T-12380 : *os pelviens, SAI* ; *bassin osseux*
T-12730 : *rotule, SAI* ; *os du genou, SAI* ; *rotulien*
T-15001 : *articulation, SAI* ; *articulaire*
T-18950 : *aponévrose plantaire, SAI* ; *fascia plantaire, SAI*
T-1A310 : *glande, SAI* ; *glandulaire* ; *glandulo-*
T-21300 : *partie interne du nez, SAI* ; *cavité nasale*
T-24470 : *commissure du larynx, SAI* ; *commissure laryngée, SAI*
T-28052 : *pneumocyte, type 2* : *pneumocyte granuleux*
T-32100 : *oreillette, SAI* ; *oreillette cardiaque* ; *auriculaire (coeur)*
T-32411 : *cloison auriculo-ventriculaire* ; *septum atrio-ventriculaire*
T-32830 : *faisceau auriculo-ventriculaire, SAI* ; *faisceau de His*
T-40500 : *circulation cérébrale, SAI* ; *vaisseaux cérébraux, SAI*
T-42340 : *crosse aortique en amont du canal artériel* ; *région préligamentaire de la crosse aortique*
T-44290 : *branche basale externe de l'artère pulmonaire droite* ; *artère latéro-basale droite*
T-44460 : *branche apicale lobaire inférieure de l'artère pulmonaire gauche* ; *artère apicale du lobe inférieur gauche*
T-46200 : *artère thoracique interne, SAI* ; *artère mammaire interne, SAI*
T-48003 : *réseau veineux, SAI* ; *plexus veineux, SAI*
T-49102 : *système veineux profond du membre supérieur* ; *veines profondes du membre supérieur*
T-49848 : *sinus basilaire* ; *sinus occipital transverse*
T-51004 : *cavité buccale, SAI* ; *cavité orale, SAI*
T-53001 : *corps de la langue* ; *portion antérieure de la langue* ; *deux tiers antérieurs de la langue*
T-55300 : *hypopharynx, SAI* ; *laryngopharynx, SAI* ; *hypopharyngé* ; *laryngo-pharyngé*
T-58082 : *follicules lymphoïdes agminés de l'intestin grêle* ; *plaques de Peyer*
T-59900 : *anus, SAI* ; *anal* ; *ano-*
T-62260 : *cellule parenchymateuse hépatique* ; *hépatocyte*
T-70250 : *appareil génital, SAI* ; *génital*
T-82000 : *vagin, SAI* ; *vaginal* ; *vagino-*
T-94000 : *testicule, SAI* ; *gonade masculine, SAI* ; *testiculaire* ; *testiculo-*
T-A0421 : *macroglie* ; *macroglial*
T-A1112 : *méninges cérébrales, SAI* ; *méninges crâniennes, SAI*
T-A4000 : *thalamus, SAI* ; *thalamique*
T-A8020 : *nerf olfactif, SAI* ; *premier nerf crânien*
T-A8780 : *nerf accessoire, SAI* ; *onzième nerf crânien* ; *nerf spinal*
T-AA200 : *cornée, SAI* ; *cornéen*
T-AB320 : *membrane du tympan, SAI* ; *tympan, SAI*
T-B6050 : *cellule parafolliculaire de la glande thyroïde* ; *cellules C*
T-C1120 : *pronormoblaste* ; *proérythroblaste*
T-C1700 : *monocyte, SAI* ; *monocytaire*
T-C4900 : *ganglions lymphatiques du corps dans son entier* ; *ganglions lymphatiques dans plusieurs territoires* ; *tous les ganglions*
T-D0AA0 : *sexe féminin, SAI* ; *féminin*
T-D2700 : *périnée, SAI* ; *périnéal*
T-D4500 : *mésentère, SAI* ; *mésentérique*
T-E0000 : *cellule, SAI* ; *cellulaire*
T-F1900 : *emplacement de nidation du foetus* ; *emplacement foeto-maternel* ; *emplacement du placenta*

A.1.2 Règles (fréquence ≥ 5)

Français (153 règles sur 567)

1140 $\epsilon|s$
290 $\epsilon|e$
143 $\epsilon|es$
74 $e|ique$
67 $e|aire$
64 $e|que$
60 $e|s$
55 $é|ation$
53 $x|se$
43 $l|ux$
42 $o|al$
42 $e|o$
40 $le|ux$
40 $e|ose$
38 $o|ite$
36 $es|aire$
36 $e|ome$
34 $\epsilon|ment$
34 $f|ve$
33 $\epsilon|ne$
33 $\epsilon|le$
32 $e|é$
31 $me|sarcome$
31 $e|atose$
28 $f|on$
26 $o|aire$
23 $se|tique$
23 $on|ve$
23 $o|ique$
23 $me|se$
23 $e|ienne$
21 $\epsilon|ux$
20 $\epsilon|se$
20 $ome|ique$
20 $eux|ose$
20 $e|ques$
20 $ée|ation$
20 $e|ale$
19 $\epsilon|use$
19 $é|ant$
19 $e|al$
18 $t|ce$
18 $s|ne$
18 $me|ïde$
18 $me|blastome$
18 $f|ves$
18 $es|ose$
18 $es|ome$
18 $é|ement$
17 $x|ses$
17 $\epsilon|me$
17 $e|ée$
17 $e|ateux$
16 $ose|euse$
16 $es|que$
16 $e|ien$
16 $e|ie$
15 $ose|aire$
15 $\epsilon|les$
15 $e|oïde$
14 $e|ité$
13 $se|ïde$
13 $ome|aire$
13 $nt|tion$
13 $al|um$
12 $e|ateuse$
11 $o|um$
11 $ose|ique$
11 $nte|tion$
11 $ïde|sarcome$
11 $e|urie$
11 $\epsilon|ement$
11 $al|ome$
10 $\epsilon|ur$
10 $o|eux$
10 $é|ure$
10 $é|es$
10 $e|ant$
9 $s|ux$
9 $o|é$
9 $\epsilon|o$
9 $\epsilon|nelle$
9 $\epsilon|nel$
9 $\epsilon|n$
9 $me|fibrome$
9 $me|blastique$
9 $\epsilon|ion$
9 $eux|aire$
9 $é|ion$
9 $é|aire$
9 $\epsilon|é$
9 $\epsilon|al$
8 $us|ale$
8 $te|enne$
8 $sarcome|blastome$
8 $sarcome|blastique$
8 $me|lipome$
8 $matose|blastome$
8 $\epsilon|l$
8 $ite|ose$
8 $ite|opathie$
8 $ire|tion$
8 $ique|oïde$
8 $\epsilon|ique$
8 $ïde|mateuse$
8 $e|ocyte$
8 $e|isation$
8 $e|ante$
8 $ale|ome$
8 $\epsilon|aire$
8 $a|e$
7 $\epsilon|x$
7 $que|sme$
7 $ome|ales$
7 $\epsilon|ome$
7 $o|ien$
7 $me|carcinome$
7 $ion|oire$
7 $euse|oïde$
7 $es|oïde$
7 $é|cation$
7 $cytome|blastome$
6 $us|ique$
6 $se|pathie$
6 $se|matose$
6 $o|us$
6 $\epsilon|isé$
6 $ïde|blastique$
6 $eux|oïde$
6 $e|émie$
6 $e|atique$
6 $a|o$
6 $al|us$
5 $x|gé$
5 $\epsilon|virus$
5 $\epsilon|ssement$
5 $ses|tique$
5 $qué|cation$
5 $ome|inique$
5 $o|ienne$
5 $\epsilon|né$
5 $me|cytaire$
5 $le|ilité$
5 $\epsilon|isation$
5 $ire|ris$
5 $ion|oires$
5 $gène|sarcome$
5 $e|yome$
5 $er|ère$
5 $ène|énique$
5 $ème|émateux$
5 $e|iase$
5 $aire|euse$

Anglais (747 règles sur 3 039)

1285 $\epsilon|s$
362 $a|us$
330 $\epsilon|e$
290 $\epsilon|i$
281 $a|um$
243 $m|s$
225 $\epsilon|a$
218 $\epsilon|ing$
214 $\epsilon|is$
193 $\epsilon|ed$
192 $\epsilon|al$
188 $a|osis$
166 $us|idae$
166 $\epsilon|l$
163 $\epsilon|us$
155 $ed|ing$
147 $\epsilon|es$
144 $a|c$
123 $\epsilon|er$
123 $a|e$
120 $ed|ion$
118 $otomy|ectomy$
118 $a|idae$
115 $ic|ate$
115 $\epsilon|d$
113 $\epsilon|y$
112 $e|us$
109 $\epsilon|ic$
105 $e|is$
103 $on|ve$
103 $i|us$
101 $us|osis$
95 $\epsilon|n$
94 $y|ic$
93 $\epsilon|um$
93 $al|um$
91 $sis|tic$
91 $er|ing$
91 $e|ing$
87 $al|us$
85 $\epsilon|r$
83 $\epsilon|ly$
78 $e|ase$
77 $\epsilon|ase$
74 $s|ing$
69 $y|ies$
69 $\epsilon|ia$
68 $\epsilon|ii$
66 $ectomy|oplasty$
65 $ng|on$
63 $e|ic$
61 $o|al$
61 $a|s$
60 $or|ion$
60 $e|um$
59 $\epsilon|osis$
59 $es|is$
58 $s|al$
58 $\epsilon|ectomy$
57 $plasty|rrhaphy$
57 $al|ectomy$
56 $\epsilon|ta$
53 $d|s$
52 $ed|ation$
52 $a|o$
51 $idae|osis$
51 $al|otomy$
51 $al|itis$
50 $dae|formes$
49 $um|osis$
48 $\epsilon|plasty$
48 $\epsilon|fishes$
47 $emia|aemia$
47 $\epsilon|ae$
46 $\epsilon|transferase$
46 $e|osis$
46 $e|al$
46 $\epsilon|ceae$
45 $o|ectomy$
45 $\epsilon|ness$
45 $\epsilon|ity$
45 $e|l$
44 $y|ia$
44 $\epsilon|t$
44 $s|dae$
44 $\epsilon|ine$
44 $ic|us$
44 $ic|osis$
44 $\epsilon|ensis$
44 $e|ion$
43 $es|osis$
42 $s|t$
42 $\epsilon|sis$
41 $tomy|stomy$
41 $m|phy$
41 $c|s$
40 $us|ion$
40 $\epsilon|in$
40 $\epsilon|head$
40 $a|ed$
39 $\epsilon|otomy$
39 $\epsilon|o$
39 $\epsilon|ne$
39 $ion|ory$
39 $e|y$
39 $e|ity$
39 $al|oplasty$
37 $\epsilon|na$
37 $\epsilon|itis$
37 $\epsilon|idae$
37 $i|al$
37 $ectomy|ostomy$
37 $a|ous$
37 $a|ectomy$
36 $t|ce$

36 *es|ing*
36 *ed|us*
35 *ϵ|tosis*
35 *ϵ|ate*
34 *se|te*
34 *i|um*
34 *ic|yl*
34 *c|n*
33 *s|ic*
33 *osis|iasis*
33 *ϵ|ers*
33 *al|ostomy*
32 *yl|ate*
32 *ma|sis*
32 *itis|opathy*
32 *ic|ectomy*
32 *ia|us*
32 *es|ic*
31 *ϵ|x*
31 *s|us*
31 *ϵ|ism*
31 *c|sm*
31 *a|ic*
31 *a|es*
30 *ing|ation*
30 *ϵ|ation*
30 *a|itis*
29 *ϵ|ted*
29 *ϵ|te*
29 *s|osis*
29 *itis|osis*
29 *ϵ|ian*
29 *ed|ive*
29 *a|otomy*
28 *y|ing*
28 *us|ectomy*
28 *s|y*
28 *o|um*
28 *o|ic*
28 *nt|tion*
28 *ϵ|ion*
28 *ic|ase*
28 *exy|lasty*
28 *e|uria*
28 *es|idae*
28 *c|tis*
28 *asis|osis*
27 *y|us*
27 *ϵ|yl*
27 *ϵ|virus*
27 *stomy|rrhaphy*
27 *ic|ostomy*
27 *a|is*
26 *ve|ons*
26 *us|ine*
26 *um|ectomy*
26 *ϵ|tion*
26 *ϵ|tic*
26 *ϵ|oma*
26 *l|tion*
26 *emia|uria*
26 *e|i*
26 *ed|ions*
26 *al|osis*
26 *al|is*
26 *a|ensis*
25 *ϵ|tous*
25 *sa|us*
25 *ϵ|less*
25 *ϵ|ization*
25 *es|us*
25 *ϵ|alis*
24 *s|ne*
24 *ϵ|ry*
24 *n|s*
24 *e|o*
24 *e|ar*
24 *ea|oea*
24 *copy|tomy*
24 *al|ed*
24 *a|iasis*
23 *um|otomy*
23 *ϵ|tis*
23 *s|um*
23 *o|us*
23 *ng|ve*
23 *ϵ|ella*
23 *ϵ|ctomy*
23 *ar|us*
23 *ar|ectomy*
23 *a|ia*
22 *ϵ|tidae*
22 *s|tis*
22 *ol|yl*
22 *ϵ|mia*
22 *l|ta*
22 *i|s*
22 *ar|itis*
22 *al|oma*
21 *yl|ine*
21 *us|oides*
21 *ϵ|ns*
21 *ma|sarcoma*
21 *id|ma*
21 *ic|oma*
21 *ϵ|ica*
21 *ϵ|ene*
21 *al|es*
20 *y|ed*
20 *us|sis*
20 *us|ation*
20 *us|ate*
20 *ϵ|uria*
20 *o|is*
20 *le|ular*
20 *i|ar*
20 *e|in*
20 *ϵ|ata*
20 *ϵ|an*
19 *us|otomy*
19 *um|idae*
19 *ϵ|ous*
19 *l|ne*
19 *in|ate*
19 *ic|oid*
19 *e|ae*
19 *c|st*
19 *ϵ|ated*
18 *ϵ|weed*
18 *us|ing*
18 *se|tion*
18 *l|te*
18 *ϵ|h*
18 *e|ta*
18 *e|r*
18 *e|ation*
18 *a|oplasty*
18 *al|ic*
18 *ae|osis*
17 *us|sus*
17 *tomy|centesis*
17 *ϵ|se*
17 *oma|itis*
17 *ϵ|ol*
17 *ϵ|line*
17 *ϵ|ive*
17 *is|ate*
17 *i|oplasty*
17 *in|um*
17 *ic|ous*
17 *ϵ|emia*
17 *ϵ|ea*
17 *a|sm*
17 *a|on*
17 *a|n*
17 *ae|otomy*
17 *ae|itis*
16 *yl|ase*
16 *us|ous*
16 *us|oplasty*
16 *um|ate*
16 *sis|tomy*
16 *s|idae*
16 *se|kinase*
16 *or|ing*
16 *ns|ting*
16 *e|ium*
16 *eal|itis*
16 *de|ne*
16 *c|um*
16 *ϵ|ar*
16 *a|oma*
16 *a|ity*
16 *a|ca*
15 *us|ism*
15 *us|inus*
15 *us|aceae*
15 *um|ation*
15 *um|ata*
15 *t|cy*
15 *s|ty*
15 *sis|pathy*
15 *sis|idea*
15 *s|ectomy*
15 *osis|ectomy*
15 *ϵ|oid*
15 *o|ar*
15 *ϵ|ized*
15 *is|osis*
15 *ϵ|ina*
15 *ϵ|icum*
15 *ic|oplasty*
15 *ϵ|gia*
15 *ϵ|gen*
15 *e|yl*
15 *emic|aemia*
15 *e|ine*
15 *c|ne*
15 *algia|odynia*
14 *y|is*
14 *us|sum*
14 *us|iformes*
14 *um|oplasty*
14 *um|aceae*
14 *ta|lis*
14 *s|less*
14 *sis|lysis*
14 *s|ctomy*
14 *ϵ|phosphate*
14 *ϵ|m*
14 *l|s*
14 *is|ectomy*
14 *in|osis*
14 *id|tomy*
14 *e|yltransferase*
14 *e|otomy*
14 *cyte|blast*
14 *ase|osis*
14 *a|oid*
14 *al|oscopy*
14 *a|ella*
13 *ϵ|water*
13 *um|tis*
13 *tomy|lithotomy*
13 *s|ia*
13 *ria|tus*
13 *ne|dae*
13 *ma|blastoma*
13 *ϵ|lla*
13 *is|eal*
13 *in|ase*
13 *ia|tic*
13 *e|idae*
13 *c|as*
13 *ar|es*
13 *a|oses*
13 *a|oidea*
13 *al|opexy*
13 *al|ography*
13 *a|cus*
12 *us|iasis*
12 *us|alis*
12 *s|otomy*
12 *sis|genic*
12 *s|dis*
12 *s|asis*
12 *ϵ|oplasty*
12 *ϵ|nic*
12 *ne|dine*
12 *is|otomy*
12 *id|us*
12 *ic|on*
12 *ic|amide*
12 *ic|aldehyde*
12 *ϵ|form*
12 *ϵ|eus*
12 *es|otomy*
12 *es|itis*
12 *e|ive*
12 *d|ment*
12 *ϵ|ary*
12 *ar|osa*
12 *an|ca*
12 *al|tis*
12 *alis|itis*
12 *a|ida*
12 *ϵ|aceae*
11 *ty|zation*
11 *sis|tica*
11 *sia|tic*
11 *o|s*
11 *ol|ine*
11 *o|ine*
11 *ϵ|nidae*
11 *l|n*
11 *itis|odynia*
11 *i|osis*
11 *i|oid*
11 *ia|osis*
11 *ϵ|foot*
11 *e|ia*
11 *d|l*
11 *a|tis*
11 *ate|ine*
11 *asis|idae*
11 *ase|ine*
11 *a|orrhaphy*
11 *a|opsis*
11 *an|um*
11 *an|ensis*
11 *al|osus*
10 *us|ostomy*
10 *us|inae*
10 *uria|aemia*
10 *um|ous*
10 *stomy|lithotomy*
10 *stomy|cutaneous*
10 *s|ng*
10 *ϵ|si*
10 *s|ation*
10 *ϵ|ral*
10 *ϵ|orum*
10 *o|ation*
10 *l|r*
10 *le|ularis*
10 *les|ular*
10 *le|ility*
10 *ϵ|lar*
10 *itis|opathies*
10 *ϵ|eum*
10 *e|rum*
10 *e|oid*
10 *ϵ|ening*
10 *ϵ|cyte*
10 *ϵ|atic*
10 *ate|okinase*
10 *ase|uria*
10 *ar|um*
10 *an|us*
10 *al|ine*
10 *al|ctomy*
9 *us|ida*
9 *um|ary*
9 *ϵ|ty*
9 *stomy|centesis*

9 *s|tic*
9 *sis|derma*
9 *s|ca*
9 *sarcoma|blastoma*
9 *ϵ|sa*
9 *ϵ|ping*
9 *ϵ|phenol*
9 *ϵ|orrhaphy*
9 *ϵ|oides*
9 *n|us*
9 *m|tic*
9 *ma|tic*
9 *is|oplasty*
9 *in|oid*
9 *id|cyte*
9 *ic|um*
9 *ic|ography*
9 *ϵ|iasis*
9 *ϵ|ful*
9 *es|ation*
9 *e|n*
9 *emic|aemic*
9 *e|ism*
9 *dae|asis*
9 *c|de*
9 *ble|tion*
9 *ar|osis*
9 *ϵ|amide*
9 *al|ovesical*
9 *al|opathy*
9 *al|ogram*
9 *al|in*
9 *a|id*
8 *yl|idase*
8 *y|ium*
8 *y|ied*
8 *x|geal*
8 *x|cis*
8 *x|cidae*
8 *us|orrhaphy*
8 *us|oidea*
8 *uria|idase*
8 *um|ostomy*
8 *s|tosis*
8 *stomy|ileostomy*
8 *s|ion*
8 *se|ting*
8 *s|didae*
8 *s|des*
8 *ry|tion*
8 *ϵ|pyruvate*
8 *ϵ|phosphotransferase*
8 *ϵ|peptidase*
8 *ϵ|ped*
8 *os|us*
8 *ϵ|ostomy*
8 *osis|uria*
8 *ϵ|olysis*
8 *o|es*
8 *nt|ting*
8 *n|tomy*
8 *n|lysis*
8 *ϵ|nal*
8 *m|phic*
8 *ϵ|lgia*
8 *ϵ|lessness*
8 *itis|osclerosis*
8 *ing|ement*
8 *ϵ|iformes*
8 *i|es*
8 *id|sa*
8 *icus|ectomy*
8 *i|ate*
8 *ia|os*
8 *ϵ|iae*
8 *ϵ|grass*
8 *es|ous*
8 *e|ous*
8 *emia|aemic*
8 *e|ii*
8 *ed|cation*
8 *ed|atus*
8 *ed|able*
8 *e|ated*
8 *eal|ostomy*
8 *ca|sm*
8 *ϵ|berry*
8 *ϵ|ating*
8 *ase|inase*
8 *a|ostomy*
8 *a|ography*
8 *ϵ|ans*
8 *al|one*
8 *al|oid*
8 *al|ocele*
7 *yl|inase*
7 *yl|ide*
7 *y|ization*
7 *y|ius*
7 *y|al*
7 *ϵ|viridae*
7 *ϵ|viral*
7 *ve|onal*
7 *us|ica*
7 *us|ary*
7 *um|orrhaphy*
7 *um|asis*
7 *stomy|gastrostomy*
7 *sis|rrhaphy*
7 *sis|pathies*
7 *sis|matosis*
7 *s|form*
7 *s|cular*
7 *r|ted*
7 *ϵ|peritoneal*
7 *ory|ions*
7 *ϵ|opathy*
7 *ol|ate*
7 *ϵ|ography*
7 *o|eal*
7 *ϵ|nuclear*
7 *ϵ|nsis*
7 *ma|fibroma*
7 *itis|omyelitis*
7 *is|tic*
7 *ϵ|ish*
7 *ion|ure*
7 *ine|amine*
7 *ic|opathy*
7 *ic|olithotomy*
7 *ϵ|iana*
7 *i|an*
7 *ia|emia*
7 *e|tion*
7 *er|ra*
7 *emia|osis*
7 *ed|ure*
7 *ed|ory*
7 *ed|ment*
7 *ϵ|cytic*
7 *ce|tial*
7 *ϵ|cal*
7 *ate|oyl*
7 *ate|ite*
7 *ase|transferase*
7 *ase|ose*
7 *a|oscopy*
7 *a|opexy*
7 *a|ogram*
7 *ant|ive*
7 *al|ous*
7 *al|odynia*
7 *al|ocentesis*
7 *ae|ous*
6 *yl|ium*
6 *y|iness*
6 *y|ii*
6 *y|ial*
6 *y|ectomy*
6 *x|gitis*
6 *us|sal*
6 *us|ian*
6 *us|ase*
6 *um|opexy*
6 *tomy|myotomy*
6 *ϵ|stone*
6 *s|tes*
6 *s|pox*
6 *s|ous*
6 *s|le*
6 *se|tase*
6 *ϵ|serine*
6 *s|daceae*
6 *s|da*
6 *s|cle*
6 *r|tum*
6 *plasia|trophy*
6 *ϵ|out*
6 *osis|atidae*
6 *ϵ|oscopy*
6 *ϵ|ocyte*
6 *n|tic*
6 *n|sis*
6 *n|de*
6 *ϵ|nase*
6 *le|ulectomy*
6 *ium|osis*
6 *i|osa*
6 *ion|ation*
6 *ion|able*
6 *ϵ|io*
6 *ϵ|inum*
6 *in|ol*
6 *ine|ylic*
6 *ide|omethane*
6 *id|blastoma*
6 *ic|ol*
6 *ia|otomy*
6 *ϵ|ial*
6 *ϵ|ette*
6 *e|ry*
6 *er|ridae*
6 *ectomy|oduodenal*
6 *ea|us*
6 *ded|sion*
6 *cytic|blastic*
6 *ϵ|cysteine*
6 *ϵ|ations*
6 *ate|ectomy*
6 *ary|ectomy*
6 *ar|opathy*
6 *a|olysis*
6 *ϵ|amphetamine*
6 *a|d*
5 *y|ous*
5 *yl|oyl*
5 *yl|one*
5 *yl|oid*
5 *y|ism*
5 *y|ie*
5 *y|eutic*
5 *y|as*
5 *xillary|ndibular*
5 *x|cal*
5 *us|sitis*
5 *us|oscopy*
5 *us|opexy*
5 *us|oenterostomy*
5 *us|odidae*
5 *um|oscopy*
5 *um|orum*
5 *tis|opathy*
5 *ϵ|snake*
5 *sm|zation*
5 *s|lar*
5 *sis|therapy*
5 *sis|poiesis*
5 *sis|matous*
5 *r|ur*
5 *ϵ|pleural*
5 *plasty|intestinal*
5 *pathy|sclerosis*
5 *pathy|myelitis*
5 *os|ine*
5 *on|ectomy*
5 *ϵ|olithotomy*
5 *ol|ene*
5 *ϵ|ogenic*
5 *ϵ|oenterostomy*
5 *ϵ|ocolic*
5 *ϵ|nut*
5 *n|rrhaphy*
5 *ne|sm*
5 *n|al*
5 *ϵ|mycosis*
5 *ma|carcinoma*
5 *lysis|megaly*
5 *l|rum*
5 *l|mine*
5 *le|ulo*
5 *ϵ|led*
5 *is|orrhaphy*
5 *in|ene*
5 *ii|ella*
5 *ϵ|iform*
5 *ϵ|ies*
5 *ϵ|ids*
5 *ide|oethane*
5 *ida|oidea*
5 *ic|ogastrostomy*
5 *ic|ocyte*
5 *ic|esis*
5 *ϵ|ianus*
5 *ϵ|hydrolase*
5 *eum|itis*
5 *e|ula*
5 *esthesia|aesthesia*
5 *ϵ|esterase*
5 *es|iasis*
5 *e|si*
5 *ery|ical*
5 *e|oplasty*
5 *e|opathy*
5 *ens|ing*
5 *emoglobin|aemoglobin*
5 *eic|oea*
5 *ed|is*
5 *d|y*
5 *cytic|penia*
5 *c|an*
5 *ϵ|buck*
5 *ϵ|body*
5 *blastic|sarcoma*
5 *as|ic*
5 *ase|izing*
5 *ϵ|arthritis*
5 *ar|oid*
5 *a|odynia*
5 *a|ocele*
5 *an|ense*
5 *an|cus*
5 *al|orum*
5 *al|on*
5 *al|ism*
5 *a|iosis*
5 *a|iform*
5 *ϵ|aciduria*

Russe (205 règles sur 834)

460 *ая|ый*
243 *е|й*
212 *є|а*
207 *ая|ое*
196 *ая|ой*
186 *ое|ый*
140 *ая|ые*
113 *ой|ый*
103 *ая|ого*
98 *о|ая*
97 *о|ый*
96 *ый|ого*
96 *а|ы*
86 *ая|ий*
81 *и|я*
79 *й|х*
74 *е|го*
73 *й|го*
69 *є|е*
67 *е|я*
62 *є|й*
58 *є|ный*
53 *ое|ые*
53 *ие|ный*
52 *ая|ых*
51 *є|ов*
51 *є|ная*
50 *ой|ые*
50 *є|ы*
49 *й|я*
49 *а|и*
40 *и|ь*
38 *ые|ого*
37 *є|м*
36 *й|м*
35 *ий|ое*
34 *є|ом*
32 *о|ые*
32 *є|го*
32 *а|о*
32 *а|ная*
30 *ых|ого*
29 *є|оз*
28 *є|тоз*
27 *а|ов*
26 *ма|саркома*
26 *ая|ие*
25 *ой|ых*
25 *ия|ный*
24 *ие|ная*
24 *а|ный*
24 *ация|ированный*
23 *є|овый*
22 *ое|ых*
22 *ие|ое*
21 *з|ма*
20 *ый|ость*
20 *а|ом*
19 *о|ых*
19 *ий|ого*
19 *и|ей*
19 *я|ей*
19 *а|овая*
18 *є|ное*
18 *ма|идная*
18 *ая|их*
18 *з|тический*
17 *є|арная*
16 *ие|ого*
16 *я|ческий*
16 *ы|ов*
16 *ы|ная*
16 *ая|оклеточная*
16 *є|арный*
16 *є|тозный*
15 *є|но*
15 *є|и*
15 *ая|ым*
14 *ома|альная*
14 *є|ные*
14 *ма|бластома*
14 *а|ной*
13 *є|овая*
13 *я|ческая*
13 *ы|альная*
12 *ка|очный*
12 *я|альная*
12 *ы|ной*
12 *є|тозная*
11 *ома|арная*
11 *ая|оклеточный*
11 *з|тическая*
10 *оз|арная*
10 *є|о*
10 *є|клеточная*
10 *є|льный*
10 *ма|фиброма*
10 *к|чная*
10 *ий|яя*
10 *ия|ная*
10 *и|ью*
10 *е|ы*
10 *ы|ный*
10 *ый|оклеточный*
10 *ым|ого*
10 *а|оз*
10 *а|ное*
10 *т|я*
10 *т|ция*
9 *о|альный*
9 *к|чный*
9 *ее|яя*
9 *є|ический*
9 *в|з*
9 *я|онная*
9 *ь|ей*
9 *я|альный*
9 *я|ческое*
9 *ы|атоз*
9 *є|альный*
9 *саркома|бластома*
8 *оз|арный*
8 *ная|ома*
8 *є|клеточный*
8 *ма|генная*
8 *их|ой*
8 *я|онный*
8 *є|ю*
8 *а|овые*
8 *а|ярный*
8 *ция|тивный*
7 *о|ный*
7 *ки|ок*
7 *ма|липома*
7 *й|альная*
7 *матоз|бластома*
7 *и|ная*
7 *ие|ные*
7 *з|идная*
7 *и|ческий*
7 *є|ирующий*
7 *ь|ный*
7 *ь|ная*
7 *ы|ные*
7 *а|овой*
7 *альная|оидная*
7 *є|х*
6 *ома|альные*
6 *є|овое*
6 *ов|арная*
6 *ки|очный*
6 *мы|идная*
6 *й|альный*
6 *ма|карцинома*
6 *з|матоз*
6 *й|т*
6 *и|ный*
6 *ей|яя*
6 *ик|ый*
6 *ие|ый*
6 *ия|ьное*
6 *ия|ьный*
6 *є|ия*
6 *и|ческое*
6 *є|ированный*
6 *яя|его*
6 *ь|ый*
6 *а|истый*
6 *ая|ости*
6 *х|ми*
6 *ция|тивная*
6 *цитома|бластома*
5 *оз|ная*
5 *ок|ковый*
5 *ома|иновая*
5 *ка|очно*
5 *є|ма*
5 *к|чного*
5 *ения|ительная*
5 *е|ный*
5 *ие|яя*
5 *зия|стический*
5 *зия|стическая*
5 *и|ая*
5 *ит|опатия*
5 *є|ирование*
5 *ь|ной*
5 *ь|но*
5 *я|альное*
5 *ы|ное*
5 *ы|но*
5 *ый|ный*
5 *ых|оклеточный*
5 *ых|оклеточная*
5 *а|озное*
5 *є|ами*
5 *а|евой*
5 *а|ический*
5 *ая|ный*
5 *ая|ее*
5 *а|ярная*
5 *т|нт*
5 *є|ся*
5 *ции|тивный*
5 *є|у*

A.1.3 Suffixes (fréquence ≥ 5)

Français (130 suffixes sur 447)

1254 *-s*
1046 *-e*
268 *-es*
243 *-o*
221 *-é*
212 *-aire*
169 *-me*
166 *-se*
165 *-ique*
143 *-ome*
141 *-ose*
114 *-ux*
101 *-al*
99 *-que*
97 *-x*
83 *-ïde*
81 *-f*
80 *-le*
80 *-ation*
79 *-sarcome*
63 *-eux*
57 *-ve*
56 *-on*
54 *-ite*
51 *-ne*
51 *-l*
51 *-blastome*
50 *-oïde*
41 *-euse*
39 *-ale*
38 *-ée*
38 *-ant*
35 *-um*
35 *-ment*
35 *-atose*
34 *-us*
34 *-tique*
34 *-ion*
33 *-cytaire*
32 *-tion*
32 *-ienne*
31 *-ement*
31 *-blastique*
29 *-ien*
24 *-ire*
22 *-ses*
22 *-ie*
22 *-ce*
21 *-ure*
21 *-ques*
21 *-matose*
19 *-use*
18 *-ves*
18 *-t*
18 *-gène*
17 *-ité*
17 *-ateux*
16 *-les*
16 *-a*
15 *-nt*
15 *-carcinome*
14 *-opathie*
14 *-isation*
13 *-ocyte*
13 *-n*
13 *-fibrome*
13 *-émie*
12 *-cation*
12 *-ateuse*
11 *-urie*
11 *-sclérose*
11 *-nte*
10 *-ur*
10 *-te*
10 *-pathie*
10 *-mateuse*
10 *-ène*
10 *-ante*

10 *-ales*
9 *-osarcome*
9 *-nelle*
9 *-nel*
9 *-gliome*
9 *-cytome*
8 *-lipome*
8 *-enne*
8 *-en*
7 *-sme*
7 *-oire*
7 *-logique*
7 *-ielle*
7 *-ère*
7 *-er*
7 *-énique*
7 *-cellulaire*
7 *-atique*
7 *-adénome*
6 *-ylique*
6 *-plasie*
6 *-oblastome*
6 *-isé*
6 *-iques*
6 *-ine*
6 *-glie*
6 *-génique*
6 *-endothéliome*
5 *-yome*
5 *-virus*
5 *-ulaire*
5 *-ssement*
5 *-ris*
5 *-qué*
5 *-oires*
5 *-né*
5 *-mateux*
5 *-logue*
5 *-kératose*
5 *-inome*
5 *-inique*
5 *-ilité*
5 *-iforme*
5 *-iel*
5 *-iase*
5 *-greffe*
5 *-gé*
5 *-ème*
5 *-émateux*
5 *-dermie*
5 *-chondrosarcome*
5 *-atif*

Anglais (460 suffixes sur 2 498)

2484 *-s*
2247 *-a*
1970 *-us*
1696 *-e*
1233 *-al*
989 *-um*
984 *-ic*
921 *-osis*
781 *-ing*
761 *-ed*
642 *-ectomy*
613 *-y*
600 *-i*
584 *-is*
531 *-es*
484 *-idae*
401 *-o*
386 *-c*
367 *-l*
365 *-otomy*
364 *-ate*
361 *-ion*
348 *-itis*
335 *-sis*
319 *-m*
308 *-ase*
262 *-n*
252 *-ia*
245 *-er*
227 *-yl*
225 *-on*
221 *-ar*
220 *-ation*
213 *-oplasty*
211 *-ine*
211 *-d*
210 *-tic*
166 *-in*
162 *-ve*
157 *-ous*
151 *-ne*
150 *-tomy*
150 *-r*
141 *-t*
139 *-ostomy*
138 *-ae*
131 *-tion*
130 *-x*
126 *-emia*
125 *-dae*
124 *-stomy*
124 *-id*
120 *-tis*
119 *-plasty*
119 *-ma*
117 *-uria*
113 *-oma*
110 *-ta*
110 *-se*
109 *-oid*
106 *-ity*
105 *-iasis*
102 *-ng*
102 *-an*
101 *-rrhaphy*
95 *-le*
90 *-ii*
83 *-te*
83 *-sm*
83 *-ly*
83 *-ensis*
82 *-or*
82 *-ol*
81 *-ism*
81 *-aemia*
78 *-asis*
76 *-opathy*
76 *-ies*
75 *-ive*
73 *-transferase*
68 *-eal*
56 *-ea*
55 *-ness*
54 *-ory*
52 *-alis*
51 *-odynia*
51 *-formes*
51 *-ella*
51 *-ary*
50 *-ca*
49 *-sarcoma*
49 *-ctomy*
48 *-fishes*
47 *-orrhaphy*
46 *-tosis*
46 *-ns*
46 *-ceae*
46 *-algia*
45 *-ography*
45 *-ian*
45 *-ce*
44 *-ty*
44 *-ry*
43 *-ted*
43 *-ene*
43 *-aceae*
42 *-sa*
42 *-ium*
42 *-cyte*
41 *-phy*
41 *-pathy*
41 *-oidea*
41 *-ide*
41 *-de*
40 *-head*
39 *-nt*
39 *-less*
38 *-na*
38 *-lysis*
38 *-ization*
38 *-as*
37 *-oscopy*
36 *-ular*
36 *-ers*
36 *-cytic*
36 *-ata*
35 *-ting*
35 *-opexy*
35 *-ated*
34 *-oides*
33 *-ions*
32 *-virus*
32 *-kinase*
32 *-ica*
30 *-oea*
30 *-lasty*
30 *-ida*
30 *-blastoma*
28 *-iformes*
28 *-form*
28 *-exy*
28 *-centesis*
28 *-amide*
27 *-tidae*
26 *-tous*
26 *-os*
26 *-ons*
26 *-lithotomy*
26 *-genic*
26 *-emic*
26 *-cal*
24 *-osa*
24 *-ial*
24 *-gen*
24 *-copy*
23 *-mia*
22 *-idea*
22 *-h*
21 *-ral*
20 *-weed*
20 *-therapy*
20 *-orum*
20 *-ment*
20 *-idase*
20 *-ical*
20 *-cus*
20 *-blastic*
19 *-zation*
19 *-yltransferase*
19 *-st*
19 *-ocyte*
19 *-nic*
19 *-inus*
19 *-inae*
19 *-ac*
18 *-sia*
18 *-les*
18 *-lar*
18 *-graphy*
17 *-sus*
17 *-one*
17 *-olysis*
17 *-olithotomy*
17 *-ogram*
17 *-ocele*
17 *-line*
17 *-inase*
17 *-gia*
17 *-en*
17 *-dystrophy*
17 *-aemic*
16 *-oses*
16 *-lla*
16 *-icum*
16 *-eum*
16 *-cy*
16 *-atic*
15 *-tus*
15 *-si*
15 *-sclerosis*
15 *-rum*
15 *-plasia*
15 *-otic*
15 *-megaly*
15 *-logic*
15 *-ized*
15 *-ina*
15 *-blast*
14 *-ure*
14 *-sum*
14 *-ria*
14 *-phosphate*
14 *-ovesical*
14 *-osus*
14 *-osclerosis*
14 *-mycosis*
14 *-lis*
14 *-lgia*
14 *-iosis*
14 *-cutaneous*
14 *-able*
13 *-ylic*
13 *-water*
13 *-tica*
13 *-sion*
13 *-ogenic*
13 *-ocentesis*
13 *-itic*
13 *-iform*
13 *-iae*
13 *-foot*
13 *-eus*
13 *-ening*
13 *-dine*
13 *-atidae*
12 *-peptidase*
12 *-oyl*
12 *-opsis*
12 *-nidae*
12 *-matosis*
12 *-dis*
12 *-ans*
12 *-aldehyde*
11 *-proctostomy*
11 *-ovirus*
11 *-ospasm*
11 *-opathies*
11 *-nal*
11 *-io*
11 *-ility*
11 *-gastrostomy*
11 *-didae*
11 *-ations*
11 *-active*
10 *-ularis*
10 *-serine*
10 *-phosphotransferase*
10 *-ose*
10 *-omegaly*
10 *-okinase*
10 *-oenterostomy*
10 *-fibroma*
10 *-da*
10 *-cular*
10 *-carcinoma*
10 *-ble*
10 *-ating*
9 *-ur*
9 *-ulo*
9 *-tropin*
9 *-tory*
9 *-surgery*
9 *-pyruvate*
9 *-poietic*
9 *-poiesis*
9 *-ping*
9 *-phenol*
9 *-ped*
9 *-ovascular*
9 *-nsis*
9 *-nase*
9 *-myotomy*
9 *-myelitis*
9 *-mine*
9 *-matous*
9 *-ite*
9 *-icus*
9 *-ful*
9 *-ement*
9 *-des*
9 *-derma*
9 *-chalasia*
9 *-cation*
9 *-body*
9 *-berry*
9 *-atus*
9 *-atase*
9 *-amine*
8 *-ulotomy*
8 *-ula*
8 *-rectal*
8 *-que*
8 *-pleural*
8 *-phic*
8 *-penia*

8 *-orrhea*
8 *-opelvic*
8 *-onal*
8 *-omyotomy*
8 *-ogenase*
8 *-ogen*
8 *-odidae*
8 *-metry*
8 *-lessness*
8 *-ish*
8 *-ileostomy*
8 *-ied*
8 *-ids*
8 *-grass*
8 *-geal*
8 *-ferase*
8 *-ens*
8 *-endothelioma*
8 *-cytosis*
8 *-cystic*
8 *-cis*
8 *-cidae*
7 *-viridae*
7 *-viral*
7 *-vascular*
7 *-trophy*
7 *-tial*
7 *-tase*
7 *-sidase*
7 *-rrhage*
7 *-ra*
7 *-pox*
7 *-potassemia*
7 *-peritoneal*
7 *-pathies*
7 *-ostenosis*
7 *-omyelitis*
7 *-omycosis*
7 *-odesis*
7 *-ocolic*
7 *-nut*
7 *-nuclear*
7 *-myoma*
7 *-izing*
7 *-ius*
7 *-intestinal*
7 *-iana*
7 *-hydrolase*
7 *-gram*
7 *-go*
7 *-genous*
7 *-esis*
7 *-ent*
7 *-enia*
7 *-emoglobin*
7 *-eic*
7 *-ded*
7 *-cle*
7 *-cele*
7 *-ant*
7 *-amphetamine*
7 *-aemoglobin*
6 *-ylyltransferase*
6 *-ural*
6 *-ulosis*
6 *-ulectomy*
6 *-tum*
6 *-tes*
6 *-stone*
6 *-snake*
6 *-sal*
6 *-ridae*
6 *-out*
6 *-osity*
6 *-omethane*
6 *-ogenesis*
6 *-ogastrostomy*
6 *-oduodenal*
6 *-ocytosis*
6 *-nae*
6 *-led*
6 *-kokinesia*
6 *-itase*
6 *-inum*
6 *-insic*
6 *-iness*
6 *-ile*
6 *-ie*
6 *-hin*
6 *-gitis*
6 *-ette*
6 *-edema*
6 *-desis*
6 *-dectomy*
6 *-daceae*
6 *-cytoma*
6 *-cysteine*
6 *-cosis*
6 *-cephaly*
6 *-cardiography*
6 *-arthritis*
6 *-activity*
5 *-zing*
5 *-xillary*
5 *-ulus*
5 *-tia*
5 *-sy*
5 *-suria*
5 *-sitis*
5 *-otripsy*
5 *-otoxin*
5 *-otherapy*
5 *-osidase*
5 *-orenal*
5 *-opia*
5 *-ola*
5 *-oethane*
5 *-oderma*
5 *-nts*
5 *-ndibular*
5 *-matosa*
5 *-lysosome*
5 *-leiomyoma*
5 *-lalia*
5 *-kinetic*
5 *-kinesia*
5 *-inoma*
5 *-ianus*
5 *-globulin*
5 *-glioma*
5 *-fer*
5 *-eutic*
5 *-esthesia*
5 *-esterase*
5 *-ery*
5 *-ense*
5 *-emphraxis*
5 *-dynia*
5 *-dism*
5 *-dermatitis*
5 *-ctic*
5 *-cotomy*
5 *-coccus*
5 *-clasis*
5 *-chondrosarcoma*
5 *-chokinesis*
5 *-chokinesia*
5 *-cautery*
5 *-buck*
5 *-asthenia*
5 *-arthrosis*
5 *-aris*
5 *-anol*
5 *-aesthesia*
5 *-aciduria*

Russe (167 suffixes sur 673)

1464 *-ая*
1011 *-ый*
625 *-а*
569 *-й*
536 *-ое*
486 *-е*
398 *-ой*
342 *-о*
335 *-я*
314 *-ого*
313 *-ые*
279 *-ы*
277 *-и*
220 *-ный*
182 *-го*
181 *-ная*
163 *-ие*
162 *-ий*
159 *-ых*
140 *-ма*
104 *-ов*
92 *-х*
91 *-з*
87 *-ь*
83 *-оз*
75 *-м*
63 *-ома*
61 *-альная*
60 *-ей*
59 *-ом*
57 *-ия*
56 *-саркома*
51 *-идная*
47 *-арная*
45 *-альный*
44 *-ное*
44 *-бластома*
40 *-яя*
38 *-ированный*
37 *-ка*
36 *-ной*
36 *-т*
35 *-овая*
34 *-к*
33 *-ация*
32 *-тоз*
30 *-их*
30 *-ческий*
29 *-но*
29 *-арный*
28 *-оклеточный*
28 *-ные*
28 *-ция*
27 *-оклеточная*
27 *-ым*
26 *-овый*
26 *-тический*
25 *-ический*
22 *-очный*
22 *-матоз*
22 *-зия*
21 *-ок*
21 *-генная*
20 *-ость*
19 *-енный*
18 *-ки*
16 *-альные*
16 *-тозный*
16 *-тическая*
15 *-клеточная*
15 *-карцинома*
15 *-ческое*
15 *-цитарный*
14 *-ее*
13 *-ческая*
13 *-тивный*
13 *-фиброма*
12 *-льный*
12 *-бластный*
12 *-ю*
12 *-чный*
12 *-чная*
12 *-тозная*
11 *-озная*
11 *-ительная*
11 *-ирующий*
11 *-ью*
11 *-ярная*
10 *-оидная*
10 *-овидный*
10 *-оцит*
10 *-плазия*
10 *-ение*
10 *-ик*
10 *-истый*
10 *-в*
10 *-ьный*
10 *-ярный*
10 *-альное*
10 *-атоз*
10 *-ции*
10 *-цитома*
10 *-цитарная*
9 *-онная*
9 *-озный*
9 *-идный*
9 *-зии*
9 *-стический*
9 *-стическая*
9 *-у*
8 *-онное*
8 *-онный*
8 *-овые*
8 *-зная*
8 *-клеточный*
8 *-пластический*
8 *-ения*
8 *-ами*
8 *-с*
7 *-овой*
7 *-овое*
7 *-зный*
7 *-нт*
7 *-липома*
7 *-мы*
7 *-евой*
7 *-генный*
7 *-изация*
7 *-бластная*
7 *-ально*
6 *-опатия*
6 *-ованный*
6 *-очно*
6 *-оцитарная*
6 *-ости*
6 *-лейкоз*
6 *-ми*
6 *-матозная*
6 *-иновая*
6 *-емия*
6 *-его*
6 *-ирование*
6 *-ьное*
6 *-эндотелиома*
6 *-эпителиома*
6 *-яющий*
6 *-ающий*
6 *-склероз*
6 *-тивная*
6 *-ус*
5 *-озное*
5 *-офилия*
5 *-ковый*
5 *-кация*
5 *-ицома*
5 *-глиома*
5 *-иальный*
5 *-еское*
5 *-ительный*
5 *-ит*
5 *-атических*
5 *-атическая*
5 *-за*
5 *-много*
5 *-хондросаркома*
5 *-ся*
5 *-цированный*

A.1.4 Préfixes (fréquence ≥ 5)

Français (88 préfixes sur 2 879)

29 *myélo-*
25 *ostéo-*
22 *fibro-*
22 *fibr-*
21 *adéno-*
19 *lympho-*
16 *lipo-*
16 *granul-*
15 *angio-*
13 *plasmocyt-*
13 *immun-*
13 *chondro-*
12 *ménin-g*
12 *mélano-*
12 *lymphocyt-*
11 *hyalin-*
11 *histiocyt-*
11 *hémangio-*
10 *neuro-*
9 *ulcér-*
9 *plan-*
9 *néphro-*
9 *lymph-*
9 *hyper-*
9 *hépat-*
9 *corticosurrénal-*
8 *synovi-*
8 *papul-*
8 *myxo-*
8 *monocyt-*
8 *hépato-*
8 *glio-*
8 *dent-*
7 *tubul-*
7 *tubercul-*
7 *synovial-*
7 *radi-*
7 *papill-*
7 *fécal-*
7 *derm-*
7 *carcino-*
7 *acid-*
6 *spontané-*
6 *réticul-*
6 *rect-*
6 *pustul-*
6 *oligodendro-*
6 *oesophag-*
6 *kérato-*
6 *épithéli-*
6 *cutané-*
6 *bronch-*
6 *artério-*
6 *amygdal-*
5 *traché-*
5 *thrombo-*
5 *segment-*
5 *protéin-*
5 *plasmocyto-*
5 *pigment-*
5 *pharyng-*
5 *pharyn-*
5 *parasit-*
5 *palm-*
5 *microcyt-*
5 *métacarp-*
5 *mélanocyt-*
5 *macul-*
5 *ligament-*
5 *leucocyt-*
5 *léiomyo-*
5 *kyst-*
5 *immédiat-*
5 *histiocyto-*
5 *glomérul-*
5 *follicul-*
5 *épididym-*
5 *épiderm-*
5 *endométrio-*
5 *endométri-*
5 *cylindr-*
5 *coup-*
5 *class-*
5 *caus-*
5 *cardi-*
5 *bronchiol-*
5 *astrocyt-*
5 *adén-*

Anglais (892 préfixes sur 10 942)

68 *hyper-*
58 *ureter-*
55 *sept-*
50 *enter-*
49 *cyst-*
48 *osteo-*
43 *fibro-*
40 *neuro-*
38 *neur-*
36 *oesophag-*
36 *limb-*
36 *esophag-*
35 *rect-*
33 *chlor-*
32 *vagin-*
32 *urethr-*
32 *arteri-*
31 *hypo-*
31 *form-*
30 *hepat-*
29 *radi-*
29 *jejun-*
29 *carp-*
29 *cardi-*
28 *nephro-*
28 *gastr-*
27 *rhin-*
27 *myelo-*
26 *pleur-*
25 *medi-*
25 *dent-*
25 *brom-*
25 *aort-*
24 *scler-*
24 *choledoch-*
24 *anti-*
24 *adeno-*
23 *phosph-*
22 *ventricul-*
22 *plan-*
22 *phospho-*
22 *duoden-*
22 *derm-*
21 *tympan-*
21 *trans-*
21 *spin-*
21 *hepatic-*
21 *crani-*
21 *colo-*
21 *acet-*
20 *sulf-*
20 *pylor-*
20 *lympho-*
20 *cost-*
20 *conjunctiv-*
20 *angio-*
19 *uretero-*
19 *prostat-*
19 *post-*
19 *poly-*
19 *phen-*
19 *pallid-*
19 *hepato-*
19 *epididym-*
19 *dermato-*
19 *cord-*
19 *aden-*
19 *acid-*
18 *trache-*
18 *stat-*
18 *pericardi-*
18 *meth-*
18 *lymph-*
18 *lipo-*
18 *hydr-*
18 *erythro-*
18 *burs-*
18 *bull-*
18 *bronch-*
17 *retin-*
17 *pigment-*
17 *orbit-*
17 *line-*
17 *laryng-*
17 *lamin-*
17 *cephal-*
17 *acetyl-*
16 *radiat-*
16 *prim-*
16 *port-*
16 *oment-*
16 *micro-*
16 *medic-*
16 *mast-*
16 *entero-*
16 *coccy-*
16 *citr-*
16 *cardio-*
15 *thym-*
15 *papill-*
15 *palat-*
15 *nucle-*
15 *methyl-*
15 *lact-*
15 *gluco-*
15 *fasci-*
15 *disc-*
15 *cervic-*
15 *atri-*
15 *aneurysm-*
14 *vulv-*
14 *tars-*
14 *rose-*
14 *reticul-*
14 *peri-*
14 *nephr-*
14 *musc-*
14 *ment-*
14 *hypoderm-*
14 *glycer-*
14 *gastro-*
14 *electro-*
14 *electr-*
14 *diverticul-*
14 *arter-*
13 *ventric-*
13 *uter-*
13 *trabecul-*
13 *test-*
13 *teno-*
13 *splen-*
13 *sever-*
13 *segment-*
13 *scrot-*
13 *pulp-*
13 *microspor-*
13 *mening-*
13 *lute-*
13 *long-*
13 *labi-*
13 *hymen-*
13 *gingiv-*
13 *chondro-*
13 *antr-*
12 *weight-*
12 *vesicul-*
12 *thyro-*
12 *thorac-*
12 *sigmoid-*
12 *radio-*
12 *oste-*
12 *organ-*
12 *mono-*
12 *laryn-*
12 *indic-*
12 *implant-*
12 *gall-*
12 *flex-*
12 *endometri-*
12 *duct-*
12 *cycl-*
12 *corona-*
12 *cere-*
12 *cement-*
12 *candid-*
12 *arthr-*
11 *vestibul-*
11 *vaccin-*
11 *toxic-*
11 *thrombo-*
11 *sulfur-*
11 *puls-*
11 *polymorph-*
11 *pelvi-*
11 *nitr-*
11 *nerv-*
11 *myxo-*
11 *meat-*
11 *mamm-*
11 *hydro-*
11 *granul-*
11 *galactos-*
11 *fund-*
11 *foet-*
11 *digit-*
11 *count-*
11 *condition-*
11 *chol-*
11 *cerebr-*
11 *capsul-*
11 *benz-*
11 *arthro-*
10 *vesic-*
10 *valv-*
10 *urethro-*
10 *ulcer-*
10 *synovi-*
10 *sphenoid-*
10 *sial-*
10 *rickettsi-*
10 *protein-*
10 *placent-*
10 *physi-*
10 *peritone-*
10 *perine-*
10 *pectin-*
10 *part-*
10 *paradox-*
10 *oxal-*
10 *oesophago-*
10 *myocardi-*
10 *mucos-*
10 *male-*
10 *lympha-*
10 *lent-*
10 *keto-*
10 *kerato-*
10 *inter-*
10 *hemo-*
10 *granulo-*
10 *glob-*
10 *esophago-*
10 *cyclo-*
10 *cross-*
10 *coat-*
10 *capillari-*
10 *born-*
10 *benzo-*
10 *atroph-*
10 *appendic-*
10 *adenyl-*
10 *aconit-*
9 *trem-*
9 *tract-*
9 *trac-*
9 *thromb-*
9 *thora-*
9 *sulfon-*
9 *sple-*
9 *sphincter-*
9 *simi-*
9 *scal-*
9 *sacr-*
9 *pyelo-*
9 *psych-*
9 *psittac-*
9 *process-*
9 *point-*
9 *photo-*
9 *pharyng-*
9 *pharyn-*
9 *pancreatic-*
9 *oophor-*
9 *nodul-*
9 *margin-*
9 *levi-*
9 *leuko-*
9 *irid-*
9 *hyalin-*
9 *helminth-*
9 *haem-*
9 *guan-*
9 *glutam-*
9 *gastri-*
9 *fascia-*
9 *extract-*
9 *epiderm-*
9 *dentat-*
9 *cysto-*
9 *cyan-*

9 *cryo-*
9 *compress-*
9 *columb-*
9 *clitorid-*
9 *cistern-*
9 *cili-*
9 *cholecyst-*
9 *cestod-*
9 *camel-*
9 *bulb-*
9 *branch-*
9 *base-*
9 *band-*
9 *arteriol-*
9 *arct-*
9 *alveol-*
9 *alternat-*
9 *alkal-*
9 *african-*
9 *adipos-*
8 *ventriculo-*
8 *vari-*
8 *tube-*
8 *theileri-*
8 *tect-*
8 *striat-*
8 *streptococc-*
8 *sleep-*
8 *sinus-*
8 *sign-*
8 *separat-*
8 *schistosom-*
8 *rubr-*
8 *rota-*
8 *rock-*
8 *rest-*
8 *reflex-*
8 *pulmon-*
8 *psycho-*
8 *polyp-*
8 *pneumon-*
8 *pneumo-*
8 *phys-*
8 *phrenic-*
8 *phenyl-*
8 *penicilli-*
8 *parasit-*
8 *pancreat-*
8 *pancrea-*
8 *ovar-*
8 *nicotin-*
8 *mycobacteri-*
8 *muco-*
8 *milk-*
8 *melano-*
8 *mark-*
8 *lupin-*
8 *lobul-*
8 *later-*
8 *lacto-*
8 *kerat-*
8 *iodi-*
8 *intermedi-*
8 *incis-*
8 *ileo-*

8 *hair-*
8 *glomerul-*
8 *fren-*
8 *fimbria-*
8 *extern-*
8 *ethmoid-*
8 *dura-*
8 *dors-*
8 *diffus-*
8 *cutane-*
8 *curv-*
8 *constrict-*
8 *condyl-*
8 *compact-*
8 *clon-*
8 *chrom-*
8 *chemo-*
8 *cervi-*
8 *cerv-*
8 *cauda-*
8 *carboxy-*
8 *carbon-*
8 *carbo-*
8 *carbam-*
8 *canth-*
8 *bromi-*
8 *bond-*
8 *bell-*
8 *ascari-*
8 *arctic-*
8 *appendi-*
8 *amino-*
8 *adenoid-*
7 *worm-*
7 *wood-*
7 *wall-*
7 *variegat-*
7 *vagina-*
7 *typh-*
7 *tubercul-*
7 *trypanosom-*
7 *trichomon-*
7 *transplant-*
7 *tolu-*
7 *thio-*
7 *strongyl-*
7 *stri-*
7 *stain-*
7 *spondyl-*
7 *spleno-*
7 *spectr-*
7 *sole-*
7 *septic-*
7 *scar-*
7 *rigid-*
7 *ricin-*
7 *quin-*
7 *pull-*
7 *product-*
7 *perioste-*
7 *perc-*
7 *para-*
7 *pain-*
7 *osteopath-*
7 *opisthorchi-*

7 *obliqu-*
7 *myel-*
7 *mito-*
7 *minim-*
7 *membran-*
7 *media-*
7 *mastoid-*
7 *malle-*
7 *lymphocyt-*
7 *lump-*
7 *luci-*
7 *linea-*
7 *limit-*
7 *laryngo-*
7 *lance-*
7 *king-*
7 *intra-*
7 *intestin-*
7 *infant-*
7 *hemi-*
7 *hall-*
7 *haemophil-*
7 *glycol-*
7 *glucuron-*
7 *glucos-*
7 *gene-*
7 *ganglion-*
7 *fung-*
7 *front-*
7 *follic-*
7 *fistul-*
7 *fibrin-*
7 *fasciol-*
7 *erect-*
7 *encephalo-*
7 *elastic-*
7 *echinostom-*
7 *disseminat-*
7 *derma-*
7 *depress-*
7 *dens-*
7 *dall-*
7 *cyto-*
7 *cyclop-*
7 *crypt-*
7 *conjugat-*
7 *color-*
7 *coli-*
7 *coco-*
7 *caus-*
7 *calcul-*
7 *bronchiol-*
7 *bovi-*
7 *black-*
7 *basi-*
7 *atrophic-*
7 *aspart-*
7 *anomal-*
7 *annulat-*
7 *amygdal-*
7 *amin-*
7 *acyl-*
7 *abdomin-*
6 *yellow-*
6 *xylos-*

6 *wart-*
6 *vulva-*
6 *vitre-*
6 *viper-*
6 *ventilat-*
6 *veno-*
6 *trigemin-*
6 *transvers-*
6 *tonsil-*
6 *tongue-*
6 *thyroid-*
6 *thick-*
6 *tear-*
6 *tann-*
6 *syring-*
6 *syndactyl-*
6 *symphysi-*
6 *sweat-*
6 *sutur-*
6 *suppress-*
6 *strang-*
6 *stomat-*
6 *stoma-*
6 *staphylococc-*
6 *splanchnic-*
6 *sorb-*
6 *serrat-*
6 *sens-*
6 *salmon-*
6 *rhomboid-*
6 *rhino-*
6 *retro-*
6 *report-*
6 *radicul-*
6 *pyro-*
6 *pustul-*
6 *pulse-*
6 *pseudo-*
6 *porocephal-*
6 *plica-*
6 *plasmodi-*
6 *plasmacyt-*
6 *plasma-*
6 *placenta-*
6 *place-*
6 *phosphatid-*
6 *periodont-*
6 *period-*
6 *peptid-*
6 *parv-*
6 *papul-*
6 *pantothen-*
6 *palm-*
6 *osteochondro-*
6 *osteochondr-*
6 *ornat-*
6 *orchi-*
6 *oesophagogastr-*
6 *ocul-*
6 *nitro-*
6 *necro-*
6 *naphth-*
6 *muta-*
6 *mold-*
6 *mercur-*

6 *melan-*
6 *mediastin-*
6 *marti-*
6 *lobo-*
6 *lobat-*
6 *ligament-*
6 *leuk-*
6 *leuc-*
6 *larva-*
6 *lamina-*
6 *labyrinth-*
6 *kill-*
6 *interrupt-*
6 *inhibit-*
6 *infundibul-*
6 *indi-*
6 *ichthyophthiri-*
6 *hyster-*
6 *hypophosphat-*
6 *hypokal-*
6 *hymenolepi-*
6 *human-*
6 *histiocyt-*
6 *histid-*
6 *hernia-*
6 *hepatico-*
6 *hemorrhoid-*
6 *hand-*
6 *guard-*
6 *grunt-*
6 *granuloma-*
6 *granulocyt-*
6 *grand-*
6 *glute-*
6 *glutamin-*
6 *function-*
6 *fructo-*
6 *fort-*
6 *formos-*
6 *formic-*
6 *foli-*
6 *fluor-*
6 *flow-*
6 *flor-*
6 *fissur-*
6 *filari-*
6 *ferr-*
6 *feli-*
6 *faint-*
6 *esophagogastr-*
6 *erythrocyt-*
6 *erythr-*
6 *equin-*
6 *endometrio-*
6 *endocardi-*
6 *emboli-*
6 *elongat-*
6 *dream-*
6 *dracuncul-*
6 *domestic-*
6 *dipylidi-*
6 *diphyllobothri-*
6 *dicrocoeli-*
6 *curl-*
6 *correct-*

6 *colour-*
6 *coll-*
6 *coit-*
6 *coagula-*
6 *cloud-*
6 *clitor-*
6 *chord-*
6 *chondr-*
6 *choledocho-*
6 *chem-*
6 *cathartic-*
6 *case-*
6 *capr-*
6 *canalicul-*
6 *calc-*
6 *bronchi-*
6 *block-*
6 *bigemin-*
6 *basophil-*
6 *basil-*
6 *balantidi-*
6 *bacteri-*
6 *azur-*
6 *audit-*
6 *atrophi-*
6 *asper-*
6 *asparagin-*
6 *ascarid-*
6 *arthropod-*
6 *argin-*
6 *agon-*
6 *aggregat-*
6 *affect-*
6 *adip-*
6 *activ-*
6 *acar-*
6 *acanthamoeb-*
6 *abscess-*
6 *abomas-*
5 *xero-*
5 *xanth-*
5 *vivipar-*
5 *virgin-*
5 *vertebra-*
5 *verruca-*
5 *verruc-*
5 *valve-*
5 *vagino-*
5 *urin-*
5 *uncinat-*
5 *tyrosin-*
5 *tung-*
5 *tumo-*
5 *tuberculo-*
5 *tuber-*
5 *tryptophan-*
5 *trichostrongyl-*
5 *trichin-*
5 *trematod-*
5 *trabecula-*
5 *thion-*
5 *thermophil-*
5 *thalam-*
5 *testi-*
5 *terminal-*

5 *telescop-*
5 *taur-*
5 *taeni-*
5 *sulfo-*
5 *stria-*
5 *stretch-*
5 *strain-*
5 *ster-*
5 *stellat-*
5 *steel-*
5 *spring-*
5 *splint-*
5 *spinos-*
5 *soap-*
5 *silver-*
5 *sigmoido-*
5 *show-*
5 *sheath-*
5 *septi-*
5 *septa-*
5 *senegal-*
5 *section-*
5 *screen-*
5 *scratch-*
5 *scop-*
5 *sclero-*
5 *scapula-*
5 *scapul-*
5 *scalen-*
5 *saur-*
5 *sarcopt-*
5 *saliva-*
5 *saccul-*
5 *sabin-*
5 *russ-*
5 *reticulat-*
5 *recto-*
5 *record-*
5 *radia-*
5 *pyroglutam-*
5 *purpose-*
5 *pupil-*
5 *punctat-*
5 *psoropt-*
5 *protect-*
5 *propionic-*
5 *prol-*
5 *polar-*
5 *poison-*
5 *poikilotherm-*
5 *pneumoni-*
5 *pneumat-*
5 *plicat-*
5 *plant-*
5 *plac-*
5 *physic-*
5 *phren-*
5 *phosphatidyl-*
5 *phosphat-*
5 *phenol-*
5 *pericardio-*
5 *perforat-*
5 *pelv-*
5 *pari-*
5 *parent-*
5 *parathyroid-*
5 *parasitic-*
5 *paragonim-*
5 *papillos-*
5 *pancre-*
5 *palpebra-*
5 *palmar-*
5 *palatin-*
5 *pack-*
5 *pacific-*
5 *oxid-*
5 *oval-*
5 *osti-*
5 *optic-*
5 *ophthalm-*
5 *ocellat-*
5 *obtus-*
5 *obstruct-*
5 *obscur-*
5 *nodos-*
5 *neutrophil-*
5 *neuron-*
5 *nematod-*
5 *neck-*
5 *nasopharyn-*
5 *narco-*
5 *mull-*
5 *morpho-*
5 *monilia-*
5 *mongol-*
5 *minut-*
5 *mili-*
5 *mexican-*
5 *metatars-*
5 *mesenter-*
5 *masto-*
5 *mass-*
5 *marsupial-*
5 *marin-*
5 *macroglobulin-*
5 *lymphangio-*
5 *lophi-*
5 *lock-*
5 *lipomatos-*
5 *lipid-*
5 *light-*
5 *leukocyt-*
5 *lenticular-*
5 *lamp-*
5 *lamella-*
5 *lacuna-*
5 *keton-*
5 *keratin-*
5 *john-*
5 *ixod-*
5 *isocitr-*
5 *involut-*
5 *invers-*
5 *invaginat-*
5 *intern-*
5 *instrument-*
5 *infarct-*
5 *induc-*
5 *imperfect-*
5 *hypoton-*
5 *hydra-*
5 *homoiotherm-*
5 *histidin-*
5 *hexamit-*
5 *hemangio-*
5 *haemolytic-*
5 *haemo-*
5 *haemangio-*
5 *glyceri-*
5 *glyc-*
5 *gloss-*
5 *globos-*
5 *gland-*
5 *gill-*
5 *genu-*
5 *gastroduoden-*
5 *gale-*
5 *galacto-*
5 *fragment-*
5 *fraction-*
5 *follow-*
5 *follicul-*
5 *fold-*
5 *flav-*
5 *fish-*
5 *fimbri-*
5 *fibros-*
5 *feed-*
5 *erythroblast-*
5 *epitheli-*
5 *epiphys-*
5 *epiglotti-*
5 *ependym-*
5 *eosinophil-*
5 *endo-*
5 *encounter-*
5 *embol-*
5 *elast-*
5 *echo-*
5 *donovan-*
5 *delay-*
5 *crur-*
5 *crotali-*
5 *crotal-*
5 *corniculat-*
5 *coordinat-*
5 *conjunctivo-*
5 *commissur-*
5 *colpo-*
5 *colon-*
5 *coccidioid-*
5 *clitori-*
5 *citrullin-*
5 *cingul-*
5 *choroid-*
5 *chordat-*
5 *chor-*
5 *chlori-*
5 *cerebell-*
5 *central-*
5 *caudat-*
5 *carnosin-*
5 *calv-*
5 *call-*
5 *calic-*
5 *calci-*
5 *burn-*
5 *breath-*
5 *borne-*
5 *blatt-*
5 *blast-*
5 *bilob-*
5 *bend-*
5 *bear-*
5 *bark-*
5 *barba-*
5 *babesi-*
5 *austral-*
5 *auri-*
5 *attenuat-*
5 *atlantic-*
5 *atlant-*
5 *aster-*
5 *argentin-*
5 *arab-*
5 *anserin-*
5 *anseri-*
5 *anser-*
5 *angi-*
5 *angel-*
5 *ancylostom-*
5 *american-*
5 *amazon-*
5 *alkyl-*
5 *agar-*
5 *aegypti-*
5 *adrenal-*
5 *acut-*
5 *acetylglucosamin-*
5 *aceton-*
5 *acetabul-*
5 *aborti-*

Russe (330 préfixes sur 2 937)

36 *фибро-*
28 *лимфо-*
24 *почечн-*
22 *мието-*
21 *легочн-*
21 *фиброз-*
19 *злокачественн-*
19 *кожн-*
19 *костн-*
18 *нейро-*
18 *кишечн-*
17 *остео-*
17 *нерв-*
17 *диффузн-*
17 *врожденн-*
17 *базальн-*
16 *эпители-*
16 *хромосом-*
15 *клеточн-*
15 *адено-*
14 *мышечн-*
14 *полов-*
14 *печеночн-*
14 *эндометри-*
14 *артери-*
14 *тяжел-*
13 *клапан-*
12 *открыт-*
12 *мембран-*
12 *лимф-*
12 *липо-*
11 *нормальн-*
11 *наружн-*
11 *молочн-*
11 *кост-*
11 *множественн-*
11 *крестцов-*
11 *полип-*
11 *подвздошн-*
11 *пузыр-*
11 *глиальн-*
11 *влагалищ-*
11 *генитальн-*
11 *женск-*
11 *желез-*
11 *дефект-*
11 *эндокринн-*
11 *ангио-*
11 *сосудист-*
11 *слизист-*
11 *фиброзн-*
11 *серозн-*
11 *сердечн-*
11 *хирургическ-*
11 *хроническ-*
11 *тубулярн-*
10 *кров-*
10 *пигмент-*
10 *переходн-*
9 *лейкоцит-*
9 *лимфоцит-*
9 *липидн-*
9 *липид-*
9 *легк-*
9 *кист-*
9 *проток-*
9 *гемангио-*
9 *бронх-*
9 *сосуд-*
9 *умеренн-*
9 *физическ-*
9 *тромб-*
8 *надгортанн-*
8 *коллаген-*
8 *копчиков-*
8 *мозг-*
8 *лобков-*
8 *клет-*
8 *мелано-*
8 *лекарственн-*
8 *лимфатическ-*
8 *меди-*
8 *кислот-*
8 *мягк-*
8 *маточно-*
8 *мужск-*
8 *поверхностн-*
8 *почечно-*
8 *плоск-*
8 *плазм-*
8 *патологически-*
8 *патологическ-*
8 *пупочн-*
8 *дополнительн-*
8 *воспалительн-*
8 *воспал-*
8 *изменен-*
8 *гликоген-*
8 *железист-*
8 *желудочн-*
8 *желудочков-*
8 *гипер-*
8 *гепато-*
8 *дистрофи-*
8 *биологическ-*
8 *анальн-*
8 *амилоид-*
8 *ацинарн-*
8 *хондро-*
8 *смешанн-*
8 *спонтанн-*
8 *тромбоцит-*
8 *срез-*
7 *остр-*
7 *нижн-*
7 *моноцит-*
7 *копчик-*
7 *клиновидн-*
7 *лекарств-*
7 *микроцит-*
7 *миксо-*
7 *поверхност-*
7 *поясничн-*
7 *плоскоклеточн-*
7 *пищеварительн-*
7 *первичн-*
7 *прав-*
7 *инфекци-*

7 желчн-
7 гипертрофи-
7 верхн-
7 гистиоцит-
7 дистрофическ-
7 дыхательн-
7 вагинальн-
7 зернист-
7 эндометриальн-
7 шейн-
7 задн-
7 артериальн-
7 частичн-
7 соедин-
7 систем-
7 церебральн-
7 хронически-
7 уретр-
7 тучн-
6 орган-
6 норм-
6 низк-
6 кожно-
6 костно-
6 клеточно-
6 клитор-
6 мелко-
6 лимфоидн-
6 легочно-
6 лево-
6 мегакариоцит-
6 кишечно-
6 металл-
6 мышечно-
6 красн-
6 позвоночн-
6 передн-
6 гортан-
6 внутренн-
6 иммунн-
6 глио-
6 измененн-
6 глаз-
6 глубок-
6 гемоглобин-
6 геморраги-
6 губчат-
6 гранул-
6 груд-
6 больш-
6 бронхиол-
6 эндотели-
6 эластическ-
6 аденоид-
6 атрофи-
6 астроцит-
6 тонк-
6 фолликул-
6 сниженн-
6 тканев-
6 тяжело-
6 средне-
6 сустав-
5 определен-
5 опухол-
5 овариальн-
5 образованн-
5 образован-
5 основн-
5 осложненн-
5 нодулярн-
5 ненормальн-
5 нейрон-
5 неполн-
5 нижне-
5 нервн-
5 надпочечник-
5 конечн-
5 коллагенов-
5 мозгов-
5 ложн-
5 коронарн-
5 морфологи-
5 клубочков-
5 мезонефр-
5 мезотели-
5 минимальн-
5 минеральн-
5 лейомио-
5 лимфоцитарн-
5 лимфа-
5 миелоидн-
5 метаболическ-
5 керат-
5 кистозн-
5 кисто-
5 мышц-
5 капилляр-
5 полост-
5 полн-
5 полиповидн-
5 полихром-
5 покров-
5 поперечн-
5 позвоно-
5 подобн-
5 подостр-
5 подкожн-
5 повышенн-
5 подагрическ-
5 повторн-
5 подслизист-
5 посмертн-
5 портальн-
5 плазмоклеточн-
5 плацентарн-
5 переходно-
5 периферическ-
5 перв-
5 панкреатическ-
5 папиллярн-
5 патологоанатомическ-
5 промежуточн-
5 приобретенн-
5 придаточн-
5 пузырн-
5 голов-
5 доброкачественн-
5 гормональн-
5 ворсинчат-
5 инородн-
5 интерстициальн-
5 глоточн-
5 иммунодефицит-
5 длинн-
5 влагалищн-
5 венозн-
5 венечн-
5 геморроидальн-
5 железисто-
5 желч-
5 веществ-
5 гиалинов-
5 височн-
5 витамин-
5 верхне-
5 вызванн-
5 высок-
5 вторичн-
5 гранулематозн-
5 истинн-
5 грудн-
5 белков-
5 бедренн-
5 бактери-
5 базально-
5 бронхиальн-
5 эозинофил-
5 эласто-
5 эпителиальн-
5 яремн-
5 эритроцит-
5 аортальн-
5 алкогольн-
5 закрыт-
5 аденом-
5 зародышев-
5 атипичн-
5 атипи-
5 ацидофильн-
5 череп-
5 фолликулярн-
5 толст-
5 сложн-
5 скелет-
5 ткан-
5 слухов-
5 спинномозгов-
5 синус-
5 симпатическ-
5 семенн-
5 семейн-
5 свищ-
5 твердо-
5 связанн-
5 термин-
5 серповидн-
5 церебикальн-
5 радиаци-
5 расширенн-
5 разрыв-
5 хромосомн-
5 уротели-
5 стеноз-
5 уреми-
5 средн-
5 хрящев-
5 хрящ-
5 фрагмент-
5 стромальн-
5 труб-
5 структур-

A.1.5 Familles

Français (187 familles sur 1 678)

abdomen, *abdominal*, *abdominale*, *abdominales*, *abdominaux*, *abdomino*
acanthomateux, *acanthome*, *acanthose*, *acanthosique*
acquisition, *acquisitions*
adhérence, *adhérences*, *adhérent*
affaissement, *affaissé*
albumine, *albumineuse*, *albuminoïde*
alvéolaire, *alvéolaires*, *alvéole*, *alvéoles*
amygdale, *amygdales*, *amygdalien*, *amygdalienne*, *amygdalite*, *amygdalo*
anesthésie, *anesthésique*, *anesthésiques*
annulaire, *annulaires*, *annulé*
antécédent, *antécédents*
apocrine, *apocrines*
argyrie, *argyrose*
artère, *artères*
asthmatique, *asthme*
atrophiante, *atrophie*, *atrophique*, *atrophiques*, *atrophodermie*
autolyse, *autolytique*
bandelette, *bandelettes*
bilatéral, *bilatérale*, *bilatérales*, *bilatéraux*
blanc, *blancs*
branche, *branches*, *branchial*, *branchiale*
bulle, *bulleuse*, *bulleuses*, *bulleux*
callosité, *callosités*
carcino, *carcinogène*, *carcinomateuse*, *carcinomatose*, *carcinome*, *carcinomes*, *carcinosarcome*, *carcinoïde*
caséeuse, *caséeux*
cervical, *cervicale*, *cervicales*, *cervicaux*, *cervicite*, *cervico*
chimique, *chimiques*
cholécystite, *cholécystites*
chyle, *chyleux*, *chylocèle*, *chylurie*
clinique, *cliniquement*, *clinoïde*
colliquation, *colliquative*
complexe, *complexes*
condylomateux, *condylome*
conservation, *conservé*
contraction, *contractions*, *contracture*, *contracté*, *contractée*
coronaire, *coronaires*, *coronarien*, *coronarienne*, *coronariens*
couleur, *couleurs*
croisé, *croisés*
cutané, *cutanée*, *cutanées*, *cutanéo*, *cutanés*
cytopathologie, *cytopathologique*
côte, *côtes*

dermatophytose, *dermatophytoses*
diamniotique, *diamniotiques*
dilatation, *dilaté*
disséminé, *disséminée*, *disséminées*
dorsal, *dorsale*, *dorsales*, *dorsaux*, *dorso*, *dorsum*
durable, *durables*
décidual, *déciduale*, *déciduaux*
déficience, *déficiences*
démyélinisante, *démyélinisantes*, *démyélinisation*
déshydratation, *déshydraté*, *déshydratée*
ectasie, *ectasié*
encapsulé, *encapsulée*
endocrine, *endocrines*, *endocrinien*, *endocrinienne*, *endocriniennes*, *endocriniens*
enkysté, *enkystée*
enzymatique, *enzymatiques*, *enzyme*, *enzymes*
excavation, *excavé*
explicatif, *explication*, *explications*
extrême, *extrêmement*, *extrêmes*
falciforme, *falciformes*
ferrocalcinose, *ferrocalcique*
fluor, *fluorure*
force, *forces*
friabilité, *friable*
fusion, *fusionné*
gangliocytaire, *gangliome*
genou, *genoux*
globule, *globules*
gonade, *gonadique*, *gonadoblastome*
gravide, *gravidique*
gémistocyte, *gémistocytique*
herniaire, *hernie*, *hernies*, *herniorrhaphie*
hormonal, *hormonale*, *hormonaux*, *hormone*, *hormones*, *hormono*
hydrocèle, *hydrocèles*
hyperkinétique, *hyperkinétiques*
hypochrome, *hypochromie*
hypostase, *hypostatique*
hématologique, *hématologiques*, *hématopoïétique*, *hématopoïétiques*
hémosidérine, *hémosidérinique*
identifiable, *identification*, *identifié*
immunoprolifération, *immunoprolifératives*
incarcération, *incarcéré*
indépendant, *indépendantes*, *indépendants*
inflammation, *inflammations*, *inflammatoire*, *inflammatoires*
instrument, *instrumental*
intermittent, *intermittente*
intoxication, *intoxications*
intrinsèque, *intrinsèques*
ischio, *ischion*
jour, *jours*
labyrinthique, *labyrinthiques*
laryngite, *laryngo*, *laryngopharynx*, *laryngé*, *laryngée*, *laryngées*, *larynx*
leucokératose, *leucoplasie*
lingula, *lingulaire*
liée, *liées*
long, *longs*
lèvre, *lèvres*
macrocytaire, *macrocyte*, *macrocytose*, *macroglial*, *macroglie*
malade, *maladie*, *maladies*
maniaque, *maniaques*
massif, *massive*
membre, *membres*
migraine, *migraines*, *migraineux*
modéré, *modérée*
moteur, *moteurs*
multiloculaire, *multilocularis*
myofibromatose, *myofibrome*, *myofibrose*
myélodysplasique, *myélodysplasiques*
média, *médiale*, *médian*, *médiane*, *médio*
mélioïdose, *mélioïdoses*
mésothélial, *mésothéliale*, *mésothéliome*, *mésothélium*
méthode, *méthodes*
neuromusculaire, *neuromusculaires*
noir, *noire*
nutrition, *nutritionnel*, *nutritionnelle*, *nutritionnelles*, *nutritionnels*
névroglie, *névroglique*, *névrogliques*
occlus, *occlusif*, *occlusion*, *occlusions*, *occlusive*
omphalite, *omphalo*
opportuniste, *opportunistes*
original, *origine*, *origines*
ostéolyse, *ostéolytique*
ovarien, *ovarienne*, *ovariennes*, *ovarite*
palpébral, *palpébrale*, *palpébraux*, *palpébro*
parakératose, *parakératosique*
parotide, *parotidien*
pemphigoïde, *pemphigoïdes*, *pemphigus*
petit, *petite*, *petites*, *petitesse*, *petits*
photographie, *photomicrographie*
pilomatricaire, *pilomatrixome*
plasmatique, *plasmatiques*
pneumatique, *pneumatose*
polykystique, *polykystiques*
population, *populations*
poumon, *poumons*
probable, *probablement*
progressant, *progressif*, *progression*, *progressive*
prothèse, *prothèses*
précipitation, *précipité*, *précipités*
prématuré, *prématurée*, *prématurées*, *prématurés*
psammomateux, *psammome*
pulpaire, *pulpe*, *pulpeux*
pyomètre, *pyométrie*
pénétrant, *pénétrante*, *pénétré*
périodontale, *périodontales*
pétéchial, *pétéchiale*, *pétéchie*
radical, *radicale*
rayonnement, *rayonnements*
requis, *requise*
rond, *rondes*, *ronds*
récurrent, *récurrente*, *récurrentes*, *récurrents*
réparation, *réparé*
rétropéritoine, *rétropéritonéal*, *rétropéritonéale*
sarcomateuse, *sarcomateux*, *sarcomatose*, *sarcomatoïde*, *sarcome*, *sarcomes*, *sarcoïde*, *sarcoïdose*, *sarcoïdosique*
second, *secondaire*, *secondaires*
sept, *septique*
simple, *simples*, *simplex*
somatoforme, *somatoformes*
sphérocytaire, *sphérocyte*, *sphérocytose*
spontané, *spontanée*, *spontanées*, *spontanément*, *spontanés*
staphylocoque, *staphylocoques*
strongyloïdes, *strongyloïdose*
substance, *substances*, *substantia*
supérieur, *supérieure*, *supérieures*, *supérieurs*, *supéro*
symptomatique, *symptomatiques*
système, *systèmes*

séquestration, *séquestre*, *séquestré*
teinte, *teinté*
thoracique, *thoraciques*, *thoraco*, *thorax*
tige, *tiges*
trachéal, *trachéale*, *trachée*, *trachéite*, *trachéo*
transport, *transports*
troisième, *troisièmement*
tumeur, *tumeurs*
ténosynovite, *ténosynovites*
uretère, *uretères*
utérin, *utérine*, *utérines*, *utérins*, *utéro*, *utérus*
varice, *varices*, *variqueuse*, *variqueuses*, *variqueux*
verre, *verres*
villeuse, *villeux*, *villo*
vocal, *vocale*, *vocales*, *vocaux*
vénérien, *vénérienne*, *vénériens*
échographie, *échographiques*
élevé, *élevée*
épicanthique, *épicanthus*
épiplo, *épiploon*, *épiploïque*
éraflure, *éraflé*
étranglement, *étranglements*, *étranglé*

Anglais (199 familles sur 6 550)

[methylcrotonyl, *[methylcrotonoyl*
accident, *accidents*, *accidental*, *accidentally*
acrocephalosyndactyly, *acroencephalosyndactyly*
adjuvant, *adjuvants*
agent, *agency*, *agents*, *agencies*
albino, *albinotic*
allocreadium, *allocreadiidae*
america, *american*, *americana*, *americaine*, *americanum*, *americanus*, *amerindian*
amphetamine, *amphetamines*
anchor, *anchoring*
anisa, *anise*
anthelmintic, *anthelmintica*, *anthelmintics*
apatemon, *apatemosis*
apyrexia, *apyrexic*
ariomma, *ariommum*, *ariommus*
asclepias, *asclepiadaceae*
ataxia, *ataxic*
aulopus, *aulopidae*
avoidant, *avoidance*
ball, *baller*
baso, *basophil*, *basophils*, *basophilia*, *basophilic*, *basophilism*
bentazon, *bentazone*
bilirubin, *bilirubinemia*, *bilirubinuria*
blastocyst, *blastocystis*
bombyx, *bombycis*
bozemanae, *bozemanii*
brewer, *brewers*
buetschlia, *buetschliidae*
butterfly, *butterflies*, *butterflyfish*, *butterflyfishes*
callistin, *callistius*
cannabis, *cannabinum*, *cannabosis*, *cannabaceae*, *cannabinoid*, *cannabinoids*, *cannabinosis*
cardinal, *cardinalis*, *cardinalfish*, *cardinalfishes*
castellani, *castellanii*
cavum, *cavus*
cercaria, *cercariae*, *cercarial*
char, *chart*, *charina*, *charity*, *charting*
chlamydoselachus, *chlamydoselachidae*
chromosome, *chromosomal*, *chromosomes*, *chromosomopathy*
circumstance, *circumstances*, *circumstantial*, *circumstantiality*
cleaved, *cleaving*
cobalt, *cobaltosis*
colliquativa, *colliquative*
compound, *composite*, *compounds*, *compositae*, *composition*
conie, *conium*, *coniine*
cookei, *cookeii*
corticovirus, *corticoviridae*
crater, *craterization*
croton, *crotonic*, *crotonaldehyde*
culdotomy, *culdocentesis*
cycad, *cycas*, *cycada*, *cycadaceae*
cytotoxic, *cytotoxin*, *cytotoxicity*
dctp, *dctpase*
deglutition, *deglutitive*
deodorant, *deodorants*
destruction, *destructive*, *destructiveness*
diaphysis, *diaphyseal*, *diaphysitis*, *diaphysectomy*
dihydro, *dihydric*, *dihydrate*, *dihydrogen*
dipper, *dipping*
distributed, *distribution*, *distributions*
dracunculus, *dracontiasis*, *dracunculosis*, *dracunculiasis*, *dracunculoidea*, *dracunculoides*
duranta, *durante*
dystonia, *dystonic*
eight, *eighth*
emilia, *emiliae*
endo, *endocrine*, *endocytic*, *endogenous*, *endocardial*, *endocardium*, *endocarditis*, *endocardiosis*, *endopeptidase*, *endocarditides*, *endocrinopathy*, *endomyocardial*
epidermoptes, *epidermoptidae*
erythema, *erythematosa*, *erythematous*, *erythematosus*
eumycotic, *eumycetoma*
execution, *executions*
exudate, *exudates*, *exudation*, *exudativa*, *exudative*, *exudativum*
fast, *fastin*, *fasting*
feto, *fetor*
flaccid, *flaccida*, *flaccidity*
fluent, *fluency*
foramina, *foraminotomy*
friction, *frictional*
furrow, *furrows*, *furrowed*
gangrene, *gangrenes*, *gangrenosa*, *gangrenous*, *gangrenosum*
genius, *genioplasty*
girdle, *girdles*, *girdlestone*
glyceria, *glyceric*, *glycerin*, *glycerol*, *glyceryl*, *glycerate*, *glyceride*, *glycerite*, *glycerone*, *glyceraldehyde*, *glycerophosphate*, *glycerolphosphate*, *glycerophosphatase*
gordian, *gordius*
grenadier, *grenadiers*
guppy, *guppyi*
haemostasis, *haemostatic*
harmine, *harmaline*, *harmidine*
hemangiectasia, *hemangiectasis*

hernia, *hernial*, *hernias*, *herniated*, *herniation*, *hernioplasty*, *herniorrhaphy*
himantolophus, *himantolophidae*
holmes, *holmesii*
howler, *howling*
hydroxymethyl, *hydroxymethylase*, *hydroxymethyltransferase*
hypersplenia, *hypersplenism*
hypothenar, *hypothenaris*
ileovesical, *ileoneocystostomy*
inclusion, *inclusions*
inflation, *inflatable*
instruction, *instructions*
invaginata, *invaginated*, *invaginatus*, *invagination*
isoleucyl, *isoleucine*, *isoleucinemia*
java, *javan*, *javanica*, *javanicum*, *javanicus*
kell, *kelly*, *keller*
kyphoscoliosis, *kyphoscoliotic*
lamp, *lamps*, *lampas*, *lampers*, *lamprey*, *lampetra*, *lampfish*, *lampreys*, *lampanyctus*
leach, *leachi*
leptomeninges, *leptomeningeal*, *leptomeningitis*
ligase, *ligases*, *ligator*, *ligation*, *ligature*
little, *littler*, *littlehead*
lophotes, *lophotus*, *lophotidae*
lutra, *lutrae*, *lutris*, *lutrensis*
macrosomia, *macrosomic*
malpresentation, *malpresentations*
marie, *mariae*
measure, *measures*, *measuring*, *measurement*, *measurements*
menispermum, *menispermaceae*
metanephros, *metanephrine*, *metanephrines*
microdontia, *microdontism*
minnow, *minnows*
mollusk, *mollusca*, *mollusks*, *molluscum*
mori, *morio*
multiceps, *multicepsosis*
mycotoxic, *mycotoxin*, *mycotoxicoses*, *mycotoxicosis*
myxoid, *myxoma*, *myxozoa*, *myxosoma*, *myxobolus*, *myxomatous*, *myxosporea*, *myxobolosis*, *myxofibroma*, *myxomatosis*, *myxosarcoma*, *myxoliposarcoma*
neck, *necked*, *necker*, *necking*
nevi, *nevin*
nonobstetric, *nonobstetrical*
nutricia, *nutriens*, *nutrient*, *nutrients*
octomitus, *octomitosis*
olivaceum, *olivaceus*
opposita, *opposite*, *opposition*, *oppositional*
orthodon, *orthodontic*, *orthodontics*, *orthodontist*
oudemansi, *oudemansiella*
palatoglossal, *palatoglossus*
paracolic, *paracolon*, *paracolostomy*
paresthesia, *paresthetica*
pathogens, *pathology*, *pathogenic*, *pathologic*, *pathological*
peltata, *peltatum*
pericoronal, *pericoronitis*
pesticide, *pesticidal*, *pesticides*
philippine, *philippinensi*, *philippinensis*
physa, *physic*, *physis*, *physeal*, *physics*, *physalis*, *physical*, *physidae*, *physitis*, *physician*, *physopsis*, *physiatric*, *physically*, *physicians*, *physiatrist*, *physiologic*, *physiologist*, *physiological*, *physiotherapy*, *physiologically*, *physiotherapeutic*
pinon, *pinos*
platyrhynchos, *platyrhynchus*
poikilotherm, *poikilothermy*, *poikilothermia*, *poikilothermic*
pomacentrus, *pomacentridae*
posture, *postural*, *posturing*
prescribed, *prescribing*, *prescription*, *prescriptions*
product, *products*, *productus*, *production*, *productive*
protan, *protanopia*, *protanomaly*
psitacci, *psittaci*, *psittacus*, *psittacine*, *psittacoid*, *psittacidae*, *psittacines*, *psittacinus*, *psittacosis*, *psittaciformes*
punicea, *puniceum*
pyrogallic, *pyrogallol*
rachialgia, *rachiodynia*
recent, *recently*
regressing, *regression*
replace, *replaced*, *replacer*, *replacing*, *replacement*, *replacements*
retrobulbar, *retroflexus*, *retrocession*, *retroflexion*, *retroorbital*, *retropyloric*
rhytides, *rhytidectomy*, *rhytidoplasty*
rolando, *rolandic*
rugosa, *rugosum*, *rugosus*
salamander, *salamandra*, *salamanders*, *salamandridae*
sarcocystis, *sarcocystidae*, *sarcocystosis*, *sarcosporidiasis*, *sarcosporidiosis*
schultz, *schultze*, *schultzi*, *schultzea*
scutata, *scutatum*, *scutatus*
seminole, *seminolis*
sesame, *sesamoid*, *sesamoidea*, *sesamoidean*, *sesamoiditis*, *sesamoidectomy*
show, *showy*, *shower*, *showing*
siren, *sirenia*, *sirenidae*, *sirenoform*, *sirenomelia*, *sirenomelus*
snip, *snipe*, *snipefish*, *snipefishes*
span, *spanish*
spigelia, *spigelian*
squatina, *squatinidae*
steno, *stensen*, *stenella*, *stenoses*, *stenosis*, *stenotic*, *stenosing*
stone, *stony*, *stones*
strophantin, *strophanthin*
subterranean, *subterraneum*, *subterraneus*
superficial, *superficialis*
sutton, *suttonii*
syndrome, *syndromes*
talc, *talcum*, *talcosis*
telephone, *telephonic*
tesselata, *tesselatus*
theo, *theox*
thyme, *thymic*, *thymin*, *thymol*, *thymus*, *thymine*, *thymoma*, *thymidine*, *thymocyte*, *thymopexy*, *thymectomy*
tooth, *toothed*, *toothache*, *toothless*
transfusion, *transfusions*
triaenophorus, *triaenophorosis*
trigonocephaly, *trigonocephalia*, *trigonocephalum*, *trigonocephalus*
trumpet, *trumpets*, *trumpeter*, *trumpetfish*, *trumpetfishes*
twin, *twins*, *twinning*
unicorn, *unicornis*
urtica, *urticaria*, *urticatus*, *urticaceae*, *urticarial*
vasoprotectant, *vasoprotectants*
verruculosa, *verruculosum*
virgin, *virginal*, *virginia*, *virginica*, *virginiana*, *virginicus*, *virginianus*
waccamaw, *waccamensis*
weybridge, *weybridgensis*
worry, *worried*, *worrisome*
zelleri, *zelleriella*

Russe (190 familles sur 1 709)

агрессивная, агрессивный
азот, азота
алейкемическая, алейкемический
амелобластная, амелобластома, амелобластическая
амфофилия, амфофильное, амфофильный
анемия, анемический, анемического
апокринная, апокринной
артериола, артериолы, артериолярный
атерома, атероматоз, атеросклероз, атероматозная, атероматозный, атеросклеротическая, атеросклеротический
аутотрансплантат, аутотрансплантант
баланс, баланса
белые, белый
бифуркации, бифуркация
болезни, болезнь, болезней
брюшная, брюшное, брюшной
васкулярная, васкулярный
вентрикулярная, вентрикулярные, вентрикулярный
веществ, вещества, вещество, веществом
включение, включения, включающий, включенный
внутрипеченочные, внутрипеченочный
воздух, воздуха, воздушная, воздушный
воск, воска, восковидный
вскрытие, вскрытии, вскрытия, вскрытием
выпадение, выпадения
гадолинит, гадолиния
гангрена, гангренозное, гангренозный
гематурии, гематурия
генерализованная, генерализованный
герпес, герпеса, герпесвирус, герпетическая, герпетический, герпетическое, герпесвирусная
гиперкератоз, гиперкератозы
гипокалемии, гипокалемия, гипокалиемия, гипокалемическая
гистоплазма, гистоплазмоз
глиально, глиальная, глиальное, глиальные, глиальный, глиальных
глюкозы, глюкоаминофосфатурией
гонада, гонадо
группа, группы, групповые
давление, давления
дегидразы, дегидразной
дерматофиброма, дерматофибросаркома
деформация, деформирующий, деформированный
диск, диска, дисков
диффузия, диффузно, диффузная, диффузное, диффузные, диффузный, диффузным, диффузных, диффузного
друг, друга, других, другой, другом, другими
желтое, желтые, желтый, желтого
жидкости, жидкость, жидкостей, жидкостью
заключительная, заключительный
затвердение, затвердевший
зрачки, зрачок
излияние, излившийся
иммуноглобулин, иммуноглобулина, иммуноглобулиновый
инволюции, инволюция, инволютивная, инволютивный
интерстициально, интерстициальная, интерстициальное, интерстициальный
инфекции, инфекция, инфекцией, инфекционная, инфекционное, инфекционный
истекший, истечение
каждое, каждый, каждого
калькулез, калькулеза, калькулезный
канализация, канализированный
карман, кармана
кислой, кислот, кислый, кислота, кислоты, кислотно, кислотное, кислотного
клиновидно, клиновидная, клиновидное, клиновидной, клиновидный
кокк, кокки
комбинация, комбинированная, комбинированный
консистенции, консистенция, консолидация, консолидированный
корка, корки
краска, красители
криптококк, криптококкоз
круг, круга
ладони, ладонная
легочно, легочная, легочное, легочной, легочные, легочный, легочным, легочных, легочного
лепра, лепры, лепрозная
линейно, линейная, линейный
лобковая, лобковой, лобковые, лобковый, лобкового
лучевая, лучевой
масса, массе, массы, массой
мегакариоцит, мегакариоциты, мегакариобласт, мегакариоцитов, мегакариоцитоз, мегакариобластный, мегакариоцитарный
межпозвоночный, межпозвоночных, межпозвоночного
мелко, мелкие, мелкое, мелкого, мелкоклеточная, мелкоклеточный, мелколимфоцитарная
меркеля, меркелома
метафиз, метафизарный
микоз, микотическая, микотический
милиарная, милиарный
миоэпителиома, миоэпителиальная, миоэпителиоидная
мозжечка, мозжечок, мозжечково, мозжечковая, мозжечковый
мононуклеоз, мононуклеоза
мошонка, мошонки, мошоночно, мошоночный
мышьяк, мышьяка
наличие, наличия
неадекватный, неадекватным
недостаток, недоразвитие, недоразвитой, недоразвитого, недостаточное, недостаточный, недостаточного, недостаточности, недостаточность, недостаточностью
нейтрофил, нейтрофилов, нейтрофильная, нейтрофильный
неокостенелый, неокостеневающий
непроходимости, непроходимость, непроходимостью
нефрит, нефрома, нефропатия, нефрогенная, нефрогенный, нефролитиаз, нефробластома, нефропатический
нисходящая, нисходящей, нисходящий, нисходящих
носоглотка, носоглотки, носоглоточная, носоглоточный
образование, образования, образованием, образованная, образованное, образованные, образованный, образующийся
обычное, обычной
одностороннее, односторонний, односторонняя
окостеневающая, окостеневающий
олигофрении, олигофрения
орган, органа, органы, органов
осложнение, осложненная, осложненное, осложненной, осложненный
остеолиз, остеолитическое
отделение, отделенный
относящий, относящийся
пазуха, пазухи
парагонимус, парагонимусом
паренхиматозная, паренхиматозноклеточная
перегиб, перегибе, переднее, передней, передние, передний, передняя, перешеек, переднего, перегнутый, переднебоковая, переднебоковой
перепонка, перепончатость
периодонта, периодонтальный
пигмент, пигментная, пигментное, пигментный, пигментация, пигментарная, пигментозная, пигментированный
питание, питания, питательный
плевра, плевры, плеврально, плевральная, плевральный
плутонит, плутония
подавление, подавленный, подавляющий
подмышечная, подмышечной, подмышечный
подход, подходов
поле, полей
половина, половины
поперечно, поперечная, поперечное, поперечный

посмертная, *посмертное*, *посмертные*, *посмертный*
правае, *правая*, *правое*, *правой*, *правый*, *правого*
прелейкоз, *прелейкемический*
прижизненная, *прижизненное*, *прижизненный*
причин, *причина*, *причины*
пролапс, *пролапса*
просвет, *просветов*
профили, *профиль*
псевдолипома, *псевдолипоматозная*
пустула, *пустуло*, *пустулярный*
рабдомиома, *рабдосаркома*, *рабдомиосаркома*
разделение, *разделенная*, *разделенные*, *разделенный*
раннее, *ранний*, *ранняя*
растяжение, *растяжения*
ревматическая, *ревматический*
резина, *резины*
ректосигмоидная, *ректосигмоидное*
ретромолярная, *ретромолярный*
родий, *родия*
рудиментная, *рудиментный*, *рудиментарный*
саркома, *саркомы*, *саркоидоз*, *саркоидная*, *саркоматоз*, *саркоидного*
связи, *связь*
седалищная, *седалищный*
семейная, *семейное*, *семейной*, *семейный*
серп, *серповидно*, *серповидная*, *серповидный*, *серповидных*
силикатов, *силикатоз*
синцитиома, *синцитиальная*, *синцитиальный*
скрытая, *скрытое*, *скрытый*
слюна, *слюнно*, *слюнная*, *слюнной*
содержащий, *содержащих*
сонная, *сонной*, *сонный*
соустье, *соустья*
спинномозговая, *спинномозговой*, *спинномозговые*, *спинномозгового*
средства, *средство*
стено, *стенсена*
стоматоцит, *стоматоцитоз*
струма, *струмозный*
сумка, *сумки*
таллий, *таллия*
тело, *телом*
тестикулярно, *тестикулярная*, *тестикулярный*
тонкая, *тонкие*, *тонкий*, *тонкой*, *тонкого*
трансплантат, *трансплантант*, *трансплантата*, *трансплантация*, *трансплантационный*, *трансплантированная*, *трансплантированный*
триад, *триада*
трофобластическая, *трофобластические*
тучная, *тучный*, *тучных*, *тучноклеточная*, *тучноклеточные*, *тучноклеточный*
удвоение, *удвоенный*
уплотнение, *уплотненная*, *уплотненный*
усиление, *усиленная*, *усиленное*, *усиленный*
фаланг, *фаланга*, *фаланги*, *фаланговый*
фетальная, *фетальный*
фиксации, *фиксация*, *фиксированный*
фруктозо, *фруктозы*, *фруктоземия*, *фруктозурия*, *фруктокиназы*
хлор, *хлора*, *хлорид*, *хлоридов*
хориоидная, *хориоидное*
царапин, *царапина*
церебральная, *церебральной*, *церебральные*, *церебральный*, *церебральных*, *цереброваскулярная*, *цереброваскулярный*
циркоий, *циркония*
частиц, *частицы*
четвертый, *четвертых*
шейка, *шейки*
экзантема, *экзантематозный*
эластоз, *эластома*, *эластоидная*, *эластическая*, *эластические*, *эластический*, *эластических*, *эластического*, *эластофиброма*
эндокард, *эндокардиальный*
эпендимома, *эпендимоцит*, *эпендимальная*, *эпендимобластома*
эритема, *эритематозная*, *эритематозный*
эфферентная, *эфферентный*
ядро, *ядром*

A.2 Apprentissage à partir de séries de synonymes étiquetées et lemmatisées

A.2.1 Jeu d'étiquettes morphosyntaxiques `Brill` INaLF

ABR – abréviation
ACJ:sg – verbe *avoir* conjugué au singulier
ADJ2PAR:pl – participe passé adjectival pluriel
ADJ2PAR:sg – participe passé adjectival singulier
ADJ:pl – adjectif pluriel
ADJ:sg – adjectif singulier
ADV – adverbe
ANCFF – verbe *avoir* à l'infinitif
ANCNT – participe présent ou gérondif du verbe *avoir*
CAR – adjectif cardinal
COO – coordonnant
DTC:pl – déterminant du groupe nominal contracté pluriel
DTC:sg – déterminant du groupe nominal contracté singulier
DTN:pl – déterminant du groupe nominal non contracté pluriel
DTN:sg – déterminant du groupe nominal non contracté singulier
ECJ:sg – verbe *être* conjugué au singulier
ENCFF – verbe *être* infinitif
EPAR:sg – participe présent du verbe *être*
NN – substantif singulier
NNP – nom propre
PFX – préfixe détaché
PREP – préposition
PRO:++ – pronom genre indéterminé
PRO:pl – pronom pluriel
PRO:sg – pronom singulier
PRV:sg – pronom personnel
PUL – particule
REL – relatif
SBC:pl – substantif pluriel
SUB$ – subordonnant, code par défaut de *que*
SUB – subordonnant
VCJ:sg – autre verbe conjugué au singulier
VNCFF – autre verbe à l'infinitif
VNCNT – participe présent ou gérondif d'autres verbes

A.2.2 Règles (fréquence ≥ 5)

Étiquetage (154 règles sur 594)

584 *ϵ/ADJ:sg|s/ADJ:pl*
519 *ϵ/NN|s/SBC:pl*
263 *ϵ/ADJ:sg|e/ADJ:sg*
128 *ϵ/ADJ:sg|es/ADJ:pl*
72 *e/NN|ique/ADJ:sg*
67 *e/NN|aire/ADJ:sg*
64 *e/NN|que/ADJ:sg*
54 *ϵ/NN|ϵ/ADJ:sg*
50 *ation/NN|é/ADJ:sg*
49 *x/ADJ:sg|se/ADJ:sg*
48 *e/ADJ:sg|s/ADJ:pl*
42 *o/PFX|al/ADJ:sg*
39 *e/NN|o/PFX*
38 *o/PFX|ite/NN*
38 *le/ADJ:sg|ux/ADJ:pl*
38 *l/ADJ:sg|ux/ADJ:pl*
37 *e/NN|ose/NN*
36 *es/SBC:pl|aire/ADJ:sg*
34 *ϵ/ADJ:sg|ment/ADV*
32 *f/ADJ:sg|ve/ADJ:sg*
31 *me/NN|sarcome/NN*
31 *e/NN|ome/NN*
31 *e/NN|atose/NN*
31 *ϵ/ADJ:sg|ne/ADJ:sg*
27 *on/NN|f/ADJ:sg*
27 *e/NN|é/ADJ:sg*
26 *ϵ/ADJ:sg|le/ADJ:sg*
25 *o/PFX|aire/ADJ:sg*
23 *se/NN|tique/ADJ:sg*
23 *me/NN|se/NN*
22 *o/PFX|ique/ADJ:sg*
22 *on/NN|ve/ADJ:sg*
21 *ϵ/NN|ux/ADJ:sg*
20 *ose/NN|eux/ADJ:sg*
20 *ome/NN|ique/ADJ:sg*
20 *ation/NN|ée/ADJ:sg*
19 *ϵ/NN|use/ADJ:sg*
19 *e/NN|ques/ADJ:pl*
18 *me/NN|blastome/NN*
18 *f/ADJ:sg|ves/ADJ:pl*
18 *e/NN|ienne/ADJ:sg*
18 *ement/NN|é/ADJ:sg*
17 *s/ADJ:pl|ne/ADJ:sg*
17 *ϵ/PFX|me/NN*
17 *ose/NN|es/SBC:pl*
17 *me/NN|ïde/ADJ:sg*
17 *e/NN|ateux/ADJ:sg*
17 *e/NN|ale/ADJ:sg*
17 *e/NN|al/ADJ:sg*
16 *x/ADJ:sg|ses/ADJ:pl*
16 *ose/NN|euse/ADJ:sg*
16 *es/SBC:pl|que/ADJ:sg*
16 *ϵ/ADJ:sg|ϵ/ADJ2PAR:sg*
15 *ϵ/PFX|se/NN*
15 *ose/NN|aire/ADJ:sg*
14 *ome/NN|es/SBC:pl*
14 *ité/NN|e/ADJ:sg*
14 *e/NN|ien/ADJ:sg*
14 *ce/NN|t/ADJ:sg*
13 *um/NN|al/ADJ:sg*
13 *se/NN|ïde/ADJ:sg*
13 *ome/NN|aire/ADJ:sg*
13 *e/NN|oïde/ADJ:sg*
13 *ϵ/ADJ:sg|les/ADJ:pl*
12 *e/NN|ateuse/ADJ:sg*
11 *sarcome/NN|ïde/ADJ:sg*
11 *ose/NN|ique/ADJ:sg*
11 *o/PFX|um/NN*
11 *ome/NN|al/ADJ:sg*
11 *ϵ/NN|e/ADJ:sg*
11 *ie/NN|e/ADJ:sg*
11 *e/NN|urie/NN*
11 *e/NN|ée/ADJ:sg*
11 *e/ADJ:sg|ienne/ADJ:sg*
10 *ure/NN|é/ADJ:sg*
10 *tion/NN|nte/ADJ:sg*
10 *o/PFX|eux/ADJ:sg*
10 *é/ADJ:sg|ant/ADJ:sg*
9 *s/SBC:pl|ux/ADJ:sg*
9 *ϵ/NN|nelle/ADJ:sg*
9 *ϵ/NN|nel/ADJ:sg*
9 *ϵ/NN|é/ADJ:sg*
9 *ϵ/NN|al/ADJ:sg*
9 *me/NN|fibrome/NN*
9 *me/NN|blastique/ADJ:sg*
9 *eux/ADJ:sg|aire/ADJ:sg*
9 *é/ADJ:sg|es/SBC:pl*
9 *ϵ/ADJ:sg|ement/ADV*
8 *tion/NN|ire/ADJ:sg*
8 *te/NN|enne/ADJ:sg*
8 *sarcome/NN|blastome/NN*
8 *sarcome/NN|blastique/ADJ:sg*
8 *o/PFX|é/ADJ:sg*
8 *ome/NN|ale/ADJ:sg*
8 *ϵ/NN|l/ADJ:sg*
8 *me/NN|lipome/NN*
8 *matose/NN|blastome/NN*
8 *ite/NN|ose/NN*
8 *ite/NN|opathie/NN*
8 *ique/ADJ:sg|oïde/ADJ:sg*
8 *e/NN|ie/NN*
7 *ur/NN|ϵ/ADJ:sg*
7 *tion/NN|nt/ADJ:sg*
7 *sme/NN|que/ADJ:sg*
7 *o/PFX|ien/ADJ:sg*
7 *ome/NN|ales/ADJ:pl*
7 *ϵ/NN|ique/ADJ:sg*
7 *ϵ/NN|aire/ADJ:sg*
7 *me/NN|carcinome/NN*
7 *ion/NN|oire/ADJ:sg*
7 *ion/NN|é/ADJ:sg*
7 *ïde/ADJ:sg|mateuse/ADJ:sg*
7 *euse/ADJ:sg|oïde/ADJ:sg*
7 *e/NN|ante/ADJ:sg*
7 *é/ADJ:sg|cation/NN*
7 *cytome/NN|blastome/NN*
6 *us/NN|ique/ADJ:sg*
6 *us/NN|ale/ADJ:sg*
6 *se/NN|pathie/NN*
6 *se/NN|matose/NN*
6 *ϵ/PFX|n/NN*
6 *ose/NN|e/ADJ:sg*
6 *ϵ/NN|o/PFX*
6 *ïde/ADJ:sg|blastique/ADJ:sg*
6 *eux/ADJ:sg|oïde/ADJ:sg*
6 *e/NN|ocyte/NN*
6 *e/NN|émie/NN*
6 *e/ADJ:sg|ien/ADJ:sg*
6 *é/ADJ:sg|aire/ADJ:sg*
5 *x/NN|gé/ADJ:sg*
5 *us/NN|al/ADJ:sg*
5 *ses/SBC:pl|tique/ADJ:sg*
5 *o/PFX|ienne/ADJ:sg*
5 *ome/NN|inique/ADJ:sg*
5 *ome/NN|e/ADJ:sg*
5 *ome/NN|ϵ/ADJ:sg*
5 *ϵ/NN|x/SBC:pl*
5 *ϵ/NN|né/ADJ:sg*
5 *me/NN|cytaire/ADJ:sg*
5 *ion/NN|oires/ADJ:pl*
5 *ion/NN|ϵ/ADJ:sg*
5 *ilité/NN|le/ADJ:sg*
5 *es/SBC:pl|oïde/ADJ:sg*
5 *er/ADJ:sg|ère/ADJ:sg*
5 *e/NN|yome/NN*
5 *e/NN|iase/NN*
5 *e/NN|atique/ADJ:sg*
5 *e/NN|ant/ADJ:sg*
5 *cation/NN|qué/ADJ:sg*
5 *a/NN|o/PFX*
5 *aire/ADJ:sg|euse/ADJ:sg*
5 *ϵ/ADJ:sg|ssement/NN*
5 *ϵ/ADJ:sg|isé/ADJ:sg*
5 *a/ADJ:sg|e/ADJ:sg*

Lemmatisation (117 règles sur 501)

77 *e|ique*
71 *e|aire*
70 *e|que*
64 *é|ation*
45 *o|al*
43 *e|ose*
41 *e|o*
40 *e|é*
39 *o|ite*
39 *f|on*
32 *e|ome*
31 *me|sarcome*
31 *e|atose*
28 *ϵ|ux*
27 *e|al*
26 *e|ien*
25 *o|aire*
24 *se|tique*
24 *o|ique*
24 *eux|ose*
24 *e|ateux*
23 *me|se*
21 *é|er*
20 *ome|ique*
19 *é|ement*
18 *t|ce*
18 *nt|tion*
18 *me|ïde*
18 *me|blastome*
18 *ϵ|me*
17 *é|ant*
16 *ϵ|se*
16 *e|ie*
15 *ose|aire*
15 *al|um*
14 *ϵ|ment*
14 *e|oïde*
14 *ϵ|al*
13 *se|ïde*
13 *ome|aire*
13 *o|é*
13 *ϵ|nel*
13 *ϵ|ement*
13 *ϵ|é*
12 *ose|ique*
12 *ite|opathie*
12 *e|ant*
12 *é|aire*
12 *al|ome*
11 *o|um*
11 *o|eux*
11 *ϵ|n*
11 *ϵ|l*
11 *ite|ose*
11 *ion|oire*
11 *ïde|sarcome*
11 *eux|aire*
11 *e|urie*
11 *é|ure*
11 *en|te*
10 *ïde|mateux*
10 *er|ation*
9 *ϵ|o*
9 *me|fibrome*
9 *me|blastique*
9 *ϵ|ique*
9 *é|ion*
8 *sarcome|blastome*
8 *sarcome|blastique*
8 *ϵ|ome*
8 *o|ien*
8 *me|lipome*
8 *matose|blastome*
8 *ϵ|isé*
8 *ire|tion*

8 *ique|oïde*
8 *ϵ|ion*
8 *e|ité*
8 *e|émie*
8 *ϵ|aire*
7 *se|pathie*
7 *que|sme*
7 *ϵ|né*
7 *me|carcinome*
7 *eux|oïde*
7 *e|ocyte*
7 *ène|énique*
7 *é|cation*
7 *e|atique*
7 *cytome|blastome*
7 *al|us*
7 *a|e*
6 *x|gé*
6 *us|ique*
6 *ϵ|ur*
6 *ϵ|ssement*
6 *se|matose*
6 *le|ilité*
6 *ïde|blastique*
6 *a|o*
5 *ϵ|virus*
5 *ϵ|ure*
5 *qué|cation*
5 *ome|inique*
5 *me|cytaire*
5 *ϵ|ité*
5 *ϵ|isation*
5 *ïde|gène*
5 *gène|sarcome*
5 *e|yome*
5 *ϵ|eux*
5 *e|osarcome*
5 *ème|émateux*
5 *e|ier*
5 *e|iase*
5 *é|atif*
5 *aire|opathie*

Étiquetage et lemmatisation (111 règles sur 513)

75 *e/SBC|ique/ADJ*
75 *ϵ/ADJ|ϵ/SBC*
71 *e/SBC|aire/ADJ*
70 *e/SBC|que/ADJ*
61 *é/ADJ|ation/SBC*
45 *o/PFX|al/ADJ*
41 *e/SBC|ose/SBC*
41 *e/SBC|o/PFX*
39 *o/PFX|ite/SBC*
38 *f/ADJ|on/SBC*
36 *e/SBC|é/ADJ*
32 *e/SBC|ome/SBC*
31 *me/SBC|sarcome/SBC*
31 *e/SBC|atose/SBC*
28 *ϵ/SBC|ux/ADJ*
25 *o/PFX|aire/ADJ*
24 *se/SBC|tique/ADJ*
24 *eux/ADJ|ose/SBC*
24 *e/SBC|ateux/ADJ*
24 *e/SBC|al/ADJ*
23 *o/PFX|ique/ADJ*
23 *me/SBC|se/SBC*
22 *e/SBC|ien/ADJ*
21 *er/V|é/ADJ*
20 *ome/SBC|ique/ADJ*
19 *é/ADJ|ement/SBC*
18 *me/SBC|blastome/SBC*
17 *ϵ/PFX|me/SBC*
17 *me/SBC|ïde/ADJ*
17 *é/ADJ|ant/ADJ*
16 *nt/ADJ|tion/SBC*
15 *t/ADJ|ce/SBC*
15 *ϵ/PFX|se/SBC*
15 *ose/SBC|aire/ADJ*
15 *al/ADJ|um/SBC*
14 *e/SBC|oïde/ADJ*
13 *se/SBC|ïde/ADJ*
13 *ϵ/SBC|nel/ADJ*
13 *ome/SBC|aire/ADJ*
13 *ϵ/ADJ|ment/ADV*
12 *ϵ/SBC|é/ADJ*
12 *ϵ/SBC|al/ADJ*
12 *ose/SBC|ique/ADJ*
12 *o/PFX|é/ADJ*
12 *ite/SBC|opathie/SBC*
12 *al/ADJ|ome/SBC*
11 *ϵ/SBC|l/ADJ*
11 *o/PFX|um/SBC*
11 *o/PFX|eux/ADJ*
11 *ite/SBC|ose/SBC*
11 *ion/SBC|oire/ADJ*
11 *ïde/ADJ|sarcome/SBC*
11 *eux/ADJ|aire/ADJ*
11 *e/SBC|urie/SBC*
11 *e/SBC|ant/ADJ*
11 *en/ADJ|te/SBC*
11 *é/ADJ|ure/SBC*
11 *e/ADJ|ie/SBC*
11 *e/ADJ|ien/ADJ*
11 *ϵ/ADJ|ement/ADV*
10 *er/V|ation/SBC*
9 *me/SBC|fibrome/SBC*
9 *me/SBC|blastique/ADJ*
9 *ïde/ADJ|mateux/ADJ*
9 *é/ADJ|ion/SBC*
9 *é/ADJ|aire/ADJ*
8 *ϵ/SBC|ique/ADJ*
8 *sarcome/SBC|blastome/SBC*
8 *sarcome/SBC|blastique/ADJ*
8 *o/PFX|ien/ADJ*
8 *me/SBC|lipome/SBC*
8 *matose/SBC|blastome/SBC*
8 *ire/ADJ|tion/SBC*
8 *ique/ADJ|oïde/ADJ*
8 *e/SBC|ie/SBC*
8 *e/SBC|émie/SBC*
7 *se/SBC|pathie/SBC*
7 *ϵ/SBC|o/PFX*
7 *ϵ/SBC|né/ADJ*
7 *ϵ/SBC|aire/ADJ*
7 *que/ADJ|sme/SBC*
7 *me/SBC|carcinome/SBC*
7 *eux/ADJ|oïde/ADJ*
7 *e/ADJ|ité/SBC*
7 *é/ADJ|cation/SBC*
7 *cytome/SBC|blastome/SBC*
7 *al/ADJ|us/SBC*
6 *x/SBC|gé/ADJ*
6 *us/SBC|ique/ADJ*
6 *se/SBC|matose/SBC*
6 *ϵ/PFX|n/SBC*
6 *le/ADJ|ilité/SBC*
6 *ïde/ADJ|blastique/ADJ*
6 *e/SBC|ocyte/SBC*
6 *e/SBC|atique/ADJ*
6 *ϵ/ADJ|ssement/SBC*
6 *ϵ/ADJ|ome/SBC*
6 *ϵ/ADJ|isé/ADJ*
5 *ϵ/SBC|ion/SBC*
5 *ϵ/SBC|eux/ADJ*
5 *qué/ADJ|cation/SBC*
5 *ome/SBC|inique/ADJ*
5 *me/SBC|cytaire/ADJ*
5 *e/SBC|yome/SBC*
5 *e/SBC|osarcome/SBC*
5 *e/SBC|iase/SBC*
5 *é/SBC|ation/SBC*
5 *ème/SBC|émateux/ADJ*
5 *e/ADJ|ose/SBC*
5 *a/SBC|o/PFX*
5 *aire/ADJ|opathie/SBC*

A.2.3 Suffixes (fréquence ≥ 5)

Étiquetage (142 suffixes sur 489)

1202 *-ϵ/ADJ:sg*
726 *-ϵ/NN*
649 *-s/ADJ:pl*
604 *-e/NN*
533 *-s/SBC:pl*
388 *-e/ADJ:sg*
234 *-o/PFX*
209 *-aire/ADJ:sg*
188 *-é/ADJ:sg*
168 *-me/NN*
158 *-ique/ADJ:sg*
141 *-ose/NN*
137 *-ome/NN*
128 *-es/ADJ:pl*
113 *-es/SBC:pl*
110 *-se/NN*
98 *-que/ADJ:sg*
98 *-al/ADJ:sg*
80 *-ation/NN*
79 *-ïde/ADJ:sg*
78 *-f/ADJ:sg*
77 *-sarcome/NN*
76 *-ux/ADJ:pl*
69 *-le/ADJ:sg*
67 *-x/ADJ:sg*
64 *-ϵ/PFX*
62 *-eux/ADJ:sg*
54 *-ve/ADJ:sg*
54 *-ite/NN*
53 *-on/NN*
51 *-blastome/NN*
49 *-se/ADJ:sg*
48 *-ne/ADJ:sg*
46 *-l/ADJ:sg*
43 *-oïde/ADJ:sg*
41 *-euse/ADJ:sg*
38 *-ienne/ADJ:sg*
35 *-ment/ADV*
35 *-atose/NN*
34 *-tique/ADJ:sg*
34 *-ale/ADJ:sg*
33 *-ien/ADJ:sg*
32 *-um/NN*
32 *-ée/ADJ:sg*
32 *-cytaire/ADJ:sg*
31 *-blastique/ADJ:sg*
30 *-ux/ADJ:sg*
30 *-ion/NN*
27 *-us/NN*
26 *-tion/NN*
24 *-ie/NN*
21 *-matose/NN*
20 *-x/NN*
20 *-ques/ADJ:pl*
20 *-ire/ADJ:sg*
20 *-ement/NN*
19 *-use/ADJ:sg*
18 *-ves/ADJ:pl*
18 *-s/NN*
17 *-ure/NN*
17 *-ce/NN*
17 *-ateux/ADJ:sg*
17 *-ant/ADJ:sg*
16 *-ses/ADJ:pl*
16 *-ité/NN*

16 *-ϵ/ADJ2PAR:sg*
15 *-carcinome/NN*
14 *-t/ADJ:sg*
14 *-opathie/NN*
13 *-les/ADJ:pl*
13 *-fibrome/NN*
13 *-émie/NN*
12 *-gène/ADJ:sg*
12 *-cation/NN*
12 *-ateuse/ADJ:sg*
11 *-urie/NN*
11 *-sclérose/NN*
11 *-ocyte/NN*
10 *-te/NN*
10 *-pathie/NN*
10 *-nte/ADJ:sg*
9 *-nt/ADJ:sg*
9 *-nelle/ADJ:sg*
9 *-nel/ADJ:sg*
9 *-mateuse/ADJ:sg*
9 *-gliome/NN*
9 *-ement/ADV*
9 *-cytome/NN*
9 *-ante/ADJ:sg*
8 *-osarcome/NN*
8 *-lipome/NN*
8 *-isation/NN*
8 *-enne/ADJ:sg*
8 *-a/NN*
7 *-ur/NN*
7 *-sme/NN*
7 *-oire/ADJ:sg*
7 *-n/ADJ:sg*
7 *-logique/ADJ:sg*
7 *-é/NN*
7 *-cellulaire/ADJ:sg*
7 *-ales/ADJ:pl*
7 *-adénome/NN*
6 *-ylique/ADJ:sg*
6 *-plasie/NN*
6 *-n/NN*
6 *-iques/ADJ:pl*
6 *-ine/NN*
6 *-ielle/ADJ:sg*
6 *-glie/NN*
6 *-génique/ADJ:sg*
6 *-er/ADJ:sg*
6 *-énique/ADJ:sg*
6 *-ène/ADJ:sg*
6 *-endothéliome/NN*
6 *-atique/ADJ:sg*
5 *-yome/NN*
5 *-x/SBC:pl*
5 *-ulaire/ADJ:sg*
5 *-ssement/NN*
5 *-ses/SBC:pl*
5 *-qué/ADJ:sg*
5 *-oires/ADJ:pl*
5 *-oblastome/NN*
5 *-né/ADJ:sg*
5 *-mateux/ADJ:sg*
5 *-logue/ADJ:sg*
5 *-kératose/NN*
5 *-isé/ADJ:sg*
5 *-inome/NN*
5 *-inique/ADJ:sg*
5 *-ilité/NN*
5 *-iforme/ADJ:sg*
5 *-iel/ADJ:sg*
5 *-iase/NN*
5 *-greffe/NN*
5 *-gé/ADJ:sg*
5 *-ère/ADJ:sg*
5 *-en/ADJ:sg*
5 *-dermie/NN*
5 *-chondrosarcome/NN*
5 *-a/ADJ:sg*

Lemmatisation (107 suffixes sur 408)

639 *-e*
265 *-é*
245 *-o*
176 *-ique*
176 *-aire*
171 *-me*
124 *-al*
122 *-ose*
113 *-se*
102 *-ome*
90 *-que*
86 *-ide*
79 *-sarcome*
78 *-eux*
77 *-ation*
62 *-ite*
53 *-blastome*
47 *-ien*
44 *-on*
41 *-f*
40 *-um*
40 *-er*
36 *-oïde*
35 *-ement*
35 *-ant*
33 *-cytaire*
32 *-atose*
31 *-tique*
31 *-ion*
31 *-blastique*
28 *-ux*
27 *-tion*
25 *-us*
24 *-ire*
24 *-ateux*
23 *-en*
22 *-ure*
22 *-ce*
21 *-matose*
20 *-x*
20 *-nt*
19 *-opathie*
19 *-mateux*
19 *-ie*
19 *-gène*
18 *-t*
16 *-émie*
15 *-s*
15 *-ocyte*
15 *-n*
15 *-carcinome*
15 *-a*
14 *-ment*
13 *-te*
13 *-nel*
13 *-ité*
13 *-fibrome*
12 *-ène*
12 *-cation*
11 *-urie*
11 *-sclérose*
11 *-pathie*
11 *-oire*
11 *-l*
10 *-osarcome*
9 *-gliome*
9 *-énique*
9 *-cytome*
8 *-lipome*
8 *-le*
8 *-isé*
8 *-ine*
8 *-atique*
7 *-sme*
7 *-né*
7 *-ier*
7 *-cellulaire*
7 *-adénome*
6 *-ylique*
6 *-ur*
6 *-ssement*
6 *-sion*
6 *-plasie*
6 *-logique*
6 *-isation*
6 *-ilité*
6 *-iforme*
6 *-glie*
6 *-génique*
6 *-gé*
6 *-endothéliome*
6 *-atif*
5 *-yome*
5 *-virus*
5 *-ulaire*
5 *-qué*
5 *-oblastome*
5 *-logue*
5 *-kératose*
5 *-inique*
5 *-iase*
5 *-greffe*
5 *-ème*
5 *-émateux*
5 *-dermie*
5 *-chondrosarcome*
5 *-ase*

Étiquetage et lemmatisation (109 suffixes sur 427)

595 *-e/SBC*
246 *-é/ADJ*
243 *-o/PFX*
215 *-ϵ/SBC*
174 *-aire/ADJ*
169 *-me/SBC*
168 *-ique/ADJ*
144 *-ϵ/ADJ*
123 *-ose/SBC*
119 *-al/ADJ*
112 *-se/SBC*
100 *-ome/SBC*
89 *-que/ADJ*
81 *-ide/ADJ*
79 *-ation/SBC*
78 *-eux/ADJ*
77 *-sarcome/SBC*
64 *-ϵ/PFX*
62 *-ite/SBC*
53 *-blastome/SBC*
50 *-ien/ADJ*
42 *-on/SBC*
40 *-f/ADJ*
40 *-e/ADJ*
39 *-um/SBC*
39 *-er/V*
34 *-oïde/ADJ*
33 *-ant/ADJ*
32 *-cytaire/ADJ*
32 *-atose/SBC*
31 *-tique/ADJ*
31 *-blastique/ADJ*
28 *-ux/ADJ*
28 *-ion/SBC*
25 *-tion/SBC*
24 *-ateux/ADJ*
23 *-us/SBC*
22 *-ie/SBC*
21 *-matose/SBC*
21 *-ement/SBC*
20 *-ire/ADJ*
19 *-ure/SBC*
19 *-opathie/SBC*
18 *-x/SBC*
18 *-nt/ADJ*
18 *-mateux/ADJ*
18 *-ce/SBC*
17 *-en/ADJ*
16 *-émie/SBC*
15 *-t/ADJ*
15 *-s/SBC*
15 *-carcinome/SBC*
14 *-ocyte/SBC*
13 *-te/SBC*
13 *-nel/ADJ*
13 *-ment/ADV*
13 *-fibrome/SBC*
12 *-gène/ADJ*
12 *-cation/SBC*
11 *-urie/SBC*
11 *-sclérose/SBC*
11 *-pathie/SBC*
11 *-oire/ADJ*
11 *-l/ADJ*
11 *-ité/SBC*
11 *-ement/ADV*
10 *-osarcome/SBC*
9 *-gliome/SBC*
9 *-é/SBC*
9 *-cytome/SBC*
8 *-n/ADJ*
8 *-lipome/SBC*
8 *-a/SBC*
7 *-sme/SBC*
7 *-né/ADJ*
7 *-ine/SBC*
7 *-cellulaire/ADJ*
7 *-atique/ADJ*
7 *-adénome/SBC*
6 *-ylique/ADJ*

6 *-ssement/SBC*
6 *-plasie/SBC*
6 *-n/SBC*
6 *-logique/ADJ*
6 *-le/ADJ*
6 *-isé/ADJ*
6 *-ilité/SBC*
6 *-iforme/ADJ*
6 *-glie/SBC*
6 *-génique/ADJ*
6 *-gé/ADJ*
6 *-énique/ADJ*
6 *-ène/ADJ*
6 *-endothéliome/SBC*
5 *-yome/SBC*
5 *-ulaire/ADJ*
5 *-sion/SBC*
5 *-qué/ADJ*
5 *-oblastome/SBC*
5 *-logue/ADJ*
5 *-kératose/SBC*
5 *-isation/SBC*
5 *-inique/ADJ*
5 *-iase/SBC*
5 *-greffe/SBC*
5 *-ème/SBC*
5 *-émateux/ADJ*
5 *-dermie/SBC*
5 *-chondrosarcome/SBC*

A.2.4 Préfixes (fréquence $\geq$ 5)

Étiquetage (83 préfixes sur 2 784)

29 *myélo-*
25 *ostéo-*
22 *fibro-*
22 *fibr-*
20 *adéno-*
19 *lympho-*
16 *lipo-*
16 *granul-*
15 *angio-*
13 *plasmocyt-*
13 *chondro-*
12 *ménin-g*
12 *lymphocyt-*
11 *immun-*
11 *histiocyt-*
11 *hémangio-*
10 *hyalin-*
9 *ulcér-*
9 *neuro-*
9 *néphro-*
9 *mélano-*
9 *lymph-*
9 *hyper-*
9 *hépat-*
8 *synovi-*
8 *papul-*
8 *myxo-*
8 *monocyt-*
8 *hépato-*
8 *glio-*
8 *dent-*
8 *corticosurrénal-*
8 *acid-*
7 *tubul-*
7 *tubercul-*
7 *radi-*
7 *papill-*
7 *derm-*
6 *spontané-*
6 *réticul-*
6 *pustul-*
6 *parasit-*
6 *oligodendro-*
6 *oesophag-*
6 *kérato-*
6 *bronch-*
6 *amygdal-*
5 *trachéo-*
5 *thrombo-*
5 *synovial-*
5 *segment-*
5 *protéin-*
5 *plasmocyto-*
5 *pigment-*
5 *pharyng-*
5 *pharyn-*
5 *palm-*
5 *microcyt-*
5 *métacarp-*
5 *mélanocyt-*
5 *macul-*
5 *localisé-*
5 *ligament-*
5 *leucocyt-*
5 *léiomyo-*
5 *kyst-*
5 *immédiat-*
5 *histiocyto-*
5 *glomérul-*
5 *froid-*
5 *follicul-*
5 *épithéli-*
5 *épididym-*
5 *épiderm-*
5 *endométrio-*
5 *cylindr-*
5 *cutané-*
5 *cholédo-*
5 *cardi-*
5 *carcino-*
5 *bronchiol-*
5 *astrocyt-*
5 *adén-*

Lemmatisation (58 préfixes sur 1 430)

30 *myélo-*
25 *ostéo-*
23 *fibro-*
21 *adéno-*
19 *lympho-*
16 *lipo-*
16 *angio-*
15 *fibr-*
13 *ménin-g*
13 *chondro-*
12 *mélano-*
12 *hémangio-*
11 *granul-*
10 *lymph-*
9 *réticul-*
9 *radi-*
9 *plasmocyt-*
9 *neuro-*
9 *néphro-*
9 *lymphocyt-*
9 *hyper-*
8 *papul-*
8 *myxo-*
8 *hépato-*
8 *hépat-*
8 *glio-*
7 *protéin-*
7 *histiocyt-*
7 *glomérul-*
7 *derm-*
7 *carcino-*
6 *ulcér-*
6 *tubul-*
6 *oligodendro-*
6 *kérato-*
6 *kérat-*
6 *bronch-*
6 *acid-*
5 *tubercul-*
5 *synovi-*
5 *pustul-*
5 *plasmocyto-*
5 *papill-*
5 *palm-*
5 *monocyt-*
5 *microcyt-*
5 *mélan-*
5 *macul-*
5 *lipid-*
5 *léiomyo-*
5 *immun-*
5 *hypertrophi-*
5 *histiocyto-*
5 *épiderm-*
5 *endométrio-*
5 *dent-*
5 *cardi-*
5 *adén-*

Étiquetage et lemmatisation (59 préfixes sur 1 451)

30 *myélo-*
25 *ostéo-*
23 *fibro-*
20 *adéno-*
19 *lympho-*
16 *lipo-*
16 *angio-*
15 *fibr-*
13 *ménin-g*
13 *chondro-*
12 *hémangio-*
11 *granul-*
10 *lymph-*
9 *réticul-*
9 *radi-*
9 *plasmocyt-*
9 *neuro-*
9 *néphro-*
9 *mélano-*
9 *lymphocyt-*
9 *hyper-*
8 *papul-*
8 *myxo-*
8 *hépato-*
8 *hépat-*
8 *glio-*
7 *protéin-*
7 *histiocyt-*
7 *glomérul-*
7 *derm-*
7 *acid-*
6 *ulcér-*
6 *tubul-*
6 *oligodendro-*
6 *kérato-*
6 *kérat-*
6 *bronch-*
5 *tubercul-*
5 *synovi-*
5 *pustul-*
5 *plasmocyto-*
5 *papill-*
5 *palm-*
5 *monocyt-*
5 *microcyt-*

5 *mélan-*	5 *léiomyo-*	5 *histiocyto-*	5 *dent-*	5 *carcino-*
5 *macul-*	5 *immun-*	5 *épiderm-*	5 *cholédo-*	5 *adén-*
5 *lipid-*	5 *hypertrophi-*	5 *endométrio-*	5 *cardi-*	

A.2.5 Familles

Étiquetage (183 familles sur 1 647)

abdomen/NN, *abdomino/PFX*, *abdominal/ADJ:sg*, *abdominale/ADJ:sg*, *abdominaux/ADJ:pl*, *abdominales/ADJ:pl*
acanthocyte/NN, *acanthrocyte/NN*
acquis/ADJ:sg, *acquise/ADJ:sg*, *acquises/ADJ:pl*
adipeux/ADJ:sg, *adipeuse/ADJ:sg*
affection/NN, *affectant/VNCNT*, *affectif/ADJ:sg*, *affectifs/ADJ:pl*, *affective/ADJ:sg*, *affections/SBC:pl*
alcali/NN, *alcalis/SBC:pl*
amiante/NN, *amiantose/NN*
amylase/NN, *amylose/NN*, *amyloïde/NN*, *amyloïdose/NN*, *amyloses/SBC:pl*, *amyloïde/ADJ:sg*
angiodysplasie/NN, *angiodysplasique/ADJ:sg*
anomalie/NN, *anomalies/SBC:pl*
antéro/PFX, *antéflexion/NN*, *antérieur/ADJ:sg*, *antérieure/ADJ:sg*, *antérieurs/ADJ:pl*
aponévrose/NN, *aponévroses/SBC:pl*, *aponévrotique/ADJ:sg*
arrachement/NN, *arraché/ADJ:sg*
artériole/NN, *artérioles/SBC:pl*, *artériolaire/ADJ:sg*
astrocyte/NN, *astrocytome/NN*, *astrogliome/NN*, *astroblastome/NN*, *astrocytes/SBC:pl*, *astrocytaire/ADJ:sg*
atteinte/NN, *atteintes/SBC:pl*
autogreffe/NN, *autologue/ADJ:sg*, *autoplastique/ADJ:sg*
bandelette/NN, *bandelettes/SBC:pl*
bilatéral/ADJ:sg, *bilatérale/ADJ:sg*, *bilatéraux/ADJ:pl*, *bilatérales/ADJ:pl*
blanc/ADJ:sg, *blancs/ADJ:pl*
branche/NN, *branches/SBC:pl*, *branchial/ADJ:sg*, *branchiale/ADJ:sg*
bulle/NN, *bulleux/ADJ:sg*, *bulleuse/ADJ:sg*, *bulleuses/ADJ:pl*
callosité/NN, *callosités/SBC:pl*
carbone/NN, *carbonate/NN*, *carbonique/ADJ:sg*
cartilage/NN, *cartilages/SBC:pl*, *cartilagineux/ADJ:sg*, *cartilagineuse/ADJ:sg*
certain/ADJ:sg, *certains/ADJ:pl*, *certaines/ADJ:pl*
chimique/ADJ:sg, *chimiques/ADJ:pl*
cholédoque/NN, *cholédocho/PFX*, *cholédoque/ADJ:sg*, *cholédocien/ADJ:sg*, *cholédocholithiase/NN*
chyle/NN, *chylurie/NN*, *chylocèle/NN*, *chyleux/ADJ:sg*
clinique/ADJ:sg, *clinoïde/ADJ:sg*, *cliniquement/ADV*
colle/NN, *colloïde/ADJ:sg*
complexe/NN, *complexe/ADJ:sg*, *complexes/ADJ:pl*, *complexes/SBC:pl*
conduit/NN, *conduits/SBC:pl*
conservation/NN, *conservé/ADJ:sg*
contrôle/NN, *contrôlé/ADJ:sg*, *contrôlé/ADJ2PAR:sg*
coronarien/ADJ:sg, *coronariens/ADJ:pl*, *coronarienne/ADJ:sg*
couleur/NN, *couleurs/SBC:pl*
croissant/NN, *croissants/SBC:pl*
culture/NN, *cultures/SBC:pl*
cytopathologie/NN, *cytopathologique/ADJ:sg*
cérébro/PFX, *cérébral/ADJ:sg*, *cérébrale/ADJ:sg*, *cérébraux/ADJ:pl*, *cérébrales/ADJ:pl*
dermatologique/ADJ:sg, *dermatopathique/ADJ:sg*
diamniotique/ADJ:sg, *diamniotiques/ADJ:pl*
directe/ADJ:sg, *directement/ADV*, *directes/ADJ:pl*
disséquant/ADJ:sg, *disséquante/ADJ:sg*
douve/NN, *douves/SBC:pl*
dysfonction/NN, *dysfonctions/SBC:pl*, *dysfonctionnel/ADJ:sg*
dédifférenciation/NN, *dédifférencié/ADJ:sg*
dégénéré/ADJ:sg, *dégénérescence/NN*, *dégénératif/ADJ:sg*, *dégénérative/ADJ:sg*, *dégénératives/ADJ:pl*
dépression/NN, *déprimée/ADJ:sg*, *dépressif/ADJ:sg*, *dépressifs/ADJ:pl*
détresse/NN, *détresses/SBC:pl*
embryon/NN, *embryonnaire/ADJ:sg*
encéphalite/NN, *encéphalopathie/NN*, *encéphalites/SBC:pl*, *encéphalitique/ADJ:sg*
enfant/NN, *enfants/SBC:pl*
entéro/PFX, *entérite/NN*, *entéropathie/NN*, *entérique/ADJ:sg*, *entérites/SBC:pl*, *entéroïde/ADJ:sg*, *entéropathiques/ADJ:pl*
exagérée/ADJ:sg, *exagérées/ADJ:pl*
exocrine/ADJ:sg, *exocrines/ADJ:pl*
extrapyramidal/ADJ:sg, *extrapyramidaux/ADJ:pl*
faible/ADJ:sg, *faibles/ADJ:pl*
fenêtre/NN, *fenêtré/ADJ:sg*
fissure/NN, *fissuré/ADJ:sg*, *fissurée/ADJ:sg*, *fissuraire/ADJ:sg*
foncé/ADJ:sg, *foncée/ADJ:sg*
fracture/NN, *fracturé/ADJ:sg*, *fractures/SBC:pl*
fumée/NN, *fumées/SBC:pl*
galactose/NN, *galactosurie/NN*, *galactosémie/NN*, *galactokinase/NN*
genou/NN, *genoux/SBC:pl*
globe/NN, *globus/NN*, *global/ADJ:sg*, *globale/ADJ:sg*
gonflement/NN, *gonflé/ADJ:sg*
griffe/NN, *griffes/SBC:pl*
général/ADJ:sg, *générale/ADJ:sg*, *généraux/ADJ:pl*, *généralement/ADV*, *générales/ADJ:pl*, *généralisé/ADJ:sg*, *généralisée/ADJ:sg*, *généralisés/ADJ:pl*, *généralisées/ADJ:pl*
hidradénite/NN, *hidrosadénite/NN*
huile/NN, *huilome/NN*, *huiles/SBC:pl*, *huileuse/ADJ:sg*
hydrops/NN, *hydropisie/NN*
hyperparathyroïdie/NN, *hyperparathyroïdies/SBC:pl*
hypoglycémies/SBC:pl, *hypoglycémique/ADJ:sg*
hypothalamus/NN, *hypothalamique/ADJ:sg*
hématie/NN, *hémato/PFX*, *hématite/NN*, *hématome/NN*, *hématies/SBC:pl*, *hématique/ADJ:sg*
hépato/PFX, *hépatite/NN*, *hépatome/NN*, *hépatico/PFX*, *hépatocyte/NN*, *hépatopathie/NN*, *hépatique/ADJ:sg*, *hépatites/SBC:pl*, *hépatiques/ADJ:pl*, *hépatoblastome/NN*, *hépatocarcinome/NN*, *hépatocholangiome/NN*, *hépatocellulaire/ADJ:sg*, *hépatocellulaires/ADJ:pl*, *hépatocholangiocarcino-*

me/NN, *hépatocholangiocellulaire/ADJ:sg*
ilio/PFX, *ilion/NN*
implant/NN, *implants/SBC:pl*
incompatibilité/NN, *incompatible/ADJ:sg*
indéterminé/ADJ:sg, *indéterminée/ADJ:sg*
inguinal/ADJ:sg, *inguinale/ADJ:sg*, *inguinaux/ADJ:pl*
intention/NN, *intentionnelle/ADJ:sg*
intervention/NN, *interventions/SBC:pl*
intramural/ADJ:sg, *intramurale/ADJ:sg*
investigation/NN, *investigations/SBC:pl*
jambe/NN, *jambes/SBC:pl*
koïlocytose/NN, *koïlocytaire/ADJ:sg*
laissé/ADJ:sg, *laissée/ADJ:sg*
lentille/NN, *lentilles/SBC:pl*
ligament/NN, *ligaments/SBC:pl*, *ligamenteux/ADJ:sg*, *ligamentaire/ADJ:sg*
liquide/NN, *liquide/ADJ:sg*, *liquides/SBC:pl*, *liquidienne/ADJ:sg*
lombosacré/ADJ:sg, *lombosacrées/ADJ:pl*
lymphorrhée/NN, *lymphorragie/NN*
lépromateux/ADJ:sg, *lépromateuse/ADJ:sg*
majeur/ADJ:sg, *majeure/ADJ:sg*, *majeures/ADJ:pl*
mamelon/NN, *mamelonnaire/ADJ:sg*
massif/ADJ:sg, *massive/ADJ:sg*
membre/NN, *membres/SBC:pl*
migraine/NN, *migraines/SBC:pl*, *migraineux/ADJ:sg*
modéré/ADJ:sg, *modérée/ADJ:sg*
mort/NN, *mort/ADJ:sg*, *morts/ADJ:pl*
multiple/ADJ:sg, *multiples/ADJ:pl*, *multifocal/ADJ:sg*, *multifocale/ADJ:sg*
myofibrome/NN, *myofibrose/NN*, *myofibromatose/NN*
myéloplaxe/NN, *myéloplaxes/SBC:pl*
média/NN, *médio/PFX*, *médian/ADJ:sg*, *médiane/ADJ:sg*
méningo/PFX, *méningite/NN*, *méningiome/NN*, *méningé/ADJ:sg*, *méninges/SBC:pl*, *méningée/ADJ:sg*, *méningées/ADJ:pl*, *méningiomatose/NN*, *méningites/SBC:pl*, *méningiomes/SBC:pl*, *méningothéliome/NN*, *méningioblastome/NN*, *méningothélial/ADJ:sg*
métacarpe/NN, *métacarpo/PFX*, *métacarpien/ADJ:sg*, *métacarpiens/ADJ:pl*, *métacarpienne/ADJ:sg*
naso/PFX, *nasopharynx/NN*, *nasopharyngé/ADJ:sg*, *nasopharyngien/ADJ:sg*
neutropénie/NN, *neutropénique/ADJ:sg*
nourrisson/NN, *nourrissons/SBC:pl*
négligence/NN, *négligences/SBC:pl*
observation/NN, *observations/SBC:pl*
odontome/NN, *odontogénèse/NN*, *odontosarcome/NN*, *odontogène/ADJ:sg*, *odontogènes/ADJ:pl*, *odontogénique/ADJ:sg*
oophoro/PFX, *oophorite/NN*
orbite/NN, *orbitale/ADJ:sg*, *orbitaire/ADJ:sg*
orthopédique/ADJ:sg, *orthopédiques/ADJ:pl*
ostéo/PFX, *ostéite/NN*, *ostéome/NN*, *ostéose/NN*, *ostéoblaste/NN*, *ostéodermie/NN*, *ostéofibrome/NN*, *ostéosarcome/NN*, *ostéoïde/ADJ:sg*, *ostéoblastome/NN*, *ostéochondrome/NN*, *ostéogénique/ADJ:sg*, *ostéofibrosarcome/NN*, *ostéoblastique/ADJ:sg*, *ostéochondromatose/NN*, *ostéochondrosarcome/NN*, *ostéocartilagineuse/ADJ:sg*
oxyphile/ADJ:sg, *oxyphiles/ADJ:pl*
papille/NN, *papillome/NN*, *papilles/SBC:pl*, *papillomatose/NN*, *papillaire/ADJ:sg*, *papillaires/ADJ:pl*, *papillifère/ADJ:sg*, *papillokératose/NN*, *papillokystique/ADJ:sg*
parasite/NN, *parasitose/NN*, *parasite/ADJ:sg*, *parasites/SBC:pl*, *parasitaire/ADJ:sg*, *parasitaires/ADJ:pl*
pathologie/NN, *pathologique/ADJ:sg*, *pathologiques/ADJ:pl*
permanent/ADJ:sg, *permanente/ADJ:sg*, *permanentes/ADJ:pl*
phlegmon/NN, *phlegmoneuse/ADJ:sg*
pian/NN, *pianique/ADJ:sg*, *pianiques/ADJ:pl*
plantes/SBC:pl, *plantaire/ADJ:sg*, *plantaires/ADJ:pl*
plié/ADJ:sg, *pliée/ADJ:sg*
polyembryome/NN, *polyembryonique/ADJ:sg*
pont/NN, *ponto/PFX*
postmaturité/NN, *postmature/ADJ:sg*
principal/ADJ:sg, *principale/ADJ:sg*, *principalement/ADV*, *principales/ADJ:pl*
profondeur/NN, *profond/ADJ:sg*, *profonde/ADJ:sg*, *profonds/ADJ:pl*, *profondes/ADJ:pl*
proportionné/ADJ:sg, *proportionnel/ADJ:sg*
proéminent/ADJ:sg, *proéminentes/ADJ:pl*
préleucémie/NN, *préleucémique/ADJ:sg*
présentation/NN, *présentations/SBC:pl*
pubis/NN, *pubien/ADJ:sg*, *pubienne/ADJ:sg*
putréfaction/NN, *putréfié/ADJ:sg*, *putrescent/ADJ:sg*
pédoncule/NN, *pédonculé/ADJ:sg*
périnée/NN, *périnéal/ADJ:sg*, *périnéale/ADJ:sg*
périapical/ADJ:sg, *périapicale/ADJ:sg*, *périapicaux/ADJ:pl*, *périodontique/ADJ:sg*, *périodontiques/ADJ:pl*
radical/NN, *radicale/ADJ:sg*
rare/ADJ:sg, *rarement/ADV*
remplacement/NN, *remplacé/ADJ:sg*
rompu/ADJ:sg, *rompue/ADJ:sg*, *rompu/ADJ2PAR:sg*
récidive/NN, *récidivant/ADJ:sg*, *récidivante/ADJ:sg*, *récidivants/ADJ:pl*, *récidivantes/ADJ:pl*
réparation/NN, *réparé/ADJ:sg*
rétrocession/NN, *rétroflexion/NN*
sarcoïde/NN, *sarcoïdose/NN*, *sarcoïdosique/ADJ:sg*
secondaire/ADJ:sg, *secondaires/ADJ:pl*
service/NN, *services/SBC:pl*
sinusite/NN, *sinusites/SBC:pl*, *sinusienne/ADJ:sg*
souche/NN, *souche/ADJ:sg*, *souches/ADJ:pl*
spina/NN, *spinal/ADJ:sg*, *spinale/ADJ:sg*, *spinales/ADJ:pl*
sport/NN, *sportif/ADJ:sg*, *sportive/ADJ:sg*, *sportives/ADJ:pl*
stature/NN, *statural/ADJ:sg*, *staturale/ADJ:sg*
structure/NN, *structurel/ADJ:sg*, *structures/SBC:pl*, *structurels/ADJ:pl*
sudoripare/ADJ:sg, *sudoripares/ADJ:pl*
surnuméraire/ADJ:sg, *surnuméraires/ADJ:pl*
synapse/NN, *synaptique/ADJ:sg*
sébacé/ADJ:sg, *sébacée/ADJ:sg*
séronégatif/ADJ:sg, *séronégative/ADJ:sg*
tendineux/ADJ:sg, *tendineuse/ADJ:sg*, *tendineuses/ADJ:pl*
thrombo/PFX, *thrombus/NN*, *thrombose/NN*, *thrombocyte/NN*, *thrombocytose/NN*, *thrombosé/ADJ:sg*, *thromboses/SBC:pl*, *thrombosant/ADJ:sg*, *thrombosées/ADJ:pl*, *thrombotique/ADJ:sg*, *thrombocytaire/ADJ:sg*
tissu/NN, *tissus/SBC:pl*, *tissulaire/ADJ:sg*, *tissulaires/ADJ:pl*
transitionnel/ADJ:sg, *transitionnelle/ADJ:sg*, *transitionnelles/ADJ:pl*
tricholeucocytose/NN, *tricholeucocyte/NN*, *tricholeucocytes/SBC:pl*, *tricholymphocytes/SBC:pl*
tronc/NN, *troncs/SBC:pl*
tuméfaction/NN, *tuméfié/ADJ:sg*, *tuméfactions/SBC:pl*
tétrachlorure/NN, *tétrachloroéthylène/NN*
uropathie/NN, *uropathies/SBC:pl*
vagin/NN, *vagino/PFX*, *vaginite/NN*, *vaginal/ADJ:sg*, *vaginale/ADJ:sg*, *vaginales/ADJ:pl*
vasculo/PFX, *vasculite/NN*, *vasculaire/ADJ:sg*, *vasculaires/ADJ:pl*

vertèbre/NN, *vertébro/PFX*, *vertèbres/SBC:pl*, *vertébral/ADJ:sg*, *vertébrale/ADJ:sg*, *vertébraux/ADJ:pl*, *vertébrales/ADJ:pl*
visuel/ADJ:sg, *visuelles/ADJ:pl*
vomissure/NN, *vomies/ADJ:sg*, *vomissures/SBC:pl*
xanthome/NN, *xanthélasma/NN*, *xanthomatose/NN*, *xanthofibrome/NN*, *xanthinique/ADJ:sg*, *xanthomateux/ADJ:sg*
élargi/ADJ:sg, *élargissement/NN*
énanthème/NN, *énanthémateux/ADJ:sg*
épididyme/NN, *épididymo/PFX*, *épididymite/NN*, *épididymes/SBC:pl*, *épididymaire/ADJ:sg*
épithélium/NN, *épithéliale/NN*, *épithélioma/NN*, *épithéliome/NN*, *épithélial/ADJ:sg*, *épithélisation/NN*, *épithéliale/ADJ:sg*, *épithéliales/ADJ:pl*, *épithéliosarcome/NN*, *épithélioïde/ADJ:sg*, *épithélioïdes/ADJ:pl*, *épithéliomateuse/ADJ:sg*
érythrémie/NN, *érythrocyte/NN*, *érythroblaste/NN*, *érythroleucose/NN*, *érythroïde/ADJ:sg*, *érythroblastose/NN*, *érythroleucémie/NN*, *érythrocytes/SBC:pl*, *érythrémique/ADJ:sg*, *érythrocytaire/ADJ:sg*, *érythroblastique/ADJ:sg*, *érythropoïétique/ADJ:sg*

Lemmatisation (178 familles sur 1 064)

abdomen, *abdomino*, *abdominal*
acanthocyte, *acanthrocyte*
actinomycose, *actinomycosique*
adolescent, *adolescence*
agglutiné, *agglutination*
allergie, *allergique*
amibe, *amibien*, *amibiase*
amputé, *amputer*, *amputation*
anastomose, *anastomotique*
anguleux, *angulation*
anormal, *anormalement*
aorte, *aorto*, *aortite*, *aortique*
arachnoïde, *arachnoïdien*
arthrite, *arthrose*, *arthropathie*, *arthropathique*
associé, *associer*, *association*
ataxie, *ataxique*
atypie, *atypique*
autologue, *autogreffe*, *autoplastique*
baso, *basocellulaire*, *basospinocellulaire*
blessé, *blessure*
bronche, *broncho*, *bronchite*, *bronchique*, *bronchectasie*, *bronchiectasie*, *bronchogénique*
béryllium, *bérylliose*
calcium, *calcaire*, *calcifié*, *calcifiant*, *calcification*
carbone, *carbonate*, *carbonique*
cartilage, *cartilagineux*
changé, *changeant*, *changement*
chlore, *chlorome*, *chlorure*
chondromalacie, *chondromalacique*
chrome, *chronique*
circulant, *circulaire*, *circulation*, *circulatoire*
coagulant, *coagulation*
colorant, *coloration*
comportement, *comportemental*
conscient, *conscience*
contaminé, *contaminer*
convergent, *convergence*
cortico, *cortical*
cristal, *cristallin*
cutané, *cutanéo*
cytopathologie, *cytopathologique*
céroïde, *céroïdo*
dermatofibrome, *dermatofibrosarcome*
diagnostic, *diagnostique*
dilaté, *dilatation*
distordu, *distorsion*
duodéno, *duodénal*, *duodénum*, *duodénite*
dystrophie, *dystrophique*
dédoublé, *dédoublement*
dépendant, *dépendance*
déshydrase, *déshydratase*
ecchondrome, *ecchondrose*
emphysème, *emphysémateux*
endocrin, *endocrine*, *endocrinien*
enfant, *enfance*
envahissant, *envahissement*
excavé, *excavation*
explicatif, *explication*
fascia, *fascial*
fibre, *fibro*, *fibreux*, *fibrine*, *fibrome*, *fibrose*, *fibroïde*, *fibrineux*, *fibrocyte*, *fibrosant*, *fibrinoïde*, *fibromyome*, *fibrillaire*, *fibrinogène*, *fibroblaste*, *fibrolipome*, *fibromateux*, *fibromatose*, *fibromyxome*, *fibroplasie*, *fibrosarcome*, *fibrokystique*, *fibroplasique*, *fibrosclérose*, *fibrosclérosé*, *fibroxanthome*, *fibroblastique*, *fibromyxolipome*, *fibroliposarcome*, *fibromyxosarcome*, *fibroaméloblastique*, *fibrochondrosarcome*
fièvre, *fièvreux*
fongique, *fongoïde*
fructose, *fructosurie*, *fructosémie*, *fructokinase*
fémoro, *fémoral*
gangrène, *gangreneux*
glio, *gliome*, *gliose*, *gliomatose*, *gliosarcome*, *glioblastome*
glucagon, *glucagonome*
gonocoque, *gonorrhoeae*
grippe, *grippal*
helminthe, *helminthiase*
histochimie, *histochimique*
hyalin, *hyalinisé*, *hyalinose*, *hyalinisation*
hyperchrome, *hyperchromie*, *hyperchromatique*
hypersécrétion, *hypersécrétoire*
hypokaliémie, *hypokaliémique*
hypostase, *hypostatique*
hématomètre, *hématométrie*
hémoglobine, *hémoglobinose*, *hémoglobinurie*, *hémoglobinopathie*
hépato, *hépatico*, *hépatite*, *hépatome*, *hépatique*, *hépatocyte*, *hépatopathie*, *hépatoblastome*, *hépatocarcinome*, *hépatocellulaire*, *hépatocholangiome*, *hépatocholangiocarcinome*, *hépatocholangiocellulaire*
identifié, *identifiable*, *identification*
implant, *implanté*, *implantation*
indenté, *indentation*
infiltrant, *infiltration*
insuffisant, *insuffisance*
invaginé, *invaginant*, *invagination*
ischémie, *ischémique*
kupffer, *kupfferome*
lactase, *lactate*, *lactose*, *lactique*
leptocyte, *leptocytose*
lichénifié, *lichénification*
lipogranulome, *lipogranulomatose*

local, *localisé*, *localiser*, *localisation*
lysosome, *lysosomial*
malade, *maladie*
marginé, *margination*
mention, *mentionné*
minéral, *minéralisé*, *minéralisation*
morphologie, *morphologique*
multiple, *multifocal*
myocyte, *myocarde*, *myocardique*
myélinolyse, *myélinoclasie*
médiat, *médiation*
médullo, *médullaire*
mésangial, *mésangium*
métacarpe, *métacarpo*, *métacarpien*
méthémoglobine, *méthémoglobinémie*
neutrophile, *neutrophilie*, *neutrophilique*
nucléole, *nucléolaire*
névrite, *névritique*
occlu, *occlusif*, *occlusion*
oligodendrocyte, *oligodendroglie*, *oligodendrocytome*, *oligodendrogliome*, *oligodendrocytaire*, *oligodendroblastome*
opération, *opératoire*
ossifié, *ossifiant*, *ossification*
ouvert, *ouverture*
pallidu, *pallidum*
paraffine, *paraffinome*
parent, *parental*
pelvi, *pelvis*, *pelvien*, *pelvipéritonite*
phaeohyphomycose, *phaeohyphomycosique*
phosphore, *phosphoré*
pilomatrixome, *pilomatricaire*
plaquette, *plaquettaire*
pneumonie, *pneumoniae*, *pneumocoque*, *pneumopathie*
polyurie, *polyurique*
porte, *portal*, *portant*, *porteur*
premier, *premièrement*
profession, *professionnel*
promyélocyte, *promyélocytaire*
protéine, *protéino*, *protéinase*, *protéinose*, *protéinique*, *protéinurie*
préleucémie, *préleucémique*
préservatif, *préservation*
pulpe, *pulpeux*, *pulpaire*
pyogène, *pyogénique*
pédoncule, *pédonculé*
périnée, *périnéal*
péroné, *péronier*
radiculaire, *radiculopathie*
recto, *rectu*, *rectal*, *rectum*
rhabdomyome, *rhabdomyosarcome*, *rhabdomyoblastique*
rouge, *rougeur*
réduit, *réductible*
répété, *répéter*, *répétition*
rétropéritoine, *rétropéritonéal*
sacro, *sacré*, *sacrum*
schizophrénie, *schizophrénique*
section, *sectionné*
sertoli, *sertolien*
sinus, *sinusal*, *sinusien*, *sinusite*
soupçonné, *soupçonneux*
spina, *spinal*
spongioblaste, *spongioblastome*
squame, *squameux*
stomatocyte, *stomatocytose*
sténose, *sténosé*, *sténosant*, *sténotique*
superinvolutif, *superinvolution*
synapse, *synaptique*
séquestre, *séquestré*, *séquestration*
teint, *teinture*
thalassémie, *thalassémique*
thécal, *thécome*
trachée, *trachéo*, *trachéal*, *trachéite*
transplant, *transplanté*, *transplantation*
trichiura, *trichuris*, *trichocéphale*, *trichocéphalose*
tubercule, *tuberculeux*, *tuberculose*, *tuberculosi*, *tuberculoïde*
ténia, *téniase*
urine, *urinaire*
utilisé, *utilisation*
valve, *valvule*, *valvulaire*
vasculo, *vasculite*, *vasculaire*, *vasculopathie*
vers, *version*
vitamine, *vitamino*
vérifié, *vérification*
échancré, *échancrure*
élargi, *élargissement*
épais, *épaissi*, *épaississement*
épiglotte, *épiglotto*, *épiglottite*, *épiglottique*
épithélial, *épithélium*, *épithéliale*, *épithélioma*, *épithéliome*, *épithélioïde*, *épithélisation*, *épithéliomateux*, *épithéliosarcome*
éruptif, *éruption*
évolutif, *évolution*

Étiquetage et lemmatisation (180 familles sur 1 078)

abdomen/SBC, *abdomino/PFX*, *abdominal/ADJ*
acanthocyte/SBC, *acanthrocyte/SBC*
actinomycose/SBC, *actinomycosique/ADJ*
administré/ADJ, *administratif/ADJ*, *administration/SBC*
affaissé/ADJ, *affaissement/SBC*
aliment/SBC, *alimentaire/ADJ*, *alimentation/SBC*
amiante/SBC, *amiantose/SBC*
amphophile/ADJ, *amphophilie/SBC*
analgésie/SBC, *analgésique/ADJ*, *analgésique/SBC*
angiome/SBC, *angiomyome/SBC*, *angiolipome/SBC*, *angiomateux/ADJ*, *angiomatose/SBC*, *angiomyxome/SBC*, *angiofibrome/SBC*, *angiosarcome/SBC*, *angioblastome/SBC*, *angiochoriome/SBC*, *angiokératome/SBC*, *angiokératose/SBC*, *angiosclérose/SBC*, *angioblastique/ADJ*, *angioléiomyome/SBC*, *angiomyolipome/SBC*, *angiomyosarcome/SBC*, *angioréticulome/SBC*, *angioendothéliome/SBC*, *angioendothéliomatose/SBC*
annulé/SBC, *annulaire/ADJ*
anémie/SBC, *anémique/ADJ*
appendice/SBC, *appendico/PFX*, *appendicite/SBC*, *appendiculaire/ADJ*
arrhénome/SBC, *arrhénoblastome/SBC*
artério/PFX, *artérioveineux/ADJ*, *artérioscléreux/ADJ*, *artériosclérose/SBC*, *artériosclérotique/ADJ*
astrocyte/SBC, *astrocytome/SBC*, *astrogliome/SBC*, *astrocytaire/ADJ*, *astroblastome/SBC*
atrophie/SBC, *atrophiant/ADJ*, *atrophique/ADJ*, *atrophodermie/SBC*

autolyse/SBC, *autolytique/ADJ*
baryum/SBC, *barytose/SBC*
bilieux/ADJ, *biliaire/ADJ*
branche/SBC, *branchial/ADJ*
bulle/SBC, *bulleux/ADJ*
calcique/ADJ, *calcinose/SBC*
canal/SBC, *canalaire/ADJ*
cardia/SBC, *cardio/PFX*, *cardite/SBC*, *cardiaque/ADJ*, *cardiopathie/SBC*, *cardiomégalie/SBC*
cavité/SBC, *cavitaire/ADJ*, *cavitation/SBC*
chimiotaxie/SBC, *chimiotactisme/SBC*
cholestérol/SBC, *cholestéatome/SBC*
choroïde/ADJ, *choroïde/SBC*, *choroïdien/ADJ*
cilié/ADJ, *ciliaire/ADJ*
clitoris/SBC, *clitorido/PFX*, *clitoridien/ADJ*
colle/SBC, *colloïde/ADJ*
complément/SBC, *complémentaire/ADJ*
confirmé/ADJ, *confirmation/SBC*
consolidé/ADJ, *consolidation/SBC*
contraceptif/ADJ, *contraceptif/SBC*, *contraception/SBC*
cornée/SBC, *cornéen/ADJ*
coupe/SBC, *coupure/SBC*
cryoglobuline/SBC, *cryoglobulinémie/SBC*
cyanure/SBC, *cyanhydrique/ADJ*
cytologie/SBC, *cytopathie/SBC*, *cytologique/ADJ*, *cytopathique/ADJ*
cérébro/PFX, *cérébral/ADJ*
derme/SBC, *dermo/PFX*, *dermite/SBC*, *dermique/ADJ*, *dermoïde/ADJ*, *dermatite/SBC*, *dermatose/SBC*, *dermatosis/SBC*
diaphragme/SBC, *diaphragmatique/ADJ*
discriminant/ADJ, *discrimination/SBC*
donner/V, *donné/ADJ*
durci/ADJ, *durcissement/SBC*
décollé/ADJ, *décollement/SBC*
dégénéré/ADJ, *dégénératif/ADJ*, *dégénérescence/SBC*
déplacé/ADJ, *déplacement/SBC*
désoxyadénosine/SBC, *désoxyadénylique/ADJ*
ectasie/SBC, *ectasié/ADJ*
enclavé/ADJ, *enclavement/SBC*
endoscope/SBC, *endoscopie/SBC*, *endoscopique/ADJ*
enroulé/ADJ, *enroulement/SBC*
esthésioneuro/PFX, *esthésioneurome/SBC*, *esthésioneurocytome/SBC*, *esthésioneuroblastome/SBC*
exfoliatif/ADJ, *exfoliation/SBC*
extrême/ADJ, *extrême/SBC*, *extrêmement/ADV*
faux/ADJ, *faux/SBC*
figure/SBC, *figuré/ADJ*
fluor/SBC, *fluoré/ADJ*, *fluorure/SBC*
fracture/SBC, *fracturé/ADJ*
fréquent/ADJ, *fréquence/SBC*, *fréquemment/ADV*
galactose/SBC, *galactosurie/SBC*, *galactosémie/SBC*, *galactokinase/SBC*
gastro/PFX, *gastrite/SBC*, *gastrique/ADJ*
glissant/ADJ, *glissement/SBC*
glycogénose/SBC, *glycogénique/ADJ*
graisse/SBC, *graisseux/ADJ*
général/ADJ, *généralisé/ADJ*, *généralement/ADV*
herpès/SBC, *herpétique/ADJ*, *herpèsvirus/SBC*
homologue/ADJ, *homogreffe/SBC*
hydrocéphale/SBC, *hydrocéphalie/SBC*
hyperimmunoglobuline/SBC, *hyperimmunoglobulinémie/SBC*
hypertensif/ADJ, *hypertension/ADJ*, *hypertension/SBC*
hypopharynx/SBC, *hypopharyngé/ADJ*
hypothalamus/SBC, *hypothalamique/ADJ*
hématologique/ADJ, *hématopoïétique/ADJ*
hémolyse/SBC, *hémolytique/ADJ*
hétérotopie/SBC, *hétérotopique/ADJ*
identifié/ADJ, *identifiable/ADJ*, *identification/SBC*
implant/SBC, *implanté/ADJ*, *implantation/SBC*
induré/ADJ, *induration/SBC*
inflammation/SBC, *inflammatoire/ADJ*
insuffisant/ADJ, *insuffisance/SBC*
invaginé/ADJ, *invaginant/ADJ*, *invagination/SBC*
ischémie/SBC, *ischémique/ADJ*
kupffer/SBP, *kupfferome/SBC*
lactase/SBC, *lactate/SBC*, *lactose/SBC*, *lactique/ADJ*
leptocyte/SBC, *leptocytose/SBC*
lichénifié/ADJ, *lichénification/SBC*
lipogranulome/SBC, *lipogranulomatose/SBC*
local/ADJ, *localiser/V*, *localisé/ADJ*, *localisation/SBC*
lysosome/SBC, *lysosomial/ADJ*
malade/ADJ, *maladie/SBC*
mandibule/SBC, *mandibulo/PFX*, *mandibulaire/ADJ*
mastite/SBC, *mastopathie/SBC*
microscope/SBC, *microscopie/SBC*, *microscopique/ADJ*, *microphotographie/SBC*
modifier/V, *modifié/ADJ*, *modification/SBC*
morphée/SBC, *morphéiforme/ADJ*
multiloculari/ADJ, *multiloculaire/ADJ*
mycose/SBC, *mycosis/SBC*, *mycosique/ADJ*
myxome/SBC, *myxoïde/ADJ*, *myxolipome/SBC*, *myxomateux/ADJ*, *myxofibrome/SBC*, *myxosarcome/SBC*, *myxoliposarcome/SBC*
médiastin/SBC, *médiastinal/ADJ*
médullo/PFX, *médullaire/ADJ*, *médullaire/SBC*
mésangial/ADJ, *mésangium/SBC*
métacarpe/SBC, *métacarpo/PFX*, *métacarpien/ADJ*
méthémoglobine/SBC, *méthémoglobinémie/SBC*
neutrophile/ADJ, *neutrophile/SBC*, *neutrophilie/SBC*, *neutrophilique/ADJ*
nutrition/SBC, *nutritionnel/ADJ*
névroglie/SBC, *névroglique/ADJ*
occlu/ADJ, *occlusif/ADJ*, *occlusion/SBC*
olécrâne/SBC, *olécrânien/ADJ*
opération/SBC, *opératoire/ADJ*
ossifié/ADJ, *ossifiant/ADJ*, *ossification/SBC*
ouvert/ADJ, *ouverture/SBC*
palmo/PFX, *palmé/ADJ*, *palmure/SBC*, *palmaire/ADJ*, *palmature/SBC*
parakératose/SBC, *parakératosique/ADJ*
pariéto/PFX, *pariétal/ADJ*
pemphigus/SBC, *pemphigoïde/ADJ*, *pemphigoïde/SBC*
phalange/SBC, *phalangien/ADJ*
photographie/SBC, *photomicrographie/SBC*
pinguicula/SBC, *pinguécula/SBC*
plasmocyte/SBC, *plasmocytome/SBC*, *plasmocytose/SBC*, *plasmocytaire/ADJ*, *plasmocytoïde/ADJ*, *plasmocytosarcome/SBC*
poil/SBC, *poilu/ADJ*
polype/SBC, *polypose/SBC*, *polypoïde/ADJ*, *polyadénome/SBC*, *polyadénomatose/SBC*
positif/ADJ, *position/SBC*
principal/ADJ, *principalement/ADV*
progresser/V, *progressif/ADJ*, *progression/SBC*
propre/ADJ, *proprement/ADV*
précipité/ADJ, *précipité/SBC*, *précipitation/SBC*
préménopause/SBC, *préménopausique/ADJ*
psycho/PFX, *psyché/SBC*, *psychose/SBC*, *psychique/ADJ*, *psychogène/ADJ*, *psychoactif/ADJ*, *psychotique/ADJ*

pustule/SBC, *pustulo/PFX*, *pustuleux/ADJ*, *pustulaire/ADJ*
pyélo/PFX, *pyélite/SBC*, *pyélique/ADJ*
péricyte/SBC, *péricarde/SBC*, *péricardique/ADJ*
péritoine/SBC, *péritonéal/ADJ*
rachis/SBC, *rachidien/ADJ*
ratatiné/ADJ, *ratatinement/SBC*
respiration/SBC, *respiratoire/ADJ*
rompre/V, *rompu/ADJ*
récent/ADJ, *récemment/ADV*
régénéré/ADJ, *régénération/SBC*
rétine/SBC, *rétinien/ADJ*, *rétinocytome/SBC*, *rétinoblastome/SBC*
révisé/ADJ, *révision/SBC*
satellite/ADJ, *satellitose/SBC*
scroto/PFX, *scrotal/ADJ*, *scrotum/SBC*
septal/ADJ, *septum/SBC*
silice/SBC, *silicate/SBC*, *silicium/SBC*, *silicatose/SBC*
souche/ADJ, *souche/SBC*
sperme/SBC, *spermatique/ADJ*, *spermatozoïde/SBC*
spirale/SBC, *spiroïde/ADJ*
spontané/ADJ, *spontanément/ADV*
squirrhe/SBC, *squirrheux/ADJ*
strie/SBC, *strié/ADJ*
subinvolutif/ADJ, *subinvolution/SBC*
susceptible/ADJ, *susceptibilité/SBC*
syringoadénome/SBC, *syringocystadénome/SBC*
sévère/ADJ, *sévèrement/ADV*
tendon/SBC, *tendineux/ADJ*
thrombopénie/SBC, *thrombopénique/ADJ*, *thrombocytopénie/SBC*
tissu/SBC, *tissulaire/ADJ*
trame/SBC, *tramatique/ADJ*
transséquer/V, *transséqué/ADJ*, *transversu/ADJ*, *transversal/ADJ*, *transneuronal/ADJ*, *transposition/SBC*, *transsynaptique/ADJ*, *transversalement/ADV*
trigone/ADJ, *trigone/SBC*
trouble/ADJ, *trouble/SBC*
tympan/SBC, *tympanique/ADJ*
uridine/SBC, *uridylique/ADJ*
urétéro/PFX, *urétéral/ADJ*, *urétérite/SBC*
vagin/SBC, *vagino/PFX*, *vaginal/ADJ*, *vaginite/SBC*
vascularisé/ADJ, *vascularisation/SBC*
verrue/SBC, *verruqueux/ADJ*
villo/PFX, *villeux/ADJ*
vulve/SBC, *vulvo/PFX*, *vulvae/ADJ*, *vulvite/SBC*, *vulvaire/ADJ*
wolff/SBP, *wolffien/ADJ*
écrasé/ADJ, *écrasement/SBC*
électron/SBC, *électronique/ADJ*
épiderme/SBC, *épidermo/PFX*, *épidermique/ADJ*, *épidermoïde/ADJ*
épiplo/PFX, *épiploon/SBC*, *épiploïque/ADJ*
éraflé/ADJ, *éraflure/SBC*
éventré/ADJ, *éventration/SBC*

A.3 Apprentissage à partir d'autres relations sémantiques encodées dans les terminologies

A.3.1 Règles (fréquence ≥ 5)

Relations hiérarchiques de la SNOMED (162 règles sur 606)

1140 $\epsilon\|s$	31 $me\|sarcome$	19 $e\|al$	13 $ome\|aire$	9 $o\|é$	7 $\epsilon\|x$
290 $\epsilon\|e$	31 $e\|ite$	18 $t\|ce$	13 $nt\|tion$	9 $\epsilon\|o$	7 $que\|sme$
143 $\epsilon\|es$	31 $e\|atose$	18 $\epsilon\|nes$	13 $es\|ite$	9 $\epsilon\|nelle$	7 $ome\|ales$
74 $e\|ique$	28 $f\|on$	18 $me\|ïde$	13 $al\|um$	9 $\epsilon\|nel$	7 $\epsilon\|ome$
67 $e\|aire$	26 $o\|aire$	18 $me\|blastome$	13 $ale\|ite$	9 $\epsilon\|n$	7 $o\|ine$
64 $e\|que$	25 $te\|que$	18 $f\|ves$	12 $s\|nes$	9 $me\|fibrome$	7 $o\|ien$
60 $e\|s$	23 $se\|tique$	18 $es\|ose$	12 $e\|ateuse$	9 $me\|blastique$	7 $me\|carcinome$
55 $é\|ation$	23 $on\|ve$	18 $é\|ement$	11 $o\|um$	9 $é\|aire$	7
53 $x\|se$	23 $o\|ique$	17 $x\|ses$	11 $ose\|ique$	9 $blaste\|cytaire$	$mateuse\|sarcome$
43 $l\|ux$	23 $me\|se$	17 $e\|iques$	11 $e\|urie$	9 $\epsilon\|al$	7 $ion\|oire$
42 $o\|al$	23 $e\|ienne$	17 $e\|ateux$	11 $al\|ome$	8	7 $ïde\|cyte$
42 $e\|o$	22 $o\|es$	16 $ose\|euse$	11 $al\|ite$	$sarcome\|blastome$	7 $es\|oïde$
40 $le\|ux$	22 $e\|aires$	16 $es\|que$	11 $\epsilon\|ale$	8 $o\|iques$	7 $es\|ité$
40 $e\|ose$	21 $\epsilon\|ux$	16 $es\|aires$	10 $um\|ale$	8 $me\|lipome$	7 $en\|te$
39 $o\|ale$	20 $\epsilon\|se$	16 $e\|ien$	10 $\epsilon\|pathie$	8 $ite\|ose$	7 $e\|iennes$
38 $o\|ite$	20 $ome\|ique$	16 $e\|ie$	10 $ite\|aire$	8 $ite\|opathie$	7 $cytome\|blastome$
36 $es\|aire$	20 $o\|ales$	15 $ose\|aire$	10 $é\|ure$	8 $ire\|tion$	7 $\epsilon\|blaste$
36 $e\|ome$	20 $eux\|ose$	15 $o\|aux$	10 $es\|ienne$	8 $ique\|oïde$	6 $us\|ique$
34 $\epsilon\|ment$	20 $es\|ique$	15 $\epsilon\|les$	10 $é\|es$	8 $e\|isation$	6 $se\|pathie$
34 $f\|ve$	20 $e\|ques$	15 $e\|oïde$	10 $e\|ant$	8 $e\|is$	6 $se\|matose$
33 $\epsilon\|ne$	20 $ée\|ation$	14 $e\|ité$	10 $cyte\|blaste$	8 $ale\|ome$	6 $o\|us$
33 $\epsilon\|le$	20 $e\|ale$	14 $\epsilon\|cyte$	9 $s\|ux$	8 $\epsilon\|aire$	6 $o\|ées$
32 $e\|é$	19 $\epsilon\|use$	13 $se\|ïde$	9 $o\|ée$	8 $a\|e$	6 $\epsilon\|ite$

6 *ïde|blaste*
6 *fibrome|sarcome*
6 *e|opathie*
6 *e|émie*
6 *e|atique*
6 *ϵ|cytaire*
6 *al|us*
5 *x|gé*
5 *ϵ|virus*
5 *ϵ|vasculaire*
5 *um|que*
5 *se|tiques*
5 *se|pénie*
5 *o|ienne*
5 *ϵ|nement*
5 *ϵ|né*
5 *me|sclérose*
5 *me|cytaire*
5 *ion|oires*
5 *e|yome*
5 *é|ures*
5 *es|ien*
5 *e|on*
5 *ène|énique*
5 *ème|émateux*
5 *ϵ|cytose*

Relations transversales de la SNOMED (136 règles sur 447)

1140 *ϵ|s*
290 *ϵ|e*
143 *ϵ|es*
74 *e|ique*
67 *e|aire*
64 *e|que*
60 *e|s*
55 *é|ation*
53 *x|se*
43 *l|ux*
42 *o|al*
42 *e|o*
40 *le|ux*
40 *e|ose*
39 *o|ale*
38 *o|ite*
36 *es|aire*
36 *e|ome*
34 *ϵ|ment*
34 *f|ve*
33 *ϵ|ne*
33 *ϵ|le*
32 *e|é*
31 *e|ite*
31 *e|atose*
28 *f|on*
26 *o|aire*
25 *te|que*
23 *se|tique*
23 *on|ve*
23 *o|ique*
23 *me|se*
23 *e|ienne*
22 *o|es*
22 *e|aires*
21 *ϵ|ux*
20 *ϵ|se*
20 *ome|ique*
20 *eux|ose*
20 *es|ique*
20 *e|ques*
20 *ée|ation*
20 *e|ale*
19 *ϵ|use*
19 *é|ant*
19 *e|al*
18 *ϵ|nes*
18 *f|ves*
18 *es|ose*
18 *es|ome*
18 *é|ement*
17 *x|ses*
17 *ϵ|me*
17 *e|ée*
17 *e|ateux*
16 *ose|euse*
16 *es|aires*
16 *e|ien*
16 *e|ie*
15 *o|aux*
15 *ϵ|ïde*
15 *e|oïde*
14 *e|ité*
13 *nt|tion*
13 *es|ite*
13 *é|ante*
13 *al|um*
12 *e|ateuse*
11 *o|um*
11 *nte|tion*
11 *e|urie*
11 *al|ite*
11 *ϵ|ale*
10 *um|ale*
10 *ϵ|pathie*
10 *o|eux*
10 *ite|aire*
10 *ine|ose*
10 *es|ienne*
10 *é|es*
10 *e|ant*
9 *s|ux*
9 *o|é*
9 *ϵ|o*
9 *ϵ|nelle*
9 *ϵ|nel*
9 *ϵ|n*
9 *fs|on*
9 *eux|aire*
9 *ϵ|al*
8 *us|ale*
8 *o|euse*
8 *ϵ|l*
8 *ite|ose*
8 *ire|tion*
8 *ique|oïde*
8 *ϵ|ique*
8 *e|is*
8 *e|ante*
8 *ϵ|aire*
7 *ϵ|x*
7 *te|ques*
7 *que|sme*
7 *ϵ|ome*
7 *o|ien*
7 *ion|oire*
7 *en|te*
7 *e|iennes*
6 *us|ique*
6 *se|matose*
6 *o|us*
6 *ϵ|lyse*
6 *ϵ|ite*
6 *e|opathie*
6 *e|atique*
6 *a|o*
6 *al|us*
6 *al|es*
5 *x|gé*
5 *ϵ|virus*
5 *um|que*
5 *ϵ|stomie*
5 *ses|tique*
5 *o|ectomie*
5 *ϵ|lithiase*
5 *ire|ris*
5 *ion|oires*
5 *es|ien*
5 *er|ère*
5 *e|olyse*
5 *ène|énique*
5 *ème|émateux*
5 *é|ite*
5 *e|iase*
5 *ϵ|cèle*
5 *aire|euse*

Relations hiérarchiques de la CIM10 (75 règles sur 247)

1140 *ϵ|s*
290 *ϵ|e*
143 *ϵ|es*
74 *e|ique*
67 *e|aire*
64 *e|que*
60 *e|s*
55 *é|ation*
53 *x|se*
43 *s|es*
43 *l|ux*
42 *o|al*
42 *e|o*
40 *le|ux*
39 *o|ale*
38 *o|ite*
36 *es|aire*
33 *ϵ|ne*
33 *ϵ|le*
31 *e|ite*
28 *f|on*
26 *ux|les*
26 *o|aire*
25 *te|que*
23 *se|tique*
23 *on|ve*
23 *e|ienne*
22 *o|es*
22 *e|aires*
21 *ϵ|ux*
20 *es|ique*
20 *ée|ation*
20 *e|ale*
19 *e|al*
18 *t|ce*
18 *s|ne*
18 *ϵ|nes*
18 *me|ïde*
18 *f|ves*
17 *x|ses*
17 *e|iques*
16 *ose|euse*
16 *o|aires*
16 *es|aires*
13 *es|ite*
13 *al|um*
12 *s|le*
12 *fs|ve*
12 *e|és*
11 *o|um*
10 *ϵ|pathie*
9 *s|ux*
9 *ϵ|nel*
9 *me|blastique*
9 *fs|on*
8 *us|ale*
8 *on|ves*
8 *o|iques*
8 *ϵ|ique*
8 *fs|ves*
8 *e|iens*
7 *ϵ|x*
7 *ϵ|uses*
7 *ϵ|pathies*
7 *es|ité*
7 *en|te*
7 *e|iennes*
6 *o|us*
6 *ite|ales*
6 *ϵ|ite*
6 *ïde|blastique*
6 *ϵ|ée*
5 *x|gé*
5 *ses|tique*
5 *ite|opathies*

A.3.2 Suffixes (fréquence $\geq$ 5)

Relations hiérarchiques de la SNOMED (131 suffixes sur 464)

1248 *-s*
1099 *-e*
346 *-o*
332 *-es*
203 *-aire*
167 *-me*
167 *-ique*
164 *-se*
155 *-é*
149 *-ite*
144 *-ose*
127 *-que*
118 *-ome*
116 *-al*
113 *-ux*
106 *-ale*
98 *-x*
83 *-ation*
80 *-f*
78 *-le*
69 *-ïde*
62 *-on*
59 *-sarcome*
57 *-ve*

50 *-blastome*
47 *-um*
46 *-aires*
43 *-l*
40 *-ienne*
37 *-eux*
37 *-ée*
36 *-ne*
35 *-te*
34 *-ment*
34 *-cyte*
34 *-blaste*
33 *-ien*
31 *-oïde*
31 *-iques*
31 *-atose*
30 *-tique*
30 *-nes*
29 *-fibrome*
29 *-cytaire*
29 *-ales*
28 *-us*
27 *-euse*
24 *-opathie*
22 *-ité*
21 *-tion*
21 *-ques*
21 *-n*
20 *-pathie*
19 *-use*
19 *-t*
19 *-ce*
18 *-ves*
18 *-ine*
18 *-ie*
18 *-ement*
17 *-ses*
17 *-ateux*
16 *-les*
15 *-ure*
15 *-en*
15 *-aux*
14 *-urie*
14 *-sclérose*
14 *-nt*
14 *-ire*
14 *-ateuse*
13 *-matose*
13 *-mateuse*
13 *-lipome*
12 *-ion*
11 *-émie*
11 *-cytome*
11 *-carcinome*
11 *-ant*
10 *-plasique*
10 *-plasie*
10 *-ologique*
10 *-isation*
10 *-génique*
10 *-blastique*
9 *-nelle*
9 *-nel*
9 *-is*
9 *-endothéliome*
9 *-cellulaire*
8 *-kératose*
8 *-ineuse*
8 *-adénome*
8 *-a*
7 *-tiques*
7 *-sme*
7 *-pénie*
7 *-oire*
7 *-mateux*
7 *-iennes*
7 *-cytose*
6 *-vasculaire*
6 *-poïétique*
6 *-euses*
6 *-étique*
6 *-ène*
6 *-ème*
6 *-ées*
6 *-atique*
5 *-yome*
5 *-virus*
5 *-ures*
5 *-ule*
5 *-ulaire*
5 *-ul*
5 *-scléreux*
5 *-oplasique*
5 *-oplasie*
5 *-oires*
5 *-nement*
5 *-né*
5 *-myome*
5 *-iome*
5 *-inome*
5 *-illaire*
5 *-ies*
5 *-gé*
5 *-énique*
5 *-émateux*
5 *-cellularité*
5 *-aque*

Relations transversales de la SNOMED (113 suffixes sur 317)

1236 *-s*
1097 *-e*
338 *-es*
330 *-o*
178 *-ique*
177 *-é*
169 *-aire*
129 *-se*
127 *-ite*
121 *-ose*
115 *-que*
114 *-ux*
114 *-al*
94 *-x*
91 *-ale*
88 *-ome*
80 *-f*
80 *-ation*
75 *-le*
64 *-on*
57 *-ve*
55 *-eux*
51 *-l*
50 *-te*
47 *-um*
45 *-euse*
44 *-me*
42 *-ée*
42 *-aires*
41 *-tique*
36 *-tion*
35 *-ien*
34 *-ment*
33 *-ne*
33 *-ienne*
31 *-us*
31 *-atose*
29 *-ant*
27 *-ques*
26 *-oïde*
26 *-ïde*
24 *-ire*
22 *-ses*
21 *-ante*
20 *-ie*
20 *-ateux*
19 *-use*
18 *-ves*
18 *-nes*
18 *-ement*
16 *-aux*
15 *-n*
15 *-ine*
14 *-urie*
14 *-ité*
13 *-nt*
13 *-en*
12 *-opathie*
12 *-matose*
12 *-ion*
12 *-ateuse*
11 *-nte*
10 *-plasie*
10 *-pathie*
10 *-olyse*
10 *-cèle*
9 *-sé*
9 *-nelle*
9 *-nel*
9 *-is*
9 *-fs*
9 *-a*
8 *-sclérose*
8 *-iennes*
8 *-ectomie*
8 *-cytes*
7 *-sme*
7 *-oire*
7 *-ocèle*
7 *-nne*
7 *-ico*
6 *-plasique*
6 *-otique*
6 *-oblastome*
6 *-mateux*
6 *-lyse*
6 *-énique*
6 *-ène*
6 *-cytaire*
6 *-atique*
5 *-virus*
5 *-ulaire*
5 *-trophie*
5 *-stomie*
5 *-sant*
5 *-ris*
5 *-osclérose*
5 *-osante*
5 *-osa*
5 *-oires*
5 *-ocytome*
5 *-lithiase*
5 *-iase*
5 *-gé*
5 *-ètre*
5 *-ère*
5 *-er*
5 *-ème*
5 *-émateux*
5 *-ctomie*
5 *-co*
5 *-aque*
5 *-ae*

Relations hiérarchiques de la CIM10 (61 suffixes sur 240)

1303 *-s*
764 *-e*
313 *-es*
257 *-o*
141 *-ux*
132 *-aire*
108 *-ite*
102 *-ique*
93 *-que*
87 *-x*
86 *-le*
81 *-al*
76 *-se*
75 *-ation*
72 *-on*
68 *-ale*
59 *-aires*
55 *-é*
53 *-ne*
46 *-f*
45 *-l*
35 *-ve*
35 *-tique*
34 *-ves*
33 *-ée*
32 *-te*
29 *-les*
29 *-fs*
27 *-me*
26 *-iques*
25 *-ïde*
24 *-um*
23 *-ses*
23 *-ienne*
22 *-ose*
18 *-t*
18 *-nes*
18 *-euse*
18 *-ce*
17 *-us*
15 *-és*
15 *-blastique*
11 *-uses*
10 *-pathie*
10 *-ion*
10 *-iennes*
10 *-ales*
9 *-nel*
8 *-pathies*
8 *-opathies*
8 *-iens*
8 *-en*
7 *-ques*
7 *-ité*
6 *-ires*
6 *-ire*
6 *-aux*
5 *-omyélite*
5 *-nnes*
5 *-ites*
5 *-gé*

A.3.3 Préfixes (fréquence ≥ 5)

Relations hiérarchiques de la SNOMED (121 préfixes sur 2 781)

1248 *-s*
1099 *-e*
346 *-o*
332 *-es*
203 *-aire*
167 *-me*
167 *-ique*
164 *-se*
155 *-é*
149 *-ite*
144 *-ose*
127 *-que*
118 *-ome*
116 *-al*
113 *-ux*
106 *-ale*
98 *-x*
83 *-ation*
80 *-f*
78 *-le*
69 *-ïde*
62 *-on*
59 *-sarcome*
57 *-ve*
50 *-blastome*
47 *-um*
46 *-aires*
43 *-l*
40 *-ienne*
37 *-eux*
37 *-ée*
36 *-ne*
35 *-te*
34 *-ment*
34 *-cyte*
34 *-blaste*
33 *-ien*
31 *-oïde*
31 *-iques*
31 *-atose*
30 *-tique*
30 *-nes*
29 *-fibrome*
29 *-cytaire*
29 *-ales*
28 *-us*
27 *-euse*
24 *-opathie*
22 *-ité*
21 *-tion*
21 *-ques*
21 *-n*
20 *-pathie*
19 *-use*
19 *-t*
19 *-ce*
18 *-ves*
18 *-ine*
18 *-ie*
18 *-ement*
17 *-ses*
17 *-ateux*
16 *-les*
15 *-ure*
15 *-en*
15 *-aux*
14 *-urie*
14 *-sclérose*
14 *-nt*
14 *-ire*
14 *-ateuse*
13 *-matose*
13 *-mateuse*
13 *-lipome*
12 *-ion*
11 *-émie*
11 *-cytome*
11 *-carcinome*
11 *-ant*
10 *-plasique*
10 *-plasie*
10 *-ologique*
10 *-isation*
10 *-génique*
10 *-blastique*
9 *-nelle*
9 *-nel*
9 *-is*
9 *-endothéliome*
9 *-cellulaire*
8 *-kératose*
8 *-ineuse*
8 *-adénome*
8 *-a*
7 *-tiques*
7 *-sme*
7 *-pénie*
7 *-oire*
7 *-mateux*
7 *-iennes*
7 *-cytose*
6 *-vasculaire*
6 *-poïétique*
6 *-euses*
6 *-étique*
6 *-ène*
6 *-ème*
6 *-ées*
6 *-atique*
5 *-yome*
5 *-virus*
5 *-ures*
5 *-ule*
5 *-ulaire*
5 *-ul*
5 *-scléreux*
5 *-oplasique*
5 *-oplasie*
5 *-oires*
5 *-nement*
5 *-né*
5 *-myome*
5 *-iome*
5 *-inome*
5 *-illaire*
5 *-ies*
5 *-gé*
5 *-énique*
5 *-émateux*
5 *-cellularité*
5 *-aque*

Relations transversales de la SNOMED (101 préfixes sur 2 671)

1236 *-s*
1097 *-e*
338 *-es*
330 *-o*
178 *-ique*
177 *-é*
169 *-aire*
129 *-se*
127 *-ite*
121 *-ose*
115 *-que*
114 *-ux*
114 *-al*
94 *-x*
91 *-ale*
88 *-ome*
80 *-f*
80 *-ation*
75 *-le*
64 *-on*
57 *-ve*
55 *-eux*
51 *-l*
50 *-te*
47 *-um*
45 *-euse*
44 *-me*
42 *-ée*
42 *-aires*
41 *-tique*
36 *-tion*
35 *-ien*
34 *-ment*
33 *-ne*
33 *-ienne*
31 *-us*
31 *-atose*
29 *-ant*
27 *-ques*
26 *-oïde*
26 *-ïde*
24 *-ire*
22 *-ses*
21 *-ante*
20 *-ie*
20 *-ateux*
19 *-use*
18 *-ves*
18 *-nes*
18 *-ement*
16 *-aux*
15 *-n*
15 *-ine*
14 *-urie*
14 *-ité*
13 *-nt*
13 *-en*
12 *-opathie*
12 *-matose*
12 *-ion*
12 *-ateuse*
11 *-nte*
10 *-plasie*
10 *-pathie*
10 *-olyse*
10 *-cèle*
9 *-sé*
9 *-nelle*
9 *-nel*
9 *-is*
9 *-fs*
9 *-a*
8 *-sclérose*
8 *-iennes*
8 *-ectomie*
8 *-cytes*
7 *-sme*
7 *-oire*
7 *-ocèle*
7 *-nne*
7 *-ico*
6 *-plasique*
6 *-otique*
6 *-oblastome*
6 *-mateux*
6 *-lyse*
6 *-énique*
6 *-ène*
6 *-cytaire*
6 *-atique*
5 *-virus*
5 *-ulaire*
5 *-trophie*
5 *-stomie*
5 *-sant*
5 *-ris*
5 *-osclérose*
5 *-osante*
5 *-osa*
5 *-oires*
5 *-ocytome*
5 *-lithiase*
5 *-iase*
5 *-gé*
5 *-ètre*
5 *-ère*
5 *-er*
5 *-ème*
5 *-émateux*
5 *-ctomie*
5 *-co*
5 *-aque*
5 *-ae*

Relations hiérarchiques de la CIM10 (80 préfixes sur 2 301)

1303 *-s*
764 *-e*
313 *-es*
257 *-o*
141 *-ux*
132 *-aire*
108 *-ite*
102 *-ique*
93 *-que*
87 *-x*
86 *-le*
81 *-al*
76 *-se*
75 *-ation*
72 *-on*
68 *-ale*
59 *-aires*
55 *-é*
53 *-ne*
46 *-f*
45 *-l*
35 *-ve*
35 *-tique*
34 *-ves*
33 *-ée*
32 *-te*
29 *-les*
29 *-fs*
27 *-me*
26 *-iques*
25 *-ïde*
24 *-um*
23 *-ses*
23 *-ienne*
22 *-ose*
18 *-t*
18 *-nes*
18 *-euse*
18 *-ce*
17 *-us*
15 *-és*
15 *-blastique*
11 *-uses*
10 *-pathie*
10 *-ion*
10 *-iennes*
10 *-ales*
9 *-nel*
8 *-pathies*
8 *-opathies*
8 *-iens*
8 *-en*
7 *-ques*
7 *-ité*
6 *-ires*
6 *-ire*
6 *-aux*
5 *-omyélite*
5 *-nnes*
5 *-ites*
5 *-gé*

A.3.4 Familles

Relations hiérarchiques de la SNOMED (201 familles sur 1 606)

Flavimonas, *Flavobacterium*
absence, *absent*
achromie, *achromique*
actinomycose, *actinomycosique*
adolescence, *adolescent*
agglutination, *agglutiné*
alcool, *alcools*
amibe, *amibien*, *amibienne*
amylase, *amylose*, *amyloses*, *amyloïde*, *amyloïdose*
anesthésie, *anesthésique*, *anesthésiques*
annulaire, *annulaires*, *annulé*
antirhumatismal, *antirhumatismaux*
anévrismal, *anévrisme*, *anévrismes*, *anévrysme*
appendice, *appendicite*, *appendico*
arthrite, *arthrites*, *arthropathie*, *arthropathies*, *arthropathique*, *arthrose*, *arthroses*
ascendant, *ascendante*
athéromateuse, *athéromateux*, *athéromatose*, *athérome*, *athéroscléreuse*, *athéroscléreux*, *athérosclérose*
aucun, *aucune*
avitaminose, *avitaminoses*
bandelette, *bandelettes*
bilirubine, *bilirubinique*
blastes, *blastique*, *blastome*
brachial, *brachiale*, *brachio*
bruit, *bruits*
caecal, *caeco*, *caecum*
capillaire, *capillaires*
carie, *caries*
causant, *cause*, *causes*, *causé*, *causée*, *causées*
champignon, *champignons*
chlore, *chlorome*, *chlorure*, *chlorures*
cholélithiase, *cholélithiases*
chéloïde, *chéloïdienne*
clinique, *cliniquement*, *clinoïde*
collagène, *collagènes*, *collagénique*
compensation, *compensé*
complet, *complets*, *composite*, *composé*, *composée*, *composés*
conjonctif, *conjonctival*, *conjonctivale*, *conjonctive*, *conjonctivite*, *conjonctivites*
consécutif, *consécutifs*, *consécutive*, *consécutives*
cordage, *cordages*
corrosif, *corrosifs*, *corrosion*, *corrosions*, *corrosive*, *corrosives*
court, *courtes*, *courts*
crural, *crurale*
cyanose, *cyanosé*
cémentant, *cémentante*
demande, *demandé*
descendant, *descendante*
diamniotique, *diamniotiques*
digestif, *digestifs*, *digestion*, *digestive*, *digestives*
disque, *disques*
donne, *donné*
duodénal, *duodénale*, *duodénite*, *duodéno*, *duodénum*
dysérythropoïèse, *dysérythropoïétique*
dédoublement, *dédoublé*
déhiscence, *déhiscent*
dépressif, *dépressifs*, *dépression*
déviation, *dévié*
embole, *embolie*, *embolies*, *embolique*
encéphale, *encéphalique*, *encéphalite*, *encéphalites*, *encéphalitique*, *encéphalomyélite*, *encéphalomyélites*, *encéphalopathie*
engorgement, *engorgé*
environnement, *environnemental*
examen, *examens*
exocrine, *exocrines*
extensif, *extension*, *externe*, *externes*
face, *faces*
fantôme, *fantômes*
fibre, *fibres*, *fibreuse*, *fibreuses*, *fibreux*, *fibrillaire*, *fibrillaires*, *fibrine*, *fibrineuse*, *fibrineux*, *fibrinoïde*, *fibro*, *fibroblaste*, *fibroblastique*, *fibroblastiques*, *fibrochondrosarcome*, *fibrocyte*, *fibrokystique*, *fibrolipome*, *fibroliposarcome*, *fibromateux*, *fibromatose*, *fibrome*, *fibromyome*, *fibromyxolipome*, *fibromyxome*, *fibromyxosarcome*, *fibroplasie*, *fibroplasique*, *fibrosante*, *fibrosarcome*, *fibrosclérose*, *fibrosclérosé*, *fibrose*, *fibrovasculaire*, *fibroxanthome*, *fibroïde*
focal, *focale*, *focales*
formation, *formations*
fragmentation, *fragmenté*
fumée, *fumées*
fémoral, *fémorale*, *fémoro*
gangreneuse, *gangreneux*, *gangrène*
gestation, *gestationnel*, *gestationnelle*, *gestationnels*
glomérulaire, *glomérulaires*, *glomérule*, *glomérulo*, *glomérulonéphrite*, *glomérulopathie*, *glomérulée*
gonflement, *gonflé*
greffe, *greffes*, *greffon*, *greffé*, *greffés*
gémistocyte, *gémistocytique*
haut, *haute*
histologique, *histologiques*
hyalin, *hyaline*, *hyalines*, *hyalinisation*, *hyalinisé*, *hyalinose*, *hyalins*
hyperchromatique, *hyperchrome*, *hyperchromie*
hypersegmentation, *hypersegmenté*
hypopharyngé, *hypopharynx*
hypothermie, *hypothermique*
hémolyse, *hémolyses*, *hémolytique*, *hémolytiques*
héréditaire, *héréditaires*
iléal, *iléales*
implantation, *implantée*
inclus, *incluses*
indéterminé, *indéterminée*
inférieur, *inférieure*, *inférieures*, *inférieurs*
instrument, *instrumental*
interlobulaire, *interlobulaires*
intoxication, *intoxications*
intrinsèque, *intrinsèques*
ischio, *ischion*
jour, *jours*
kératine, *kératinique*, *kératinisant*, *kératinisation*, *kératinisée*, *kératinocyte*, *kérato*, *kératome*, *kératose*, *kératosique*, *kératosiques*
lame, *lames*
leptocyte, *leptocytose*
limite, *limité*, *limitée*, *limitées*, *limités*
liquide, *liquides*, *liquidienne*

local, *locale*, *localisation*, *localisations*, *localisé*, *localisée*, *localisées*, *localisés*, *locaux*
luxation, *luxations*
légère, *légères*
macula, *maculaire*, *macule*, *maculo*
maligne, *malignes*, *malignité*, *malin*
manipulation, *manipulations*
massif, *massive*
membre, *membres*
micro, *microcytaire*, *microcyte*, *microcytose*, *microcytémie*
mixte, *mixtes*
mort, *morts*
multifocal, *multifocale*
myocarde, *myocardique*, *myocardite*, *myocardites*
myéline, *myélinique*
mécanique, *mécanisme*
médulloépithéliale, *médulloépithéliome*
ménopause, *ménopausique*, *ménopausiques*, *ménorragie*
métaphysaire, *métaphyse*
nasal, *nasale*, *nasales*
neutrophile, *neutrophiles*, *neutrophilie*, *neutrophilique*
nombre, *nombres*
nutrition, *nutritionnel*, *nutritionnelle*, *nutritionnelles*, *nutritionnels*
névrite, *névritique*, *névrotique*, *névrotiques*
occlusif, *occlusion*, *occlusions*, *occlusive*
oligodendroblastome, *oligodendrocytaire*, *oligodendrocyte*, *oligodendrocytome*
opaque, *opaques*
orbitaire, *orbitale*, *orbite*
orteil, *orteils*
ostéite, *ostéo*, *ostéoblaste*, *ostéoblastique*, *ostéoblastome*, *ostéocartilagineuse*, *ostéochondromatose*, *ostéochondrome*, *ostéochondrosarcome*, *ostéoclaste*, *ostéoclastique*, *ostéoclastome*, *ostéocyte*, *ostéofibrome*, *ostéofibrosarcome*, *ostéogénique*, *ostéome*, *ostéomyélite*, *ostéomyélites*, *ostéopénie*, *ostéosarcome*, *ostéose*, *ostéoïde*
oxydant, *oxyde*, *oxydes*
pansement, *pansements*
paranasal, *paranasaux*
partiel, *partielle*, *partiellement*, *partielles*
pelvi, *pelvien*, *pelvienne*, *pelviennes*, *pelviens*, *pelvipéritonite*, *pelvis*
perte, *pertes*
phobique, *phobiques*
pigment, *pigmentaire*, *pigmentation*, *pigmenté*, *pigmentée*
plaquettaire, *plaquette*, *plaquettes*
plume, *plumeuse*
polychromatophile, *polychromatophilie*
ponction, *ponctionné*
positif, *positifs*, *position*, *positions*, *positive*, *positives*
poussière, *poussières*
probable, *probablement*
progressif, *progression*, *progressive*
prothèse, *prothèses*, *prothétique*
préauriculaire, *préauriculaires*
prélèvement, *prélèvements*
préservatif, *préservation*
pubien, *pubienne*, *pubis*
pustulaire, *pustule*, *pustuleuse*, *pustuleux*, *pustulo*
pénicilline, *pénicillines*
période, *périodes*, *périodique*, *périodiquement*, *périodiques*
pétéchial, *pétéchiale*, *pétéchie*
radial, *radiale*, *radiation*, *radio*, *radioactivité*, *radiologique*, *radiologiques*, *radique*, *radium*, *radius*, *radiée*
recanalisation, *recanalisé*
respiration, *respiratoire*, *respiratoires*
ribonucléoside, *ribonucléotide*
réactif, *réaction*, *réactionnel*, *réactionnelle*, *réactionnelles*, *réactions*, *réactive*, *réactivité*
réparation, *réparé*
rétropéritoine, *rétropéritonéal*, *rétropéritonéale*
sarcomateuse, *sarcomateux*, *sarcomatose*, *sarcomatoïde*, *sarcome*, *sarcomes*, *sarcoïde*, *sarcoïdose*, *sarcoïdosique*
second, *secondaire*, *secondaires*
sept, *septale*, *septum*
significatif, *signification*
social, *sociale*, *sociales*, *sociaux*
soudure, *soudures*, *soudé*, *soudés*
splénique, *spléno*
spécifique, *spécifiquement*, *spécifiques*
stercoral, *stercorale*
structurel, *structurels*
subluxation, *subluxations*
superinvolutif, *superinvolution*
suture, *sutures*, *suturé*
synovite, *synovites*
sélectif, *sélective*
séropositif, *séropositive*, *séropositives*
température, *températures*
thymine, *thymique*, *thymome*, *thymus*
tissulaire, *tissulaires*
trabéculaire, *trabécule*
transmis, *transmise*, *transmises*
tricuspide, *tricuspidienne*
trouvé, *trouvée*, *trouvées*, *trouvés*
tympan, *tympanique*, *tympanite*
ulcère, *ulcères*
urothélial, *urothéliales*, *urothélium*
vaisseau, *vaisseaux*
ventriculaire, *ventriculaires*, *ventricule*, *ventricules*, *ventriculo*
vestige, *vestiges*, *vestigielle*
vitamine, *vitamines*, *vitamino*
vomies, *vomissure*, *vomissures*
zygomatique, *zygomato*
élastique, *élastiques*, *élastofibrome*, *élastome*, *élastose*
éosinophile, *éosinophiles*, *éosinophilie*, *éosinophilique*
épididymaire, *épididyme*, *épididymes*, *épididymite*, *épididymo*
épisode, *épisodes*, *épisodique*
éraflure, *éraflé*
étendu, *étendue*

Relations transversales de la SNOMED (204 familles sur 1 625)

abdomen, *abdominal*, *abdominale*, *abdominales*, *abdominaux*, *abdomino*
accidentelle, *accidentellement*, *accidentelles*
acquis, *acquise*, *acquises*
adhérence, *adhérences*, *adhérent*
affaissement, *affaissé*
aigu, *aigus*, *aiguë*, *aiguës*
altération, *altérations*, *altéré*
amnésie, *amnésique*
anal, *anale*, *anales*
angiomateux, *angiomatose*, *angiome*, *angiosclérose*
anomalie, *anomalies*
antécédent, *antécédents*
apnée, *apnées*
argyrie, *argyrose*
artère, *artères*
astrocytaire, *astrocyte*, *astrocytes*, *astrocytome*
atteinte, *atteintes*
auto, *autolyse*, *autolytique*
bacille, *bacilles*
bilatéral, *bilatérale*, *bilatérales*, *bilatéraux*
bipolaire, *bipolaires*
bourse, *bourses*
bruit, *bruits*
bénigne, *bénignes*, *bénignité*
canaliculaire, *canalicules*
cardiopulmonaire, *cardiopulmonaires*
caséeuse, *caséeux*
certain, *certaines*, *certains*
chimique, *chimiques*
cholédocho, *cholédocholithiase*, *cholédocien*, *cholédoque*
chronique, *chroniques*
classe, *classes*, *classique*, *classé*, *classée*, *classées*, *classés*
cognitif, *cognitives*
commun, *commune*, *communs*
comportement, *comportements*
confluent, *confluents*
constriction, *constrictive*
contrôle, *contrôlé*
corporel, *corporelle*
couche, *couches*
cristallin, *cristalline*, *cristallines*
cuisse, *cuisses*
cystique, *cystite*, *cystites*, *cystome*
céroïde, *céroïdes*, *céroïdo*
dense, *densité*
desquamation, *desquamative*
difficile, *difficiles*
discriminant, *discrimination*
distal, *distale*, *distales*
doré, *dorés*
durable, *durables*
décidual, *déciduale*, *déciduaux*
défectueuse, *défectueux*
démyélinisante, *démyélinisantes*, *démyélinisation*
déshydratation, *déshydraté*, *déshydratée*
ectopie, *ectopique*, *ectopiques*
encapsulé, *encapsulée*
endoste, *endostose*
entorse, *entorses*
essence, *essences*
excessive, *excessives*
explicatif, *explication*, *explications*
extrémité, *extrémités*
faible, *faibles*
fente, *fentes*
filaire, *filarienne*
foetal, *foetale*, *foetales*, *foeto*, *foetus*
fort, *forte*, *fortement*
froid, *froide*, *froids*
fébrile, *fébriles*
ganglioneuromatose, *ganglioneurome*
genou, *genoux*
globule, *globules*
goitre, *goitres*
grand, *grande*, *grandes*
guéri, *guérie*
hallucinogène, *hallucinogènes*
histiocytaire, *histiocytaires*, *histiocyte*, *histiocytes*, *histiocytome*, *histiocytose*, *histiocytoïde*
humain, *humaine*, *humains*
hydropisie, *hydrops*, *hydrosalpinx*
hyperphénylalaninémie, *hyperphénylalaninémies*
hypokaliémie, *hypokaliémique*
hypothalamique, *hypothalamus*
hématie, *hématies*, *hématique*, *hématite*, *hémato*, *hématome*
héréditaire, *héréditaires*
iléal, *iléales*
immun, *immune*, *immunes*, *immunité*, *immuno*, *immunoblastique*, *immunocyte*, *immunocytome*, *immuns*
incarcération, *incarcéré*
indépendant, *indépendantes*, *indépendants*
infestation, *infestations*
insecte, *insectes*
intention, *intentionnelle*
interstitiel, *interstitielle*, *interstitielles*
intrahépatique, *intrahépatiques*
invasif, *invasive*
isolé, *isolée*, *isolées*, *isolés*
jumeau, *jumeaux*
lactate, *lactose*
laryngite, *laryngo*, *laryngé*, *laryngée*, *laryngées*, *larynx*
leuco, *leucokératose*, *leucoplasie*
limitante, *limite*, *limité*, *limitée*, *limitées*, *limités*
lipoprotéine, *lipoprotéines*
liénal, *liéno*
long, *longs*
lymphadénome, *lymphangite*, *lymphatique*, *lymphatiques*, *lymphe*, *lympho*, *lymphoblaste*, *lymphoblastique*, *lymphocytaire*, *lymphocyte*, *lymphocytes*, *lymphocytique*, *lymphocytome*, *lymphocytopathie*, *lymphocytose*, *lymphocèle*, *lympholyse*, *lymphomateuse*, *lymphomateux*, *lymphome*, *lymphomes*, *lymphorragie*, *lymphorrhée*, *lymphoïde*, *lymphoïdes*
lépromateuse, *lépromateux*
main, *mains*
malocclusion, *malocclusions*
marge, *marges*
maternel, *maternelle*, *maternelles*, *maternels*
mental, *mentale*, *mentales*, *mentaux*
migraine, *migraines*, *migraineux*
modification, *modifications*
mort, *morts*
mullerien, *mullerienne*, *mulleroblastome*
mycobactérie, *mycobactérienne*, *mycobactéries*
myoépithélial, *myoépithéliale*, *myoépithéliales*
myéline, *myélinique*, *myélinolyse*, *myélite*, *myélites*, *myélocytaire*, *myélocyte*, *myélocèle*, *myélomateux*, *myélomatose*, *myélome*, *myélosclérose*, *myélose*, *myéloïde*, *myéloïdes*
médical, *médicale*, *médicales*, *médicaux*, *médico*

méningocoque, *méningocoques*
mésothélial, *mésothéliale*, *mésothélium*
méthyl, *méthylique*
neurinomatose, *neurinome*
nocif, *nocifs*, *nocive*, *nocives*
noyau, *noyaux*
négligence, *négligences*
observation, *observations*
oculaire, *oculaires*, *oculo*
omphalite, *omphalo*
opioïde, *opioïdes*
organisation, *organisations*
osmiophile, *osmiophilie*
ostéomyélite, *ostéomyélites*
oxalate, *oxalates*, *oxalose*
pancréas, *pancréatico*, *pancréatique*, *pancréatiques*, *pancréatite*, *pancréatites*, *pancréato*
paralytique, *paralytiques*
parotide, *parotidien*, *parotidite*
pelvi, *pelvien*, *pelvienne*, *pelviennes*, *pelviens*, *pelvipéritonite*, *pelvis*
pesticide, *pesticides*
phobique, *phobiques*
pilier, *piliers*
plasmatique, *plasmatiques*
pneumocoque, *pneumocoques*
polydipsie, *polydipsique*
pont, *ponto*
postural, *posturale*
primaire, *primaires*
production, *productive*
prolongement, *prolongé*, *prolongée*, *prolongées*
protéine, *protéines*, *protéino*, *protéinose*, *protéinurie*
précérébrale, *précérébrales*, *précérébraux*
prénatal, *prénatals*
psoriasique, *psoriasiques*
pupillaire, *pupillaires*, *pupille*
pyramidal, *pyramidale*
péricarde, *péricardique*, *péricardite*
périphérique, *périphériques*
pétéchial, *pétéchiale*, *pétéchie*
radiologique, *radiologiques*
ratatinement, *ratatiné*
relâchement, *relâché*
rhumatismal, *rhumatismale*, *rhumatismales*, *rhumatisme*
rupture, *ruptures*, *rupturé*
régénération, *régénéré*
rétropéritoine, *rétropéritonéal*, *rétropéritonéale*
sang, *sanguinolent*
scléreuse, *scléreux*, *sclérosant*, *sclérosante*, *sclérosants*, *sclérose*, *sclérosé*, *sclérotique*
semaine, *semaines*
sigmoïde, *sigmoïdectomie*, *sigmoïdes*, *sigmoïdienne*, *sigmoïdite*
siège, *sièges*
soudure, *soudures*
spiradénome, *spiroïde*
spécial, *spéciale*, *spéciales*, *spéciaux*
stercoral, *stercorale*
structure, *structurel*, *structurels*, *structures*
substance, *substances*
suppurée, *suppurées*
symbiose, *symbiotique*
syphilis, *syphilitique*
sénile, *séniles*
tacheté, *tachetées*
terminal, *terminale*, *terminalis*, *terminaux*, *termino*
thrombo, *thrombophlébite*, *thrombosant*, *thrombose*, *thromboses*, *thrombosé*, *thrombosées*, *thrombotique*, *thrombus*
tique, *tiques*
trabéculaire, *trabécule*
transmissible, *transmissibles*
tricuspide, *tricuspidienne*
trouble, *troubles*
tylome, *tylose*
ulcératif, *ulcération*, *ulcéreuse*, *ulcéreux*, *ulcéro*, *ulcéré*, *ulcérée*, *ulcérées*
urticaire, *urticarienne*
vaisseau, *vaisseaux*
veinule, *veinules*
vertèbre, *vertèbres*
villeuse, *villeux*, *villo*
voie, *voies*
vénérien, *vénérienne*, *vénériens*
éclatement, *éclaté*
émanation, *émanations*
épicanthique, *épicanthus*
épiphysaire, *épiphysaires*, *épiphyse*
équilibre, *équilibré*, *équilibrée*, *équilibrés*
érythème, *érythémateuse*, *érythémateux*
évoquant, *évoqué*

Relations hiérarchiques de la CIM10 (190 famille sur 1 520)

Ankylostomiase, *Ankylostomose*
Glomérulopathie, *Glomérulopathies*
Ostéonécrose, *Ostéopathie*
absence, *absent*
achromie, *achromique*
actif, *action*, *active*, *actives*, *activité*, *activités*
adénocarcinome, *adénocarcinoïde*
agent, *agents*
algie, *algies*
amnios, *amniotique*, *amniotiques*, *amniotite*
ancien, *ancienne*
annexiel, *annexielle*
antérieur, *antérieure*, *antérieurs*
apocrine, *apocrines*
argentaffine, *argentaffines*
artérioscléreuse, *artérioscléreux*, *artériosclérose*, *artériosclérotique*
asymétrie, *asymétrique*
atypie, *atypique*, *atypiques*
auxiliaire, *auxiliaires*
balanique, *balanite*
bilirubine, *bilirubinique*
blastes, *blastique*
branche, *branches*
buccal, *buccale*, *buccaux*, *bucco*
calcifiant, *calcifiante*, *calcification*, *calcifications*
canal, *canaux*
carcino, *carcinome*, *carcinomes*, *carcinoïde*

cataracte, *cataractes*
certain, *certaines*, *certains*
chaîne, *chaînes*
cholinergique, *cholinergiques*
chorion, *chorionique*, *chorées*
cicatrisation, *cicatrisé*
cloisonnement, *cloisonnements*
colo, *colonne*, *colonnes*
comportementales, *comportementaux*
congestion, *congestive*
constriction, *constrictive*
cordage, *cordages*
corpuscule, *corpuscules*
coup, *coupe*, *coups*
croûte, *croûteux*
cyanure, *cyanures*
cytoplasme, *cytoplasmique*, *cytoplasmiques*
d'algie, *d'algies*
dermatophytose, *dermatophytoses*
dialyse, *dialyses*, *dialysés*
dilatation, *dilaté*
disséquant, *disséquante*
douloureuse, *douloureux*
dysfonction, *dysfonctionnel*, *dysfonctionnement*, *dysfonctionnements*, *dysfonctions*
décoloration, *décoloré*
défini, *définie*, *définies*, *définis*
dépendance, *dépendant*, *dépendante*
désir, *désirée*
effet, *effets*
encéphale, *encéphalique*, *encéphalite*, *encéphalites*, *encéphalitique*, *encéphalomyélite*, *encéphalomyélites*
endothélial, *endothéliale*, *endothéliales*, *endothélium*
envahissant, *envahissante*, *envahissants*
exanthémateuse, *exanthémateux*
expectoration, *expectorations*
extrinsèque, *extrinsèques*
faible, *faibles*
fermier, *fermiers*
fissuré, *fissurée*
foetal, *foetale*, *foetales*, *foeto*, *foetoplacentaire*, *foetus*
fosse, *fosses*
fruit, *fruits*
fécondation, *fécondé*
gastrique, *gastriques*, *gastrite*, *gastrites*, *gastro*
glandulaire, *glandulaires*, *glandulo*
gonococcique, *gonococciques*
gravide, *gravidique*
géant, *géante*, *géantes*
halogéné, *halogénés*
histochimie, *histochimique*
hyalinisation, *hyalinisé*
hyperkinétique, *hyperkinétiques*
hypersegmentation, *hypersegmenté*
hypophysaire, *hypophyse*, *hypophyso*
hématogène, *hématogènes*
hémosidérine, *hémosidérinique*
idiopathique, *idiopathiques*
immédiat, *immédiate*, *immédiates*, *immédiats*
inconnu, *inconnue*, *inconnues*
infantile, *infantiles*
inguinal, *inguinale*, *inguinaux*
insuffisance, *insuffisances*, *insuffisant*, *insuffisante*, *insuffisantes*, *insuffisants*
intermédiaire, *intermédiaires*
intralobulaire, *intralobulaires*
inversion, *inversé*, *inversée*
jaune, *jaunes*
kyste, *kystes*, *kystique*, *kystiques*
lactase, *lactose*
laryngite, *laryngo*
leur, *leurs*
lingual, *linguale*, *linguales*, *linguo*
listérienne, *listériennes*, *listériose*
loisir, *loisirs*
lymphadénite, *lymphadénites*
légal, *légale*
macroglial, *macroglie*
malformatifs, *malformation*, *malformations*
manipulation, *manipulations*
mastoïde, *mastoïdectomie*, *mastoïdienne*, *mastoïdiens*, *mastoïdite*, *mastoïdites*
membre, *membres*
migraine, *migraines*, *migraineux*
modification, *modifications*
mort, *morts*
multifocal, *multifocale*
myocarde, *myocardique*, *myocardite*, *myocardites*
myélodysplasique, *myélodysplasiques*
médecin, *médecine*
mélanocytaire, *mélanocyte*, *mélanocytes*
mésenchymateuse, *mésenchymateux*
métaplasie, *métaplasique*
naturel, *naturelle*
nicotine, *nicotinique*
nourrisson, *nourrissons*
nécrose, *nécrotique*, *nécrotiques*
névrotique, *névrotiques*
occlusif, *occlusion*, *occlusions*, *occlusive*
omphalite, *omphalo*
opportuniste, *opportunistes*
organisation, *organisations*
ossifiant, *ossifiante*, *ossification*, *ossifications*
ouvert, *ouverte*, *ouvertes*
palpébral, *palpébrale*, *palpébraux*, *palpébro*
parakératose, *parakératosique*
parodonte, *parodontite*
peinture, *peintures*
personnelle, *personnels*
pharyngé, *pharyngée*, *pharyngées*, *pharynx*
pigment, *pigmentation*, *pigmenté*, *pigmentée*
plaque, *plaques*
pneumocoque, *pneumocoques*
polynucléaire, *polynucléaires*
portal, *portale*, *porte*, *portes*, *portion*
primaire, *primaires*
produit, *produits*
prolymphocytaire, *prolymphocyte*, *prolymphocytes*
protéinase, *protéinose*
précision, *précisé*, *précisée*, *précisées*, *précisés*, *précoce*, *précoces*
préménopause, *préménopausique*
psycho, *psycholeptiques*, *psychologique*, *psychologiques*, *psychose*, *psychosexuel*, *psychosociales*, *psychostimulants*, *psychotique*, *psychotiques*, *psychotrope*, *psychotropes*
purulent, *purulente*
pénien, *pénienne*
périoste, *périostique*, *périostite*
pétéchial, *pétéchiale*, *pétéchie*
radial, *radiale*, *radio*, *radium*, *radius*

recanalisation, *recanalisé*
rhabdomyoblastique, *rhabdomyome*
rompu, *rompue*
récepteur, *récepteurs*
résistance, *résistant*
saignement, *saignements*
sclérosant, *sclérosante*, *sclérosants*
sellaire, *selle*, *selles*
sidéroblaste, *sidéroblastes*, *sidéroblastique*, *sidéroblastiques*
siège, *sièges*
soufre, *soufrés*
spondylite, *spondylo*, *spondylopathies*
squame, *squameuse*, *squameuses*
streptococcique, *streptocoque*, *streptocoques*
subit, *subite*
sujet, *sujets*
surrénal, *surrénale*, *surrénales*, *surrénalienne*, *surrénaliennes*, *surrénaliens*
synapse, *synaptique*
séborrhéique, *séborrhéiques*
séropositif, *séropositive*, *séropositives*
tendineuse, *tendineuses*, *tendineux*, *tendon*, *tendons*
thrombosé, *thrombosées*
tissulaire, *tissulaires*
traitement, *traitements*
transport, *transports*
trompe, *trompes*
tumeur, *tumeurs*
ulcère, *ulcères*
uropathie, *uropathies*
vacuo, *vacuolaire*, *vacuole*
vasculaire, *vasculaires*, *vasculite*, *vasculo*, *vasculopathies*
vertèbre, *vertèbres*
villosité, *villosités*
voisine, *voisins*
vésical, *vésicale*, *vésicales*, *vésico*
élevé, *élevée*
épendymaire, *épendyme*
épileptique, *épileptiques*
équin, *équine*
établissement, *établissements*

A.4 Ajustement et maximisation des suffixes

A.4.1 Règles (fréquence $\geq$ 5 : 117 règles sur 501)

77 *e|ique*
71 *e|aire*
70 *ie|ique*
64 *é|ation*
45 *o|al*
43 *e|ose*
41 *e|o*
40 *e|é*
39 *o|ite*
39 *if|ion*
32 *e|ome*
31 *ome|osarcome*
31 *e|atose*
28 *e|eux*
27 *e|al*
26 *e|ien*
25 *o|aire*
24 *se|tique*
24 *ome|omateux*
24 *o|ique*
24 *eux|ose*
23 *ome|ose*
21 *é|er*
20 *ome|ique*
19 *é|ement*
18 *o|ome*
18 *ome|oïde*
18 *ome|oblastome*
18 *nt|nce*
18 *ant|ation*
17 *é|ant*
16 *ϵ|se*
16 *e|ie*
15 *ose|aire*
15 *al|um*
14 *ϵ|ment*
14 *e|oïde*
14 *ϵ|al*
13 *ose|oïde*
13 *ome|aire*
13 *o|é*
13 *ion|ionnel*
13 *ϵ|ement*
13 *ϵ|é*
12 *ose|ique*
12 *ite|opathie*
12 *e|ant*
12 *é|aire*
12 *al|ome*
11 *tion|toire*
11 *o|um*
11 *oïde|osarcome*
11 *o|eux*
11 *ϵ|n*
11 *ϵ|l*
11 *ite|ose*
11 *ien|ite*
11 *eux|aire*
11 *e|urie*
11 *é|ure*
10 *oïde|omateux*
10 *er|ation*
9 *ome|ofibrome*
9 *ome|oblastique*
9 *ϵ|o*
9 *ϵ|ique*
9 *é|ion*
8 *osarcome|oblastome*
8 *osarcome|oblastique*
8 *ome|olipome*
8 *ϵ|ome*
8 *omatose|oblastome*
8 *o|ien*
8 *ϵ|isé*
8 *ique|oïde*
8 *ϵ|ion*
8 *e|ité*
8 *e|émie*
8 *aire|ation*
8 *ϵ|aire*
7 *ose|opathie*
7 *on|onné*
7 *ome|ocarcinome*
7 *ocytome|oblastome*
7 *ique|isme*
7 *ifié|ification*
7 *gène|génique*
7 *eux|oïde*
7 *e|ocyte*
7 *e|atique*
7 *al|us*
7 *a|e*
6 *us|ique*
6 *ose|omatose*
6 *oïde|oblastique*
6 *i|issement*
6 *e|eur*
6 *ble|bilité*
6 *arynx|aryngé*
6 *a|o*
5 *ϵ|virus*
5 *ϵ|ure*
5 *qué|cation*
5 *ome|omyome*
5 *ome|ocytaire*
5 *ome|inique*
5 *oïde|ogène*
5 *ogène|osarcome*
5 *ϵ|ité*
5 *ϵ|isation*
5 *ϵ|eux*
5 *e|osarcome*
5 *ème|émateux*
5 *e|ier*
5 *e|iase*
5 *é|atif*
5 *aire|opathie*

A.4.2 Suffixes (fréquence $\geq$ 5 : 94 suffixes sur 582)

556 *-e*
288 *-ome*
268 *-o*
254 *-ique*
239 *-é*
189 *-aire*
179 *-ose*
120 *-al*
114 *-oïde*
103 *-ation*
97 *-eux*
86 *-ie*
81 *-osarcome*
73 *-ite*
72 *-ion*
58 *-ien*
57 *-oblastome*
50 *-ant*
41 *-um*
41 *-omateux*
41 *-if*
40 *-se*
38 *-er*
35 *-ement*
33 *-tique*
31 *-atose*
29 *-ocytaire*
29 *-oblastique*
26 *-opathie*
20 *-omatose*
19 *-us*
19 *-ure*
18 *-nt*
18 *-nce*
17 *-a*
15 *-ogène*
15 *-ocarcinome*
15 *-émie*
14 *-ment*
13 *-ofibrome*
13 *-ocyte*
13 *-l*
13 *-ité*
13 *-ionnel*
12 *-on*
11 *-urie*
11 *-toire*
11 *-tion*
11 *-n*
10 *-osclérose*
10 *-isé*
9 *-isme*
9 *-arynx*
8 *-olipome*
8 *-is*
7 *-onné*
7 *-ogreffe*
7 *-ocytome*
7 *-ifié*
7 *-ification*
7 *-ier*
7 *-iase*
7 *-génique*
7 *-gène*
7 *-atique*
6 *-sion*
6 *-roglie*
6 *-rocytaire*
6 *-ocellulaire*
6 *-issement*
6 *-ine*
6 *-iforme*
6 *-i*
6 *-eur*
6 *-ble*
6 *-bilité*
6 *-aryngé*
6 *-angioendothéliome*
5 *-virus*
5 *-u*
5 *-rogliome*
5 *-qué*
5 *-omyome*
5 *-ologue*
5 *-ologique*
5 *-odermie*
5 *-isation*
5 *-inique*
5 *-fibreux*
5 *-en*
5 *-ème*
5 *-émateux*
5 *-cation*
5 *-atif*

A.4.3 Préfixes (fréquence $\geq$ 5 : 54 préfixes sur 1 251)

32 *fibr-*
26 *myél-*
25 *lymph-*
22 *osté-*
21 *adén-*
16 *lip-*
16 *granul-*
16 *chondr-*
14 *mélan-*
14 *angi-*
13 *plasmocyt-*
12 *hépat-*
11 *lymphocyt-*
11 *histiocyt-*
11 *hémangi-*
9 *réticul-*
9 *radi-*
9 *papul-*
9 *myx-*
9 *kérat-*
8 *néphr-*
8 *glomérul-*
8 *épithéli-*
8 *carcin-*
7 *tubul-*
7 *protéin-*
7 *neur-*
7 *gli-*
7 *épendym-*
7 *derm-*
7 *acid-*
6 *ulcér-*
6 *tubercul-*
6 *pustul-*
6 *psych-*
6 *macul-*
6 *hémat-*
6 *endométri-*
5 *vulv-*
5 *segment-*
5 *placent-*
5 *pigment-*
5 *pharyng-*
5 *oesophag-*
5 *monocyt-*
5 *microcyt-*
5 *mésonéphr-*
5 *lipid-*
5 *léiomy-*
5 *érythr-*
5 *épiderm-*
5 *dent-*
5 *amygdal-*
5 *affect-*

Index

CIM, 19–21, 69, 85, 105
MeSH, 15–18
SNOMED, 18–19, 66, 69, 84, 85, 105, 118
UMLS, 21–23, 60, 68, 69

`Brill`, 55, 78, **78**
`Celex`, 28, 56, 60, 65
`Flemm`, 55, 79, 103
`MorTAL`, 56, 60, 65
`TreeTagger`, 55, 103
`lvg`, 22, 24, 68
`norm`, 22

aboutness, **17**, 137, 166
adjectivations concurrentes, 115–120
affixation, 11, 22, 25, **40**, 40–42, 60, 68, 72, 167
affixation isocatégorielle, **42**
affixation transcatégorielle, **41**
affixe, 40
allomorphie, 38–39, 56, 74, 107, 116, 168
ambiguïté, **54**, 73, 117
antonymie, **138**

base, 11, 25, 36, 37, **39**, 40, 51
base autonome, **36**
base non autonome, **37**
base supplétive, **37**, 39
baseline, **26**
bruit, **9**, 26, 103, 107

classification, **15**
co-hyponyme, **22**, 138, 143, 149
cognat, **94**
composant, 12
composition, 3, 12, 43–46, 60, 167
composition populaire, **43**
composition savante, **45**
concept, **14**, 66, 124
conversion, 10, 42–43, 166

dérivation, 3, **40**
désuffixation, 25, **41**, 61, 128, 130, 169
désuffixeur, **25**, 56

étiquetage morphosyntaxique, 30, 33, **54**, 78, 103, 166
explosion de concept, 17, **124**, 135, 137

famille morphologique, 47–48, 61–64, 68
flexion, 3, 60, 68, 72, 166

gain d'information, 121

hapax, 91, **97**
holonyme, **137**
hypenonyme, **137**
hyponyme, **137**
hyponymie, 17

interfixe, **45**
interopérabilité sémantique, 13

langue de spécialité, **60**
langue générale, **60**
lemmatisation, **54**, 78, 79, 103, 166
lemme, 25, **46**
lexème, **10**, 36

marqueur, **141**, 148
méronyme, **137**
méronymie, 17
morphologie constructionnelle, 10, 39–46
morphologie flexionnelle, 46
mot plein, **149**

mot-forme, 10
mots grammaticaux, 10
mots vides, **10**, 10

nomenclature, **14**

occurrence, **96**, 97, 104, 128
ordre des mots, 9

patron lexico-syntaxique, 29, 30, **142**, 148
précision, **27**, 68

racinisation, **41**
rappel, **27**, 68
règle d'association, 67, 105
règle de désuffixation, 25, 128, 130, 169
relation hiérarchique, **137**
relation partitive, **137**
relation transversale, **138**

sac de mots, **9**, 144
silence, **26**, 72, 107, 109
stem, **25**
stemming, **41**
supplétion, **11**, 56
synonymie, **137**

terme, **14**
thème verbal, **44**
thesaurus, **14**
type, **96**, 104, 128

variante orthographique, 56
variation morphosyntaxique, 12
variation terminologique, 7–13

Glossaire

Aboutness, relation sémantique entre termes. Elle relie des termes proches et est alors considérée comme souhaitable en recherche d'information. Elle peut être rapprochée de la relation `voir-aussi`. (*p. 17*)
Affixation. L'affixation correspond à un ensemble de procédés de formation de *lexèmes*. À travers la préfixation et la suffixation, elle consiste en la combinaison d'une *base* et d'un affixe : {*anévrisme*, *anévrismal*}. (*p. 40*)
Affixation isocatégorielle. Dans une affixation isocatégorielle le *lexème* affixé a la même catégorie syntaxique que sa *base*. (*p. 42*)
Affixation transcatégorielle. Dans une affixation transcatégorielle le *lexème* affixé n'a pas la même catégorie syntaxique que sa *base*. (*p. 41*)
Allomorphie. Les allomorphies sont des variations phonologiques et donc formelles des éléments morphologiques, déclenchées par des contextes morphologiques. (*p. 38*)
Ambiguïté. Il y a ambiguïté là où il existe plus d'un trait en compétition pour caractériser un mot. Avec l'ambiguïté syntaxique, il existe plus d'une étiquette syntaxique et/ou trait morphologique disponibles pour un mot. Avec l'ambiguïté sémantique, il existe plus d'une signification pour un mot. Dans les applications automatiques, le système doit résoudre ces ambiguïtés et assigner les traits les plus pertinents dans un contexte donné. (*p. 54*)
Antonymie. L'antonymie relie des notions contraires ou opposées. Formellement, elle met en relation des termes *co-hyponymes* qui ne sont pas *synonymes*. (*p. 138*)
Base. Le terme *base* regroupe, en morphologie constructionnelle, les lexèmes et les bases non autonomes, deux signes linguistiques qui peuvent servir de base lors de la formation des lexèmes affixés. (*p. 36*)
Baseline. La baseline correspond à la non-application des traitements ou bien à l'application d'un traitement simple, 'de base'. Dans notre travail, la *désuffixation* et les traitements linguistiques de la variation morphologique des termes sont opposés à la baseline, la non-application de ces traitements. (*p. 26*)
Base non autonome. Les bases non autonomes n'apparaissent pas à l'état autonome et ne constituent pas un matériau direct pour la syntaxe. Elles peuvent se substituer aux bases autonomes dans la formation des lexèmes : {foie*/Nom*, hépat*ique/Adj*}. On les appelle également bases supplétives, archéoconstituants, composants néolatins, bases savantes ou primitifs supplétifs. (*p. 37*)
Base supplétive, voir *base non autonome*. (*p. 37*)

Bruit. Le bruit correspond aux données indésirables et/ou fausses proposées par un système automatique. Le bruit est opposé au *silence*. (*p. 9*)
CIM, Classification internationale des maladies et des problèmes de santé connexes. (*p. 19*)
Circumfixation. La circumfixation est propre aux langues comme l'allemand et le néerlandais, elle consiste en un ajout simultané d'un préfixe et d'un suffixe. (*p. 64*)
CISMeF, Catalogue et Index des Sites Médicaux Francophones. `www.chu-rouen.fr/cismef` (*p. 101*)
Classification. Les classifications sont utilisées pour l'encodage des documents dans des buts statistiques. À la différence des *nomenclatures*, les termes contenus dans les classifications que nous avons rencontrées ne sont pas des créations naturelles, mais correspondent à un métalangage artificiel. Ils servent à spécifier des classes des documents. (*p. 15*)
Cognat. Les cognats sont des mots graphiquement et sémantiquement proches dans différentes langues : *sténose* en français *vs stenosis* en anglais. (*p. 94*)
Co-hyponymes, termes qui ont le même *hyperonyme* et qui peuvent, dans certains cas, être substituables. (*p. 22*)
Composition. La composition est un procédé de formation des *lexèmes* qui permet de combiner au moins deux composants, que ce soient des lexèmes ou bases non autonomes : {*anévrisme*, *anévrismorraphie*}. (*p. 43*)
Composition cachée. Dans une composition cachée (*concealed compounding*), la forme tronquée d'un *lexème* est combinée avec d'autres *lexèmes* tronqués ou non. Les deux gardent leur contenu sémantique. Par exemple, *handi*place, *handi*vie, *handi*mag sur la base de *handicap*. (*p. 40*)
Composition populaire. Dans une composition populaire française, l'élément recteur se trouve en première position : *porte-fenêtre*. (*p. 43*)
Composition savante. Dans une composition savante les composants sont souvent des bases non autonomes et proviennent du latin ou du grec. Le composant recteur se trouve généralement en dernière position : {*anévrisme*, *anévrismorraphie*}. (*p. 45*)
Concept. Un concept signifie des notions et des objets d'une manière abstraite. (*p. 14*)
Conversion. La conversion est un procédé de formation de *lexèmes*. La conversion peut prendre en entrée un seul élément de n'importe quelle catégorie. Elle produit en sortie un seul lexème de catégorie majeure. Les catégories syntaxiques des lexème base et lexème convert sont différentes : {*muqueuse/Adj*, *muqueuse/Nom*}. (*p. 42*)
Désuffixation. La désuffixation correspond à un procédé linéaire de suppression des affixes ou des chaînes finales en général. Chez certains auteurs, elle s'oppose à la *racinisation*. (*p. 41*)
Désuffixeur. Les désuffixeurs permettent de réduire les variantes morphologiques des mots à leur *stem*, qui correspond, dans l'idéal, à une *base* ou un *lemme*, selon que l'affixation et/ou la flexion sont traitées. (*p. 25*)
Disponibilité. La disponibilité d'un affixe concerne sa possibilité de construire des lexèmes. Elle prend en compte la nature de l'affixe et les contraintes de son appli-

cation. Cette notion correspond aux approches qualitatives, qui montrent une image binaire de la *productivité* des affixes : ils sont productifs ou non productifs. (*p. 95*)
Étiqueteur morphosyntaxique. Les étiqueteurs morphosyntaxiques sont des outils automatiques de traitement de données textuelles. Ils proposent d'assigner aux mots une étiquette grammaticale (nom, adjectif, etc.), munie parfois de traits morphologiques (nombre, genre, etc.). (*p. 54*)
Explosion de concept. En science de l'information et de communication, on parle de l'explosion d'un concept lorsqu'une notion générique est étendue avec ses notions spécifiques. Pour ceci, on utilise les *relations hiérarchiques* d'une terminologie structurée. (*p. 124*)
Famille morphologique. Une famille morphologique couvre l'ensemble des *lexèmes* reliés par des relations formelles et sémantiques à travers des règles de formation. (*p. 47*)
Hapax, type possédant une seule *occurrence* dans un corpus. (*p. 97*)
Holonyme, terme qui désigne le tout dans une *relation partitive*. Voir *Méronyme*. (*p. 137*)
Hyperonyme, terme générique dans une *relation hiérarchique* ou hyperonymique. (*p. 137*)
Hyperonymie, chez certains auteurs, le versant lexical de la *relation hiérarchique*. Cette dernière étant d'ordre conceptuel. (*p. 137*)
Hyponyme, terme spécifique dans une *relation hiérarchique* ou hyperonymique. (*p. 137*)
Interfixe, une voyelle, le plus souvent *-o-* ou *-i-*, qui s'intercale entre deux composants pour améliorer la phonologie du lexème. (*p. 45*)
Langue de spécialité. Une langue de spécialité couvre un domaine scientifique ou technique. Elle comporte un lexique spécifique à ce domaine, sa terminologie, mais elle peut également présenter des particularités syntaxiques, morphologiques ou autres. (*p. 60*)
Langue générale. La langue générale n'est pas consacrée à un domaine de spécialité précis. Elle reflète souvent l'usage quotidien de la langue, tel que le fait par exemple la presse grand public. Les articles des journaux ne sont évidemment pas complètement neutres de ce point de vue : ils concernent différents domaines de spécialités (agriculture, cuisine, politique, etc.), mais on considère généralement que le niveau de leur spécialisation n'est pas élevé et le mélange de différents domaines les neutralise les uns par rapport aux autres. (*p. 60*)
Lemmatisation. La lemmatisation consiste à définir les *lemmes* des mots. (*p. 54*)
Lemme. Un lemme couvre une notion proche de celle de *lexème*. Il correspond aux formes 'citationnelles' des mots : infinitif pour les formes verbales, singulier pour les noms, singulier masculin pour les adjectifs et les pronoms. (*p. 46*)
Lexème. Un lexème est une unité linguistique abstraite dans le sens où elle est dépourvue de marques flexionnelles. Le lexème s'oppose au mot-forme, qui est pourvu de marques flexionnelles. (*p. 10*)
Lexème lacunaire. Dans les travaux de (Corbin, 1987), un lexème lacunaire est

possible et prédictible morphologiquement mais n'est pas attesté par l'usage ni dans les dictionnaires. (*p. 47*)

Lexique morphologique. Un lexique morphologique se présente souvent sous forme de couples {*lemme*, *forme*} pour les flexions et sous forme de couples {*base*, *dérivé*} pour les dérivations. (*p. 56*)

log, ou journal de bord. Un fichier qui garde les traces des événements produits par un programme. Un log d'un moteur de recherche contient des traces des connexions au serveur : adresse Internet (IP) des machines connectées, mots clés des utilisateurs, durée des connexions, etc. (*p. 125*)

Marqueurs d'une relation sémantique. En structuration de termes, les marqueurs d'une relation sémantique sont des unités textuelles caractéristiques de cette relation. *is a* et *including* sont des marqueurs de l'*hyperonymie*, et *entraîne* est un marqueur de la causalité. (*p. 141*)

Méronyme, terme qui désigne la partie dans une *relation partitive*. Voir *Holonyme*. (*p. 137*)

Méronymie, voir *Relations partitives*. (*p. 137*)

MeSH, Medical Subject Headings. (*p. 16*)

Morphologie constructionnelle. La morphologie constructionnelle décrit les procédés de formation de *lexèmes* : l'*affixation*, la *conversion* et la *composition* parmi les procédés centraux. La *composition cachée* et les *mots-valises* font partie des procédés périphériques. (*p. 39*)

Morphologie dérivationnelle, voir *morphologie constructionnelle*. (*p. 39*)

Morphologie flexionnelle. La morphologie flexionnelle décrit la variation des *lexèmes* lors de leur insertion dans les énoncés. Elle consiste en l'application d'affixes flexionnels (désinences ou flexions) à un *lexème* ou un *lemme*. (*p. 46*)

Mot. Ce terme est employé dans le sens informatique : il s'agit d'une chaîne de caractères obtenue suite à la segmentation d'une proposition. Quant à la segmentation, elle peut être faite sur les espaces et/ou la ponctuation et/ou les symboles ($, *, etc) et/ou les nombres. (*p. 8*)

Mots pleins. Les mots pleins s'opposent aux *mots vides*. Il s'agit des catégories majeures : noms, adjectifs, verbes et adverbes. (*p. 149*)

Mots-valises. La formation des mots valises est soumise à des patrons basés sur des segments phonologiquement identiques ou similaires dans les lexèmes combinés. Par exemple : *métropolitique*, *potimarron*, *hippidémie*. (*p. 40*)

Mots vides. Les mots vides peuvent regrouper les mots grammaticaux (articles, prépositions, pronoms, adverbes, etc.) et, souvent, les mots très fréquents d'un domaine. (*p. 10*)

NLM, U. S. National Library of Medicine. (*p. 16*)

Nomenclature. Les nomenclatures sont utiles dans les traitements informatiques des documents numériques. Elles recensent les termes réels et aident ainsi l'accès aux informations contenues dans ces documents. (*p. 14*)

Occurrence. En linguistique de corpus, la notion d'occurrence correspond au nombre total d'apparitions d'un type (forme) dans le corpus. L'occurrence est souvent opposée

au *type*. (*p. 96*)
OMS, Organisation Mondiale de la Santé. (*p. 19*)
Partie-tout, voir *Relations partitives*. (*p. 137*)
Patron lexico-syntaxique. Les patrons lexico-syntaxiques utilisés pour la détection de relations sémantiques entre termes reposent sur la présence d'indices linguistiques (*such as*, *including*, *tel que*, *certain*) dans les textes. À la différence des *marqueurs*, les patrons identifient la relation recherchée plus précisément en définissant à la base les contraintes syntaxiques ou typographiques du contexte, la position des termes en relation, etc. (*p. 142*)
Précision. La précision indique le pourcentage de documents corrects parmi ceux qui ont été trouvés par un système automatique : il s'agit de l'exactitude des résultats. La précision est souvent évaluée en même temps que le *rappel*. (*p. 27*)
Préfixe. Un préfixe est un affixe dérivationnel. Suite aux opérations d'affixation, il apparaît avant la base. Dans les traitements automatiques, un préfixe peut signifier 'la chaîne initiale' d'un mot. Voir aussi *suffixe*. (*p. 38*)
Productivité morphologique. La productivité morphologique est globalement reflétée par la fréquence d'emploi des affixes. Deux acceptions de productivité se dégagent : productivité qualitative (*disponibilité*) et productivité quantitative (*rentabilité*). (*p. 95*)
Racinisation. Chez certains auteurs, la racinisation s'oppose à la *désuffixation*. La racinisation sous-entend alors une analyse linguistique et morphologique des mots. Sinon la racinisation couvre la même notion que *désuffixation* ou *stemming*. (*p. 41*)
Rappel. Le rappel indique le pourcentage de documents trouvés par rapport à ce qui est attendu : il s'agit de la complétude des résultats fournis par un système automatique. Le rappel est souvent évalué en même temps que la *précision*. (*p. 27*)
Règles contextuelles. Les outils d'étiquetage morphosyntaxique s'appuient sur les règles contextuelles pour définir les traits morphosyntaxiques du mot courant grâce à l'analyse de ses mots proches. (*p. 55*)
Règles lexicales. Les outils d'étiquetage morphosyntaxique s'appuient sur les règles lexicales pour définir les traits morphosyntaxiques du mot courant grâce à la forme (souvent sa chaîne de caractères finale). (*p. 55*)
Relations hiérarchiques. Les relations hiérarchiques relient deux *concepts* dont l'un recouvre une notion 'plus large' que l'autre. Elles permettent d'organiser les termes en un arbre hiérarchique et d'établir ainsi une arborescence du domaine. Les relations hiérarchiques couvrent parfois les *relations partitives*. Il s'agit d'une relation structurante. Elle met en relation des *hyperonymes* et des *hyponymes*. (*p. 137*)
Relations partitives. Les relations partitives sont aussi appelées `partie-de` ou méronymie. Elles identifient les parties d'un objet et servent à décrire les objets manufacturés ou artefacts. Elles sont parfois regroupées avec la *relation hiérarchique* et participent alors aussi à la constitution de l'arborescence du domaine. Elles mettent en relation des *méronymes* et des *holonymes*. (*p. 137*)
Relations transversales. Les relations transversales correspondent aux relations entre des termes situés dans les différentes branches de la hiérarchie. Elles permettent

d'identifier des rapports entre termes que les *relations synonymiques*, *antonymiques* ou *hiérarchiques* ne peuvent pas refléter. Il s'agit par exemple des rôles thématiques (objet, instrument, agent, patient, etc.), des relations causales, temporelles, spatiales, mais surtout des relations spécifiques à un domaine. (*p. 138*)

Rentabilité. La rentabilité d'un affixe concerne la possibilité qu'il a de s'appliquer à un grand nombre de bases et de former un grand nombre de lexèmes construits. Cette notion correspond aux approches quantitatives de la *productivité morphologique* et consiste souvent en un recensement des formations d'un affixe. Dans ce cas, la productivité d'un procédé est une variable qui se situe sur un continuum qui va du non productif au très productif. (*p. 95*)

Sacs de mots. Dans les sacs de mots la structure originale des termes ou expressions n'est pas considéré. (*p. 9*)

Segmentation d'une proposition. La segmentation d'uen proposition en mots peut être faite sur les espaces et/ou la ponctuation et/ou les symboles ($, *, etc) et/ou les nombres. (*p. 8*)

Silence. Le silence correspond aux résultats désirables mais non fournis par un système. C'est le complémentaire du *bruit*. (*p. 26*)

SNOMED, Nomenclature systématique des médecines humaine et vétérinaire. (*p. 18*)

Sous-domaine. Les sous-domaines d'une langue de spécialité correspondent à ses différentes spécialités qui peuvent se recouvrir partiellement. Dans le domaine médical, on peut distinguer environ 70 sous-domaines médicaux (cardiologie, néphrologie, neurologie, hématologie, rhumatologie, pédiatrie, etc.). Le sous-domaine d'un texte peut être décrit à l'aide de mots clés pris par exemple dans un thesaurus de référence. (*p. 100*)

Stem. Un stem correspond, dans l'idéal, à une *base* ou un *lemme*, selon que l'affixation et/ou la flexion sont traitées. (*p. 25*)

Stemming. Le stemming correspond à un procédé linéaire de suppression des affixes ou des chaînes finales en général. Selon les auteurs, le stemming couvre la *désuffixation* ou la *désuffixation* et la *racinisation*. (*p. 41*)

Suffixe. Un suffixe est un affixe dérivationnel. Suite aux opérations d'affixation, il apparaît après la base. Dans les traitements automatiques, un suffixe peut signifier 'la chaîne finale' d'un mot. Voir aussi *préfixe*. (*p. 38*)

Supplétion. La supplétion prend en charge la formation de *lexèmes* qui présentent des *bases* sémantiquement équivalentes mais dont les langues d'origine sont différentes. Leurs formes graphiques sont donc également différentes. (*p. 11*)

Synonymie. La synonymie met en relation des expressions qui ont, dans un contexte donné, des sens équivalents. Sortie de ce contexte, la relation de synonymie peut s'atténuer ou bien perdre de la pertinence. (*p. 137*)

TAL, Traitement automatique des langues. (*p. 1*)

Techniques d'appariements lexicaux, traitements au niveau des caractères, au niveau morphologique, abstraction de l'ordre des mots, suppression des mots vides, etc. (*p. 7*)

Terme. Un terme se rapporte aux expressions linguistiques qui sont utilisées pour

dénommer les *concepts*. Les termes se retrouvent dans les textes. (*p. 14*)
Thème verbal, forme verbale non fléchie. (*p. 44*)
Thesaurus. Le thesaurus est utilisé pour l'indexation des connaissances et la recherche d'information dans les bases de données et les portails. Un thesaurus assure l'accès à la littérature scientifique et technique du domaine. Les termes d'un thesaurus sont, le plus souvent, des créations artificielles dont le but est de cerner aussi précisément que possible le sens d'un concept. (*p. 14*)
Transaction log analysis. Il s'agit d'une analyse des *logs* des moteurs de recherche par exemple. Cette analyse permet de collecter des informations sur le comportement des utilisateurs et de tirer les conclusions utiles à l'amélioration de l'interface, du fonctionnement du système, etc. (*p. 125*)
Type. En linguistique de corpus, la notion de type correspond à chaque forme (ou vocable) différente qui apparaît dans ce corpus. Le type est souvent opposé à l'*occurrence*. (*p. 96*)
UMLS, Unified Medical Language System. (*p. 21*)
Variation au niveau des caractères. Ce type de variations englobe : la variation de casse, l'omission d'accents, les variantes orthographiques, les accents erronés, et les fautes de frappe. (*p. 9*)