



Méthode d'enrichissement et d'élargissement d'une ontologie à partir de corpus de spécialité multilingues

Yuliya Korenchuk

► To cite this version:

Yuliya Korenchuk. Méthode d'enrichissement et d'élargissement d'une ontologie à partir de corpus de spécialité multilingues. Autre [cs.OH]. Université de Strasbourg, 2017. Français. NNT : 2017STRAC014 . tel-01794940

HAL Id: tel-01794940

<https://theses.hal.science/tel-01794940>

Submitted on 18 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE 520 « Humanités »

EA 1339 LiLPa (Linguistique, Langues, Parole)

THÈSE présentée par :

Yuliya KORENCHUK

soutenue le : 11 juillet 2017

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Sciences du Langage / Linguistique informatique

**Méthode d'enrichissement et
d'élargissement d'une ontologie à partir
de corpus de spécialité multilingues**

THÈSE co-dirigée par :

Mme TODIRAȘCU Amalia
Mme BERNHARD Delphine

Professeur, Université de Strasbourg
Maître de conférences, Université de Strasbourg

RAPPORTEURS :

M. HEID Ulrich
Mme ZANNI-MERK Cécilia

Professeur Dr., Universität Hildesheim
Professeur, INSA Rouen Normandie

À ma mère qui m'a appris à dépasser mes limites.

Remerciements

Je remercie avant tout ma directrice de thèse Prof. Amalia Todiraşcu et Mme Delphine Bernhard qui a assuré le co-encadrement. Grâce à leur expertise, conseil et encouragements, je n'ai pas perdu le Nord tout au long de ce chemin.

Je tiens à remercier également M. Jean Marc Zuber, fondateur de Rebus SAS, d'avoir accepté le projet CIFRE.

Je suis reconnaissante aux membres du jury, Prof. Ulrich Heid et Prof. Cécilia Zanni-Merk, pour l'intérêt porté à mon travail.

Je remercie toute l'équipe LiLPa et son Directeur Prof. Rudolph Sock pour l'accueil bienveillant et le soutien qui ont marqué mon parcours.

J'adresse mes remerciements à l'équipe LexiCon de l'Université de Grenade, et personnellement au Prof. Pamela Faber, pour l'accès à la ressource EcoLexicon.

La définition du sujet est loin d'être une tâche simple et je suis très reconnaissante au Prof. François Rousselot et au Prof. Thierry Grass pour leur aide à cette étape.

Un grand merci au Dr Gaël Lejeune pour ses conseils et pour les relectures, au Dr Elena Vulpe pour son aide dans l'évaluation des résultats et au Dr Mirabela Navlea pour son soutien et ses encouragements.

Je pense également à tous mes collègues de LiLPa et de Rebus SAS. Vous étiez toujours là pour moi et je vous remercie pour tous les échanges qui m'ont beaucoup appris.

Le bon déroulement du projet se doit aussi à ma famille. Je suis particulièrement reconnaissante à ma mère qui a su être proche malgré la distance, et à Sylvain qui, par son exemple, m'a motivé à travailler rigoureusement et passionnément. J'adresse également une pensée à M. Marcel Knack pour son soutien inestimable à la fin de la thèse.

Je voudrais remercier tous mes amis qui ont été avec moi avant le début de la thèse, pendant ces années et qui sont toujours là. Leur présence a marqué ce temps par des moments inoubliables qui m'ont sans cesse rappelé que la vie ne se limite pas à une thèse.

Enfin, ce projet n'aurait pas été possible sans financement. Il a été mené grâce à la convention CIFRE 2013/0744 de l'ANRT. Je suis très reconnaissante pour ce financement qui m'a motivé jusqu'à l'aboutissement du projet.

Table des matières

Introduction	1
I Processus d'élargissement et d'enrichissement d'une ontologie	15
1 Extraction terminologique	19
1.1 Termes et terminologie	19
1.1.1 Le terme dans la linguistique	20
1.1.2 Le terme dans le TAL	23
1.2 Méthodes d'extraction terminologique	28
1.2.1 Méthodes statistiques	29
1.2.2 Méthodes linguistiques	31
1.2.3 Méthodes mixtes	32
1.3 Alignement de terminologie multilingue	38
1.3.1 Alignement des termes composés monolexicaux	38
1.3.2 Alignements des termes polylexicaux	43
1.4 Conclusion	47
2 Enrichissement et élargissement d'une ontologie	49
2.1 Concepts, relations sémantiques et ontologies	49
2.1.1 Concept	50
2.1.2 Vers un choix pragmatique	52
2.1.3 Relations sémantiques	54
2.2 Ressources sémantiques et ontologies	56
2.2.1 Ontologie dans la philosophie	57
2.2.2 Ontologie dans l'intelligence artificielle	57
2.2.3 Autres types de ressources	58

TABLE DES MATIÈRES

2.3	Méthodes d'enrichissement et d'élargissement des ontologies	59
2.3.1	Méthodes manuelles	60
2.3.2	Méthodes statistiques	65
2.3.3	Méthodes utilisant des ressources structurées	68
2.3.4	Méthodes mixtes	72
2.4	Conclusion	75

II Utilisation des ressources endogènes pour l'élargissement et l'enrichissement d'une ontologie 77

3 Ressources 83

3.1	Ontologies et thésaurus spécialisés	83
3.1.1	Ontologie EcoLexicon	84
3.1.2	Agrovoc et GEMET	86
3.1.3	Statistiques liées aux ressources ontologiques	88
3.2	Corpus PatTR	89
3.2.1	Caractéristiques techniques	90
3.2.2	Choix des sous-corpus pour l'évaluation	91
3.3	Prétraitement et annotation du corpus	95
3.3.1	Tokenisation et lemmatisation	95
3.3.2	Annotation par mots informatifs et vides	95
3.3.3	Annotation manuelle du corpus pour évaluation	101
3.4	Conclusion	104

4 Extraction terminologique multilingue à partir des ressources endogènes 105

4.1	Méthode d'extraction de termes monolexicaux à l'aide de ressources morphologiques endogènes	106
4.1.1	Construction de la liste initiale de termes	107
4.1.2	Construction des ressources morphologiques endogènes	111
4.1.3	Extraction des candidats monolexicaux	114
4.2	Méthode d'extraction de termes polylexicaux à base de patrons morphosyntaxiques endogènes	115
4.2.1	Apprentissage des patrons	115
4.2.2	Sélection des mots de schéma	116
4.3	Évaluation	117

TABLE DES MATIÈRES

4.3.1	Ressources	117
4.3.2	Mesures d'évaluation	117
4.3.3	Système de référence	118
4.3.4	Résultats	120
4.4	Discussion	128
4.4.1	Influence de l'annotation par les mots informatifs et vides . . .	128
4.4.2	Candidats extraits par des mesures statistiques	129
4.4.3	Candidats extraits par les ressources morphologiques endogènes	130
4.4.4	Patrons morphosyntaxiques endogènes	131
4.5	Conclusion	133
5	Exploitation des n-grammes de caractères pour la recherche des traductions	135
5.1	Méthode de recherche de traductions à base de n-grammes de caractères	137
5.1.1	Modèle de traductions à base de n-grammes de caractères . . .	137
5.1.2	Alignement de la terminologie bilingue	142
5.2	Évaluation	144
5.2.1	Comparaison avec des méthodes d'identification des cognats .	145
5.2.2	Comparaison avec TTC TermSuite	150
5.3	Discussion	154
5.3.1	Influence du paramètre n sur les résultats d'alignement	154
5.3.2	N-grammes de caractères et formants néo-classiques	155
5.3.3	Identification de cognats	156
5.3.4	Traduction des termes composés	157
5.4	Conclusion	158
6	Construction des familles morphologiques endogènes	159
6.1	Définition des familles morphologiques endogènes	160
6.2	Méthodologie	161
6.2.1	Création des index de n-grammes	162
6.2.2	Constitution des familles morphologiques	162
6.2.3	Optimisation de la présentation de résultats	164
6.3	Évaluation	165
6.3.1	Familles pour les libellés de l'ontologie	165
6.3.2	Recherche des néologismes	168
6.4	Discussion	175

TABLE DES MATIÈRES

6.4.1	Structure des familles morphologiques endogènes	175
6.4.2	Pertinence des familles morphologiques endogènes	176
6.4.3	Aspect multilingue des familles morphologiques	177
6.4.4	Identification des termes à faible fréquence	177
6.4.5	Fonctionnalités de l'interface utilisateur	178
6.4.6	Limites d'application	178
6.5	Conclusion	178
7	Enrichissement et élargissement de l'ontologie à l'aide des familles morphologiques endogènes	181
7.1	Élargissement et enrichissement d'une ontologie	181
7.1.1	Pré-traitement du corpus	182
7.1.2	Extraction des termes	184
7.1.3	Recherche des traductions	188
7.1.4	Enrichissement et élargissement de l'ontologie	189
7.1.5	Résumé de l'exemple français	193
7.2	Évaluation	194
7.2.1	Ressources	195
7.2.2	Résultats	196
7.3	Discussion	197
7.3.1	Composition des familles morphologiques endogènes	198
7.3.2	Influence de la structure des termes	199
7.3.3	Alignement multilingue des familles morphologiques endogènes	202
7.4	Conclusion	203
	Conclusion et perspectives	205
	Annexes	211
A	Guide de lecture	211
B	Catégories de la Classification Internationale de Brevets (CIB)	213
C	Statistiques du corpus PatTR	215
D	Échantillons des termes candidats	217

TABLE DES MATIÈRES

Bibliographie	219
Liste de publications	233

Liste des tableaux

3.1	Statistiques des ressources ontologiques	89
3.2	Nombre de segments dans le corpus PatTR ¹	90
3.3	Sélection manuelle des catégories de la CIB ²	91
3.4	Sélection manuelle des concepts généraux	92
3.5	Couverture du corpus et le pourcentage des concepts généraux (CG) .	93
3.6	Statistiques des sous-corpus retenus	94
3.7	Statistiques des sous-corpus annotés	95
3.8	Mots informatifs triés par fréquence absolue (Corpus C02F, FR) . . .	99
3.9	Influence du seuil pour les mots informatifs sur le tri par fréquence absolue (Corpus C02F, FR)	100
3.10	Termes annotés (corpus lemmatisés)	104
4.1	Top-10 résultats pour la fréquence absolue maximale et TFXIDF ob- tenus sur le corpus lemmatisé B09C (FR)	109
4.2	Top-5 de n-grammes dans GEMET par fréquence décroissante	112
4.3	Top-5 de n-grammes par score terminologique décroissant (B09C lem- matisé, GEMET)	114
4.4	Top-5 de patrons morphosyntaxiques endogènes par fréquence dé- croissante (B09C lemmatisé, GEMET)	116
4.5	Évaluation TTC TermSuite 2.2 : termes monolexicaux	119
4.6	Évaluation TTC TermSuite 2.2 : termes polylexicaux	119
4.7	Évaluation TTC TermSuite 1.6 : termes monolexicaux	120
4.8	Évaluation des termes monolexicaux sur les corpus lemmatisés, mo- dèles à partir des ressources, $n = 3$	122
4.9	Évaluation des termes monolexicaux sur les corpus tokenisés, modèles à partir des ressources, $n = 3$	123
4.10	Évaluation des termes monolexicaux sur les corpus tokenisés, modèles à partir de Top-90 % de statistiques, $n = 3$	125

LISTE DES TABLEAUX

4.11	Les top-10 mots non-informatifs triés par fréquence décroissante, corpus lemmatisé	125
4.12	Les top-10 mots non-informatifs triés par fréquence décroissante, corpus tokenisé	126
4.13	Évaluation des termes polylexicaux (en utilisant les termes monolexicaux validés)	127
4.14	Évaluation des termes polylexicaux (en utilisant les candidats monolexicaux, modèle GEMET)	128
4.15	Top-10 candidats monolexicaux pour les ressources morphologiques endogènes, corpus lemmatisé B09C	130
4.16	Exemples d'étiquetage morphosyntaxique erroné, format token-étiquette-lemme, TreeTagger, corpus B09C	131
5.1	Sortie d'Anymalign (EN-DE) à partir de GEMET	140
5.2	N-grammes après l'application de filtres (EN-DE, GEMET)	141
5.3	Impact des filtres (jeu de développement, $n = 4$)	142
5.4	Éléments du score de traduction	143
5.5	Résultats obtenus sur EcoLexicon, modèle entraîné sur GEMET	148
5.6	Résultats obtenus sur GEMET, modèle entraîné sur EcoLexicon	149
5.7	Nombre de termes traduits sur Top-50 candidats selon le score de spécificité (TermSuite)	153
5.8	Exemples de formants néo-classiques dans les n-grammes alignés à partir de GEMET	156
5.9	Éléments pour la reconnaissance de cognats	156
5.10	Alignement des éléments des termes composés (GEMET)	157
6.1	Taille et couverture des familles morphologiques endogènes	166
6.2	Répartition du sous-corpus C25B PatTR par période	169
6.3	Précision des termes candidats (%) par rapport à la liste triée dans l'ordre décroissant des scores (corpus 2000)	171
6.4	Famille morphologique pour le terme <i>aérobie</i>	171
6.5	Familles morphologiques dans le temps	172
6.6	Résultats retenus de Google Ngrams Viewer	174
7.1	N-grammes permettant d'extraire les termes candidats	185
7.2	Traductions pour les termes candidats français	188
7.3	Traductions pour les termes candidats anglais	189
7.4	Traductions pour les termes candidats allemands	190

LISTE DES TABLEAUX

7.5	Statistiques d'EcoLexicon	195
7.6	Statistiques des sous-corpus C02F et B09C	195
7.7	Libellés (termes) d'EcoLexicon pris en compte dans le traitement . .	196
7.8	Couverture des familles morphologiques endogènes	196
7.9	Taille des familles morphologiques endogènes	197
B.1	Corpus PatTR : les catégories de la CIB sélectionnées	214
C.1	Corpus PatTR : le nombre de segments parallèles	216
D.1	Échantillons des termes candidats avec leurs fréquences (corpus 2000)	218

Table des figures

1	Exemple d'une ontologie pour la recherche d'information multilingue .	2
2	Degré de formalisation des ressources (Navigli et Velardi, 2008)	4
3	Étapes de traitement	6
4	Enrichissement et élargissement d'une ontologie	17
1.1	Exemple de la résolution probabiliste de segments (Brooke <i>et al.</i> , 2014)	30
1.2	La construction des mots composés non-connus à l'entraînement Cap <i>et al.</i> (2014)	39
2.1	Triangle sémiotique (Staab et Studer, 2009)	51
2.2	Exemple des relations hiérarchiques du concept <i>STORM</i> (EcoLexicon)	53
2.3	Exemple des relations méronymiques du concept <i>STORM</i> (EcoLexicon)	55
2.4	Exemple des relations non-hiérarchiques du concept <i>STORM</i> (Eco- Lexicon)	55
2.5	Le nombre de relations sémantiques renseignées dans les différentes versions de KnowNet (Cuadros et Rigau, 2008)	58
2.6	Les trois niveaux de connaissance et les trois étapes de la méthodologie de TERMINAE (Omrane <i>et al.</i> , 2011b)	63
2.7	Dernières assertions apprises par NELL (11 février 2017)	66
2.8	Les synsets de WordNet pour la requête <i>electricity</i>	70
2.9	Exemples de définitions en langue naturelle Velardi <i>et al.</i> (2006) . . .	72
2.10	Architecture de notre système	80
2.11	Modules implémentés de notre système	81
3.1	Le schéma sémantique d'événement naturel ³	85
3.2	Extrait des informations du concept <i>OSMOSIS</i> d'Agrovoc en RDF .	89
3.3	Densité moyenne de termes dans les mots informatifs dans les corpus complets	98
3.4	Réduction du nombre moyen des mots informatifs et des termes en fonction du seuil (FR)	99

TABLE DES FIGURES

4.1	Extraction terminologique dans notre méthode	105
4.2	Extraction de termes monolexicaux	106
4.3	Évolution de l'intersection de la fréquence absolue et TFxIDF par valeur de seuil Top-X % scores	110
4.4	Place de la recherche des termes polylexicaux dans notre méthode . .	115
4.5	Évaluation du paramètre n (valeurs moyennes pour trois ressources et deux corpus)	121
4.6	Évaluation du seuil Top-X % pour les mesures statistiques, corpus lemmatisé	124
4.7	Évaluation du nombre de mot de schéma pour l'extraction de termes polylexicaux	126
5.1	Place de la recherche des traductions dans notre méthode	135
5.2	Processus de recherche de traductions	136
5.3	Apprentissage du modèle de traduction	138
5.4	Alignement mot-à-mot pour la paire de langues anglais - français (Navlea, 2014)	139
5.5	Identification des traductions	142
5.6	Évaluation du seuil pour la baseline (valeurs moyennes par rapport à toutes les paires de langues dans les deux directions)	147
5.7	Exemple de traductions EN-DE proposées par TermSuite pour le can- didat <i>contaminated soil</i> (corpus B09C)	150
5.8	Influence du paramètre n , où $n \in (3 : 5)$	155
6.1	Place de la constitution des familles morphologiques dans notre méthode	159
6.2	Graphe bipartite reliant les mots et les n -grammes de phonèmes (Ha- thout, 2011)	160
6.3	Création des index	162
6.4	Construction des familles morphologiques endogènes	163
6.5	Création de l'index	169
6.6	Calcul des scores	170
6.7	Construction des familles morphologiques endogènes	170
6.8	Google Ngrams Viewer, EN, 1980 - 2008	173
6.9	Fréquence des termes sur SciFinder	175
7.1	Concept <i>CHLORINE</i> (chlore) dans EcoLexicon	191
7.2	Concept <i>PURIFICATION</i> dans EcoLexicon	192
7.3	Concept <i>SAND</i> dans EcoLexicon	202

Introduction

Depuis l'apparition de l'écriture, l'humanité n'a pas cessé d'accumuler des données textuelles. Autrefois enregistrées sous forme de papyrus, de parchemins, de manuscrits ou de livres, les données textuelles sont à notre époque largement disponibles sous forme de documents électroniques. Une partie de ces documents, par exemple des articles scientifiques, des brevets ou des manuels, contient des connaissances sur des domaines de spécialité¹. L'accès à ces connaissances est possible grâce aux systèmes de recherche d'information (dont les moteurs de recherche) qui permettent aux utilisateurs de formuler des requêtes et identifient ensuite des documents correspondants.

Une des limites importantes dans la recherche d'information est la langue. En effet, la recherche est actuellement effectuée dans la langue de la requête et ne permet donc pas de retrouver des documents écrits dans d'autres langues. La recherche d'information multilingue vise à dépasser cette limite et à identifier des documents pertinents en plusieurs langues. Cela est possible grâce à une analyse de la requête au niveau sémantique, donc celui des sens qui existent indépendamment d'une langue. Ainsi, le mot français *eau*, le mot allemand *Wasser* et le mot anglais *water* correspondent au même sens.

L'analyse sémantique de la requête nécessite des ressources qui répertorient des sens et des mots correspondants en plusieurs langues. Il s'agit des ressources sémantiques où les sens sont appelés concepts et les mots – leur libellés. Dans le contexte d'un domaine spécialisé, ces libellés correspondent aux termes du domaine en question. La constitution de ce type de ressources est particulièrement compliquée, car il faut non seulement répertorier des termes d'un domaine en plusieurs langues, mais aussi les organiser au niveau des concepts. Cette complexité constitue un verrou technologique, notamment pour des entreprises qui développent des systèmes d'analyse sémantique².

1. Un domaine de spécialité correspond à un champ de connaissances lié à des activités spécialisées, qui peut être assez large (par exemple, *l'environnement* ou *l'informatique*) ou plus précis (*traitement des eaux usées* ou *calcul à hautes performances*).

2. Le contexte industriel est important, car cette thèse a été effectuée dans le cadre d'une convention CIFRE (http://www.anrt.asso.fr/fr/espace_cifre/accueil.jsp) en partenariat avec l'entreprise Rebus SAS (société spécialisée en analyse sémantique de textes pour des domaines de spécialité).

L'objectif de cette thèse est d'exploiter des collections de textes d'un domaine en plusieurs langues (appelées corpus spécialisés multilingues) pour ajouter des nouveaux éléments dans une ressource sémantique destinée à la recherche d'information multilingue. Dans le cadre de cette thèse, nous avons trois langues de traitement : le français, l'anglais et l'allemand³. Le recours à un corpus multilingue permet d'en extraire des termes et leurs traductions pour ajouter des nouveaux libellés et concepts dans la ressource visée, ce qui doit améliorer la recherche d'information multilingue.

Nous allons maintenant illustrer l'utilisation d'une ressource sémantique dans la recherche d'information multilingue avec un exemple.

Exemple de recherche d'information multilingue

Prenons l'exemple de la requête *installation d'osmose inverse*⁴ (un système de purification d'eau) qui appartient au domaine de traitement de l'eau.

La recherche traditionnelle par mots clés permet d'identifier des documents contenant les mots *installation*, *osmose* et *inverse*. Toutefois, cette recherche se limite à des documents en français et ne permet pas de prendre en compte d'éventuels synonymes. Ces limites peuvent être dépassées au niveau sémantique. Pour cela, il est nécessaire de recourir à une ressource sémantique du domaine du traitement de l'eau.

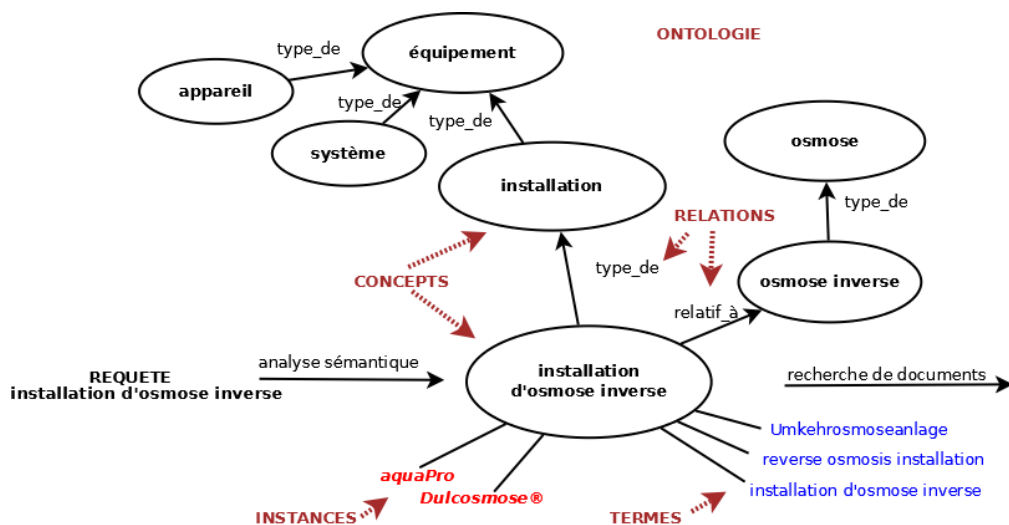


FIGURE 1 – Exemple d'une ontologie pour la recherche d'information multilingue

La Figure 1 illustre un exemple d'une ressource sémantique qui s'appelle une

3. Ci-après, nous utilisons les abréviations FR pour le français, EN pour l'anglais et DE pour l'allemand.

4. https://fr.wikipedia.org/wiki/Osmose_inverse

ontologie. Au sein d'une ontologie spécialisée, les concepts définissent les sens des termes, donc du lexique propre au domaine en question. Un terme, qui peut être un mot simple (par exemple, *osmose*) ou composé (comme le terme allemand *Umkehrosmoseanlage*), sert à dénommer un concept. Certaines ontologies peuvent également contenir des instances qui correspondent à des entités, c'est-à-dire à des objets du monde réel (*aquaPro* et *Dulcosmose*). Les concepts d'une ontologie sont liés par des relations sémantiques qui peuvent être hiérarchiques (*installation* est un *équipement*) ou non-hiérarchiques (*installation d'osmose inverse* est relative à *osmose inverse*).

L'ensemble de ces informations est nécessaire pour traiter la requête *installation d'osmose inverse* au niveau sémantique. Grâce à l'ontologie, le traitement de cette requête prend en compte la synonymie, les instances, les termes et leurs traductions en d'autres langues.

Gestion de la synonymie Dans notre exemple, le concept *installation d'osmose inverse* est lié par la relation hiérarchique *type_de* avec le concept plus général *installation*, qui est à son tour lié à *équipement*. Les concepts *installation*, *système* et *appareil* sont également des types d'équipements et se situent au même niveau. Les libellés de ces concepts sont des synonymes du terme *installation*, c'est-à-dire des mots proches au niveau du sens⁵. Ces synonymes permettent d'étendre la requête initiale à ses variantes *système d'osmose inverse* et à *appareil d'osmose inverse*.

Prise en compte des instances Les instances ne sont pas des termes du domaine en question, mais des objets réels appartenant à un concept (Noy et McGuinness, 2001b). Par exemple, les instances *aquaPro* et *Dulcosmose* sont des installations d'osmose inverse spécifiques. Elles peuvent également être utilisées pour identifier des documents pertinents par rapport à la requête de l'utilisateur. De cette façon, un document peut être retrouvé s'il contient les mots *aquaPro* ou *Dulcosmose*, même en absence des mots de la requête initiale *installation d'osmose inverse*.

Gestion de la terminologie multilingue Les concepts existent indépendamment d'une langue et peuvent avoir des libellés en plusieurs langues. Par exemple, le concept *installation d'osmose inverse* a le libellé correspondant en français et ses traductions en allemand (*Umkehrosmoseanlage*) et en anglais (*reverse osmosis installation*). Ces libellés multilingues peuvent alors servir à l'identification des documents pertinents par rapport à la requête en allemand et en anglais, étant donné qu'ils indiquent le même concept.

5. En fonction de la structure d'une ontologie développée pour un projet concret, les termes *appareil*, *installation*, *équipement* et *système* peuvent également désigner un seul concept *équipement*. Le choix de la structure appartient aux créateurs d'une ressource et il n'existe pas de critères universaux pour définir le degré de précision sémantique à adopter.

Cet exemple montre comment la recherche d'information peut être étendue à des documents qui ne contiennent pas de mots exacts de la requête initiale et qui peuvent être rédigés en d'autres langues. Le fonctionnement de tels systèmes d'analyse repose sur les ressources sémantiques, notamment les ontologies. Leur place parmi les ressources représentant le vocabulaire d'un domaine est discutée dans la section suivante.

Ressources sémantiques pour la recherche d'information multilingue

La représentation des connaissances spécifiques à un domaine peut prendre plusieurs formes. En effet, le choix de la forme dépend largement de l'utilisation visée et des moyens mis à disposition.

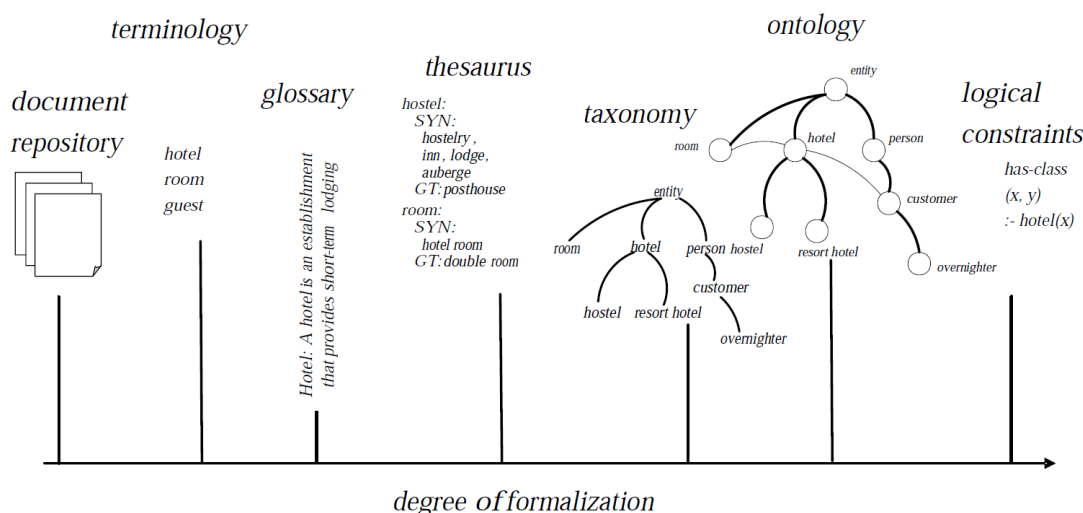


FIGURE 2 – Degré de formalisation des ressources (Navigli et Velardi, 2008)

La Figure 2 illustre la diversité des ressources représentant le vocabulaire d'un domaine en fonction du degré de formalisation des connaissances (Navigli et Velardi, 2008) :

- Une collection de documents non-structurés d'un domaine s'appelle un corpus spécialisé. Si cette collection contient des textes en une seule langue, il s'agit d'un corpus monolingue. Dans le cas contraire, le corpus est multilingue. Lorsqu'il contient des textes en langue source alignés avec leurs traductions, ce type de corpus s'appelle parallèle. Enfin, si les textes ne sont pas des traductions directes, le corpus est comparable.
- Une liste de termes répertorie des mots spécifiques au domaine en question. Dans l'exemple de Navigli et Velardi (2008) que nous utilisons ici, la liste de

termes du domaine d'hôtellerie contient les termes *hôtel*⁶, *chambre* et *hôte*.

- Un glossaire terminologique fournit des définitions pour les termes d'un domaine (*Un hôtel est un établissement qui fournit un hébergement de courte durée*).
- Un thésaurus contient des synonymes : par exemple, un des synonymes du terme *hostel*⁷ est *auberge de jeunesse*, et des termes généraux (*auberge*).
- Une taxonomie organise des termes dans une structure hiérarchique. Par exemple, un *client* est une *personne*, un *hostel* est un type d'*hôtel*, etc.
- Une ontologie correspond aux niveaux supérieurs de la formalisation. En effet, elle opère au niveau des sens, donc celui des concepts, et contient des relations sémantiques entre les concepts. Ces relations peuvent être hiérarchiques (le concept *client* est lié à son concept général *personne*) et non-hiérarchiques (par exemple, *client* est lié à *hôtel* par la relation *se loger dans*).

Les ontologies utilisées dans la recherche multilingue d'information fournissent, selon Shvaiko et Euzenat (2013), un vocabulaire décrivant un domaine de spécialité et une spécification des sens des termes constituant ce vocabulaire.

Ces ontologies sont appelées légères, par opposition à des ontologies lourdes qui possèdent des niveaux d'abstraction supérieurs, notamment, des règles et des axiomes. Ces contraintes logiques s'appliquent à des branches entières de l'ontologie. Par exemple, une axiome peut indiquer que deux concepts sont dissociés : dans une branche des objets géographiques, une rivière ne peut pas être une montagne (Cimiano, 2006). Les règles peuvent être illustrées avec l'exemple de Buitelaar *et al.* (2003) selon lequel si deux instances du concept *Personne* sont en relation *est_marié_à*, il est possible de définir la règle qu'elles sont également liées par la relation *est_amoureux_de*.

Les ontologies lourdes correspondent à la définition de Gruber (1993) selon laquelle une ontologie « est une spécification formelle et explicite d'une conceptualisation partagée d'un domaine d'intérêt ». Les notions-clés utilisées dans cette thèse, dont celle d'ontologie, seront discutées dans la Partie I.

Les domaines de spécialité sont en constante évolution qui entraîne le besoin de maintenir les ressources à jour. Le haut degré de formalisation des ontologies qui font l'objet de cette thèse, implique des processus particuliers qui s'appellent enrichissement et élargissement.

Enrichissement et élargissement d'une ontologie

Pour qu'une ontologie reflète bien un domaine de spécialité, il est important d'ajouter des nouveaux éléments dès leur apparition. L'ajout des nouveaux élé-

6. Les termes de l'exemple sont traduits par nos soins.

7. Nous avons légèrement modifié l'exemple pour garder une cohérence en français.

ments sémantiques, c'est-à-dire des concepts et des relations⁸ s'appelle enrichissement d'une ontologie.

Le peuplement d'une ontologie consiste à ajouter des instances à des concepts existants (Petasis *et al.*, 2011). Toutefois, il faut également prendre en compte l'ajout des nouveaux libellés, donc des termes, et de leurs traductions.

Le terme élargissement, venu de la terminologie de l'entreprise Rebusz, englobe l'ajout des nouveaux éléments lexicaux, qu'il s'agisse des instances ou des termes, à des concepts existants. Si l'élargissement est effectué par un expert humain, la décision d'ajouter un élément, instance ou terme, lui appartient. Elle peut alors être prise tenant compte des besoins du projet en question.

Les processus d'enrichissement et d'élargissement sont étroitement liés et il est difficile de définir une limite qui les sépare. En effet, ces processus partagent les mêmes étapes de traitement et la différence entre eux se manifeste au moment de la validation des résultats. C'est à ce moment qu'un expert humain décide de créer un nouveau concept ou d'ajouter un terme ou une entité à un concept existant.

Étapes d'enrichissement et d'enrichissement multilingue

L'enrichissement et l'élargissement d'une ontologie multilingue comprennent plusieurs étapes et chaque étape nécessite des outils et des ressources, y compris des ressources humaines.

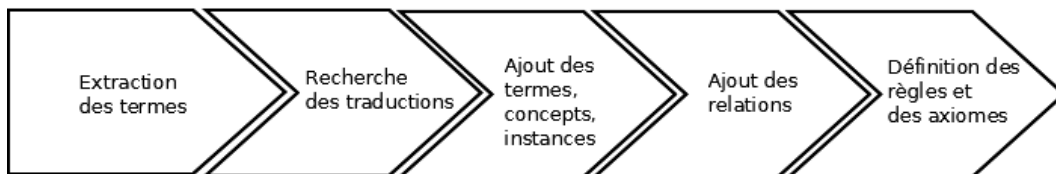


FIGURE 3 – Étapes de traitement

La Figure 3 présente les étapes d'enrichissement et d'élargissement d'une ontologie à partir de corpus. La première étape consiste à extraire des termes à partir d'un corpus. Ensuite, il est nécessaire d'identifier leurs traductions en d'autres langues. Lors de la troisième étape, l'expert humain exploite les termes multilingues pour créer des nouveaux concepts ou pour ajouter des termes et des instances. Ensuite, il peut lier des concepts par des relations sémantiques et, si l'ontologie contient des niveaux d'abstraction avancés, ajouter des contraintes logiques.

Dans le cadre de cette thèse, nous traitons uniquement les trois premières étapes du processus. Ces traitements nécessitent des ressources et des outils spécifiques

8. L'extraction et l'ajout des relations sémantiques ne sont pas traités dans le cadre de cette thèse.

qui constituent, comme nous allons le montrer pour chaque étape, des limites des approches existantes.

Extraction des termes

L'extraction des termes repose sur des corpus de textes spécialisés. Ces corpus doivent être de bonne qualité, c'est-à-dire rédigés par des experts du domaine en question, et en quantité suffisante pour effectuer des traitements statistiques. Cependant, en fonction du domaine et de la langue, ces corpus peuvent être difficiles à constituer. Cette contrainte est prise en compte depuis le début de notre projet qui vise à extraire des termes sans dépendre de la taille de corpus. De plus, les traitements statistiques se focalisent sur le repérage des éléments récurrents, ce qui empêche l'identification des termes qui viennent d'apparaître (les néologismes), car le nombre de leurs occurrences dans le corpus peut être assez faible. Si notre méthode doit fonctionner sur des corpus indépendamment de leur taille, il devient important de repérer ces termes rares.

L'extraction terminologique peut également recourir à des lexique du domaine pour identifier des termes composés. Par exemple, si le terme *eau* est présent dans un lexique relatif à l'environnement, il peut aider à identifier le terme *eaux usées* dans le corpus. Dans certains cas, les lexiques peuvent être substitués par des listes de formants néo-classiques (Estopà *et al.*, 2000b), c'est-à-dire des morphèmes d'origine grecque ou latine qui font partie des nomenclatures en médecine, biologie et d'autres domaines. Ainsi, si une telle liste contient le morphème *-chlor-*, il peut servir à la découverte des termes comme *chlorinité* ou encore *chlorure*. Or, les ressources lexicales ou morphologiques ne sont pas toujours disponibles. De plus, il est nécessaire d'avoir une version des ressources par langue de traitement. Un des objectifs de notre projet est de substituer ces ressources extérieures par les informations déjà disponibles dans l'ontologie et le corpus spécialisé.

L'extraction des termes nécessite également un étiqueteur morphosyntaxique qui associe les mots d'une phrase avec leurs parties du discours, par exemple, *installation* (NOM) *de* (PREP⁹) *osmose* (NOM) *inverse* (ADJ¹⁰). Il est alors possible d'utiliser des patrons prédéfinis (Daille, 2003; Bourigault et Fabre, 2000; Frantzi *et al.*, 2000) permettant d'extraire des groupes nominaux : NOM-PREP-NOM (*installation d'osmose*), NOM-ADJ (*osmose inverse*), etc. Toutefois, les étiqueteurs morphosyntaxiques ne sont pas disponibles pour toutes les langues et dépendent du lexique injecté. De plus, ces outils introduisent un risque d'erreur d'étiquetage sur des textes de spécialité, car les termes du domaine en question peuvent être inconnus du lexique utilisé dans l'outil, ce qui influence les résultats de l'extraction terminologique. Nous avons identifié la méthode de Vergne (2003, 2005) qui propose un étiquetage de corpus par mots informatifs, donc des termes potentiels, et vides (des articles, des

9. Préposition.

10. Adjectif.

prépositions, etc.) sans recours à des ressources extérieures. Les traitements qui ne nécessitent pas de ressources complémentaires s'appellent endogènes et présentent un grand intérêt pour notre projet.

Recherche des traductions

La recherche des traductions est souvent conduite sur des corpus multilingues parallèles où chaque phrase source est alignée avec sa traduction en langue cible. L'alignement des termes est alors effectué à l'aide d'un aligneur statistique, outil qui met en correspondance des mots en langue source et cible présents dans un grand nombre de segments parallèles. Toutefois, ces corpus sont rares en ne couvrant pas l'ensemble des domaines, ni toutes les paires de langues.

On recourt alors à des corpus comparables qui contiennent des textes portant sur un même domaine et sujet, sans que ces textes soient des traductions. Ce type de corpus est plus facile à constituer. L'absence de lien de traduction complique la tâche d'alignement, ce qui nécessite des ressources complémentaires. Ainsi, des dictionnaires multilingues d'un domaine peuvent servir à l'identification semi-distributionnelle ou compositionnelle des traductions (Cap *et al.*, 2014; Fraser *et al.*, 2012). Par exemple, si un dictionnaire franco-allemand contient des traductions *installation* – *Anlage* et *inverse* – *umkehr*, il est possible de suggérer la traduction allemande *Umkehrosmoseanlage* pour le terme français *installation d'osmose inverse*. Il est également possible d'exploiter des alignements au niveau des morphèmes (Delpech *et al.*, 2012; Weller *et al.*, 2011). En effet, si le préfixe anglais *post-* est aligné avec le mot français *après*, cette information peut être utile pour proposer la traduction française *après-ménopause* pour le terme anglais *post-menopause*. La disponibilité de ces ressources dépend également des domaines et des langues. Notre réflexion porte sur les façons de constituer ces alignements lexicaux et morphologiques en utilisant les informations disponibles dans l'ontologie traitée, sans avoir recours à des dictionnaires complémentaires.

Ajout des concepts, des termes et des instances

Les corpus spécialisés peuvent également servir à la construction des taxonomies qui aident à alimenter une ontologie. Une autre possibilité est d'utiliser des ressources connectées libres d'accès *Linked Open Data (LOD)*¹¹ qui contiennent des ressources sémantiques spécialisées ou générales (notamment, DBpedia¹² dérivée de Wikipédia¹³) publiées sur Internet. Chacune de ces ressources suit les conventions de format commun au LOD et possède un identifiant uniforme de ressource (*Uniform Resource Identifier* ou URI), une chaîne de caractères unique qui permet d'identifier

11. <http://linkeddata.org/>

12. <http://wiki.dbpedia.org/>

13. www.fr.wikipedia.org

une ressource de façon non-ambiguë et lisible pour une machine. Des concepts faisant partie du LOD sont également accessibles via leurs URIs et contiennent des renvois à des concepts équivalents ou plus spécifiques dans d'autres ressources. De cette façon, si un concept, par exemple *osmose inverse* d'une ressource est présent dans une autre ressource, il est possible d'exploiter l'ensemble des informations concernant ce concept disponibles dans les deux ressources. Toutefois, cette démarche nécessite une disponibilité des données pour le domaine en question. Nous avons choisi de ne pas exploiter ces ressources, car il est difficile d'assurer un traitement de qualité à partir des ressources hétérogènes et difficiles à maîtriser.

La complexité de mise en place de ces traitements et l'absence de garantie de résultats satisfaisants favorisent le recours à des experts humains. Leur travail nécessite toutefois du temps, des outils et des ressources financières. De plus, l'aspect multilingue peut compliquer la communication entre des équipes d'experts. En effet, il est préférable de confier la création des libellés à des experts en fonction de leurs langues maternelles. Toutefois, la coordination d'une telle équipe nécessiterait une maîtrise d'une langue commune par tous les membres. Cet objectif est difficile à atteindre pour plusieurs raisons, dont la disponibilité des experts et les coûts liés à la constitution d'une telle équipe. Notre projet a pour vocation de préparer des suggestions de termes, de leurs traductions et de leurs regroupements pour faciliter leur ajout ultérieur dans l'ontologie par un expert humain.

L'analyse plus approfondie de chaque étape est présente dans les chapitres de la Partie I. L'ensemble des contraintes liées à des différents traitements rendent l'enrichissement et l'élargissement difficiles et coûteux sont au cœur de notre problématique.

Problématique

La problématique de notre projet est liée aux limites des traitements actuels que nous avons décrits ci-dessus.

Premièrement, la disponibilité des données et des outils dépend de la langue et du domaine visé. Les approches existantes d'extraction terminologique, ainsi que d'autres étapes d'élargissement et d'enrichissement, sont souvent développées pour l'anglais et, au mieux, pour quelques langues européennes. Elles dépendent des outils, comme des étiqueteurs morphosyntaxiques, et s'adaptent difficilement à des langues pour lesquelles il existe moins de ressources. La création d'une méthode multilingue et multidomaine nécessite une minimisation des apports en ressources et en outils, car chaque nouvel apport signifie une contrainte potentielle pour une ou plusieurs langues ou domaines.

Deuxièmement, les domaines évoluent et il est nécessaire de repérer des nouveaux termes et concepts. Ces nouveaux éléments peuvent avoir peu d'occurrences dans les corpus (textes de spécialité) et ne sont pas identifiés par des méthodes statistiques. Il

est alors souhaitable de combiner des calculs statistiques avec d'autres types de traitements. Par exemple, ces nouveaux termes peuvent avoir des liens morphologiques avec des termes connus (par exemple, être dérivés de ceux-ci ou contenir une même base). Le défi consiste à repérer ces liens morphologiques sans outils spécifiques, ni ressources complémentaires.

Troisièmement, les connaissances peuvent être exprimées de manières différentes dans le corpus. La synonymie, les variantes orthographiques et les fautes de frappe compliquent la tâche d'extraction terminologique à partir de corpus. Repérer ces variations et les ajouter dans l'ontologie permettra ultérieurement à un système de recherche d'information d'analyser plus de documents. En effet, si les corpus peuvent contenir des imperfections et des variations, la recherche d'information doit être capable de traiter ces corpus. Toutefois, afin de maintenir la qualité de l'ontologie, il faut indiquer qu'un libellé est une variante ou un terme mal orthographié.

Quatrièmement, dans un corpus multilingue, le même concept peut s'exprimer par plusieurs termes en plusieurs langues. Identifier les traductions des termes à partir de corpus est une tâche difficile. D'une part, les corpus parallèles sont rares et, d'autre part, les traitements sur des corpus comparables nécessitent d'utiliser des ressources complémentaires. Afin de rendre nos traitements vraiment multilingues, nous visons à utiliser pour cette tâche uniquement des outils indépendants de la langue et éviter des ressources autres que l'ontologie source et le corpus.

Enfin, il est difficile d'organiser les connaissances identifiées dans le corpus. Traditionnellement, l'extraction terminologique fournit des résultats organisés sous forme de liste. Toutefois, une telle représentation n'est pas adaptée au travail avec une ontologie, car elle ne met pas en évidence des liens existant entre les termes, notamment au niveau morphologique. Tout comme dans le cas de l'extraction des termes à faibles fréquences, la difficulté du regroupement des termes morphologiquement proches consiste à éviter une analyse morphologique fine et des ressources extérieures.

À cette problématique s'ajoute la question de l'applicabilité de la méthode dans le contexte industriel. Notamment, il faut tenir compte des éventuelles évolutions de la structure d'une ontologie et donc rendre les traitements indépendants de celle-ci. Un autre point important concerne les coûts engendrés par la méthode. En plus des coûts de licences, chaque nouvelle ressource ou outil entraîne des modifications dans le programme (gestion des entrées et des sorties, du format, influence sur la précision globale, etc.), ce qui impacte le temps de travail du personnel. Ainsi, l'usage des outils et des ressources extérieurs doit être minimisé.

Compte tenu des difficultés énumérées ci-dessus, il est nécessaire de développer une méthode robuste (peu dépendante des données et des outils) et multilingue basée sur des corpus comparables.

Hypothèses

Les problèmes évoqués précédemment sont en grande partie dûs à l'utilisation d'informations linguistiques fines et à la nécessité de grands volumes de données. Or, l'ontologie source et le corpus d'un domaine contiennent déjà des connaissances sur le domaine en question et définissent un contexte suffisant pour le traitement de la terminologie.

Selon notre hypothèse, il est possible de remplir notre objectif en utilisant les textes bruts sans informations linguistiques avancées. Les ressources morphologiques, lexicales et morphosyntaxiques peuvent être substituées par des informations générées en cours d'analyse.

Ces informations peuvent notamment être générées grâce aux n -grammes de caractères, séquences de caractères d'une longueur n fixée par l'utilisateur. Par exemple, si $n = 4$, le mot *osmose* peut être segmenté en *osmo*, *smos* et *mose*. Situés au niveau sous-lexical, ces séquences de caractères peuvent correspondre à des morphèmes, voire à des mots faisant partie d'un mot composé. Nous étudions donc leur capacité de remplacer des ressources traditionnelles.

D'autre part, les connaissances de l'expert sont représentées dans l'ontologie du domaine et il est possible d'utiliser ces connaissances pour en découvrir des nouvelles.

Notre approche repose donc sur une ontologie et un corpus multilingue à partir desquelles nous allons apprendre les ressources morphologiques et morphosyntaxiques nécessaires pour appliquer la chaîne de traitements.

Contributions

L'apport principal de notre travail est le développement d'une méthode robuste et multilingue utilisant des corpus comparables et basée sur des ressources endogènes. Ces ressources sont appelées endogènes, car elles sont apprises à partir du corpus et de l'ontologie en cours d'analyse. La robustesse de notre méthode se justifie par sa faible sensibilité au volume de données, tandis que l'aspect multilingue se traduit par l'utilisation des outils indépendants de la langue.

Parmi les avantages des ressources endogènes, nous pouvons énumérer leur disponibilité, ce qui se traduit par un coût faible de l'application, et leur relative indépendance vis-à-vis à la langue traitée. Nous étudions leur applicabilité tout au long du processus de l'élargissement et de l'enrichissement des ontologies légères, ce qui résulte en trois contributions correspondant aux niveaux de traitements.

1. Extraction terminologique à partir de ressources endogènes

Dans le cadre de l'extraction terminologique, les termes présents dans l'ontologie source servent à apprendre des ressources morphologiques et morphosyntaxiques endogènes. Selon Vergne (2003), un outil terminologique endogène n'utilise aucune ressource à l'exception du corpus à traiter. Dans notre système, les ressources disponibles sont le corpus et l'ontologie.

Notre méthode exploite des n-grammes de caractères extraits au début ou au milieu des mots pour apprendre des termes candidats monolexicaux. Ces n-grammes de caractères s'approchent des préfixes et des radicaux, et peuvent remplacer entre autres des listes de morphèmes d'origine grecque ou latine utilisées pour identifier des termes candidats.

Ensuite, les n-grammes de caractères de la fin de mot sont utilisés pour former des patrons morphosyntaxiques endogènes dans le but d'apprendre les termes candidats polylexicaux. Les termes de l'ontologie peuvent être complétés par les termes candidats les plus fréquents et les plus représentatifs du corpus.

Notre approche permet aussi d'identifier les termes candidats à faible fréquence et est peu sensible à la taille de corpus. Elle facilite la maintenance de l'ontologie, car permet d'intégrer les nouveaux termes (des néologismes) dès leur apparition.

2. Construction des ressources endogènes pour la recherche des traductions

Notre méthode de recherche des traductions est applicable à des corpus multilingues parallèles ou comparables. Nous utilisons les traductions disponibles dans l'ontologie source pour apprendre des alignements au niveau des n-grammes de caractères en utilisant l'aligneur statistique Anymalign (Lardilleux et Lepage, 2009b, 2008; Lardilleux *et al.*, 2013).

Dans le cas où l'ontologie source ne contient pas de libellés en plusieurs langues, une ressource terminologique multilingue externe peut être utilisée. Les alignements de n-grammes les plus fiables sont ensuite utilisés comme modèle pour identifier les traductions possibles entre les termes candidats extraits à partir du corpus en deux langues.

3. Enrichissement et élargissement de l'ontologie à l'aide des familles morphologiques endogènes

La représentation des résultats sous forme de liste de termes candidats ne suffit pas pour la tâche d'alimentation d'une ontologie. Les listes de candidats, même triées selon des différents paramètres, ne regroupent pas des termes morphologiquement proches.

Nous avons donc opté pour les familles morphologiques, des regroupements de mots partageant un ou plusieurs éléments qui peuvent être, en fonction de l’approche, des morphèmes (Bernhard, 2006) ou des séquences de caractères (Schone et Jurafsky, 2000). Cette représentation nous permet de proposer des candidats par rapports aux termes de l’ontologie. Par exemple, la famille du terme *aérobie* contient les candidats *aérosol*, *aérodynamique*, *anaérobie*, *aérien* et *aération*.

Les familles morphologiques endogènes sont construites en exploitant des n-grammes de caractères représentatifs du domaine appris lors des étapes précédentes. Mono ou multilingues, elles représentent un potentiel important pour l’enrichissement et l’élargissement d’une ontologie.

Nous allons retrouver ces contributions dans l’organisation de cette thèse.

Organisation du manuscrit

La présente thèse se compose de deux parties. La première est consacrée à l’état de l’art relatif à l’élargissement et à l’enrichissement des ontologies. La seconde présente notre méthode à base de ressources morphologiques endogènes.

Dans la première partie, nous étudions notamment la notion du terme, de la terminologie et les méthodes d’extraction de termes à partir de corpus. En effet, l’extraction terminologique constitue une des étapes-clés de la création et de l’alimentation des ontologies. De la notion de terme, nous passons ensuite à celle de concept, de relation sémantique et d’ontologie. L’analyse des approches existantes nous permet de constituer une base théorique du processus d’élargissement et d’enrichissement des ontologies.

La deuxième partie débute par un chapitre portant sur les ressources utilisées dans notre projet et les traitements préalables effectués notamment sur les corpus. Ensuite, nous reprenons nos trois contributions décrites précédemment dans l’ordre de traitement, à savoir l’extraction terminologique suivie de la recherche de traductions et de l’organisation des résultats dans des familles morphologiques. Chaque étape est décrite et évaluée séparément afin de fournir des résultats objectifs non-biaisés par les étapes précédentes. Enfin, notre système est évalué dans son intégralité sur une ontologie et un corpus complet, ce qui permet d’estimer ses performances dans des conditions réelles.

Pour finir, une conclusion permet de rappeler les objectifs fixés dans notre projet et de positionner les résultats obtenus par rapport à ces objectifs. De plus, notre domaine de recherche étant à la fois vaste et dynamique, la conclusion présente les perspectives de recherche qui découlent de notre méthode et de nos résultats.

Première partie

Processus d'élargissement et d'enrichissement d'une ontologie

La première partie de cette thèse est consacrée aux notions principales et à l'état de l'art lié à l'enrichissement et à l'élargissement multilingue d'une ontologie spécialisée. Rappelons que notre travail se situe dans le contexte de la recherche d'information multilingue qui consiste à identifier des documents pertinents en plusieurs langues par rapport à une requête émise par un utilisateur. L'ontologie est alors considérée comme une ressource spécialisée multilingue permettant d'interpréter la requête au niveau sémantique et effectuer la recherche en tenant compte de la synonymie, des instances, des termes et de leurs traductions en d'autres langues.

Afin d'assurer une bonne performance de la recherche d'information, doivent non seulement être complètes au moment de leur création, mais aussi suivre l'évolution permanente des domaines de spécialité. Nous pouvons alors parler du cycle de vie d'une ontologie qui comprend plusieurs étapes, de sa création à la limite de son application. L'enrichissement et l'élargissement d'une ontologie font partie de son cycle de vie et permettent de la maintenir à jour par rapport au domaine en question.

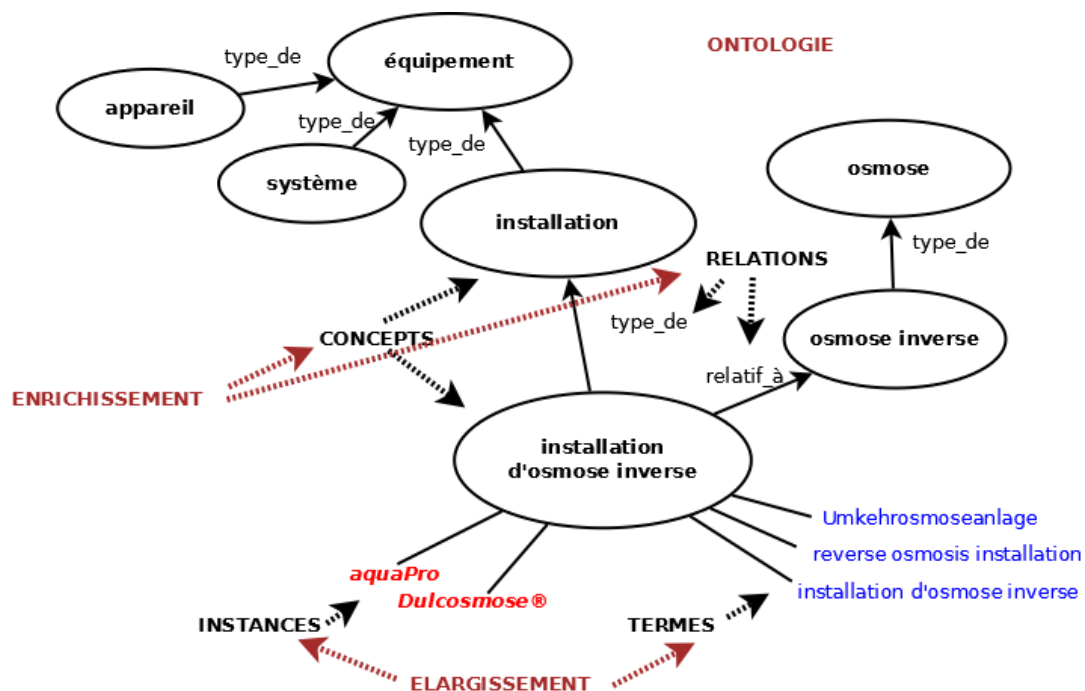


FIGURE 4 – Enrichissement et élargissement d'une ontologie

La Figure 4 illustre la portée de chaque processus : l'enrichissement d'une ontologie consiste en ajout des nouveaux concepts et des nouvelles relations, tandis que l'élargissement englobe l'ajout des instances, donc le peuplement d'une ontologie, et des termes à des concepts existants. Ces étapes sont cruciales pour améliorer la performance d'un système qui utilise l'ontologie en question ou pour prolonger la durée de vie de la ressource. Ces étapes diffèrent de l'apprentissage de l'ontologie, car la structure initiale existe déjà et impose certaines contraintes.

Le processus d'élargissement et d'enrichissement n'est pas linéaire, mais cyclique. Il doit être conduit en permanence pour assurer la complétude et la qualité d'une ressource.

Un des moyens pour suivre l'évolution d'un domaine consiste à analyser des textes de spécialité récents. Il s'agit donc d'exploiter des corpus spécialisés pour identifier des termes qui sont absents dans l'ontologie du domaine traité. L'extraction de terminologie multilingue constitue ainsi la base de l'élargissement et de l'enrichissement d'une ontologie. Le premier chapitre de cette thèse a pour objectif de définir le statut du terme et fournit un panorama des approches d'extraction terminologique multilingue.

Les résultats d'extraction terminologique peuvent être ensuite analysés en vue d'ajout des nouveaux éléments lexicaux (des libellés ou des instances) ou sémantiques (des concepts) dans l'ontologie cible. Le deuxième chapitre se concentre alors sur les notions de concept, de relation et d'ontologie, ainsi que sur les méthodes de son élargissement et de son enrichissement.

Chapitre 1

Extraction terminologique

L'extraction terminologique a pour objectif d'identifier les termes d'un domaine dans le corpus spécialisé, constituant ainsi la base pour l'élargissement et l'enrichissement d'une ontologie. C'est un processus dont la complexité varie en fonction de différents paramètres, parmi lesquels avant tout :

- l'application à une ou à plusieurs langues ;
- l'utilisation de corpus parallèles ou comparables ;
- le volume de données nécessaire.

Nous pouvons également ajouter des critères secondaires, comme la prise en charge de termes polylexicaux, composés de plusieurs mots, ou la capacité de repérage de candidats à faible fréquence (dont des néologismes du domaine en question).

Ce chapitre débute par l'exposé de notions théoriques liées à la notion de terme et à la terminologie. Il se poursuit avec l'analyse des approches existantes pour l'extraction terminologique monolingue. Ensuite, nous examinons les méthodes existantes permettant une recherche des traductions pour les termes candidats extraits, ainsi que des néologismes.

1.1 Termes et terminologie

La terminologie est la clé qui ouvre l'accès aux connaissances d'un domaine. Pour cette raison, elle est nécessairement présente dans les ontologies spécialisées. Les connaissances d'un domaine sont exprimées dans le corpus spécialisé à travers la terminologie. Même si les termes doivent transmettre ces connaissances sans ambiguïté, le terme *terminologie* est lui même polysémique. Ici nous nous concentrons sur les deux sens qui sont en relation au sujet de la recherche :

1. la terminologie comme l'ensemble des termes qui représentent le système de notions d'un domaine particulier (ISO 1087) ;
2. la terminologie comme l'étude systémique de la dénomination des notions appartenant à des domaines spécialisés de l'expérience humaine et considérées

dans leur fonctionnement social¹.

Dans le premier sens du terme, une terminologie est un ensemble de désignations (termes) dont le champ d'utilisation (l'extension) est délimité ou, au moins, limité et spécifique. Cette délimitation peut porter un caractère arbitraire (Gouadec, 1990).

Dans le second sens, la terminologie étudie les aspects théoriques et pratiques de la constitution, gestion, organisation, évolution de terminologies de domaines de spécialité.

Dans les deux cas, le terme est l'unité de base opérée par la terminologie. Comme nous le démontrons ci-dessous, il n'existe pas une seule notion de terme qui puisse satisfaire toutes les théories et convenir à tous les usages. Il revient à chaque projet de choisir une définition qui convient au mieux à ses objectifs.

1.1.1 Le terme dans la linguistique

Trouver une définition cohérente aide à identifier les traits de termes qui facilitent leurs extraction à partir du corpus. Pour cela, nous allons examiner les deux grandes tendances dans la définition du terme (L'Homme, 2005) : le terme comme réalisation d'un concept et le terme comme signe linguistique.

1.1.1.1 Le terme désigne un concept

La définition de la terminologie par le *Le grand dictionnaire terminologique* (GDT) suggère de définir les termes comme les dénominations des notions (concepts²) d'un domaine de spécialité. Elle s'inscrit dans la Théorie Générale de la Terminologie (TGT, fondée par Eugen Wüster, école de Vienne). On peut citer d'autres définitions qui s'inscrivent dans le même cadre :

- « Le terme est une désignation verbale d'un concept général dans un domaine spécifique » (ISO 1087-1 :2000, dans (Van Campenhoudt, 2010)).
- « Le terme est [...] une unité syntaxique qui peut être composée d'un ou plusieurs mots mais qui est relative à un domaine spécifique dont elle décrit une notion. Les termes ont généralement un sens précis dans le domaine auquel ils appartiennent, ils sont moins ambigus que les mots courants » (Omrane *et al.*, 2011a).
- « Les termes sont des objets linguistiques utilisés dans la littérature technique et scientifique et visant à faire référence à des concepts de façon non ambiguë et consensuelle » (Jacquemin *et al.*, 1997).
- « Le terme est l'unité de désignation d'un concept ou d'un objet ou d'un processus de la réalité perçue ou conçue » (Gouadec, 1990).

1. Le grand dictionnaire terminologique (GDT), <http://www.granddictionnaire.com/>

2. Nous allons traiter la notion de concept dans le chapitre suivant.

Si le terme est une étiquette d'un concept comme le prescrit la TGT, l'extraction automatique doit tenir compte de facteurs, comme la définition du concept général et la place de ce concept dans la structure conceptuelle générale et dans celle spécifique à un domaine.

En effet, le concept doit être défini au préalable, ce qui n'est pas toujours le cas dans des domaines en innovation et en développement. Il doit prendre sa place dans une ontologie générale (vu que la terminologie fait partie de la connaissance universelle) et dans une ontologie du domaine. Défini une fois et de manière claire, il prend sa place et établit ses relations avec d'autres concepts selon une logique stricte et des preuves suffisantes.

L'apparition de nouveaux termes est alors difficile à gérer, car il faut d'abord identifier ou créer un concept, trouver sa place dans l'ontologie et s'assurer que ce terme soit une étiquette unique, standardisée et non-ambiguë du concept. La gestion de la synonymie, de la variation lexicale et morphosyntaxique et des paraphrases est également compliquée par les exigences envers la précision sémantique du terme.

1.1.1.2 Le terme est un signe linguistique, le résultat d'une association entre un sens et une forme

Les difficultés imposées par la TGT ont conduit à des approches alternatives (Cabré Castellví, 2003; L'Homme, 2005; L'Homme et Polguère, 2008; Van Campenhoudt, 2010) qui rejettent complètement ou partiellement la nature conceptuelle du terme :

- « Le terme est une unité de connaissance spécialisée » (Cabré Castellví, 2003).
- « Le terme est une unité lexicale monosémique propre à un domaine de spécialité. Cette unité peut être simple ou complexe » (Van Campenhoudt, 2010).

Selon Van Campenhoudt (2010), « il n'est nul besoin de présupposer l'existence ou la nécessité d'un concept universel pour justifier de l'existence du terme. Loin d'être une étiquette apposée sur un référent, il est une manière de dire et d'exprimer au mieux la connaissance, dans une forme récursive et souvent transparente ».

L'Homme et Polguère (2008), à leur tour, cherchent à démontrer que les termes n'ont pas de nature linguistique propre comparés à des unités lexicales générales (ULG), c'est-à-dire, les mots de la langue générale par opposition à celle de spécialité. Pour cela, ils avancent deux postulats théoriques :

1. Il n'existe pas de différence de nature fondamentale entre termes et ULG pour ce qui est de leurs propriétés structurales centrales, qui sont :
 - leur sémantisme (= leur définition) ;
 - leur forme (signifiants linguistiques associés) ;
 - leur combinatoire lexicale et grammaticale.
2. Les termes sont des unités lexicales particulières. Ils se distinguent des ULG — les unités lexicales non marquées — uniquement par le fait que leur maîtrise

1.1. TERMES ET TERMINOLOGIE

linguistique est associée à la maîtrise d'un domaine de connaissance donné (scientifique, technique, etc.).

Ces principes ont une influence directe sur la façon d'organiser les ressources. Ainsi, Van Campenhoudt (2010) insiste sur le besoin d'un « dégroupement homonymique » des entrées monolexicales dans les dictionnaires qui était jadis fondé par le souci d'économie d'espace sur le support papier. Quant aux syntagmes terminologiques non lexicalisés, L'Homme et Polguère (2008) se réfèrent à la pratique des dictionnaires de langue générale où les collocations sont fournies sous l'entrée du terme phare.

De plus, L'Homme et Polguère (2008), Bernier-Colborne et L'Homme (2015) et L'Homme (2016) remettent en question l'approche de l'extraction terminologique fondée sur les patrons morphosyntaxiques nominaux. Ils considèrent que leur portée doit s'étendre également sur les verbes, adjectifs et adverbes. Cette approche est en lien avec la théorie de la terminologie basée sur les schémas sémantiques (Faber *et al.*, 2005, 2006, 2009; Faber, 2011; Castro *et al.*, 2014; Bernier-Colborne et L'Homme, 2015) d'après les schémas sémantiques de Fillmore (1982, 1985) où le verbe prend la place centrale dans la représentation des processus propres au domaine de spécialité en question.

Cependant, définir le terme comme une unité lexicale monosémique propre à un domaine de spécialité ou, de manière plus large, comme une unité de connaissance spécialisée, n'apporte pas de solution complète.

Premièrement, si la référence à un domaine de spécialité est la seule caractéristique du terme comme unité lexicale, le terme doit être décrit dans chaque domaine de spécialité à part, même si son sens est identique. Dans le cas contraire, il est nécessaire d'explicitier les autres particularités de ces unités de connaissance afin de les rendre utilisables pour le traitement automatique.

De plus, si le terme se réfère à un domaine de spécialité, comment décrire ce domaine de manière suffisante et non-ambiguë de façon à couvrir toute variation du sens du terme? Si l'on considère qu'il existe une taxonomie de domaines de spécialité où les domaines plus généraux se divisent en sous-domaines allant jusqu'à une précision maximale, deux questions se posent :

- Comment refléter les spécifications possibles du sens au fur et à mesure de la descente dans cette taxonomie?
- Quel est le rapport entre les domaines de connaissances qui se trouvent au même niveau dans la taxonomie et partagent une partie de la terminologie?

Deuxièmement, si le terme est toujours monosémique, les variations du sens induites par ses contextes mènent à la création de nouveaux termes. Dans le cas contraire, le sens du terme n'est pas une valeur fixe et indissociable. La synonymie est alors un lien qui s'établit au niveau d'un sens du terme dans un contexte particulier. Dans les deux cas, un système de classification de termes (unités lexicales) est nécessaire.

Dans cette sous-section, nous avons montré les deux grandes tendances dans la définition de terme du point de vue linguistique. D'un côté, la nature conceptuelle de termes permet d'imaginer une organisation de connaissances allant des notions universelles vers celles plus spécialisées. Cette même nature impose toutefois des restrictions préalables à la création des nouveaux termes. De l'autre côté, le regard sur le terme en tant que signe linguistique rend la gestion des terminologies plus flexible. C'est alors l'usage spécialisé qui définit une unité lexicale en tant que terme. Nous verrons par la suite que le traitement automatique des langues (TAL) adopte une approche pragmatique et fusionne en quelque sorte les deux définitions.

1.1.2 Le terme dans le TAL

Dans le domaine du traitement automatique de langues (TAL), la définition du terme est étroitement liée avec l'usage que l'on en fait. De plus, le recours à des traitements automatiques nécessite des précautions complémentaires, ce qui mène vers l'adoption de la notion de terme candidat. Ces notions se positionnent à l'écart du débat théorique présenté ci-dessus qui donne peu de réponses quant à solutions pratiques. En même temps, une certaine liberté dans la définition du terme favorise la diversité des ressources terminologiques qui varient en forme, en contenu, en organisation et en applications. La définition du terme est surtout utile à l'étape de validation des résultats, ce que nous allons illustrer par les critères de jugement des termes candidats issues d'un projet européen. Nous allons également fournir une typologie de termes réunissant les critères utiles pour notre projet.

1.1.2.1 Vers un choix pragmatique

En effet, le choix en faveur d'une structure d'organisation de la terminologie est fait pour chaque projet en fonction des besoins rencontrés. Ces structures peuvent être très différentes : bases de données terminologiques, thésaurus, dictionnaires mono ou multilingues, glossaires, etc. De plus, malgré les normes et les conventions, il n'existe pas de solution idéale qui saurait satisfaire tout projet.

Évidemment, dans le cadre de notre projet, nous allons opter pour une définition du terme en relation avec l'ontologie. Nous préférons la définition donnée par la terminologie basée sur des schémas sémantiques (*Frame Based Terminology*) de Faber *et al.* (2005, 2006, 2009); Faber (2011) : le terme est un mot simple ou composé qui sert à dénommer un concept. Toutefois, le terme n'est pas égal à un concept, car le concept peut regrouper plusieurs dénominations (libellés) qui se réfèrent à une notion. Ces dénominations peuvent être des synonymes (par exemple, dans l'ontologie EcoLexicon³ qui est développée dans le contexte de la terminologie basée sur des schémas sémantiques, les termes anglais *rain* et *rainfall* se réfèrent au concept

3. <http://ecolexicon.ugr.es/en/index.htm>

1.1. TERMES ET TERMINOLOGIE

*RAIN*⁴) ou encore des traductions, comme c'est le cas du terme espagnol *lluvia* qui désigne le même concept. Il faut remarquer qu'il s'agit d'une ontologie spécialisée dans les domaines de l'environnement (écologie, agriculture, etc.). Le terme est donc un mot simple ou composé qui sert à dénommer un concept dans un ou plusieurs domaines de spécialité.

1.1.2.2 Notion de terme candidat

La notion de terme candidat est utilisée en extraction terminologique pour désigner les unités lexicales repérées par les systèmes automatiques (Drouin et Langlais, 2006). Cette notion sécurise le travail avec les systèmes d'extraction terminologique, car le fait d'insister sur le caractère incomplet de résultats du traitement incite à valider ces résultats avant de les réutiliser.

Cette notion est liée à celle du potentiel terminologique (*termhood*), comme l'indiquent Drouin et Langlais (2006). Ce terme désigne l'ensemble de critères qui peuvent être mis en œuvre afin de filtrer la liste des termes candidats. Ces critères peuvent s'appuyer aussi bien sur des mesures statistiques que sur des indices morphosyntaxiques, ou encore sur le contexte.

Dans cette thèse, nous allons utiliser des notions de termes candidats, de traductions candidates et de concepts candidats pour désigner les unités extraites automatiquement et en attente de validation par un expert humain. Cette validation nécessite des critères selon lesquels un terme candidat est évalué. En absence de tels critères, toute validation est subjective, ce qui ne convient pas pour le travail sur des ressources terminologiques. Nous allons maintenant analyser les critères proposés par un projet d'extraction terminologique où la validation a également été faite manuellement.

1.1.2.3 Critères de jugement des termes candidats du projet TTC

Le projet TTC est un projet européen⁵ dont un des objectifs était de créer un outil d'extraction terminologique multilingue, TermSuite (Rocheteau et Daille, 2011; Cram et Daille, 2016; Daille *et al.*, 2011). Ce projet a réuni plusieurs partenaires, tels que l'Université de Nantes (laboratoire LINA), l'Université de Stuttgart, l'Université de Leeds, ainsi que les partenaires industriels SOGITEC, SYLLABS, TILDE et EURRINOV. La collaboration au sein du projet portait sur des problématiques d'actualité, comme l'utilisation des corpus comparables, l'extraction et la gestion des terminologies, mais aussi sur des outils d'aide à la traduction.

Dans le cadre de ce projet, des listes de termes ont été créées à partir des corpus du projet pour l'évaluation du système d'extraction terminologique. L'implication

4. <http://ecolexicon.ugr.es/visual/statistics.html>

5. <http://www.ttc-project.eu/>

de plusieurs partenaires nécessitait un guide précis pour cette tâche (Syllabs, 2012). Ce guide décrit les critères qui doivent être pris en compte dans la validation des termes candidats. Nous adoptons les lignes directrices de ce guide dans notre projet. De plus, TTC TermSuite est utilisé en tant qu'un outil de référence pour certaines étapes de notre système, car il permet d'extraire des termes et leurs traductions à partir des corpus comparables et fonctionne pour nos langues d'évaluation (le français, l'anglais et l'allemand).

Selon ce guide de validation (Syllabs, 2012), un terme candidat monolexical est valide :

- s'il est dérivé d'un terme (p. ex. : *rotor* – *rotorique*, *rotation* ; *pompe* – *pompage*) ;
- s'il existe un lien paradigmatique avec un terme du domaine, c'est-à-dire que le candidat apparaît dans la définition d'un terme (comme *rotor* qui figure dans la définition de *hélice* dans le Grand Dictionnaire Terminologique) ;
- si les cooccurrences du candidat diffèrent de celle identifiées dans le corpus de langue générale (la collocation *arbre lent* exclut l'interprétation non terminologique de ses composants) ;
- s'il s'agit d'un emprunt ou d'un terme en langue étrangère.

Le guide distingue les termes et leurs variantes, qui peuvent être orthographiques, par exemple, *mineralisation* – *mineralization* (EN), flexionnelles ou dérivationnelles (*Atmosphäre* – *Atmosphären* en allemand). Pour les termes polylexicaux, le guide (Syllabs, 2012) préconise deux principes : dans le cas où le choix se situe entre deux candidats polylexicaux où l'un inclut l'autre (p. ex. : *énergie éolienne* et *production de l'énergie éolienne*), le terme polylexical le plus court est considéré comme terme et celui plus long comme sa variante, sauf si cette dernière forme une unité de sens à part (comme *Windanlage* (DE) et *Offshore-Windanlage* qui se traduit par *éolienne offshore* ou *éolienne en mer*).

Les listes de termes⁶ du projet TTC sont disponibles pour deux corpus et sept langues. Ces listes contiennent les termes monolexicaux avec un sous-ensemble de termes monolexicaux néoclassiques (c'est-à-dire, contenant des formants grecs ou latins, comme *aérodynamique*) et les termes polylexicaux. Les termes sont présentés sous leurs formes lemmatisées et accompagnés de leurs formes fléchies, par exemple, le terme *stockage* est associé à la forme au pluriel *stockages*. Certains termes sont accompagnés de leurs variantes graphiques (*aérodynamique* – *aerodynamique*), morphologiques (*synchrone* – *asynchrone*), syntaxiques (*énergie éolien* – *énergie électrique éolien*) ou paradigmatiques (*génératrice* par rapport au *générateur*). Les usages collocationnels sont également représentés, par exemple, *système de stockage*.

Ce guide combiné avec d'autres travaux sur l'extraction terminologique, permet de dresser une typologie de termes que nous avons adopté dans notre projet.

6. <http://www.lina.univ-nantes.fr/?listes-terminologiques-de.html> ; les exemples sont tirés de la liste du domaine de l'énergie éolienne.

1.1.2.4 Typologie de termes

Les termes peuvent être classés et organisés selon plusieurs critères, dont les critères structurels, morphologiques, morphosyntaxiques et sémantiques.

Classification structurelle Les termes constitués d'un seul mot sont appelés monolexicaux (par exemple, *énergie*) et ceux qui en comptent plus – polylexicaux (comme *énergie éolienne*). Les termes monolexicaux peuvent avoir une structure simple (c'est-à-dire être constitués d'un seul noyau sémantique, comme *vent*, *air*, *mât*, *etc.*) ou composée, lorsqu'ils sont formés à partir de plusieurs éléments sémantiques (*biodiversité* = *bio-* + *diversité*).

Les termes comportant des morphèmes grecs ou latins, sont appelés néoclassiques : *aérodynamique*, *mégawatt*, *photovoltaïque*, *aérogénérateur*, *etc.* En fonction du domaine, ils peuvent jouer un rôle plus ou moins important. Par exemple, des sciences dites dures ont tendance à employer des nomenclatures basées sur les éléments néoclassiques, tandis que d'autres domaines (la mode, l'informatique, *etc.*) en utilisent moins.

Bien que ces critères structurels et morphologiques portent plutôt sur la forme que sur le sens, ils sont importants pour la définition des méthodes d'extraction. L'information morphosyntaxique que nous allons discuter ci-après complète ses critères.

Termes et prédictivité L'information sur la partie du discours exprimée via les étiquettes morphosyntaxiques s'avère utile pour l'extraction des termes mono et polylexicaux. Toutefois, les ressources terminologiques se limitent souvent aux noms et aux adjectifs, en excluant les verbes. Leur statut n'étant pas clairement défini, les instructions pour l'annotation des termes (Bernier-Colborne, 2012; Syllabs, 2012) suivent la tendance consistant à les omettre. Pourtant, il existe un certain nombre de verbes, comme *broyer*, *filtrer*, *dissoudre*, *etc.*, qui désignent des actions spécialisées d'un domaine.

Selon L'Homme (2004), la distinction entre l'usage terminologique et usuel d'un verbe se base sur le sens que ce dernier possède lorsqu'il se combine avec des termes du domaine et les mots de langue générale. Par exemple, le verbe *comporter* ne mérite pas d'être inclus dans la terminologie, car son sens est identique dans les contextes spécialisés et généraux.

L'Homme (2016) distingue trois types de termes :

1. prédictifs, qui comprennent des verbes (*boiser*, *planter*, *surexploiter*), des noms prédictifs (*agriculture*, *déboisement*) et certains adjectifs (*durable*, *bio-dégradable*, *etc.*) ;
2. quasi-prédictifs, en suivant Polguère (2012), comme *aire de Y utilisée par X*, par exemple, *aire de répartition de la baleine grise* ;

3. non-prédicatifs, désignant des objets (*dioxyde de carbone*).

Cette classification permet d'inclure les verbes et adjectifs au même titre que les noms, ce qui contribue à la richesse des ressources terminologiques. Les verbes pourraient trouver leur place dans la catégorie sémantique des processus et les adjectifs – dans celle des attributs, selon la classification d'EcoLexicon décrite ci-après.

Catégories sémantiques Selon L'Homme (2016), « les termes peuvent dénoter des éléments et composés chimiques (*dioxyde de carbone, méthane*), des ressources naturelles (*bois, eau, pétrole*), des dispositifs techniques (*éolienne, turbine*), des acteurs (*agriculteur, environnementaliste*), des lieux (*atmosphère, cryosphère, océan*), mais également des propriétés (*potable, toxique, vert*), des états (*climat, condition, temps*), des faits (*amincissement, éroder, recul*) et des activités (*déversement, pollution, recyclage*) ».

Rappelons que cette typologie est proposée dans le contexte des schémas sémantiques. De ce fait, il est intéressant de la comparer avec les catégories sémantiques du projet EcoLexicon (Faber *et al.*, 2005, 2006, 2009; Faber, 2011) qui comprend :

- les agents naturels et artificiels ;
- les processus naturels, artificiels et les instruments ;
- les patients et les résultats ;
- et les descriptions qui regroupent les attributs, les mesures, les disciplines, etc.

Ces deux classifications se rejoignent et se complètent. Elles permettent non seulement de valider un terme selon son appartenance à l'une des catégories, mais aussi de le situer dans le schéma sémantique de son domaine. Toutefois, certaines catégories sémantiques, notamment les acteurs ou les agents selon une classification, peuvent être exprimés via les entités nommées. Il convient donc de définir la place de cette catégorie du lexique.

Termes et entités nommées Les entités nommées incluent généralement les noms propres (personnes, lieux, organisations), mais aussi les expressions temporelles (dates, durées, horaires) et les noms de quantités (monétaires, unités de mesure, pourcentages) (Omrane *et al.*, 2011b). Sanchez-Cisneros et Gali (2013) y ajoutent les noms des substances chimiques et des médicaments. Selon les projets, le nombre de types des entités nommées peut aller d'une dizaine à plusieurs centaines (Cimiano et Völker, 2005).

Les entités nommées n'étant pas des étiquettes de concepts, peuvent toutefois être utiles pour alimenter les instances de concepts dans des ressources d'un domaine. Dans notre projet, aucun traitement des entités nommées n'est prévu, mais certaines unités de mesure ou encore des substances chimiques apparaissent dans les résultats de l'extraction terminologique. Nous rejoignons Omrane *et al.* (2011b) qui considèrent que les entités nommées peuvent jouer le rôle des étiquettes de concepts. En effet, dans leur méthode TERMINAE, le spécialiste humain opère sur le niveau

semi-formel, celui des termino-concepts. C'est à ce niveau qu'il peut décider de créer un concept pour une entité nommée ou de l'assigner en tant qu'une instance d'un concept.

Cette section a été consacrée à la définition des notions de terme et terminologie en fonction du domaine de recherche. À la différence de la linguistique qui propose deux définitions principales d'un termes : d'une part, un terme désigne un concept (la Théorie Générale de la Terminologie) et d'autre part, un terme est considéré comme un signe linguistique, associant un sens et une forme. Le traitement automatique de langues (TAL) adopte une approche pragmatique à la définition de terme, car elle sert surtout à valider un terme candidat extrait par un système automatique à partir d'un corpus spécialisé. Nous adoptons la définition du terme liée à la sémantique où un mot simple ou composé qui sert à dénommer un concept.

Nous avons donc analysé les critères d'évaluation des termes candidats proposées par le projet TTC. Un terme candidat valide doit répondre au minimum à un des critères suivants : être dérivé d'un terme, être défini moyennant un terme, avoir des cooccurrences différentes dans un corpus de spécialité par rapport à un corpus de langue générale ou enfin être un emprunt. Nous avons également synthétisé les principales calcifications des termes en fonction de leurs propriétés structurelles, morphologiques, morphosyntaxiques et sémantiques. Ainsi, les termes peuvent être des mots simples ou composés, appartenant à des différentes parties du discours (des noms, des verbes, des adjectifs, etc.) et désignant des object, phénomènes et processus d'un domaine de spécialité.

L'ensemble des propriétés des termes analysées dans cette section peuvent être exploitées par les méthodes d'extraction terminologique que nous allons étudier dans la section suivante.

1.2 Méthodes d'extraction terminologique

En général, les classifications des méthodes d'extraction terminologique (Bernhard, 2006) distinguent les méthodes basées sur des mesures statistiques, les méthodes basées sur des éléments linguistiques et les méthodes mixtes. Cela correspond aux propriétés de termes analysées ci-dessus. En effet, si les termes sont largement présents dans des corpus spécialisés, il est possible d'exploiter les informations sur leurs fréquences. Les propriétés structurelles et morphosyntaxiques sont à leur tour utilisées par les méthodes linguistiques. Enfin, les méthodes mixtes combinent plusieurs propriétés.

1.2.1 Méthodes statistiques

Les mesures statistiques, telles que la fréquence absolue ou pondérée, TFxIDF, l'information mutuelle ou l'indice de Jaccard mettent en évidence des unités mono ou polylexicales caractéristiques d'un document ou d'un corpus. Ces méthodes sont indépendantes de la langue et nécessitent uniquement un ou plusieurs corpus. Toutefois, leurs résultats sont meilleurs sur des corpus de taille importante. Certaines méthodes permettent d'évaluer le potentiel terminologique de termes candidats (Drouin et Langlais, 2006), ce qui peut servir de filtre pour d'autres méthodes.

Fréquence absolue La fréquence absolue (*Term Frequency* ou *TF*) correspond au nombre d'occurrences d'une unité lexicale dans le corpus. La fréquence absolue est difficilement exploitable en tant que le seul paramètre, car les mots les plus fréquents sont généralement les mots grammaticaux ou les mots de langue générale. Toutefois, combinée avec d'autres traits, elle constitue un paramètre intéressant. Les hapax, ou les éléments ayant une seule occurrence dans le corpus, sont un cas intéressant qui relève de la fréquence absolue. Souvent rejetés à cause du bruit excessif, les hapax peuvent toutefois contenir quelques termes du domaine.

Fréquence relative La fréquence relative permet de situer des unités lexicales sur une échelle normalisée $0 < freq \leq 1$. Elle peut être calculée selon la formule suivante : $tf_rel_m = \frac{tf_m}{tf_{max}}$ où la fréquence d'une unité m est divisée par la fréquence maximale constatée dans l'échantillon. Il est également possible de normaliser par le nombre de mots dans le corpus. Les deux calculs permettent d'identifier les mots les plus fréquents.

TFxIDF La mesure TFxIDF *Term Frequency-Inverse Document Frequency* permet d'identifier les unités fréquentes dans peu de documents d'une collection. En effet, certains termes peuvent apparaître dans un nombre réduit de documents du corpus, tandis que les mots de langue générale ont tendance à être présents dans l'ensemble de documents. Tout comme la fréquence, pour être efficace cette mesure doit être accompagnée par d'autres traitements. TFxIDF est calculée comme suit : $TFxIDF_m = \frac{tf_m}{tf_{max}} \times \log \frac{D}{d}$ où tf_{max} est la fréquence maximale, d est le nombre de documents contenant l'unité m dans l'ensemble de documents D .

Information mutuelle Cette mesure est utilisée pour identifier des termes candidats polylexicaux. L'indice d'information mutuelle mesure l'association entre deux unités lexicales dans un corpus, en se servant de leurs fréquences relatives indépendantes : $MI_{(i,k)} = \log_2 \frac{freq_rel_{(i,k)}}{freq_rel_i * freq_rel_k}$ où $freq_rel_{(i,k)}$ est le nombre relatif de cooccurrences communes et $freq_rel_i$ et $freq_rel_k$ – des occurrences distinctes.

Indice de Jaccard Cet indice permet également de vérifier si une paire d'unités lexicales peut être un terme polylexical. L'indice de Jaccard est calculé selon la formule suivante : $J_{(i,k)} = \frac{freq_{(i,k)}}{freq_i + freq_k - freq_{(i,k)}}$ où les occurrences communes des unités i et k sont divisées par le nombre total de leurs occurrences moins les occurrences communes. Si l'indice de Jaccard dépasse un seuil voulu, la paire des mots analysée n'est pas une simple cooccurrence, mais une collocation et, potentiellement, un terme polylexical.

Résolution probabiliste des termes polylexicaux Brooke *et al.* (2014) segmentent le corpus afin d'identifier les expressions polylexicales sans lien direct avec la terminologie, ce qui n'empêche pas de l'utiliser pour l'extraction des termes candidats polylexicaux.

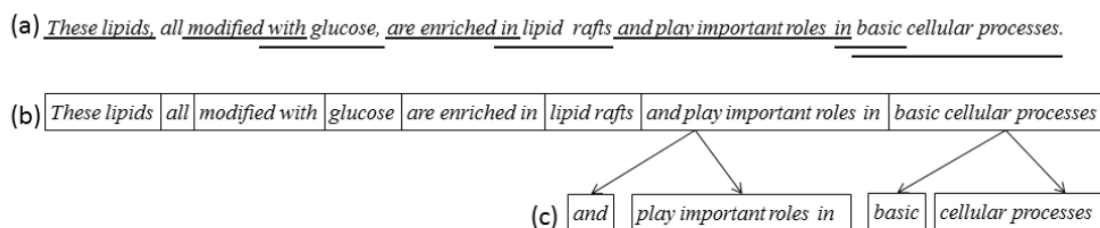


FIGURE 1.1 – Exemple de la résolution probabiliste de segments (Brooke *et al.*, 2014)

Leur méthode comprend trois étapes consécutives (Figure 1.1) :

- l'extraction de n-grammes de mots où $2 < n < 10$;
- la résolution probabiliste des chevauchements entre les segments ayant pour objectif de choisir les limites des segments les plus probables (par exemple, le segment *modified with* est plus probable que celui *with glucose*) ;
- la segmentation affinée à l'aide du lexique défini préalablement.

Cette méthode présente un double intérêt, car elle est multilingue et permet d'identifier des expressions polylexicales en exploitant des termes ou des termes candidats monolexicaux.

Il existe d'autres mesures statistiques applicables à l'extraction terminologique. Le trait commun à toutes ces méthodes est le rappel, donc le nombre de candidats identifié sur l'ensemble de candidats à identifier, important. Toutefois, l'ensemble de ces mesures nécessite des traitements complémentaires ou une combinaison avec d'autres méthodes pour arriver à un degré de précision acceptable. La précision est le nombre de candidats valides divisé par le nombre de candidats extraits.

Nous allons maintenant analyser les éléments linguistiques qui permettent d'extraire des termes candidats et d'atteindre une meilleure précision.

1.2.2 Méthodes linguistiques

La deuxième catégorie de méthodes fait appel à des éléments linguistiques parmi lesquels on peut distinguer deux groupes : les formants de langues classiques et les patrons morphosyntaxiques.

Formants néo-classiques L'identification des termes de certains domaines peut utiliser des traits morphologiques spécifiques à leurs nomenclatures. Tel est le cas de la nomenclature en chimie, en physique, en biologie ou encore en médecine. Cette approche est adaptée pour des termes qui contiennent des composants néo-classiques, c'est-à-dire, des morphèmes provenant du grec ou du latin. En effet, ces morphèmes sont très productifs ; par exemple, le radical *AZOT* est à base des termes suivants : *azote, azotate, azoté, azoteux, azotique, azotite, azotémie, azothydrique, azoture, azoturie, etc.* Estopà *et al.* (2000a) ont développé une méthode d'extraction terminologique qui utilise une liste de formants néo-classiques fournie par l'utilisateur pour un domaine spécifique. Appliquée aux domaines listés ci-dessus, cette méthode a permis d'identifier 87,70 % (Estopà *et al.*, 2000a,b) de termes contenant des formants néo-classiques dans l'échantillon.

Les listes de formants grecs et latins sont disponibles sur Internet. Cependant, elles nécessitent une vérification manuelle et une adaptation de format avant d'être introduites dans le programme. Les méthodes qui comparent des fréquences de formants dans un corpus de spécialité et dans un corpus de langue générale sont applicables pour affiner les listes existantes (Bernhard, 2006). La limite évidente des formants est leur productivité dans le domaine analysé : certains domaines ont très peu de recours à ces morphèmes, en préférant des emprunts ou la néologie.

Patrons morphosyntaxiques Les patrons morphosyntaxiques représentent des séquences d'étiquettes attribuant une partie du discours à chaque mot, par exemple, un nom suivi d'un adjectif constitue le patron NOM-ADJ⁷. Comme nous l'avons mentionné lors de la discussion sur les propriétés d'un terme, les termes monolexicaux sont souvent des noms, des adjectifs ou des verbes. Les termes polylexicaux suivent eux aussi un certain nombre de combinaisons syntagmatiques qui dépendent de la langue traitée. Ainsi, les termes en français sont souvent identifiables grâce aux patrons NOM-PREP-NOM (p.ex., *traitement des eaux*) ou bien NOM-ADJ (*eaux usées*). En anglais, nous pouvons observer des patrons NON-NOM (*water pollution*), etc.

Cette approche est utilisée dans les méthodes de Daille (2003); Morin et Daille (2012), de Bourigault et Fabre (2000); Bourigault (2002). Les patrons morphosyntaxiques sont également exploités dans la C-value/NC-value (Frantzi *et al.*, 2000), combinés avec des calculs statistiques. Notamment, la C-value correspond au score

7. Les jeux d'étiquettes diffèrent d'un outil à l'autre et, dans certains cas, d'une langue à l'autre.

terminologique calculé à base de la fréquence d'un terme candidat polylexical dans le corpus, de sa fréquence dans d'autres termes candidats plus longs, le nombre de ces candidats plus longs et le nombre de mots du candidat en question. Cela permet d'extraire des termes candidats polylexicaux qui montrent une certaine indépendance des autres candidats, donc apparaissent dans le corpus tels quels, mais peuvent également être inclus dans des termes plus longs. La NC-value complète la mesure précédente en favorisant des candidats qui ont plus de termes dans leurs contextes dans le corpus (Frantzi *et al.*, 2000). Orobinska *et al.* (2013) proposent une approche plus souple en apprenant les patrons morphosyntaxiques caractéristiques à partir d'un corpus étiqueté. Cette méthode permet d'éviter l'étape de constitution manuelle des patrons. L'un des avantages des patrons morphosyntaxiques est la possibilité d'identifier la variation au sein des termes candidats polylexicaux et de définir des règles de transformation afin d'améliorer l'organisation des résultats (Bourigault et Jacquemin, 1999). Par exemple, il est possible de rapprocher des candidats où seul l'article change : *traitement des eaux* et *traitement d'eau*.

Dans les deux cas, les éléments linguistiques peuvent être complétés par des traitements statistiques pour augmenter la qualité des résultats. Il faut remarquer que les formants néo-classiques fournissent, selon les domaines, une excellente précision. Le rappel, quant à lui, dépend étroitement de la richesse de la liste de formants utilisée dans l'extraction. Dans la section suivante, nous allons analyser des méthodes qui combinent plusieurs traits et que nous appelons mixtes.

1.2.3 Méthodes mixtes

Les méthodes mixtes se caractérisent par la prise en compte de plusieurs éléments linguistiques et statistiques. Dans notre travail, nous reprenons certains éléments des méthodes mixtes qui ne nécessitent pas de pré-traitement de corpus.

La première méthode que nous allons analyser est l'étiquetage par des mots informatifs et vides (Vergne, 2003, 2004, 2005). Ensuite, vient la notion du lexique transdisciplinaire (Jacquey *et al.*, 2013; Tutin, 2007a). La troisième méthode s'appelle Apprentissage Naturel Automatique Enguehard (1993) qui permet d'identifier des termes candidats polylexicaux. Nous allons également mentionner les ressources nécessaires à l'application de ces méthodes.

Dans cette section, nous allons nous baser sur un exemple de deux phrases tokenisées⁸ issues des descriptions de brevets⁹ :

8. La tokénisation consiste à séparer tous les mots d'une phrase les uns des autres, y compris la ponctuation.

9. Le corpus PatTR (Wäschle et Riezler, 2012a,b) sera présenté plus loin dans la partie dédiée aux ressources utilisées.

1.2. MÉTHODES D'EXTRACTION TERMINOLOGIQUE

Phrase 1 :

La présente invention concerne des dispositifs de connexion , du type comportant au moins un élément tubulaire de connexion , conducteur de l' électricité , supporté par un support électriquement isolant , élément dont la première et la seconde partie extrême , opposées l' une de l' autre , comportent respectivement une première et une seconde fente transversale constituant des moyens de connexion autodénudants susceptibles de recevoir chacun au moins un conducteur , un passage étant prévu dans le support isolant , vis-à-vis d' au moins la seconde fente transversale , pour y loger le conducteur , selon le brevet FR-A-2.541.049 (= GB-A-2 135 530) .

Phrase 2 :

Plus précisément, l' invention concerne un compresseur à prélèvement d' air centripète , du type comportant plusieurs tubes d' orientation radiale , régulièrement répartis circonférentiellement , lesdits tubes étant installés entre deux viroles coaxiales du rotor dudit compresseur , respectivement une virole extérieure et une virole intérieure , l' espace à l' extérieur de ladite virole extérieure communiquant avec la veine dudit compresseur , caractérisé en ce qu' une extrémité de chaque tube est engagée dans un trou de ladite virole extérieure , en ce que chaque tube est engagé dans un trou de ladite virole intérieure et en ce que des moyens de verrouillage sont définis entre un tel tube et ladite virole intérieure .

1.2.3.1 Étiquetage par des mots informatifs et vides

La méthode proposée par Vergne (2003, 2004, 2005) permet d'annoter le corpus par des mots informatifs et vides pour ensuite en extraire des termes à structure contrôlée.

L'idée de discriminer les mots outils et les mots de langue générale afin de mettre en valeur les termes candidats n'est pas nouvelle. En effet, la technique de l'annotation de mots pleins, lexicaux, est largement utilisée dans les moteurs de recherche qui utilisent les anti-dictionnaires (les listes de mots grammaticaux à exclusion de l'indexation). Un exemple d'anti-dictionnaire est celui du projet Snowball¹⁰. L'anti-dictionnaire est également utilisé dans la méthode de l'extraction terminologique ANA (Apprentissage naturel Automatique) développée par Enguehard (1993), où il prend la forme de quatre listes séparées qui sont exploitées pour identifier des termes candidats polylexicaux. La différence d'approche est claire : l'indexation tend à discriminer les mots vides, tandis que l'extraction terminologique doit s'en servir.

Il est intéressant de comparer les définitions de mots vides données par ces deux perspectives. Dans la recherche d'information, le mot vide est « tout mot à faible contenu informatif » (Chibout *et al.*, 2000). Parmi les exemples de mots vides, Chibout *et al.* (2000) cite les déterminants, les formes de verbes modaux (*devoir*) et

10. snowball.tartarus.org/algorithms/

les mots fréquents et peu représentatifs pour le domaine en question. Pour illustrer ces derniers, il donne l'exemple du mot *avion* par rapport au corpus sur l'avionique. Amini et Gaussier (2013) incluent également dans les mots vides « tous les mots des parties de discours autres que nom, adjectif, verbe et adverbe ».

Dans le TAL, Vergne adopte la définition des mots vides de Tesnière (1959) et de Debili (1982), fondée également sur l'appartenance des mots à des parties de discours, en y ajoutant les mots lexicaux non discriminants (Vergne, 2003). Il propose alors la combinaison de trois propriétés, la fréquence, la longueur et l'entourage proche, pour détecter les mots vides à partir de corpus brut sans être contraint de renseigner la langue de ce dernier (Vergne, 2003, 2004, 2005). Selon sa méthode, le mot informatif est sélectionné selon trois critères :

1. longueur importante ;
2. fréquence réduite ;
3. cooccurrences avec par des mots plus courts et plus fréquents.

Le reste des mots sont considérés comme des mots vides. Cependant, les résultats de l'étiquetage ne sont pas assez homogènes pour formuler la notion de mots informatifs.

L'application de la méthode de Vergne veut que les mots informatifs s'approchent de termes candidats. Or, les exemples de résultats obtenus pour cette méthode sur des corpus constitués d'articles de presse (Vergne, 2005) contiennent peu de termes candidats tels que nous pouvons en exploiter dans le système de gestion de la terminologie. Par exemple, dans *une nouvelle résolution de l'ONU*, les mots informatifs sont : *nouvelle*, *résolution*, *ONU*. Les mots *ONU* et *résolution* peuvent effectivement être des termes candidats et même former un terme candidat polylexical *résolution de l'ONU*, mais le mot *nouvelle* ne s'inscrit pas dans cette terminologie. Dans d'autres exemples, les mots informatifs incluent également des mots de la langue générale, comme *cherche*, *utiliser*, *tonnes*, *l'or*, etc.

Nous allons illustrer cela sur l'exemple des deux phrases annotées manuellement pour illustrer les résultats attendus, en mettant les mots vides en gris :

Phrase 1 :

La présente invention concerne des dispositifs de connexion , du type comportant au moins un élément tubulaire de connexion , conducteur de l' électricité , supporté par un support électriquement isolant , élément dont la première et la seconde partie extrême , opposées l' une de l' autre , comportent respectivement une première et une seconde fente transversale constituant des moyens de connexion autodénudants susceptibles de recevoir chacun au moins un conducteur , un passage étant prévu dans le support isolant , vis-à-vis d' au moins la seconde fente transversale , pour y loger le conducteur , selon le brevet FR-A-2.541.049 (= GB-A-2 135 530) .

Phrase 2 :

Plus précisément , l' invention concerne un compresseur à prélèvement d' air centripète , du type comportant plusieurs tubes d' orientation radiale , régulièrement répartis circonférentiellement , lesdits tubes étant installés entre deux viroles coaxiales du rotor dudit compresseur , respectivement une virole extérieure et une virole intérieure , l' espace à l' extérieur de ladite virole extérieure communiquant avec la veine dudit compresseur , caractérisé en ce qu' une extrémité de chaque tube est engagée dans un trou de ladite virole extérieure , en ce que chaque tube est engagé dans un trou de ladite virole intérieure et en ce que des moyens de verrouillage sont définis entre un tel tube et ladite virole intérieure .

L'avantage de cette annotation est évident : les éléments potentiellement intéressants pour l'extraction terminologique et pour la recherche des entités nommées sont plus visibles. Cependant, il nous reste à filtrer le lexique qui est propre à la structure et au style des brevets. Nous allons détailler l'application de cette méthode dans la Section 3.3.

1.2.3.2 Lexique transdisciplinaire

Le lexique transdisciplinaire peut être utilisé comme indicateur et délimiteur des unités terminologiques (Jacquey *et al.*, 2013; Tutin, 2007a). En effet, des mots comme *concept*, *méthode*, *technologie*, *analyser*, *etc.* servent souvent à introduire des termes d'un domaine sans appartenir à ce domaine. La détection de ces mots ou expressions et leur classification facilitent l'extraction terminologique. Jacquey *et al.* (2013) combinent le lexique transdisciplinaire projeté sur le corpus avec l'analyse syntaxique pour identifier les cas où une unité du lexique transdisciplinaire introduit un terme candidat, par exemple *le concept de diachronie* où le mot *concept* qui appartient au lexique transdisciplinaire, aide à identifier le terme *diachronie*. Leur hypothèse a été confirmée dans environ 74 % des cas (Jacquey *et al.*, 2013).

En même temps, Hatier (2013) souligne le fait que la frontière entre le lexique transdisciplinaire et la terminologie peut être assez floue, surtout pour les sciences humaines comparées à des sciences exactes. Blumenthal (2007) a illustré les divergences lexicales existantes entre ces domaines scientifiques à travers une étude statistique sur les collocations et les contextes présents dans deux corpus correspondants.

Le lexique transdisciplinaire est défini par Tutin (2007b) comme « le lexique partagé par la communauté scientifique mis en œuvre dans la description et la présentation de l'activité scientifique. [...] un lexique de genre, n'intégrant pas la terminologie du domaine, mais renvoyant aux concepts mis en œuvre dans l'activité scientifique (examiner, prouver, réfuter, concluant, hypothèse, examen, encourageant ...) ».

Hatier (2013) définit le lexique scientifique transdisciplinaire comme transversal aux différentes disciplines et spécifique à l'écrit scientifique (par l'opposition au

1.2. MÉTHODES D'EXTRACTION TERMINOLOGIQUE

lexique général). Dans le même ordre d'idée, Da Sylva (2010) opère la notion du vocabulaire savant de base (VSB) qui contient des « mots abstraits », appartenant au « vocabulaire savant » et dotés d'un « sens général et non spécifique à un domaine disciplinaire ».

Comme nous pouvons l'observer, l'attention a été mise davantage sur les domaines scientifiques. Cependant, le travail de Biber *et al.* (2004) souligne la nécessité de tenir compte du genre de document. Ils étudient le lexique et les expressions propres à des textes utilisés dans l'enseignement de l'anglais, dont les manuels, les cahiers de travaux dirigés et les livres pour les enseignants. L'analyse est menée à la fois au niveau structurel et pragmatique. Les auteurs fournissent une analyse statistique comparative de l'utilisation de ces expressions entre les différents types de textes.

L'identification du lexique transdisciplinaire et structurel peut constituer une étape complémentaire de l'extraction terminologie, notamment pour empêcher ces éléments d'entrer dans la liste des termes candidats. Dans l'exemple suivant, nous allons surligner en jaune les mots et les expressions caractéristiques à des brevets :

Phrase 1 :

La présente invention concerne des dispositifs de connexion, du type comportant au moins un élément tubulaire de connexion, conducteur de l'électricité, supporté par un support électriquement isolant, élément dont la première et la seconde partie extrême, opposées l'une de l'autre, comportent respectivement une première et une seconde fente transversale constituant des moyens de connexion autodénudants susceptibles de recevoir chacun au moins un conducteur, un passage étant prévu dans le support isolant, vis-à-vis d'au moins la seconde fente transversale, pour y loger le conducteur, selon le brevet FR-A-2.541.049 (= GB-A-2 135 530).

Phrase 2 :

Plus précisément, l'invention concerne un compresseur à prélèvement d'air centripète, du type comportant plusieurs tubes d'orientation radiale, régulièrement répartis circonférentiellement, lesdits tubes étant installés entre deux viroles coaxiales du rotor dudit compresseur, respectivement une virole extérieure et une virole intérieure, l'espace à l'extérieur de ladite virole extérieure communiquant avec la veine dudit compresseur, caractérisé en ce qu'une extrémité de chaque tube est engagée dans un trou de ladite virole extérieure, en ce que chaque tube est engagé dans un trou de ladite virole intérieure et en ce que des moyens de verrouillage sont définis entre un tel tube et ladite virole intérieure.

Le lexique restant après le double filtrage par les mots vides et par le lexique transdisciplinaire et structurel, est considérablement saturé par les termes candidats mono et polylexicaux. On peut donc constater que le filtrage facilite la tâche de l'extraction terminologique et d'entités nommées.

1.2.3.3 Apprentissage Naturel Automatique

La méthode ANA (Apprentissage Naturel Automatique), a été développée par Enguehard (1993) dans l'objectif d'apprendre des termes candidats mono et polylexicaux en partant d'un vocabulaire limité d'un domaine de spécialité.

A l'étape de la familiarisation, le logiciel parcourt le corpus et en extrait les listes des mots fonctionnels qui « sont, typiquement, des articles, des pronoms, quelques adverbes » (*a, alors, après, etc.*), des mots fortement liés qui « sont des chaînes de caractères comprenant des caractères blancs mais pourtant considérés comme des mots » (*de la, de l', etc.*) et des mots de schéma, donc « des mots fonctionnels structurant des syntagmes » (*en, de, du, d, de la, des*) (Enguehard, 1993). Ces deux derniers groupes servent de liaison dans les termes candidats polylexicaux et permettent de les identifier dans le texte.

Ensuite, le programme analyse le corpus de manière itérative, en identifiant les mots qui apparaissent fréquemment avec les termes de bootstrap (un ensemble de quelques termes du domaine prédéfinis manuellement dont il est question dans le corpus de textes), et en les rajoutant dans le bootstrap.

Nous allons retenir de cette méthode la notion des mots de schéma. Ils peuvent être définis comme un vocabulaire restreint d'éléments acceptés au sein de termes polylexicaux qui assurent la liaison entre leurs composants.

Les méthodes mixtes combinent les points forts des méthodes statistiques et linguistiques, ce qui les rend plus précis et performants. Ce type de méthodes est le mieux adapté à la tâche d'extraction terminologique dans un contexte multilingue et multi-domaine.

Dans cette section, nous avons présenté les différentes méthodes à l'extraction des termes à partir de corpus. Ces méthodes sont classées en statistiques, linguistiques et mixtes. Nous avons constaté qu'il n'existe pas de méthode véritablement multi-domaine et multilingue, notamment à cause des limites imposés par chaque type de méthodes.

Ainsi, les méthodes statistiques sont très dépendantes du volume de corpus et détectent essentiellement des termes fréquents dans un corpus spécialisé. Les méthodes linguistiques et mixtes, à leur tour, nécessitent des outils et des ressources complémentaires, qui ne sont pas disponibles pour toutes les langues et domaines.

Par exemple, les ressources morphosyntaxiques, notamment les patrons d'identification de candidats polylexicaux, sont développées pour chaque langue à part. Elles dépendent de l'étiqueteur morphosyntaxique disponible et reposent sur les résultats de ce dernier. Cependant, l'étiqueteur risque de perdre en précision lorsqu'il s'agit d'un domaine très spécifique, comme la médecine ou la chimie. Les méthodes basées sur ces patrons sont performantes, mais limitées à des langues qui ont déjà un étiqueteur disponible.

Les ressources lexicales, comme des anti-dictionnaires (*stop words lists*) ou le lexique transdisciplinaire (Tutin, 2007a; Jacquy *et al.*, 2013), sont évidemment liées à la langue pour laquelle elles sont développées. De ce point de vue, les travaux de Drouin (2007), qui portent sur l'identification automatique du lexique transdisciplinaire, sont très intéressants. Une autre méthode permettant d'éviter l'utilisation d'un anti-dictionnaire est l'annotation de Vergne (2003, 2004, 2005). En effet, les mots étiquetés comme vides sont dans la majorité des cas éligibles pour un anti-dictionnaire. Enfin, les termes de bootstrap pour la méthode ANA doivent également être sélectionnés manuellement avant la lancement du système.

Il est donc nécessaire de diminuer la quantité de ressources et d'outils afin de rendre l'extraction terminologique multilingue et applicable à un grand nombre de domaines.

L'extraction terminologique est faite pour chaque langue à part. Il est donc nécessaire d'ajouter un traitement pour identifier les traductions de ces termes pour chaque combinaison de langues.

1.3 Alignement de terminologie multilingue

L'alignement des termes en plusieurs langues constitue une étape incontournable dans l'extraction terminologique multilingue. Les approches utilisées varient en fonction des langues et des ressources disponibles, mais aussi en fonction de la structure des termes. Notamment, les termes monolexicaux, constitués d'un seul mot, et polylexicaux, qui en comptent plus d'un, sont souvent traités séparément.

1.3.1 Alignement des termes composés monolexicaux

Les termes composés monolexicaux peuvent contenir des unités lexicales (par exemple, le mot anglais *fireplace* est composé de *fire* et de *place*) ou des morphèmes, comme le terme *biology* formé de *bio-* et *-logy*. Certains mots composés peuvent combiner des éléments des deux types, comme dans le cas de *biocarburant* construit à partir du morphème *bio-* et du mot *carburant*. Ces deux catégories d'éléments sont souvent traitées avec des méthodes différentes, ce que nous allons illustrer dans cette sous-section.

1.3.1.1 Alignement des composants lexicaux

Les termes composés des unités lexicales sont généralement segmentés en mots et traduits à l'aide des dictionnaires. Le résultat final dépend de la qualité de cette segmentation, ainsi que de la disponibilité des ressources nécessaires.

Segmentation et fusion des mots composés Cap *et al.* (2014) ont étudié la compositionnalité du point de vue de son utilité pour la traduction automatique. En effet, leur expérience montre que la pré-traduction des mots composés améliore la qualité du texte produit par un système de la traduction automatique. Cependant, cette approche peut être intéressante pour la recherche des traductions de termes candidats dans des corpus parallèles et comparables.

Les mots composés sont caractéristiques pour certaines langues (par exemple, pour l'allemand) plus que pour d'autres (le français et l'anglais), d'où découle la différence dans les fréquences et dans la distribution de ces mots et de leurs traductions dans les corpus bilingues.

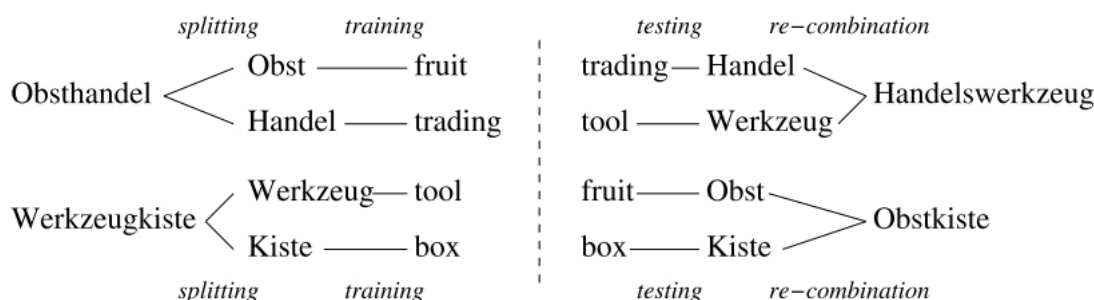


FIGURE 1.2 – La construction des mots composés non-connus à l'entraînement Cap *et al.* (2014)

Le découpage des mots composés est relativement bien étudié, tandis que leur ré-composition ne l'est pas. La particularité de la méthode proposée par Cap *et al.* (2014) (Figure 1.2) consiste à amener les parties des mots composés à leurs formes "de base" (*underspecified representations*) (Fritzinger et Fraser, 2010), c'est à dire au lemme et à l'étiquette morphosyntaxique de la partie du discours, complétée par l'information sur le genre et le nombre. De cette façon, la position du mot simple au sein du mot composé n'est plus importante, car ce dernier est formé à partir des traductions selon les règles de la langue cible (Fraser *et al.*, 2012). Par exemple, les mots *Obsthandel* (commerce de fruits) et *Werkzeugkiste* (boîte à outils) présents dans le corpus d'entraînement permettent de générer les traductions pour un outil de commerce (*Handelswerkzeug*) et une boîte à fruits (*Obstkiste*), qui n'y apparaissent pas.

Il faut noter que les formes de base sont complétées par l'information sur la forme qu'elle prennent dans les mots composés en fonction de la position (un ajout de morphèmes complémentaires peut être nécessaire, p. ex. : *Ort* (endroit)+*s*+*Zeit* (temps) = *Ortszeit* (l'heure locale)). La décision si un mot composé doit être généré dans la langue cible ou pas est prise en fonction des règles syntaxiques définies pour chaque paire de langue. Par exemple, si deux mots simples consécutifs en allemand sont alignés à deux mots faisant partie d'un syntagme nominal en anglais, un mot

composé allemand doit être construit (Cap *et al.*, 2014).

Modélisation de la flexion Dans certaines paires de langues, une langue contient plus d’informations morphologiques que l’autre. Par exemple, dans la paire anglais-allemand, le mot anglais *blue* (bleu) peut être traduit en allemand par *blau*, *blaue* (bleue, genre féminin), *blauer* (comparatif, plus bleu), etc. (Fraser *et al.*, 2012). Les langues qui ont peu d’informations morphologiques comme l’anglais sont appelées « morphologiquement pauvres », tandis que les langues qui incluent les informations sur le genre, le nombre, les cas, etc. – « morphologiquement riches », comme l’allemand. Fraser *et al.* (2012) séparent la traduction de mots d’une langue morphologiquement pauvre vers une langue morphologiquement riche en trois étapes :

- les mots sont décomposés morphologiquement (Fritzinger et Fraser, 2010) et désuffixés (*stemming* (EN)) ;
- les mots désuffixés (*stems* (EN)) sont traduits vers la langue cible ;
- les mots cibles sont fléchis selon les règles de transformation prédéfinies pour la paire de langues en question. Ces règles comprennent les étiquettes morphosyntaxiques et leurs séquences.

Dans la continuité de ce travail, Weller *et al.* (2013) cherchent à améliorer la prédiction du cas pour les syntagmes nominaux en langue cible. Ils reprennent les deux premières étapes énumérées ci-dessus et y ajoutent la prédiction du cas en langue cible. Elle s’effectue soit en analysant les fonction syntaxiques de ces syntagmes en langue source, soit en fonction du verbe qui régit le syntagme nominal en question. Cap *et al.* (2014) reprennent cette méthode pour fléchir les mots composés en langue cible dans leur chaîne de traitements.

Une méthode similaire est développée pour les paires EN-DE (Weller et Heid, 2012) et EN-FR (Weller *et al.*, 2014) : les combinaisons générées dans la langue cible sont transformées en utilisant des patrons morphosyntaxiques prédéfinis et ensuite filtrés à l’aide d’une liste de mots en langue cible. La flexion est alors également traitée via la méthode de Fraser *et al.* (2012).

Traduction compositionnelle par langues pivots Garera et Yarowsky (2008) proposent d’utiliser une collection de dictionnaires bilingues de paires de langues différentes comme corpus parallèle. Pour traduire un mot composé d’une langue source vers une langue cible, ils procèdent d’abord au découpage du mot composé en mots simples et cherchent à les ré-composer et traduire le mot composé obtenu dans chaque langue pivot disponible (les paires de langues source-pivot pour lesquelles les dictionnaires bilingues sont disponibles). Ensuite, ils choisissent les traductions les plus fréquentes obtenues pour ces mots composés de des langues pivots vers la langue cible.

Par exemple, le mot albanien *hekurudhë* signifiant *chemin de fer* est segmenté en *hekur* (*fer*) et *udhë* (*chemin*) ; ensuite, leurs traductions sont recherchées dans

toutes les langues disponibles. Ainsi, *hekur* peut être aligné avec le mot italien *ferro* et *udhë* avec *via*, dont la combinaison est *ferrovia* qui peut être traduite par le mot composé anglais *railroad* Garera et Yarowsky (2008) .

Ainsi, même si le mot composé de la langue source ne se traduit pas par un mot composé dans la langue cible, il suffit qu’il se traduise par un mot composé dans une langue pivot pour que sa traduction dans la langue cible puisse être identifiée. La précision pour le top-1 candidat est de 20 % et peut atteindre 50 % pour les top-10 candidats. Toutefois, cette méthode traite uniquement les mots composés des éléments lexicaux, et n’est pas applicable à la compositionnalité morphologique, ce qui exclut un grand nombre de mots composés.

Dans le même esprit, Claveau et Kijak (2011) proposent de segmenter les mots et d’aligner les morphèmes obtenus avec les mots simples en japonais (plus précisément, en kanjis).

Cet aperçu de techniques d’alignement des composants lexicaux permet de constater leur dépendance de la segmentation et des dictionnaires disponibles. Nous allons maintenant analyser les approches à l’alignement des termes au niveau des morphèmes.

1.3.1.2 Traitement morphologique des mots composés

Dans le cas des mots composés contenant des morphèmes, les listes de traductions de ces derniers sont construites manuellement pour chaque paire de langues, et la qualité de segmentation des mots composés en morphèmes reste un point important. De ce point de vue, seulement le méthode de Bollegala *et al.* (2015) n’est pas sujette à ces problèmes.

Traduction morpho-compositionnelle Dans l’approche compositionnelle, Delpech *et al.* (2012) identifient les équivalences de traduction au niveau des morphèmes à partir de corpus comparables.

Leur démarche comprend trois étapes :

1. décomposition du terme en morphèmes (*post-menopause* EN $\rightarrow$ *post-* et *menopause*)
2. traduction de chaque morphème vers un morphème ou vers un mot simple (*post-* $\rightarrow$ FR *post-* ou *après* ; *menopause* $\rightarrow$ FR *ménopause*)
3. reconstruction du terme en langue cible (*post-ménopause* ou *après la ménopause*)

Cette approche permet de gérer les traductions *fertiles*, c’est-à-dire, les cas où le nombre de mots du terme source est inférieur à celui cible (par exemple, *post-menopause* EN $\rightarrow$ FR *après la ménopause*). La traduction fertile est possible grâce

aux deux principes. Selon le premier principe, un mot peut être segmenté en composants "libres" (qui peuvent former les mots tout seuls, comme "*cancer*") ou "liés" (qui n'apparaissent qu'en liaison avec d'autres éléments, comme le préfixe "ré-"). Le deuxième principe permet de traduire chaque composant par son équivalence traductionnelle "libre" ou "liée". Ainsi, peu importe si le terme source est mono ou polylexical, le système pourra générer ses traductions mono ou polylexicales dans la langue cible.

Les traductions obtenues sont ensuite classées à l'aide de paramètres utilisés séparément ou en combinaison :

- la fréquence ;
- la probabilité de l'étiquette morphosyntaxique ;
- la similarité du contexte ;
- le score de confiance des ressources utilisées.

Les auteurs utilisent plusieurs ressources. Notamment, ils s'appuient sur un dictionnaire bilingue de langue générale, un dictionnaire des cognats¹¹ construit automatiquement à partir de corpus comparables et une table de correspondance des morphèmes établie manuellement. Cette table présente des éléments sous-lexicaux alignés pour les deux langues, par exemple *-cardio-* (EN) peut se traduire en français par *-cardio-* ou *cœur*, ou bien par *cardiaque*. Ces ressources sont complétées par un dictionnaire de synonymes (Xelda¹²) et une liste de familles morphologiques générée à partir des mots désuffixés à partir du corpus comparable.

Alignement des formants néo-classiques Weller *et al.* (2011) proposent une méthode similaire appliquée aux mots composés ayant au moins un formant néo-classique, c'est-à-dire, des morphèmes d'étymologie grecque ou latine, appelés aussi des formants savants. Par exemple, le mot *biologie* contient deux formants néo-classiques : *bio-* (du grec *bios*, la vie) et *-logie* (*logos*, l'étude). La méthode fait appel à une liste de racines néo-classiques et à un dictionnaire bilingue. Dans le même ordre d'idée, Harastani *et al.* (2012) utilisent une liste bilingue de formants néo-classiques construite manuellement. Cette liste est ensuite projetée sur le corpus comparable afin d'en extraire les listes de mots contenant ces formants. Les mots de la liste en langue source sont segmentés et chaque segment est traduit à l'aide de la liste des formants ou d'un dictionnaire bilingue (pour les mots simples obtenus après la segmentation). Les traductions candidates ré-composées sont recherchées dans la liste de mots en langue cible et si un mot est retrouvé, la traduction est validée.

Mesure de similarité crosslingue Bollegala *et al.* (2015) distinguent deux types de paramètres pour la détection des traductions pour des termes biomédicaux :

11. Selon la définition de Navlea (2014), les cognats sont des mots bilingues similaires au niveau orthographique et/ou phonétique, susceptibles de partager un sens commun. Par exemple, les mots français *ménopause* et anglais *menopause* sont des cognats.

12. <http://www.xrce.xerox.com/About-XRCE/History/Historical-projects/XeLDA>

- les paramètres intrinsèques représentés par des uni-grammes et des bi-grammes de caractères extraits à partir des termes ;
- les paramètres extrinsèques, c'est-à-dire, des des uni-grammes et des bi-grammes de caractères extraits à partir du contexte des termes.

Ils utilisent une projection vectorielle prototype qui consiste à réduire le nombre de dimensions de l'espace vectoriel et à garder uniquement une partie des vecteurs, appelés alors *prototypes*. Ces vecteurs contiennent trois composants : le vecteur de contexte en langue source, un vecteur de contexte en langue cible et un vecteur contenant les traits crosslingues (donc les uni- et les bi-grammes décrits ci-dessus).

Les liens entre les espaces vectoriels en langue source et cible sont détectés en utilisant la régression des moindres carrés partiels (Formule 1.1). Cette approche statistique permet de rapprocher des éléments similaires. Elle est surtout utile lorsque la quantité d'éléments à comparer est importante.

$$Y = ThC'h + Eh = XWh * C'h + Eh = XWh(P'hWh) - 1C'h + Eh \quad (1.1)$$

où Y est la matrice des variables dépendantes, X celle des variables explicatives, et où Th , Ch , W^*h , Wh et Ph , sont des matrices générées par l'algorithme PLS, et où Eh est la matrice des résidus¹³.

L'expérience a été conduite sur les termes anglais biomédicaux traduits vers quatre langues : le français, le japonais, le grec et l'espagnol. La précision varie en fonction de la paire de langues, et il n'est pas clair si la méthode est capable de retrouver des traductions monolexicales pour des termes polylexicaux. De surcroît, la méthode dépend de la taille de la ressource d'entraînement, ne prend pas en compte les positions des n-grammes dans le terme et ne fonctionne que pour les n-grammes présents dans la ressource d'entraînement. Les paramètres extrinsèques diminuent la précision de manière significative.

Nous avons analysé des méthodes permettant d'aligner des termes monolexicaux au niveau des éléments lexicaux et morphologiques. Nous allons maintenant fournir un aperçu des approches permettant d'aligner des termes polylexicaux.

1.3.2 Alignements des termes polylexicaux

Les stratégies utilisées pour les termes polylexicaux sont similaires, car, dans certaines paires de langues, ils sont alignés à des termes composés monolexicaux. Dans ce cas, la segmentation en mots est simple, mais la recomposition en langue cible représente un défi.

13. <https://www.xlstat.com/fr/solutions/fonctionnalites/regression-par-moindres-carres-partiels-pls>

Les méthodes de traduction des termes polylexicaux sont basées soit sur des dictionnaires, soit sur les vecteurs de contexte qui aident quand la traduction n'est pas présente dans un dictionnaire.

1.3.2.1 Méthodes basées sur des dictionnaires

Baldwin et Tanaka (2004) proposent de mener la traduction des syntagmes nominaux du type Nom-Nom (NN) dans une étape à part de la traduction automatique d'un texte. Pour cela, ils sélectionnent les bi-grammes NN à partir de gros corpus (BNC, Reuters pour l'anglais et Mainichi Shimbun Corpus pour le japonais). Puis, les syntagmes sont séparés et l'ensemble de traductions possibles de chaque mot est retrouvé dans des dictionnaires bilingues anglais-japonais. Ces traductions servent à générer les syntagmes NN dans la langue cible qui sont ensuite évalués dans le corpus et les meilleures traductions sont retenues.

La méthode a été évaluée pour les trois catégories de syntagmes classés selon leur fréquence (haute, moyenne et basse). Les résultats obtenus ont été supérieurs à d'autres méthodes.

Dans le même ordre d'idées, Weller et Heid (2012) et Weller *et al.* (2014) procèdent à l'alignement des termes en trois étapes qu'ils illustrent sur l'exemple du terme anglais *glass fibre* :

1. chaque mot du terme polylexical est traduit à l'aide des dictionnaires
 - N1 : *glass* - verre, loupe
 - N2 : *fibre* - fibre
2. les mots traduits sont placés dans des patrons de traduction prédéfinis
 - *N1 N2* - N2 de N1
 - fibre de verre, fibre de loupe
3. les traductions polylexicales obtenues sont filtrées à l'aide du corpus en langue cible
 - fibre de verre, ~~fibre de loupe~~

Comme Fraser *et al.* (2012), Weller *et al.* (2014) transmettent l'information morphosyntaxique (p.ex. le nombre pour les noms et les adjectifs) de la langue source afin de générer les formes correctes en langue cible.

1.3.2.2 Utilisations des vecteurs de contextes

Daille et Morin (2012) suggèrent d'enrichir l'approche basée sur des dictionnaires par les vecteurs de contexte, lorsque le dictionnaire ne contient pas de traduction pour un des termes recherchés. Un vecteur de contexte est constitué des mots co-occurents d'un terme dans le corpus avec leurs scores. Par exemple, le vecteur de contexte pour le terme *antécédent* contient *familial*, *personnel*, *cancer* et d'autres

mots. Dans cette approche, les vecteurs sont construits en langues source et cible, et la similarité des traductions possibles est calculée.

Laroche et Langlais (2010) décrivent l'approche standard de l'extraction de lexique bilingue à partir de corpus comparable en cinq étapes.

Lors de la première étape, un corpus comparable est construit pour chaque terme à traduire. Dans le cadre du projet en question, les auteurs ont extrait des pages correspondantes de Wikipédia.

Ensuite, un vecteur de contexte est construit pour chaque terme. Afin de filtrer les contextes, un anti-dictionnaire peut être nécessaire. En plus, les mots retenus sont validés par un score d'une mesure d'association (la probabilité que les mots soient en relation avec le terme du vecteur).

La troisième étape consiste à projeter des mots du vecteur dans la langues cible à l'aide d'un dictionnaire bilingue ; ceux qui n'ont pas de traduction renseignée, sont ignorés. Un vecteur de contexte est alors calculé pour chaque terme candidat en langue cible (y compris pour les bi-grammes).

Enfin, les scores de la similarité sont calculés pour les vecteurs en deux langues et les traductions candidates dont les vecteurs sont les plus proches de celui source sont retenues.

Alignement des termes polysémiques Bouamor *et al.* (2013) améliorent la méthode standard d'extraction bilingue de termes monolexicaux à base de vecteurs de contextes à partir de corpus comparables. La motivation principale de leur recherche est de proposer les traductions les plus pertinentes pour les termes polysémiques en passant par la désambiguïsation des vecteurs de contexte à l'aide de WordNet.

Les éléments clés de l'approche proposée sont les corpus comparables bilingues spécialisés (FR-EN), un dictionnaire bilingue de langue générale et la ressource sémantique WordNet. D'abord, les vecteurs de contexte sont construits pour la langue source, ensuite les mots des vecteurs sont traduits à l'aide du dictionnaire bilingue.

Les mots monosémiques (qui ont une seule traduction) d'un vecteur, servent à désambiguïser ceux qui sont polysémiques. À cette étape,

« pour chaque entrée polysémique de chaque vecteur, nous (Bouamor *et al.*, 2013) mesurons la similarité sémantique entre toutes les traductions qui lui sont associées et toutes les unités non polysémiques du même vecteur. En fonction des valeurs de similarité, nous obtenons une liste ordonnée de sens ou traductions pour chaque mot polysémique ».

Les auteurs illustrent leur approche par le vecteur "bénéfice" qui

« est construit à partir de corpus comparable spécialisé et contient les mots *action*, *dividende*, *liquidité* et d'autres unités. Lors du transfert de ce vecteur de la langue source (français) à celle cible (anglais), le dictionnaire bilingue propose les traductions suivantes *act*, *stock*, *ac-*

tion, deed, lawsuit, fact, operation, plot, share, dividend et *liquidity* pour traduire respectivement les mots *action, dividende* et *liquidité*. Nous utilisons les unités lexicales non polysémiques *dividende* et *liquidité* pour désambiguïser le mot *action* » (Bouamor *et al.*, 2013).

Dans les résultats obtenus, les traductions les plus appropriées sont classées en tête de la liste : *share* : 0,5236, *stock* : 0,5236, *action* : 0,4256, *act* : 0,2139, *operation* : 0,2045, *plot* : 0,2011, *fact* : 0,1934, *deed* : 0,1594, *lawsuit* : 0,1212.

Alignement par classification Aker *et al.* (2013) créent un système de recherche des traductions combinant deux approches : l'utilisation d'une ressource bilingue pour l'apprentissage des traductions au niveau lexical (y compris, des mots composés) et le recours à des mesures de similarité pour identifier des cognats. Navlea et Todirascu (2011) définissent les cognats comme :

« des équivalences de traduction ayant une forme identique ou partageant des similarités orthographiques ou phonétiques (à cause de leur étymologie commune ou des emprunts). »

L'identification des cognats inclut plusieurs méthodes : le ratio de la sous-chaîne la plus longue, le ratio de la sous-séquence la plus longue, la similarité de Dice, la distance de Needleman Wunsch et la distance de Levenshtein. Ces mesures opèrent au niveau de la similarité graphique des mots. Notamment, les ratio de la sous-chaîne et de la sous-séquence les plus longues, ainsi que la similarité de Dice qui est basée sur cette dernière, prennent en compte des suites de caractères continues (la sous-chaîne) ou discontinues (la sous-séquence) communes entre le terme source et la traduction candidate. La distance de Levenshtein permet de sélectionner les traductions candidates nécessitant le minimum d'opérations (suppression, insertion et remplacement) pour se transformer en terme source.

La méthode nécessite un ensemble de règles de translittération pour traiter des différents alphabets, tandis que notre approche d'apprentissage des n-grammes à partir d'une ressource permet d'éviter cette étape. Nous avons sélectionné cette méthode comme baseline dans l'évaluation de notre système en matière de la recherche des traductions. Nous allons donc détailler les calculs associés à cette méthode dans la description de notre évaluation.

Dans notre travail la préférence est donnée aux méthodes applicables à des corpus comparables où les textes ne sont pas alignés au niveau des segments. En effet, les corpus parallèles sont limités en paires de langues et ne sont pas disponibles en quantité suffisante pour tous les domaines.

Les méthodes d'alignement bilingue des termes décrites dans cette section exploitent plusieurs traits de traduction : la similarité graphique des termes en langues source et cible (le cas des cognats), la proximité des contextes en deux langues, les

correspondances au niveaux des morphèmes (formants néo-classiques) et la présence des mots faisant partie d'un terme composé dans des dictionnaires bilingues.

Les méthodes basées sur la similarité graphique permettent d'identifier des traductions lorsque les langues source et cible sont proches. Elles fonctionnent moins bien pour des langues ayant des alphabets différents, nécessitant des règles de translittération.

Les approches compositionnelles exploitant des morphèmes ou des mots simples nécessitent des ressources, souvent constituées manuellement, et des outils pour segmenter les mots composés en langue source et les reconstruire en langue cible. Ce besoin de ressources pour chaque nouvelle paire de langues rend ces méthodes difficilement exploitables dans le contexte industriel où les langues de traitement et les domaines sont peu prévisibles.

1.4 Conclusion

Dans ce chapitre, nous avons d'abord étudié les notions de terme et terminologie. Nous avons démontré que la notion de terme peut être définie de plusieurs façons, en fonction du contexte théorique. Ainsi, le débat principal se concentre sur la nature de terme : selon la Théorie Générale de la Terminologie, un terme désigne un concept, tandis que dans des approches alternatives, un terme est considéré comme un signe linguistique, associant un sens et une forme. Le domaine du traitement automatique de langues (TAL) se positionne à l'écart du débat théorique présenté ci-dessus qui donne peu solutions pratiques. Étant donné que notre projet se situe dans la perspective sémantique, nous adoptons la définition du terme où il est considéré comme un mot simple ou composé qui sert à dénommer un concept.

Une des questions primordiales du point de vue du TAL est la définition des critères permettant de valider un terme candidat extrait par un système automatique à partir d'un corpus spécialisé. Nous avons donc analysé la grille d'évaluation des termes candidats proposée par le projet TTC. Cette grille nous servira dans l'évaluation de nos résultats.

Ensuite, nous avons analysé les principales méthodes d'extraction terminologique. Il existe en fonction des moyens utilisés, des méthodes statistiques, linguistiques et mixtes. Chaque type de méthodes a ses avantages et inconvénients.

Ainsi, méthodes statistiques permettent d'extraire des termes à partir d'un corpus pour toute langue et tout domaine de spécialité. Toutefois, leur efficacité dépend largement de la qualité du corpus et du volume de données qui peuvent varier en fonction d'une langue et d'un domaine. De plus, ce type de méthodes se concentre sur des termes récurrents, ce qui le rend inexploitable pour détecter des nouveaux termes d'un domaine (des néologismes).

Les méthodes utilisant des ressources linguistiques permettent de pallier aux li-

1.4. CONCLUSION

mites des calculs statistiques. Elles sont moins sensibles au volume de données et sont mieux adaptées pour identification des termes peu fréquents. Malgré ces avantages, les méthodes linguistiques ont besoin des ressources (par exemple, des listes de formants néo-classiques ou des patrons morphosyntaxiques) et des outils, dont des étiqueteurs morphosyntaxiques. Or, ces ressources et outils peuvent être indisponibles pour certaines langues et s'avérer peu performants pour certains domaines.

Les méthodes mixtes, bénéficient des avantages mais aussi héritent les inconvénients des méthodes statistiques et linguistiques.

Nous pouvons constater les mêmes difficultés par rapport à la recherche des traductions pour les termes extraits à l'étape précédente. L'alignement des termes par des méthodes statistiques fournit des meilleurs résultats lorsqu'il est effectué sur des corpus parallèles d'un volume significatif. Or, la disponibilité de ce type de corpus dépend de la paire de langues et du domaine. L'utilisation des corpus comparables nécessite à son tour des ressources complémentaires (des dictionnaires, des listes de formants néo-classiques alignés, etc.).

Pour résumer, l'extraction et l'alignement multilingue des termes à partir de corpus sont actuellement très dépendants des outils et de ressources, et par conséquent, des langues et des domaines. Cela est dû à l'utilisation des informations linguistiques fines, comme des morphèmes et les étiquettes morphosyntaxiques. Pour développer un système d'extraction terminologique véritablement multilingue, il faut s'abstraire de ces informations et minimiser l'utilisation des ressources et des outils. Nous visons donc à développer une méthode d'extraction terminologie robuste, car peu sensible au volume de données, et multilingue, utilisant des corpus comparables.

Une des pistes pour arriver à cela est montrée par (Vergne, 2003, 2005) qui propose un étiquetage endogène de corpus qui ne nécessite pas d'outils ni de ressources, et convient à une large palette de langues.

Dans le chapitre suivant, nous allons détailler les notions et les méthodes relatives à l'ingénierie des ontologies.

Chapitre 2

Enrichissement et élargissement d'une ontologie

La différence entre l'enrichissement et l'élargissement d'une ontologie est assez subtile et ces deux processus sont souvent confondus. Selon les définitions de Petasis *et al.* (2011) et de Suárez-Figueroa (2010), l'enrichissement (*ontology enrichment*) consiste en l'ajout de nouveaux concepts et des nouvelles relations. Nous avons adopté le terme élargissement d'une ontologie (venu de la terminologie interne de l'entreprise Rebus SAS) pour désigner l'ajout des nouveaux éléments lexicaux, libellés ou instances. Ainsi, l'élargissement inclut le peuplement d'une ontologie (*ontology population*) qui signifie l'ajout des instances, donc des entités, aux concepts existants d'une ontologie (Petasis *et al.*, 2011). En réalité, les processus d'enrichissement et d'élargissement sont liés et se complètent.

Dans ce chapitre, nous allons d'abord présenter les notions théoriques liées à ces processus, notamment les concepts avec leur libellés et instances, les relations sémantiques et enfin l'ontologie. Ensuite, nous allons analyser des approches existantes d'enrichissement et d'élargissement des ontologies. Cette analyse nous permet de positionner notre méthode par rapport à l'état de l'art dans le domaine.

2.1 Concepts, relations sémantiques et ontologies

Le développement des technologies sémantiques a favorisé la création des ontologies. Ces structures complexes dans lesquelles les concepts sont liés entre eux par des relations sémantiques hiérarchiques et non-hiérarchiques, peuvent servir à plusieurs objectifs :

- conceptualisation des connaissances d'un domaine ;
- modélisation des processus ;
- raisonnement automatique, etc.

Dans notre travail, les ontologies sont étudiées du point de vue de la recherche d'information multilingue. Nous allons commencer ce chapitre par les notions de concept, de relation sémantique et enfin d'ontologie.

2.1.1 Concept

Le terme *concept* est utilisé dans de nombreuses disciplines. Dans cette partie, nous allons aborder uniquement les définitions relatives à notre domaine de recherche.

La notion de concept joue un rôle clé dans tout projet qui touche à des ontologies. La façon dont les concepts sont définis dans une ontologie conditionne le travail avec cette ontologie, ce que nous essaierons de démontrer par des exemples.

2.1.1.1 Concept dans l'ingénierie des connaissances

L'ingénierie des connaissances utilise la notion de concept pour désigner les unités permettant de développer des modèles qui reflètent des domaines de connaissance ou des scénarios particuliers. De ce point de vue, les concepts servent à représenter des classes d'objets qui participent dans ces modèles.

La définition de concept la plus abstraite est fournie par le W3C dans le projet SKOS (*Simple Knowledge Organization System*) : le concept est une idée ou une notion abstraite ; une unité de pensée¹. L'avantage de cette définition est son applicabilité à une large palette de situations, mais elle fournit très peu d'informations pour quelqu'un qui voudrait créer un concept en pratique.

En même temps, l'utilité pratique de la conceptualisation est évidente : en remontant en niveau d'abstraction, le système réduit le nombre d'unités nécessaires pour créer une représentation de connaissances.

Par exemple, une base de connaissances contient une centaine de modèles de voitures et on veut lier ces modèles aux carburants qu'ils utilisent. Au lieu de créer tous les liens entre les différentes voitures et leurs carburants, nous allons, en premier lieu, réunir tous les modèles de voitures (appelés les instances ou les individus) sous le concept *voiture* et tous les types de carburants sous le concept *carburant*, et on lie ces deux concepts par la relation sémantique *consomme/alimente*.

1. <http://www.w3.org/2004/02/skos/core#Concept>

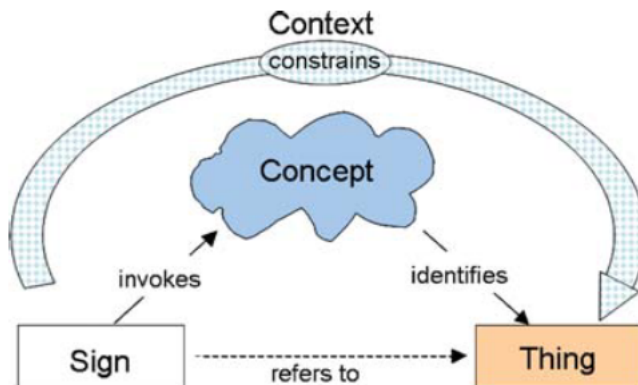


FIGURE 2.1 – Triangle sémiotique (Staab et Studer, 2009)

Il convient de distinguer les libellés et les instances d'un concept. Pour cela, prenons l'exemple du triangle sémiotique (Figure 2.1) proposé par Staab et Studer (2009). Le concept est une notion, un sens. Un libellé est un signe qui désigne ce sens. Par exemple, le mot *chat* désigne la notion de cet animal domestique. Cette notion peut être désignée par des différents mots en plusieurs langues : *cat* (EN), *Katze* (EN), *gato* (en espagnol), etc. Dans le cadre des ontologies spécialisées, les libellés sont des termes du domaine en question. Les objets identifiés par un concept sont ses instances. Les instances peuvent avoir un nom, par exemple, un chat peut s'appeler Marcel. Dans des contextes spécialisés, les instances peuvent correspondre à des produits, des références, des articles, etc. Elles sont présentes dans des textes spécialisés et peuvent apparaître dans les résultats d'extraction terminologique.

Ensuite, nous pouvons regrouper les modèles (instances) qui roulent à l'essence dans un concept *voiture à essence* qui sera lié au concept *voiture* par la relation hyponymique *est-un*, et lier ce nouveau concept par la même relation *consomme/alimente* à celui de l'*essence*. Ainsi, les quatre concepts et deux relations suffisent pour lier un certain nombre d'individus de la classe voiture à des individus de carburant (par exemple, des types d'essence).

Certes, la notion de concept est étroitement liée à celle de synonymie, car un concept peut être désigné par plus d'un mot (libellé). Par exemple, le concept *voiture* peut avoir les étiquettes *voiture(fr)* et *automobile(fr)*. Parfois, un regroupement des synonymes dans un concept est appelé un *synset* (Navigli, 2013; Fellbaum, 1998).

Cañas et Novak (2009) soulignent, à leur tour, l'importance des relations entre les concepts pour la caractérisation de l'un d'entre eux. Les propriétés communes exprimées par ces relations sont nécessaires pour définir un concept particulier. Selon eux, un concept n'existe pas seul, il doit obligatoirement faire partie d'un système conceptuel où les éléments sont liés entre eux (Cañas et Novak, 2009).

Cette définition nous rapproche de la notion de concept dans l'analyse de concepts formels.

2.1.1.2 Notion de concept dans l'analyse de concepts formels

Le concept est la notion centrale de l'analyse de concepts formels (*FCA* ou *Formal Concept Analysis*) décrite par Buitelaar *et al.* (2003) et Blohm (2010).

Blohm (2010) définissent un concept comme une classe d'objets. Un objet peut être, entre autres, une entité (nommée) ou une chaîne de caractères renvoyant à une valeur caractérisant l'objet. Une entité est une chose à laquelle on peut se référer dans le texte (une personne, un objet du monde réel ou une valeur numérique). Les entités peuvent avoir plusieurs formes (dans le sens chaînes de caractères) associées (synonymie) et une même chaîne de caractères peut se référer à plusieurs entités (polysémie). Les objets sont réunis en classes, sur lesquelles il est possible de faire un raisonnement.

Par exemple, dans la phrase *Les tomates sont rouges*, on évoque deux classes : les tomates et les choses rouges (Blohm, 2010). Un objet (une tomate) peut appartenir à plusieurs classes et en être une instance. Une classe a une cardinalité qui correspond au nombre d'instances qui constituent cette classe.

Selon Buitelaar *et al.* (2003), un concept doit avoir trois types d'informations :

1. une définition intentionnelle ;
2. une extension ;
3. un ensemble de réalisations linguistiques (p.ex., des termes multilingues).

L'intention d'un concept est l'ensemble de critères qui permettent d'associer un objet à ce concept, tandis que son extension est l'ensemble des instances associées (Blohm, 2010; Buitelaar *et al.*, 2003).

Ces définitions sont difficilement exploitables dans le contexte d'une ontologie légère destinée à la recherche d'information multilingue.

2.1.2 Vers un choix pragmatique

Dans notre travail, nous allons utiliser deux notions : celle de concept pour parler des sens renseignés dans une ontologie et de concept candidat pour désigner les concepts appris par un système informatique.

2.1.2.1 Concept dans des ontologies spécialisées

Au sein d'une ontologie spécialisée, les concepts définissent les sens des termes, donc du lexique propre au domaine en question. Pour cette raison, il est important de donner une définition terminologique à ces concepts. La présence d'une définition diminue l'ambiguïté et facilite la gestion d'un concept, car elle fournit une base par rapport à la quelle il est possible d'ajouter des libellés en d'autres langues, des synonymes, des abréviations et des instances.

2.1. CONCEPTS, RELATIONS SÉMANTIQUES ET ONTOLOGIES

Prenons l'exemple de l'ontologie EcoLexicon² qui est développée en suivant les principes de la terminologie basée sur des schémas sémantiques (Faber *et al.*, 2005, 2006, 2009; Faber, 2011). Cette ressource est orientée vers l'usage terminologique et destinée à l'analyse sémantique multilingue de corpus. Le concept est alors défini comme une représentation verbale d'une classe de choses ou de phénomènes dans le monde³. En plus de libellés en plusieurs langues, les concepts dans EcoLexicon ont une définition.

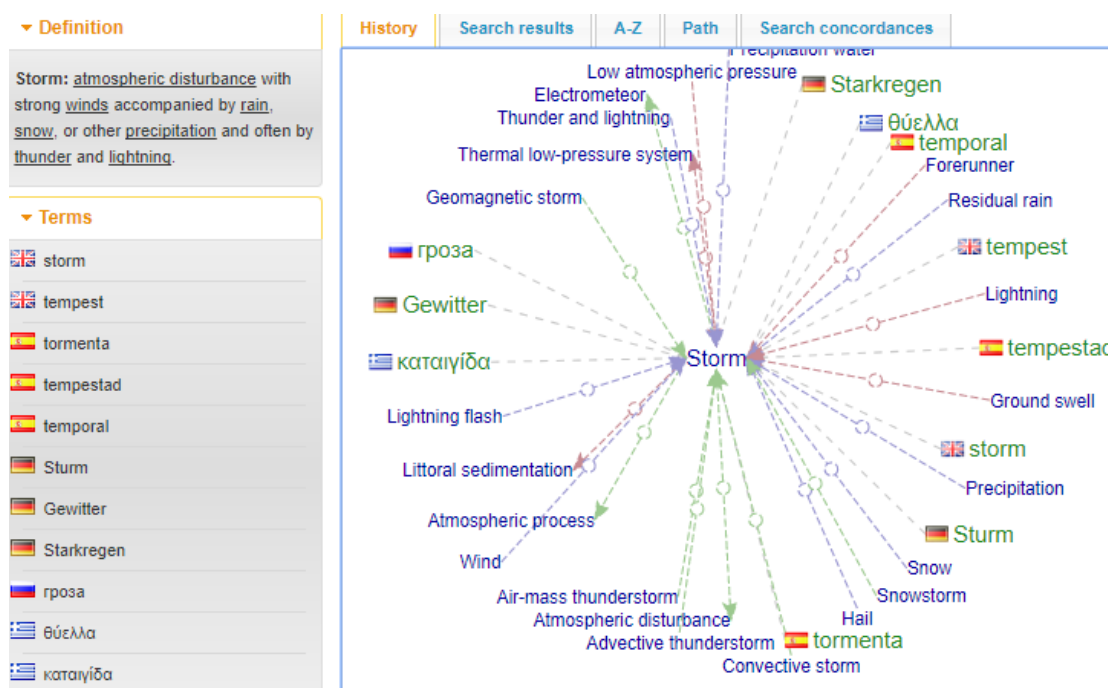


FIGURE 2.2 – Exemple des relations hiérarchiques du concept *STORM* (EcoLexicon)

Par exemple, le concept *STORM* (Figure 2.2) est défini comme « une perturbation atmosphérique avec des vents forts accompagnés de pluie, de neige, ou d'autres précipitations et souvent par le tonnerre et la foudre ». Ce concept regroupe deux libellés en anglais, *tempest* et *storm*, quatre termes en espagnol, trois en allemand, deux en grec et un en russe.

Dans les descriptions du projet, le terme *concept* est souvent associé au *terme*, par exemple *concept/term* (Faber et Castro, 2014), mais le concept désigne une notion qui appartient à une des catégories conceptuelles (Agent, Processus, Patient/Résultat, etc.) définies en accord avec la théorie des schémas sémantiques (Faber *et al.*, 2005; Araúz *et al.*, 2011). Ces catégories conceptuelles se réfèrent au schéma du processus naturel que nous allons détailler dans la section dédiée à EcoLexicon en tant que ressource dans notre projet (Section 3.1.1).

2. <http://ecolexicon.ugr.es>

3. <http://ecolexicon.ugr.es/visual/statistics.html>

2.1.2.2 Notion de concept candidat

Tout comme la notion de terme candidat que nous avons définie dans la Section 1.1.2.2, celle de concept candidat vise à distinguer les concepts ajoutés dans la ressource par un module informatique (Omrane *et al.*, 2011b) et qui n'ont pas encore été validés par un expert humain.

Cette séparation est importante à prévoir, car elle permet d'assigner un statut particulier à ces objets en construction avant de les injecter dans la ressource cible.

Dans les ontologies, les concepts sont liés par des relations sémantiques qui font l'objet de la section suivante.

2.1.3 Relations sémantiques

Les relations sémantiques lient les concepts de l'ontologie de façon hiérarchique (les relations de subsomption) et non-hiérarchique.

2.1.3.1 Relations hiérarchiques

Petasis *et al.* (2011) définissent les relations hiérarchiques comme les relations du type *est-un* qui peuvent être lues dans deux sens : si un concept *A* est en relation d'hyponymie avec le concept *B*, alors le concept *B* est forcément son hyperonyme. Ces relations définissent la structure d'une ontologie entière ou de ses sous-arbres et permettent à la machine d'effectuer des opérations logiques⁴.

Par exemple, dans EcoLexicon (Figure 2.2), le concept *STORM* est lié à son hyperonyme *STORM is type-of ATMOSPHERIC_PROCESS* (une tempête est un type de processus atmosphérique). En même temps, *SNOWSTORM is type-of STORM*, une tempête de neige est un type de tempête.

2.1.3.2 Relations non-hiérarchiques

Les relations non-hiérarchiques englobent toutes les autres relations entre les concepts d'une ontologie (Kister *et al.*, 2011) et diffèrent d'une ressource à l'autre en fonction des besoins associés à la ressource en question. La liste non exhaustive des relations non-hiérarchiques contient les liens de méronymie-holonymie, de cause-effet et d'autres liens conditionnés par l'application de la ressource.

Dans EcoLexicon, les relations méronymiques sont dans une catégorie à part. Pour le concept *STORM* (Figure 2.3), elles incluent la neige, le vent, etc.

4. Pour voir des exemples d'inférence : <http://geneontology.org/page/ontology-relations>

2.1. CONCEPTS, RELATIONS SÉMANTIQUES ET ONTOLOGIES

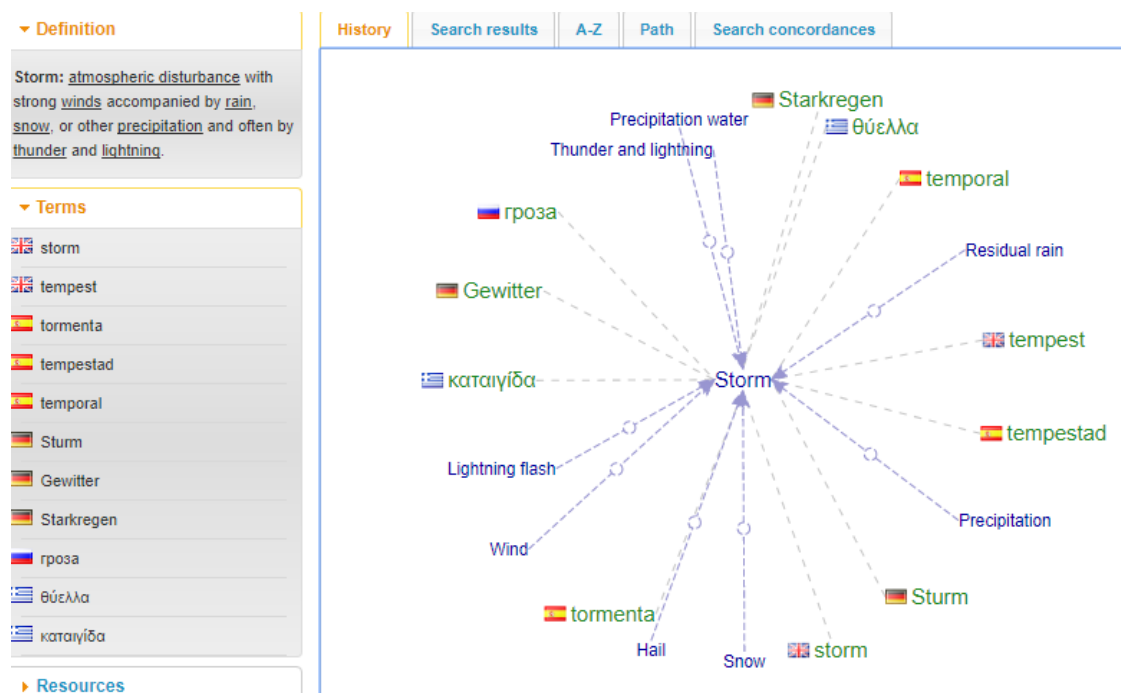


FIGURE 2.3 – Exemple des relations méronymiques du concept *STORM* (EcoLexicon)

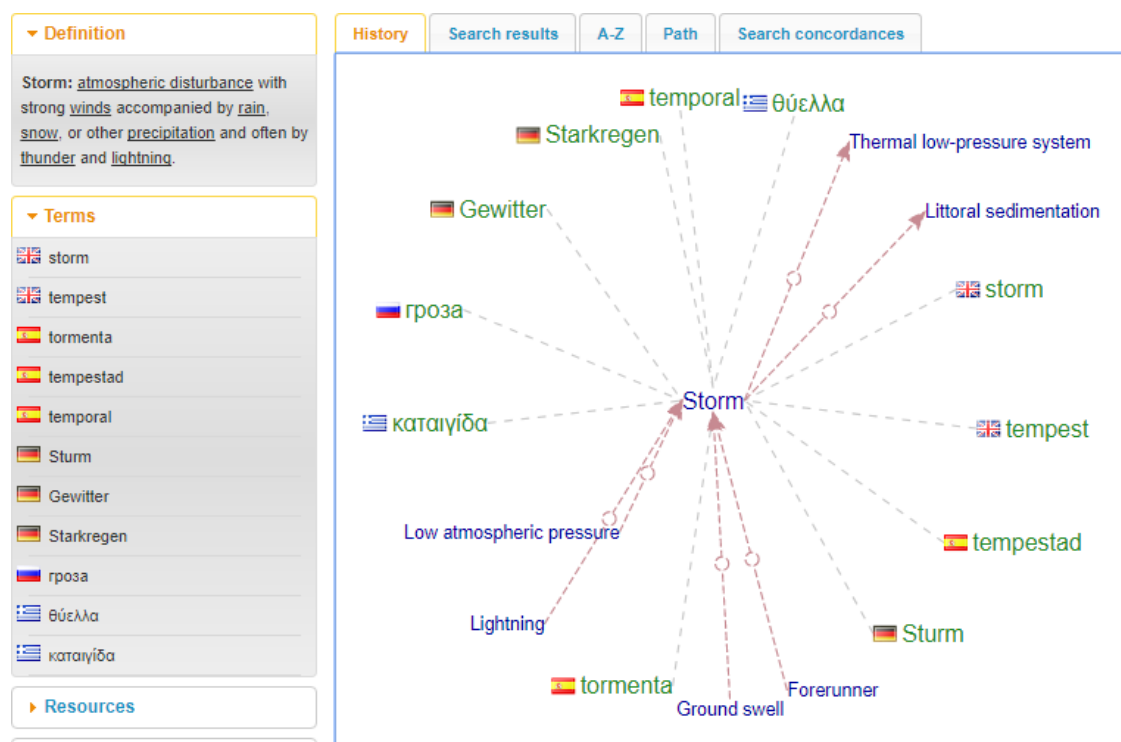


FIGURE 2.4 – Exemple des relations non-hiérarchiques du concept *STORM* (EcoLexicon)

Le concept *STORM* est lié par d'autres relations sémantiques non-hiérarchiques (Figure 2.4), dont :

- le résultat, comme *STORM is result-of THERMAL_LOW-PRESSURE_SYSTEM* (une tempête est un résultat d'un système thermal de basse pression) ou encore la foudre qui résulte d'une tempête (*LIGHTNING is result-of STORM*) ;
- l'attribut, par exemple, *STORM is attribute-of LOW_ATMOSPHERIC_PRESSURE* (une tempête est un attribut d'une pression atmosphérique basse, car c'est la baisse de pression atmosphérique qui en est une des causes) ;
- ou encore d'influence – *STORM affects LITTORAL_SEDIMENTATION* (une tempête influence la sédimentation littorale).

De plus, EcoLexicon contient des relations⁵ de composition (*made-of*), de lieu (*located-at*, *delimited-by*, *takes-place-in*), de mesure (*measures*), de fonction (*has-function*), etc.

2.1.3.3 Instances de relations

Les relations peuvent être extraites à partir d'un corpus (Hristovski *et al.*, 2015). Par exemple, si l'on considère la phrase *Une tempête est un processus atmosphérique qui résulte d'un système thermal de basse pression*, il est possible de détecter deux relations sémantiques : *type-of* et *result-of*. Ces séquences peuvent être retrouvées à l'aide des patrons, comme *X est un Y* pour la relation *type-of* et *X résulte de Z* pour *result-of*. Plus précisément, les instances peuvent être associées avec des patrons lexicaux ou lexicosyntaxiques qui incluent des étiquettes morphosyntaxiques (Maynard *et al.*, 2009; Maynard et Bontcheva, 2014) : *NOM est un NOM* et *NOM résulte de NOM*. L'extraction des instances des relations sémantiques constitue une étape importante qui permet non seulement d'élargir l'ontologie, mais aussi de mieux identifier les concepts et leurs instances dans le corpus. En effet, plus une relation contient d'instances, plus il est possible d'extraire de libellés et d'instances de concepts liés par cette relation à partir du corpus. De cette façon, le patron lexicosyntaxique *NOM est un NOM* appliqué à une phrase *Une tornade est une tempête [...]* peut aider à identifier le libellé *tornado* qui peut soit être ajouté au concept correspondant s'il existe, soit motiver la création d'un nouveau concept.

2.2 Ressources sémantiques et ontologies

Le dictionnaire Trésor de la Langue Française informatisé (TLFi)⁶ distingue l'usage du terme ontologie comme mot abstrait (l'ontologie) et comme nom concret (une ontologie). Le premier cas d'usage désigne l'ontologie comme une branche de la philosophie, tandis que le second traite des réalisations particulières au sein de la

5. <http://ecolexicon.ugr.es/visual/statistics.html>

6. <http://www.cnrtl.fr/lexicographie/ontologie>

théorie philosophique. Guarino *et al.* (2009) ont décrit la même distinction des sens et des usages de la traduction anglaise du terme.

L'informatique et notamment l'intelligence artificielle (IA) ont basé leur définitions d'une ontologie sur le deuxième cas d'usage philosophique.

2.2.1 Ontologie dans la philosophie

La première définition donnée par le TLFi se réfère au XVII^e siècle et à la philosophie aristotélicienne. Selon cette définition, l'ontologie est une « partie de la philosophie qui a pour objet l'étude des propriétés les plus générales de l'être, telles que l'existence, la possibilité, la durée, le devenir ».

Selon la deuxième définition du TLFi, l'ontologie est dans la pensée contemporaine une « partie de la philosophie qui a pour objet l'élucidation du sens de l'être considéré simultanément en tant qu'être général, abstrait, essentiel et en tant qu'être singulier, concret, existentiel ».

La troisième définition du TLFi concerne l'usage du terme *ontologie* en tant que nom désignant un objet concret. Ainsi, une ontologie est une « conception ou visée particulière de l'ontologie ; système ontologique particulier ou problématique ontologique particulière ».

2.2.2 Ontologie dans l'intelligence artificielle

La définition la plus adoptée dans la communauté de l'intelligence artificielle est celle de Buitelaar *et al.* (2003) :

« une ontologie est une spécification formelle et explicite d'une conceptualisation partagée d'un domaine d'intérêt (Gruber, 1993)⁷, où formelle signifie que l'ontologie doit être lisible pour une machine, et partagée nécessite son approbation par un groupe ou une communauté. »

La nécessité de partage est également accentuée par Petasis *et al.* (2011) qui définit les ontologies comme « un moyen de partage et de la réutilisation de connaissances, un contenant pour capter des informations sémantiques sur un domaine ».

Noy et McGuinness (2001a) fournissent une définition plus technique : une ontologie est une description formelle explicite des concepts dans un domaine du discours (classes, appelées parfois concepts), des propriétés de chaque concept décrivant des caractéristiques et attributs du concept (attributs, appelés également rôles ou propriétés) et des restrictions sur les attributs (facettes, ou restrictions de rôles). Cette définition convient surtout aux ontologies qui comprennent plusieurs niveaux d'abstraction, dites lourdes. Par opposition, les ontologies légères, dotées de moins de niveaux, fournissent un « vocabulaire décrivant un domaine de spécialité et une

7. Citation initialement incluse dans la définition de Buitelaar *et al.* (2003).

spécification des sens des termes constituant ce vocabulaire » (Shvaiko et Euzenat, 2013).

2.2.3 Autres types de ressources

Les ressources sémantiques ne sont pas limitées aux ontologies. Il existe également des bases de connaissances, des thésaurus et, dans certaine mesure, des folksonomies.

Bases de connaissances Noy et McGuinness (2001a) définissent une base de connaissances comme une ontologie avec l'ensemble des instances individuelles de ses classes. De ce point de vue, la ressource Prolex (Maurel, 2008) est une base de connaissances, car elle est dotée d'un cœur constitué de concepts et de l'ensemble de noms propres qui en constituent les instances.

KnowNet⁸, un exemple de base de connaissances, contient un grand nombre de concepts annotés par des mots de domaine (*topic words*) et liés par des relations sémantiques. Les termes de domaine proviennent des signatures de domaine (*Topic Signatures*) (Cuadros et Rigau, 2008) qui se rapprochent des sacs-de-mots spécifiques au domaine en question.

Source	#relations
Princeton WN3.0	235,402
Selectional Preferences from SemCor	203,546
eXtended WN	550,922
Co-occurring relations from SemCor	932,008
New KnowNet-5	231,163
New KnowNet-10	689,610
New KnowNet-15	1,378,286
New KnowNet-20	2,358,927

FIGURE 2.5 – Le nombre de relations sémantiques renseignées dans les différentes versions de KnowNet (Cuadros et Rigau, 2008)

En fonction du nombre de mots figurant dans chaque signature de domaine, il existe quatre versions de KnowNet : KnowNet-5, KnowNet-10, KnowNet-15 et KnowNet-20. Par conséquent, le nombre de relations sémantiques s'accroît d'une version à l'autre (Figure 2.5).

Thésaurus Selon Giudicelli (2013), un thésaurus relie des termes entre eux selon des relations précises : synonyme, homonyme, hiérarchie, terme associé ; il normalise l'usage de ces termes et peut définir le terme préconisé. Une ontologie ajoute des règles et des outils de comparaison sur et entre les termes, groupes de termes et relations : équivalence, symétrie, contraire, cardinalité, transitivité.

8. <http://adimen.si.ehu.es/web/KnowNet>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

Folksonomies Les folksonomies sont le produit de systèmes d'annotation collaborative (Delpech *et al.*, 2013), tels que Flickr ou Delicious, qui permettent à une communauté d'utilisateurs d'annoter manuellement des ressources (des pages Web, des photos, etc.) à l'aide de descripteurs (étiquettes ou *tags*). Ces étiquettes sont normées par l'usage fait par la communauté : les tags les plus importants et informatifs sont utilisés plus souvent, tandis que des étiquettes moins importantes peuvent être abandonnées.

Dans notre travail, l'accent est fait sur l'alimentation des ontologies légères en concepts et en termes. Toutefois, notre système peut également servir pour d'autres ressources sémantiques, à part les folksonomies, ou pour des dictionnaires d'un domaine.

Les principales notions décrites dans cette section sont le concept avec les libellés multilingues et instances associées ; les relations sémantiques qui lient des concepts et enfin l'ontologie. Notre travail concerne surtout des ontologies légères, car nous ne prenons pas en compte d'éventuels niveaux d'abstraction logiques (comme des axiomes). Un autre critère important pour notre projet est la présence des libellés, donc termes, en plusieurs langues. Cela est possible grâce l'indépendance des concepts par rapport à des langues : le sens (concept) existe sans dépendre d'une langue et peut être désigné par plusieurs libellés en plusieurs langues.

Au sein d'une ontologie spécialisée, les concepts définissent les sens des termes, donc du lexique propre au domaine en question et les termes, appelés libellés, servent à dénommer des concept. Certaines ontologies peuvent également contenir des instances qui correspondent à des entités, c'est-à-dire à des objets du monde réel. Les concepts d'une ontologie sont liés par des relations sémantiques qui peuvent être hiérarchiques ou non-hiérarchiques.

Dans la section suivante, nous allons analyser les méthodes permettant d'ajouter les éléments sémantiques et lexicaux dans une ontologie.

2.3 Méthodes d'enrichissement et d'élargissement des ontologies

Les ontologies sont des ressources complexes qui contiennent des informations sémantiques et lexicales de plusieurs niveaux, dont obligatoirement des concepts, des relations sémantiques et des libellés de concepts. Elle peut également contenir des instances de concepts et des contraintes logiques (règles et axiomes). Chaque niveau d'une ontologie a ses spécificités qui influences l'ajout de nouveaux éléments dans la structure. Dans le cadre de cette thèse, une attention particulière est accordée au niveau lexical correspondant aux libellés et les instances de concepts (élargissement), et à l'ajout de nouveaux concepts (enrichissement).

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

Par rapport à ces éléments, il existe un certain nombre de méthodes que nous pouvons classer selon les critères suivants :

- nécessité d'assistance par un expert humain ;
- recours à d'autres ressources sémantiques ;
- utilisation de traitements statistiques.

Le premier critère conditionne largement l'applicabilité d'une méthode à l'échelle industrielle. En effet, le recours à un expert (ou plutôt à une équipe d'experts) coûte cher, nécessite du temps et ne garantit pas un résultat cohérent, car il introduit une certaine subjectivité. Au contraire, un processus entièrement automatisé, sans post-traitement, est également assez risqué.

Le recours aux ressources encyclopédiques et sémantiques existantes, comme Wikipédia⁹ ou WordNet¹⁰, n'est pas rare. Ces ressources peuvent aider à désambiguïser certains mots ou concepts et trouver des traductions.

Enfin, des traitements statistiques permettent de gérer un gros volume de données (*Big Data*) et de sélectionner des portions de données plus réduites pour des traitements plus fins (et plus consommateurs de temps). Toutefois, ces méthodes peuvent nécessiter des capacités de calcul importantes et des volumes de données hors de portée des petites et moyennes entreprises.

En partant de ces critères, nous allons présenter les méthodes manuelles, statistiques, ayant recours à des ressources extérieures et mixtes. Certes, aucune méthode n'est limitée à une seule catégorie, mais la classification par l'aspect principal facilite la présentation de l'ensemble.

2.3.1 Méthodes manuelles

L'avantage incontestable des méthodes manuelles est le contrôle permanent du processus et du résultat. En effet, la décision d'ajouter ou de supprimer des concepts ou des relations est prise en fonction du besoin et des priorités du projet. Le travail manuel permet d'obtenir une structure qui correspond aux spécifications, modifiable si besoin.

Cette approche est surtout utile, voire nécessaire, au début du travail sur une ontologie. Souple et rapide à mettre en place, la création manuelle d'une ontologie y intègre des concepts et des relations clés d'un domaine permettant ainsi de tester le système pour lequel cette ontologie est conçue. Notre démarche s'inscrit dans cette catégorie de méthodes, car elles sont, à notre avis, les mieux adaptées à la création des ressources spécialisées nécessitant une haute qualité et répondant à des objectifs précis.

Un des gages de succès pour les méthodes manuelles est le recours à une méthodologie ou à un logiciel permettant la création et la maintenance d'une ontologie. Un

9. <https://www.wikipedia.org/>

10. wordnet.princeton.edu

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

exemple de tels outils est l'éditeur d'ontologies Protégé¹¹, un outil gratuit et open source qui suit les spécifications du W3C (*World Wide Web Consortium*)¹². Protégé est conçu pour une communauté large des utilisateurs comprenant des étudiants, des chercheurs, des entreprises et des organismes publiques. Protégé permet de créer une ontologie, d'y ajouter des concepts et des instances, ainsi que d'introduire des relations et certaines restrictions logiques (règles et axiomes). Par exemple, il est possible de dissocier une classe des autres (une pizza ne peut pas être une glace¹³) ou bien définir que la propriété *poids* d'une pizza doit avoir une valeur > 0 . Il existe également une version en ligne permettant une édition collaborative des ontologies.

Il existe d'autres outils permettant de manipuler une ontologie. Le site¹⁴ de Mike Bergman¹⁵, dirigeant de deux entreprises spécialisés en technologies sémantiques et passionné de ce domaine, recense 185 outils de construction, de gestion et de modification des ontologies. Cette liste constitue un excellent point de départ pour ceux qui cherchent un outil adapté à des besoins d'un projet. Les outils sont classés selon leur complétude et l'usage, par exemple, il sépare des plateformes de développement et des extracteurs terminologiques. Il indique également les systèmes propriétaires et open-source, et avertit lorsqu'un système n'est plus utilisé. Toutefois, nous pouvons constater que l'accent est fait sur la compatibilité des formats. De plus, il ne propose pas d'outil d'élargissement et d'enrichissement multilingue, car la majorité d'outils d'extraction traite une langue à la fois.

Cimiano *et al.* (2006) fournissent un tableau comparatif des outils fondé sur des critères et des fonctionnalités, telles que la gestion des termes, des synonymes, des concepts ou encore la construction d'une hiérarchie des concepts, l'ajout des relations, des axiomes, etc. Dans le même esprit, Petasis *et al.* (2011) proposent une classification des systèmes et des méthodes de l'enrichissement et du peuplement selon les caractéristiques comme :

- le nombre de niveaux d'une ontologie (Buitelaar *et al.*, 2003) concernés (libellés, instances, concepts, relations, etc.) ;
- les pré-requis (des sources de données, des outils) ;
- la méthode de traitement (automatique ou manuelle) ;
- le degré d'autonomie lors du traitement ;
- le maintien de la fiabilité de la structure et l'élimination des redondances ;
- l'applicabilité à d'autres domaines ;
- et le degré d'indépendance vis-à-vis des corpus.

Notre projet vise à travailler avec les couches de termes, de leurs traductions et de concepts. Du point de vue des ressources et des corpus qui en font partie,

11. <http://protege.stanford.edu/>

12. Ce consortium définit les standards du Web sémantique pour favoriser le développement et la mutualisation des technologies qui y sont associées.

13. Nous faisons référence à l'exemple d'ontologie *pizza.owl* disponible avec l'outil.

14. <http://www.mkbergman.com/904/listing-of-185-ontology-building-tools/>

15. <http://www.mkbergman.com/about-mike/>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

nous optons pour un système minimaliste. Cela nous permet d'être indépendants du corpus, du domaine et même de la langue. L'utilisateur reçoit des données pour valider les termes candidats et leurs traductions, et pour les insérer dans l'ontologie. Notre méthode est alors semi-manuelle, avec une intervention limitée et bien outillée.

Dans la section suivante, nous allons faire un point sur quelques outils parmi les plus référencés dans les publications du domaine.

2.3.1.1 NeOn Toolkit

La suite d'outils NeOn Toolkit¹⁶ a été développée dans le cadre du projet NeOn (Haase *et al.*, 2008) qui a pour objectif de recenser l'ensemble d'activités associées au cycle de vie d'une ontologie (Pérez *et al.*, 2008; Suárez-Figueroa, 2010; Suárez De Figueroa Baonza *et al.*, 2012). L'avantage de NeOn Toolkit est son intégration dans la plate-forme Eclipse¹⁷, ce qui facilite son utilisation.

La méthodologie NeOn (Suárez-Figueroa, 2010) présuppose qu'un expert humain, un ingénieur des connaissances, conduise le développement de l'ontologie de A à Z. Ce fonctionnement est avantageux pour autant qu'il soit question d'une ontologie de taille raisonnable conçue pour une tâche précise. Toutefois, cet outil ne contient pas d'extracteur terminologique, ni module de recherche de traductions. C'est à l'expert humain d'annoter le corpus et d'insérer les données dans l'ontologie. Cette absence d'automatisation constitue un frein pour l'utilisation de NeOn Toolkit dans le contexte spécialisé, car le processus de création manuelle d'une ontologie spécialisée multilingue peut s'avérer long et fastidieux. De plus, l'impossibilité d'analyser un corpus du domaine en question peut conduire à des omissions de concepts et de termes, et ne permet pas de suivre les changements du domaine traité.

2.3.1.2 MoKi

MoKi¹⁸ (Bosca *et al.*, 2014) est une plateforme d'édition des ontologies basée sur MediaWiki. Chaque concept y est associé à une ou plusieurs pages de Wikipédia.

La plateforme permet une édition collaborative des ontologies, une gestion des versions et une publication de l'ontologie obtenue en format compatible avec le *Linked Open Data*. De plus, il est possible d'associer des termes en plusieurs langues soit en mode manuel, soit en utilisant le traducteur automatique MS Bing. L'outil propose également des fonctionnalités de recherche de concepts similaires (internes et externes), d'extraction de termes et d'enrichissement de l'ontologie à partir de textes.

16. http://neon-toolkit.org/wiki/Main_Page.html

17. <https://eclipse.org>

18. <https://moki.fbk.eu/website/index.php>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

2.3.1.3 DaFOE

Nous pouvons également mentionner DaFOE¹⁹ (Charlet *et al.*, 2009, 2010), plateforme de construction des ontologies, qui prévoit quatre étapes principales :

- la visualisation des résultats des outils de TAL, tels qu'un extracteur terminologique ;
- l'étude linguistique des résultats, qui permet de sélectionner des candidats retenus ;
- la création d'un réseau termino-conceptuel par l'expert humain ;
- et enfin la création d'une ontologie basée sur ce réseau.

Cette plateforme open-source est conçue pour permettre à l'utilisateur de créer une ontologie à partir de textes ou à partir d'un thésaurus existant.

2.3.1.4 TERMINAE

Omrane *et al.* (2011b) proposent TERMINAE²⁰, une plateforme de construction d'une ontologie à partir de corpus de textes. Les étapes d'extraction terminologique et d'extraction des entités nommées sont suivies d'un tri et d'un ajout manuel des termino-concepts dans le système (Figure 2.6).

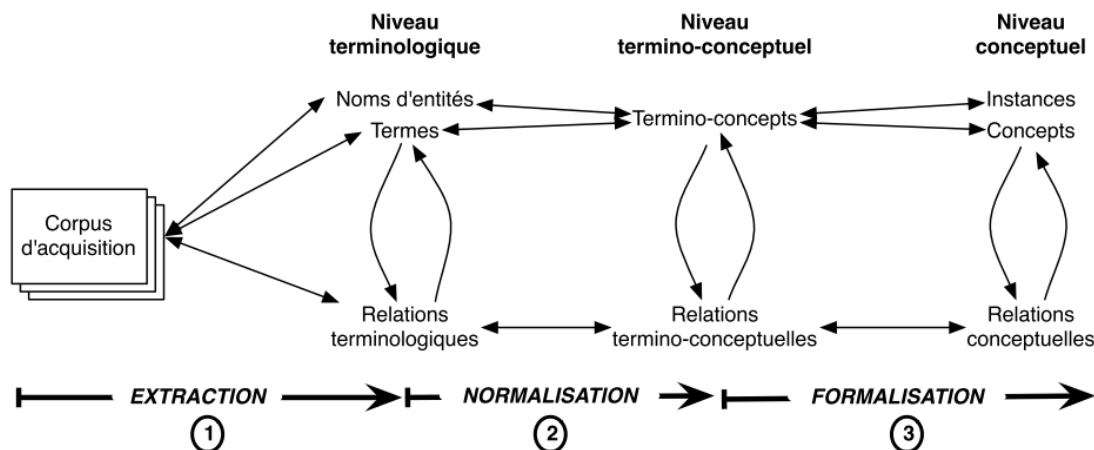


FIGURE 2.6 – Les trois niveaux de connaissance et les trois étapes de la méthodologie de TERMINAE (Omrane *et al.*, 2011b)

On pourrait classer cette méthode comme semi-supervisée, car l'expert travaille sur des données pré-traitées et son rôle consiste à les disposer dans une structure adaptée. De plus, Simperl (2014) suggère d'utiliser la méthode OntoClean (Guarino

19. <https://lipn.univ-paris13.fr/fr/rcln-3/logiciels>

20. http://lipn.fr/terminae/index.php/Main_Page

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

et Welty, 2009) afin d'améliorer la structure de l'ontologie obtenue. La méthode OntoClean permet d'évaluer la cohérence de la structure de l'ontologie en utilisant trois notions philosophiques : le rigidité, l'identité et l'unité (Guarino et Welty, 2009). La rigidité correspond au caractère obligatoire d'une propriété pour l'identification d'un concept. L'identité est la capacité de reconnaître un concept en fonction de ses propriétés. Enfin l'unité désigne l'aptitude à reconnaître toutes les parties appartenant à une entité individuelle. Cette méthode présente plutôt des pistes de questionnement que des calculs informatiques. Elle est destinée à offrir aux ingénieurs des ontologies un moyen d'évaluer l'impact de leurs choix structurels.

2.3.1.5 Analyse de concepts formels

Également dans le cadre d'un raisonnement logique, Cimiano *et al.* (2005) utilisent l'analyse de concepts formels (*Formal Concept Analysis, FCA*) dans laquelle le concept formel est défini par rapport à son contexte formel. Selon leurs définitions, le contexte formel est un triplet des ensembles G, M, I où les éléments de G sont des objets, ceux de M des attributs et I est l'incidence du contexte (relation binaire entre G et M). Le concept formel, à son tour, est un couple (A, B) où l'ensemble des attributs partagés par les objets de A est identique à ceux de B et où l'ensemble des objets de A retrouvent leurs attributs en B . Pour un concept (A, B) , on dit que A est son extension et B son intension. Les concepts sont liés par une relation hiérarchique par rapport à leur contexte formel.

Les objets et les attributs sont placés dans un treillis binaire à partir duquel les concepts formels et les relations hiérarchiques sont identifiés. Le nombre de concepts est ensuite réduit par la suppression des nœuds ayant une extension identique. Après cette étape appelée la transformation, la structure peut être nettoyée de concepts qui sont en trop, pour devenir une hiérarchie compacte de concepts (*Compacted Concept Hierarchy*).

L'expérience a été menée sur des corpus des deux domaines : le tourisme et la finance (Cimiano *et al.*, 2005). Les corpus ont été lemmatisés et les verbes ont été extraits avec leurs dépendances syntaxiques. Les mesures de similarité permettent d'identifier des paires verbe-argument où le verbe constitue le concept de l'ontologie et l'argument (sur une position syntaxique donnée) une instance du concept. Afin d'augmenter le nombre de résultats, les auteurs présupposent que le corpus ne contient pas toutes les paires qui pourraient exister. Pour cette raison, ils assignent une valeur supérieure à zéro pour les paires absentes dans le corpus.

L'avantage incontestable des méthodes manuelles ou manuelles est la qualité de l'ontologie obtenue et le contrôle absolu du processus par les experts. Toutefois, lorsque le volume de données à traiter est important, ces méthodes doivent être outillées afin d'accélérer le processus, sans pour autant perdre en qualité. Certains calculs statistiques peuvent alors être mis en place.

2.3.2 Méthodes statistiques

Dans la famille des approches statistiques, nous regroupons les méthodes d'apprentissage automatique qui ne nécessitent pas d'intervention humaine et qui opèrent sur un volume de données important, par exemple, le Web. Souvent critiquées pour leur opacité et leur besoin important en matière de ressources, les méthodes statistiques constituent toutefois une branche bien établie et qui a de belles perspectives. Il est en effet assez difficile à ce jour d'appliquer cette famille de méthodes dans des entreprises de taille modeste, mais cela pourrait changer dans l'avenir si les capacités de calcul et les ressources deviennent plus accessibles.

Les ressources utilisées par ces méthodes varient : des pages Web (Carlson *et al.*, 2010; Mitchell *et al.*, 2015; Cuadros et Rigau, 2008), des corpus spécialisés (Cimiano *et al.*, 2005), des données issues des réseaux sociaux (Rettinger *et al.*, 2009) ou encore des textes segmentés en arbres syntaxiques (Poon et Domingos, 2010).

2.3.2.1 NELL

Le projet NELL²¹ (*Never-Ending Language Learner*) (Carlson *et al.*, 2010; Mitchell *et al.*, 2015) cherche à apprendre des faits sur le Web et à construire une taxonomie. La particularité de ce système est constituée de son développement continu : lancé en 2010, le système apprend des faits et améliore ses capacités d'apprendre. En effet, le système s'appuie sur une ontologie initiale qui contenait très peu d'instances associés à des concepts et des relations. Par exemple, la catégorie des insectes comptait seulement 10 noms d'insectes (Suchanek, 2014).

Le système complète son ontologie au fur et à mesure de l'apprentissage. NELL utilise trois extracteurs de connaissance qui partagent les informations et se renforcent mutuellement :

- CPL (*Coupled Pattern Learner*) qui opère sur des textes non-structurés à l'aide des patrons lexicosyntaxiques ;
- CSEAL (*Coupled SEAL*) qui extrait les informations à partir de documents structurés avec les patrons lexicosyntaxiques qui peuvent contenir des éléments XML ou HTML ;
- MBL (*Meta-Bootstrap-Learner*) qui coordonne les deux autres extracteurs entre eux et avec l'ontologie centrale.

21. <http://rtw.ml.cmu.edu/rtw/>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES



instance	iteration	date learned	confidence	
pink_bubble_gum is a kind of candy	1036	18-jan-2017	97.7	 
linda_yellen is a fashion model	1034	05-jan-2017	100.0	 
right_internal_jugular is a vein	1036	18-jan-2017	99.4	 
richie_foran is a professor	1034	05-jan-2017	96.0	 
value_percentages is a cheese	1034	05-jan-2017	92.2	 
jimmie_johnson has spouse jeff	1039	07-feb-2017	93.8	 
westchester is a city that lies on the river hudson	1036	18-jan-2017	99.8	 
fema is headquartered in the country us	1039	07-feb-2017	100.0	 
texas_state_aquarium is an aquarium in the city corpus_christi	1039	07-feb-2017	100.0	 
system001 is a subpart of the body within respiratory_tract	1036	18-jan-2017	99.6	 

FIGURE 2.7 – Dernières assertions apprises par NELL (11 février 2017)

En 2015, NELL avait déjà appris et lié entre eux environ 80 millions de faits (Mitchell *et al.*, 2015). Parmi les dernières assertions apprises par NELL (Figure 2.7), on retrouve des faits comme *pink_bubble_gum is a kind of candy* ou encore *right_internal_jugular is a vein*. Toutefois, nous pouvons remarquer que certaines assertions sont fausses : Linda Yellen n'est pas une mannequin, mais une écrivaine, productrice et scénariste.

2.3.2.2 Méthode KnowNet

Cuadros et Rigau (2008) proposent une méthode automatique de construction de ressources sémantiques à partir du Web. La méthode en question est basée sur des signatures de domaine (*Topic Signatures*) qui sont en fait des contextes lexicaux vectoriels utilisés pour référencer un sens d'un mot appartenant à un synset (Lin et Hovy, 2000). Par exemple, la signature *Visite-au-restaurant* peut contenir des mots comme *menu*, *manger*, *payer* et éventuellement *serveur* (Lin et Hovy, 2000). Ces mots sont extraits à partir d'un corpus préalablement trié par domaines et les signatures ainsi construites peuvent s'appliquer à l'identification des documents d'un domaine dans des corpus non-triés.

2.3.2.3 Sémantique distributionnelle

La sémantique distributionnelle cherche à induire une information sémantique à partir des distributions des mots dans des textes. Beltagy *et al.* (2014) utilisent cette approche pour construire des parties d'une ontologie "*on-the-fly*", en partant des phrases similaires pour identifier des relations sémantiques entre les mots.

Un exemple de cette approche est fourni par Beltagy *et al.* (2014) : "pour les phrases *A man is driving a car* et *A guy is driving a vehicle*, on aimerait induire les

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

règles comme $\forall x.\text{man}(x) \Leftrightarrow \text{guy}(x) \mid w1$ qui signifie que les mots *man* et *guy* sont des synonymes avec une certitude $w1$, et $\forall x.\text{car}(x) \Rightarrow \text{vehicle}(x) \mid w2$ qui indique que le mot *car* est un hyponyme du mot *vehicle* avec une certitude $w2$. Dans leur approche, la sémantique distributionnelle a été combinée avec la logique probabiliste qui diffère de la logique descriptive par le fait d'assigner des valeurs dans l'intervalle du 0 à 1, tandis que cette dernière est binaire.

Pour plus d'informations sur les différentes logiques applicables au niveau des axiomes, on peut se référer à Krötzsch *et al.* (2014); Beltağy *et al.* (2014); Lehmann et Fanizzi (2014).

2.3.2.4 Modèles probabilistes relationnels (*PRMs*)

Les modèles probabilistes relationnels (Ławrynowicz et Tresp, 2014; Getoor *et al.*, 2007) basés sur les réseaux bayésiens trouvent leur utilité dans l'extraction des relations sémantiques. Ces modèles permettent d'estimer la probabilité d'un phénomène en tenant compte de plusieurs variables du contexte et diffèrent des réseaux bayésiens par l'exploitation directe de la structure relationnelle de la ressource (Getoor *et al.*, 2007). En d'autres mots, cette approche exploite les relations d'une classe dans la ressource (par exemple, une base de données) pour vérifier leur importance sur un jeu de données. Un des exemples est la classification des films en fonction du critère de leur popularité : la popularité des films d'action dépend du budget, tandis que les documentaires dépendront plutôt de la popularité du réalisateur (Getoor *et al.*, 2007).

2.3.2.5 Modèle caché relationnel infini (*IHRM*)

Ce type de modèles est utilisé pour la recherche de relations entre des classes en fonction d'un certain nombre de paramètres (Xu *et al.*, 2006; Rettinger *et al.*, 2009; Ławrynowicz et Tresp, 2014). Rettinger *et al.* (2009) illustrent l'application de ce modèle à la tâche de prédiction de liens entre les personnes dans une ontologie construite à partir des données issues des réseaux sociaux (FOAF *Friend of a Friend* data). Cette prédiction se fait à la base des propriétés comme la localisation, la présence d'une photo de profil, le nombre d'amis en commun, etc.

Les calculs statistiques présentés ci-dessus permettent d'exploiter des grands volumes de données. Leurs résultats doivent toutefois être validés par des experts humains, car il est difficile de prédire si un algorithme fonctionne bien sur un jeu de données particulier. Dans la sous-section suivante, nous allons analyser les méthodes qui s'appuient sur des ressources existantes afin d'améliorer la précision des résultats.

2.3.3 Méthodes utilisant des ressources structurées

Les ressources linguistiques peuvent servir de base pour la création et l'alimentation des ontologies. Fruit du travail des experts ou d'une communauté, ces ressources sont structurées, riches et, dans certains cas, fournissent des informations de qualité. La qualité d'une ressource dépend de ses créateurs : plus les experts d'un domaine sont impliqués, meilleure est la qualité obtenue.

Parmi les principaux types de ressources et les méthodes qui les utilisent, nous pouvons citer :

- Wikipédia²² et DBpédia (Savary *et al.*, 2013; Chrizman et Itai, 2014; Wu et Weld, 2008);
- WordNet (Navigli, 2013; Wu et Weld, 2008; Mahdisoltani *et al.*, 2015), VerbNet et cadres sémantiques (Coppola *et al.*, 2014);
- thésaurus et dictionnaires spécialisés;
- documents pré-annotés (Kergosien *et al.*, 2009);
- folksonomies et autres types de crowdsourcing (Benz et Hotho, 2014).

2.3.3.1 Utilisation de Wikipédia et de DBpédia

Wikipédia est sans doute la ressource la plus utilisée dans la construction des ontologies. Riche, multilingue, maintenue à jour par sa communauté, cette encyclopédie est utilisée comme :

- un corpus général ou spécialisé, parallèle ou comparable;
- une ressource structurée via la classification de ses catégories;
- une source de relations et d'attributs contenus dans les infoboxes.

DBpedia, quant à elle, réunit quatre taxonomies (Suchanek, 2014) : les catégories de Wikipédia, l'ontologie YAGO, l'UMBEL (*Upper Mapping and Binding Exchange Layer*) de OpenCyc et la taxonomie de la communauté DBpedia.

Toutes les deux ressources sont gratuites et disponibles pour une exploitation industrielle.

Peuplement multilingue d'une ontologie des entités nommées La méthode de Savary *et al.* (2013) porte sur le peuplement multilingue d'une ontologie des entités nommées à partir des sources libres. L'ontologie à enrichir est Prolexbase (Maurel, 2008) où le français est la langue la plus représentée. La méthode utilise Wikipédia (en français, en anglais et en polonais) et la base multilingue de toponymes GeoNames²³.

L'intégration des données commence par l'identification manuelle des correspondances entre les *pivots*, des variantes de noms propres indépendantes d'une langue

22. Bien qu'elle soit plutôt semi-structurée, cette encyclopédie est souvent utilisée pour extraction des connaissances.

23. <http://www.geonames.org/>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

particulière (contrairement aux *prolexèmes* qui correspondent à ce nom propre en plusieurs langues (Maurel, 2008)) de Prolexbase avec les catégories et les infoboxes de Wikipédia et avec les classes de GeoNames. Les articles (avec les liens entre les trois langues) et les instances appartenant aux catégories sélectionnées sont ensuite rattachées aux pivots de Prolexbase. Les auteurs insistent sur l'effet bénéfique du travail simultané sur les trois langues à cette étape, car la chance qu'un pivot soit identifié pour une des trois langues est plus élevée.

Les candidats sont ensuite validés par un lexicographe humain. Ce travail coûteux en matière de temps est nécessaire pour préserver la qualité de Prolexbase en tant que ressource lexicographique fiable.

Construction d'une ontologie du domaine du crime organisé à partir de Wikipédia Chrizman et Itai (2014) proposent d'utiliser deux corpus (un corpus propre au domaine du crime organisé et l'autre général, compilé à partir de Wikipédia), WordNet et DBpédia.

Les auteurs ont essayé d'identifier les expressions polylexicales propres au domaine à l'aide des mesures statistiques appliquées sur les corpus, mais les résultats ont été trop imprécis. Ils ont également rejeté la recherche des concepts et des relations à partir de WordNet, car la base contient trop peu de concepts du domaine.

L'exploitation de Wikipédia n'est pas non plus une garantie de qualité : les auteurs ont observé des imprécisions au niveau de l'annotation des articles par catégorie (seulement 55 % des articles associés à la catégorie O_CRIME ont été réellement pertinents). Les meilleurs résultats de recherche du lexique propre au domaine ont été obtenus par l'entraînement du système sur un corpus de résumés d'articles annotés (manuellement) comme pertinents ou, à l'inverse, comme non relatifs au sujet.

L'utilisation de Wikipédia peut paraître une solution facile et peu coûteuse. Toutefois, dans l'analyse des méthodes citées ci-dessus, nous retrouvons des inconvénients liés à cette ressource. Notamment, il est difficile de construire des corpus spécialisés à partir des articles de Wikipédia, car leur pertinence n'est pas garantie. Même s'il est possible d'extraire des articles en plusieurs langues, les traductions peuvent varier considérablement en termes de qualité et de volume d'une langue à l'autre.

Le recours à des ressources sémantiques que nous allons présenter ci-après présente moins d'incertitude, car ces ressources sont mieux structurées. Toutefois, la disparité de la quantité de données disponible pour des langues riches en ressources (comme l'anglais ou le français) d'une part et peu dotées (comme le bulgare ou le kazakh) d'autre part persiste également dans cette catégorie de ressources.

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

2.3.3.2 Exploitation des ressources sémantiques

La base lexicale WordNet²⁴, développée depuis 1986 à Princeton, est un réseau sémantique dans lequel les mots anglais sont classés dans des groupes de synonymes appelés synsets. Construite manuellement, la base compte à ce jour environ 117 000 synsets (Figure 2.8). Ils sont liés entre eux par les liens hiérarchiques (hypéro- et hyponymie) et par les liens non-hiérarchiques (méronymie, antonymie).

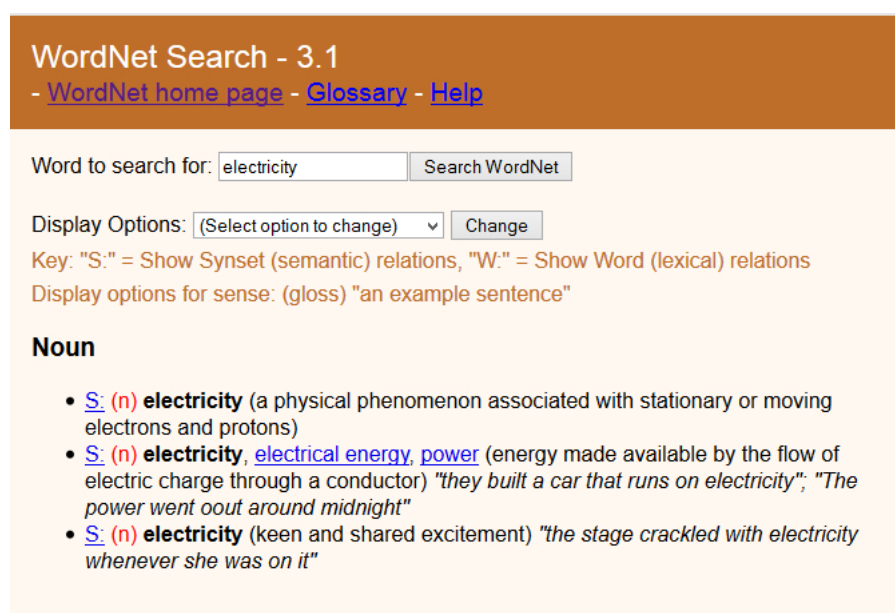


FIGURE 2.8 – Les synsets de WordNet pour la requête *electricity*

Bien que certains synsets puissent être étiquetés dans un domaine spécifique (par exemple, sport), la base est très générale. WordNet est disponible uniquement pour l'anglais. Toutefois, il existe une famille de ressources pour de nombreuses langues qui suivent le modèle de WordNet²⁵, dont Open Multilingual Wordnet, BalkaNet, EuroWordNet, Russian Wordnet, etc. WordNet est disponible pour la consultation en ligne et pour le téléchargement, y compris pour un usage commercial, mais ce n'est pas toujours le cas des ressources pour les autres langues.

Des ressources sémantiques, comme WordNet ou VerbNet, servent souvent à la compilation d'autres ressources similaires. WordNet est souvent combiné avec Wikipédia.

C'est notamment le cas de BabelNet²⁶ qui, selon un des ses créateurs, (Navigli, 2013), est un dictionnaire encyclopédique à large échelle. BabelNet est construit à partir de plusieurs ressources, dont Wikipédia et WordNet. Après un alignement, les

24. <http://wordnet.princeton.edu/>

25. <http://globalwordnet.org/wordnets-in-the-world/>

26. <http://babelnet.org>

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

synsets de WordNet et les pages de Wikipédia ont donné les concepts de BabelNet. La ressource représente les connaissances sous forme de graphes qui contiennent les nœuds (concepts) et les relations entre ces nœuds. Chaque nœud contient des unités lexicales pour les langues traitées (50 langues) qui sont réunies en synsets Babel.

Dans le même esprit, le générateur des ontologies Kylin *Kylin Ontology Generator (KOG)* construit les ontologies à partir des infoboxes de Wikipédia et associe sa hiérarchie à celle de WordNet (Wu et Weld, 2008). Nous pouvons également citer les exemples de YAGO, une autre ontologie construite automatiquement à partir de Wikipédia, WordNet, Universal WordNet et Geonames (Suchanek, 2008; Mahdisoltani *et al.*, 2015); Coppola *et al.* (2014) emploient les cadres sémantiques définis dans les ressources comme Wordnet ou VerbNet pour apprendre des ontologies.

2.3.3.3 Collection d'annotations de documents et thésaurus

Kergosien *et al.* (2009) ont abordé le problème en deux temps. Tout d'abord, ils construisent le noyau de l'ontologie à partir d'une collection d'annotations de documents (en format XML) où les balises contenues dans une balise-parent sont supposées avoir une relation sémantique. Les relations d'hypo et d'hyponymie sont extraites à partir de textes des documents à l'aide de patrons lexico-syntaxiques (*est un*, etc.).

Le travail est effectué sur les textes appartenant au domaine de la géographie. Les entités nommées présentes dans les textes, ont été identifiées et les mots qui les précèdent ont été considérés comme termes candidats qui peuvent donner lieu à de nouveaux concepts. L'ontologie créée au départ peut être enrichie par ces concepts. Ils sont vérifiés dans le thésaurus RAMEAU²⁷ (Répertoire d'autorité-matière encyclopédique et alphabétique unifié), utilisé dans l'indexation des documents par les bibliothèques françaises et belges, qui sert aussi à enrichir l'ontologie par ses sous-arbres.

Il faut tout de même remarquer que des documents annotés sont assez rares et la méthode est donc difficilement applicable à d'autres domaines et ressources.

Les méthodes utilisant des ressources structurées peuvent fournir des résultats plus fiables que les approches statistiques pures. Toutefois, la faiblesse de ces méthodes est leur dépendance à ces ressources. Ainsi, elle ne peuvent être multilingues que si les ressources le permettent. Si cela ne pose pas de problème pour les langues bien outillées, les langues peu dotées sont moins concernées par ces méthodes. Notre objectif est de réduire le volume de données nécessaire et de permettre ainsi l'utilisation de notre système pour les langues peu dotées en ressources.

Dans la section suivante, nous allons présenter le dernier groupe de méthodes qui repose sur une mixité d'approches et de ressources.

27. <http://rameau.bnf.fr/informations/rameauenbref.htm>

2.3.4 Méthodes mixtes

Les méthodes mixtes combinent plusieurs traitements et ressources. Ainsi, la méthode OntoLearn (Velardi *et al.*, 2006, 2013) a recours à un corpus, à des ressources sémantiques, comme WordNet, et à l'expertise humaine. BOEMIE (Petasis *et al.*, 2011) à son tour présente une mixité de ressources : la méthode fonctionne sur des corpus textuels et multimédia. Nous avons également inclus dans cette catégorie les patrons lexicosyntaxiques qui, combinés avec d'autres moyens, permettent d'alimenter des ontologies par des nouvelles relations et concepts.

2.3.4.1 OntoLearn

Velardi *et al.* (2006) ont créé la méthode OntoLearn qui permet de construire une ontologie du domaine à partir de textes. Le traitement commence par l'extraction des termes candidats polylexicaux à partir de corpus contenant des textes spécialisés comparés à des textes généraux. Ensuite, le système cherche des définitions pour les termes retrouvés. La recherche se fait dans des ressources en ligne et dans les documents du corpus (extraction de définitions). Si une définition est identifiée, le système teste sa pertinence et applique des patrons morphosyntaxiques pour extraire des relations, notamment la relation d'hyponymie. Si toutefois le système ne trouve pas de définition pour le terme complet, il essaie de trouver des définitions pour ses composantes, y compris dans WordNet. Il est alors nécessaire de résoudre des cas ambigus. Si aucune stratégie n'a donné de résultat, la définition est fournie par un expert humain. Les relations hiérarchiques permettent de regrouper les termes en arborescences qui doivent être rattachées à l'ontologie de base.

<H> be the syntactic head of $w_h w_k$ (e.g. *festival*, *debt*)
 <M> be the syntactic modifier of $w_h w_k$ (e.g. *jazz*, *long-term*)
 <GNC> be the gloss of the new complex concept S^{hk}
 <HYP> be the selected senses of <H> (e.g. respectively, *festival#1* and *debt#1*)
 <MSGHYP> be the main sentence¹⁴ of the gloss of <HYP>
 <MSGM> be the main sentence of the gloss of the selected sense for <M>

Two examples of rules for generating GNCs are:

- If $\text{sem-rel} = \text{Topic}$, $\langle \text{GNC} \rangle ::= \text{a kind of } \langle \text{HYP} \rangle, \langle \text{MSGHYP} \rangle$, relating to the $\langle \text{M} \rangle$, $\langle \text{MSGM} \rangle$.
 e.g.: $\text{GNC}(\text{jazz festival}) = \text{"a kind of festival, a day or period of time set aside for feasting and celebration, relating to the jazz, a style of dance music popular in the 1920s"}$.
- If $\text{sem-rel} = \text{Attribute}$, $\langle \text{GNC} \rangle ::= \text{a kind of } \langle \text{HYP} \rangle, \langle \text{MSGHYP} \rangle, \langle \text{MSGM} \rangle$.
 e.g.: $\text{GNC}(\text{long term debt}) = \text{"a kind of debt, the state of owing something (especially money), relating to or extending over a relatively long time"}$.

FIGURE 2.9 – Exemples de définitions en langue naturelle Velardi *et al.* (2006)

2.3. MÉTHODES D'ENRICHISSEMENT ET D'ÉLARGISSEMENT DES ONTOLOGIES

Velardi *et al.* (2006) ont facilité le travail de validation manuelle des résultats en générant des définitions en langue naturelle pour chaque concept (Figure 2.9). Ces définitions sont utilisées par des experts pour évaluer la performance du système OntoLearn.

Velardi *et al.* (2013) proposent une nouvelle pondération qui améliore les résultats du système. Ils appellent cette méthode OntoLearn Reloaded. Le nouveau système permet d'identifier les concepts et les relations hiérarchiques à partir de textes, et propose une nouvelle pondération pour la création des arborescences. Notamment, ils proposent un module d'extraction des relations hiérarchiques indépendant d'un domaine. Pour cela, les termes candidats extraits à partir du corpus sont associés avec les concepts de WordNet les plus généraux indépendants d'un domaine. Les définitions pour ces termes candidats sont extraites à partir du corpus, ce qui permet ensuite d'induire des taxonomies qui, après une étape d'élagage, deviennent les arborescences de l'ontologie.

2.3.4.2 BOEMIE

Le projet BOEMIE (*Bootstrapping Ontology Evolution with Multimedia Information Extraction*) (Petasis *et al.*, 2011) vise à extraire les objets à partir de ressources hétérogènes, telles qu'un corpus de textes ou un corpus multimédia. Pour cela, le système sépare les concepts de leurs réalisations. La particularité de la méthode BOEMIE consiste à utiliser deux types de concepts :

- les concepts de niveau intermédiaire (*mid-level concepts* ou *MLCs*), des concepts "primitifs" qui n'ont pas de réalisations directes ;
- les concepts de haut niveau (*high-level concepts* ou *HLCs*) ou des concepts « composés » qui sont construits à partir de MLCs. Ces concepts représentent des objets ou des événements réels et possèdent des réalisations directes.

Cette structure de l'ontologie implique une approche particulière pour l'enrichissement et le peuplement. En effet, le module qui extrait des instances de concepts et des relations ne fonctionne que sur le niveau des concepts "primitifs". Cela rend le système plus robuste par rapport à des changements possibles dans l'ontologie, car ce module a besoin d'être adapté uniquement suite à des évolutions dans les concepts « primitifs ». Les concepts « composés » sont enrichis à leur tour par l'inférence de la meilleure explication (ou l'abduction) à partir des concepts « primitifs » et des relations entre ces concepts.

Cette approche a deux avantages. Premièrement, les concepts « composés » sont toujours en cohérence avec les définitions formelles des concepts "primitifs" associés. Deuxièmement, l'enrichissement des concepts "composés" permet de répondre au besoin de maintenir la consistance de l'ontologie, par exemple, en respectant certaines contraintes (des règles, etc.).

2.3.4.3 Patrons lexicosyntaxiques

Les patrons lexicosyntaxiques regroupent les variations autour des patrons de Hearst (1992). Ces patrons sont appelés lexicosyntaxiques car ils mélangent les éléments lexicaux et les étiquettes morphosyntaxiques. Par exemple, la relation *dbpedia:author* contient les patrons suivants :*[[adj]] book by, [[det]] novel by, was written to, etc..* Les patrons de la ressource PATTY sont enrichis par l'information de l'ontologie, notamment par le concept (*person*), et par l'indice de confiance. (Dragos et Jaulent, 2010) illustrent cela sur l'exemple de relation d'association entre les concepts *endofin* et *early endosome* (EN) dans la phrase *Endofin is associated with early endosomes*. Le patron *is/VER associated/VER with/PREP* peut servir pour l'identification d'autres couples de concepts entretenant une relation similaire.

Utilisés seuls ou combinés avec des patrons appris du corpus (Etzioni *et al.*, 2005; Carlson *et al.*, 2010), les patrons lexicosyntaxiques sont un moyen efficace pour apprendre des relations hiérarchiques et non-hiérarchiques (Maynard *et al.*, 2009; Maynard et Bontcheva, 2014).

Les patrons lexicosyntaxiques peuvent être mis en correspondance avec les patrons de conception des ontologies (*Ontology design patterns*) (De Cea *et al.*, 2008; Maynard *et al.*, 2009) afin de faciliter la réutilisation des ressources. Cette approche est implémentée dans le système SPRAT (*Semantic Pattern Recognition and Annotation Tool*) (Maynard *et al.*, 2009) qui permet par ailleurs d'extraire des synonymes, des instances et des propriétés.

Les patrons lexicosyntaxiques peuvent être appris et organisés sous forme d'une taxonomie (Nakashole *et al.*, 2012). Les résultats de leur système sont accessibles dans la ressource PATTY²⁸ qui réunit 350 000 relations extraites de Wikipédia et organisées en synsets.

Pour générer ces patrons, les auteurs utilisent les dépendances syntaxiques fournies par *Stanford parser* (De Marneffe *et al.*, 2006) qui donnent un meilleur résultat que les étiquettes morphosyntaxiques (Nakashole *et al.*, 2012). Les entités nommées sont prises à partir de ressources comme YAGO2 (Hoffart *et al.*, 2011) et servent à la détection des syntagmes nominaux susceptibles de former les patrons de type *sujet-relation-objet*.

Tout comme l'extraction des termes, l'ajout des nouveaux libellés et instances, ainsi que des nouveaux concepts ou relations peut se baser sur des calculs statistiques ou sur des ressources, y compris sur des ressources ouvertes connectées *Linked Open Data*. Les approches hybrides qui utilisent des ressources, des calculs et l'expertise humaine sont sans doute les mieux adaptées aux projets applicatifs.

Toutefois, l'automatisation d'ajout des nouveaux éléments dans une ontologie spécialisée est une tâche extrêmement difficile, étant donné le nombre de niveaux

28. <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/patty/>

d'information qui peuvent être présents. En effet, chaque niveau d'information nécessite des traitements particuliers : des patrons lexicosyntaxiques pour les relations ou encore l'analyse distributionnelle pour des concepts liés par des relations hiérarchiques. Cette complexité et diversité de traitements favorise le recours à des experts humains, dont le travail nécessite du temps, des outils et des ressources (y compris, financières).

Il est alors particulièrement difficile d'ajouter des libellés en plusieurs langues. D'une part, l'extraction terminologique multilingue qui constitue la base pour élargissement et enrichissement, est très dépendantes des langues et des domaines (comme nous l'avons montré dans le chapitre précédent). D'autre part, constituer une équipe d'experts qui maîtrisent plusieurs langues et qui peuvent communiquer en une langue commune est une tâche difficile à atteindre, notamment en raison des coûts et de la disponibilité des experts.

2.4 Conclusion

Dans ce chapitre, nous avons défini les notions de concept avec ses libellés et ses instances, des relations entre les concepts et, finalement, l'ontologie en tant que ressource sémantique spécialisée. Nous avons démontré que la complexité des ontologies peut atteindre des niveaux très abstraits, comme c'est le cas des axiomes dans les ontologies dites *lourdes*. Afin d'éviter toute confusion, nous limitons la portée de notre analyse à des ontologies appelées *légères* destinées à la recherche d'information multilingue. Ces ontologies contiennent (au minimum) des concepts, des libellés multilingues qui y sont associés et les relations sémantiques²⁹. En fonction de la structure d'une ontologie, elle peut également contenir le niveau des instances.

Une ontologie doit refléter les connaissances sur un domaine de spécialité. Elle doit alors intégrer des nouvelles notions dès leur apparition. Ce besoin résulte en processus de l'élargissement des ontologies et de leur enrichissement. Rappelons que l'enrichissement consiste en ajout de nouveaux concepts et de nouvelles relations, tandis que l'élargissement se traduit par l'ajout des libellés et des instances aux concepts existants. Nous avons constaté que les méthodes d'élargissement et d'enrichissement des ontologies peuvent être classées en quatre catégories : manuelles, statistiques, utilisant des ressources structurées et enfin les méthodes mixtes.

Toutefois, chaque catégorie de méthodes présente plusieurs limites. Notamment, les méthodes statistiques manquent de fiabilité et nécessitent une validation par un expert humain. Des méthodes qui utilisent des ressources structurées sont applicables uniquement aux domaines et aux langues pour lesquels de telles ressources existent. De plus, dans le contexte industriel, des ressources peuvent engendrer des coûts complémentaires (notamment, à cause des licences). Même le recours à des ressources

29. L'extraction des relations sémantiques constitue un riche sujet de recherche que nous ne pouvons pas inclure dans cette thèse.

2.4. CONCLUSION

externes ne peut garantir une bonne qualité de d'enrichissement et d'élargissement. L'expert humain reste alors au centre de la problématique. Lorsqu'il s'agit des traitements multilingues, il doit également parler plusieurs langues ou être au sein d'équipe plurilingue. Les outils que nous avons évoqué dans ce chapitre, sont essentiellement conçus pour des traitements monolingues. Le plus souvent, les systèmes sont évalués en anglais, car pour il existe un grand nombre de ressources pour cette langue. De plus, une méthodes est souvent adapté à une ontologie cible, car elle dépend de la structure de cette ontologie. Cela rend difficile son application à d'autres ressources.

De plus, les systèmes actuels ne tiennent pas compte des liens morphologiques qui existent entre des termes sémantiquement liés. Par exemple, certains morphèmes peuvent designer des relations d'antonymie (par exemple, *constitutionnel* et *anti-constitutionnel*, d'autres termes peuvent être dérivés d'un libellé d'un concept avec lequel il gardent une certaine proximité (*purifier* – *purification*, etc.). Tout comme dans le cas de l'extraction des termes à faibles fréquences, la difficulté du regroupement des termes morphologiquement proches consiste à éviter une analyse morphologique fine et des ressources extérieures.

L'ensemble de ces limites permet de conclure que l'élargissement et l'enrichissement multilingue d'une ontologie spécialisée est loin d'être un problème résolu. En vu de l'émergence des systèmes de recherche sémantiques et multilingues, ce sujet de recherche a une grande importance. Actuellement, la création et la maintenance des ontologies constituent un vrai verrou technologique pour des entreprises qui développent des systèmes d'analyse sémantique, surtout multilingue.

Il est nécessaire d'alléger le processus, en utilisant moins de ressources et d'outils et en privilégiant des solutions véritablement multilingues et multi-domaines.

Deuxième partie

Utilisation des ressources endogènes pour l'élargissement et l'enrichissement d'une ontologie

Notre projet de recherche vise à développer une méthode permettant d'élargir et d'enrichir une ontologie de spécialité à partir d'un corpus multilingue. L'analyse de l'état de l'art présentée dans la partie précédente a mis en évidence les difficultés intrinsèques de ces processus et les limites des méthodes existantes. Dans l'ensemble des étapes de traitement, nous avons identifié cinq verrous à lever par rapport auxquels est définie notre méthode :

1. L'applicabilité à plusieurs langues et domaines.
2. La capacité d'intégrer des nouvelles notions et termes dès leur apparition.
3. La prise en compte des variantes orthographiques, des fautes de frappe et des synonymes.
4. L'intégration de la terminologie multilingue.
5. La pré-organisation des résultats pour le traitement par un expert humain.

Le besoin d'une méthode multilingue et multi-domaine peut être satisfait par la minimisation d'utilisation des outils et des ressources externes, car ils ne sont pas disponibles pour toutes les langues ni tous les domaines.

L'intégration des nouveaux termes et concepts dépend à son tour de la capacité d'une méthode à repérer des termes à faible fréquence à partir de corpus, sans contrainte quant à taille de celui-ci.

Le troisième besoin nécessite également l'identification des termes à faible fréquences, mais cette fois-ci, ils doivent être mis en relation avec leur formes de base (bien orthographiées ou termes existants).

Le besoin suivant, c'est-à-dire l'intégration de la terminologie multilingue dans l'enrichissement et l'élargissement d'une ontologie, permettrait d'éviter la répétition de ces processus pour chaque langue à part. Toutefois, satisfaire le besoin d'applicabilité à plusieurs paires de langues et des domaines, il est nécessaire d'utiliser des corpus comparables sans lien de traduction.

Enfin, il est nécessaire de faciliter le traitement des résultats par un expert. Pour cela, la méthode visée doit proposer des regroupements des termes extraits à partir d'un corpus spécialisé par rapport aux concepts de l'ontologie, sans toutefois dépendre de la structure de celle-ci.

Nous avons défini une ontologie source et un corpus multilingue comparable comme deux ressources minimales nécessaires pour l'élargissement et l'enrichissement d'une ontologie. Pour répondre à notre problématique, notre méthode utilise des textes bruts sans informations linguistiques avancées. Les ressources morphologiques, lexicales et morphosyntaxiques sont substituées par des informations générées en cours d'analyse.

L'exploitation d'un corpus spécialisé multilingue implique deux étapes de traitements : l'extraction des termes et la recherche de leurs traductions. Ensuite, les termes alignés en plusieurs langues doivent être organisés par rapport aux concepts définis dans l'ontologie afin d'aider l'expert humain à ajouter des nouveaux concepts,

des libellés multilingues à chaque concept ou encore des instances.

Chaque étape répond à un problème particulier et peut être exploitée séparément. Pour cette raison, nous avons développé des modules permettant d'effectuer des traitements de manière indépendante. Assemblés, ces modules constituent un système complet remplissant ainsi l'objectif de la thèse.

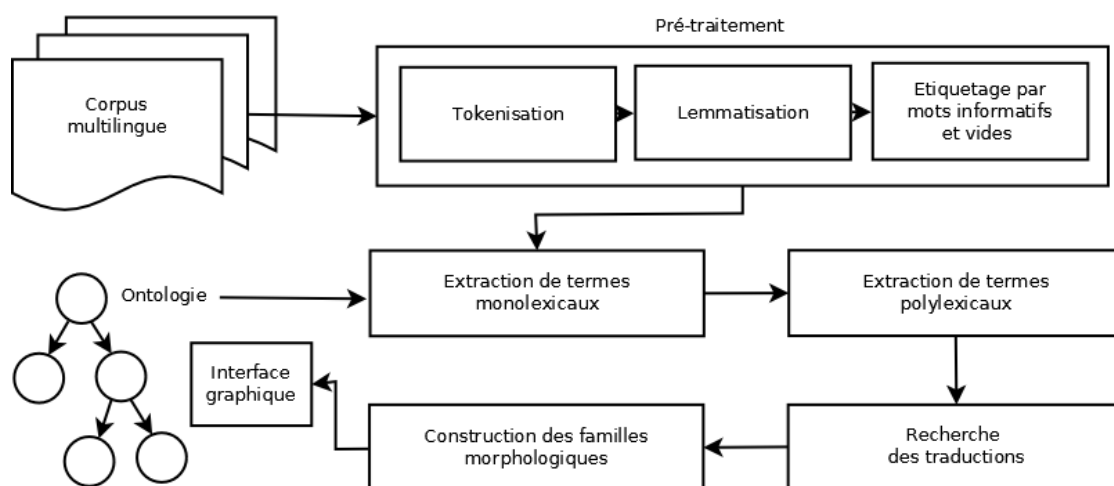


FIGURE 2.10 – Architecture de notre système

La Figure 2.10¹ illustre notre système : après le pré-traitement incluant la tokenisation, la lemmatisation et l'annotation par des mots informatifs et vides (Vergne, 2005), le corpus multilingue parallèle ou comparable est prêt pour être exploité par les modules d'extraction terminologique. Dans un premier temps, nous identifions des termes monolexicaux. Pour cela, nous utilisons les libellés des concepts de l'ontologie spécialisée, qui correspondent aux termes du domaine en question. Si toutefois l'ontologie ne contient pas de libellés en langue de traitement, nous pouvons utiliser des termes extraits à partir de corpus à l'aide des méthodes statistiques. Cette étape nous fournit la matière nécessaire pour l'identification des termes polylexicaux. Les listes de termes extraits séparément pour chaque langue sont transmises au module qui, en exploitant des traductions déjà présentes dans l'ontologie, y identifie des traductions pour les nouveaux termes et pour ceux déjà connus par l'ontologie. Enfin, les termes sont regroupés dans des familles morphologiques, c'est-à-dire, construites à base de proximité de la structure des mots, facilitant leur association à des termes déjà présents dans l'ontologie ou la création de nouveaux concepts.

1. Afin de faciliter la lecture des schémas, nous avons créé un guide de lecture (Annexe A).

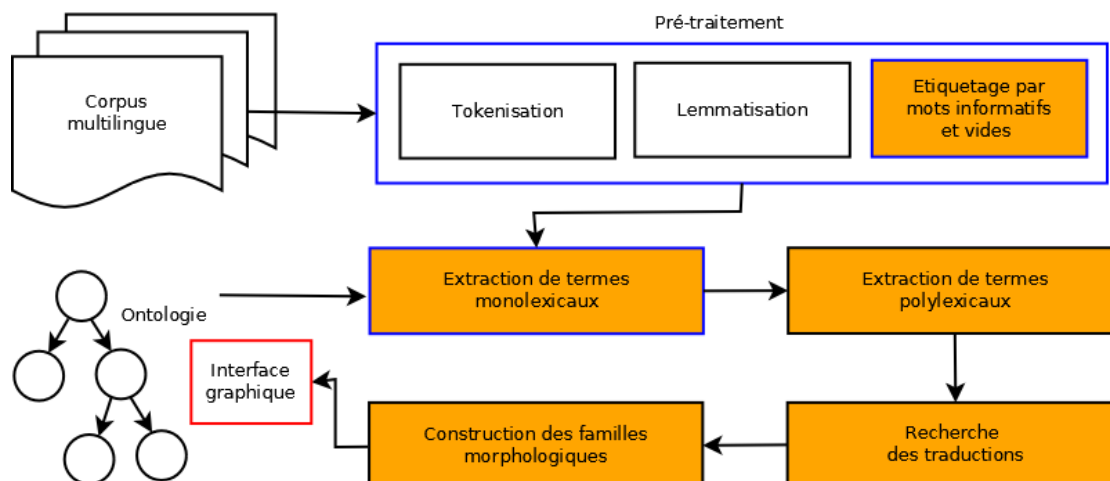


FIGURE 2.11 – Modules implémentés de notre système

Le premier prototype (les modules avec le contour bleu sur la Figure 2.11) a été développé en langage de programmation Java (version 1.7), avec le framework Spring² et une base de données MySQL³. Ce prototype suivait des conventions de développement définies au sein de l'entreprise Rebuz SAS et enregistrait les données issues des traitements du corpus dans une base de donnée relationnelle.

Toutefois, pour des raisons d'une faible vitesse de traitement causée par la sauvegarde dans la base de données, nous avons décidé de développer le deuxième prototype en scripts (Python 2.7) avec des entrées et des sorties en fichiers *.txt*. Les modules concernés sont en orange. Les résultats des évaluations dans les chapitres suivants sont obtenus avec ce deuxième prototype. L'interface graphique pour utilisateur est le seul élément du système qui n'est pas développé dans le cadre de cette thèse.

Dans cette partie, nous allons d'abord présenter les ressources utilisées pour l'évaluation à savoir un corpus multilingue et des ressources sémantiques. Nous allons également expliquer le pré-traitement du corpus. Ensuite, chaque module sera détaillé dans un chapitre dédié et son évaluation sur les ressources indépendantes y sera présentée. En effet, il est nécessaire d'évaluer la performance de chaque étape indépendamment des autres afin de mesurer sa précision et son impact sur les résultats finaux. Enfin, nous fournirons les résultats pour la méthode appliquée dans sa totalité sur une ontologie. Ces résultats permettront de vérifier la performance de notre système et de confirmer son applicabilité pour l'industrie, ainsi que de mieux cerner les limites d'application.

2. <https://spring.io/>

3. <https://www.mysql.com/>

Chapitre 3

Ressources

Cette thèse a été menée en collaboration avec l'entreprise Rebus SAS où la structure et le contenu des ressources sémantiques sont confidentielles. Nous avons donc opté pour des ressources non-confidentielles disponibles pour la recherche. Les objectifs du projet définissent les exigences envers les ressources qui doivent alors être multilingues, spécialisées et de bonne qualité.

En partant de ces critères, nous avons sélectionné deux ressources principales : l'ontologie EcoLexicon¹ (Faber et Castro, 2014; Araúz *et al.*, 2011) et le corpus parallèle des segments de brevets PatTR (Wäschle et Riezler, 2012b,a). Les deux ressources sont issues de projets de recherche et sont bien documentées. Il s'agit des principales ressources utilisées pour le développement et l'évaluation du processus dans son intégralité.

Afin de mieux évaluer certaines étapes du processus, nous allons recourir à deux ressources complémentaires : les thésaurus Agrovoc² et GEMET³, et deux sous-corpus PatTR annotés manuellement par les termes mono et polylexicaux.

3.1 Ontologies et thésaurus spécialisés

Dans le chapitre précédent consacré aux méthodes d'enrichissement et d'élargissement, nous avons constaté que de nombreuses méthodes ont été évaluées sur une seule ressource (ontologie, taxonomie ou thésaurus). Or, une méthode doit être applicable à un grand nombre de ressources, sans contrainte de structure de celles-ci. À notre avis, cet objectif nécessite l'évaluation d'une méthode sur plusieurs ressources, langues et domaines.

Pour cette raison, nous avons retenu trois ressources multilingues traitant des

1. <http://ecolexicon.ugr.es/en/index.htm>

2. <http://aims.fao.org/standards/agrovoc/about>

3. <http://www.eionet.europa.eu/gemet/about?langcode=en>

domaines de l'environnement : l'ontologie EcoLexicon et les thésaurus Agrovoc et GEMET.

3.1.1 Ontologie EcoLexicon

La base de connaissances terminologique EcoLexicon est développée par le groupe de recherche LexiCon⁴ à l'Université de Grenade.

La ressource réunit 3 547 concepts et 19 712⁵ termes du domaine de l'environnement en sept langues : l'espagnol, l'anglais, l'allemand, le français, le grec moderne, le russe et le danois. Dotée d'une interface graphique et accessible gratuitement pour consultation, elle s'adresse aux traducteurs, aux terminologues, aux spécialistes et aux étudiants travaillant dans les domaines de l'environnement.

Dans EcoLexicon, les concepts sont dotés de multiples informations linguistiques et sémantiques (Faber et Castro, 2014; Castro et Faber, 2016) :

- une définition terminologique ;
- des termes (libellés) en plusieurs langues ;
- des informations morphosyntaxiques sur les libellés (partie du discours, genre, nombre, etc.) ;
- des types de libellés (terme principal ou alternatif, variante, acronyme, etc.) ;
- des contextes dans les corpus utilisés par le projet (concordancier) ;
- des ressources associées (documentaires, visuelles, etc.) ;
- des catégories sémantiques (par exemple, l'*éventail fluvial* appartient aux catégories des côtes et des côtes modifiées) prédéfinies et hiérarchisées dans la ressource ;
- des relations sémantiques hiérarchiques (par exemple, une rivière est un type de cours d'eau)⁶, méronymiques (les rives font partie de la rivière) et non-hiérarchiques (fluvial est un attribut de rivière) avec d'autres concepts.

EcoLexicon correspond aux critères d'appartenance à une ontologie spécialisée *légère* : ses concepts reflètent les éléments d'un domaine de spécialité et sont liés par des relations hiérarchiques et non-hiérarchiques décrites ci-dessus, sans règles ni axiomes que contiennent les ontologies dites *lourdes*.

EcoLexicon se distingue des autres ressources par son interface interactif ergonomique qui facilite son utilisation par des utilisateurs de tous horizons, dont les traducteurs et des étudiants. Cette interface permet une navigation dynamique entre des différents concepts et leurs relations, en fournissant l'ensemble des informations pour chaque concept sélectionné.

La ressource qui est également originale à cause de ses fondements théoriques,

4. <http://lexicon.ugr.es/>

5. La version utilisée dans notre projet date du mai 2016. En novembre 2016, le nombre de concepts s'élève à 3 601 et celui de termes à 20 192 (<http://ecolexicon.ugr.es/visual/statistics.html>)

6. Les exemples sont extraits du concept *RIVER* dans l'EcoLexicon.

3.1. ONTOLOGIES ET THÉSAURUS SPÉCIALISÉS

a été conçue manuellement par des terminologues et est de très haute qualité. Elle fait l'objet des développements constants (Castro et Faber, 2016) et des nouveaux concepts et libellés y sont ajoutés manuellement au fur et à mesure.

Fondements théoriques La ressource suit les principes de la terminologie basée sur les schémas sémantiques et de la gestion de terminologie orientée processus (Faber *et al.*, 2005, 2006, 2009; Faber, 2011). La théorie des schémas sémantiques remonte aux travaux de (Fillmore, 1982, 1985) et est implémentée dans FrameNet (Fillmore et Atkins, 1998).

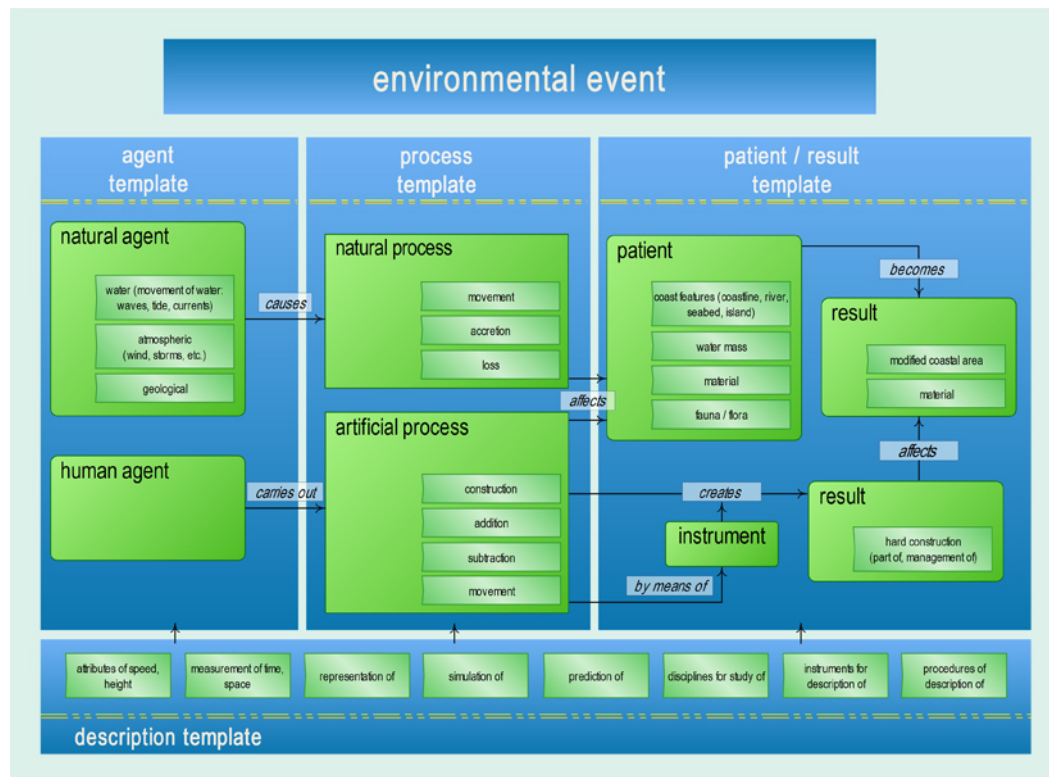


FIGURE 3.1 – Le schéma sémantique d'événement naturel⁷

Les schémas sémantiques qui représentent des processus généraux et spécifiques d'un domaine, sont organisés sur plusieurs niveaux, où les niveaux abstraits sont réutilisables pour d'autres processus. Le schéma d'un événement (Figure 3.1) contient les rôles comme *agent*, *processus*, *patient/résultat* (un agent effectue une action qui fait partie d'un processus et affecte un patient ou produit un résultat). La partie de la *description* qui fournit des informations complémentaires (l'espace, le temps, etc.). Un schéma général permet donc de décrire un événement quel que soit le domaine.

7. <http://lexicon.ugr.es/fbt>

3.1. ONTOLOGIES ET THÉSAURUS SPÉCIALISÉS

EcoLexicon est développé pour le domaine de l'environnement, repose sur le schéma sémantique d'un événement environnemental, où le rôle *agent* donne lieu à deux sous-classes : *agent naturel* (géologique, atmosphérique, etc.) et *agent humain*. Un *agent naturel* agit dans un *processus naturel*, tandis qu'un *agent humain* participe à un *processus artificiel* en utilisant éventuellement un *instrument*. Les processus affectent des *patients* et créent des *résultats* qui peuvent interagir entre eux (par exemple, un bâtiment peut influencer le sol). La *description* quant à elle recense les relations sémantiques complémentaires que les concepts peuvent avoir dans le schéma et il faut noter que ces relations peuvent être spécifiques à un domaine (Faber *et al.*, 2009). La description se rapproche alors des méta-données relatives à un processus, à un agent ou à un patient.

Développement de la ressource EcoLexicon a été développé à partir de deux types de ressources : des corpus sélectionnés manuellement (l'approche *bottom-up*) et une collection des ressources terminologiques du domaine (l'approche *top-down*) (Faber *et al.*, 2006). Les corpus ont été construits manuellement à partir de sources fiables. Le processus de composition des corpus prenait en compte quatre critères : la quantité, la qualité, la simplicité et la documentation. La quantité se réfère à la taille et la complétude des textes, tandis que la qualité des textes est définie par le prestige de l'auteur et de son travail. La simplicité garantit la compréhension du texte par l'expert et enfin la documentation contient des méta-données sur la collection. Les ressources terminologiques, dont les thésaurus et les dictionnaires, ont été comparées et évaluées afin d'obtenir un jeu de données représentatif pour le domaine.

3.1.2 Agrovoc et GEMET

Les ressources complémentaires sont utilisées dans l'évaluation indépendante des différents modules du système afin d'exclure la dépendance des résultats d'une ressource particulière, notamment lors de l'extraction terminologique et de la recherche de traductions.

Nous avons choisi deux ressources, GEMET et Agrovoc, qui correspondent aux exigences en matière de qualité, de disponibilité, du domaine spécialisé et de la présence des libellés multilingues. Les deux ressources sont classées dans la catégorie des thésaurus, c'est-à-dire du vocabulaire multilingue spécialisé contenant des synonymes, des termes généraux ou spécifiques, sans toutefois arriver à une structure taxonomique. Toutefois, nous avons constaté que GEMET et Agrovoc contiennent des concepts, donc des sens, et que les termes multilingues représentent leurs libellés. C'est en effet la seule exigence structurelle imposée dans notre projet, car nous allons utiliser ces libellés multilingues dans notre méthode. L'absence des relations sémantiques possibles dans une ontologie, ne constitue pas d'obstacle pour l'utilisation de ces ressources dans nos traitements, ni pour l'évaluation.

3.1. ONTOLOGIES ET THÉSAURUS SPÉCIALISÉS

Thésaurus GEMET Le thésaurus GEMET (*The GEneral Multilingual Environmental Thesaurus*) contient plus de 5 000 concepts avec des termes associés en 19 langues dans environ 30 domaines liés à l'environnement. Il a été développé par l'Agence européenne d'environnement et son réseau (Eionet) dans l'objectif de fournir une ressource terminologique multilingue de référence facilitant la coopération internationale dans ses domaines de compétence.

Les informations d'un concept, par exemple *OSMOSIS* (osmose), contiennent :

- l'URL⁸ du concept : <http://www.eionet.europa.eu/gemet/concept/5949> ;
- sa définition accompagnée d'une source ;
- les libellés (termes) multilingues ;
- les concepts liés, généraux (*PHYSICOCHEMICAL PROCESS*, processus physicochimique) ou plus spécifiques (*REVERSE OSMOSIS*, osmose inverse) ;
- les domaines (*biologie*, *chimie*) ;
- le groupe, *chimie*, *substances*, *processus* ;
- les concepts équivalents dans d'autres ressources connectées ouvertes *Linked Open Data* (Agrovoc⁹ et UMTES¹⁰) et le lien vers un article de Wikipédia.

Cette liste n'est pas exhaustive et peut varier d'un concept à l'autre. Il faut noter que les thèmes et les groupes dans GEMET sont prévus pour préserver la cohérence de la ressource du point de vue terminologique. Il s'agit donc des catégories propres à GEMET qui ne sont pas forcément destinées à être exploitées par des applications extérieures. Cela illustre que les structures des ressources sémantiques peuvent varier. Il est donc important de s'abstraire d'une ressource concrète pour éviter une dépendance d'une méthode à cette ressource. Dans notre projet, nous utilisons les concepts et leurs libellés multilingues, sans dépendre des catégories sémantiques, ni relations.

Thésaurus Agrovoc Le thésaurus Agrovoc est développé par l'Organisation des Nations Unies pour l'alimentation et l'agriculture. Il contient 32 000 concepts avec des termes en 23 langues issus des domaines différents, tels que l'agriculture, la protection de l'environnement, l'alimentation, etc.

La ressource contient également plusieurs renseignements pour les concepts. Prenons encore une fois l'exemple du concept *OSMOSIS* (osmose) :

- l'identifiant du concept dans Agrovoc (http://aims.fao.org/aos/agrovoc/c_5442) ;
- une définition ;
- des libellés multilingues, en distinguant les termes préférentiels (*osmosis*, EN) et alternatifs (*osmotic flow*, EN) ;
- le lien vers le concept plus générique (*DIFFUSION*, diffusion ou perméat) et

8. L'identifiant permettant d'accéder aux informations de façon non-ambiguë sur Internet et dans les ressources connectées ouvertes *Linked Open Data*.

9. http://aims.fao.org/aos/agrovoc/c_5442

10. https://sns.uba.de/umthes/de/concepts/_00018782.html

3.1. ONTOLOGIES ET THÉSAURUS SPÉCIALISÉS

- des concepts proches, sans préciser s'il s'agit des sens plus spécifiques, comme *REVERSE OSMOSIS* (osmose inverse) ;
- les liens sémantiques vers d'autres concepts (osmose *influence* stress osmotique ou encore membrane semi-perméable *utilise* osmose) ;
- les correspondances exactes ou partielles dans d'autres ressources *Linked Open Data* ;
- d'autres méta-données.

Encore une fois, la ressource dépasse les limites d'un thésaurus. En effet, un effort considérable a été fourni pour transformer un thésaurus en un service des ontologies spécialisées en agriculture (*Agricultural Ontology Service*) (Liang *et al.*, 2006). Notamment, le passage vers le niveau conceptuel a nécessité une analyse de cohérence entre les libellés en plusieurs langues, car certains termes n'avaient pas de traductions exactes en d'autres langues. Ce problème était dû à l'utilisation de l'anglais comme langue principale dans le thésaurus original (Liang *et al.*, 2006).

Agrovoc implémente un schéma d'extension du *Simple Knowledge Organisation System* (SKOS)¹¹ (Miles *et al.*, 2005), plus précisément SKOS-XL¹². SKOS fournit un modèle de données centré sur les concepts qui sont définis par des identifiants uniformes de ressources (*Unique Resource Identifier* ou URI). Cela permet de l'intégrer dans *Linked Open Data*. Les données sont organisées en triplets RDF *Resource Description Framework*¹³, ce qui permet leur consultation via le langage SPARQL¹⁴.

L'aperçu du concept *OSMOSIS* en format RDF (Figure 3.2) permet de visualiser les différentes informations citées précédemment¹⁵ : les libellés multilingues préférentiels (*Ósmosis* en espagnol, *osmosis* en anglais) avec leurs références uniques, les libellés alternatifs (*osmotic flow*, EN) avec leurs références, les renvois aux concepts identiques ou similaires dans d'autres ressources, etc. Il faut noter que ce format n'est pas destiné à la lecture par un utilisateur humain, mais sert surtout à la consultation par une machine.

Nous allons maintenant réunir les informations statistiques sur les trois ressources sémantiques.

3.1.3 Statistiques liées aux ressources ontologiques

Pour résumer, les trois ressources sont relatives au domaine de l'environnement et contiennent les termes et leurs traductions en plusieurs langues, dont le français, l'anglais et l'allemand. La Table 3.1 permet de comparer ces ressources en matière de volume : le nombre de termes varie entre 622 (EcoLexicon, FR) et 40 635 (Agrovoc,

11. <https://www.w3.org/2004/02/skos/>

12. <https://www.w3.org/TR/skos-reference/skos-xl.html>

13. <https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/>

14. <https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>

15. Pour consulter l'intégralité des informations, veuillez consulter le lien suivant : http://aims.fao.org/aos/agrovoc/c_5442.rdf

3.2. CORPUS PATTR

```

- <rdf:RDF>
- <rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_5442">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <prefLabel xml:lang="es">Osmosis</prefLabel>
  <prefLabel xml:lang="en">osmosis</prefLabel>
  <prefLabel xml:lang="fr">Osmose</prefLabel>
  <prefLabel xml:lang="de">Osmose</prefLabel>
  <prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_es_1299510781827"/>
  <prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_en_1299510781808"/>
  <prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_de_1299510781790"/>
  <prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_fr_1299510781870"/>
  <related rdf:resource="http://aims.fao.org/aos/agrovoc/c_26870"/>
  <altLabel xml:lang="de">Osmotischer Fluss</altLabel>
  <altLabel xml:lang="en">osmotic flow</altLabel>
  <altLabel xml:lang="es">Flujo osmótico</altLabel>
  <inScheme rdf:resource="http://aims.fao.org/aos/agrovoc"/>
  <exactMatch rdf:resource="http://aims.fao.org/aos/asfa/c_7655"/>
  <exactMatch rdf:resource="http://www.eionet.europa.eu/gemet/concept/5949"/>
  <closeMatch rdf:resource="http://dbpedia.org/resource/Osmosis"/>
  <broader rdf:resource="http://aims.fao.org/aos/agrovoc/c_24936"/>
  <related rdf:resource="http://aims.fao.org/aos/agrovoc/c_26870"/>
  <created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2012-02-03T16:57:45Z</created>
  <modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2013-08-27T21:01:47Z</modified>
  <altLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_es_1299510784281"/>
  <altLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_de_1299510784688"/>
  <inDataset rdf:resource="http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc"/>
  <makeUseOf rdf:resource="http://aims.fao.org/aos/agrovoc/c_27605"/>
  <hasProperty rdf:resource="http://aims.fao.org/aos/agrovoc/c_35835"/>
  <hasStatus>Published</hasStatus>
</rdf:Description>
</rdf:RDF>

```

FIGURE 3.2 – Extrait des informations du concept *OSMOSIS* d’Agrovoc en RDF

EN).

Ressource	Nombre de termes			Nombre de traductions		
	FR	EN	DE	FR-EN	FR-DE	EN-DE
EcoLexicon	622	4 180	3 601	666	1 009	5 594
GEMET	5 113	5 241	5 094	5 022	5 129	5 111
Agrovoc	38 362	40 635	30 433	51 724	62 399	57 171

TABLE 3.1 – Statistiques des ressources ontologiques

Cette différence nous permettra d’évaluer l’impact de la taille de ressource sur la performance de notre méthode.

3.2 Corpus PatTR

Le corpus des segments parallèles de brevets PatTR (*Patent Translation Resource*) est construit à l’Université de Heidelberg dans le cadre d’un projet sur

3.2. CORPUS PATTR

l'amélioration de la traduction automatique des brevets via l'entraînement des modèles de traduction à part pour les différentes sections d'un brevet (titre, résumé, description et revendications) (Wäschle et Riezler, 2012b,a).

3.2.1 Caractéristiques techniques

Source de données Le corpus des segments parallèles de brevets PatTR est compilé à partir du corpus MAREC¹⁶. PatTR contient des segments de brevets alignés automatiquement par l'outil GARGANTUA¹⁷ (Braune et Fraser, 2010). Par conséquent, les segments alignés peuvent présenter des différences d'une langue à l'autre. La précision de l'alignement varie d'une section de brevet à l'autre. Ainsi, les titres et les revendications sont mieux alignés que les segments de descriptions et des résumés (Wäschle et Riezler, 2012a).

Organisation Le corpus est organisé d'abord par paires de langues et ensuite par partie du brevet. Dans chaque dossier (par exemple, *PatTR/EN-DE/titles/*), on trouve deux fichiers de segments parallèles, à raison d'un fichier par langue, et un fichier de méta-données.

Méta-données Les méta-données associées à chaque segment contiennent plusieurs informations :

- le numéro du brevet et, le cas échéant, celui de la famille de brevets ;
- la date du dépôt du brevet ;
- un ou plusieurs codes de la classification internationale des brevets¹⁸ (CIB) de l'Organisation Mondiale de la Propriété Intellectuelle (OMPI). La liste de catégories sélectionnées est disponible dans l'Annexe B.

Paire de langues	Nombre de segments
EN-DE	22 998 357
EN-FR	18 764 038
FR-DE	5 110 262

TABLE 3.2 – Nombre de segments dans le corpus PatTR¹⁹

Statistiques Le volume de données varie pour les paires de langues (Table 3.2) et pour les catégories de la CIB (cf Annexe C).

16. <http://www.ifs.tuwien.ac.at/imp/marec.shtml>

17. <http://gargantua.sourceforge.net/>

18. <http://www.wipo.int/classifications/ipc/fr/>

19. <http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/>

Défis pour notre projet Le corpus PatTR représente deux défis principaux : sa taille et la diversité de domaines. Il est en effet difficile d’estimer le volume de données disponibles pour chaque catégorie de la CIB sans analyser les méta-données. Même si les méta-données facilitent le travail, un certain volume de travail manuel est inévitable. À moins d’être un spécialiste de la CIB, il est nécessaire de se renseigner et de choisir les catégories qui peuvent correspondre à une ontologie en particulier.

Cette opération manuelle est guidée par l’intuition et il convient de valider le choix des sous-corpus fait à ce stade. Toutefois, le corpus est trop grand et n’est pas destiné à être lu par un humain. Toute validation manuelle semble alors insuffisante, ce qui nous a conduit vers le développement d’une approche de sélection semi-automatique.

3.2.2 Choix des sous-corpus pour l’évaluation

Afin de réduire le volume de travail manuel et de valider les sous-corpus, nous avons analysé leur couverture par les termes associés aux concepts de l’ontologie et vérifié que leurs occurrences ne soient pas limitées à des concepts généraux.

Notre démarche consiste à sélectionner manuellement quelques catégories de la CIB et extraire les sous-corpus correspondants. Ensuite, nous pouvons calculer la couverture du corpus par les concepts de l’ontologie et procéder à une évaluation semi-automatique des occurrences des concepts.

CIB	EcoLexicon
C02F Traitement des eaux usées	3.2.5.1 Traitement des déchets
B09C Traitement des déchets solides et remise en état des terres contaminées	3.2.5.2 Traitement des eaux
H01(G,M) Éléments électriques de base ,	3.2.5.4 Gestion de la qualité des sols
C01G Chimie inorganique, H02(J,M) Génération, conversion, ou distribution de l’énergie électrique, C25(B,C,D,F) Processus électrolytiques ou électrophorétiques et appareillage associé	3.5 Ingénierie énergétique

TABLE 3.3 – Sélection manuelle des catégories de la CIB ²⁰

Sélection manuelle des catégories de la CIB La difficulté principale est de sélectionner les catégories de la CIB permettant d’élargir et d’enrichir l’ontologie EcoLexicon. Vu la taille importante du corpus, il serait difficile d’évaluer la couverture pour toutes les catégories de la CIB. Nous avons donc commencé par regarder

²⁰. Les titres complets des catégories étant très complexes, nous avons pris les titres généraux (par exemple, *Éléments électriques de base* est le titre général de la catégorie H01)

3.2. CORPUS PATTR

les domaines définis dans l'ontologie EcoLexicon. Les domaines sélectionnés sont énumérés dans la Table 3.3. Ensuite, nous avons sélectionné les catégories de la CIB qui pourraient correspondre à ces domaines. Cette démarche de rapprochement manuel étant à la fois subjective et opaque, il nous faut donc la valider de façon automatique.

Calcul de la couverture Pour valider nos sous-corpus, nous avons utilisé le calcul de la couverture du corpus par les concepts de l'ontologie. Une approche similaire a été utilisée par Oostdijk *et al.* (2010) pour évaluer la couverture d'un domaine par une ontologie à partir d'un corpus de spécialité.

Nous avons calculé le pourcentage des occurrences des concepts dans le nombre total de tokens pour chaque sous-corpus. Par exemple, le sous-corpus en anglais pour la catégorie C02F compte 1 339 946 occurrences de concepts pour 7 806 687 tokens, la couverture est alors égale à 17 % (le maximum dans notre collection). Le sous-corpus le moins couvert est celui en français pour la catégorie H02M avec une couverture égale à 1 % (55 803 occurrences pour 4 359 434 tokens).

La disparité de couverture entre les sous-corpus de langues différentes (17,16 % maximum pour l'anglais, 3,67 % pour l'allemand et 3,60 % pour le français) pourrait s'expliquer par les différences en nombre de termes connus pour ces langues dans l'ontologie EcoLexicon (Table 3.1). Vu que notre corpus est parallèle, nous pouvons baser notre analyse des domaines sur la couverture des sous-corpus en anglais et nous donner l'objectif de retrouver les traductions qui manquent à partir du corpus.

Nous avons fait l'hypothèse que les sous-corpus contenant plus de concepts d'une ontologie conviennent mieux pour l'élargissement et l'enrichissement de cette dernière. Pour cette raison, nous allons commencer le travail par les sous-corpus les plus couverts.

Validation semi-automatique A cette étape, l'objectif est à la fois de voir quels concepts apparaissent dans le corpus, mais également de s'assurer que leur sens dans le corpus correspond à celui décrit dans l'ontologie.

Concept	Labels
Méthode	method@en, méthode@fr, Methode@de
Processus	process@en, processus@fr, Prozess@de
Traitement	treatment@en, traitement@fr, Verarbeitung@de, Behandlung@de
Outil	device@en, outil@fr, Mechanismus@de
Système	system@en, système@fr, System@de

TABLE 3.4 – Sélection manuelle des concepts généraux

3.2. CORPUS PATTR

Nous avons remarqué qu’une partie des occurrences appartient à des concepts généraux qui se rapprochent de la notion du lexique transdisciplinaire (Tutin, 2007a; Jacquy *et al.*, 2013), par exemple *méthode*, *outil*, *processus*, etc. Cela est certainement dû au fait que le corpus est construit à partir de brevets. Nous devons maintenant nous assurer que la couverture du corpus n’est pas seulement due à ces concepts généraux. Pour cela, nous avons sélectionné cinq concepts généraux récurrents et les termes associés à ces concepts dans EcoLexicon (Table 3.4). Cela nous a permis de calculer le pourcentage des occurrences de ces concepts généraux par rapport au nombre total des occurrences.

CIB	EN		FR		DE	
	Couverture	CG	Couverture	CG	Couverture	CG
C02F	17,16	11,86	1,63	1,19	3,36	7,29
B09C	16,17	13,40	3,60	0,99	3,67	6,66
C25C	12,54	11,91	3,33	0,88	3,12	2,66
C25D	11,66	14,88	2,63	0,98	2,48	4,41
C01G	11,57	14,72	3,10	0,46	2,57	2,91
C25B	11,18	13,43	2,93	0,79	2,25	5,31
C25F	11,04	19,00	2,18	1,09	2,55	6,75
H01M	10,32	10,73	2,69	0,55	1,86	4,17
H02J	9,57	15,49	1,57	0,94	1,68	9,14
H01G	8,15	12,12	1,79	1,13	1,94	2,70
H02M	8,08	9,54	1,28	0,45	1,39	4,03

TABLE 3.5 – Couverture du corpus et le pourcentage des concepts généraux (CG)

La combinaison de la couverture du corpus et du pourcentage des concepts généraux dans cette couverture (Table 3.5) permet de sélectionner les meilleurs sous-corpus pour la suite des traitements. Le pourcentage maximal de concepts généraux, égal à 19 %, est observé pour le sous-corpus en anglais pour la catégorie C25F. Cela signifie que chaque cinquième occurrence appartient à un concept général. Sans résultats complets de l’enrichissement et de l’élargissement de l’ontologie, il est assez difficile de juger si ce pourcentage est trop élevé. Les pourcentages maximaux pour les concepts généraux pour l’allemand et le français sont respectivement de 9,14 % et de 1,19 %.

Nous avons également sélectionné cinq segments au hasard pour dix termes (également sélectionnés de manière aléatoire) afin de s’assurer qu’ils sont employés dans leur sens terminologique. Ce survol rapide nous a aidé à confirmer que les sous-corpus conviennent à des traitements ultérieurs prévus par notre méthode.

Quant à la structure des brevets, nous avons remarqué que la partie des revendications se démarque par des expressions spécifiques, comme *method as in claim X*, *product accord to one of the claim X*, *a process along the line of claim*, etc. Toutefois, cela ne remet pas en question les sens des termes contenus dans cette partie.

3.2. CORPUS PATTR

3.2.2.1 Sous-corpus retenus pour l'évaluation

L'étape précédente a permis de choisir les sous-corpus correspondant au mieux à l'ontologie EcoLexicon. Il s'agit des catégories C02F et B09C de la CIB. Nous avons retenu les fichiers correspondant aux revendications et aux résumés des brevets, car les fichiers de descriptions n'étaient pas toujours disponibles, tandis que ceux des titres contenaient essentiellement des mots en majuscules (sans accents).

Corpus pour l'évaluation complète L'intégralité des fichiers de revendications et de résumés pour les domaines choisis, c'est-à-dire C02F (traitement des eaux usées) et B09C (traitement des sols), représente nos corpus d'évaluation du processus complet.

Langue	C02F			B09C		
	Tokens	Types	Lemmes	Tokens	Types	Lemmes
FR	18 414 204	51 890	36 919	1 864 871	15 137	9 970
EN	12 316 949	45 932	36 479	1 250 593	12 524	9 515
DE	4 789 889	150 079	134 881	480 707	28 021	23 375

TABLE 3.6 – Statistiques des sous-corpus retenus

La Table 3.6 fournit les informations statistiques sur la taille des corpus. Nous pouvons constater que le corpus C02F est environ dix fois plus important que celui B09C. Nous allons tenir compte de cette différence dans l'évaluation finale des résultats. L'observation des deux corpus permet de remarquer une régularité au niveau de nombre de tokens, types (formes fléchies uniques des mots) et lemmes²¹ : les corpus en allemand ont trois fois moins de tokens par rapport à l'anglais et quatre fois moins par rapport au français. En même temps, le nombre de types et de lemmes est beaucoup plus important, ce qui peut s'expliquer par la préférence de la langue allemande pour les mots composés.

Sous-corpus annoté en termes Afin d'évaluer l'extraction terminologique, nous avons manuellement annoté des termes mono et polylexicaux dans deux sous-corpus du corpus décrit précédemment. Le processus de cette annotation sera décrit dans la section suivante.

Le corpus est construit à partir des 100 premières lignes de chaque fichier où chaque ligne est un segment du corpus parallèle, ce qui résulte en 400 segments par langue et par corpus. Les segments dans les corpus peuvent inclure plus d'une phrase. La longueur moyenne d'un segment dépasse 200 tokens. La Table 3.7 fournit

21. Lemmatisation faite par l'outil de l'étiquetage morphosyntaxique TreeTagger (Schmid, 1994), cf Section 3.3.

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

Langue	C02F			B09C		
	Tokens	Types	Lemmes	Tokens	Types	Lemmes
FR	25 199	2 549	1 849	26 316	2 794	2 111
EN	21 791	2 442	1 922	20 181	2 386	1 900
DE	23 328	3 997	3 396	21 390	3 840	3 305

TABLE 3.7 – Statistiques des sous-corpus annotés

des informations relatives à la taille des corpus annotés pour l'évaluation de notre projet.

3.3 Prétraitement et annotation du corpus

L'étape de pré-traitement comprend la tokenisation, la lemmatisation (optionnelle) et l'annotation du corpus par les mots informatifs et vides.

3.3.1 Tokenisation et lemmatisation

Le système fait appel au script de tokenisation du projet Europarl²² et à l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1994) qui est utilisé en tant que lemmatiseur. Rappelons que la tokenisation sert à séparer les mots (tokens) dans le texte. Par exemple, la phrase *L'invention concerne un matériau filtrant microporeux.* sera tokenisée comme suit : *L' invention concerne un matériau filtrant microporeux .* (les espaces séparent *L'* et *invention*, *microporeux* et *."*). La lemmatisation, à son tour, ramène les tokens à leur forme dictionnaire : *le invention concerner un matériau filtrant microporeux .*

La lemmatisation est en général liée à l'étiquetage morphosyntaxique et nécessite des outils spécifiques disponibles pour les langues bien dotées (dont le français, l'anglais et l'allemand), mais manquants pour un grand nombre de langues moins outillées. Il est donc intéressant de pouvoir contourner cette étape dans l'objectif de rendre la méthode utilisable pour une plus grande palette de langues. Afin de comparer l'efficacité de notre méthode sur le corpus tokenisé et celui lemmatisé, la suite des traitements est effectuée sur les deux versions du corpus.

3.3.2 Annotation par mots informatifs et vides

Le corpus est traité par l'algorithme d'annotation par des mots informatifs et vides de Vergne (2003, 2004, 2005) que nous avons déjà évoqué dans la Section 1.2.3.1. Rappelons que cet algorithme combine deux propriétés, la fréquence et la

22. <http://www.statmt.org/europarl/v7/tools.tgz>

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

longueur de mots, pour détecter les mots vides à partir de corpus brut indépendamment de la langue de ce dernier. L'algorithme a été implémenté selon l'algorithme²³ fourni par Vergne (2005) :

- le type de la différence (la différence est orientée) :
 - si $f1^a > f2$ & $l1 < l2$, alors `type_diff` = nI (non informatif - Informatif)
 - sinon si $f1 < f2$ & $l1 > l2$, alors `type_diff` = In (Inform. - non informatif)
 - sinon `type_diff` = contradictoire (contradiction entre les deux critères)
 - la mesure de la différence :
 - mesure = $(\max(f1, f2) / \min(f1, f2)) * (\max(l1, l2) / \min(l1, l2))$
 - le seuil appartient à l'intervalle^b entre :
 - borne basse = 1 (égalité sur les 2 critères pour les 2 mots)
 - borne haute = $\frac{\sqrt{(f_max * l_max)}}{2}$
 - le calcul :
 - si `type_diff` $\neq$ contradictoire et mesure > seuil,
 - alors `étiq_mot2` = étiquette 2 du `type_diff`
 - si `étiq_mot1` indéterminée,
 - alors `étiq_mot1` = étiquette 1 du `type_diff`
 - sinon `étiq_mot2` = `étiq_mot1`
- a.* $f1$ et $f2$ sont des fréquences des mots ; $l1$ et $l2$ sont leurs longueurs.
b. Selon Vergne (2005), la valeur optimale du seuil est proche de 0.4, la valeur que nous retenons dans nos expériences.

Cependant, l'application de la méthode, telle qu'elle est présentée par Vergne, rapproche les mots informatifs des termes candidats. Or, les exemples de résultats obtenus pour cette méthode sur des corpus constitués d'articles de presse Vergne (2005) contiennent peu de termes candidats tels que nous pouvons exploiter dans un système de gestion de la terminologie. Par exemple, dans *une nouvelle résolution de l'ONU*, les mots informatifs sont : *nouvelle*, *résolution*, *ONU*. Les mots *ONU* et *résolution* peuvent effectivement être des termes candidats et même former un terme candidat poly lexical *résolution de l'ONU*, mais le mot *nouvelle* ne s'inscrit pas dans cette terminologie. Dans d'autres exemples, les mots informatifs incluent également des mots de la langue générale, comme *cherche*, *utiliser*, *tonnes*, *L'or*, etc.

Nous avons remarqué que cette annotation n'est pas homogène²⁴ : un même mot peut être annoté comme informatif s'il est entouré par des mots plus courts ou comme non informatif s'il est à côté d'un mot plus long et encore moins fréquent. Par exemple, dans [...] *représente*_(*n*)²⁵ *un*_(*n*) *halogène*_(*I*) ,_(*n*) *avec*_(*I*) *la*_(*n*)

23. Implémenté par nos soins en Java 1.7 et en Python 2.7

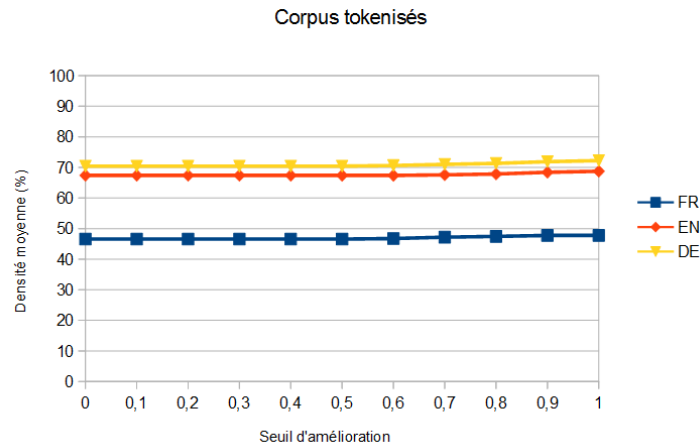
24. Il est difficile d'évaluer la performance de cette annotation, car la frontière entre les mots vides et informatifs n'est pas clairement définie. Les résultats de l'évaluation faite par Vergne (2003) ne sont pas comparables à l'usage que l'on fait de sa méthode. En effet, ils contiennent des candidats pour plusieurs langues mélangées et ne permettent pas de déduire les critères de validation.

25. Où *I* signifie un mot informatif et *n* – un mot vide.

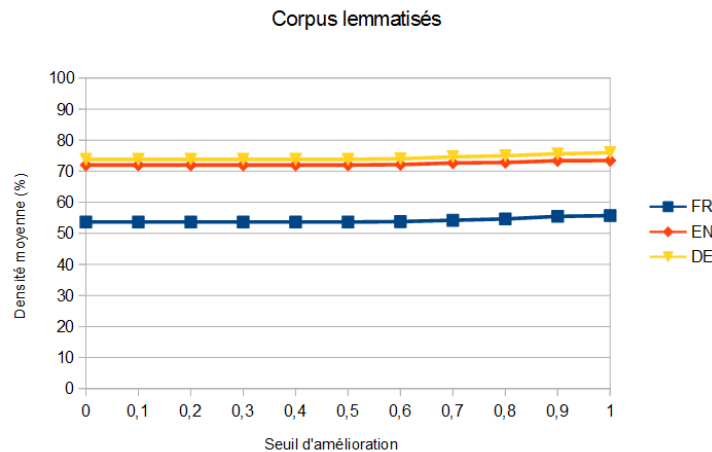
3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

*condition_(I) supplémentaire_(I) que_(n) a_(n))_(n) **si_(I)** l_(I) signifie_(I) un_(n) atome_(I) de_(n) chlore_(I) [...]*, le mot *si* est étiqueté comme informatif, tandis que dans *[...] procédé_(n) pour_(n) déterminer_(I) **si_(n)** un_(n) état_(I) connecté_(I) d'_(n) une_(n) pluralité_(I) de_(n) fils_(I) connectés_(I)[...]*, il est annoté comme vide. Cela concerne notamment les mots courts, les mots plus longs étant généralement étiquetés de manière stable.

Pour cette raison, nous avons appliqué un seuil pour corriger l'annotation. Le mot est validé comme informatif si le pourcentage d'occurrences annotées comme telles dépasse ce seuil. Nous avons testé les valeurs du seuil entre 0,1 et 1. Par exemple, si un mot a été annoté comme informatif dans 8 cas sur 10, il sera informatif pour les seuils inférieurs ou égaux à 0,8 et deviendra vide pour les seuils 0,9 et 1.

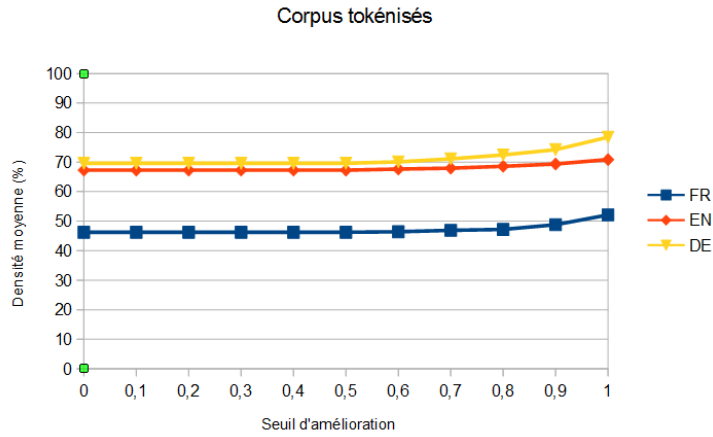


(a) Corpus tokenisés partiels

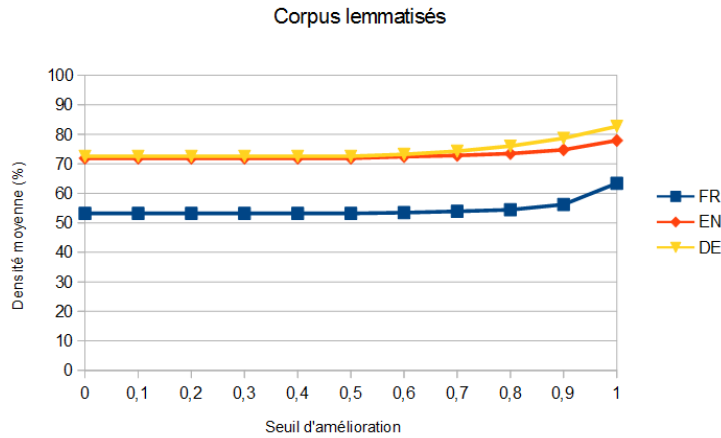


(b) Corpus lemmatisés partiels

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS



(c) Corpus tokenisés complets



(d) Corpus lemmatisés complets

FIGURE 3.3 – Densité moyenne de termes dans les mots informatifs dans les corpus complets

La densité de termes dans les mots informatifs a été calculée en divisant le nombre de termes dans les mots informatifs du corpus et multipliée par 100 pour obtenir le pourcentage. Nous avons remarqué que l'annotation par des mots informatifs et vides fournit des meilleurs résultats sur le corpus lemmatisé, avec un écart d'environ 5 % par rapport à celui tokenisé. Nous avons également testé la performance de cette annotation sur les corpus de petite et de grande taille : les Figures 3.3a et 3.3b illustrent la densité de termes dans les mots informatifs des corpus pré-annotés décrits dans la Section 3.2.2.1. Rappelons que ces corpus contiennent entre 20 000 et 26 000 tokens par langue. Les Figures 3.3c et 3.3d montrent l'augmentation de la densité des termes dans les mots informatifs lorsque le corpus est annoté dans sa totalité (cf Section 3.6) et évalué à la base du même-sous corpus. La taille des sous-corpus complets varie entre 1,2 million et plus de 18 millions de tokens.

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

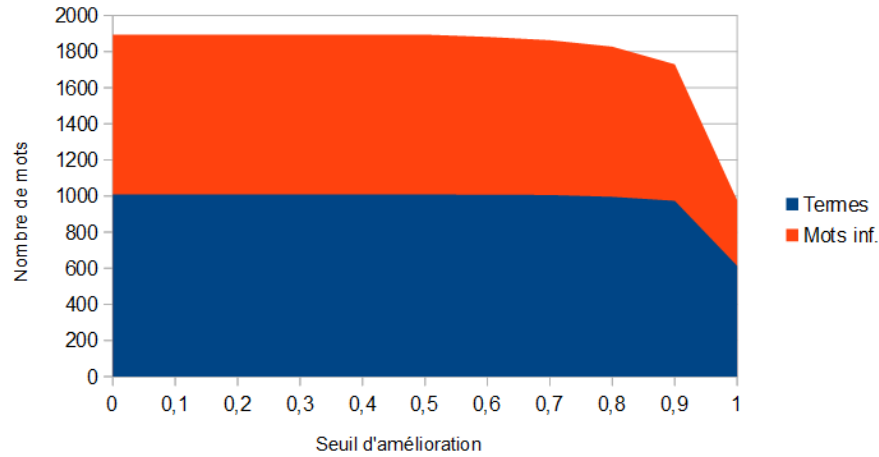


FIGURE 3.4 – Réduction du nombre moyen des mots informatifs et des termes en fonction du seuil (FR)

L'augmentation de la densité est due à la réduction du nombre de mots annotés comme informatifs. Comme les termes en font partie, certains sont également filtrés (cf Figure 3.4). Cela signifie que le seuil plus élevé permettrait d'améliorer la précision de l'extraction terminologique, tout en diminuant le rappel. On peut remarquer que l'application du seuil influence les résultats à partir de la valeur 0,6. C'est à partir de cette valeur que le filtre sélectionne l'étiquette de la majorité des occurrences du mot dans le corpus.

Corpus	Sans annotation	Baseline (Vergne, 2005)
Tokenisé	de	selon
	,	procédé
	)	revendication
	(	invention
	la	qui
Lemmatisé	de	selon
	le	revendication
	@card@	lequel
	,	signal
	un	premier

TABLE 3.8 – Mots informatifs triés par fréquence absolue (Corpus C02F, FR)

Sans annotation par mots informatifs et vides, les éléments les plus fréquents de corpus en français sont les signes de ponctuation, des articles et des prépositions (Table 3.8). L'annotation *baseline*²⁶ obtenue en suivant l'algorithme de Vergne

26. Une baseline (base de référence) représente des résultats d'une configuration de départ ou d'un

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

(2005) décrit au début de cette section, permet d'enlever la plupart des prépositions, articles et signes de ponctuation, mais les mots informatifs contiennent encore des mots comme *selon*, *qui*, *avoir* qui ne peuvent pas être des termes.

Corpus	Seuil				
	0,6	0,7	0,8	0,9	1
Tokenisé	procédé	revendication	revendication	revendication	instruction
	revendication	invention	invention	invention	diverse
	invention	revendications	revendications	revendications	instructions
	revendications	signal	signal	signal	excédant
	signal	eaux	dispositif	dispositif	constitutifs
Lemmatisé	selon	signal	signal	signal	instruction
	revendication	premier	premier	invention	manière
	signal	comprendre	procédé	élément	composante
	premier	moyen	invention	caractériser	quatrième
	comprendre	procédé	élément	dispositif	excéder

TABLE 3.9 – Influence du seuil pour les mots informatifs sur le tri par fréquence absolue (Corpus C02F, FR)

La Table 3.9 illustre l'amélioration apportée aux résultats en tête de la liste triée dans l'ordre décroissant de fréquence absolue. L'application du seuil égal à 0,6 améliore les résultats de la baseline en enlevant ces mots apparemment vides. L'observation des résultats, dont seulement une partie est présentée dans la Table 3.9 montre que la qualité de l'annotation s'améliore jusqu'au seuil 0,9 (par exemple, le mot *premier* est encore présent dans les résultats obtenus avec le seuil égal à 0,8). Ensuite, le seuil égal à 1 devient trop contraignant et discrimine des candidats valides, comme *signal*, *dispositif* ou encore *traitement*. Nous allons donc utiliser le seuil égal à 0,9. Toutefois, le seuil reste paramétrable afin de laisser la main à l'utilisateur.

Toutefois, cette annotation, ainsi que son amélioration via le seuil, discrimine certains termes, comme c'est le cas du terme *eau* dans le corpus C02F portant sur les eaux usées. En effet, ce terme est court et très fréquent, ce qui contribue à son annotation en tant qu'un mot vide même dans la baseline. Ce problème pourrait être résolu via une rapide correction manuelle par un expert, toutefois nous ne l'avons pas incluse dans notre chaîne de traitements pour éviter une complexification de la méthode. L'application d'un seuil neutralise également le caractère local du calcul qui permet de tenir compte des homographes (Vergne, 2004). Toutefois, le seuil permet d'harmoniser l'étiquetage, ce qui contribue à la qualité l'extraction terminologie.

Les corpus pré-traités et annotés par les mots informatifs et les mots vides serviront de base pour l'extraction de termes candidats monolexicaux. Nous allons

système de référence par rapport auxquels sont évalués les résultats obtenus par nos traitements.

illustrer les résultats de l'extraction terminologique sur les corpus lemmatisés et non-lemmatisés afin d'évaluer l'impact de la lemmatisation.

3.3.3 Annotation manuelle du corpus pour évaluation

Nous avons pré-annoté les termes mono et polylexicaux dans les sous-corpus décrits dans la section précédente. Pour cela, nous avons lancé l'extracteur TTC TermSuite 2.2 (que nous allons utiliser comme baseline) et notre extracteur, et avons manuellement validé les termes candidats fournis par les deux outils. Dans le processus de la validation des termes, nous avons suivi les instructions du guide du projet TTC énumérées dans la Section 1.1.2.3. Ce guide est globalement en accord avec la méthodologie de Bernier-Colborne (2012). Rappelons les points essentiels de ce guide (Syllabs, 2012) :

- un candidat monolexical est valide s'il est dérivé d'un terme, s'il est présent dans la définition d'un terme, si son contexte diffère de celui dans la langue générale ou encore s'il s'agit d'un emprunt ou d'un internationalisme ;
- un candidat polylexical peut être inclus dans un terme plus long, dans ce cas, pour valider le candidat plus court, il faut vérifier si leurs nombres d'occurrences sont différents.

À l'instar de ce guide, nous avons organisé les termes attendus de façon à les regrouper avec leurs variantes et leurs usages collocationnels. Il faut quand même noter que les listes du projet TTC²⁷ sont fondées sur les patrons nominaux et adjectivaux. Cela nous semble trop limitatif, surtout que les variantes et les collocations offrent une possibilité d'inclure les verbes et les patrons verbaux dans les résultats, comme le suggèrent (Castro *et al.*, 2014; Bernier-Colborne et L'Homme, 2015; L'Homme, 2016).

3.3.3.1 Annotation des termes monolexicaux

Dans l'annotation des termes monolexicaux, nous avons traité des noms, des adjectifs et des adverbes, ainsi que des verbes auxquels nous avons prêté une attention particulière.

Annotation des noms et des adjectifs Parmi les termes monolexicaux annotés, on retrouve les noms désignant des substances chimiques (*photocatalyseur*, *polyalkylènepolyamines*, *c2h5ch*), des objets naturels (*sédiment*, *roche*), des processus (*acidification*, *rinçage*) ; les adjectifs caractérisant des propriétés (*microbiotique*, *bactériophage*, *rectangulaire*, *etc.*) et certains adverbes dérivés des adjectifs spécialisés (*électrostatiquement*).

27. <http://www.lina.univ-nantes.fr/?listes-terminologiques-de.html>

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

Parmi les difficultés de l'annotation, figurent les éléments du lexique transdisciplinaire (Jacquey *et al.*, 2013; Tutin, 2007a; Hatier, 2013) discutés dans la Section 1.2.3.2 : *méthode, procédé, outil, traitement, dispositif*, etc. En effet, ces mots ne sont pas propres à un domaine précis, mais sont largement utilisés dans les écrits techniques. Nous avons décidé de les garder dans la liste de termes, surtout qu'ils sont déjà présents dans les ressources terminologiques à notre disposition et qu'ils forment des termes polylexicaux valides : *dispositif de désionisation capacitif, moyen de gazéification, bain de traitement aqueux, procédé de traitement anaérobie*, etc.

Contrairement au lexique transdisciplinaire, les mots propres aux brevets, comme *revendication, invention, concerner*, n'ont pas été retenus comme termes, car ils ne forment pas de termes polylexicaux valides.

Annotation des verbes Dans l'annotation des verbes, nous avons suivi le principe préconisé par L'Homme (2004) : le verbe est un terme si son sens dans le domaine de spécialité diffère de celui usuel. Par exemple, le verbe *stabiliser* peut être employé dans plusieurs contextes, comme *stabiliser la situation politique*, mais son sens dans le domaine de spécialité est différent : « rendre stable un produit tant au point de vue physico-chimique que microbiologique²⁸ ». De plus, certains verbes ont un sens spécialisé et sont peu employés dans les textes généraux, par exemple *ioniser, chlorer, oxygéner, dépolluer*, etc. Nous avons également annoté comme termes les verbes désignant des actions propres à un domaine de spécialité, mais employés dans la langue générale, comme *filtrer, distiller, perforer, fondre*, etc.

Au total, nous avons retenu comme termes 123 verbes (en français), ce qui représente en peu moins de 8 % de termes monolexicaux que nous avons annotés. Parmi les verbes qui n'ont pas été retenus, on peut citer *générer, incorporer, substituer, concentrer, transférer*, etc., car leur sens reste globalement stable dans tous les contextes.

3.3.3.2 Annotation des termes polylexicaux

La majorité des termes polylexicaux²⁹ en français correspond à des syntagmes nominaux : *système de assainissement in-situ, solution de aluminium acide*³⁰, etc.

28. La définition du Grand Dictionnaire Terminologique pour le domaine laitier http://www.granddictionnaire.com/fiche0qlf.aspx?Id_Fiche=17599212.

29. Nous avons défini les limites de longueur pour les termes polylexicaux entre deux et quatre mots.

30. Les exemples de termes issus de nos résultats sont présentés tels qu'ils sont extraits. Notre système ne prévoit pas de la mise en forme complémentaire : si un terme a été identifié dans sa forme lemmatisée, il est présenté ainsi à l'utilisateur. Pour les termes allemands, la majuscule au début des noms est perdue au cours des traitements. Pour les termes français, l'apostrophe devant les mots commençant par une voyelle ou par un *h* muet est également enlevé au cours de la lemmatisation. Nous avons choisi de garder les exemples bruts afin d'éviter toute confusion sur les résultats de notre système. Il est possible d'ajouter des étapes de traitement dédiées, mais cela

3.3. PRÉTRAITEMENT ET ANNOTATION DU CORPUS

Le guide du projet TTC Syllabs (2012) indique qu'il est préférable de garder les termes polylexicaux les plus courts, tandis que la méthodologie de Bernier-Colborne (2012) donne la préférence aux termes les plus longs. Nous considérons que les deux approches sont valides et que c'est au terminologue de définir quel terme sera le terme de base, les autres étant ses variantes. Plus concrètement, dans le cas du terme *système de assainissement in-situ*, nous allons également annoter *système de assainissement* et *assainissement in-situ*. Toutefois, pour les termes commençant par le même mot et ayant le même nombre d'occurrences, comme *compartiment perméable* et *compartiment perméable au liquide*, où un terme plus court est inclus dans un terme long, on conserve le terme le plus long.

Les termes polylexicaux contenant un verbe peuvent être classés dans deux catégories : soit le verbe est à la tête de l'expression et désigne une action (par exemple, *extraire le radium*, *aspirer du gaz*, *désintégrer le élément organique*), soit il se trouve à une autre position (*matrice à dépolluer*, *flocon coagulé*, etc.) et sa catégorie sémantique est alors différente (usage, résultat, propriété, etc.). Notons quand même qu'en français dans la majorité des cas, le participe passé du verbe se trouve derrière un nom et désigne une propriété de la notion exprimée par ce dernier.

Nous considérons qu'un système d'extraction terminologique rencontre des difficultés dans la distinction entre un terme polylexical et ses variantes. Par exemple, dans la liste de termes du projet TTC, on retrouve le terme *stockage de énergie* accompagné de deux variantes : *stockage de energie* (variante graphique) et *stockage inertiel de énergie* (variante syntaxique). Le seul indice permettant au système informatique de préférer le terme de base est sa fréquence dans le corpus (égale à 116), supérieure à celles de ses variantes (7 et 1). Dans l'annotation du corpus, nous allons donc considérer les termes et les variantes comme termes valides pour l'évaluation de l'extraction terminologique.

L'allemand représente un cas à part : les termes monolexicaux contiennent des noms composés correspondant aux termes polylexicaux en français et en anglais. Étant donné que les termes polylexicaux annotés pour ces deux langues contiennent entre deux et quatre mots, il est peu utile d'annoter les termes polylexicaux en allemand. En effet, si un terme polylexical allemand contient un terme monolexical composé de deux ou trois termes, nous ne serons pas en mesure de trouver la traduction du terme polylexical allemand en français ni en anglais du fait qu'il correspondra à 5 ou 6 mots cibles. De plus, le nombre de termes polylexicaux à valider serait trop important pour être fait dans le cadre du projet de thèse, car la liste de termes monolexicaux annotés contient 4 097 termes (Table 3.10).

La Table 3.10 récapitule les résultats de l'annotation manuelle des termes. La différence en nombre de termes entre les langues s'explique avant tout par la présence des termes monolexicaux composés qui sont traduits par des termes polylexicaux ou monolexicaux simple en d'autres langues. Par exemple, le terme *Ammoniakkon-*

n'entre pas dans les limites de notre projet de thèse.

3.4. CONCLUSION

Langue	Termes monolexicaux	Termes polylexicaux	Total
FR	1 835	2 442	4 277
DE	4 213	-	4 213
EN	2 094	2 517	4 611

TABLE 3.10 – Termes annotés (corpus lemmatisés)

zentrations (DE) est traduit par *concentration d'ammoniaque* (FR). La deuxième raison est la différence dans la lemmatisation fournie par TreeTagger pour des différentes langues : en allemand, les participes des verbes sont repris en tant que lemmes (par exemple, *destilliert* au lieu de *destillieren*), tandis que pour le français et pour l'anglais ils sont lemmatisés par l'infinitif du verbe correspondant (*distillé(e)* est lemmatisé par *distiller*).

3.4 Conclusion

Dans ce chapitre, nous avons présenté les ressources utilisées dans notre projet. Notamment, nous avons décrit les ressources sémantiques, à savoir l'ontologie Eco-Lexicon et les thésaurus Agrovoc et GEMET. Ces ressources permettent d'évaluer les différentes étapes de notre système. Nous avons également présenté le corpus PatTR et ses sous-corpus sélectionnés pour l'évaluation.

Les sous-corpus ont été pré-traité par des outils de lemmatisation et par l'étiquetage en mots informatifs et vides de Vergne (2005). De plus, nous avons annoté des parties des sous-corpus en termes mono et polylexicaux qui seront utilisés pour évaluer les performances de notre système dans les chapitres suivants.

Chapitre 4

Extraction terminologique multilingue à partir des ressources endogènes

L'extraction terminologique constitue la première étape dans l'élargissement et l'enrichissement des ontologies spécialisées. L'objectif de cette étape est d'identifier les nouveaux termes à partir de corpus du domaine. L'aspect innovant de notre méthode est l'utilisation de ressources endogènes, c'est-à-dire générées directement au cours de l'analyse à partir du corpus et de l'ontologie, qui permettent de minimiser l'apport extérieur en termes de ressources nécessaires. Plus précisément, notre méthode demande très peu de connaissances et de ressources lexicales extérieures.

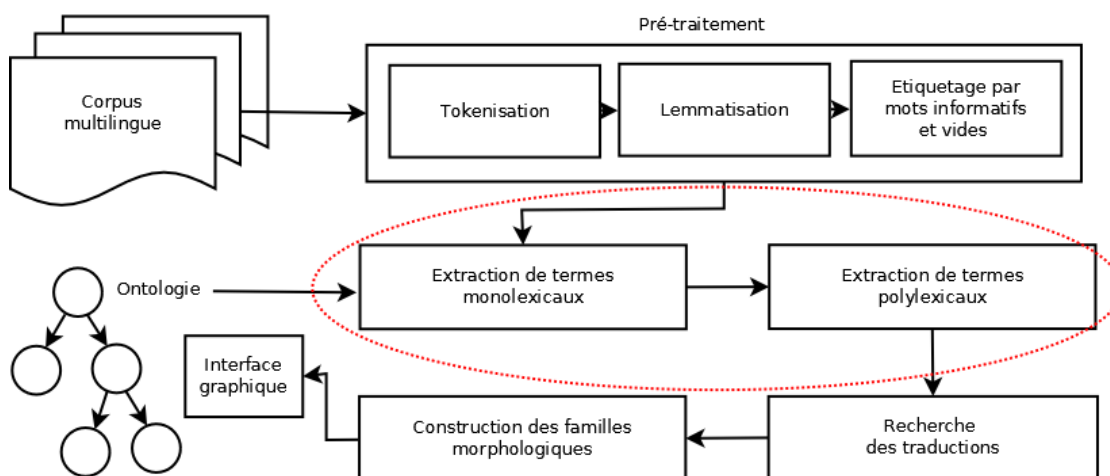


FIGURE 4.1 – Extraction terminologique dans notre méthode

La Figure 4.1 contextualise l'extraction terminologique dans notre système : le corpus pré-traité (tokenisé, étiqueté et éventuellement lemmatisé) et l'ontologie à

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

enrichir sont les seules ressources utilisées à cette étape. Notre approche est conçue pour être applicable à plusieurs langues et indépendante du domaine de spécialité. De plus, elle combine des stratégies pour extraire des termes candidats monolexicaux, constitués d'un seul mot, et polylexicaux qui en comptent plus d'un.

En effet, les méthodes de l'extraction terminologique qui se montrent efficaces, nécessitent des ressources externes assez importantes, comme par exemple des listes de formants néo-classiques, des patrons morphosyntaxiques, des dictionnaires de référence, etc. Les systèmes basés sur de telles méthodes sont dépendants vis-à-vis de la langue et de ces ressources. Or, ces ressources ne font que refléter des régularités linguistiques. De ce point de vue, il doit être possible de générer ces ressources à partir du corpus et de l'ontologie source.

Cet objectif a été atteint par le biais de patrons morphologiques et morphosyntaxiques endogènes qui résultent d'une approche multi-étapes à base de n-grammes de caractères. Un n-gramme est une séquence de n éléments d'une chaîne ; dans notre cas, il s'agit de séquences de caractères extraites à partir des termes candidats du domaine. Les patrons sont appris à partir de corpus tokenisé (ou lemmatisé pour atteindre une meilleure précision) et étiqueté selon l'algorithme de Vergne (2005).

4.1 Méthode d'extraction de termes monolexicaux à l'aide de ressources morphologiques endogènes

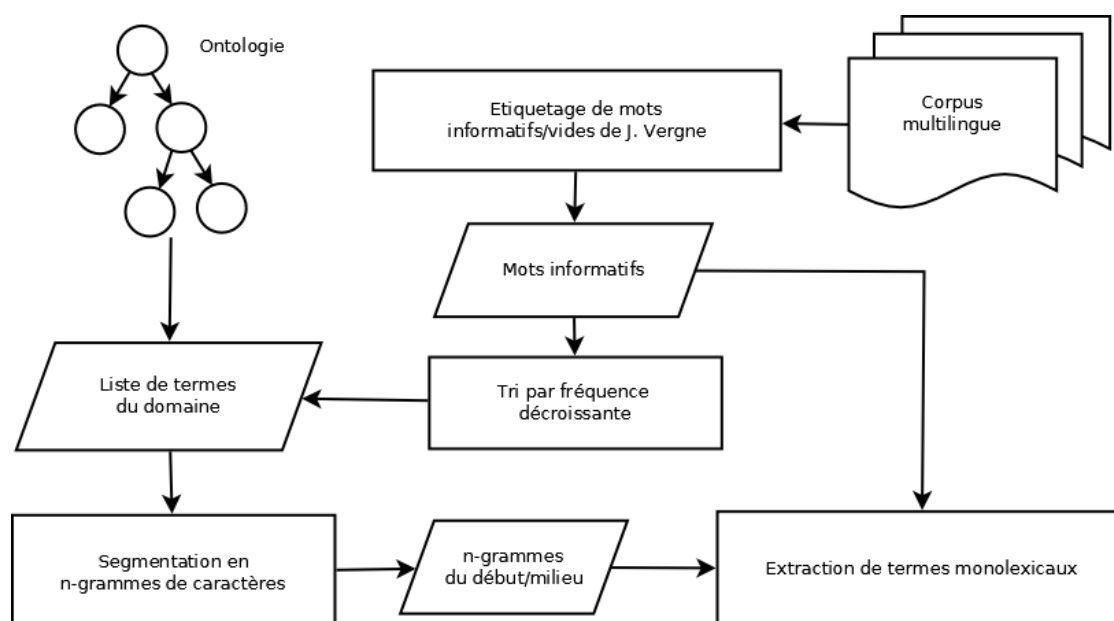


FIGURE 4.2 – Extraction de termes monolexicaux

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

Le premier objectif est d'identifier les termes monolexicaux dans le corpus (Figure 4.2). Pour cela, le corpus est pré-traité comme décrit dans la Section 3.3, c'est-à-dire, tokenisé, lemmatisé¹ et annoté par les étiquettes de mots informatifs et vides selon la méthode de Vergne. Nous considérons que tous les termes sont des mots informatifs, mais pas tous les mots informatifs sont des termes. Pour cette raison, nous constituons la liste de mots informatifs du corpus utilisée pour chercher des termes candidats. L'extraction des termes monolexicaux nécessite également les libellés d'une ontologie. Ces libellés constituent la liste de termes qui servent à construire la ressource morphologique endogène composée des n-grammes de caractères du début et du milieu des mots. Ces n-grammes permettent de sélectionner des termes candidats parmi les mots informatifs.

4.1.1 Construction de la liste initiale de termes

Notre approche nécessite une liste de termes du domaine de départ pour générer des ressources morphologiques endogènes. En fonction des contraintes du projet, il est possible d'utiliser deux sources de termes :

- les termes présents dans l'ontologie source ;
- les termes candidats extraits à partir du corpus à l'aide de méthodes statistiques.

4.1.1.1 Utilisation des libellés présents dans l'ontologie

Dans le cas où l'ontologie à enrichir possède déjà des termes (libellés) en langue du traitement, ces termes monolexicaux sont utilisés pour la génération du modèle des n-grammes de caractères qui servira à l'identification des nouveaux termes candidats. Cette option a l'avantage d'assurer la qualité des termes initiaux. Rappelons que l'objectif du projet est l'enrichissement et l'élargissement de l'ontologie, ce qui présuppose l'existence d'un certain nombre de concepts et de termes du départ. Cependant, étant dans le contexte multilingue, nous devons faire face à des cas où l'ontologie est développée pour une seule langue ou elle contient plus de données pour une langue que pour d'autres. C'est notamment le cas d'EcoLexicon qui contient 4 180 termes anglais contre seulement 622 termes français.

L'utilisation des termes de l'ontologie source reste dans le cadre des ressources endogènes dans la mesure où elle ne nécessite pas de ressources autres que celles présentes dans le projet. Si le nombre de termes s'avère insuffisant, il est possible de compléter la liste par les candidats identifiés par des méthodes statistiques.

1. La lemmatisation est optionnelle, mais elle permet d'obtenir une meilleure précision.

4.1.1.2 Extraction des termes monolexicaux par fréquence et TFxIDF

Si toutefois l'ontologie ne contient pas ou très peu de termes, on peut recourir aux mots informatifs du corpus triés par fréquence décroissant. Comme l'illustre la Table 3.9, les mots informatifs, surtout après l'application du seuil d'amélioration, contiennent des termes du domaine. Comme nous l'avons discuté dans la Section 3.3.2, ce seuil est fixé à 0,9. En effet, si le seuil est plus bas, les mots informatifs contiennent plus de bruit, tandis qu'avec le seuil égal à 1, des bons termes candidats sont discriminés.

Les résultats obtenus par différentes mesures statistiques ne sont pas identiques, surtout ceux de tête de la liste triée. Pour cette raison, nous avons basé l'extraction de termes candidats monolexicaux sur l'intersection de la fréquence absolue et de la mesure TFxIDF par palier de valeurs. Le calcul tient compte uniquement des mots informatifs.

Fréquence absolue

$$tf_{m_i} = \sum_{j=1}^n tf_{m_i j} : d_j \in D \quad (4.1)$$

où $tf_{m_i j}$ est la fréquence d'un mot informatif m_i dans d_j , un élément de l'ensemble de documents D et n est le nombre de documents.

La fréquence absolue (Formule 4.1) est le nombre d'occurrence d'un mot (lemme ou token) dans le corpus.

TFxIDF

$$TFxIDF(m_i) = \frac{tf(m_i)}{tf(m_{max})} \times \log \frac{D}{d} \quad (4.2)$$

où m_{max} est le mot informatif le plus fréquent, d est le nombre de documents contenant le mot informatif m_i et D est l'ensemble des documents.

La Formule 4.2 appartient à la famille de TFxIDF, mais en représente une version adaptée pour être appliquée sur une autre unité textuelle (le corpus). Elle retrouve des termes fréquents qui apparaissent dans un nombre réduit de documents. L'idée d'utiliser le coefficient TFxIDF pour l'extraction des termes n'est pas neuve : Witschel (2005) suggère d'identifier pour chaque document les termes fréquents qui apparaissent dans peu de documents du corpus. Nous avons modifié cette méthode en calculant la fréquence pondérée d'un mot lemmatisé par rapport à la fréquence maximale d'un mot informatif dans le corpus entier. Cette modification s'explique par la volonté d'éviter un bruit possible si un document du corpus ne porte pas vraiment sur le domaine donné, risquant ainsi de polluer les résultats de TFxIDF calculés pour chaque document.

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

Paramètre	Top-10
MaxFreq	revendication, eau , caractériser, dispositif , procédé , sol , matière , groupe , invention, contaminer
TFxIDF	revendication, formule , image, alkyl , trou , concerner, écran , essieu , sol , fructosane

TABLE 4.1 – Top-10 résultats pour la fréquence absolue maximale et TFxIDF obtenus sur le corpus lemmatisé B09C (FR)

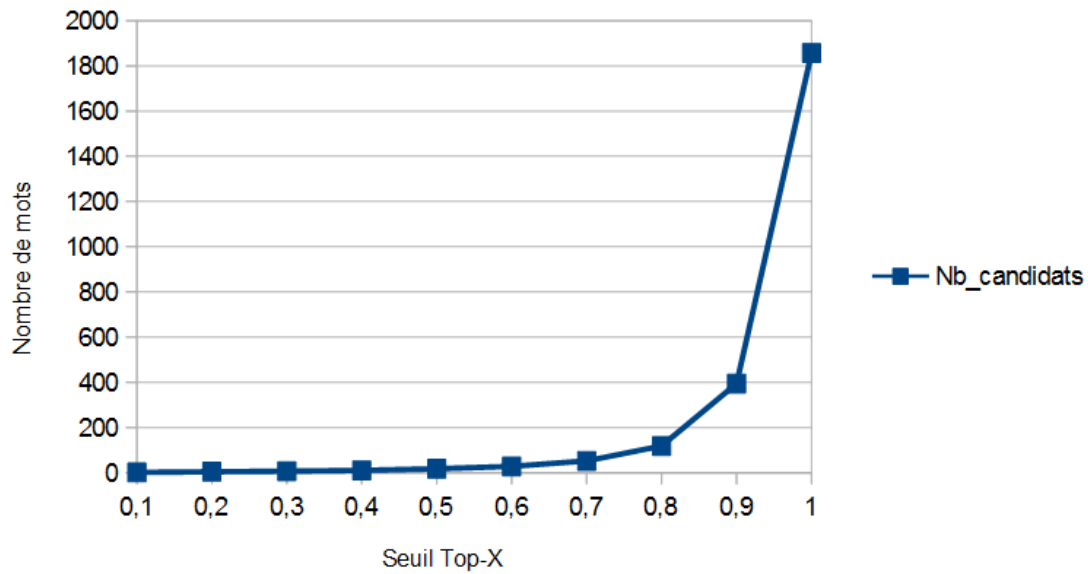
Comparaison des deux méthodes La table 4.1 illustre la différence d'approche : la fréquence absolue maximale met en valeur les mots les plus fréquents dans les textes du corpus, tandis que TFxIDF identifie les mots fréquents, mais présents dans un nombre restreint de documents. Nous avons constaté que les mots qui sont en tête de la liste TFxIDF se trouvent au milieu de la liste triée par fréquence décroissante.

Les deux listes sont triées par ordre décroissant du score (fréquence ou TFxIDF) et rangées par déciles du nombre de valeurs que prend le score en question. L'objectif est alors de sélectionner les termes candidats monolexicaux qui serviront à la génération des n-grammes pour l'identification d'autres termes candidats. Pour cela, nous gardons les termes candidats qui se trouvent à la tête de ces deux listes. Nous prenons en compte les premiers déciles de ces listes que nous appelons les Top-X % de candidats. Il est important de souligner que la sélection du Top-X % de candidats en fonction du nombre de valeurs que prend le score, et non pas en fonction du nombre de candidats, fournit des résultats stables : les valeurs de scores ont des listes de candidats correspondants fixes, tandis que l'ordre interne de ces candidats par valeur de score peut varier.

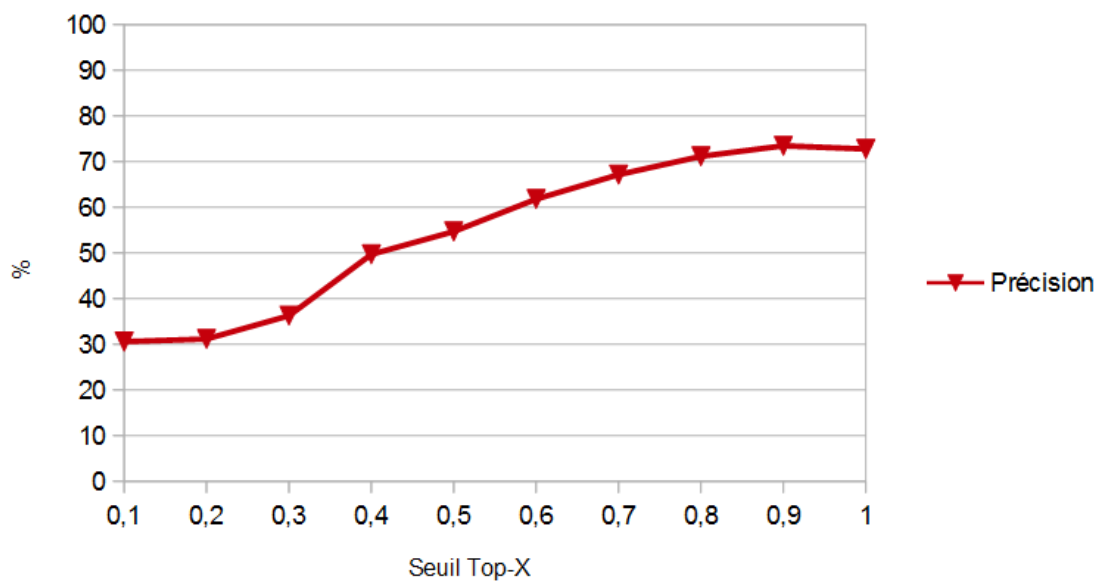
Par exemple, si la liste compte 100 candidats ayant 20 valeurs de fréquence différentes, ces données sont réparties en déciles à raison de deux valeurs de fréquence par tranche. Ainsi, le premier décile contient les deux valeurs de fréquence maximales et le dernier – les deux valeurs minimales (probablement, 2 et 1). En même temps, les premiers déciles contiennent peu de candidats, car le nombre de mots fréquents est moins important que celui des mots rares.

Les scores de la tête de liste contiennent peu de termes candidats et surtout peu de termes en commun entre les deux méthodes. Il est donc nécessaire de définir le seuil Top-X % qui réunira l'intersection de candidats entre les deux méthodes pour les valeurs de tous les paliers jusqu'au seuil défini.

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES



(a) Nombre de candidats moyen²



(b) Précision moyenne

FIGURE 4.3 – Évolution de l'intersection de la fréquence absolue et TFxIDF par valeur de seuil Top-X % scores

La Figure 4.3 permet de visualiser les moyennes du nombre moyen de candidats et de la précision par seuil Top-X % scores : par exemple, la liste de candidats

2. Ici et après, les pourcentages sont normalisés entre 0 et 1 qui correspond à 100 %.

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

obtenue avec le seuil égal à 0,8 contient environ 100 termes, dont environ 70 sont des candidats valides, tandis que le seuil égal à 0,9 regroupe environ 400 candidats avec une précision légèrement plus élevée. Nous allons évaluer les valeurs du seuil en question dans la section 4.3.4.2.

Une fois la liste de termes candidats constituée, nous pouvons en apprendre les n-grammes de caractères qui servent à extraire d'autres termes candidats à partir du corpus.

4.1.2 Construction des ressources morphologiques endogènes

Notre méthode prévoit la génération de ressources morphologiques basées sur les n-grammes de caractères extraits à partir de termes candidats monolexicaux ou à partir de termes présents dans l'ontologie de base. Ces n-grammes sont ensuite exploités pour identification d'autres termes candidats à partir des mots informatifs du corpus, en remplaçant ainsi des listes de formants classiques et des unités lexicales au sein des mots composés. Par exemple, si le paramètre n est égal à quatre, à partir du terme candidat *nanostucture*, nous allons alors apprendre les quadri-grammes suivants : *nano*, *anos*, *nost*, *ostr*, *stru*, *truc*, *ruct*, *uctu*, *ctur* et *ture*.

En effet, les n-grammes correspondent à des quasi-morphèmes selon leurs positions, ce qui a notamment été exploité pour la tâche de tokenisation (McNamee *et al.*, 2009; McNamee, 2008). Dans ces travaux conduits sur les langues indiennes (hindi, bengali et marathi), mais aussi sur l'anglais, la longueur des n-grammes est fixée à 4 et à 5. Les auteurs utilisent également des skip-grammes qui sont des n-grammes où un ou plusieurs caractères sont remplacés par un signe acceptant tout caractère : les formes conjuguées du verbe espagnol *contar* (compter) *cuento*, *cuentas*, *cuenta*, *contamos* peuvent être regroupées sous le skip-gramme c^*nt (McNamee, 2008).

De plus, Bojanowski *et al.* (2017) ont évalué les performances des n-grammes et skip-grammes de caractères pour la classification de documents dans le cadre de FastText³ en cinq langues. FastText vise une classification rapide de documents rendue possible grâce à une représentation hiérarchique des classes. Ainsi, les documents sont d'abord repartis dans des classes générales pour être ensuite triés dans des classes plus spécifiques à l'intérieur de chaque classe générale. FastText utilise et améliore la méthode word2vec de Mikolov *et al.* (2013), qui représente les mots des documents dans un espace vectoriel où les mots sémantiquement proches (par exemple, des capitales : Paris, Ankara, Moscou) ou ayant un lien sémantique (Paris et la France, Ankara et la Turquie) se regroupent. Pour représenter les mots en tant que vecteurs, word2vec utilise des skip-grammes de mots, c'est-à-dire les mots qui précèdent et qui suivent le mot en question. Dans FastText, Bojanowski *et al.* (2017) ont inclus l'information au niveau sous-lexical en utilisant des n-grammes de

3. <https://research.facebook.com/blog/fasttext/> et <https://research.fb.com/fasttext/>, développé par Facebook

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

caractères, ce qui permet de traiter des langues morphologiquement riches (comme le tchèque ou le finnois). Pour illustrer cette amélioration, Bojanowski *et al.* (2017) ont comparé les mots les plus proches identifiés par leur approche basée sur les n-grammes de caractères et par les skip-grammes de mots : pour le mot anglais *english-born* (un natif anglais), les skip-grammes suggèrent les mots *most-capped* (probablement, il s'agit des joueurs les plus sélectionnés dans des tournois sportifs) et *ex-scotland* (dont le sens est ambigu sans contexte complémentaire), tandis que les n-grammes de caractères permettent de le rapprocher avec *british-born* (un natif britannique) et *polish-born* (un natif polonais). Dans FastText, un vecteur n'est alors plus associé à chaque mot, mais à chaque n-gramme de caractères, et les mots sont considérés comme des sommes de ces n-grammes (Bojanowski *et al.*, 2017).

Les travaux de (Sennrich *et al.*, 2016; Luong et Manning, 2016) démontrent l'efficacité des n-grammes pour la traduction automatique, notamment dans le cadre de la traduction automatique par des réseaux de neurones. En même temps, (Ling *et al.*, 2015) ont employé les n-grammes pour améliorer l'étiquetage morphosyntaxique.

A la différence de ces méthodes, dans notre approche, les mots sont segmentés en n-grammes de caractères au paramètre n fixe et les trois positions de n-grammes sont prises en compte : au début du mot, au milieu et à la fin.

Les listes obtenues confirment l'hypothèse que ces n-grammes se rapprochent des morphèmes de la langue. Les résultats de cette comparaison permettent de faire quelques observations sur le comportement des n-grammes selon leur position.

Position n	FR			EN			DE		
	2	3	4	2	3	4	2	3	4
Début (marque _)	_co _pr _pa _po _re	_con _pro _com _pou _car	_cont _comp _part _cara _méta	_th _co _an _pr _re	_con _pro _met _com _wat	_cont _wate _proc _comp _char	_di _ei _un _be _vo	_ein _ver _wer _sch _nac	_durc _verf _kenn _wass _bode
Milieu	ar ca ti an rd	car ard tio ati con	atio emen icat comp cati	th at ar an ti	tio car ard con ate	card cont teri atio mina	ei er di in un	ein sch ver ich ode	durc erde werd asse ungs
Début (marque _)	on_ ie_ re_ nt_ té_	ion_ gie_ ent_ age_ ité_	tion_ ogie_ ment_ ique_ ture_	on_ ng_ gy_ ty_ ry_	ion_ ing_ ogy_ ity_ ics_	tion_ logy_ ment_ ting_ lity_	ng_ ie_ en_ er_ on_	ung_ ion_ gie_ cht_ eit_	rung_ tung_ ogie_ tion_ gung_

TABLE 4.2 – Top-5 de n-grammes dans GEMET par fréquence décroissante

En début de mots (cf Table 4.2), les bi-grammes et les tri-grammes contiennent

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

des préfixes qui ne semblent pas être spécifiques à un domaine, ce sont donc les éléments de la langue générale. Les quadri-grammes, au contraire, laissent deviner des termes du domaine, par exemple, *méta* (FR) fait partie des termes *métaboliquement*, *métabolite*, *métal* ou *métallique*, *wass* (DE) provient certainement du terme *wasser* qui est placé au début de 72 termes monolexicaux et est inclus dans 158 termes monolexicaux au total.

Au milieu, les bi-grammes ne représentent pas d'intérêt, tandis que les tri-grammes et les quadri-grammes peuvent être utiles pour retrouver des termes. Toutefois, le tri par fréquence décroissante permet de constater que cette seule mesure ne suffit pas pour détecter les n-grammes particuliers d'un domaine. Nous allons calculer dans la section suivante le score terminologique des n-grammes afin d'identifier les plus spécifiques à un domaine.

A la fin, la quasi-totalité des n-grammes est constituée des suffixes propres aux substantifs (*-tion*, *-ment* (FR), *-logy*, *-ity* (EN) et *-ung*, *-gie* (DE)) et aux adjectifs (*-ique* (FR)). Certes, une partie de ces suffixes peut correspondre aux adverbes (*-ment* (FR)) ou aux participes présent (*-ing* (EN)), mais l'usage que l'on en fait minimise l'influence de ces éléments.

Dans notre méthode, les n-grammes du début et du milieu de mots servent à générer un modèle pour extraire des termes candidats monolexicaux. Les n-grammes de la fin des mots qui jouent le rôle de suffixes sont utilisés pour l'apprentissage de patrons morphosyntaxiques endogènes.

4.1.2.1 Construction du modèle

Une fois la liste initiale de termes constituée, nous pouvons segmenter ces termes en n-grammes de caractères. Nous allons évaluer les valeurs optimales du paramètre n dans la section 4.3.4.1.

L'objectif de cette étape de traitement est d'identifier les n-grammes propres aux termes du domaine. Notre idée consiste à sélectionner les n-grammes du début et du milieu qui sont présents dans un grand nombre de termes et de mots informatifs, sans que ces mots informatifs soient les plus fréquents du corpus. Pour cela, nous avons calculé le score terminologique des n-grammes selon la Formule 4.3.

$$n_gram_score = \frac{nb_termes_ressource + nb_mots_corpus}{\sum freq_mots_corpus} \quad (4.3)$$

où $nb_termes_ressource$ est le nombre de termes de la ressource qui contiennent le n-gramme en question ; nb_mots_corpus est le nombre de mots informatifs qui contiennent le n-gramme et $\sum freq_mots_corpus$ est la somme de fréquences de ces mots informatifs dans le corpus.

4.1. MÉTHODE D'EXTRACTION DE TERMES MONOLEXICAUX À L'AIDE DE RESSOURCES MORPHOLOGIQUES ENDOGÈNES

Position n	FR			EN			DE		
	2	3	4	2	3	4	2	3	4
Début (marque _)	_sc	_cri	_phyt	_cr	_agr	_agro	_ti	_ind	_umwe
	_oc	_mar	_supr	_sw	_geo	_phyt	_ök	_umw	_natu
	_go	_agr	_fina	_ru	_inf	_flow	_mu	_leb	_lebe
	_fu	_soc	_attr	_ec	_mar	_term	_ag	_öko	_wirt
	_hé	_sal	_croi	_ga	_dem	_expl	_zi	_alt	_geze
Milieu	ié	yto	ogra	rk	hyt	colo	ök	öko	welt
	-é	hyt	phyt	oe	ono	phyt	mw	hut	umwe
	go	xic	icit	nk	geo	alis	sü	mwe	mwel
	rk	asi	oxic	az	nom	clim	sn	keh	anze
	-i	été	hyto	sw	inf	lima	ud	klu	klun

TABLE 4.3 – Top-5 de n-grammes par score terminologique décroissant (B09C lemmatisé, GEMET)

La Table 4.3 contient les top-5 n-grammes après le calcul du score terminologique. Les bi-grammes ne semblent pas appartenir à un domaine, tandis que les tri et les quadri-grammes rappellent les termes du domaine :

- *phyt* (FR) est présent dans *phytomasse*, *phytosanitaire*, *phytoplankton*, *etc.* ;
- *clim* (EN) dans *climate*, *microclimate*, *palaeoclimatology*, *etc.* ;
- et enfin *umwe* (DE) fait partie de *umweltverschmutzung*, *umweltkosten* et *umweltplan*⁴.

4.1.3 Extraction des candidats monolexicaux

Une fois les n-grammes générés, le modèle est appliqué pour identifier des termes monolexicaux dans le corpus. Le score pour chaque mot informatif est calculé selon l'équation 4.4 et la liste des termes candidats est construite.

$$terme_score = \frac{\sum scores_n_grammes}{longueur_mot} \quad (4.4)$$

où $\sum scores_n_grammes$ est la somme des scores des n-grammes d'un mot présents dans le modèle et $longueur_mot$ est la longueur du mot en question.

Cette formule permet d'identifier les candidats qui contiennent plus de n-grammes par rapport à leur longueur. Ainsi, pour être sélectionnés, les candidats longs doivent contenir plus de n-grammes du modèle que les candidats courts. Nous avons fixé le seuil pour les candidats supérieur à zéro. Il est possible de définir un seuil supérieur pour accélérer le traitement des résultats. Toutefois, pour éviter des heuristiques, il est préférable de définir les déciles des résultats à garder. C'est-à-dire, au lieu de définir un seuil, l'utilisateur peut garder, par exemple, top-60 % des résultats triés dans l'ordre décroissant du score.

4. Pollution, coût environnemental et plan environnemental.

4.2. MÉTHODE D'EXTRACTION DE TERMES POLYLEXICAUX À BASE DE PATRONS MORPHOSYNTAXIQUES ENDOGÈNES

La liste de termes candidats monolexicaux est utilisée à l'étape suivante pour l'extraction des termes polylexicaux.

4.2 Méthode d'extraction de termes polylexicaux à base de patrons morphosyntaxiques endogènes

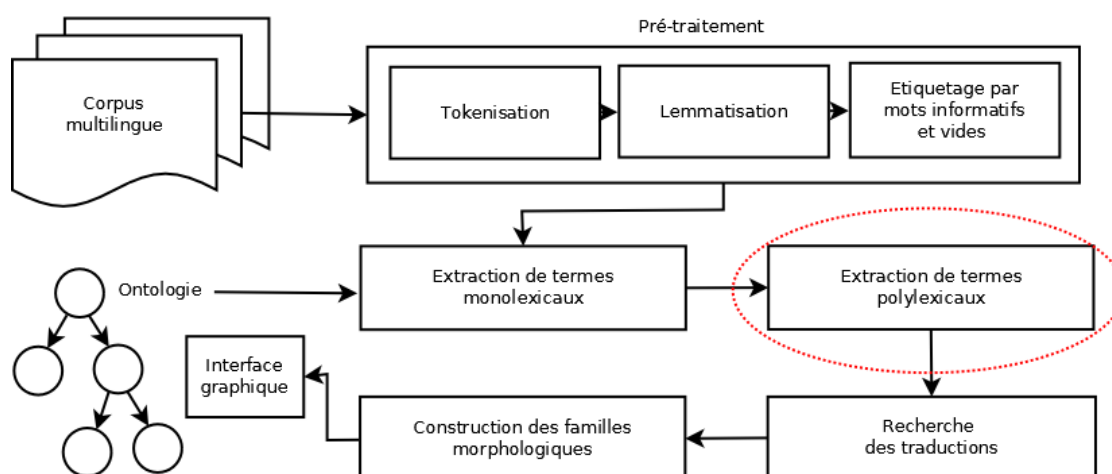


FIGURE 4.4 – Place de la recherche des termes polylexicaux dans notre méthode

L'extraction des termes polylexicaux s'effectue soit après la validation des termes monolexicaux (Figure 4.4), soit directement à partir de libellés monolexicaux présents dans l'ontologie source.

4.2.1 Apprentissage des patrons

Nous avons observé deux types de patrons pour l'extraction de termes candidats polylexicaux : les patrons morphosyntaxiques, c'est-à-dire, basés sur les étiquettes morphosyntaxiques, comme *NOM (PREP?(DET?)) NOM*, et les structures contrôlées de Vergne (2005) du type *I n I*, où *I* est un mot informatif et *n* – un mot vide. Les deux types de patrons représentent des inconvénients. Les patrons morphosyntaxiques sont dépendants de la langue et l'efficacité de la méthode dépend de la richesse de la liste de patrons fournie en avance, tandis que dans les structures contrôlées basées sur les annotations mot informatif – mot vide, tout mot informatif peut se trouver en tête de l'expression.

Notre méthode permet de profiter des points forts des deux approches sans avoir à résoudre les problèmes mentionnés ci-dessus. Tout d'abord, la structure de patrons endogènes n'est pas figée : ils sont appris dans la fenêtre de N^5 mots. Les patrons validés satisfont les conditions de commencer et de terminer par des mots informatifs, et de ne pas contenir de chiffres, ni de signes de ponctuation.

5. Afin de limiter le coût de la validation des résultats, nous avons fixé la longueur de termes polylexicaux à 4 mots.

4.2. MÉTHODE D'EXTRACTION DE TERMES POLYLEXICAUX À BASE DE PATRONS MORPHOSYNTAXIQUES ENDOGÈNES

La deuxième particularité de ces patrons est d'utiliser les n-grammes de la fin de mots informatifs pour remplacer les étiquettes morphosyntaxiques. Il faut noter que cette démarche prend en compte uniquement les termes dont la longueur est supérieure ou égale à la valeur du paramètre n . Par exemple, si le paramètre $n = 4$, le terme français *eau* ne sera pas traité et, par conséquent, les termes polylexicaux *aspiration de eau*, *réception de eau* ou encore *canalisation de eau* ne seront pas identifiés.

FR		EN	
3	4	3	4
ion_ que_	cédé_ n ⁶ tion_	cal_ ion_	ical_ tion_
édé_ n ion_	tion_ ique_	ium_ ate_	tion_ cess_
ent_ n ion_	oyen_ n tion_	ion_ ess_	ical_ ment_
ide_ que_	cide_ ique_	cal_ ent_	tion_ thod_
ion_ n eau_	ment_ n tion_	tic_ ion_	tion_ stem_

TABLE 4.4 – Top-5 de patrons morphosyntaxiques endogènes par fréquence décroissante (B09C lemmatisé, GEMET)

En effet, la Table 4.4 démontre que les patrons les plus fréquents correspondent à des syntagmes nominaux et adjectivaux et rappellent les patrons morphosyntaxiques *NOM (PREP ?(DET ?)) NOM*, *NOM ADJ* ou encore *VER (PREP / DET) NOM*.

4.2.2 Sélection des mots de schéma

L'apprentissage de patrons endogènes ne nécessite pas d'étiquettes morphosyntaxiques : les n-grammes de fin de mots remplissent cette fonction. Certes, les patrons endogènes n'excluent pas un certain niveau de bruit. Afin d'éviter ce bruit, nous pouvons restreindre les mots vides valides à la liste de mots de schéma générée à partir du corpus à l'étape de l'acquisition des patrons. Pour simplifier, nous avons choisi ce terme pour regrouper les mots fonctionnels (*pour* en français ou *for* en anglais) et les mots de schéma qui sont, selon la définition de Enguehard (1993), "des mots fonctionnels structurant des syntagmes" (*de*, *le*, *à* en français ou bien *of*, *the* en anglais). Les mots fortement liés sont définies comme "des chaînes de caractères comprenant des caractères blancs mais pourtant considérés comme des mots" (Enguehard, 1993). Dans notre approche, ces mots ne sont pas traités, mais ils apparaissent dans les résultats, car ils correspondent aux séquences de mots non-informatifs. Par exemple, le patron *-ion n n -ide* peut correspondre au terme candidat *sécrétion de l'acide*, où nous retrouvons le mot fortement lié *de l'*.

Pour construire cette liste, nous avons repéré les mots non-informatifs les plus fréquents qui apparaissent entre les mots informatifs à l'intérieur des patrons, en excluant la ponctuation et les chiffres. Les mots de schéma influencent la précision des résultats : plus leur liste est longue, plus la méthode extrait des candidats erronés. Il est alors intéressant de limiter cette liste à quelques mots non-informatifs les plus fréquents.

6. Où n est un mot non-informatif.

4.3. ÉVALUATION

En appliquant les patrons endogènes combinés avec les mots de schéma, nous arrivons à extraire les syntagmes susceptibles d'être des termes candidats polylexicaux. La précision de la méthode dépend de la quantité des mots de schéma : plus leur liste est longue, plus la méthode propose de candidats.

Les mots de schéma complètent les patrons morphosyntaxiques endogènes. Par exemple, le patron *-cédé n -tion* devient *-cédé de -tion*, ce qui restreint davantage les candidats qu'il peut extraire. Dans la section suivante, nous procédons à l'évaluation des résultats de l'extraction terminologique.

4.3 Évaluation

Avant de détailler les résultats de l'extraction terminologique, nous allons rappeler les ressources et les mesures utilisées dans l'évaluation, ainsi que l'outil de référence TTC TermSuite. Les résultats de ce dernier servent de baseline dans notre projet.

4.3.1 Ressources

Nous avons évalué l'étape de l'extraction terminologique sur deux corpus pré-annotés décrits dans la section 3.2.2.1. Les modèles ont été construits à partir de trois ressources sémantiques (EcoLexicon, GEMET et Agrovoc) et des termes candidats extraits à l'aide de fréquence absolue et de TFxIDF.

4.3.2 Mesures d'évaluation

La performance est évaluée via le rappel, la précision et la F-mesure calculés par rapport à la liste des termes attendus, disponible pour chaque langue et chaque corpus (cf Section 3.3.3.1).

Rappel

$$rappel = \frac{\text{nombre de termes retrouvés}}{\text{nombre de termes attendus}} * 100 \quad (4.5)$$

Le rappel (Formule 4.5) montre combien de termes sont retrouvés par rapport au nombre total de termes attendus.

Précision

$$précision = \frac{\text{nombre de termes retrouvés}}{\text{nombre de candidats}} * 100 \quad (4.6)$$

La précision (Formule 4.6) évalue le nombre de termes retrouvés correctement par rapport au nombre total de candidats.

4.3. ÉVALUATION

F-mesure

$$f_mesure = 2 * \frac{rappel * precision}{rappel + precision} \quad (4.7)$$

La F-mesure (Formule 4.7) est la moyenne harmonique du rappel et de la précision. Elle permet d'évaluer les résultats de point de vue de leur application : dans le cas de l'extraction terminologique, le système doit fournir la liste de candidats complète et précise. Un score bas de la F-mesure signifie qu'au moins une de ces deux exigences n'est pas satisfaite.

4.3.3 Système de référence

Afin de se positionner par rapport aux méthodes existantes, nous avons utilisé TTC Term Suite 2.2⁷ (Rocheteau et Daille, 2011; Daille *et al.*, 2011; Cram et Daille, 2016).

4.3.3.1 Présentation

L'outil est conçu pour extraire la terminologie à partir de corpus comparables en anglais, français, allemand, espagnol, letton, chinois et russe. Il s'appuie sur l'étiqueteur morpho-syntaxique TreeTagger et exploite les lemmes et les étiquettes morphosyntaxiques. Plus précisément, TTC TermSuite considère que les termes monolexicaux sont des noms ou des adjectifs et laisse à l'utilisateur l'option d'y inclure les verbes et les adverbes. Pour les termes polylexicaux, il extrait des syntagmes nominaux, adjectivaux et, éventuellement, verbaux.

Paramètres Nous avons utilisé TTC TermSuite 2.2 avec les paramètres suivants :

- inclure les verbes et les adverbes dans les termes ;
- seuil d'occurrences égal à 1 ;
- format de sortie - *.json*.

4.3.3.2 Résultats du système de référence

Nous avons évalué la baseline⁸ sur le corpus de référence et les termes mono et polylexicaux présents dans les résultats qui ont été inclus dans les listes de termes attendus.

Dans le cadre de notre projet, nous avons fixé la longueur maximale d'un terme polylexical à 4 mots. Afin de pouvoir comparer notre méthode et TTC TermSuite 2.2, nous avons pris en compte les candidats répondant à ce critère de longueur⁹.

TTC TermSuite est un outil très précis (Tables 4.5 et 4.6), surtout dans l'identification de termes monolexicaux. Toutefois, dans certains cas, la validation de candidats est

7. <http://termsuite.github.io/getting-started/#download-termsuite>

8. La baseline est le résultat d'un système par rapport auquel notre méthode est évaluée.

9. L'application de ce critère a écarté 292 candidats polylexicaux en français (12 %) et 55 en anglais (10 %).

10. Le total correspond à l'ensemble de termes candidats validés manuellement dans le corpus de référence.

4.3. ÉVALUATION

Corpus	Lang.	Candidats lemmes				Candidats tokens			
		Total ¹⁰	Rappel	Préc.	F-mes.	Termes att.	Rappel	Préc.	F-mes.
B09C	FR	1 275	51,48	87,98	64,96	1 524	48,92	88,82	63,09
	EN	1 356	36,58	91,01	52,18	1 626	33,76	91,81	49,37
	DE	2 327	50,60	87,01	63,99	2 594	49,08	86,67	62,67
C02F	FR	1 075	53,35	87,90	66,40	1 267	51,58	88,38	65,14
	EN	1 347	38,75	91,26	54,40	1 597	35,25	91,84	50,95
	DE	2 349	50,49	88,14	64,20	2 639	48,42	87,97	62,46

TABLE 4.5 – Évaluation TTC TermSuite 2.2 : termes monolexicaux

Corpus	Lang.	Candidats lemmes				Candidats tokens			
		Total	Rappel	Préc.	F-mes.	Termes att.	Rappel	Préc.	F-mes.
B09C	FR	1 478	58,22	51,48	54,64	1 522	51,08	44,13	47,35
	EN	1 317	52,50	66,92	58,84	1 362	52,79	66,95	59,03
C02F	FR	984	60,85	45,15	51,84	1 012	50,00	36,38	42,11
	EN	1 374	50,58	67,08	57,68	1 423	51,23	67,81	58,37

TABLE 4.6 – Évaluation TTC TermSuite 2.2 : termes polylexicaux

difficile. Par exemple, TTC TermSuite identifie les candidats polylexicaux contenant les conjonctions de coordination *et*, *ou* et même les virgules :

- *matrice solide, semi-solide ou liquide* ;
- *tuyau de aspiration et de conduite* ;
- *pentachlorure de phosphore ou du chlorure*, etc.

Nous avons décidé d’invalider ces candidats, car ils nécessitent un traitement complémentaire par l’utilisateur les séparant en plusieurs termes polylexicaux : *matrice solide*, *matrice semi-solide* et *matrice liquide*.

Une autre particularité de TTC TermSuite est l’omission des articles ou des pronoms pour regrouper les variantes de termes : l’occurrence *concentration of a covalent oxygen* (EN) est transformée en candidat *concentration of covalent oxygen, agent of this type* en *agent of type*, etc. Cette option, certes bénéfique pour l’exploitation des résultats par un utilisateur, complique notre validation. En effet, il est difficile d’identifier toutes les transformations faites par TTC TermSuite 2.2, même si elles sont mentionnées dans (Cram et Daille, 2016). Nous avons pris la décision de réintroduire les articles (*a*, *an* et *the* pour l’anglais ; *la*, *le*, *les*, *un*, *une* pour le français) dans leur forme lemmatisée (par exemple, *la* est transformé en *le*). Les autres cas seront exclus de l’évaluation. Au total, nous avons modifié 92 termes (B09C, FR) et 70 termes (C02F, FR).

4.3. ÉVALUATION

4.3.3.3 Comparaison de versions TTC TermSuite

Au début du projet, nous avons utilisé la version 1.6¹¹ de TTC TermSuite. Cette version fonctionnait sous format *.xmi*.

Le passage vers la version 2.2 est motivé à la fois par la volonté de tester la version la plus récente de l'outil et par des problèmes techniques avec le module *Indexer* qui ne générât pas de fichier de sortie nécessaire pour les alignements bilingues.

Corpus	Lang.	Candidats lemmes				Candidats tokens			
		Total	Rappel	Préc.	F-mes.	Total	Rappel	Préc.	F-mes.
B09C	FR	1 275	89,49	73,28	80,58	1 524	87,86	74,68	80,74
	EN	1 356	82,96	82,30	82,63	1 626	77,12	83,38	80,13
	DE	2 327	82,29	77,12	79,63	2 594	82,23	76,04	79,01
C02F	FR	1 075	91,07	71,93	80,38	1 267	90,69	73,28	81,06
	EN	1 347	84,48	82,58	83,52	1 597	78,96	83,18	81,02
	DE	2 349	82,63	77,30	79,88	2 639	82,53	76,05	79,16

TABLE 4.7 – Évaluation TTC TermSuite 1.6 : termes monolexicaux

Toutefois, nous avons remarqué que la version TTC TermSuite 1.6, est meilleure pour l'extraction des termes monolexicaux : elle fournit un meilleur rappel et F-mesure (Table 4.7). Dans l'évaluation de la version 1.6, nous avons appliqué exactement la même stratégie que pour la version 2.2. Il est difficile d'expliquer cette différence de résultats entre les deux versions.

4.3.4 Résultats

Les résultats de notre méthode dépendent de la valeur du paramètre n . Nous allons donc commencer par les tests de ce paramètre. Ensuite, nous fournirons les résultats de l'extraction des termes mono et polylexicaux sur les corpus pré-annotés par les mots informatifs et vides, sans seuil d'amélioration.

4.3.4.1 Choix des valeurs du paramètre n

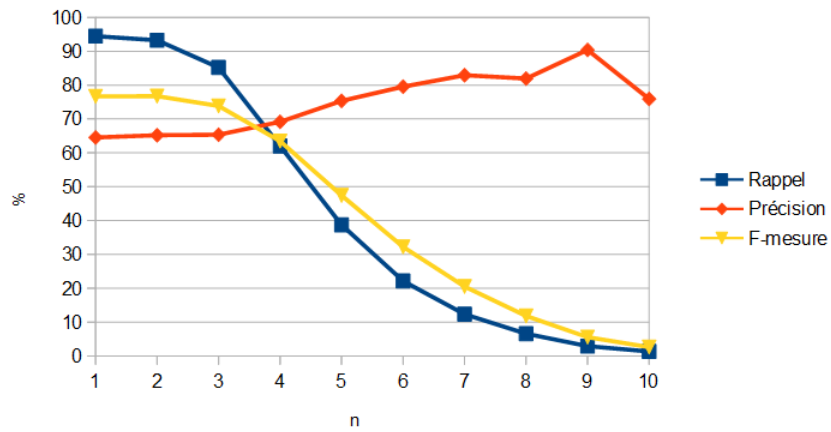
Notre approche d'extraction terminologique, ainsi que celle de la recherche de traductions, repose sur les n -grammes de caractères. Il est donc nécessaire de définir les valeurs de n les plus performantes pour chaque langue.

Nous avons généré un modèle de n -grammes pour chaque valeur de n où $n \in (1 : 10)$ à partir de chaque ressource¹² et l'avons appliqué à l'extraction des termes monolexicaux à partir du corpus. Les termes d'une ressource sont alors segmentés en n -grammes de longueur n .

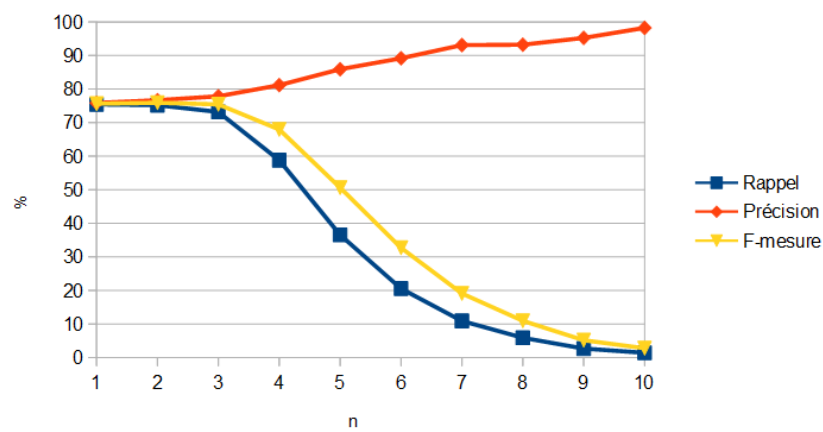
11. <https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite/wiki/Readme>

12. Pour chaque ressource, on exclut les termes présents sur la liste des termes attendus. De cette manière, le processus est évalué sur trois ressources et les modèles générés ne contiennent pas les termes attendus en résultat.

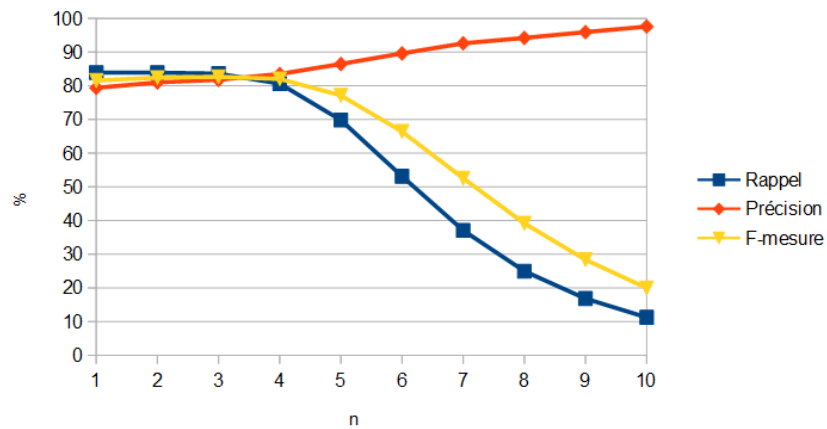
4.3. ÉVALUATION



(a) FR



(b) EN



(c) DE

FIGURE 4.5 – Évaluation du paramètre n (valeurs moyennes pour trois ressources et deux corpus)

4.3. ÉVALUATION

La Figure 4.5 montre que les valeurs optimales sont comprises entre 2 et 4 pour toutes les langues. En effet, plus la valeur de n est grande, plus le modèle contient de n-grammes. En même temps, le nombre de termes servant à leur génération se réduit, car seuls les mots longs peuvent contenir des n-grammes de taille importante. Les bi-grammes ne sont pas assez spécifiques pour identifier les termes d'un domaine et leur modèle est très réduit. Le choix entre les tri-grammes et les quadri-grammes n'est pas simple : d'un côté il est possible que les quadri-grammes puissent donner plus de précision, mais le nombre de termes candidats sera assez réduit ; les tri-grammes peuvent générer plus d'imprécision, mais le nombre de termes candidats sera plus élevé. Dans l'évaluation de notre méthode, nous allons utiliser $n = 3$.

4.3.4.2 Termes monolexicaux

Le modèle de n-grammes pour l'extraction des termes monolexicaux est généré soit à partir de termes déjà présents dans l'ontologie source, soit à partir des candidats extraits par fréquence et TFxIDF.

C. ¹³	L.	Total	EcoLexicon			GEMET			Agrovoc		
			R.	P.	F.	R.	P.	F.	R.	P.	F.
B09C	FR	1 275	78,59	64,56	70,89	91,45	63,96	75,27	93,80	63,48	75,72
	EN	1 356	90,56	77,82	83,71	90,27	77,57	83,44	94,17	77,07	84,77
	DE	2 327	84,23	81,43	82,81	84,14	81,31	82,70	84,57	81,29	82,90
C02F	FR	1 075	78,88	62,58	69,79	92,74	62,20	74,46	95,16	61,78	74,92
	EN	1 347	91,61	76,03	83,10	90,72	75,71	82,54	94,80	75,25	83,90
	DE	2 349	82,80	82,31	82,56	82,93	82,16	82,54	82,97	82,13	82,55

TABLE 4.8 – Évaluation des termes monolexicaux sur les corpus lemmatisés, modèles à partir des ressources, $n = 3$

Modèles à partir de ressources terminologiques Les résultats obtenus (Table 4.8) pour les trois modèles confirment la stabilité de notre méthode. En effet, la taille de ressource utilisée pour la génération du modèle varie considérablement : EcoLexicon compte seulement 622 termes français, contre plus de 3 000 pour GEMET et plus de 33 000 pour Agrovoc (cf Table 3.2.2.1). Le nombre réduit de termes français dans EcoLexicon influence le rappel, mais pas la précision. Ce modèle permet d'identifier un peu plus de 78 % de termes du corpus, contre plus de 90 % avec d'autres modèles.

Notre méthode est meilleure en matière de rappel et de F-mesure par rapport au TTC TermSuite 2.2 (Table 4.5) et est très proche des résultats de TTC TermSuite 1.6 (Table 4.7). Nous pouvons également constater que les meilleurs résultats sont atteints pour l'allemand, tandis qu'en français la précision est plus basse.

4.3. ÉVALUATION

C.	L.	Total	EcoLexicon			GEMET			Agrovoc		
			R.	P.	F.	R.	P.	F.	R.	P.	F.
B09C	FR	1 524	78,67	60,04	68,11	92,39	59,61	72,47	94,75	59,40	73,02
	EN	1 626	91,70	77,01	83,72	91,33	76,63	83,33	94,53	76,13	84,33
	DE	2 594	96,53	74,77	84,27	96,65	74,64	84,23	97,11	74,57	84,36
C02F	FR	1 267	80,43	57,38	66,97	94,08	56,63	70,70	95,90	56,17	70,85
	EN	1 597	93,17	75,46	83,38	92,11	75,13	82,76	95,49	74,75	83,86
	DE	2 639	97,88	74,03	84,30	98,18	73,84	84,29	98,22	73,72	84,22

TABLE 4.9 – Évaluation des termes monolexicaux sur les corpus tokenisés, modèles à partir des ressources, $n = 3$

Une des particularités de notre approche est son applicabilité sur des corpus non-lemmatisés. En effet, les outils de lemmatisation et d'étiquetage morphosyntaxique ne sont pas disponibles pour toutes les langues, et la possibilité de contourner cette étape peut être intéressante pour certains projets. La Table 4.9 permet de constater que la méthode préserve ses performances quant à l'extraction des termes monolexicaux. Il est toutefois nécessaire de remarquer que l'absence de lemmatisation signifie que c'est à l'utilisateur de définir les formes de base pour les insérer dans la structure sémantique cible.

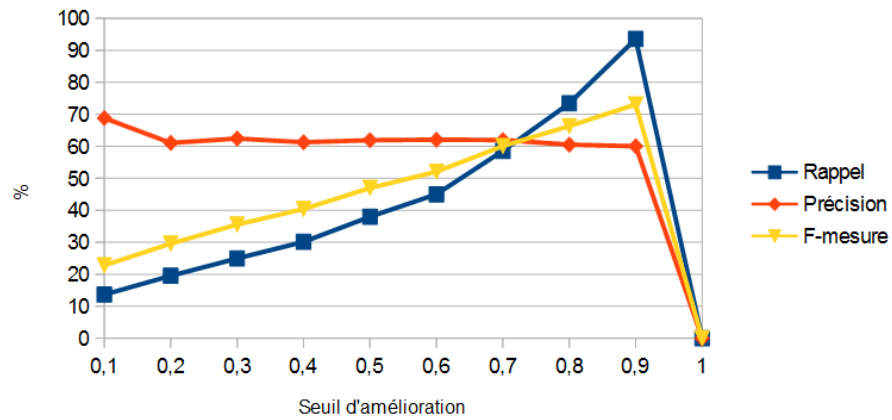
Modèle à partir de candidats extraits par la fréquence absolue et TFXIDF

Le modèle construit à partir des top-candidats des méthodes statistiques peut s'appliquer si l'ontologie d'origine ne contient pas de termes ou très peu de termes une langue. Comme ces candidats ne font pas l'objet d'une validation manuelle préalable, cette approche peut être considérée comme moins sûre par rapport à la liste de termes du domaine issue d'une ressource validée.

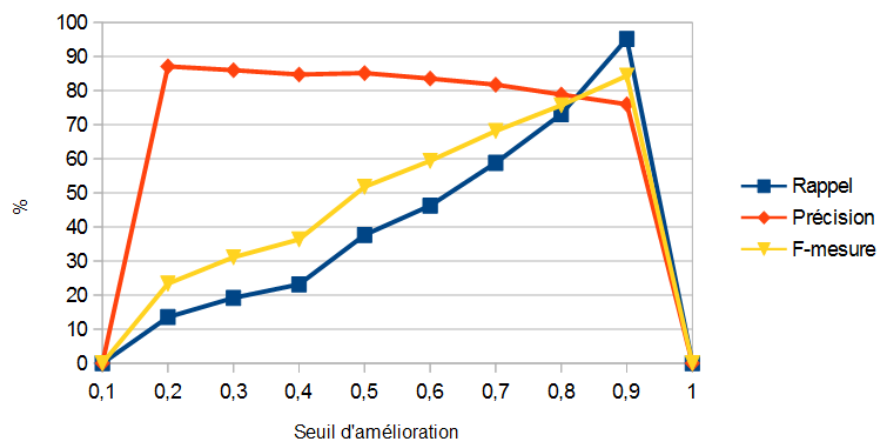
Pour pallier ce risque en augmentant la densité de termes dans les mots informatifs, nous avons augmenté le seuil d'amélioration de Vergne jusqu'à 0,9 à l'étape de calcul de la fréquence et de TFXIDF. Rappelons que ce seuil aide à harmoniser l'étiquetage par des mots informatifs et vides. Il correspond au pourcentage des cas où le mot en question doit être étiqueté comme informatif pour garder son étiquette (90 % dans ce cas).

13. C. est l'abréviation de corpus, L. de langues, R. du rappel, P. de la précision et F. de la F-mesure.

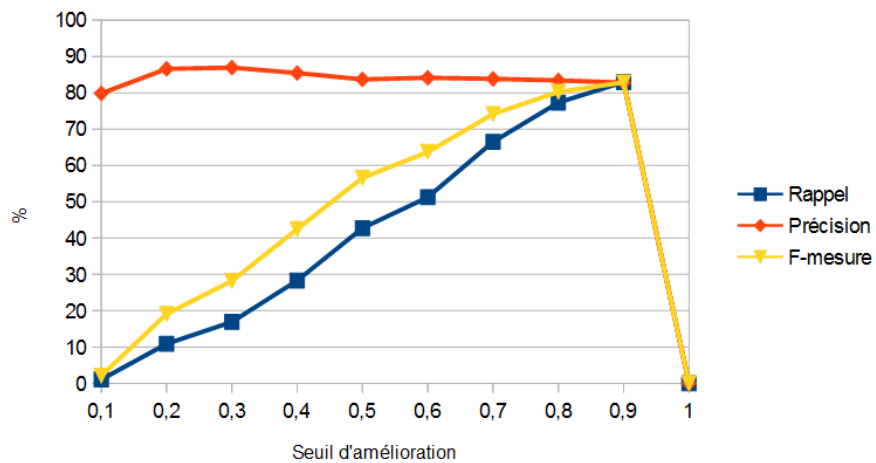
4.3. ÉVALUATION



(a) FR



(b) EN



(c) DE

FIGURE 4.6 – Évaluation du seuil Top-X % pour les mesures statistiques, corpus lemmatisé

4.3. ÉVALUATION

Les meilleurs résultats (Figure 4.6) sont atteints lorsque le seuil Top-X % est égal à 90. C'est-à-dire, le modèle est généré à base de l'intersection entre les Top-90 % de candidats pour fréquence et pour TfxIDF. Rappelons que les candidats utilisés dans la génération du modèle ne font pas partie de la liste extraite par approche endogène.

Corpus	Lang.	Total	Corpus lemmatisé			Total	Corpus tokenisé		
			R.	P.	F.		R.	P.	F.
B09C	FR	955	91,10	60,63	72,80	1 179	92,54	56,15	69,89
	EN	1 010	91,39	78,89	84,68	1 244	93,49	77,53	84,77
	DE	1 943	84,20	82,38	83,28	2 226	98,07	74,61	84,74
C02F	FR	752	96,01	59,42	73,41	933	96,36	53,32	68,65
	EN	768	98,96	73,08	84,07	922	100,00	72,01	85,32
	DE	2 105	81,71	83,21	82,45	2 231	100,00	73,90	85,09

TABLE 4.10 – Évaluation des termes monolexicaux sur les corpus tokenisés, modèles à partir de Top-90 % de statistiques, $n = 3$

Le modèle basé sur l'intersection des Top-90 % candidats identifiés par les mesures statistiques fonctionne aussi bien que ceux générés à partir des mesures statistiques (Table 4.10). Le nombre de mots utilisés pour le modèle est proche de 400, dont plus de 70 % de termes valides (cf Figure 4.3). Cette configuration dépasse également la baseline en termes de rappel et de F-mesure.

4.3.4.3 Choix du nombre de mots de schéma

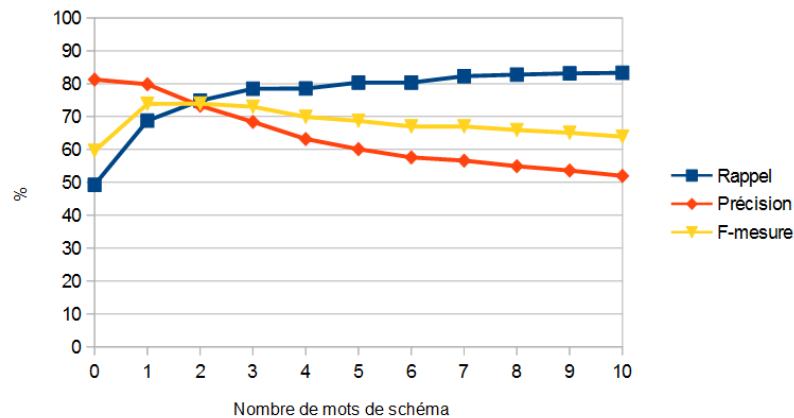
Les performances de notre approche dépendent du nombre de mots vides pris en compte (Figure 4.7). Plus leur nombre est grand, plus le système propose de termes candidats polylexicaux. Cela joue sur la précision, car certains mots de schéma, par exemple, *ou*, *et*, etc., sont moins pertinents pour la création de termes.

Corpus	Lang.	Mots non-informatifs									
		de	le	du	et	un	être	à	ou	en	dans
B09C	FR										
	EN	the	of	and	be	in	a	to	or	from	with
C02F	FR	de	le	un	et	du	être	à	dans	en	ou
	EN	the	of	and	be	in	a	to	water	from	or

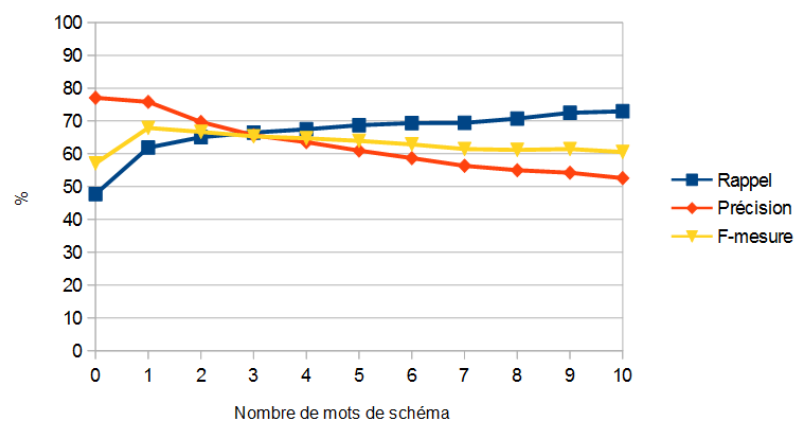
TABLE 4.11 – Les top-10 mots non-informatifs triés par fréquence décroissante, corpus lemmatisé

Pour les langues où les articles changent en nombre et en genre, comme c'est le cas du français et de l'allemand, l'approche fonctionne mieux sur le corpus lemmatisé. Les meilleurs résultats sont atteints sur le corpus lemmatisé avec deux mots de schéma, qui sont *de* et *le* pour la français et *the* et *of* pour l'anglais (cf Table 4.11), ce qui correspond au génitif.

4.3. ÉVALUATION



(a) Corpus lemmatisé



(b) Corpus tokenisé

FIGURE 4.7 – Évaluation du nombre de mot de schéma pour l'extraction de termes polylexicaux

Corpus	Lang.	Mots non-informatifs									
B09C	FR	de	et	la	un	des	à	ou	une	du	en
	EN	the	of	and	in	a	to	is	or	from	with
C02F	FR	de	et	la	à	une	un	dans	des	est	en
	EN	the	of	and	in	is	a	to	water	from	or

TABLE 4.12 – Les top-10 mots non-informatifs triés par fréquence décroissante, corpus tokenisé

La détérioration des résultats sur le corpus non-lemmatisé s'explique par une plus grande variation au sein des mots de schéma, surtout en français (Table 4.12).

4.3. ÉVALUATION

4.3.4.4 Termes polylexicaux

L'extraction des termes polylexicaux s'effectue après la constitution de la liste de termes monolexicaux. Elle peut également tenir compte des termes monolexicaux déjà présents dans l'ontologie.

Notre méthode exploite l'annotation par les mots informatifs et vides et apprend des patrons morphosyntaxiques endogènes où les mots informatifs sont représentés par leurs n-grammes situés à la fin de mot, et les mots vides sont réduits à ceux les plus fréquents dans le corpus. De plus, le terme candidat polylexical doit commencer et se terminer par un terme monolexical. Dans l'objectif de limiter la validation, nous avons fixé la longueur maximale d'un terme polylexical à 4 mots.

Corpus	Lang.	Total	Lemmatisé, 2 m. s. ¹⁴			Total	Tokenisé, 1 m. s.		
			R.	P.	F.		R.	P.	F.
B09C	FR	1 478	72,80	79,88	76,18	1 522	54,66	80,54	65,13
	EN	1 359	78,81	68,92	73,53	1 405	68,26	72,98	70,54
C02F	FR	984	71,65	72,09	71,87	1 012	55,63	71,54	62,59
	EN	1 397	76,09	72,26	74,13	1 446	68,95	78,20	73,28

TABLE 4.13 – Évaluation des termes polylexicaux (en utilisant les termes monolexicaux validés)

Liste de termes monolexicaux validée manuellement Notre méthode fonctionne nettement mieux sur le corpus lemmatisé (Table 4.13). Toutefois, il est possible de modifier manuellement la liste des mots de schéma autorisés en y ajoutant toutes les formes pour les articles (*le, la, les, un, une, des,* etc.) et améliorer ainsi le rappel.

Dans cette configuration, la méthode dépasse les performances de TTC TermSuite 2.2¹⁵, bien que cela pourrait être dû à la différence dans la politique de regroupement des termes contenant les termes moins longs et dans la limitation de la longueur de termes à 4 mots. Rappelons que dans le cadre de notre méthode, un terme polylexical inclus dans un terme plus long, est enregistré séparément, sauf si les fréquences de ces deux termes sont identiques. Dans ce cas, seul le terme plus long est enregistré.

14. 2 m. s. signifie 2 mots de schéma.

15. Dans cette évaluation, nous n'avons pas évalué la significativité statistique des résultats. Cette évaluation fera partie des futures expériences dans le cadre du développement ultérieur du système.

4.4. DISCUSSION

Corpus	Lang.	Total	Lemmatisé, 2 m. s.			Total	Tokenisé, 1 m. s.		
			R.	P.	F.		R.	P.	F.
B09C	FR	1 478	55,35	31,09	39,82	1 522	43,56	34,30	38,38
	EN	1 359	61,44	33,02	42,95	1 405	53,31	33,17	40,90
C02F	FR	984	57,62	25,39	35,25	1 012	45,16	25,73	32,78
	EN	1 397	56,34	32,39	41,13	1 446	52,21	35,17	42,03

TABLE 4.14 – Évaluation des termes polylexicaux (en utilisant les candidats monolexicaux, modèle GEMET)

Liste de termes monolexicaux obtenue par l’extracteur monolexical L’utilisation de la liste de candidats monolexicaux sans validation préalable par un expert humain entraîne une détérioration des résultats (Table 4.14), surtout en matière de précision. La baisse du rappel est à son tour due à des termes monolexicaux non identifiés par l’extracteur.

4.4 Discussion

L’évaluation quantitative des résultats présentée dans la section précédente, nous a permis de comparer notre méthode avec la baseline et de juger de sa conformité aux objectifs du projet. Toutefois, un examen plus détaillé de certains aspects contribuera à la compréhension de notre approche et mettra l’accent sur ses particularités.

4.4.1 Influence de l’annotation par les mots informatifs et vides

L’annotation par mots informatifs et vides joue un rôle très important dans notre méthode. En effet, les termes candidats monolexicaux sont choisis parmi les mots informatifs, tandis que les mots vides sont utilisés en tant que mots de schéma dans les candidats polylexicaux.

Cette annotation donne les meilleurs résultats sur les textes en allemand, car la différence de longueurs entre les mots est très prononcée. En français, il arrive que certains mots sont annotés à tort en tant que informatifs, par exemple le mot *dudit* dans *production dudit matériau filtrant*. Tout comme dans le cas de patrons morphosyntaxiques traditionnels, il est alors difficile de limiter le résultat aux bons termes candidats.

Il n’existe pas de référentiel concret pour évaluer si un mot est informatif ou vide par rapport à la tâche. Le seul moyen de valider ou invalider certaines annotations est de suivre son intuition.

4.4. DISCUSSION

FR : pendant_(I) le_(n) électroréparation_(I) de_(n) champ_(I) électrique_(I) de_(n) sol_(I) générer_(I) par_(n) deux_(I) électrode_(I) (_(n) @card@_(n) ,_(n) @card@_(n))_(n) situer_(I) dans_(n) ce_(n) sol_(I) ,_(n) on_(n) utiliser_(I) pour_(n) le_(n) décontamination_(I) du_(n) agent_(I) complexants_(I) ,_(n) du_(n) réactif_(I) de_(n) chimie_(I) et_(n) /_(n) ou_(n) du_(n) nutriments_(I) à_(n) travers_(I) le_(n) sol_(I) ou_(n) pour_(n) chauffer_(I) le_(n) sol_(I) ._(n)

EN : a_(n) method_(I) of_(n) process_(I) mineral_(I) oil-containing_(I) material_(I) ,_(n) which_(I) be_(n) spread_(I) on_(n) a_(n) support_(I) and_(n) be_(n) sprinkle_(I) with_(n) a_(n) lime-containing_(I) reactant_(I) which_(I) have_(I) be_(n) adjust_(I) to_(n) be_(n) hydrophobic_(I) ,_(n) whereafter_(I) the_(n) mineral_(I) oil-containing_(I) material_(I) be_(n) intimately_(I) mix_(I) with_(n) the_(n) reactant_(I) so_(I) that_(I) a_(n) fine-grained_(I) water-repellent_(I) reaction_(I) product_(I) be_(n) obtain_(I) after_(I) a_(n) reaction_(I) time_(I) [...]

DE : bei_(n) eine_(n) verfahren_(n) zu_(n) arbeiten_(n) von_(n) mineral_(n) ölhaltigem_(I) gut_(n) ,_(n) die_(n) auf_(n) eine_(n) auflage_(I) aufgebracht_(I) ,_(n) mit_(n) eine_(n) hydrophob_(I) eingestellt_(I) ,_(n) kalk_(n) enthaltend_(n) reaktionsmittel_(I) bestreuen_(I) und_(n) dann_(n) mit_(n) die_(n) reaktionsmittel_(I) innig_(I) vermischen_(n) werden_(n) ,_(n) werden_(n) nach_(n) eine_(n) reaktionszeit_(I) eine_(n) feinkörni_(I) ges_(I) ,_(n) wasserabweisend_(I) reaktionsprodukt_(I) erhalten_(n) ._(n)

Dans les exemples ci-dessus, nous considérons que les mots en gras sont mal étiquetés en tant que informatifs :

- FR – *pendant* et *travers* ;
- EN – *which*, *have*, *so*, *that* et *after*.

En allemand, au contraire, certains mots sont étiquetés en tant que vides : *arbeiten*, *kalk*, *gut*, etc. Les deux cas de figures introduisent des erreurs dans le repérage de candidats mono et polylexicaux.

4.4.2 Candidats extraits par des mesures statistiques

La particularité de notre approche est d'appliquer les mesures statistiques sur des mots étiquetés comme informatifs par la méthode de Vergne (2005). Selon cette méthode, les mots informatifs sont plus longs et moins fréquents que les mots vides. Ce traitement préalable améliore significativement les résultats des mesures statistiques et permet de construire une liste de candidats suffisamment précise pour générer un modèle de n-grammes.

Les candidats mis en valeur par la fréquence absolue et par le TFXIDF sont différents. Ainsi, dans la liste de fréquences, les positions en tête de liste sont occupées par les termes assez généraux, comme *matière*, *traitement* ou encore *élément*, et par le lexique propre aux brevets : *revendication*, *caractériser*, *invention*. Nous avons décidé de ne pas considérer le

4.4. DISCUSSION

lexique de brevets en tant que termes valides (cf Section 3.3.3.1). Les mots vides n'étant pas annotés comme tels se trouvent également en tête de la liste de fréquence, par exemple *autre*, *moins* ou encore *dass*, *dadurch* (DE) et *any*, *first*, *more* (EN).

La liste des candidats triés par score TFxIDF décroissant commence par des termes plus spécifiques, comme *formule*, *alkyl*, *sol*, *fructosane*, etc. De plus, la tête de la liste contient peu de bruit.

Nous avons décidé d'appliquer l'intersection entre ces deux méthodes au lieu de prendre le Top-X % de résultats (Korechuk, 2014) pour deux raisons. D'abord, il est difficile d'évaluer le seuil Top-X %, car sa valeur optimale peut différer entre les deux méthodes et, ensuite, à cause de la précision accrue fournie par l'intersection. De plus, nous avons déjà mentionné la présence des candidats non termes dans la liste triée par fréquence.

Cette intersection est composée de termes hétérogènes :

- termes techniques (*tuyau*, *coque*, *curtain* (EN), *pool* (EN), *Perforationslocher* (DE), *Langsamfilter* (DE), etc.) ;
- termes chimiques (*oxydant*, *halogène*, *polyol* (EN), *Pyrimidin-derivate* (DE)) ;
- termes liés au traitement des sols et de l'eau (*terrain*, *grain*, *liquefy* (EN), *fluidize* (EN), *Biomasse* (DE) ou encore *Dekontamination* (DE)).

Cette hétérogénéité représente un avantage pour la méthode, car le modèle obtenu à partir de ces listes cible ces mêmes catégories de termes.

4.4.3 Candidats extraits par les ressources morphologiques endogènes

Dans la section des résultats, nous avons présenté les évaluations de l'extraction des termes monolexicaux dans leur globalité. Il est toutefois intéressant d'analyser la liste de candidats triée dans l'ordre décroissant du score.

FR	EN	DE
marché	phyto	Umweltbelastung
phytoréparation	economical	Umweltverträglich
socle	geoelectrical	Umweltfreundlich
bois	influence	indem
physique	demand	Umwandelnden
coton	synthetic	Waldbrand-
saleté	must	wirtschaftlich
agricole	agro-alimentary	umgekehrt
diagonal	municipal	lebend
herbicide	marsh	forstwirtschaftlich

TABLE 4.15 – Top-10 candidats monolexicaux pour les ressources morphologiques endogènes, corpus lemmatisé B09C

Les candidats ayant les scores maximaux sont presque tous des candidats valides (Table 4.15). Certains parmi les meilleurs candidats français et anglais sont formés à partir des

4.4. DISCUSSION

formants classiques :

- *phytoréparation, physique, agricole*, etc. (FR) ;
- *phyto, economical, geoelectrical*, etc. (EN).

Les termes allemands sont plutôt repérés grâce à des mots inclus dans le terme monolexical composé, comme c'est le cas du n-gramme *_umwe*, lié à *Umwelt*, qui a certainement influencé les candidats *Umweltbelastung*, *Umweltverträglich* et *Umweltfreundlich*. Il faut remarquer que l'allemand, comparé avec le français et l'anglais, emploie moins de formants classiques.

4.4.4 Patrons morphosyntaxiques endogènes

Les patrons morphosyntaxiques endogènes, intéressants du point de vue de leur structure, constituent une alternative aux patrons morphosyntaxiques classiques. Leurs résultats, tout comme ceux dans patrons classiques, peuvent être présentés de plusieurs manières.

4.4.4.1 Alternative aux patrons morphosyntaxiques classiques

Le recours à des patrons combinant l'annotation par mots informatifs et vides avec les n-grammes de fin des mots informatifs constitue une particularité de notre méthode. Les patrons sont appris à partir du corpus et ont l'avantage d'être indépendants de la qualité de l'étiquetage morphosyntaxique. Ce dernier dépend du vocabulaire couvert par le modèle disponible, et les termes d'un domaine peuvent y manquer.

FR	EN	DE
une DET :ART un	a DT a	o3-gasgemisch NN o3-gasgemisch
solution NOM solution	geopolymeric JJ geopolymeric	schüttgüter NN schüttgüter
de PRP de	barium NN barium	sehr ADV sehr
polymère ADJ polymère	sulphate NN sulphate	effektive ADJA effektiv
cationique NOM cationique	granulate VVP granulate	oxidieren NN oxidieren
ou KON ou	. SENT .	
anionique ADJ anionique		

TABLE 4.16 – Exemples d'étiquetage morphosyntaxique erroné, format token-étiquette-lemme, TreeTagger, corpus B09C

Les erreurs d'étiquetage peuvent réduire la précision des approches basées sur des patrons morphosyntaxiques prédéfinis. Par exemple, en français, les adjectifs sont le plus souvent positionnés après le nom qu'ils caractérisent. L'exemple d'étiquetage en français (Table 4.16) pourrait alors ne pas être repéré. Dans l'exemple anglais, le nom *granulate* est confondu avec son homographe verbal, alors que le patron ADJ-NOM-NOM-VER ne correspond pas a priori ni aux syntagmes nominaux, ni à ceux adjectivaux. Le cas illustré pour l'allemand, au contraire, pourrait introduire un faux candidat : le verbe *oxidieren* est étiqueté en tant que nom, ce qui rend possible l'association avec le patron ADJ-NOM.

Les jeux d'étiquettes morphosyntaxiques, même pour le même outil TreeTagger (Schmid, 1994)) diffèrent d'une langue à l'autre (Table 4.16) et les patrons morphosyntaxiques nécessitent d'être adaptés en conséquence. Par rapport à cet aspect des approches basées sur patrons prédéfinis, notre méthode présente un grand avantage : les patrons sont extraits à partir du texte sans aucune analyse préalable. Ainsi, nous arrivons à extraire des termes candidats assez complexes, comme :

- FR – *polystyrène fonctionnalisé réticulé, surstructure cristalline inverse, enzyme protoporphyrinogène oxydase*, etc.
- EN – *protoporphyrinogen oxidase enzyme-inhibiting herbicide, strong electrolysis solution, 2-sulfato ethyl methacrylate ammonium*, etc.

4.4.4.2 Présentation des résultats

La manière la plus simple d'exploiter la liste de patrons obtenus est de la trier par fréquence décroissante. Il faut noter également que ce ne sont pas toujours les patrons les plus fréquents qui donnent le plus de résultats pertinents. La variation de résultats pour un même patron peut être assez intéressante.

Par exemple, nous pouvons présenter les résultats pour le patron le plus fréquent dans le corpus B09C (FR) *ent_ n ion_* de manière suivante :

- agent de oxydation ;
- agent de lixivation ;
- agent de extraction ;
- élément de aspiration ;
- traitement de extraction ;
- auvent de aspiration ;
- agent de réaction ;
- élément de construction ;
- instrument de excavation, etc.

Toutefois, cette présentation a l'inconvénient de mélanger des termes n'ayant pas de relations entre eux. Cela peut ralentir le travail d'un spécialiste humain qui sera amené à changer constamment de sujet.

Il est également possible de regrouper les candidats par mot informatif (positionné en tête, à la fin ou présent dans le candidat polylexical). Ainsi, la liste de candidats pour le terme *agent* (FR) comprend :

- agent tensio-actif anionique ;
- agent de conditionnement ;
- agent de oxydation ;
- agent complexants ;
- agent de oxydation écologique¹⁶ ;
- agent composer de poudre ;
- agent décollants ;

16. Les candidats faisant partie des candidats plus longs sont retenus si la fréquence du candidat court et différente de celle du candidat long. Ainsi, les deux candidats, *agent de oxydation* et *agent de oxydation écologique*, sont retenus, car leurs fréquences respectives sont égales à 5 et à 1.

4.5. CONCLUSION

- agent de lixivation, etc.

Cette présentation est plus cohérente, car l'utilisateur se concentre sur la description des différents agents et peut les insérer dans l'ontologie en un seul bloc de travail. Dans la même logique, il est intéressant de proposer les termes candidats polylexicaux lors de la validation d'un terme candidat monolexical, car cela fournit à l'utilisateur plus d'informations sur ce dernier.

Certains patrons sont très productifs, d'autres le sont moins. En général, plus un patron est long, moins il fournit de résultats. Ainsi, le patron *ion_ que_* a 17 candidats, dont *opération arithmétique*, *ion métallique*, *consommation électrique*, *oxydation chimique* ou encore *tension électrique*, tandis que celui *ère_ que_ que_ ser_* n'a qu'un seul : *matière organique toxique volatiliser*.

4.5 Conclusion

La méthode proposée permet d'extraire les termes candidats mono et polylexicaux à partir d'un corpus de spécialité. L'avantage principal de la méthode est son applicabilité à une large palette de langues et de corpus. Elle ne nécessite aucune ressource morphologique ou morphosyntaxique extérieure. Ainsi, cette méthode peut être appliquée sur un corpus de tout domaine de spécialité et toutes les ressources sont alors générées par le programme.

La particularité de la méthode est d'utiliser des n-grammes de caractères pour remplacer des listes de formants classiques et des étiquettes morphosyntaxiques. Le modèle de n-grammes de début et du milieu de mots peut être généré soit à partir des termes déjà présents dans l'ontologie source, soit à partir de termes candidats monolexicaux qui se trouvent dans le Top-X % de l'intersection entre la fréquence absolue et TFxIDF. Ce modèle sert à l'identification des termes monolexicaux. La méthode se montre meilleure en matière de rappel et de F-mesure par rapport au TTC TermSuite 2.2 et est très proche des résultats de TTC TermSuite 1.6.

Les termes polylexicaux sont identifiés grâce aux patrons morphosyntaxiques endogènes qui combinent l'annotation par mots informatifs et vides de J. Vergne et les n-grammes de fin de mots informatifs. Il faut toutefois remarquer que l'étiquetage des mots informatifs a ses particularités. Par exemple, des termes courts d'un domaine, comme *air* ou *eau* peuvent être étiquetés comme des mots vides. Ou, au contraire, des mots longs, comme *particulièrement* sont étiquetés comme informatifs et se retrouvent parmi les termes candidats. Les mots vides sont alors restreints à des mots de schéma les plus fréquents appris à partir du corpus et les étiquettes des mots informatifs sont remplacées par les n-grammes de fin de mot. De plus, un terme candidat polylexical doit commencer et se terminer par un terme monolexical. Appliquée sur les corpus lemmatisés avec la liste de termes monolexicaux validés, la méthode dépasse les performances de TTC TermSuite 2.2.

La méthode peut être utilisée pour d'autres langues que le français, l'anglais et l'allemand à condition qu'un tokeniseur et un lemmatiseur soient disponibles pour la langue ciblée. Même si la tâche de lemmatisation nécessite un étiquetage morphosyntaxique, notre méthode ne l'utilise pas. De cette manière, la méthode exploite l'outil d'étiquetage morphosyntaxique différemment des autres méthodes de l'extraction terminologique. De plus, elle

4.5. CONCLUSION

peut fonctionner sur un corpus non lemmatisé et fournit des résultats avec une perte de précision d'environ 5 points (de 73,81 % à 69,43 %) en moyenne pour les termes monolexicaux par rapport aux corpus lemmatisés et presque sans perte pour les termes polylexicaux.

Chapitre 5

Exploitation des n-grammes de caractères pour la recherche des traductions

L'étape de recherche des traductions est incontournable lorsqu'il s'agit de l'extraction terminologique multilingue. En effet, une fois que nous avons constitué les listes de termes candidats pour chaque langue, l'objectif est d'y trouver des traductions.

Cette tâche peut être réalisée à l'aide d'un corpus parallèle où chaque phrase est accompagnée de sa traduction. Toutefois, ce type de corpus est assez rare et disponible pour un nombre de domaines limités. Les corpus comparables qui contiennent des textes relatifs à un domaine particulier en deux ou plusieurs langues sont alors utilisés. Cela rend l'alignement des traductions de termes encore plus difficile : il n'est plus possible de se baser sur les occurrences des termes dans les phrases parallèles et les traductions peuvent se trouver dans des parties différentes du corpus comparable.

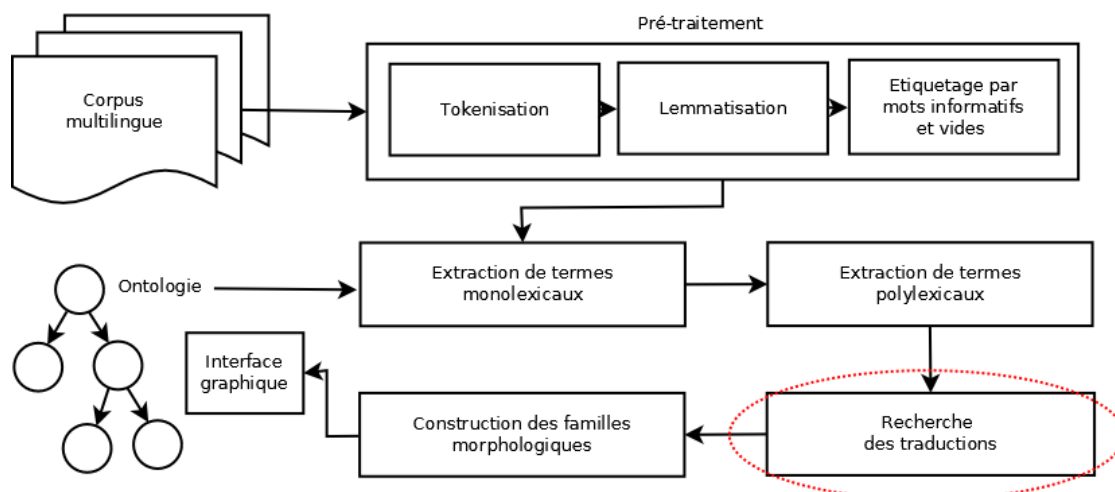


FIGURE 5.1 – Place de la recherche des traductions dans notre méthode

Dans notre système, l'extraction terminologique est suivie par la recherche de traductions (Figure 5.1) à la fois pour les termes connus par l'ontologie source et pour les termes candidats identifiés à partir du corpus.

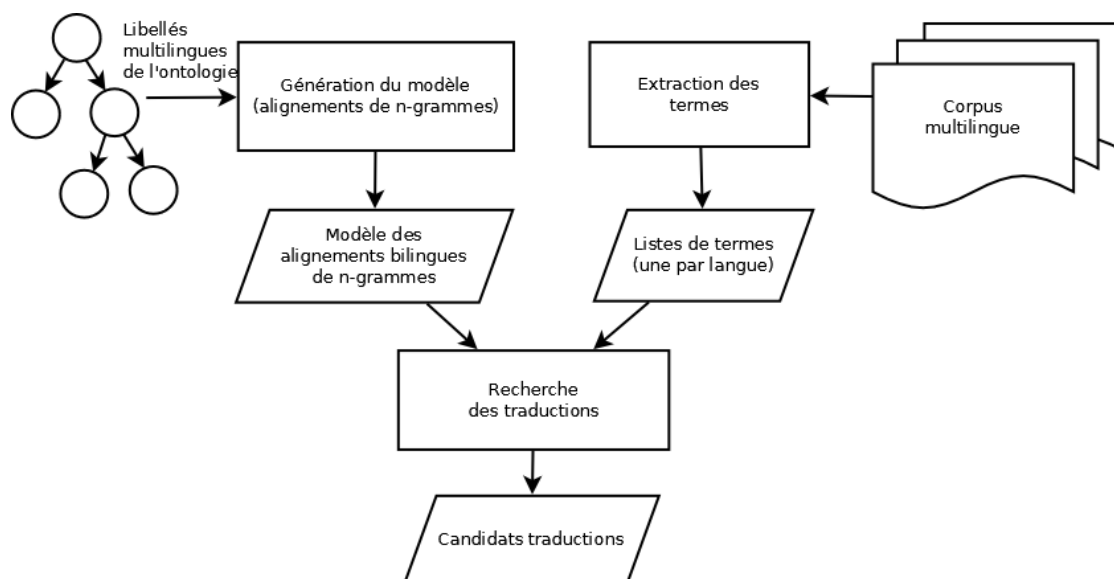


FIGURE 5.2 – Processus de recherche de traductions

Notre méthode (Figure 5.2) utilise une terminologie bilingue existante (par exemple, les termes déjà connus par l'ontologie source) pour générer des alignements au niveau sous-lexical, c'est-à-dire, au niveau des n-grammes de caractères. Ces alignements aident ensuite à aligner les termes extraits à partir d'un corpus spécialisé, indépendamment de leur type. Dans le contexte de la recherche des traductions, le modèle de traduction à base des n-grammes de caractères remplace les dictionnaires de morphèmes et, plus précisément, de formants néo-classiques, c'est-à-dire, des morphèmes d'étymologie grecque ou latine, appelés aussi des formants savants. À la différence de ces derniers, les n-grammes de caractères sont générés automatiquement à partir de données disponibles, indépendamment de la langue¹ et de la taille de ressource.

Notre objectif est de retrouver les candidats traductions pour les termes identifiés par le module d'extraction terminologique² décrit dans les chapitres précédents et pour les termes connus de l'ontologie d'évaluation. Toutefois, nous avons besoin d'évaluer la performance de cette étape indépendamment du reste du système. Pour cela, nous avons recours à deux types d'évaluation : la comparaison avec des méthodes d'identification des cognats, c'est-à-dire des mots graphiquement proches en deux langues, et avec le système TTC TermSuite (Cram et Daille, 2016; Daille *et al.*, 2011). Ces évaluations sont effectuées sur trois paires de langues FR-EN, FR-DE et DE-EN dans les deux directions. Les listes de termes se

1. À condition que les mots de cette langue puissent être décomposés en n-grammes de caractères.

2. Rappelons que les termes sont extraits à partir du corpus pour chaque langue séparément.

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

substituent aux listes de termes candidats, reproduisant ainsi le contexte du résultat de l'extraction terminologique à partir du corpus comparable.

Ce chapitre commence par la présentation détaillée de notre méthode. Les résultats obtenus sont discutés dans la section correspondante, suivis de la discussion sur le rôle des n-grammes de caractères dans l'alignement des traductions. Ces résultats permettent également de cerner les limites de notre approche.

5.1 Méthode de recherche de traductions à base de n-grammes de caractères

Notre méthode comprend deux étapes-clés : tout d'abord, nous construisons un modèle de traduction à base de n-grammes de caractères à partir de la ressource d'entraînement pour appliquer ensuite ce modèle à la recherche des traductions dans les listes de termes en langue source et cible.

5.1.1 Modèle de traductions à base de n-grammes de caractères

Les n-grammes de caractères représentent une ressource performante et accessible pour l'alignement des termes en plusieurs langues. Dans le contexte de la traduction automatique statistique, Tiedemann (2012) a exploité des uni et des bi-grammes de caractères pour améliorer les résultats, surtout pour les langues peu dotées. Nakov et Tiedemann (2012) ont utilisé les n-grammes pour la recherche des traductions entre les mots des langues proches à savoir le macédonien et le bulgare. Dans leur expériences, ils ont combiné les alignements au niveau lexical avec des n-grammes de caractères où $n \in [1 : 3]$. Parmi les avantages des n-grammes, les auteurs citent leur disponibilité, leur applicabilité à des langues peu dotées et l'absence de segmentation en morphèmes (Tiedemann, 2012; Nakov et Tiedemann, 2012).

Le système CLIR (Vilares *et al.*, 2007b,a, 2016) repose également sur l'alignement des mots en utilisant des n-grammes de caractères. Le modèle des n-grammes est généré à partir de corpus parallèle. Ce modèle est donc appliqué à l'ensemble de vocabulaire, et non aux termes d'un domaine. Les n-grammes sont alignés à l'aide de Giza++ (Och et Ney, 2003) et la probabilité d'alignement est complétée par des mesures d'association. Au départ, ce système fonctionnait pour la paire de langues EN-FR (Vilares *et al.*, 2007a), mais l'évaluation conduite sur un nombre important de paires de langues (Vilares *et al.*, 2016) permet de confirmer l'efficacité des n-grammes pour l'alignement multilingue.

Les n-grammes sont également utilisés pour traduire des mots rares dans les systèmes de traduction par des réseaux de neurones (Sennrich *et al.*, 2016; Luong et Manning, 2016). Sennrich *et al.* (2016) utilisent des uni-grammes et des bi-grammes, combinés avec des dictionnaires lexicaux, pour améliorer les résultats de traduction. Leurs expériences ont montré une amélioration dans toutes les paires de langues par rapport aux modèles sans n-grammes, mais la performance varie en fonction des langues (la paire anglais-allemand

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

fournit des meilleurs résultats que celle anglais-russe). Toutefois, l'expérience avec le russe confirme la capacité des n-grammes à effectuer la translittération. Luong et Manning (2016) ont mené des expériences semblables sur la paire de langue anglais-tchèque et ont également démontré l'amélioration de la qualité de traduction, surtout en ce qui concerne les mots rares.

Compte tenu des avantages que les n-grammes de caractères présentent pour la traduction automatique, nous allons les exploiter pour identifier les traductions candidates pour les termes extraits à partir du corpus. Dans notre cas, il s'agit d'un corpus comparable, il n'est donc pas possible d'aligner les phrases entre-elles. Nous avons contourné ce problème en utilisant des alignements multilingues présents dans notre ontologie à élargir. Ainsi, nous exploitons ces n-grammes comme une ressource morphologique endogène.

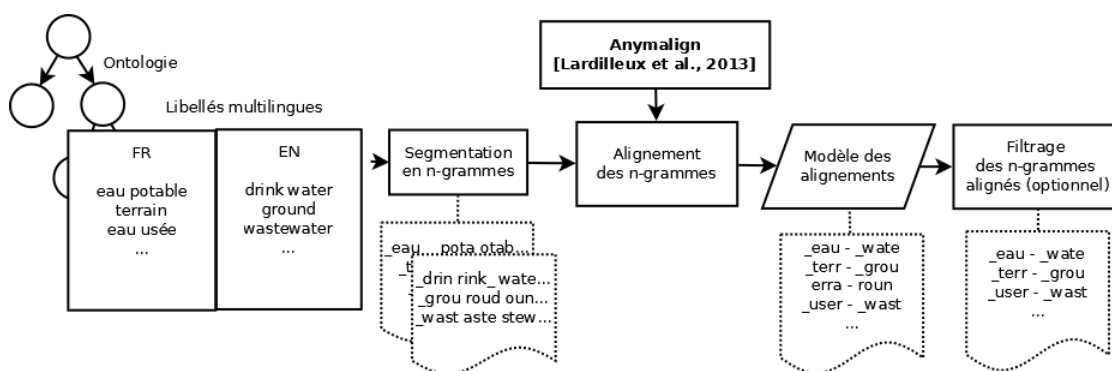


FIGURE 5.3 – Apprentissage du modèle de traduction

Notre méthode est basée sur les n-grammes de caractères obtenus à partir d'une ressource terminologique multilingue, par exemple une ontologie spécialisée ou un thésaurus électronique. Pour générer le modèle (Figure 5.3), on construit un corpus parallèle de la terminologie présente dans la ressource (tous les termes dans la langue source avec leur traductions en langue cible). Par exemple, si le terme français *draperies aurorales* a deux traductions en anglais *auroral draperies* et *auroral curtains*, on obtient deux alignements (FR-EN) :

- *draperies aurorales* - *auroral draperies* ;
- *draperies aurorales* - *auroral curtains*.

5.1.1.1 Extraction des n-grammes de caractères

Nous considérons chaque terme comme un ensemble de n-grammes de caractères au niveau sous-lexical. Pour construire le modèle, les termes de la ressource sont segmentés en n-grammes de la longueur n , où la valeur de n peut être adaptée à une langue en particulier, si nécessaire.

Nous avons testé la méthode avec $n \in [1 : 10]$ sur un jeu de développement et les meilleurs résultats ont été obtenus pour $3 \leq n \leq 5$, ce qui est en accord avec la démarche de Bojanowski *et al.* (2017) qui utilise $3 \leq n \leq 6$. Nous avons décidé de baser

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

notre évaluation sur $n = 4$ et donner plus de détails sur l'influence de n dans la discussion des résultats.

Nous avons ainsi obtenu deux fichiers parallèles pour chaque paire de langues (FR-DE, FR-EN, EN-DE) dans les deux directions. Afin de différencier les n-grammes de caractères au début (des préfixes potentiels) et à la fin (des suffixes ou des terminaisons) de mots, nous entourons chaque mot par les marqueurs " _ " (par exemple *conformation anatomique* devient *_ conformation _ _ anatomique _*).

5.1.1.2 Alignement des n-grammes de caractères

Dans le domaine de la traduction automatique, la création des modèles de traduction utilise des corpus bilingues parallèles où les textes sont alignés avec leurs traductions au niveau des phrases. Il existe des outils qui identifient des traductions potentielles, des mots en langue source et cible qui apparaissent dans un grand nombre de segments alignés. Il s'agit des aligneurs statistiques, dont Anymalign³ (Lardilleux et Lepage, 2009a,b) et Giza++ (Och et Ney, 2003). Ces outils calculent des probabilités qu'un mot en langue cible soit la traduction d'un mot en langue source et vice versa.



FIGURE 5.4 – Alignement mot-à-mot pour la paire de langues anglais - français (Navlea, 2014)

La Figure 5.4 illustre un exemple d'alignement mot-à-mot des deux phrases parallèles : chaque mot a un alignement correspondant. Toutefois, dans certains cas, un mot en langue source peut correspondre à plusieurs mots en langue cible (par exemple, *He often eats potatoes. – Il mange souvent des pommes de terre.*).

Le calcul de probabilité d'alignement se fait dans deux directions : source-cible et cible-source. Prenons l'exemple d'un corpus parallèle contenant quelques phrases en anglais et en français. Si dans l'ensemble de phrases en anglais qui contiennent le mot *often*, il est toujours aligné avec le mot français *souvent* dans la direction de langues EN-FR, la probabilité d'alignement de ces deux mots dans la direction source-cible est égale à 100 % (1,00). En même temps, si une des phrases en français contient le mot *souvent* qui est aligné avec un autre mot correspondant en anglais (par exemple, *frequently*), la probabilité dans la direction cible-source sera inférieure à 100 %. Plus les deux probabilités sont hautes, plus l'alignement est précis.

3. <https://anymalign.limsi.fr/>

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

Dans notre projet, nous avons utilisé l'outil Anymalign pour aligner les n-grammes de caractères un-à-un. Cet outil a été choisi, car, évalué sur des alignements lexicaux, il atteint le niveau d'exactitude de Giza++ (Och et Ney, 2003) en le dépassant dans la tâche d'alignements un-à-un (Lardilleux *et al.*, 2013). Un aligneur statistique est destiné à aligner des éléments de deux ou plusieurs séquences parallèles : dans le cas d'alignements au niveau lexical, les séquences sont les phrases en langue source et cible et les éléments – leurs mots. Nous avons effectué les alignements au niveau sous-lexical où les séquences sont les termes et leurs traductions segmentés en n-grammes de caractères et les éléments alignés – ces n-grammes de caractères.

Anymalign est un outil multilingue par excellence, car non seulement il traite les unités lexicales sans aucune corrélation avec la langue, mais aussi permet un alignement simultané en plusieurs langues. Par exemple, il est possible d'aligner des mots en anglais, en français et en allemand en une seule commande. Une autre particularité d'Anymalign consiste en la séparation du corpus en plusieurs sous-corpus pour chaque itération. Les alignements sont alors calculés au sein de ces sous-corpus et fusionnés avec les itérations précédentes. L'alignement s'effectue en cinq étapes. D'abord, l'outil regroupe des mots de chaque langue. Puis, il constitue les groupes de mots qui apparaissent sur les mêmes lignes (des alignements parfaits). Ensuite, pour chaque groupe, des séquences de mots et des contextes sont extraits à partir des lignes où ces mots sont rencontrés. Ces étapes sont répétées tant que l'utilisateur n'arrête pas le processus. Les auteurs conseillent de maintenir l'exécution tant que le système retrouve de nouveaux alignements (Lardilleux et Lepage, 2009a). Afin d'assurer la reproductibilité des résultats, nous avons lancé l'outil pour 60 minutes pour chaque paire de langues.

Par défaut, les alignements un-à-plusieurs sont activés (un mot de la langue source peut être aligné avec plusieurs mots de la langue cible). Pour éviter cela, nous avons désactivé cette option en autorisant ainsi uniquement les alignements un-à-un. Par conséquent, un quadri-gramme de caractères peut être aligné à un autre (par exemple, *foto* à *phot*), mais pas à un quadri-gramme suivi d'un autre quadri-gramme (comme *foto* à *phot+hoto*).

Source	Cible	Score source-cible	Score cible-source	Fréquence
_biol	_biol	0,87	0,98	8 491
_dese	_wüst	1,00	1,00	8 476
_bios	_bios	1,00	1,00	8 430
iner	iner	0,90	0,86	8 393
rope_	ropa_	1,00	1,00	8 378
rice_	reis_	0,97	0,99	8 304
smic_	smis	1,00	1,00	8 146
tech	otec	0,99	0,99	7 670

TABLE 5.1 – Sortie d'Anymalign (EN-DE) à partir de GEMET

Dans l'échantillon de la sortie d'Anymalign (Table 5.1), le n-gramme source est suivi par sa traduction potentielle en langue cible et par les scores de probabilité dans les deux directions : source-cible et cible-source. Par exemple, les scores de l'alignement *_dese* – *_wüst* indiquent que ces deux n-grammes ont été alignés dans la totalité des échantillons,

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

tandis que dans la paire *_biol - _biol*, le n-gramme anglais a été aligné dans 13 % (100 %-87 %) de cas avec d'autres n-grammes en allemand, tandis que le n-gramme allemand a été presque toujours (98 % de cas) aligné avec cet n-gramme français. La dernière valeur représente la fréquence de l'alignement en question dans les échantillons de segments alignés.

Nous transformons ce fichier de résultats en dictionnaire où les n-grammes source sont associés avec leur traductions et le score de traduction (la moyenne des scores de probabilité dans les deux directions). Par exemple, le score pour la traduction allemande *_wüst* pour le n-gram anglais *_dese* est égal à 1 $((1+1)/2)$, tandis que l'alignement *rice_* et *reis_* a le score de 0,98 résultant de $(0,97+0,99)/2$.

5.1.1.3 Filtrage des n-grammes de caractères alignés

Dans l'objectif de garder uniquement les meilleurs alignements des n-grammes, nous avons développé des filtres pour éliminer ceux qui ne satisfont pas une des conditions suivantes :

- si le n-gramme source contient le marqueur du début ou de la fin, sa traduction doit aussi contenir ce marqueur ;
- la probabilité moyenne est inférieure au seuil *avrg_{min}* fixé par l'utilisateur (par exemple, 0,5 correspondant à 50 % de probabilité qu'un alignement soit correct dans les deux directions, source-cible et cible-source) ;
- aucune des deux scores de probabilité n'atteint le seuil *score_{keep}* également fixé par l'utilisateur (par exemple, il peut être utile de garder un alignement qui a une probabilité de 90 % d'être correct pour une direction, donc un des scores égal à 0,9).

L'application du seuil *avrg_{min}* permet de garder uniquement les alignements dont la moyenne des scores calculés par Anymalign est supérieure à la valeur fixée par l'utilisateur : $avrg_{min} < \frac{score_{source_cible} + score_{cible_source}}{2}$. Le *score_{keep}* permet de garder l'alignement si au moins un des deux scores (source-cible ou l'inverse) est supérieur à la valeur souhaitée. Il est en effet intéressant d'exclure les alignements peu probables à l'aide du filtre *avrg_{min}* et de garder ceux qui ont au moins un score important (via le filtre *score_{keep}*).

Source	Cible	Score source-cible	Score cible-source	Fréquence
<i>_dese</i>	<i>_wüst</i>	1,00	1,00	8 476
<i>_bios</i>	<i>_bios</i>	1,00	1,00	8 430
<i>rope_</i>	<i>ropa_</i>	1,00	1,00	8 378
<i>smic_</i>	<i>smis</i>	1,00	1,00	8 146
<i>tech</i>	<i>otec</i>	0,99	0,99	7 670
<i>meri</i>	<i>meri</i>	1,00	1,00	7 582
<i>_ecot</i>	<i>_ökot</i>	1,00	1,00	7 385
<i>ogeo</i>	<i>ogeo</i>	1,00	1,00	7 006
<i>_hunt</i>	<i>_jagd</i>	1,00	1,00	6 938

TABLE 5.2 – N-grammes après l'application de filtres (EN-DE, GEMET)

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

Les filtres gardent les meilleurs alignements (Table 5.2) et réduisent la taille du modèle (de 32 217 à 2 237 alignements dans le cas de GEMET, EN-DE).

Afin de choisir les meilleures valeurs de ces filtres, nous avons créé un jeu de données de développement à partir de la ressource Agrovoc. Il compte 6 000 paires de traductions séparées en deux parties : 1 000 paires pour générer le modèle et 5 000 paires pour évaluer les résultats.

Filtre	Meilleure valeur	Rappel	Précision	F-mesure
Marques de position	-	+0,34	+0,36	+0,38
$score_{keep}$	0,99	+1,29	+2,33	+1,66
avr_{gmin}	0,75	+1,36	+2,49	+1,76

TABLE 5.3 – Impact des filtres (jeu de développement, $n = 4$)

La Table 5.3 présente les améliorations apportées par les filtres associés avec leur meilleure valeur (en points). Nous avons testé les valeurs entre 0 et 1. Les améliorations sont moins importantes que nous avons prévu, mais nous avons décidé de garder ces filtres afin rendre le système paramétrable.

5.1.2 Alignement de la terminologie bilingue

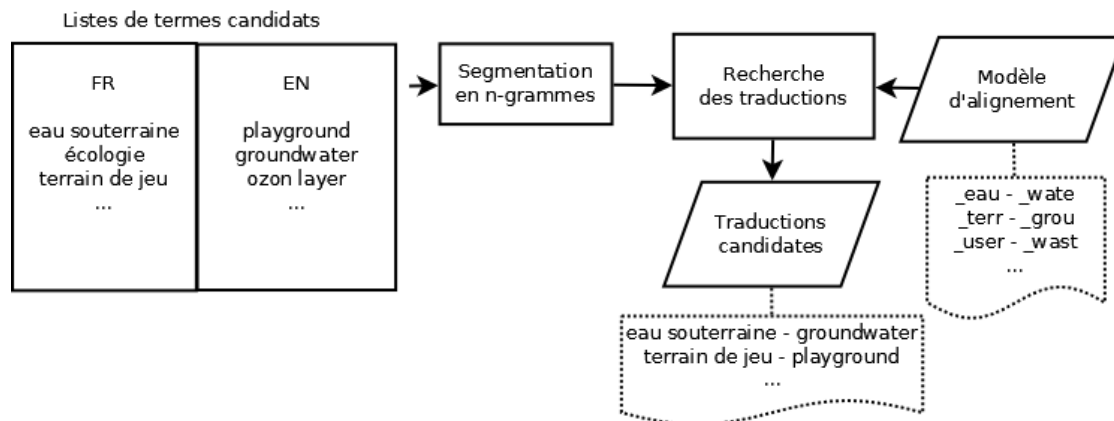


FIGURE 5.5 – Identification des traductions

Après la génération du modèle de traduction, il est appliqué à la recherche des traductions (Figure 5.5) entre les termes en langue source et cible. Le processus comprend plusieurs étapes. D’abord, les termes candidats sont segmentés en n-grammes. Ensuite, les traductions pour ces n-grammes sont recherchées dans le modèle, ce qui permet d’identifier les candidats traduction contenant ces n-grammes cibles. Enfin, les résultats sont filtrés pour garder les meilleurs candidats.

5.1. MÉTHODE DE RECHERCHE DE TRADUCTIONS À BASE DE N-GRAMMES DE CARACTÈRES

5.1.2.1 Traduction basée sur les n-grammes de caractères

Afin de trouver les traductions, chaque terme source est segmenté en n-grammes de caractères. Pour chaque n-gramme, nous recherchons ses traductions dans le modèle et vérifions également si ce n-gramme existe dans la liste des n-grammes cibles. De cette manière, nous prenons en compte des n-grammes identiques entre les deux langues. Cela permet d'identifier les cognats même si les n-grammes correspondants ne sont pas dans le modèle de traduction.

Source	Position du n-gramme source	
	Début/milieu	Fin
Modèle de traduction	$\frac{t_s}{pos_a}$	$\frac{t_s}{pos_b}$
N-gramme identique	$\frac{avg_s}{pos_a}$	$\frac{avg_s}{pos_b}$

TABLE 5.4 – Éléments du score de traduction

où t_s est le score de traduction, pos_a et pos_b sont des constantes définies en fonction de la position du n-gramme, avg_s est le score moyen dans le modèle de traduction. En effet, dans les cas où les n-grammes identiques n'ont pas de score dans le modèle, il est remplacé par la moyenne afin de les intégrer dans le calcul sans les favoriser ni les pénaliser.

Le score de traduction d'un n-gramme t_s est égal à la moyenne des scores d'un alignement dans les deux directions divisée par le seuil correspondant à la position du n-gramme en question (Table 5.4). En effet, les n-grammes situés à la fin de mots sont généralement bien alignés par Anymalign, leurs scores sont donc importants. Nous avons besoin de réduire leur influence sur le choix des traductions en divisant leurs scores par un nombre entier pos_b .

En fonction de la langue, les n-grammes importants peuvent se situer au début ou au milieu des mots (comme c'est le cas du français, de l'anglais et de l'allemand) où à la fin. Les valeurs des paramètres peuvent être revues en conséquence. Nous avons obtenu les meilleurs résultats avec les valeurs $pos_a = 1$ et $pos_b = 5$.

Le score associé à chaque traduction potentielle résulte de l'addition de tous les scores des n-grammes. Par exemple, le terme anglais *water-soluble* est traduit par le terme allemand *wasserlöslich*⁴ de manière suivante :

- les limites des mots sont marquées par _ (*_water-soluble_*, *_wasserlöslich_*) ;
- les mots sources et cibles sont segmentés en n-grammes (par exemple, si $n = 4$, *_wate*, *wate*, *ater*, etc.) ;
- les alignements du modèle (comme *_wate-lwas*, *_wate-_wass*, *olub-ösli*) et les n-grammes identiques entre les mots des deux langues (*ater*, *ter*-) servent à identifier les mots cibles qui contiennent les n-grammes recherchés ;
- les scores sont calculés pour chaque traduction candidate :
 - par exemple, le score de la traduction *_wasser* est égal à la somme des scores n-grammes traduits, donc *_wate* - *_wass* (avec le score 0,66 correspondant à

4. Modèle à partir de GEMET.

5.2. ÉVALUATION

- la moyenne des scores d'Anymalign dans les deux directions multipliée par la constante de position $pos_a = 1$, c'est-à-dire, $\frac{0,37+0,94}{2} * 1 = 0,66$;
- les traductions *_wasserunlöslich_* (insoluble par l'eau) et *_wasserlöslich_* (soluble par l'eau) ont le score égal à la somme des scores de *_wate_* - *_wass_* (0,66) et de *olub-ösli* (0,73), donc $0,66 + 0,73 = 1,39$, etc.

5.1.2.2 Filtrage des traductions

L'étape précédente donne plusieurs traductions potentielles pour chaque terme source. Pour cette raison, nous avons développé un système de filtres complémentaires qui s'appliquent les uns après les autres en gardant uniquement les meilleurs candidats à chaque étape. Ces filtres sont destinés à augmenter la précision des résultats.

Filtrage par le score de traduction Le premier filtre garde tous les candidats ayant le score de traduction maximal.

Filtrage par le nombre de n-grammes traduits La deuxième étape prend les candidats sélectionnés par le premier filtre et garde tous les candidats ayant le nombre des n-grammes traduits le plus important. Ce filtre appliqué sur le jeu de développement avec $n = 4$, baisse le rappel moyen de 1,09 points (31,46 % sans filtre contre 30,37 % avec le filtre), mais augmente la précision moyenne de 5,99 points (de 36,14 % à 41,41 %) et la F-mesure moyenne de 1,49 points (de 33,63 % à 35,12 %).

Filtrage par le ratio de longueur

$$ratio_longueur = \frac{score}{1 + |longueur_{source} - longueur_{cible}|} \quad (5.1)$$

Nous avons également expérimenté un filtre qui sélectionne les traductions potentielles les plus proches du terme source en matière de longueur. Le ratio de longueur est calculé selon la Formule 5.1. Ce filtre est utile dans les cas où un terme source, par exemple *water* (EN), a deux traductions potentielles, comme *Wasser* et *Wasserpotential* (DE), permettant d'en garder la plus courte.

En reprenant notre exemple d'avant, ce filtre permet de garder la traduction *_wasserlöslich_* (soluble par l'eau) pour le terme *_water-soluble_*. Le score final du candidat *_wasserlöslich_* = $\frac{1,39}{1+|15-15|} = 1,39$, tandis que le candidat *_wasserunlöslich_* a le score moindre ($\frac{1,39}{1+|15-17|} = 0,46$).

Toutefois, nous avons constaté que ce filtre baisse le rappel moyen (30,36 %), la précision moyenne (41,81 %) et la F-mesure moyenne (35,12 %) d'environ 5 points.

5.2 Évaluation

Nous avons choisi deux types d'évaluation pour cette étape de traitement : premièrement, la comparaison avec des méthodes de recherche des cognats (Aker *et al.*, 2013) et,

5.2. ÉVALUATION

deuxièmement, avec l'outil TTC TermSuite (Cram et Daille, 2016; Daille *et al.*, 2011).

5.2.1 Comparaison avec des méthodes d'identification des cognats

Étant donné que notre méthode est basée sur les n-grammes de caractères, elle permet dans certain cas d'identifier des traductions similaires du point de vue graphique appelés *cognats*. Navlea (2014) définit les cognats comme « des mots bilingues similaires au niveau orthographique et/ou phonétique, susceptibles de partager un sens commun ». Par exemple, le mot anglais *plate* est proche du mot français *plateau*, ou encore le mot français *concentration* est similaire au mot allemand *konzentration*.

Nous comparons notre méthode avec les méthodes d'identification de cognats suggérées par Aker *et al.* (2013)⁵

- le ratio de sous-chaîne de caractères commune la plus longue (*Longest Common Substring Ratio*);
- le ratio de sous-séquence commune la plus longue (*Longest Common Subsequence Ratio*);
- la similarité de Dice;
- et la distance normalisée de Levenshtein.

Sous-chaîne de caractères commune la plus longue⁶ Comme son nom l'indique, ce ratio permet d'identifier les mots en langue cible qui partagent une sous-chaîne de caractères avec le terme en langue source. Le calcul de ce ratio se fait à l'aide de la formule suivante :

$$LCSTR(X, Y) = \frac{\text{len}[LCST(X, Y)]}{\max[\text{len}(X), \text{len}(Y)]} \quad (5.2)$$

où *LCST* est la sous-chaîne de caractères commune la plus longue, *X* et *Y* sont des mots en langue source et cible.

Par exemple, dans le couple de termes *concentration* (FR) et *konzentration* (DE), il s'agit de la sous-chaîne *entration* et la $LCSTR(\text{concentration}, \text{konzentration}) = \frac{9}{13} = 0,69$. Toutefois, cette méthode ne permet pas d'identifier plus de similarité entre ces termes, à savoir les deux caractères *on* également communs.

Ratio de sous-séquence commune la plus longue⁷ Ce ratio dépasse l'étendue de la sous-chaîne de caractères la plus longue, car il permet d'identifier la sous-séquence la plus longue, même discontinue.

$$LCSR(X, Y) = \frac{\text{len}[LCS(X, Y)]}{\max[\text{len}(X), \text{len}(Y)]} \quad (5.3)$$

5. Les formules ci-dessous sont fournies par Aker *et al.* (2013).

6. http://www.bogotobogo.com/python/python_longest_common_substring_lcs_algorithm_generalized_suffix_tree.php

7. https://rosettacode.org/wiki/Longest_common_subsequence#Python

5.2. ÉVALUATION

où LCS est la sous-séquence commune la plus longue, X et Y sont des mots en langue source et cible.

Ce ratio se calcule selon la formule 5.3. Dans le cas de notre couple *concentration* (FR) et *konzentration* (DE), il permet d'identifier la sous-séquence partagée la plus longue *on*entrution* : $LCSR(concentration, konzentration) = \frac{11}{13} = 0,85$.

Similarité de Dice Cette mesure permet également de mesurer la similarité entre deux ensemble, à savoir des chaînes de caractères dans notre cas.

$$dice(X, Y) = \frac{2 * LCST(X, Y)}{len(X) + len(Y)} \quad (5.4)$$

où $LCST$ est la sous-chaîne de caractères commune la plus longue, X et Y sont des mots en langue source et cible.

Elle est calculée selon la Formule 5.4.

Distance normalisée de Levenshtein⁸ L'algorithme calcule le nombre d'opérations singulières (suppression, insertion ou remplacement) nécessaire pour transformer la chaîne de caractères source en celle cible. Deux chaînes identiques ont un score égal à zéro (et ne sont pas prises en compte dans notre évaluation). En reprenant notre exemple, la distance entre les termes *concentration* (FR) et *konzentration* (DE) est égale à deux, car il y a une différence de deux caractères. L'avantage de la distance de Levenshtein est sa capacité à identifier les chaînes similaires sans aucun entraînement, ce qui la rend intéressante pour la recherche de traductions dans les termes extraits à partir de corpus comparables.

$$DL_{norm}(X, Y) = 1 - \frac{DL(X, Y)}{\max[len(X), len(Y)]} \quad (5.5)$$

où DL est la distance de Levenshtein, X et Y sont des mots en langue source et cible (Aker *et al.*, 2013).

La normalisation s'effectue à l'aide de la Formule 5.5 et permet d'obtenir les valeurs dans l'intervalle entre 0 et 1.

Ces méthodes sont sélectionnées en tant que baseline, car elles sont capables d'identifier des chaînes de caractères similaires sans entraînement préalable.

5.2.1.1 Paramètres de la baseline

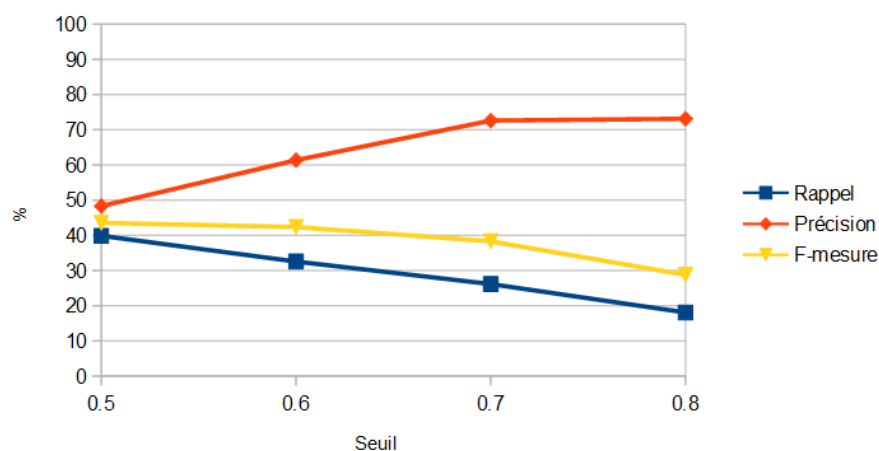
La baseline est appliquée comme suit : pour chaque terme en langue source, nous parcourons la liste des termes en langue cible et calculons les mesures énumérées ci-dessus.

Ensuite, les traductions candidates qui ont au moins un des scores énumérés précédemment supérieur ou égal à 0,7 (en suivant (Aker *et al.*, 2013)), sont sélectionnées et gardées pour l'évaluation.

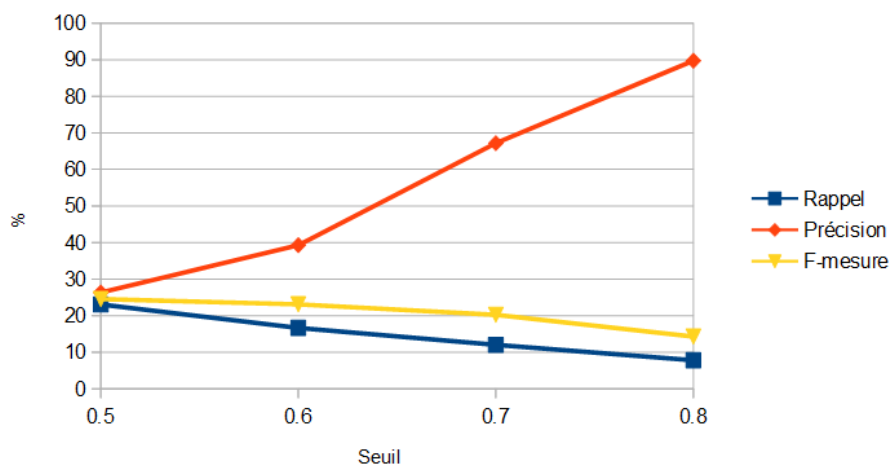
8. <http://hetland.org/coding/python/>

5.2. ÉVALUATION

Valeur du seuil Nous avons testé les valeurs du seuil pour vérifier si la valeur proposée par Aker *et al.* (2013) est adaptée à nos données. La Figure 5.6 montre la sensibilité de la baseline au seuil appliqué. Plus le seuil est important, plus les résultats sont précis. En même temps, l'augmentation de seuil réduit le nombre de traductions candidates, ce qui se traduit par une baisse de rappel. Il faut souligner qu'en absence de seuil, les méthodes d'identification de cognats considèrent chaque terme en langue cible en tant que traduction potentielle et le rappel est donc de 100 %. Compte tenu de cette particularité, le seuil de 0,7 nous semble adapté, car il limite le nombre de traductions suggérées pour chaque terme.



(a) EcoLexicon



(b) GEMET

FIGURE 5.6 – Évaluation du seuil pour la baseline (valeurs moyennes par rapport à toutes les paires de langues dans les deux directions)

Dans l'évaluation, Aker *et al.* (2013) prend en compte des traductions partielles (par exemple, *visible graphical interface* — *grafische Benutzerschnittstelle*). Étant donné que notre évaluation s'effectue sur des ressources terminologiques, ce qui facilite notre tâche,

5.2. ÉVALUATION

nous considérons dans cette évaluation que seules les traductions complètes présentes dans les ressources sont correctes.

5.2.1.2 Ressources

L'évaluation a été menée sur deux ressources : GEMET et EcoLexicon. Chaque ressource a servi à la construction d'un modèle de traduction utilisé pour trouver les traductions dans les listes de termes de l'autre ressource.

Nous avons utilisé GEMET et EcoLexicon pour prouver que la méthode peut fonctionner même avec un nombre de traductions limité dans la ressource, comme dans le cas des paires FR-DE et DE-FR dans EcoLexicon. L'intersection moyenne entre GEMET et EcoLexicon pour toutes les paires de langues est de 3,61 %.

L'évaluation est effectuée en comparant les traductions identifiées par la méthode à partir des listes de termes en langue source et cible, avec celles présentes dans la ressource d'évaluation. On exclut de l'évaluation les termes qui sont traduits par la même chaîne de caractères (par exemple, *production* en français et *production* en anglais). Dans ces cas, la distance de Levenshtein est égale à zéro.

5.2.1.3 Résultats

Nous avons évalué notre méthode sur l'exemple de deux configurations de ressources présentant le cas où le nombre de termes à traduire est faible et le modèle est généré à partir d'une ressource riche (comme GEMET) et la situation inverse où le nombre de termes est important, tandis que le modèle est construit à partir de peu de paires de traductions (le cas d'EcoLexicon). L'évaluation prend en compte les résultats de référence (la baseline), les résultats de notre méthodes et la combinaison des deux méthodes (lorsque les listes des traductions candidates obtenues par les deux méthodes sont fusionnées). En effet, les deux méthodes identifient des candidats différents sans ressources extérieures et sans dépendance de la paire de langues, ce qui permet d'envisager de les utiliser ensemble.

Lang.	Termes	Inter. ⁹ %	Rappel			Précision			F-mesure		
			BL	NG	Mix	BL	NG	Mix	BL	NG	Mix
FR-EN	570	36,77	29,12	40,70	51,05	62,88	83,15	69,12	39,81	54,65	58,73
FR-DE	460	45,45	26,96	32,17	40,65	82,12	74,75	74,21	40,59	44,98	52,53
EN-DE	3 833	46,20	23,72	28,57	35,69	71,57	55,03	58,36	35,63	37,61	44,29
EN-FR	648	35,18	25,46	38,58	47,38	61,34	89,93	72,24	35,99	54,00	57,22
DE-FR	473	43,48	25,58	30,23	38,90	81,76	89,94	84,79	38,97	45,25	53,33
DE-EN	3 390	47,32	26,08	31,56	39,03	75,88	68,24	68,87	38,81	43,16	49,82

TABLE 5.5 – Résultats obtenus sur EcoLexicon, modèle entraîné sur GEMET

9. Inter. signifie intersection des traductions correctes identifiées par les deux méthodes, BL – baseline, NG – n-grammes; les colonnes Mix ont été calculées à partir d'une liste de traductions candidats qui fusionne les résultats des deux méthodes.

5.2. ÉVALUATION

Lang.	Termes	Inter. %	Rappel			Précision			F-mesure		
			BL	NG	Mix	BL	NG	Mix	BL	NG	Mix
FR-EN	4 864	17,23	12,52	42,74	47,14	63,84	58,43	60,61	20,94	49,37	53,04
FR-DE	4 898	22,06	7,78	18,54	21,56	63,18	31,56	34,26	13,85	23,36	26,47
EN-DE	4 935	41,59	15,74	20,95	25,92	70,25	36,28	39,59	34,54	26,08	31,33
EN-FR	4 880	18,36	12,46	43,81	47,54	58,86	63,76	65,13	20,56	51,94	54,96
DE-FR	4 898	20,95	7,82	20,70	23,58	67,55	52,27	54,40	14,02	29,66	32,90
DE-EN	4 918	36,73	15,72	22,73	28,12	79,36	56,01	61,99	26,24	32,34	38,69

TABLE 5.6 – Résultats obtenus sur GEMET, modèle entraîné sur EcoLexicon

Dans les deux cas ((Tables 5.5 et 5.6), notre méthode (les colonnes NG) a un meilleur rappel et une meilleure F-mesure que la baseline (BL) : 33,64 % et 28,24 % de rappel moyen obtenu par les n-grammes contre 26,15 % et 12,01 % par la baseline ; 46,61 % et 36,87 % de F-mesure moyenne contre 38,30 % et 21,69 % respectivement. Cet écart s'explique par la capacité de notre approche d'identifier les traductions qui ne sont pas assez similaires pour être traitées en tant que cognats.

En même temps, l'intersection moyenne entre les traductions retrouvées par les deux méthodes est de 42,40 % et 26,15 % pour les deux jeux de données. Elle contient essentiellement des cognats, par exemple, le terme anglais *stochastic hydrology* et sa traduction allemande *stochastische hydrologie*.

En termes de précision, la baseline est très performante 72,59 % et 68,01 %. Les n-grammes offrent une meilleure précision (76,84 % en moyenne) lorsque le modèle est généré à partir de GEMET (donc, une ressource plus riche et équilibrée), mais semblent avoir plus de difficultés avec le modèle généré à partir d'EcoLexicon (49,72 %).

Dans le cas où la ressource disponible contient peu de traductions, les n-grammes offrent toujours un meilleur rappel par rapport à la baseline, mais leur précision est considérablement plus basse (49,72 % en moyenne contre 68,01 % sur le même jeu de données). Nous pouvons également constater que les deux méthodes fournissent de meilleurs résultats lorsque la quantité de termes à traduire et, donc, le nombre de traductions possibles, est faible.

Enfin, il est possible de combiner notre méthode avec la baseline : les colonnes Mix dans les Tables 5.5 and 5.6 montrent que la fusion des deux listes de candidats obtenues séparément par chaque méthode, permet d'atteindre un meilleur rappel et F-mesure. Cependant, le temps d'exécution varie considérablement entre quelques minutes pour notre approche et plusieurs heures pour la baseline. En effet, étant donné que la baseline compare des chaînes de caractères d'une liste avec celles d'une autre, son temps d'exécution augmente linéairement avec l'augmentation du jeu de données. Notre méthode, au contraire, limite la recherche de candidats cible à des mots contenant les traductions de n-grammes contenus dans le modèle ou partageant un ou plusieurs n-grammes avec le mot source.

Les méthodes d'alignement statistique des cognats n'utilisent pas de ressources complémentaires et prennent en compte uniquement la similarité graphique des mots. De ce fait, elles sont incapables d'identifier les traductions dépourvues de toute ressemblance. Cette faiblesse peut leur être reprochée. Pour cette raison, nous voulons comparer notre méthode à un outil de référence qui procède à des traitements plus évolués.

5.2. ÉVALUATION

5.2.2 Comparaison avec TTC TermSuite

Nous avons également comparé les résultats d'alignement fournis par TTC TermSuite¹⁰. En effet, fruit d'un projet de recherche européen, TTC TermSuite est une forte référence dans l'extraction terminologique multilingue.

TTC TermSuite suit le même chemin : d'abord, les termes candidats sont extraits pour chaque langue et, ensuite, le système identifie les candidats traductions. Dans les traitements, le système se base sur l'étiquetage morphosyntaxique, les calculs statistiques (notamment, la distribution des occurrences et les contextes), les listes de formants classiques et les dictionnaires de taille importante (entre 100 000 et 1 000 000 paires de traductions en fonction de paire de langues).

Si un terme peut être traduit à l'aide des dictionnaires bilingues du système, il s'agit de la traduction compositionnelle. Dans certains cas, c'est une partie de terme qui est traduite à l'aide d'un dictionnaire. Cette approche est appelée semi-distributionnelle. Enfin, si aucune traduction n'est disponible dans les dictionnaires, TTC TermSuite calcule la probabilité distributionnelle pour proposer des candidats. Plus précisément, le système construit des vecteurs de contexte (les mots cooccurrent avec les candidats dans le corpus) pour les candidats en langues source et cible. Ensuite, il utilise les dictionnaires pour identifier les traductions entre les mots faisant partie des vecteurs en deux langues et retient les traductions candidates dont les vecteurs contiennent le plus de traductions identifiées. La probabilité distributionnelle qu'il s'agit de traductions potentielles accroît avec le nombre de traductions identifiées dans les vecteurs.

Il faut remarquer que la précision des résultats diminue lorsque les termes ne sont pas présents dans les dictionnaires.

#	Pilot	Pattern	Spec.	Fr.	D-Freq	Spotting rule	an
1	contaminated soil	A N	4,37	24	4	Pattern	A N
2	extractant	N	4,31	21	1	Pilot	contamini
3	aqueous	A	4,29	20	4	Frequency	24
4	preceding claims	A N	4,22	17	2	Specificity	4.3687024
5	micro-organisms	N	4,16	15	4	Document Frequency	4
6	sediments	N	4,16	15	4		
7	fructan	N	4,16	15	1		
8	groundwater	N	4,13	14	3		
9	decomposition	N	4,13	14	4		
10	contaminated resin particles	A N N	4,13	14	1		
11	oxidizer	N	4,13	14	1		
12	ions	N	4,13	14	4		
13	inorganic	A	4,10	13	4		
14	rotary	A	4,10	13	3		
15	treatment stage	N N	4,10	13	2		
16	contaminated area	A N	4,07	12	1		
17	ground water	N N	4,07	12	3		

#	Term	Score	Method
1	gesättigte bodenzone	5,6	semi-dist
2	temperatur im bereich	5,4	semi-dist
3	wassers in den boden	5,4	semi-dist
4	gereinigtem boden	5,4	semi-dist
5	beseitigung von verunr...	5,1	semi-dist
6	beseitigung von untergr...	5,1	semi-dist
7	grundwassers in der ges...	5,1	semi-dist
8	bodenbestandteilen	5,1	semi-dist
9	derartigen bereich	5,1	semi-dist

FIGURE 5.7 – Exemple de traductions EN-DE proposées par TermSuite pour le candidat *contaminated soil* (corpus B09C)

10. Version fr.univnantes.termsuite.ui_2.3.1.201610061343 du 17 novembre 2016.

5.2. ÉVALUATION

Compte tenu des différences dans l'identification de termes entre TermSuite et notre système, nous avons basé cette évaluation sur les Top-50 candidats identifiés par TermSuite dans l'ordre décroissant du score de spécificité (Figure 5.7). Le score de spécificité est attribué aux candidats lors de leur extraction à partir du corpus. Il reflète l'appartenance du candidats en question au corpus d'un domaine par rapport au corpus de langue générale¹¹. L'interface de TermSuite permet la recherche de traductions pour un terme candidat à la fois. Nous avons donc vérifié les traductions en deux langues pour chaque candidat des Top-50.

5.2.2.1 Paramètres de TTC TermSuite

TTC TermSuite permet de varier les paramètres de traitement. Nous avons globalement suivi les conseils du guide d'utilisateur¹². Plus précisément, nous avons lancé le traitement avec les paramètres suivants :

- choix de l'étiqueteur TreeTagger ;
- conservation des occurrences de termes dans la mémoire pendant le repérage ;
- utilisation de l'analyse morphosyntaxique ;
- utilisation des variantes syntaxiques ;
- utilisation des variantes graphiques avec un seuil fixé à 0,9 ;
- utilisation des contextes avec une fenêtre de 3 mots, termes monolexicaux uniquement ;
- fixation du seuil de fréquence à 1.

Nous avons fixé le seuil de fréquence égal à 1 pour rendre les résultats comparables avec notre approche qui prend en compte les termes ayant une seule occurrence dans le corpus.

Nous avons utilisé les dictionnaires proposés par TTC TermSuite¹³, sans ressources complémentaires. Nous avons également prévu de tester l'outil avec les ressources que nous utilisons dans notre système (Agrovoc, GEMET ou EcoLexicon), sans les dictionnaires de base afin de comparer les performances du TermSuite avec une faible quantité d'alignements.

5.2.2.2 Corpus

Nous avons utilisé les mêmes sous-corpus que pour l'évaluation de l'extraction terminologique : B09C et C02F constitués de 100 premières lignes de chaque fichier (cf Section 3.2.2.1) et annotés en termes attendus.

Nous avons créé une liste de termes par langue (FR, EN et DE) comprenant les termes mono et polylexicaux (sauf pour l'allemand, où nous nous sommes limités aux termes monolexicaux pour les raisons expliquées dans la Section 3.3.3.2).

Étant donné que le corpus est constitué de trois paires de langues, certains termes peuvent avoir une traduction dans une seule langue, ce qui rapproche les conditions à

11. <http://termsuite.github.io/documentation/properties/>

12. <http://termsuite.github.io/gui/#run-pipeline>, consulté en novembre 2016.

13. https://bintray.com/termsuite/generic/download_file?file_path=dictionaries-free.zip, consulté en novembre 2016.

5.2. ÉVALUATION

l'utilisation d'un corpus comparable. Par exemple, lorsqu'un terme français apparaît uniquement dans le sous-corpus FR-EN, le système ne pourra trouver sa traduction allemande que si elle est présente dans le sous-corpus EN-DE et si les dictionnaires (TermSuite) ou le modèle de n-grammes permettent l'alignement FR-DE.

5.2.2.3 Méthode d'évaluation

Pour chaque candidat, nous avons vérifié les traductions proposées par TermSuite (Figure 5.7) et par notre système. Cette vérification a révélé cinq catégories de résultats :

1. traduction complète (par exemple, *treatment stage* (EN) - *étage de traitement* (FR)) ;
2. traduction partielle (*contaminated soil* (EN) - *particules de sol* (FR) où seul le mot *soil* est traduit) ;
3. traductions proposées ne contenant pas d'éléments traduits (elles sont donc fausses) ;
4. absence de proposition ;
5. erreur signalée par le programme (TermSuite).

Les traductions complètes couvrent la totalité du terme source, tandis que celles partielles correspondent à une partie de terme (par exemple, *well shaft* -> *puits*, où *puits* est traduction de *well*).

Nous avons décidé de les prendre en compte dans l'évaluation, car une traduction incomplète est meilleure qu'une traduction erronée. Dans TermSuite, les traductions partielles sont souvent identifiées par l'approche semi-distributionnelle, tandis que les traductions complètes correspondent fréquemment à la méthode compositionnelle ou à celle qui applique le dictionnaire.

Dans le cas où un système n'est pas en mesure d'identifier une traduction, une absence de proposition est préférable, car elle évite une analyse infructueuse des candidats par un spécialiste humain.

5.2.2.4 Résultats

Nous avons évalué les traductions candidates pour les 50 termes candidats ayant le score de spécificité maximal dans les résultats de TTC TermSuite 2.2 pour les deux corpus dans toutes les paires de langues et dans toutes les directions (FR-EN et EN-FR, etc.). Pour la plupart des termes candidats, les deux systèmes, TTC TermSuite 2.2 et notre méthode, ont fourni plusieurs traductions candidates. Nous avons validé les traductions complètes et partielles, ainsi que celles erronées et absentes, comme décrit ci-dessus.

Nous avons également calculé le P-score¹⁴ qui évalue la significativité statistique des résultats, c'est-à-dire, la probabilité que la différence entre les résultats obtenus soit le fruit du hasard. Généralement, le seuil du P-score est fixé à 0,05 qui limite à 5 % la probabilité du résultat hasardeux. Ce score est inférieur à ce seuil dans la moitié de cas, et le résultat est statistiquement significatif dans l'ensemble de résultats avec le P-score égal à 0,00001.

14. <http://www.socscistatistics.com/pvalues/chidistribution.aspx>

5.2. ÉVALUATION

Dans notre cas, le P-score a été calculé à base de χ^2 (chi-carré)¹⁵. Nous avons choisi cette mesure car les résultats sont obtenus sur un même échantillon et sont évalués par rapport à quatre catégories, à savoir le nombre de traductions complètes, partielles, erronées et absentes (Table 5.7).

Le χ^2 est calculé selon la Formule 5.6 :

$$\chi^2 = \sum_{k=1}^{nb_cat} \frac{(O_k - E_k)^2}{E_k} \quad (5.6)$$

où k est la catégorie de résultats ; nb_cat est le nombre de catégories égal à 4 dans notre cas ; O est le nombre de traductions obtenues par la méthode des n -grammes et E – par TermSuite.

Dans le calcul du P-score, nous avons utilisé le degré de liberté (*degree of freedom*, DF) égal à 3, car nous avons 4 catégories d'évaluation et le $DF = nb_cat - 1$.

Corpus	Lang.	TermSuite				N-grammes				P-score
		C. ¹⁶	P.	F.	A/P	C.	P.	F.	A/P	
B09C	FR-EN	20	5	21	4	18	13	9	10	0,000017
	EN-FR	19	12	15	4	23	8	8	11	0,004
	EN-DE	16	8	20	6	17	9	17	7	0,839
	DE-EN	3	9	32	6	7	8	19	16	0,0005
	FR-DE	10	8	25	7	15	8	19	8	0,30
	DE-FR	10	11	24	5	9	10	21	10	0,37
C02F	FR-EN	19	7	14	10	16	7	8	19	0,025
	EN-FR	19	9	17	5	19	9	15	7	0,84
	EN-DE	12	4	29	5	16	10	16	10	0,00099
	DE-EN	4	11	21	14	7	13	18	12	0,49
	FR-DE	13	2	24	11	12	6	16	16	0,04
	DE-FR	7	12	17	14	9	14	13	14	0,58
Total		152	98	259	91	168	115	179	138	0,00001

TABLE 5.7 – Nombre de termes traduits sur Top-50 candidats selon le score de spécificité (TermSuite)

Selon ces critères, notre approche identifie plus de traductions complètes dans 7 cas sur 12 (Table 5.7), plus de traductions complètes et partielles dans 9 échantillons et plus de traductions dans l'ensemble de résultats. Dans un cas, les deux systèmes sont en égalité. En même temps, notre méthode propose moins de fausses traductions, en leur préférant une absence de proposition.

Le système TTC TermSuite se montre plus performant lorsque le candidat ou un des mots qui le composent, est présent dans les dictionnaires. Dans ces cas, TermSuite propose

15. <http://www.socscistatistics.com/tests/goodnessoffit/Default2.aspx>

16. C . signifie des traductions complètes, P . – partielles. Les résultats sont faux (F .) lorsque une ou plusieurs traductions sont proposées, mais ne contiennent pas d'éléments corrects et enfin A/P signifie une absence de proposition.

5.3. DISCUSSION

moins de traductions candidates et fournit une traduction complète ou partielle. Si toutefois il est contraint à se baser sur les vecteurs de contexte, les suggestions sont plus nombreuses et moins précises. Pour notre système, ce sont les candidats ayant des composants savants ou des n-grammes qui ont les propositions les plus précises.

Les deux systèmes ne possèdent pas de critères permettant d'éviter des fausses suggestions. Cet aspect est assez intéressant et peut être développé dans la perspective du projet. Encore un point faible dans les résultats est l'incapacité des deux systèmes de choisir une bonne traduction parmi plusieurs traductions candidates. Par exemple, pour le terme français *moléculaire*, notre système propose quatre traductions allemandes : *molekularem*, *molekulares*, *molekulare* et *molekulargewicht* (poids moléculaire). Cette dernière traduction candidate est nettement moins plausible et doit être éliminée des résultats. Nous avons essayé de choisir les traductions les plus proches au terme source en matière de longueur, mais ce filtre n'a pas démontré d'amélioration attendue. Toutefois, cet aspect pourrait être exploité davantage dans les futurs développements du projet.

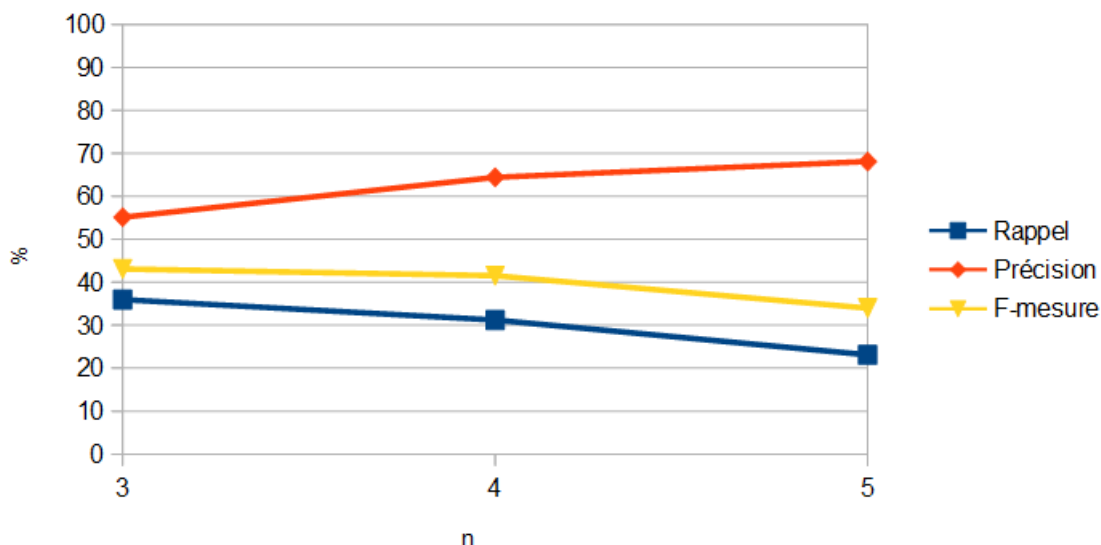
Malgré les écarts faibles dans certains échantillons, les résultats obtenus sont satisfaisants, car les moyens mis en œuvre par les deux systèmes ne sont pas les mêmes. En effet, nous avons déjà mentionné que TermSuite s'appuie sur des ressources linguistiques plus riches, ce qui lui permet d'identifier des termes par approche dictionnaire, compositionnelle (lorsque tous les mots d'un terme polylexical sont traduits à l'aide d'un dictionnaire) ou encore semi-distributionnelle (au moins un mot est traduit à l'aide d'un dictionnaire).

5.3 Discussion

Les résultats décrits dans la section précédente nous ont permis de faire quelques observations portant sur le comportement de notre système. Notamment, nous avons essayé plusieurs longueurs de n-grammes afin de choisir la valeur optimale de n . De plus, nous avons analysé la capacité des n-grammes de se substituer aux formants classiques, d'identifier des cognats et les traductions des termes complexes.

5.3.1 Influence du paramètre n sur les résultats d'alignement

Nous avons essayé de varier les valeurs du paramètre n sur le jeu de développement et sur les données d'évaluation. D'un côté, nous avons remarqué que les bi-grammes, étant très courts et fréquents, ne sont pas efficaces pour notre tâche. De l'autre côté, l'augmentation de n au-delà de penta-grammes limite la couverture, car plusieurs éléments des termes composés sont assez courts et fréquents, comme *luft* (DE) ou *_poly* (EN, FR, DE).

FIGURE 5.8 – Influence du paramètre n , où $n \in (3 : 5)$

La Figure 5.8 illustre l'évolution des moyennes du rappel, de la précision et de la F-mesure en fonction de la valeur de n sur le jeu de développement. L'augmentation du n permet d'obtenir une meilleure précision, mais diminue le rappel. Notre méthode atteint une précision proche de baseline avec $n = 4$ et $n = 5$.

Il est également possible d'utiliser simultanément des valeurs différentes de n , par exemple, combiner des tri-grammes, des quadri-grammes et les penta-grammes. Dans ce cas, il faut générer des fichiers d'alignement séparés et les combiner par la suite dans le modèle de traduction. En effet, Anymalign tend à aligner des n -grammes plus courts et plus fréquents, ce qui l'empêche d'aligner les n -grammes longs, naturellement moins fréquents.

5.3.2 N-grammes de caractères et formants néo-classiques

Le modèle de traduction contient un grand nombre de formants néo-classiques, c'est-à-dire des morphèmes d'origine grecque ou latine. Nous avons effectué une validation manuelle des paires de n -grammes alignés en prenant en compte les 100 premières lignes du modèle (généré à partir de GEMET, $n = 4$) trié dans l'ordre décroissant de la fréquence d'alignements fournie par Anymalign (le tri par défaut de l'outil). Pour la paire de langues EN-DE, nous avons relevé 19 alignements contenant au moins un formant classique, par exemple, *_biol* une partie de formant, comme *xide_*, dont 18 ont été alignés avec des éléments correspondants en langue cible. Pour la paire FR-DE, nous avons noté 32 alignements, dont 29 plausibles, et enfin pour la paire EN-FR tous les 34 alignements identifiés ont été bien alignés. Il faut remarquer que les n -grammes alignés pour cette dernière paire de langues sont les plus proches graphiquement. Les bonnes performances d'alignement des formants dans nos ressources sont dues aux termes chimiques et biologiques présents en quantité et serait peut-être différent pour des ressources dans des domaines utilisant moins de formants

5.3. DISCUSSION

néo-classiques dans leur nomenclature.

Langues	N-grammes alignés	
	Langue source	Langue cible
FR-EN	_poly	_poly
	_oxyd	_oxid
	_stér	_ster
FR-DE	_phot	_phot
	_vect	_vect
	_géom	_geom
EN-DE	_elec	_elek
	_aero	_aero
	_myco	_myko

TABLE 5.8 – Exemples de formants néo-classiques dans les n-grammes alignés à partir de GEMET

La Table 5.8 illustre quelques exemples de ces formants alignés. De plus, on observe fréquemment l'apparition des formants néo-classiques au début des mots (comme l'indique le marqueur "_").

5.3.3 Identification de cognats

L'alignement des n-grammes de caractères peut aider à la recherche des cognats, c'est-à-dire, des mots très proches dans les deux langues, par exemple *anthropogeography* (EN) et *anthropogeographie* (DE).

Langues	N-grammes alignés	
	Langue source	Langue cible
FR-EN	_écho	_echo
	_péri	_peri
	_écol	_ecol
FR-DE	_accu	_akku
	_ecol	_ökol
	ectr	ektr
EN-DE	_card	_kard
	_ecol	_ökol
	_soci	_sozi

TABLE 5.9 – Éléments pour la reconnaissance de cognats

Dans les alignements de n-grammes illustrés dans les Tables 5.8 et 5.9, les chaînes de caractères source et cible sont très proches. Ce phénomène est observé sur les formants classiques, mais aussi pour certains suffixes, par exemple *logy_* (EN) – *ogie_* (DE).

5.3. DISCUSSION

La prise en compte des n-grammes identiques dans la langue source et cible permet à notre méthode d'identifier les cognats même si leurs n-grammes ne sont pas dans le modèle de traduction : par exemple la traduction allemande *Navier-stokes-gleichungen* pour le terme anglais *navier-stokes equations*.

5.3.4 Traduction des termes composés

À la différence des cognats, les termes composés ne sont pas toujours proches graphiquement. Leur traduction constitue un des principaux défis pour notre projet. Afin de pouvoir travailler avec l'allemand, nous voulons pouvoir retrouver les traductions des mots composés sans avoir à les segmenter via un dictionnaire spécialisé.

Langues	N-grammes alignés	
	Langue source	Langue cible
FR-DE	_sabl	_sand
	_oxyg	_saue
	_frui	_obst
EN-DE	_solv	_lösu
	mill	ühle_
	_dese	_wüst
FR-EN	_bonn	_good
	_étiq	_labe
	_pâtu	_graz

TABLE 5.10 – Alignement des éléments des termes composés (GEMET)

Comme l'illustre la Table 5.10, les alignements des éléments constituant des termes composés ne sont pas parfaits. Par exemple, *étiq* (FR) peut à la fois se référer à *étiquette* ou à *étique* en français. Toutefois, nous n'avons pas besoin d'avoir des correspondances parfaites : lorsqu'un terme est traduit en utilisant les n-grammes de caractères, c'est le nombre des n-grammes traduits qui aidera à trouver la traduction correcte.

Notre méthode identifie des traductions assez complexes, comme par exemple *acide acrylique* (FR) – *Acrylsäure* (DE), *prévention des risques naturels* (FR) – *Naturrisikoverhütung* (DE), *Interessenausgleich* (DE) – *balancing of interests* (EN), *Grundwasserschutz* (DE) – *groundwater protection* (EN), *déchets hospitaliers* (FR) – *hospital waste* (EN) et *protection des réserves d'eau souterraines* (FR) – *groundwater protection* (EN).

Dans le couple de traductions *acide acrylique* (FR) – *Acrylsäure* (DE), le mot *acrylique* correspond à son cognat *Acryl*, tandis que *acide* est loin de *säure*. Ce couple a été identifié grâce aux n-grammes *_acry*, *acry*, *cryl* et *säur*. Ainsi, les n-grammes de caractères permettent d'éviter la segmentation des mots composés, tout en identifiant les traductions des éléments lexicaux. Il est également intéressant de noter que certaines traductions peuvent

16. *bensm* est très probablement extrait à partir de *lebensmittel* (DE) qui signifie *aliments* ; *ater_* de *water* (EN).

5.4. CONCLUSION

être partielles, sans pour autant être sémantiquement incomplètes. Par exemple, dans le cas de *protection des réserves d'eau souterraines* (FR) – *groundwater protection* (EN), le terme anglais ne contient pas de mot correspondant à *réserve*, mais le sens des deux termes est le même.

5.4 Conclusion

Notre méthode à base de n-grammes de caractères identifie les traductions des termes extraits à partir de corpus parallèles et comparables. La méthode requiert seulement un petit nombre de traductions des termes d'un domaine en deux langues pour préparer les alignements des n-grammes avec l'outil Anymalign. L'avantage principal de la méthode est sa capacité à identifier les traductions pour les termes composés monolexicaux, contenant des mots simples et les morphèmes, et polylexicaux, sans avoir à les segmenter en éléments constituants. La méthode prend en compte la position des n-grammes dans les mots et permet d'identifier les cognats même si les n-grammes qui les composent ne sont pas connus dans le modèle.

La méthode a été évaluée sur les paires de langues incluant l'anglais, le français et l'allemand, dans les deux directions (FR-DE et DE-FR, etc.). Elle a obtenu les meilleurs rappel (30,94 % en moyenne) et F-mesure (41,7 %) par rapport aux méthodes d'identification de cognats (19,08 % et 30 % respectivement), avec une légère différence en termes de précision (63,13 % et 70,30 %). Nous avons également démontré que la méthode reste performante en termes de rappel et de F-mesure même si le modèle a été obtenu à partir d'un nombre limité de traductions, la précision se montre toutefois plus sensible à la taille du modèle.

Nous avons également comparé les résultats de cette deuxième contribution avec TTC TermSuite 2.2. Notre approche identifie plus de traductions complètes et partielles (lorsque seulement une partie de terme est traduite) dans 9 échantillons sur 12 et plus de traductions dans l'ensemble de résultats. En même temps, notre méthode propose moins de fausses traductions, en leur préférant une absence de proposition.

Dans la perspective, il serait possible d'optimiser l'application combinée de notre méthode et des méthodes d'identification de cognats utilisés en tant que baseline. Cela permettrait de gagner en rappel et en F-mesure. Le blocage actuel se situe dans la différence significative de temps d'exécution entre les deux approches.

Chapitre 6

Construction des familles morphologiques endogènes

Les résultats de l'extraction terminologique mono ou multilingue sont généralement présentés sous forme de listes. Cette présentation permet de les trier dans l'ordre alphabétique ou en fonction de scores utilisés pour leur identification (par exemple, de la fréquence ou TFxIDF, ou encore d'un autre paramètre). Toutefois, cette forme de présentation ne permet pas de visualiser les termes proches si leurs scores ne le sont pas. Par exemple, si le terme *géométrie* est beaucoup plus fréquent que *géométrique*, il est possible qu'ils soient éloignés sur la liste de résultats. Cela entraîne une perte de temps pour un spécialiste humain qui se charge du traitement de ces résultats, par exemple, en les insérant dans un dictionnaire spécialisé. La présentation des résultats de l'extraction terminologique sous forme de familles morphologiques permet de pallier à ce problème. En effet, les termes morphologiquement proches le sont certainement aussi sur le plan sémantique.

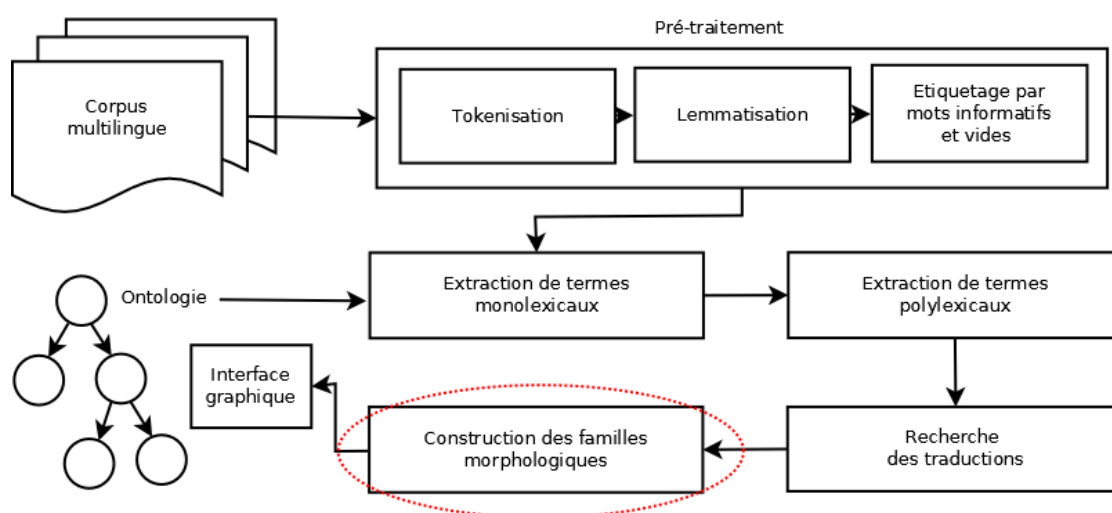


FIGURE 6.1 – Place de la constitution des familles morphologiques dans notre méthode

6.1. DÉFINITION DES FAMILLES MORPHOLOGIQUES ENDOGÈNES

La Figure 6.1 montre que la construction des familles morphologiques intervient après l'extraction terminologique et la recherche des traductions. En fonction de la configuration choisie, les familles peuvent être mono ou multilingues. Dans ce chapitre, les familles sont évaluées par rapport :

- à leur capacité de regrouper des termes candidats pour les termes de l'ontologie ;
- à leur utilité pour suivre les changements diachroniques dans la terminologie ;
- et à l'utilité de regroupements multilingues pour améliorer les suggestions de traductions.

6.1 Définition des familles morphologiques endogènes

Les familles morphologiques sont des regroupements de mots partageant un ou plusieurs éléments qui peuvent être, en fonction de l'approche, des morphèmes (Bernhard, 2006) ou des séquences de caractères (Schone et Jurafsky, 2000). Hathout (2011) définit les familles dérivationnelles comme « les groupements de mots très proches les uns des autres dans le sens où ils partagent le plus de propriétés possible, ces propriétés sémantiques et formelles étant les plus spécifiques possible ». Cette spécificité des propriétés oppose les familles à des séries dérivationnelles qui sont construites à base des propriétés les plus générales possibles et contiennent des mots des différentes familles morphologiques (Hathout, 2011). Dans l'exemple fourni par Hathout (2011), la famille du mot *produire* contient *produit*, *production*, *reproduire*, *productif*, etc., tandis que la série dérivationnelle du suffixe *-age* regroupe les mots comme *lavage*, *façonnage*, etc.

Dans le cadre de notre projet, les familles sont construites grâce aux n-grammes du début et du milieu des mots, qui correspondent préfixes et aux racines. Nous pouvons imaginer comme exemple une famille de mots partageant le préfixe *_nano* : *nanoparticule*, *nanostucture*, *nanotube*, etc. Dans ce cas, il faut segmenter les mots en morphèmes, ce qui introduit une dépendance à la qualité d'une telle segmentation, ce qui constitue le problème majeur des systèmes existants.

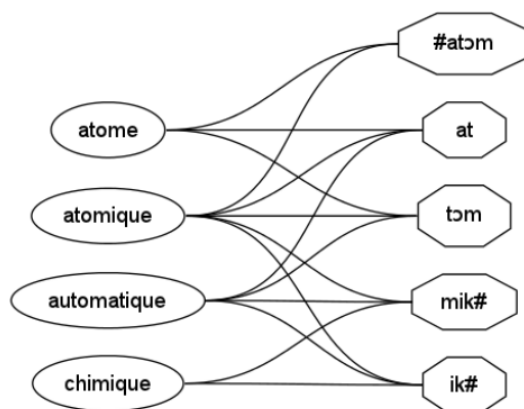


FIGURE 6.2 – Graphe bipartite reliant les mots et les n-grammes de phonèmes (Hathout, 2011)

Hathout (2011) a contourné cette segmentation en utilisant les n-grammes de phonèmes¹, comme illustre la Figure 6.2. La longueur de ces n-grammes est variable de 1 à la longueur maximale d’un mot dans la ressource. L’information phonétique n’est pas disponible dans la plupart des ressources, ni dans les corpus. On pourrait toutefois imaginer contourner ce problème avec un système graphème vers phonème.

Comme dans le reste de notre projet, nous utilisons les n-grammes de caractères. Ces n-grammes ayant une longueur n fixe, ne correspondent pas exactement à des morphèmes, qui ont une longueur variable. Par exemple, le préfixe *_bio* compte quatre caractères, contre cinq caractères pour *_nano*. Si l’on veut regrouper des mots autour de ces n-grammes, il faut sélectionner les n-grammes présentant un intérêt pour le domaine. Par exemple, pour créer une famille de mots contenant le n-gramme *nano*, il faut d’abord dresser la liste de n-grammes à traiter et y inclure *nano*.

Afin d’éviter la sélection des n-grammes, nous construisons les familles par rapport aux termes déjà présents dans l’ontologie. De cette façon, si l’ontologie contient un concept avec le terme *nanomètre*, les candidats *nanoparticule*, *nanostucture*, *nanotube* sont proposés pour ce terme. La construction d’une famille pour un terme permet à l’utilisateur de définir sa propre stratégie de travail. Par exemple, il peut se concentrer sur une branche de l’ontologie et y ajouter des termes ou des nouveaux concepts. Les termes candidats peuvent être présents dans plusieurs familles et être introduits dans la structure à partir d’une de ces familles.

De manière générale, la présence de termes d’une ontologie ne constitue pas un pré-requis absolu. Si le système est exploité dans l’objectif unique d’extraire la terminologie d’un domaine, il est possible de constituer les familles pour les termes candidats.

Les familles morphologiques produites par notre système peuvent être considérées comme endogènes, dans la mesure où elle s’appuient sur les n-grammes appris à partir de termes présents dans l’ontologie et du corpus.

6.2 Méthodologie

L’apprentissage des familles morphologiques endogènes suit l’extraction terminologique et utilise la liste de termes candidats mono et polylexicaux, ainsi que l’index des n-grammes avec leurs scores d’importance calculés à l’étape de l’extraction terminologique monolexicale. Rappelons que nous calculons leur score d’importance en prenant en compte le nombre de termes dans la ressource terminologique contenant le n-gramme en question, ainsi que le nombre de lemmes associés dans le corpus et leur fréquence (cf Équation 4.3).

1. Cette information phonétique est disponible dans le Trésor de la Langue Française informatisé (TLFi), la ressource utilisée pour la construction de Morphonette (Hathout, 2011).

6.2.1 Création des index de n-grammes

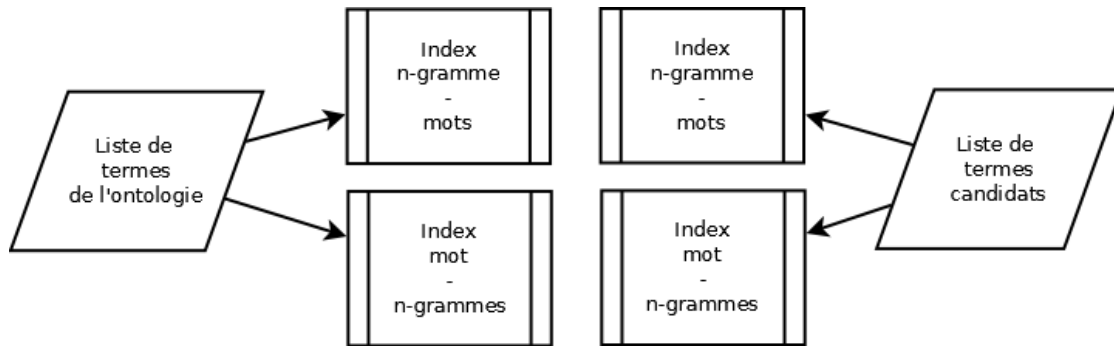


FIGURE 6.3 – Création des index

Au début du traitement, nous avons deux listes de mots (Figure 6.3) :

- la liste de termes de l'ontologie ;
- la liste de termes candidats extraits à partir de corpus.

Les mots sur ces deux listes sont segmentés en n-grammes de caractères² et deux index sont générés pour chaque liste :

- les mots indexés par n-grammes ;
- les n-grammes indexés par mots.

Ce traitement rend possible la recherche des n-grammes communs entre les mots des deux listes et il est possible, le cas échéant, d'inverser les traitements et de créer des familles de termes de l'ontologie pour les termes candidats.

6.2.2 Constitution des familles morphologiques

L'objectif de cette étape est d'identifier parmi les termes candidats ceux qui ont des traits communs suffisamment forts avec les termes de l'ontologie.

2. Il faut noter que les n-grammes associés à un terme polylexical résultent de la segmentation de chaque mot faisant partie de ce terme.

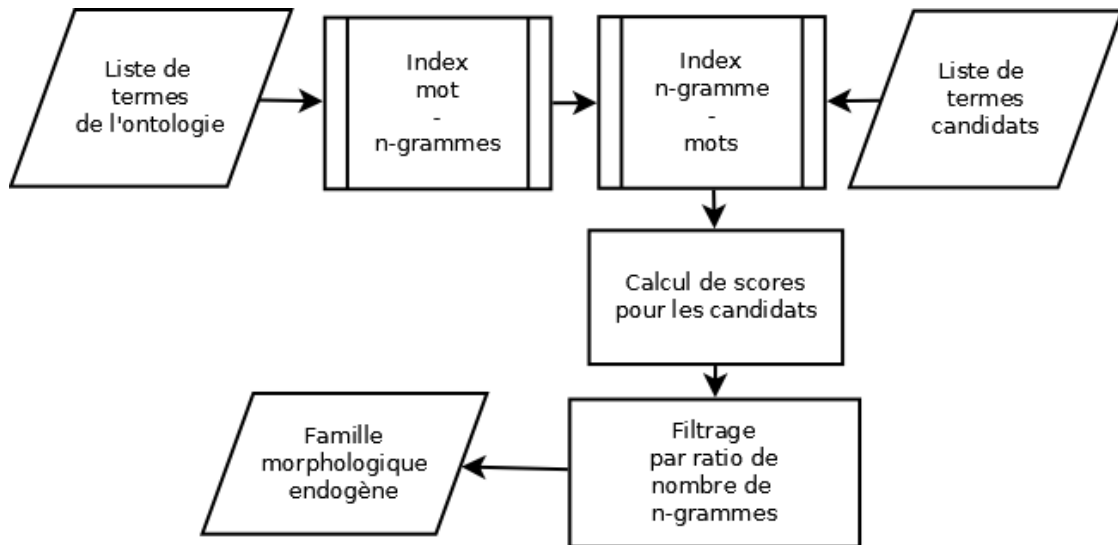


FIGURE 6.4 – Construction des familles morphologiques endogènes

Ainsi, pour chaque terme, une liste de ses n-grammes est récupérée à partir de l'index correspondant (Figure 6.4). Ensuite, pour chaque n-gramme, on récupère tous les candidats qui y sont associés.

Pour chaque candidat identifié ainsi, on calcule son score d'appartenance à la famille. Ce score prend en compte le nombre de n-grammes partagés par le terme et le candidat. Afin d'éviter de classer dans une famille des candidats trop éloignés, nous avons introduit un filtre qui augmente ou diminue le nombre de n-grammes nécessaire en fonction de leur longueur n . En effet, plus les n-grammes sont courts, moins ils sont fiables. Il en faut donc plus pour confirmer l'appartenance d'un candidat à une famille. Si, au contraire, les n-grammes sont longs, il ne peuvent pas être nombreux dans un terme, mais leur qualité est meilleure. Il en faut donc moins.

Plus précisément, le filtre est basé sur le ratio du nombre de n-grammes du mot le moins long des deux (le terme ou le candidat) par rapport à la longueur de n-grammes n . Deux cas sont possibles :

- si le terme est plus long que le candidat, comme dans le cas du terme *chlorinité* et du candidat *chlore*,
 - $ratio = \frac{\text{nombre_ngrammes_candidat}}{n}$;
- si la longueur du candidat est égale ou supérieure à celle du le terme, par exemple le candidat *tissu biologique* et le terme *biologie*,
 - $ratio = \frac{\text{nombre_ngrammes_terme}}{n}$.

Pour passer à travers de ce filtre, le nombre de n-grammes du candidat communs avec le terme doit être supérieur ou égal à la valeur de ce ratio.

Exemple 1 Si le paramètre $n = 3$, les n-grammes communs entre le terme *chlorinité* et le candidat *chloré* sont *_chl*, *hlo* et *lor*. Le nombre total de *nombre_ngrammes_candidat* pris en compte est égal à 3. Rappelons que les n-grammes de fin de mots sont exclus de ce traitement. En effet, ces n-grammes tendent à correspondre à des suffixes et à des

6.2. MÉTHODOLOGIE

terminaisons, et n'apportent donc pas de précision dans l'alignement des candidats. Le ratio est égal à 1 : $ratio = \frac{\text{nombre_ngrammes_candidat}}{n} = \frac{3}{3} = 1$. Le nombre de n-grammes communs est supérieur à la valeur du ratio et le candidat *chloré* est retenu pour la famille du terme *chlorinité*.

Exemple 2 Prenons maintenant le candidat *chlorure de aluminium* qui est plus long que le terme *chlorinité*. Dans ce cas, c'est le nombre de n-grammes dans le terme qui est pris dans le calcul du ratio. Les trois n-grammes communs sont les mêmes que dans l'exemple précédent. Les n-grammes du terme *chlorinité* sont *_chl*, *hlo*, *lor*, *ori*, *rin*, *ini* et *nit*. Leur nombre est égal à 7. Alors, la valeur du ratio est égale à 2,33 ($ratio = \frac{\text{nombre_ngrammes_terme}}{n} = \frac{7}{3} = 2,33$). Le candidat *chlorure de aluminium* est également éligible à la famille du terme *chlorinité*.

Exemple 3 Pour illustrer un contre-exemple, nous avons choisi le candidat *aminique*. Il est plus court que le terme *chlorinité*. Ses n-grammes sont *_ami*, *min*, *ini*, *niq* et *iqu*, dont seulement un n-gramme en commun avec le terme. La valeur du ratio est égale à 1,67 ($ratio = \frac{\text{nombre_ngrammes_candidat}}{n} = \frac{5}{3} = 1,67$). Ce candidat ne satisfait pas à la condition d'avoir le nombre de n-grammes en commun supérieur ou égal à la valeur du ratio.

6.2.3 Optimisation de la présentation de résultats

La traitement précédent résulte en une famille morphologique endogène pour un terme de l'ontologie, où les candidats sont présentés dans l'ordre décroissant de leurs scores d'appartenance à la famille en question :

chlorinité : *chlorure*, *chlorure de chlorométhyle triéthylammonium*, *chlorurer*, *ion chloruré*, *sel de chlorure*, *chlorure ferrique*, *solution aqueuse de chlorure*, *chlorométhyle*, *chlorure de chlorométhyle triéthylammonium*, *chlorométhyle triéthylammonium*, *chloré*, *hydrocarbure chloré*, *chlorhydrique*, *acide chlorhydrique*, *assainir les biphényles polychlorés*, *polychloré*, *biphényles polychlorés*.

Cette présentation a l'inconvénient de contenir des sous-familles sans les mettre en évidence. Notre solution à ce problème est de regrouper des termes ayant le même score dans des sous-familles, si les termes plus longs contiennent ceux plus courts :

chlorinité :

- *chlorure* :
 - *chlorure de chlorométhyle triéthylammonium*
- *chlorurer* :
 - *ion chloruré*
 - *sel de chlorure*
 - *chlorure ferrique*
 - *solution aqueuse de chlorure*
- *chlorométhyle* :
 - *chlorométhyle triéthylammonium* :
 - *chlorure de chlorométhyle triéthylammonium*
- *chloré* :
 - *hydrocarbure chloré*
- *chlorhydrique* :
 - *acide chlorhydrique*
- *polychloré* :
 - *biphényles polychlorés* :
 - *assainir les biphényles polychlorés*

L’optimisation de la famille met en évidence les liens sémantiques, dont ceux hiérarchiques (par exemple, *polychloré* et *biphényles polychloré*) ou non-hiérarchiques (*biphényles polychloré* et *assainir le biphényles polychloré*), qui existent entre les candidats extraits à partir du corpus. Ainsi, l’utilisateur peut optimiser son travail et insérer des nouveaux concepts avec leur micro-hiérarchie ou ajouter des liens avec d’autres concepts.

6.3 Évaluation

À la différence des étapes précédentes, il est difficile de définir une baseline pour évaluer les familles morphologiques endogènes. Les mesures d’évaluation, comme le rappel, la précision ou encore la F-mesure, ne conviennent pas dans le cas des familles, car elles ne sont pas définies en avance.

Pour cette raison, l’évaluation de cette étape est plus qualitative que quantitative. Nous allons analyser les cas d’application des familles morphologiques endogènes par rapport aux deux objectifs :

- regrouper les candidats par rapport aux termes (libellés) de l’ontologie ;
- étudier l’évolution diachronique des familles.

6.3.1 Familles pour les libellés de l’ontologie

Nous avons évalué la construction de familles morphologiques endogènes pour les termes des trois ontologies : EcoLexicon, GEMET et Agrovoc. Afin d’évaluer cette étape de traitement indépendamment des autres, les listes de termes candidats mono et polylexicaux ont été constituées des termes attendus pour nos corpus pré-annotés (cf Section 3.3.3). La première étape consiste à segmenter les termes de l’ontologie et les termes candidats

6.3. ÉVALUATION

en n-grammes de caractères pour ensuite regrouper les termes candidats dans les familles morphologiques pour les termes de l'ontologie.

Langue	Couverture (%)	Taille moyenne (mots)	Taille médiane (mots)
FR	93,42	32,28	18,33
EN	95,41	40,91	22,33
DE	95,30	27,40	16,33

TABLE 6.1 – Taille et couverture des familles morphologiques endogènes

Le système construit des familles pour 94,7 % de termes dans les ressources en moyenne (Table 6.1). Ce pourcentage élevé s'explique par le fait que le modèle de n-grammes est généré à partir de termes d'une ressource. La taille moyenne des familles fluctue entre 19 et 48 candidats, tandis que la taille médiane est entre 10 et 30 candidats selon les corpus et les ontologies.

6.3.1.1 Exemples de familles morphologiques endogènes

Les familles sont organisées sous forme d'arborescences, où les candidats plus longs sont placés dans les sous-arbres correspondant à ceux plus courts qu'ils contiennent :

aquifère semi-captif :

- *aquifère* :
 - *zone aquifère*
- *aqueux*
- *capteur* :
 - *capteur d'ions*
- *captage* :
 - *captage d'air*
- *semi-précieuse*
- *capot*
- *semi-perméable*
- *semi-solide*
- *capter*

Cette organisation permet de préparer des sous-arbres à insérer dans l'ontologie cible. Dans l'exemple ci-dessus, l'utilisateur peut d'abord créer le concept *aquifère* et le connecter au concept *aquifère semi-captif* par la relation d'hypéronymie ; ensuite, créer le concept *zone aquifère* et le lier par une relation sémantique d'association avec le concept *aquifère*.

La famille pour le terme *aquifère semi-captif* contient les candidats en lien avec l'eau (*aquifère*, *zone aquifère*, *aqueux*), l'action de capter (*capter*, *capteur*, *captage*, etc.), et la notion de moitié (*semi-précieuse*, *semi-perméable* et *semi-solide*).

6.3. ÉVALUATION

Indépendance à la langue Les familles sont indépendantes de la langue dans la mesure où celle-ci permet une segmentation des mots en n-grammes de caractères. Les familles les plus grandes sont construites pour l'anglais et celles les plus courtes sont générées pour l'allemand.

En allemand, un grand nombre de familles contient très peu de mots, comme dans le cas du terme *Ökoentwicklung* (développement écologique) qui n'a qu'un seul membre dans sa famille : *Weiterentwicklung* (futur développement). Ce phénomène s'explique par la performance du filtre selon le ratio du nombre de n-grammes (cf Section 6.2.2).

Taille de familles Il faut noter que, même si le nombre total de candidats regroupés dans une famille peut être important, le nombre d'éléments du premier niveau reste limité. En effet, les familles de taille importante contiennent des candidats comme *procédé*, *produit*, etc. faisant partie d'un grand nombre de candidats composés. Dans l'exemple suivant, la famille du terme *processus* contient seulement 10 candidats du premier niveau d'arborescence, dont deux (*processus* et *procédé*) ont des sous-arbres riches en candidats :

processus : – <i>processus</i> : – <i>processus catalytique</i> – <i>processus d'épuration</i> – <i>processus électrolytique</i> – <i>processus biocatalytiques</i> : – <i>accélération du processus biocatalytique</i> – <i>optimisation de processus</i> – <i>processus de recyclage</i> – <i>donnée de processus</i> – <i>processus de lavage</i> – <i>processus de nitrification</i> – <i>processus de ventilation</i> – <i>processus de contrôle</i> – <i>simulation de processus</i>	– <i>procédé</i> – <i>procédé de production</i> – <i>procédé électrochimique</i> – <i>procédé de dégradation</i> – <i>procédé pour la dégradation</i> – <i>procédé servant à extraire</i> – <i>procédé de détoxification</i> – <i>procédé de immobilisation</i> – <i>procédé de contrôle</i> – <i>procédé de élimination</i> – <i>procédé de fermentation</i> : – <i>procédé de fermentation anaérobie</i> – <i>procédé de fermentation aérobie</i> – <i>procédé de séparation</i> , etc.
---	--

Ainsi, l'exploitation des familles pourrait être facilitée grâce à une interface utilisateur permettant une navigation dynamique dans cette arborescence. Par exemple, l'utilisateur pourrait ouvrir les sous-arbres pour les candidats du premier niveau qu'il juge pertinents.

6.3.1.2 Pertinence des familles

Les familles morphologiques endogènes fonctionnent bien pour les termes composés et pour les termes monolexicaux faisant partie d'un certain nombre de termes polylexicaux. Pour ces termes, les familles comportent des éléments utiles à l'élargissement et à l'enrichissement des sous-arbres correspondants de l'ontologie source.

Toutefois, presque toutes les familles contiennent du bruit, des candidats sans relation avec le terme de tête. De manière générale, ces candidats ont des scores peu élevés et se trouvent à la fin de la liste. Cependant, certaines familles sont entièrement constituées de

6.3. ÉVALUATION

candidats sans relation au terme en question ou ne contiennent pas de candidats. Cela peut être dû à plusieurs raisons :

- le corpus ne contient pas de candidats qui pourraient correspondre au terme de l'ontologie (par exemple, c'est le cas du terme *zoologie* qui n'a pas de relation avec les corpus en traitement des eaux usées ou des sols pollués, et dont la famille est remplie de candidats comme *écologie*, *biologie*, etc.) ;
- l'étymologie commune avec des mots qui ont pris des sens différents au cours de l'histoire (la famille du terme *effet de serre* contient le candidat *serrure*, ce qui est dû à leur origine : le verbe *serrer* du latin *serare*³) ;
- la longueur du terme est inférieure à la longueur *n* du n-gramme (*co2*, *cr*, etc.) ;
- les candidats ne sont retrouvés par le module d'extraction terminologique (des cas difficiles à illustrer, mais toutefois nécessaires à prévoir).

Cette évaluation permet de juger l'utilité des familles morphologiques endogènes pour l'organisation des résultats d'extraction terminologique multilingue. Nous allons maintenant présenter les résultats d'une expérience qui vise à utiliser ces familles dans l'analyse diachronique de la terminologie d'un domaine. Plus concrètement, il s'agit de comparer les familles construites à des périodes différentes pour mettre en évidence les néologismes d'un domaine.

6.3.2 Recherche des néologismes

Détecter des termes du domaine à partir du moment où ils apparaissent est compliqué, car leurs fréquences ne sont pas élevées. De plus, il existe rarement des corpus traitant de dernières avancées dans un domaine et leur taille serait modeste comparée aux corpus du domaine complets.

Notre méthode d'extraction terminologique à base de ressources endogènes permet d'identifier les candidats, même si leurs fréquences sont faibles, sans exigences particulières envers la taille du corpus. Le regroupement des résultats dans des familles morphologiques endogènes permet d'associer les candidats aux termes existants et de comparer les résultats entre les corpus appartenant à des périodes différentes.

Cette expérience a été menée indépendamment des autres évaluations, sur des corpus différents de ceux présentés dans le chapitre dédié et en utilisant des paramètres particuliers dans l'extraction terminologique. Toutefois, les résultats de cette expérience sont intéressants pour illustrer l'évolution diachronique des familles morphologiques endogènes et, notamment, l'apparition des nouveaux termes. Pour cette raison, les ressources et les paramètres d'expérience sont inclus dans cette section.

6.3.2.1 Ressources

Dans le cadre de cette expérience, nous avons recours à deux types de ressources : l'ontologie EcoLexicon qui sert à la génération du modèle des n-grammes et au regroupement des familles par terme, et le corpus subdivisé en trois périodes pour identifier les évolutions des familles dans le temps.

3. <https://fr.wiktionary.org/wiki/serre> et <https://fr.wiktionary.org/wiki/serrure>

6.3. ÉVALUATION

Corpus Dans cette expérience, nous avons utilisé une partie du corpus PatTR correspondante à la catégorie C25B⁴ de la CIB.

Afin d'étudier la terminologie du point de vue du temps, nous avons séparé le sous-corpus en trois périodes (en identifiant les dates à partir des méta-données) :

- les années 1980 (1979-1989)
- les années 1990 (1990-1999)
- et les années 2000 (2000-2008)

Période	Lemmes		
	FR	EN	DE
1980s	101 237	29 968	86 294
1990s	598 075	793 199	481 315
2000s	540 027	662 297	437 158
Total	1 239 339	1 485 464	1 004 767

TABLE 6.2 – Répartition du sous-corpus C25B PatTR par période

La quantité de données par période n'est pas homogène (cf. Table 6.2), car il est difficile d'avoir un corpus équilibré par période et par langue.

6.3.2.2 Paramètres utilisés dans l'expérience

Rappelons brièvement les principales étapes de l'extraction terminologique monolexicale⁵ et de la construction des familles morphologiques endogènes.

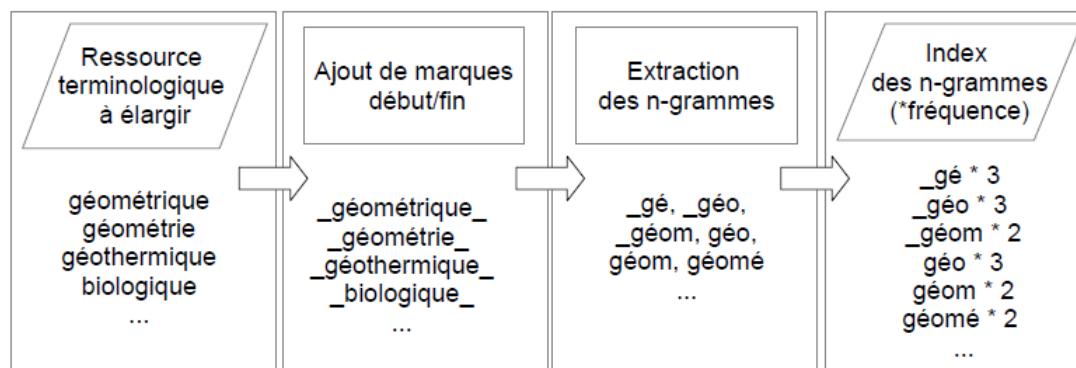


FIGURE 6.5 – Création de l'index

4. Procédés électrolytiques ou électrophorétiques pour la production de composés organiques ou minéraux, ou de non-métaux ; appareillages à cet effet.

5. Dans cette expérience, seuls les candidats monolexicaux sont pris en compte.

6.3. ÉVALUATION

Création des index de n-grammes de caractères La première étape du traitement est la création des index des n-grammes de caractères appris à partir des termes d'EcoLexicon. Dans cette expérience, nous avons utilisé simultanément les n-grammes où $n \in [3 : 5]$ ⁶ (Figure 6.5), par exemple *_gé*, *_géo* et *_géom* avec leurs fréquences (3, 3 et 2 respectivement).

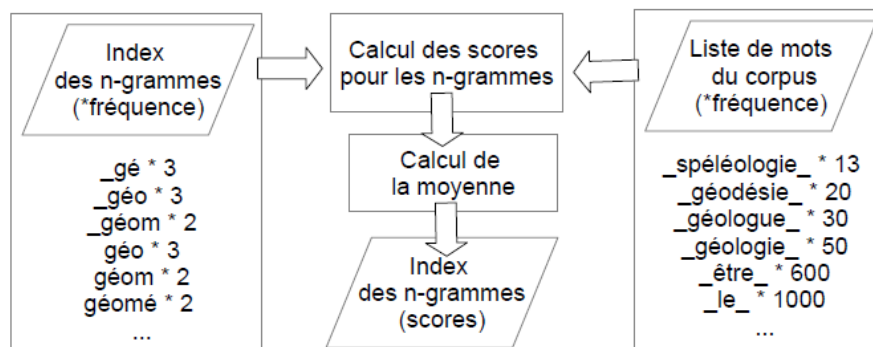


FIGURE 6.6 – Calcul des scores

Sélection des n-grammes importants Le calcul du score terminologique d'un n-gramme est fait de la même manière que dans l'extraction terminologique monolexicale (cf Formule 4.3). Les n-grammes dont le score dépasse la moyenne⁷ sont considérés importants pour le domaine et serviront à l'identification des termes candidats et à la construction des familles morphologiques. Le recours à la moyenne diminue le nombre de candidats, en augmentant ainsi la précision. Toutefois, ce paramètre est utilisé uniquement dans cette expérience.

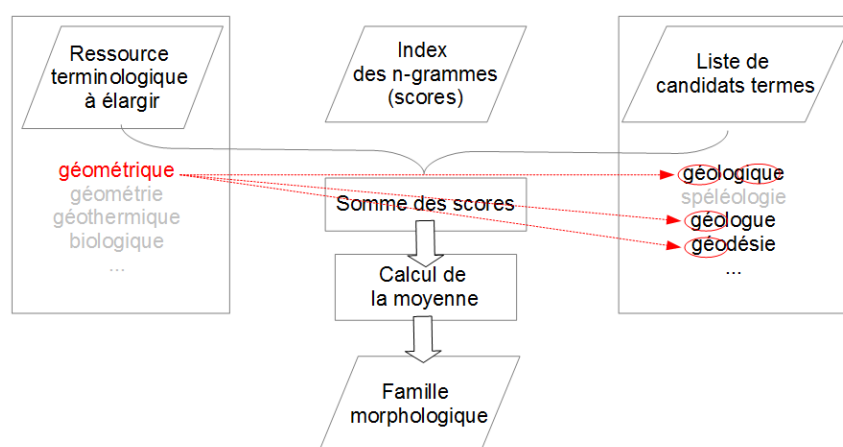


FIGURE 6.7 – Construction des familles morphologiques endogènes

6. Dans cette expérience, la marque du début et de fin de mot a été comprise dans le calcul de n .

7. $\frac{\sum n_grams_scores}{nb_n_grams}$

6.3. ÉVALUATION

Sélection des termes candidats et construction des familles À cette étape, les termes candidats monolexicaux sont identifiés à partir du corpus, tout comme dans le cadre de l'extraction terminologique monolexicale (cf Section 4.1.3). Enfin, les familles sont créées pour chaque libellé d'EcoLexicon (Figure 6.7).

6.3.2.3 Résultats

Extraction des termes candidats Notre approche a pour ambition d'identifier les termes du domaine à fréquences faibles indépendamment de la langue et de la taille du corpus.

Langue	Précision (%) par rapport à la position dans la liste			Total termes	Termes connus (sur 300)
	Tête	Milieu	Fin		
FR	100	97	71	2 151	11
EN	100	96	96	2 666	31
DE	100	97	90	6 937	5

TABLE 6.3 – Précision des termes candidats (%) par rapport à la liste triée dans l'ordre décroissant des scores (corpus 2000)

Notre méthode a permis d'identifier 2 151 candidats pour le français (cf Table 6.3), qui ont été évalués manuellement sur des échantillons de 100 termes issus de la tête, du milieu et de la fin de la liste.

Nous avons également vérifié si les termes candidats sont présents dans la ressource terminologique d'origine. Pour le corpus en français, sur les 100 premiers candidats de la liste, seuls 4 étaient présents dans l'ontologie EcoLexicon⁸ (5 au milieu et 2 à la fin). La Table D.1 (dans les Annexes) donne une meilleure idée des candidats obtenus avec leurs fréquences dans le corpus. Ces échantillons mettent en avant le fait que notre approche permet d'identifier les termes candidats rares.

Langue	Famille
aérobic (FR)	<i>aérosol, aérodynamique, anaérobic, aérien, aération</i>
aerosol (DE)	<i>Aerosole, Aerosol-abgabebehältern, Aerosol-abgabebehälter, Aerogelpartikel, Aerosolbesprühen</i>
aerobic (EN)	<i>aerosol, aerogel, aerial, anaerobic</i>

TABLE 6.4 – Famille morphologique pour le terme *aérobic*

8. Version utilisée dans cette évaluation remonte au printemps 2015.

6.3. ÉVALUATION

Familles morphologiques Pour illustrer nos propos, prenons l'exemple de la famille construite pour le mot *aérobic* en français (Table 6.4) : *_aérosol_*, *_aérodynamique_*, *_anaérobic_*, *_aérien_*, *_aération_*. Tous les candidats suggérés se réfèrent à l'air et sont potentiellement intéressants pour être ajoutés à l'ontologie en liaison avec le concept *air*. De plus, le terme candidat *_anaérobic_* est un antonyme du terme source. Présentés ainsi, ces termes candidats pourront faciliter le travail de l'expert humain sur l'ontologie source (ou sur un autre type de ressource terminologique).

En allemand, cette méthode permet d'identifier les mots composés qui contiennent un concept du domaine, par exemple, le terme *Wasser* présent dans l'ontologie EcoLexicon a attiré 303 candidats dans sa famille construite à partir du corpus des années 2000.

Évolution diachronique des familles morphologiques endogènes Étudier l'apparition des néologismes au sein de leurs familles morphologiques présente plusieurs avantages : d'un côté, cette approche permet d'extraire les termes candidats néologiques même si leur fréquence est faible et de l'autre côté, de les positionner dans la terminologie existante. Par ailleurs, le recours à la morphologie endogène dans le contexte multilingue rend possible la recherche des traductions des néologismes à partir des équivalences apprises au niveau des morphèmes.

La terminologie évolue dans le temps avec son domaine. Pour cette raison, il est important de suivre ses évolutions et mettre les ressources terminologiques à jour. Nous avons comparé les familles morphologiques construites à partir du corpus des années 80 et de celui des années 2000.

Terme	1980		2000		$\cap$
	Nb cand	Précision	Nb cand	Précision	
phosphore (FR)	26	34,61 %	99	51,51 %	4
phosphor (DE)	5	100 %	137	100 %	1
phosphorus (EN)	18	38,89 %	99	52,52 %	2

TABLE 6.5 – Familles morphologiques dans le temps

Prenons l'exemple de familles construites pour le terme *phosphore* en français illustré par la Table 6.5. Le corpus des années 80 en permet d'identifier 26 candidats dans la famille, dont 9 sont corrects⁹ : *_phosphonique_*, *_phosphure_*, *_phosphoniques_*¹⁰, *_phosphines_*, *_phosphorique_*, *_phosphate_*, *_peroxydiphosphate_*, *_polyphosphazènes_*, *_2-éthylhéxylphosphite_*. La famille construite à partir du corpus des années 2000 contient 99 candidats, dont 51 valides. L'intersection des candidats valides entre les deux corpus contient seulement 4 termes.

9. On entend ici par candidats valides ceux qui sont effectivement dérivés du terme source ou qui lui sont morphologiquement proches

10. La marque du pluriel est due à l'erreur d'étiquetage fait par TreeTagger, ce qui est souvent le cas lorsqu'il s'agit des termes techniques

6.3. ÉVALUATION

Dans le cas de l'allemand, les familles pour le terme *phosphor* contiennent 5 candidats pour le corpus des années 80 et 137 candidats pour les années 2000 avec la précision de 100 % dans les deux cas avec un seul candidat partagé. En anglais, les familles pour le terme *phosphorus* réunissent 18 candidats, dont 7 valides (les années 80), et 99 candidats avec 52 valides (les années 2000).

Afin de suivre l'évolution des termes candidats dans le temps, nous avons sélectionné des termes candidats de la famille *phosphorus* (EN) et visualisé leur fréquences sur Google Ngrams Viewer¹¹, la ressource qui exploite un large corpus de livres numérisés essentiellement en anglais.



FIGURE 6.8 – Google Ngrams Viewer, EN, 1980 - 2008

Les résultats ont été peu représentatifs (cf Figure 6.8), car cet outil connaît peu la terminologie des domaines de spécialité. Nous avons quand même identifié les termes candidats inconnus à cette ressource ou ayant des fréquences très faibles (cf Tableau 6.6).

11. <https://books.google.com/ngrams>, consulté le 14 décembre 2015

6.3. ÉVALUATION

Peu d'occurrences	Pas d'occurrences
hexafluorophosphates, nickel-phosphorous, diphosphite, diphenylphosphine	polyvinylphosphoric, diarylalkylphosphines, polyvinylphosphonic, phosphonooxy, phosphonomethyl-iminodiacetic, hexafluorophosphates, alkylphosphanes, fluorophosphorano, fluoralkylphosphates, hydroxysilylphosphonic, alkylphosphorane, dihydrogenphosphates, diarylphosphine, perfluoroalkylfluorophosphoranes, perfluoroalkylfluorophosphoranes, fluoroalkylphosphates, alkylphosphoranes

TABLE 6.6 – Résultats retenus de Google Ngrams Viewer

Nous avons continué l'évaluation en utilisant le portail SciFinder¹² qui recense des publications scientifiques en chimie sur la période allant jusqu'aux années 1900. La base permet la recherche d'un terme par langue, mais contient majoritairement des publications en anglais. Nous nous sommes donc concentrés sur la recherche monolingue en anglais. Les fonctionnalités de l'interface permettent d'exporter les données sur le nombre de publications par année contenant le terme recherché. Cette opération reste manuelle, nous avons donc choisi de faire l'évaluation sur une partie des termes candidats n'ayant pas d'occurrences sur Google Ngrams Viewer.

La Figure 6.9 montre que seul le terme *hydroxysilylphosphonic* n'a pas d'occurrences sur SciFinder. Les termes *hexafluorophosphates* et *phosphonooxy* sont largement attestés déjà dans les années 1970. Pour les autres candidats, on peut observer l'évolution croissante de leur usage, même si pour beaucoup d'entre-eux la fréquence par décennie ne dépasse pas le seuil de 10 occurrences.

12. <http://www.cas.org/products/scifinder>

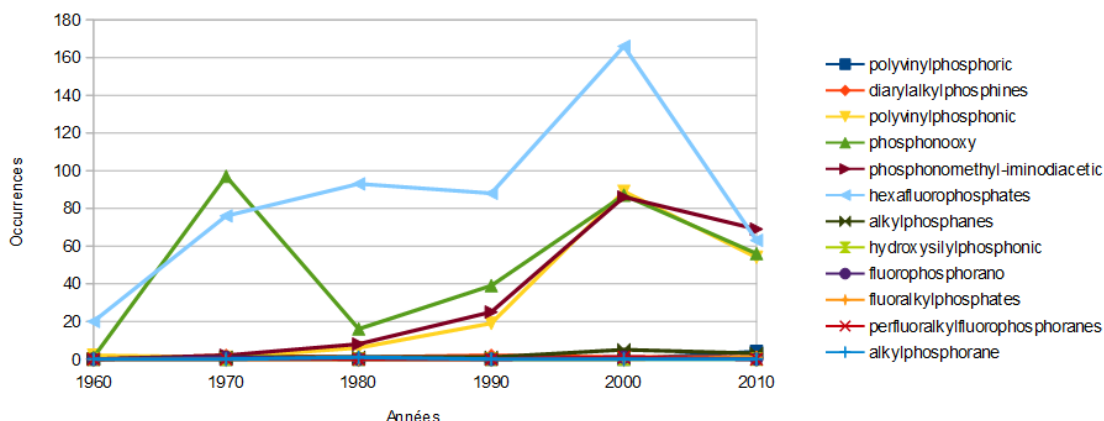


FIGURE 6.9 – Fréquence des termes sur SciFinder

Cette évaluation permet de constater que notre méthode repère les termes valides à faible fréquence, dont les termes qui viennent d'apparaître dans le domaine.

6.4 Discussion

Dans ce chapitre, nous avons présenté deux expériences liées à l'utilisation des familles morphologiques endogènes : leur exploitation pour présenter les résultats de l'extraction terminologique multilingue et l'identification des néologismes d'un domaine de spécialité.

Nous allons maintenant aborder les points qui nous semblent importants pour la compréhension de notre méthode et de ses résultats. Notamment, nous allons présenter la structure des familles morphologiques obtenues. Ensuite, nous allons discuter la pertinence des candidats faisant partie des familles morphologiques endogènes. L'organisation des résultats multilingues est également importante pour notre projet. Nous allons revenir encore une fois sur la place des termes candidats à faible fréquence dans les résultats de notre système. Enfin, nous proposerons des idées concernant l'interface utilisateur et les limites de notre approche.

6.4.1 Structure des familles morphologiques endogènes

Les familles morphologiques résultant de notre traitement diffèrent en fonction du terme de l'ontologie par rapport auquel elles sont constituées. Elle peuvent être plus ou moins peuplées. Du point de vue structurel, c'est donc l'élément de tête, donc le terme de l'ontologie, et la taille qui méritent d'être discutés.

6.4.1.1 Choix d'élément de tête pour les familles

Au début de travail sur notre projet, nous avons prévu de construire des familles à partir des n-grammes de caractères. Par exemple, on pourrait regrouper ainsi les termes

6.4. DISCUSSION

contenant le préfixe *poly-* ou encore le n-gramme *chlor* correspondant au radical *chlore*.

Toutefois, cette idée s'est avérée peu utile. Les n-grammes ne correspondent pas strictement aux morphèmes, ce qui ne constitue pas d'obstacle dans nos traitements, mais est perturbant à l'étape de présentation des résultats.

Pour cette raison, nous avons opté pour la construction des familles par rapport aux termes de l'ontologie ou par rapport aux termes candidats. La première option est utile lorsque l'utilisateur veut ajouter des termes ou des concepts similaires à ceux déjà connus par l'ontologie. Il s'agit alors de l'élargissement de l'ontologie en question. La deuxième option est plutôt destinée à traiter les termes candidats restants en vue d'ajout des nouveaux concepts dans l'ontologie (son enrichissement).

6.4.1.2 Taille des familles

La taille d'une famille dépend de la structure du terme de l'ontologie qui se trouve à sa tête. Si ce terme est composé de morphèmes ou des éléments lexicaux fréquemment utilisés dans le domaine en question, par exemple, *-chlor-*, *photo-*, *poly-*, etc., les familles ont tendance d'avoir plusieurs membres et branches. Le rôle de l'interface utilisateur est alors important pour faciliter leur traitement. Si, au contraire, le terme est un mot simple, par exemple, *pluie*, sa famille contient moins de candidats (un candidat, *eau de pluie*, en l'occurrence).

6.4.2 Pertinence des familles morphologiques endogènes

Dans cette évaluation, la pertinence des candidats est jugée du point de vue pragmatique, c'est-à-dire, par rapport à leur utilité dans le processus d'alimentation d'une ontologie.

photo-carte :

- *photocatalyse :*
 - *photocatalyse pour dégrader*
 - *photocatalyseur :*
 - revêtement de photocatalyseur
- *photoactivé*
- *photocatalytique :*
 - *destruction photocatalytique :*
 - *destruction photocatalytique de contaminants*
- *photonique :*
 - *cristaux photonique*
- *carte :*
 - *carter*

Cette famille illustre un exemple type de nos résultats. Elle regroupe les candidats liés à son élément *photo-*, comme *photocatalyse*, *photoactivé*, etc., et le candidat *carte*. Ainsi, sans segmentation explicite, le système arrive à proposer des candidats pour les morphèmes et les éléments lexicaux constituant les mots composés. L'utilisateur peut donc se concentrer sur les éléments relatifs à la lumière (*photo-*) pour alimenter la ressource cible.

accéléromètre :

- *micromètre*
- *accélération* :
 - *accélération du processus biocatalytiques*
- *kilomètre*

Prenons l'exemple de la famille du terme *accéléromètre* (ci-dessus). Nous pouvons constater que deux candidats, *micromètre* et *kilomètre*, sont des unités de mesure, tandis que le terme *accéléromètre* désigne un outil de mesure. L'utilisateur peut décider d'ajouter ces candidats dans l'ontologie (si cela n'est pas encore le cas) ou de se concentrer uniquement sur le candidat *accélération*. Il faut également remarquer que les candidats *micromètre* et *kilomètre* appartiennent également à d'autres familles : *microbiologie et climat microthermique* pour *micromètre* et *photomètre* pour *kilomètre*. Ce deuxième exemple illustre le cas, où la famille contient des résultats qui ne sont pas sémantiquement liés au terme de l'ontologie, mais contiennent des éléments communs avec celui-ci. En l'occurrence, il s'agit d'une même racine *mètre* qui peut avoir deux sens : une unité de mesure, comme dans *kilomètre* et un appareil de mesure (*accéléromètre*). Distinguer les deux sens est très compliqué à cette étape-ci. Une validation humaine est donc indispensable.

6.4.3 Aspect multilingue des familles morphologiques

Les familles morphologiques sont construites après l'extraction des termes candidats et l'identification des candidats traductions. Nous avons donc à cette étape les informations sur les traductions des candidats faisant partie d'une famille.

Ces traductions peuvent être présentées à l'utilisateur sur une interface dédiée, car leur inclusion dans les familles rend la lecture des résultats plus difficile :

oxyde de magnésium :

- *magnésium* (*magnesium@en, magnesium@de*) :
- *sel de magnésium* (*magnesium salt@en, magnesiumsalze@de*)
- *magnésium* *nh₄mgpo₄*
- *magnétoscope*
- *magnétique* (*magnetismus@de, magnetic@en*), etc.

Dans cet exemple, les langues des traductions sont indiquées par l'annotation @lang. Étant donné qu'un candidat peut avoir plus d'une traduction possible, cette présentation risque d'être trop lourde.

6.4.4 Identification des termes à faible fréquence

Les hapax, des mots ayant une seule occurrence dans le corpus, sont souvent considérés comme bruit dans les résultats de l'extraction terminologique. Toutefois, ces mots peuvent contenir des termes candidats valides, surtout si le corpus est de taille modeste.

Traiter ces termes candidats à faible fréquence dans une liste de termes candidats est assez coûteux en matière de temps, car la part de bruit peut être effectivement importante. Les familles morphologiques présentent alors un moyen efficace d'intégration de ces

6.5. CONCLUSION

candidats. En effet, même si un terme candidat, par exemple, *5-chlore-2-méthylthio-6,7,8,9-tétrahydro-1,2,4-1,2,4-triazolo* a une seule occurrence dans le corpus, il peut être validé au sein d'une des familles dont il fait partie. Dans le cas de cet exemple, le terme candidat se trouve dans les familles des termes *chlore*, *chlorinité*, *méthane* et *chlorosité*.

Nous avons également démontré que les familles morphologiques peuvent servir à l'identification des néologismes dans des domaines de spécialité. Cette fonctionnalité est intéressante en vue de la maintenance d'une ontologie, car elle permet de proposer des nouveaux termes et concepts dès leur apparition.

6.4.5 Fonctionnalités de l'interface utilisateur

Les familles morphologiques endogènes sont organisées sous forme d'arborescences qui peuvent atteindre une taille assez importante. Le confort utilisateur dépend alors de l'interface qui doit être conçue pour leur exploitation¹³. En effet, si l'interface permet d'ouvrir ou de fermer certaines branches, l'utilisateur peut naviguer dans les données et se focaliser sur une branche en particulier.

Il serait également utile de fournir des traductions proposées pour les termes d'une famille afin d'élargir l'ontologie simultanément en plusieurs langues. Inclure ces informations dans l'arborescence pourrait l'alourdir. Il conviendrait donc d'opter pour une fenêtre pop-up ou pour une zone spécifique, comme c'est le cas dans TTC TermSuite.

Enfin, l'interface en question doit permettre de manipuler l'ontologie : créer, modifier et supprimer des concepts, des relations et des termes. Ces opérations doivent être corrélées avec les familles morphologiques de manière ergonomique.

6.4.6 Limites d'application

La méthode décrite dans cette partie a ses avantages, comme l'indépendance à la langue, à la taille du corpus et à son domaine. Cependant, nous devons constater que l'efficacité de notre approche est maximale lorsque le domaine a une nomenclature bien développée qui emploie un certain nombre de composants savants ou lorsque la langue en elle-même a tendance à construire un grand nombre de mots composés, comme le fait l'allemand.

Les résultats des familles morphologiques dépendent de la structure du terme et de son potentiel néologique. Dans tous les cas, une validation manuelle est nécessaire. Ce traitement a pour vocation de faciliter le travail de l'expert humain, et non pas de le remplacer.

6.5 Conclusion

Dans ce chapitre, nous avons présenté la méthode de construction des familles morphologiques endogènes. La méthode a recours aux n-grammes de caractères appris de la

13. Nous tenons à préciser que cette interface n'a pas été développée dans le cadre du projet de thèse et fera l'objet du développement ultérieur.

6.5. CONCLUSION

ressource terminologique à enrichir (une ontologie du domaine, un dictionnaire ou des résultats de l'extraction terminologique) et du corpus pour l'identification des familles morphologiques. Comme ces n-grammes sont acquis directement à partir du corpus analysé sans apport extérieur, nous parlons de « morphologie endogène ».

Les familles morphologiques endogènes peuvent être construites pour les termes de l'ontologie source ou pour les termes candidats issus de l'extraction terminologique. Elle varie en taille en fonction de la longueur des n-grammes utilisés dans la construction, mais aussi selon la structure du terme de l'ontologie. Ainsi, les termes composés de formants néo-classiques rassemblent plus de candidats dans leurs familles.

Les familles morphologiques endogènes offrent un cadre pratique pour la détection des néologismes dans des corpus de spécialité. Nous avons démontré l'évolution des familles dans le temps en comparant les familles obtenues à partir du corpus des années 1980 avec celui des années 2000. La différence entre les résultats illustre la capacité de l'approche à suivre l'évolution de la terminologie dans le temps.

Nous avons relevé des cas où les familles contiennent des termes sémantiquement proches qui correspondent à nos attentes et permettent d'alimenter une ontologie par des nouveaux termes et concepts. Toutefois, il existe également des familles basées sur une ressemblance partielle, par exemple, partageant un même suffixe, qui n'apportent pas de forte valeur ajoutée. À ce stade, il est difficile de distinguer ces cas avant la validation par un expert humain. Dépasser cette limite serait un objectif des futurs développements.

Les limites des familles morphologiques sont très proches de celles évoquées dans d'autres étapes de traitement. La méthode fonctionne mieux lorsqu'il existe une nomenclature du domaine en question. Bien qu'elle soit applicable à une très large palette de langues, il faut tenir compte de la segmentation en n-grammes de caractères. Ainsi, le traitement peut être plus compliqué pour des langues asiatiques utilisant des idéogrammes.

Chapitre 7

Enrichissement et élargissement de l'ontologie à l'aide des familles morphologiques endogènes

Dans les chapitre précédents, nous avons détaillé les étapes principales de notre méthode. Chaque étape a été présentée dans son contexte, mais évaluée indépendamment des autres étapes. En effet, dans les flux de traitement, la qualité des résultats d'une étape influence significativement celle de l'étape suivante. Par exemple, si les résultats de l'extraction terminologique ne sont pas suffisamment bons, la recherche de traductions pour les termes et la construction des familles morphologiques ne pourront pas atteindre leurs performances maximales.

Nous avons eu recours à plusieurs ressources, ce qui présente l'avantage de confirmer l'applicabilité de la méthode à plusieurs domaines, corpus et des ressources à taille variable. Toutefois, cela génère un volume important de résultats à évaluer.

L'objectif de ce chapitre final est de reprendre l'ensemble de traitements présenté au début de cette partie et les appliquer à une ontologie et à un corpus en trois langues pour illustrer un cas d'utilisation par un expert humain. Nous avons déjà évoqué que l'organisation des résultats influence le travail de l'utilisateur en termes d'efficacité, qui se traduit par le temps et le coût, et du confort.

7.1 Élargissement et enrichissement d'une ontologie

Les processus d'élargissement et d'enrichissement d'une ontologie à partir d'un corpus nécessitent plusieurs étapes que nous avons déjà décrites dans les chapitres précédents. Néanmoins, il est utile de les reprendre dans l'ordre afin de reconstruire le flux des traitements et en rappeler les objectifs :

1. pré-traitement du corpus ;
2. extraction des termes mono et polylexicaux ;
3. recherche des traductions ;

4. construction des familles pour les libellés de l'ontologie.

Nous allons illustrer ci-après ces étapes sur des exemples tirés du corpus en régénération des sols pollués (sous-corpus B09C du PatTR).

7.1.1 Pré-traitement du corpus

Le développement de notre méthode a été mené sur les données en trois langues. Nous allons donc prendre une phrase exemple par langue pour illustrer les évolutions au cours des traitements.

Avant tout, le corpus mono ou multilingue doit être tokénisé (1), puis lemmatisé (2) et étiqueté par des mots informatifs et vides (3) selon la méthode de Vergne (2005).

FR :

L'invention est particulièrement efficace pour assainir les biphényles polychlorés.

- (1) L' invention est particulièrement efficace pour assainir les biphényles polychlorés .
- (2) l^a invention être particulièrement efficace pour assainir le biphényles polychloré .
- (3) l_(n)^b invention_(n) être_(n) particulièrement_(I) efficace_(n) pour_(n) assainir_(I) le_(n) biphényles_(I) polychloré_(I) ._(n)

a. Aucun traitement n'étant parfait, des erreurs peuvent être introduites à des différentes étapes. Ici, Le lemmatiseur n'a pas pu transformer *L'* en *le*, ni deviner le lemme *biphényle*, gardant sa forme erronée au pluriel.

b. Rappelons que *n* est l'étiquette d'un mot non-informatif et *I* – de celui informatif.

EN :

A method for remediating matter by removing phosphorus and oxyanion pollutants therefrom, the matter being such as sediments in waterways and catchments, or effluent from sewage treatment plants, [...]

- (1) A method for remediating matter by removing phosphorus and oxyanion pollutants therefrom , the matter being such as sediments in waterways and catchments , or effluent from sewage treatment plants , [...]
- (2) a method for remediating matter by remove phosphorus and oxyanion pollutant therefrom , the matter be such as sediment in waterway and catchment , or effluent from sewage treatment plant , [...]
- (3) a_(n) method_(I) for_(n) remediating_(I) matter_(I) by_(n) remove_(I) phosphorus_(I) and_(n) oxyanion_(I) pollutant_(I) therefrom_(I) ,_(n) the_(n) matter_(I) be_(n) such_(I) as_(n) sediment_(I) in_(n) waterway_(I) and_(n) catchment_(I) ,_(n) or_(n) effluent_(I) from_(n) sewage_(I) treatment_(I) plant_(I) , [...]

DE :

Die Erfindung betrifft ein Verfahren und eine Vorrichtung zur Aufbereitung und Dekontamination von Material, vorzugsweise von Sedimentationsablagerungen vom Grund eines Gewässers oder Ölabscheidern, Sandfängen oder dergleichen.

(1) Die Erfindung betrifft ein Verfahren und eine Vorrichtung zur Aufbereitung und Dekontamination von Material , vorzugsweise von Sedimentationsablagerungen vom Grund eines Gewässers oder Ölabscheidern , Sandfängen oder dergleichen .

(2) die erfindung betreffen ein verfahren und eine vorrichtung zu aufbereitung und dekontamination von material , vorzugsweise von sedimentationsablagerungen von grund eine gewässer oder ölabscheidern , sandfängen oder dergleichen .

(3) die_(n) erfindung_(I) betreffen_(I) ein_(n) verfahren_(n) und_(n) eine_(n) vorrichtung_(n) zu_(n) aufbereitung_(I) und_(n) dekontamination_(I) von_(n) material_(n) ,_(n) vorzugsweise_(I) von_(n) sedimentationsablagerungen_(I) von_(n) grund_(I) eine_(n) gewässer_(I) oder_(n) ölabscheidern_(I) ,_(n) sandfängen_(I) oder_(n) dergleichen_(I) ._(n)

Les étapes de la tokénisation et de la lemmatisation permettent de réduire le nombre de mots (tokens ou lemmes) à traiter. Sans tokénisation qui sépare les mots des signes de ponctuation, le mot *polychlorés* pourrait être considéré comme un élément différent de *polychlorés*. se terminant par un point. Une simple suppression de la ponctuation ne constitue pas de solution fiable, car les règles de tokénisation diffèrent d'une langue à autre. Par exemple, en français le point doit rester intacte lorsqu'il s'agit d'une forme abrégée de *Monsieur*, comme dans *M. Untel*. Bien que la tokénisation nécessite des règles spécifiques à une langue, ces outils sont relativement simples à concevoir et sont disponibles pour plusieurs langues. Pour cette raison, nous considérons que le recours à un tokéniseur ne nuit pas au caractère multilingue de notre système.

La lemmatisation permet d'identifier la forme de base pour chaque mot, par exemple *polychlorés* → *polychloré*, *est* → *être*. Cela contribue à la réduction du nombre de formes à traiter : le sous-corpus C02F en français contient plus de 18 millions de tokens qui constituent plus de 51 mille de formes uniques (types). La lemmatisation permet de réduire le nombre d'éléments à traiter à environ 36 mille lemmes. Les lemmatiseurs sont des outils plus complexes qui nécessitent des modèles pour chaque langue de traitement. Ces modèles s'appuient sur des vocabulaires de langue générale et n'arrivent pas toujours à bien lemmatiser les termes d'un domaine, car ces derniers sont absents de leurs modèles. Les outils de lemmatisation sont liés à l'étiquetage morphosyntaxique (l'identification de la partie du discours d'un mot) et ne sont pas disponibles pour toutes les langues. Nous avons évoqué la possibilité d'appliquer notre méthode sur des corpus tokénisés. Cette option, intéressante dans le cadre de travail avec des langues peu dotées en ressources, entraîne toutefois une baisse de rappel et de précision.

La troisième étape, l'étiquetage par des mots informatifs et vides, permet d'exclure des mots grammaticaux et de langue générale qui ne sont pas des termes candidats. Ainsi, les futures étapes de traitement se focalisent sur les lemmes étiquetés comme informatifs. Les mots vides peuvent faire partie des termes polylexicaux en tant que mots de schéma (Enguehard, 1993). Cet étiquetage n'utilise pas de ressources ni d'outils complémentaires et

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

peut s'effectuer sur des textes bruts (de préférence, transformés en minuscules), tokenisés ou lemmatisés. Étant donné que l'étiquetage tient compte de la fréquence des mots, il est plus performant sur des corpus lemmatisés. L'étiquetage peut contenir des erreurs, par exemple en anglais certains mots vides sont étiquetés comme informatifs (*therefrom* et *such*). Cela peut résulter de la longueur du mot *therefrom* (à partir de) et de sa fréquence relativement basse (120 occurrences). Le mot *such* est nettement plus fréquent (900 occurrences), mais il est systématiquement entouré par des éléments très courts et très fréquents (virgule et le mot *as*). En allemand, le problème peut être inverse : certains mots (par exemple, *material*, matière) qui doivent être étiquetés comme informatifs, ne le sont pas à cause de la longueur des mots généralement importante. Si un mot vide est étiqueté comme informatif, cela peut ajouter des faux résultats dans les termes candidats monolexicaux (par exemple, le mot allemand *vorzugsweise* qui signifie *de préférence*), ou encore *dergleichen* (*comme*). Dans le cas contraire, un terme mal étiqueté comme un mot vide n'apparaîtra pas dans les résultats d'extraction terminologique. Cela peut concerner des termes importants, par exemple *eau* et *sol* en français, car ces mots sont courts et fréquents. Il est possible de prévoir une modification de l'étiquetage par l'utilisateur (mais cette fonctionnalité ne fait pas partie de notre thèse).

Après ces étapes du pré-traitement, le corpus est prêt à être exploité pour l'extraction de termes.

7.1.2 Extraction des termes

À cette étape, nous avons besoin d'un corpus pré-traité et d'une ontologie à élargir. Dans ce chapitre, l'évaluation est faite sur l'ontologie EcoLexicon. Le traitement se fait en deux étapes : d'abord, nous identifions des candidats monolexicaux et, ensuite, polylexicaux (composés de plusieurs mots).

7.1.2.1 Extraction des termes monolexicaux

Lors de l'extraction des termes monolexicaux, les termes déjà connus de l'ontologie sont utilisés pour générer des ressources morphologiques endogènes. Plus précisément, ces termes sont segmentés en n-grammes de caractères, dont on garde ceux du début et du milieu de mots. La longueur de ces n-grammes est paramétrable. Nous avons déjà démontré (Section 4.3.4.1) que les configurations où $n = 3$ ou $n = 4$ sont les plus équilibrées en termes de rappel, de précision et de F-mesure. Dans notre exemple, nous fixons la longueur $n = 4$.

Reprenons les phrases en trois langues issues du pré-traitement :

FR : l_(n) invention_(n) être_(n) particulièrement _(I) efficace_(n) pour_(n) assainir_(I) le_(n) biphényles _(I) polychloré _(I) ._(n)

EN : a_(n) method _(I) for_(n) remediating _(I) matter _(I) by_(n) remove _(I) phosphorus _(I) and_(n) oxyanion _(I) pollutant _(I) therefrom _(I) ,_(n) the_(n) matter _(I) be_(n) such _(I) as_(n) sediment _(I) in_(n) waterway _(I) and_(n) catchment _(I) ,_(n) or_(n) effluent _(I) from_(n) sewage _(I) treatment _(I) plant _(I) , [...]
--

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

DE : die_(n) **erfindung_(I)** **betreffen_(I)** ein_(n) verfahren_(n) und_(n) eine_(n) vorrichtung_(n) zu_(n) **aufbereitung_(I)** und_(n) **dekontamination_(I)** von_(n) material_(n) ,_(n) **vorzugsweise_(I)** von_(n) **sedimentationsablagerungen_(I)** von_(n) **grund_(I)** eine_(n) **gewässer_(I)** oder_(n) **ölabscheiden_(I)** ,_(n) **sandfängen_(I)** oder_(n) **dergleichen_(I)** ._(n)

En français, la recherche des termes monolexicaux se limite à quatre mots : *particulièrement*, *assainir*, *biphényles* et *polychloré*. Tous les quatre mots sont identifiés comme des termes candidats.

terme candidat FR	N-gramme	Score du n-gramme	Score candidat
particulièrement	ière	0,19	0,01
	emen	0,06	
	icul	0,02	
assainir	sain	0,03	0,01
	ssai	0,03	
biphényles	phén	0,40	0,04
polychloré	chlo	0,20	0,03
	hlor	0,20	

TABLE 7.1 – N-grammes permettant d'extraire les termes candidats

Nous pouvons constater que certains n-grammes, comme *phén*, *chlo* et *hlor* ont des scores supérieurs et sont donc plus spécifiques au domaine en question. Ces n-grammes permettent de mieux noter les candidats qui les contiennent, à savoir *biphényles* et *polychloré*. Le mot *particulièrement* fait partie du bruit de cette étape de traitement à cause de sa longueur qui permet d'y retrouver plusieurs n-grammes pris en compte. Le terme *assainir* contient deux n-grammes du modèle, mais leurs scores sont moins importants que ceux des termes *biphényles* et *polychloré*.

En anglais, les termes *method*, *matter*, *pollutant*, *sediment*, *waterway*, *effluent*, *sewage* et *plant* sont déjà présents dans EcoLexicon. En effet, EcoLexicon contient plus de 4 mille libellés en anglais, contre environ 600 en français. Les termes candidats repérés par notre système sont *phosphorus*, *oxyanion*, *therefrom*, *catchment*, *treatment* et *remediating*. Le mot mal étiqueté *therefrom* (à *partir de*) est un candidat à éliminer. Les termes candidats *phosphorus* et *oxyanion* sont repérés grâce à leur origine néo-classique, tandis que le candidat *catchment* (*bassin versant*) est identifié grâce aux n-grammes générés à partir des deux termes présents dans EcoLexicon qui le contiennent : *catchment area* et *pilot catchment area*. Il est de même pour *treatment* (*traitement*) qui fait partie de plusieurs termes polylexicaux. Il faut remarquer que nous ne considérons pas chaque mot faisant partie des libellés polylexicaux comme un terme. En effet, il serait difficile de distinguer entre les termes, comme c'est le cas de *catchment* et des mots vides (comme *of*, *for*, *and*, etc. Les mots *remove*, *such* ne sont pas inclus dans les termes candidats.

En allemand, les termes candidats suivants ont été identifiés grâce à leur présence dans les libellées composés dans EcoLexicon : *aufbereitung* (*traitement*), *grund* (*souterrain*) et

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

gewässer (eaux). Les termes candidats sont *erfindung* (invention), *betreffen* (affecter), *dekontamination* (décontamination), *vorzugsweise* (de préférence), *sedimentationsablagerungen* (dépôt de sédiment), *ölabscheidern* (séparateurs d'huile) et *sandfängen* (dessableurs). Nous pouvons noter que les mots *betreffen* (affecter) et *vorzugsweise* (de préférence) ne peuvent pas être validés en tant que termes. Le terme candidat est certainement identifié grâce à la présence du terme *kontamination* dans EcoLexicon. *Ölabscheidern* (séparateurs d'huile) a été trouvé à l'aide des n-grammes du libellé *fettabscheider* (séparateur de graisse) présent dans EcoLexicon. Il faut remarquer que notre système permet d'utiliser les termes composés de l'ontologie sans segmentation en mots, ce qui évite l'utilisation des outils et des ressources complémentaires.

Dans cet exemple, le n est fixé à 4. Il est toutefois possible d'utiliser les longueurs au-delà de cette valeur. Dans les chapitres précédents, nous avons démontré que l'augmentation de la longueur n entraîne une hausse de précision, mais diminue le rappel et, par conséquent, la F-mesure. Selon notre hypothèse, l'augmentation de n entraîne les mêmes conséquences à toutes les étapes de traitement. La solution envisagée dans les perspectives du projet est de configurer un processus à plusieurs itérations qui partirait de la valeur n maximale (par exemple, par apport au mot le plus long du corpus ou de l'ontologie), en la réduisant à chaque itération. De cette façon, la validation de termes peut être facilitée et accélérée. C'est notamment cette approche que nous avons utilisée lors de l'annotation du corpus par les termes pour les évaluations précédentes.

Dans ce chapitre, nous ne testons pas ce mode itératif, car l'objectif visé est d'évaluer l'efficacité des familles morphologiques. L'économie du temps et la meilleure efficacité doivent alors être atteintes via l'organisation des résultats dans ces familles.

Une fois les termes candidats monolexicaux sont extraits, nous pouvons procéder à la recherche des termes polylexicaux.

7.1.2.2 Extraction des termes polylexicaux

L'extraction des termes polylexicaux se base sur les patrons morphosyntaxiques endogènes appris à partir du corpus. Ces patrons sont composés des mots informatifs et vides, comme dans la méthode de Vergne (2004, 2005), par exemple, *I n I*, où deux mots informatifs sont séparés par un mot vide. Afin de rendre ces patrons plus précis, nous avons introduit les conditions complémentaires :

- un patron doit commencer et se terminer par un mot informatif ;
- la longueur du patron est paramétrable et est limitée à quatre éléments dans nos expériences.

Ensuite, les étiquettes des mots informatifs sont remplacées par les n-grammes de fin de mot. En reprenant le patron *I n I*, nous retrouvons dans les patrons identifiés dans le corpus les exemples comme *ent n ion*, *édé n ion*, *yen n ion*, etc. Dans ces trois cas, la majorité des candidats regroupés sous ces patrons correspondent à des syntagmes nominaux du type nom-préposition-nom : *agent de oxydation*, *procédé de séparation* ou encore *moyen de alimentation*. Cela permet de rapprocher ces patrons de séquences des étiquettes morphosyntaxiques habituelles (NOM-PREP-NOM) (Daille, 2003; Weller et Blancafort, 2011; Morin et Daille, 2012; Bourigault et Fabre, 2000; Bourigault, 2002; Frantzi *et al.*, 2000;

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

Orobinska *et al.*, 2013).

Ensuite vient l'application de ces patrons sur le corpus. À cette étape, deux restrictions peuvent être appliquées dans l'objectif d'obtenir une liste plus précise de termes candidats polylexicaux. Premièrement, le candidat doit commencer et se terminer par un terme ou un terme candidat monolexical. Cette restriction vise à limiter le nombre de termes candidats polylexicaux qui peut devenir assez important, en améliorant en même temps la qualité du résultat.

Deuxièmement, nous pouvons limiter les mots vides à l'intérieur des termes polylexicaux pour viser certaines structures, par exemple, des syntagmes nominaux. Les mots vides qui jouent le rôle des mots de schéma (Enguehard, 1993), peuvent être limités à quelques uns les plus fréquents (excluant la ponctuation). Nous avons constaté que, pour être à la fois efficaces et précis, les mots de schéma doivent permettre l'identification du génitif (Section 4.3.4.3), ce qui correspond à la préposition *de* (FR) et à *of* (EN), éventuellement suivies d'articles. Il est également intéressant d'inclure certains groupes prépositionnels, comme *pompe à chaleur* ou encore *procédé pour la dégradation photocatalytique*.

Enfin, nous identifions des termes candidats contenus dans des termes candidats plus longs et comparons leurs fréquences dans le corpus. Lorsque leurs fréquences sont égales, seul le candidat plus long est retenu. Considérons de nouveau les exemples en trois langues :

FR : l_(n) invention_(n) être_(n) particulièrement_(I) efficace_(n) pour_(n) assainir_(I) le_(n) biphényles_(I) polychloré_(I) ._(n)

EN : a_(n) method_(I) for_(n) remediating_(I) matter_(I) by_(n) remove_(I) phosphorus_(I) and_(n) oxyanion_(I) pollutant_(I) therefrom_(I) ,_(n) the_(n) matter_(I) be_(n) such_(I) as_(n) sediment_(I) in_(n) waterway_(I) and_(n) catchment_(I) ,_(n) or_(n) effluent_(I) from_(n) sewage_(I) treatment_(I) plant_(I) , [...]

DE : die_(n) erfindung_(I) betreffen_(I) ein_(n) verfahren_(n) und_(n) eine_(n) vorrichtung_(n) zu_(n) aufbereitung_(I) und_(n) dekontamination_(I) von_(n) material_(n) ,_(n) vorzugsweise_(I) von_(n) sedimentationsablagerungen_(I) von_(n) grund_(I) eine_(n) gewässer_(I) oder_(n) ölabscheidern_(I) ,_(n) sandfängen_(I) oder_(n) dergleichen_(I) ._(n)

L'extraction des termes polylexicaux à partir de l'exemple français (ci-dessus), résulte en deux candidats : *biphényles polychloré* et *assainir le biphényles polychloré*. Le candidat *assainir le biphényles* a été éliminé lors de la dernière vérification. En effet, compte tenu de la faute d'orthographe dans le mot *biphényles*, c'est la seule occurrence de cette forme dans le corpus. Par conséquent, *assainir le biphényles* a perdu face à *assainir le biphényles polychloré* qui est un terme plus long.

En anglais, le système a identifié le candidat *sewage treatment plant* (*station d'épuration*) correspondant à la séquence des trois mots informatifs. Nous n'avons pas effectué de recherche de termes polylexicaux en allemand, car les termes monolexicaux composés correspondent déjà à des termes polylexicaux en anglais et en français.

Les candidats mono et polylexicaux extraits à partir de corpus sont présentés sous forme d'une liste commune qui est ensuite exploitée pour la recherche des traductions. En

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

effet, un terme monolexical en langue source peut être traduit par un terme polylexical en langue cible. La liste obtenue pour notre exemple en français contient quatre candidats monolexicaux et deux candidats polylexicaux : *particulièrement*, *assainir*, *biphényles*, *polychloré*, *biphényles polychloré* et *assainir le biphényles polychloré*. En anglais, elle contient les candidats *phosphorus*, *oxyanion*, *therefrom*, *catchment*, *treatment*, *remediating* et *sewage treatment plant*; en allemand – *erfindung* (*invention*), *betreffen* (*affecter*), *dekontamination* (*décontamination*), *vorzugsweise* (*de préférence*), *sedimentationsablagerungen* (*dépôt de sédiment*), *ölabscheidern* (*séparateurs d'huile*) et *sandfängen* (*dessableurs*).

7.1.3 Recherche des traductions

Une fois que les termes candidats sont identifiés pour chaque langue, nous procédons à la recherche des traductions, appelée également l'alignement des terminologies. Le défi principal à cette étape est l'utilisation des corpus comparables. En effet, les corpus parallèles sont rarement disponibles pour des domaines de spécialité et leur taille peut être assez modeste. Notre méthode fonctionne aussi bien pour les corpus comparables que pour ceux parallèles, car les termes sont extraits avant cette étape.

La recherche des traductions s'appuie sur les alignements bilingues au niveau des n-grammes de caractères. Ces alignements sont générés à partir des traductions disponibles dans l'ontologie source (ou dans un dictionnaire spécialisé, si l'ontologie ne dispose pas de termes en plusieurs langues). Les termes de l'ontologie en langues source et cible sont segmentés en n-grammes de caractères qui sont ensuite alignés à l'aide d'Anymalign (Lardilleux et Lepage, 2008, 2009a,b).

Ces alignements bilingues permettent d'identifier les candidats traductions dans les listes de termes candidats, ainsi que de proposer des traductions pour les termes de l'ontologie qui ne sont pas encore traduits. Les alignements des n-grammes ont des scores de probabilité calculés par Anymalign, ce qui permet de calculer les scores pour les candidats traductions et d'en proposer les plus probables.

terme candidat FR	Candidats traductions	
	EN	DE
particulièrement	particularly	<i>fausse traduction</i>
assainir	<i>fausses traductions</i>	<i>fausse traduction</i>
biphényles	halobiphenyls polyhalogenatedbiphenyls	biphenyle biphenyl
polychloré	polychlorinate polychlorinate biphenyl	polychlorierte polychlorierten polychlorierter
biphényles polychloré	polychlorinate biphenyl	polychlorbiphenyl polychlorobiphenyl
assainir le biphényles polychloré	polychlorinate biphenyl	polychlorbiphenyl polychlorobiphenyl

TABLE 7.2 – Traductions pour les termes candidats français

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

La Table 7.2 illustre les résultats de recherche des traductions pour les termes candidats français extraits à l'étape précédente. Certaines traductions candidates, dont celles des mots *particulièrement* et *assainir*, ont été fausses. Le candidat *biphényles polychloré* a retrouvé deux traductions exactes en allemand : *polychlorbiphenyl* et *polychlorobiphenyl*.

terme candidat EN	Candidats traductions	
	FR	DE
phosphorus	phosphore	phosphor
oxyanion	oxyanion	oxyanion
therefrom	<i>fausses traductions</i>	<i>fausses traductions</i>
catchment	<i>fausses traductions</i>	<i>fausses traductions</i>
treatment	traitement	wasserdampf-behandlung
remediating	remédiation	phyto-remediation
		bioremediation
		schlammphasen-bioremediations-behandlung

TABLE 7.3 – Traductions pour les termes candidats anglais

Pour les termes candidats anglais (Table 7.3), le système a identifié plusieurs traductions, dont certaines ont été exactes (*phosphorus*, *oxyanion*, *remediating*). Pour les termes *treatment* et *remediating*, le système a proposé plusieurs traductions en français, dont une partie sans relation avec le terme recherché. Pour les candidats *therefrom* et *catchment*, toutes les propositions ont été fausses.

En allemand (Table 7.4), Nous pouvons observer les traductions partielles, comme *sedimentationsablagerungen* → *sedimentation* (EN) ou encore *sandfängen* → *sandstone* (EN). Ces traductions sont proposées sans outils de segmentation morphologique ni lexicale, ce qui rend notre système intéressant pour les paires de langues où une telle segmentation est problématique à cause du manque d'outils.

Tout comme à l'étape précédente, nous envisageons dans les perspectives du projet de mettre en place des traitements itératifs afin d'ajouter des nouvelles traductions dans l'ontologie et, par conséquent, des nouveaux alignements dans le modèle de n-grammes. L'inconvénient évident de ces traitements est la nécessité de relancer l'Anymalign, et surtout de l'arrêter manuellement ¹.

Les termes candidats et leurs traductions constituent la base pour l'élargissement et l'enrichissement de l'ontologie. La prochaine étape est l'organisation de ces données en familles morphologiques endogènes.

7.1.4 Enrichissement et élargissement de l'ontologie

L'élargissement et l'enrichissement d'une ontologie sont liés et il est difficile de délimiter une frontière entre ces deux processus. Pour illustrer notre propos, reprenons l'exemple des

1. Toutefois, il doit être possible de contourner les interventions manuelles via des paramètres du module.

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

terme candidat DE	Candidats traductions	
	FR	EN
erfindung	<i>fausse traduction</i>	<i>fausses traductions</i>
betreffen	<i>fausses traductions</i>	<i>fausses traductions</i>
dekontamination	décontamination	decontamination
vorzugsweise	-	-
sedimentationsablagerungen	installation industriel de sédimentation resédimentation sédimentation	sedimentation
ölabscheiden	<i>fausses traductions</i>	<i>fausses traductions</i>
sandfängen	sandwicher	sandstone sand-sized sandwich sand-like sand-like product

TABLE 7.4 – Traductions pour les termes candidats allemands

termes candidats extraits précédemment. Dans la version d'EcoLexicon utilisée dans nos expériences qui date du 2014, aucun de ces candidats n'était présent en français. Toutefois, le terme anglais *polychlorinated biphenyl* et sa traduction allemande *polychlorierte Biphenyle* y étaient déjà. Nous allons analyser les familles dont nos termes candidats font partie. Ainsi, le terme candidat *polychloré* est présent dans les familles pour les termes suivants d'EcoLexicon : *chlorophylle*, *chlorosité*, *chlore* et *chlorinité*.

Parmi ces termes d'EcoLexicon, *chlore* semble être le plus général. Sa famille est construite comme suit :

chlore :

- *chlorobenzène*
- [...]
- *chloré* :
 - *hydrocarbure chloré*
 - *polychloré* :
 - *biphényles polychloré* :
 - *assainir le biphényles polychloré*
 - *biphényle polychloré*
 - *dichlorodiphényltrichloroéthane*
 - *chlorophosphate*, etc.

Le terme candidat *polychloré* est inclus dans la branche du terme candidat *chloré* et emmène dans sa propre branche les candidats *biphényles polychloré* et sa version bien orthographiée *biphényle polychloré*. Le candidat *assainir le biphényles polychloré* est à son tour rattaché à la branche du *biphényles polychloré*.

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

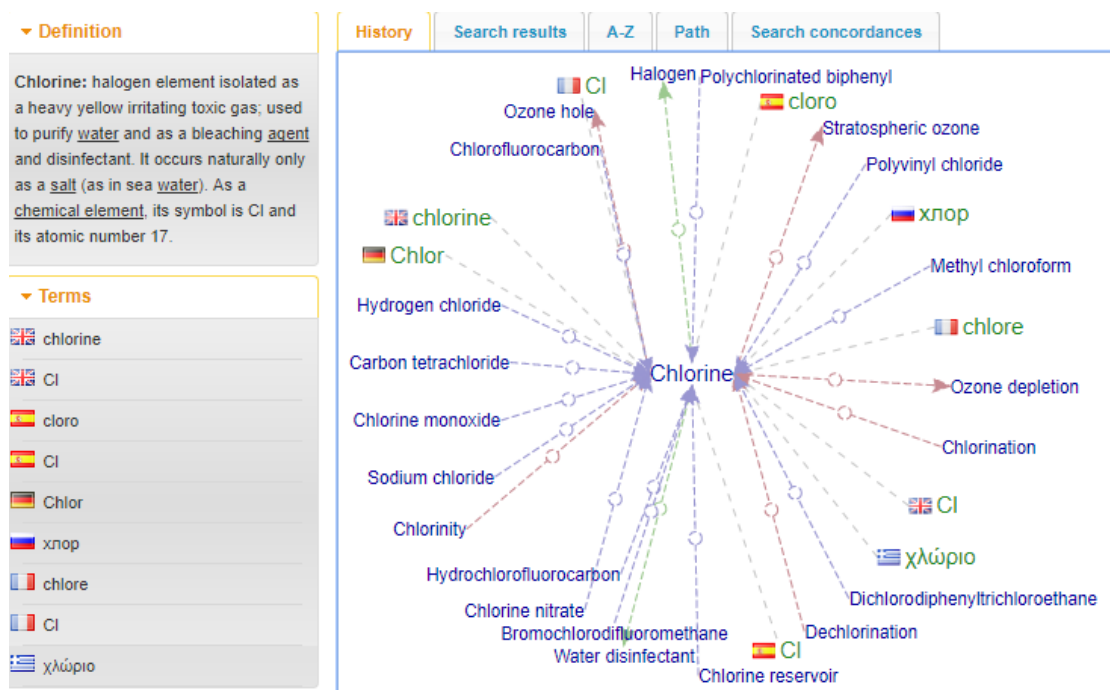


FIGURE 7.1 – Concept *CHLORINE* (chlore) dans EcoLexicon

En examinant les informations déjà disponibles dans EcoLexicon par rapport au concept *CHLORINE* (Figure 7.1), nous pouvons constater qu'un certain nombre de substances contenant du chlore, dont *POLYCHLORINATED BIPHENYL* et *DICHLORODIPHENYLTRICHLOROETHANE*, sont déjà liées avec ce concept par des relations méronymiques *made-of*. Nous pouvons donc le lier par la même relation avec autres substances présentes dans la famille (après la création des concepts correspondants), comme *chlorophosphate*.

Nous pouvons également élargir le concept *POLYCHLORINATED BIPHENYL* par deux libellés en allemand : *polychlorbiphenyl* et *polychlorobiphenyl*. Ce concept contient déjà le libellé *polychlorierte biphenyle*.

La famille en question contient également le terme *hydrocarbure chloré*. La requête d'EcoLexicon retourne le concept *CHLOROFLUOROCARBON* qui a trois traductions en français : *chlorofluorocarbure*, *hydrocarbure chlorofluoré* et *CFC*. Notre candidat *hydrocarbure chloré* (*chlorinated hydrocarbon* en anglais) est évidemment proche de ce concept et un concept similaire peut donc être créé.

Ces familles doivent mener vers des concepts du domaine en question qui manquent dans l'ontologie source. Par exemple, le terme candidat *assainir* présent dans la famille au sein du candidat *assainir le biphényles polychloré* n'a pas d'équivalent dans EcoLexicon. En même temps, l'ontologie contient le concept *PURIFICATION* (Figure 7.2). Nous pouvons donc y ajouter le concept *REMEDiate* qui correspond à *assainir* et *REMEDiation* qui est la traduction anglaise d'*assainissement*, donc l'action d'*assainir* ou le processus lié.

Le concept *PURIFICATION* est défini dans EcoLexicon comme *un traitement de l'eau*

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

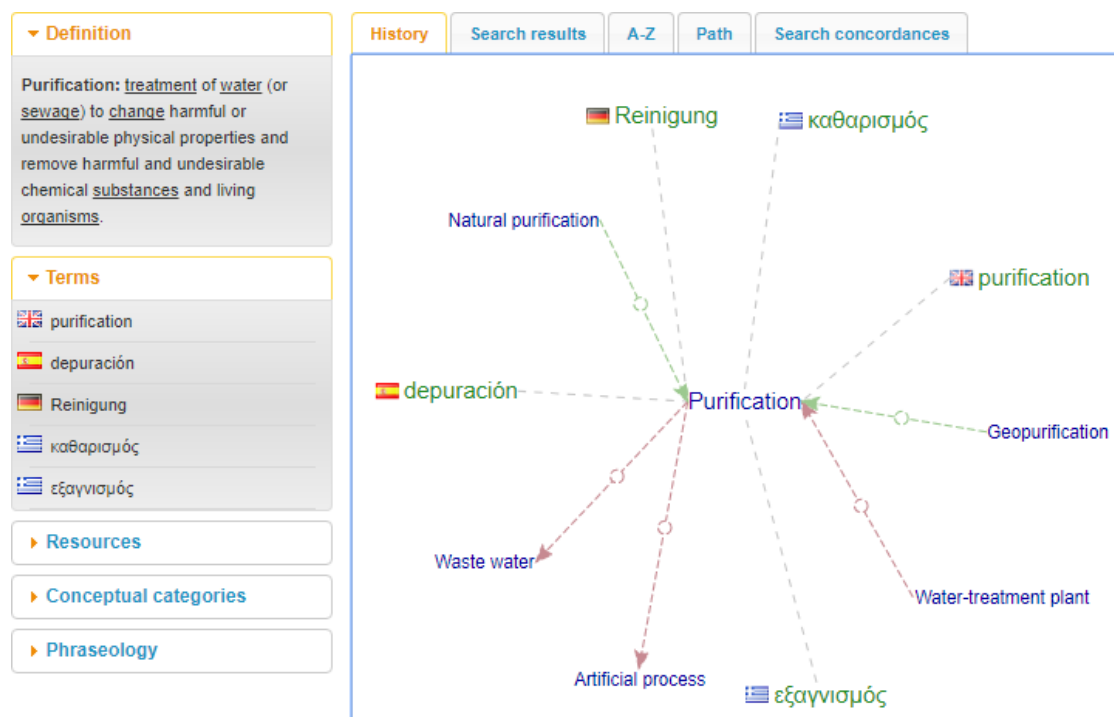


FIGURE 7.2 – Concept *PURIFICATION* dans EcoLexicon

(ou de la boue) visant à modifier ses propriétés physiques nuisibles ou indésirables et à enlever des substances chimiques et des organismes vivants nuisibles et indésirables². Selon le dictionnaire *Oxford Living Dictionaries*³, le terme *remediation* signifie l'action de réparer quelque chose, en particulier d'inverser ou d'arrêter les dommages environnementaux⁴.

Les deux sens étant très proches, nous pouvons placer le concept *REMEDIATION* au même niveau que *PURIFICATION*. Ainsi, *REMEDIATION* est également le résultat d'un processus artificiel (*result-of ARTIFICIAL PROCESS*, Figure 7.2), nous pouvons donc ajouter cette relation. Concernant le lien avec les eaux usées (*PURIFICATION affects WASTE WATER*), *REMEDIATION* semble être un synonyme moins fréquent. La requête "*waste water remediation*" fournit 46 500 occurrences sur Google⁵, contre environ 444 000 pour "*waste water purification*". Nous pouvons donc maintenir cette relation pour le concept *REMEDIATION*.

Dans notre corpus, le sens d'assainissement est souvent associé avec les sols pollués. Pour cette raison, nous avons vérifié la combinaison "*soil purification*" et "*soil remediation*" : la deuxième requête a donné 570 000 résultats, ce qui est nettement supérieur à 10 400 pour la première. EcoLexicon ne contient pas de concept *POLLUTED SOIL* qui

2. Purification : treatment of water (or sewage) to change harmful or undesirable physical properties and remove harmful and undesirable chemical substances and living organisms.

3. <https://en.oxforddictionaries.com/definition/remediation>

4. The action of remedying something, in particular of reversing or stopping environmental damage.

5. www.google.com

7.1. ÉLARGISSEMENT ET ENRICHISSEMENT D'UNE ONTOLOGIE

mériterait d'être ajouté au même titre que *WASTE WATER*, mais nous y trouvons *SOIL POLLUTION*. Il est donc possible de créer la relation *REMEDIATION affects SOIL POLLUTION*.

Dans notre exemple, nous avons identifié le terme candidat *assainir le biphényles polychloré*, qui est égal à *assainir le biphényle polychloré*. Le concept *POLYCHLORINATED BIPHENYL* est défini comme un *POLLUTANT*. Nous pouvons donc ajouter la relation *REMEDIATION affects POLLUTANT*.

Ce processus peut être conduit plus loin, en exploitant plus de liens et de familles identifiées à partir du corpus. Afin de simplifier l'assimilation des exemples en trois langues et rappeler le processus dans son intégralité, nous allons résumer les résultats obtenus pour le français dans la section suivante.

7.1.5 Résumé de l'exemple français

Après le pré-traitement du corpus, nous avons identifié six termes candidats en français : : *particulièrement*, *assainir*, *biphényles*, *polychloré*, *biphényles polychloré* et *assainir le biphényles polychloré*. Le candidat *particulièrement* est le seul faux positif dans cet exemple. Nous avons également une variante graphique du terme *biphényles* qui contient une faute d'orthographe (la forme correcte est *biphényle*).

Ensuite, nous avons lancé la recherche des traductions en anglais et en allemand. Cette recherche a permis d'identifier les traductions en anglais pour cinq termes candidats (sauf *assainir*) :

- *particulièrement* est traduit par *particularly@EN*⁶ ;
- *biphényles* par *biphenyle@DE* et *biphenyl@DE* ;
- *polychloré* par *polychlorinate@EN*, *polychlorierte@DE*, *polychlorierter@DE* et *polychlorierten@DE* ;
- *biphényles polychloré* par *polychlorinate biphenyl@EN*, *polychlorbiphenyl@DE* et *polychlorobiphenyl@DE* ;
- et enfin *assainir le biphényles polychloré* par *polychlorinate biphenyl@EN*, *polychlorbiphenyl@DE* et *polychlorobiphenyl@DE*.

L'étape suivante était de retrouver ces termes dans une famille contruite par rapport à un terme de l'ontologie source EcoLexicon. Nous avons donc choisi celle du terme *chlore*. Si l'on superpose les couches des traitements, nous avons la famille multilingue suivante :

6. Dans le but de simplification, nous avons adopté le format des traductions terme@langue, par exemple, *polychlorinate@EN*.

chlore@FR :

- [...]
 - *chloré@FR* (*chlorine-containing@EN*) :
 - *hydrocarbure chloré@FR*
 - *polychloré* (*polychlorinate@EN*, *polychlorierte@DE*, *polychlorierter@DE*, *polychlorierten@DE*) :
 - *biphényles polychloré@FR* (*polychlorinate biphenyl@EN*, *polychlorbiphenyl@DE* et *polychlorobiphenyl@DE*) :
 - *assainir le biphényles polychloré* (*polychlorinate biphenyl@EN*, *polychlorbiphenyl@DE* et *polychlorobiphenyl@DE*)
 - *biphényle polychloré@FR*
 - *dichlorodiphényltrichloroéthane* (*dichlorodiphenyltrichloroethane@EN*)
 - *chlorophosphate* (*chlorophosphate@EN*), etc.

La dernière étape est l'enrichissement de l'ontologie source par les concepts *REMEDIE* et *REMEDIATION* venant du terme candidat *assainir* et du terme dérivé *assainissement*. Ces nouveaux concepts peuvent suivre le modèle du concept existant *PURIFICATION* et hériter la relation *result-of ARTIFICIAL PROCESS*. Nous avons également démontré qu'il est possible de créer les relations *PURIFICATION affects WASTE WATER*, *REMEDIATION affects SOIL POLLUTION* et *REMEDIATION affects POLLUTANT*. De plus, la traduction allemande du concept *REMEDIATION*, *Bodensanierung*, inclut le mot *Boden* signifiant sol, donc *assainissement des sols*. Cette famille permet également d'enrichir l'ontologie source par deux concepts relatifs à des substances chimiques contenant chlore à partir des candidats *hydrocarbure chlorée* et *dichlorodiphényltrichloroéthane*. À ce moment, nous pouvons également élargir ce concept par la traduction anglaise *dichlorodiphenyltrichloroethane* de son libellé.

Cet exemple illustre le processus d'élargissement et d'enrichissement à base des familles morphologiques endogènes. Il peut certainement être complété par d'autres termes et concepts de la même famille, mais nous avons choisi de démontrer les étapes-clés. Ce processus est assez complexe et peut être conduit en continu ou par des itérations.

Une des manières de faciliter ce processus et de le rendre plus efficient est de commencer l'enrichissement par les familles les plus riches ou par les familles des candidats les plus fréquents. La première approche permet d'enrichir l'ontologie par le maximum de nouveaux concepts potentiellement liés, tandis que la deuxième assure une meilleure couverture des corpus du domaine.

7.2 Évaluation

À la différence des chapitres précédents, où chaque étape a été évaluée séparément du point de vue quantitatif et qualitatif, l'objectif de ce chapitre est de fournir une vue d'ensemble sur le système. Étant donné que, à notre connaissance, il n'existe pas de ressource organisée sous forme de familles morphologiques, nous ne pouvons pas fournir une comparaison quantitative pour les résultats. Nous allons donc surtout nous concentrer sur l'évaluation qualitative et analyser des différents cas rencontrés dans nos résultats.

7.2. ÉVALUATION

Mais avant, nous rappelons les ressources utilisées dans ce chapitre et fournissons quelques éléments quantitatifs permettant d'évaluer la performance du système.

7.2.1 Ressources

Dans ce chapitre, nous tenons à être le plus proche des conditions réelles d'exploitation. Pour cette raison, nous avons choisi une ontologie, EcoLexicon, que nous avons présentée dans la Section 3.1.1, et deux sous-corpus PatTR également décrits auparavant (Section 3.2.2.1). Rappelons qu'il s'agit de corpus relatifs aux deux catégories de la CIB : B09C, traitement des sols pollués, et C02F qui porte sur les eaux usées.

Paramètre	Langues	Quantité
Nombre de termes	FR	622
	EN	4 180
	DE	3 601
Nombre de traductions	FR-EN	666
	FR-DE	1 009
	EN-DE	5 594

TABLE 7.5 – Statistiques d'EcoLexicon

EcoLexicon La Table 7.5 rappelle les paramètres clés de l'ontologie EcoLexicon, à savoir le nombre de termes et de traductions de la ressource. Le français est la langue la moins traitée des trois, avec seulement 622 termes renseignés contre 3 601 pour l'allemand et 4 180 pour l'anglais. Nous pouvons donc espérer d'un côté de trouver les traductions françaises pour les termes anglais et allemands, et d'autre côté identifier des termes permettant de créer des nouveaux concepts.

Corpus Les corpus choisis pour l'évaluation finale sont pré-traités comme décrit dans la Section 3.3, c'est-à-dire, tokenisés, lemmatisés par TreeTagger (Schmid, 1994) et étiquetés par les mots informatifs et vides (Vergne, 2004, 2005).

Langue	C02F		B09C	
	Tokens	Lemmes	Tokens	Lemmes
FR	18 414 204	36 919	1 864 871	9 970
EN	12 316 949	36 479	1 250 593	9 515
DE	4 789 889	134 881	480 707	23 375

TABLE 7.6 – Statistiques des sous-corpus C02F et B09C

Comme le montre la Table 7.6, les deux corpus sont de taille différente : le corpus B09C est d'environ quatre fois plus petit que celui C02F. Cette différence nous permet d'évaluer

7.2. ÉVALUATION

l'influence de la taille de corpus sur les performances de notre système et sur l'expérience utilisateur.

7.2.2 Résultats

Dans l'évaluation quantitative des résultats, nous prenons en compte deux paramètres : la couverture des libellés (termes) de l'ontologie source et la taille des familles obtenues.

Couverture des termes de l'ontologie Par couverture des termes de l'ontologie source, nous entendons le nombre de libellés pour lesquels le système construit des familles.

Langue	Termes d'EcoLexicon		
	$n = 3$	$n = 4$	$n = 5$
FR	535	519	464
EN	3 861	3 625	3 239
DE	3 361	3 285	3 178

TABLE 7.7 – Libellés (termes) d'EcoLexicon pris en compte dans le traitement

La longueur des n-grammes limite le nombre de termes de l'ontologie servant à l'identification des termes candidats (Table 7.7). En effet, si le n est supérieur à la longueur d'un terme, celui-ci n'est pas pris en compte.

Corpus	Langue	Couverture (%)		
		$n = 3$	$n = 4$	$n = 5$
B09C	FR	96,82	83,43	68,10
	EN	96,76	84,52	66,59
	DE	97,02	88,55	75,80
C02F	FR	96,82	83,43	68,10
	EN	96,81	84,72	66,69
	DE	97,11	88,74	76,12

TABLE 7.8 – Couverture des familles morphologiques endogènes

La Table 7.8 illustre l'évolution de la couverture en fonction de la longueur des n-grammes de caractères pris en compte dans la construction des familles. Les n-grammes plus courts permettent de couvrir presque tous les termes de l'ontologie, tandis que le $n = 5$ réduit le nombre de familles proposées.

Il est également intéressant de noter que pour chaque valeur de n la couverture reste stable entre les deux corpus, malgré une différence de taille importante entre eux. Cela signifie que la méthode est relativement indépendante du volume de corpus.

Taille des familles morphologiques endogènes La longueur des n-grammes de caractères influence non seulement la couverture, mais aussi la composition des familles construites. Dans notre analyse, nous tenons compte de deux mesures de taille : la taille moyenne et la taille médiane.

La taille moyenne est calculée à partir de la somme du nombre de candidats inclus dans les familles divisée par le nombre de familles. La taille médiane permet d’avoir une idée de nombre de termes dans les familles qui se trouvent au point milieu de la liste ordonnée. Par exemple, si une liste de 10 éléments contient les valeurs [10, 10, 9, 7, 3, 1, 1, 1, 1, 1], la valeur moyenne est égale à 4,4. La valeur médiane est égale à 2, car c’est la moyenne des éléments sur les positions 5 et 6 : $\frac{3+1}{2} = 2$.

Corpus	Langue	Taille moyenne			Taille médiane		
		$n = 3$	$n = 4$	$n = 5$	$n = 3$	$n = 4$	$n = 5$
B09C	FR	34,58	22,96	18,76	22	14	11
	EN	45,64	36,82	24,25	25	17	14
	DE	31,87	24,97	17,85	21	13	8
C02F	FR	34,58	22,97	18,76	22	14	11
	EN	45,62	36,74	24,22	45	17	14
	DE	31,84	24,92	17,80	21	13	8

TABLE 7.9 – Taille des familles morphologiques endogènes

L’augmentation de la longueur des n-grammes entraîne également une baisse du nombre de termes dans les familles (Table 7.9). Toutefois, ces observations quantitatives ne permettent pas de juger si les résultats du système s’améliorent avec la modification de n .

Dans la section suivante, nous allons examiner des exemples de familles, pour identifier la longueur de n optimale et pour juger si réduire le nombre de termes dans une famille signifie une amélioration de résultats.

7.3 Discussion

Les familles morphologiques endogènes obtenues au cours de cette expérience soulèvent des questions aussi bien sur le plan de leur composition que sur celui de l’influence de la structure des termes.

Notamment, il est intéressant de comparer la composition de familles construites avec les longueurs de n différentes et d’analyser l’impact de la taille des familles sur leur qualité. Nous avons également évalué le potentiel de ces familles pour l’élargissement et pour l’enrichissement d’une ontologie.

La structure des termes se manifeste aussi bien au niveau des libellés de l’ontologie pour lesquels les familles sont construites, qu’au niveau des candidats qui les composent.

7.3.1 Composition des familles morphologiques endogènes

La construction des familles dépend essentiellement de la longueur du paramètre n . Nous avons observé que plus les n -grammes pris en compte sont longs, moins de termes candidats sont sélectionnés dans une famille.

En même temps, les familles les plus riches en candidats ne sont pas les plus précises. Il est donc important d'analyser le lien entre la longueur des n -grammes et la taille des familles morphologiques endogènes et leur efficacité pour l'alimentation d'une ontologie.

7.3.1.1 Influence de la longueur des n -grammes

Afin d'observer l'effet de la longueur des n -grammes de caractères sur les familles, nous avons comparé les familles d'un terme construites avec des n -grammes où n est égal à 3, 4 et 5.

En français, nous avons comparé les familles pour le terme *couche d'ozone* :

couche d'ozone :		
$n=3$	$n=4$	$n=5$
– ozone :	– ozone :	– couche
– ozoneur	– ozoneur	– intérieur de la
– couche	– couche	couche
– intérieur de la	– intérieur de la	– couche imperméable
couche	couche	au liquide
– couche imperméable	– couche imperméable	– couche de végétation
au liquide	au liquide	– couche de mousse
– couche de végétation	– couche de végétation	– couche de revête-
– couche de mousse	– couche de mousse	ment
– couche de revête-	– couche de revête-	– couche à action ca-
ment	ment	talytique
– encore 10 candidats	– couche à action ca-	– couche de substrat
– souche	talytique	– couche de gravier
– couplage	– couche de substrat	– couche de floc
– courbure	– encore 8 candidats	– couche de drainage
– encore 10 candidats	– souche	– encore 5 candidats

Les familles morphologiques construites avec $n = 4$ contiennent moins de bruit que celles où $n = 3$ et plus de candidats corrects qu'avec $n = 5$. En effet, dans la première colonne, les candidats *souche*, *couplage*, *courbure* et les 10 suivants n'ont pas de relation avec le terme *couche d'ozone* et ne doivent pas être inclus dans sa famille.

Dans la dernière colonne où $n = 5$, en revanche, les bons termes candidats *ozone* et *ozoneur* ne sont pas présents, car *ozone* ne contient que 5 caractères. Toutefois, certaines langues ont des mots plus longs et cela peut modifier la longueur optimale de n .

7.3. DISCUSSION

strukturanalyse :		
$n=3$	$n=4$	$n=5$
– strukturell	– strukturell	– strukturell
– strukturieren	– strukturieren	– strukturieren
– strukturmateriel	– strukturmateriel	– strukturmateriel
– strukturformel	– strukturformel	– strukturformel
– analyse	– analyse	– struktur
– struktur	– struktur	– überstruktur
– überstruktur	– überstruktur	– halterungsstruktur
– halterungsstruktur	– halterungsstruktur	– wanderwellenstruktur
– wanderwellenstruktur	– wanderwellenstruktur	– analyse
– granalien	– konstruktiv	– konstruktiv
– analog	– analog	– granalien
– kanal		
– et encore 10 candidats		

L'exemple ci-dessus montre qu'en allemand les familles peuvent être composées à l'aide des n -grammes assez longs, ce qui est logique compte tenu de la longueur importante des mots. Ainsi, les familles où n est égale à 4 et à 5, contiennent les mêmes candidats. La longueur $n = 4$ s'est avérée optimale dans notre expérience.

Cette expérience est basée sur les observations des résultats de notre système. Toutefois, il est difficile de tirer des conclusion sans avoir pris en compte l'ensemble de familles. Nous avons donc prévu de mener une analyse plus approfondie sur plus de familles.

7.3.1.2 Influence de la taille de familles

La taille des familles morphologiques varie entre un et des dizaines de candidats. Les familles les plus grandes contiennent des termes très fréquemment utilisés dans le domaine, par exemple, *traitement* qui compte plus de 30 candidats dans sa branche. Ces branches sont toutefois pertinentes. Ainsi, parmi les candidats liés au terme *traitement*, nous retrouvons *traitement enzymatique*, *bain de traitement aqueux* et d'autres.

Les familles de petite taille sont également intéressantes à prendre en compte. Par exemple, la famille du terme *volcan* contient un seul membre *volcanique*, mais ce candidat est tout à fait pertinent. Un autre exemple, la famille du terme *climat boréal* qui contient les candidats *climatiser* et *acclimater*, également pertinents pour notre ontologie.

Dans les grandes familles, tout comme dans celles peu peuplées, il existe des familles pertinentes. C'est donc plutôt la structure du terme qui conditionne sa capacité à regrouper des candidats de façon pertinente.

7.3.2 Influence de la structure des termes

L'enrichissement d'une ontologie peut être conduit en même temps que son élargissement ou après celui-ci. En effet, lorsque l'utilisateur travaille sur une famille, il peut décider d'élargir certains concepts qui existent déjà et d'en créer d'autres à partir des candidats présents dans la famille. Nous n'allons pas dessiner de frontière nette entre ces deux

7.3. DISCUSSION

processus, car, dans notre expérience, ils sont fortement liés.

Nous allons analyser les éléments de mots favorisant l'élargissement et l'enrichissement d'une ontologie. Ce sont notamment des morphèmes issus d'une nomenclature qui permettent de former plusieurs termes d'un domaine, mais aussi l'influence de la position des ces formants dans des mots. Étant donné que certains termes ne contiennent pas de formants néo-classiques, il est également intéressant de voir les familles construites pour ces termes.

7.3.2.1 Formants néo-classiques

Les formants néo-classiques, c'est-à-dire, les morphèmes venus du grec ancien ou du latin, sont largement présents dans les terminologies. Notamment, ces formants jouent un rôle important dans les sciences dures, comme la physique ou la chimie, et les sciences de la vie (biologie, médecine).

Sens partiel Dans certains cas, les formants néo-classiques peuvent rapprocher des termes n'ayant pas de lien direct. C'est notamment le cas de la famille morphologique construite pour le terme *xérophile* (FR) ci-dessous.

xérophile :

- hydrophile
- membrane hydrophile
- nucléophile

Le terme de base, *xérophile* (du grec ancien *xeros* – sec et *philos* – ami)⁷, désigne, selon EcoLexicon, une plante capable de supporter un manque d'humidité. La famille construite pour ce terme, s'appuie évidemment sur la deuxième composante. Toutefois, les termes *hydrophile* et *nucléophile* se réfèrent à des composés chimiques, ce qui les éloigne du terme de base. Il est nécessaire de tenir compte de ces particularités et éviter la création de liens entre les concepts sur la ressemblance formelle et étymologique des termes.

Position du formant L'observation précédente peut être corrélée avec la position du formant dans le mot.

phytologie :

- *phytoréparation* :
 - *phytoréparation électrode*
- *physiologique* :
 - *rinçage physiologique, etc.*
- *biologique* :
 - *biologiquement, etc.*

Dans l'exemple ci-dessus, le membre le plus pertinent, *phytoréparation*, est lié avec le terme *phytologie* plutôt par le début du mot *phyto-* que par son deuxième élément *-logie*. Ce cas est répandu dans les résultats. Toutefois, cela dépend également de l'appréciation de l'utilisateur. Ce dernier peut, par exemple, alimenter les concepts des disciplines liés

7. <https://fr.wikipedia.org/wiki/X%C3%A9rophile>

7.3. DISCUSSION

au domaine en question. Dans cet exemple, les deux éléments sous-lexicaux étaient des morphèmes.

- agriculture :**
- *agricole :*
 - *exploitation agricole :*
 - *exploitation agricole ou forestier*
 - *grain agricole :*
 - *utilisation de grain agricole*
 - *pisciculture*
 - *culture :*
 - *culture mixte microbiotique*
 - *culture de terre contaminer*
 - *fraction de un culture*
 - *culture de micro-organisme*

Le terme *agriculture* illustre un autre cas, où le mot est composé d’une morphème *agri-* et d’un élément lexical *culture*⁸ Nous pouvons alors observer la richesse des candidats regroupés par rapport au deuxième élément *culture* : *pisciculture*, *culture mixte microbiotique*, etc.

7.3.2.2 Familles pour des termes simples

Nous avons évoqué à l’étape d’extraction terminologique que l’existence d’une nomenclature du domaine favorise notre approche. Tout au long de nos traitements, la nomenclature continue à jouer un rôle important. Toutefois, les capacités de notre système ne s’arrêtent pas sur l’identification des termes dont l’étymologie remonte au latin ou à l’ancien grec.

Par exemple, dans la famille du terme *tourbillon de sable*, nous retrouvons des candidats relatifs au mot *tourbillon*, comme *tourbillonner*, *tourbillonnaire*, *écoulement rotationnel ou tourbillonnaire*, *tourbillonnement*, mais aussi *turbine*. Cette famille contient également des candidats relatifs au mot *sable* : *sableux*, *sablage*, *sablonneux*, *sablier*, *sable fin*, *sable géogène* et *sabler*. Cela atteste l’importance du concept *SABLE* pour le domaine de traitement des sols pollués, ce qui se confirme par la présentation de ce concept dans EcoLexicon (Figure 7.3).

Il faut quand même remarquer que certaines familles regroupent des candidats sans relation au libellé provenant de l’ontologie :

- batardeau :**
- *retardement :*
 - *élément de retardement*
 - *retarder*
 - *retardé*
 - *garder*

8. Même si étymologiquement *biologie* (du grec *bio+logos*) et *agriculture* (du latin *ager+colere*) sont formés sur un même patron, l’élément *culture* constitue un mot en français, tandis que *-logie* reste un morphème.

7.3. DISCUSSION

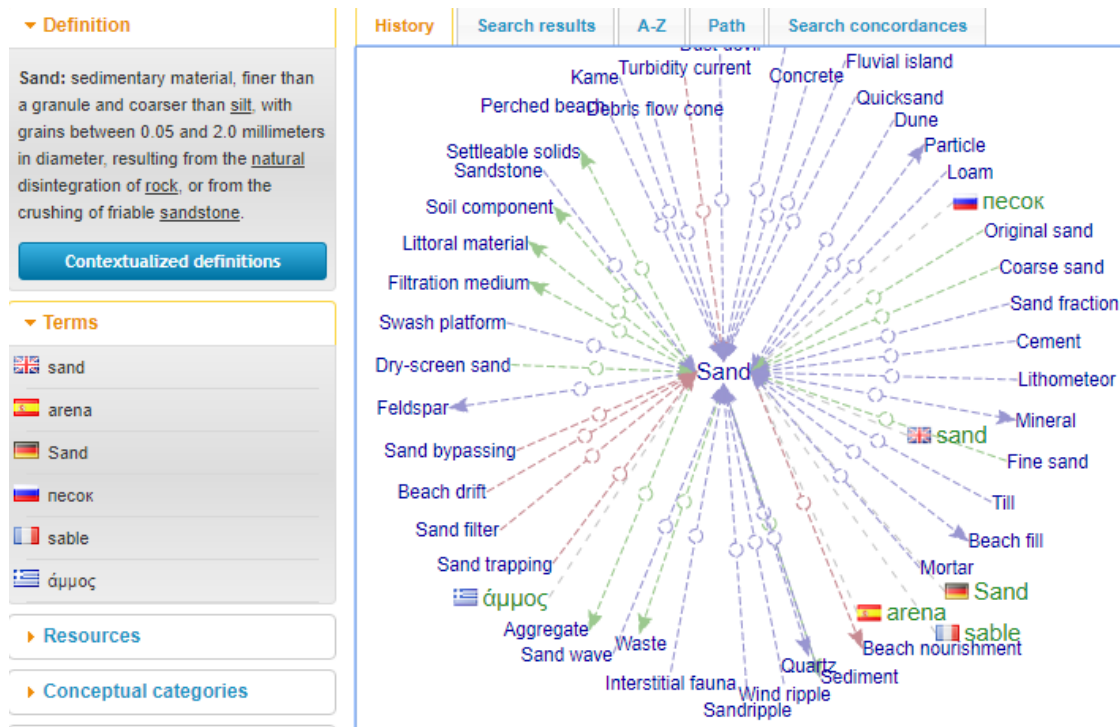


FIGURE 7.3 – Concept *SAND* dans EcoLexicon

La ressemblance formelle a rapproché le terme *batardeau*⁹ des candidats dérivés du verbe *retarder* et au verbe *garder*. Le verbe *retarder* est présent dans EcoLexicon en anglais dans les concepts *RETARD LITTORAL DRIFT*¹⁰ et *RETARD BEACH EROSION* qui signifie *retarder de l'érosion des plages*.

Même imprécise, cette famille peut aider l'utilisateur dans l'alimentation de l'ontologie en regroupant des candidats proches pour faciliter leur insertion. Par exemple, il est possible de créer un concept *RETARDEMENT* pour y attacher par la suite tous les concepts proches, comme les deux cités dans la paragraphe précédent. Cela se rapproche de l'enrichissement de l'ontologie.

7.3.3 Alignement multilingue des familles morphologiques endogènes

Dans notre système actuel, les familles sont construites pour chaque langue de traitement à part. Cela signifie que l'utilisateur est sensé effectuer l'élargissement et l'enrichissement à partir d'une langue, même s'il peut consulter les candidats traductions pour des termes candidats.

9. Un batardeau est un barrage destiné à la retenue d'eau provisoire en un lieu donné sur une surface donnée. (<https://fr.wikipedia.org/wiki/Batardeau>)

10. Retarder la dérive du littoral, c'est-à-dire, le déplacement des masses de sables par les vagues et le courant.

7.4. CONCLUSION

Nous avons imaginé, sans toutefois pouvoir le mettre en pratique dans le cadre de ce projet, des familles morphologiques alignées au niveau de leurs candidats. Cette présentation permettrait à l'utilisateur de mener un processus multilingue par excellence. Il pourrait ainsi ajouter des termes et des concepts en plusieurs langues, même si certains termes sont présents que dans une partie de langues de traitement.

macro-permeability@en, Makropermeabilität@de :

- *macronutrient@en*
- *low-permeability@en*
- *permeability@en, permeabilität@de, perméabilité@fr :*
 - *permeability of contaminated soil@en*
 - *impermeability@en, imperméabilité@fr*
 - *permeabilitätscharakteristik@de*
 - *permeabilitätskoeffizient@de*
 - *permeabilitätsverringersschritt@de*
- *perméable@fr :*
 - *ion-permeable@en*
 - *fluid-permeable@en, perméable au liquide@fr*
 - *water-permeable@en, wasserpermeabilität@de, perméable à le eau@fr*
- *semi-perméable@fr :*
 - *semi-permeable membrane@en, membrane semi-perméable@fr*
- etc.

Cette présentation permet de regrouper des termes proches à travers des langues. Nous avons sélectionné un exemple basé sur des cognats, mais il est tout à fait possible de créer des familles multilingues alignées même si les traductions ne sont pas proches graphiquement. Dans cet alignement, le système doit utiliser les candidats traductions identifiées à l'étape d'alignement des termes candidats.

Compte tenu des contraintes du projet de thèse, nous n'avons pas pu implémenter l'alignement multilingue des familles morphologiques, mais cela est prévu dans les perspectives.

7.4 Conclusion

Ce chapitre vise à rappeler notre méthodologie dans son intégralité et sur des données complètes. Nous avons illustré la chaîne des traitements par des exemples en français, en anglais et en allemand. Nous avons mené l'expérience sur les sous-corpus B09C et C02F dans leur intégralité. Cela a permis d'évaluer les performances de l'ensemble du système.

Nous avons constaté que le regroupement des termes candidats en familles morphologiques facilite le traitement des résultats par un expert humain. En effet, lorsque les termes proches sont regroupés, il est possible d'ajouter des nouveaux concepts dans une branche de l'ontologie et d'associer des libellés correspondants à ces concepts. Il faut noter que certaines familles peuvent contenir des termes candidats n'ayant pas de relation avec le libellé de l'ontologie en question. Toutefois, cela ne constitue pas de problème majeur, car l'expert peut choisir les branches de l'ontologie et d'une famille sur lesquelles il veut

7.4. CONCLUSION

travailler.

Les familles morphologiques endogènes varient en taille et en nombre en fonction de la longueur des n -grammes. Ainsi, lorsque cette longueur est fixée à 3, les familles sont construites presque pour chaque terme de l'ontologie et contiennent du bruit. La longueur n égale à 4 permet équilibrer la composition des familles, mais réduit le nombre de familles construites, car ne permet pas de traiter des termes courts. Enfin, $n = 5$ paraît excessif, car limite trop le nombre de familles et le nombre de candidats qui y sont rattachés. Il existe cependant une exception : en allemand, $n = 4$ et $n = 5$ fournissent des résultats quasi-identiques, ce qui s'explique par la longueur élevée des mots.

Un autre critère qui influence la construction d'une famille est la structure du terme de l'ontologie qui se trouve à sa tête. Si celui-ci contient des morphèmes ou des éléments lexicaux qui appartiennent à la nomenclature du domaine ou aux formants classiques, la famille de ce terme a plus de chances à être peuplée. Cette particularité influence le processus d'élargissement d'une ontologie, car permet d'ajouter plus de termes à partir des familles riches en candidats. Toutefois, les familles construites pour des termes simples qui ne contiennent pas de formants classiques sont également capables de regrouper des candidats pertinents.

Le travail sur les familles morphologiques sera poursuivi. Notamment, nous visons à aligner les familles morphologiques entre les différentes langues de traitement en partant des alignements au niveau des termes candidats. Cela permettra à l'utilisateur d'ajouter des nouveaux termes et concepts en plusieurs langues à la fois. Compte tenu des délais du projet à respecter, cet alignement multiligüe n'a pas pu être intégré dans cette thèse.

Conclusion et perspectives

Conclusion

L'objectif principal de notre thèse est de proposer une méthode d'élargissement et d'enrichissement multilingues d'une ontologie spécialisée à partir de corpus comparables. Ces processus sont composés de plusieurs étapes successives. D'abord, les termes candidats sont extraits à partir du corpus pour chaque langue. Ensuite, les traductions potentielles sont identifiées dans les listes de termes des langues différentes. Enfin vient l'ajout des nouveaux libellés et instances (l'élargissement) et des nouveaux concepts et relations (l'enrichissement) dans l'ontologie. Toutefois, nous n'avons pas traité l'ajout des relations sémantiques dans le cadre de notre projet.

Par ailleurs, nous avons visé une approche applicable à une large palette de langues et de domaines. De plus, nous avons limité l'apport extérieur en matière de ressources et d'outils, afin de rendre la méthode facilement applicable dans le contexte industriel et, plus précisément, celui des PME.

Nous voulons souligner que notre méthode est avant tout destinée à traiter des ontologies appelées *légères*. En effet, les ontologies peuvent atteindre des niveaux très abstraits, notamment via des règles ou des axiomes dans les ontologies dites *lourdes*. Plus précisément, nos traitements visent les termes et les concepts. Les relations sémantiques ne sont pas traitées dans le cadre de notre projet, mais nous en tenons compte dans l'analyse des résultats.

Nous avons évalué notre méthode sur trois ressources sémantiques (EcoLexicon, Agrovoc et GEMET) et deux corpus multilingues en traitement des eaux usées et en pollution des sols. L'évaluation a été menée sur le français, l'anglais et l'allemand, et sur toutes les six combinaisons de ces langues.

Nos objectifs ont été atteints par le biais des ressources endogènes, générées directement en cours d'analyse à partir du corpus et de l'ontologie. Nos contributions s'appuient sur ces ressources endogènes appliquées à chaque étape de traitement énumérées ci-dessus.

La première contribution porte sur l'utilisation des ressources morphologiques et morphosyntaxiques endogènes pour extraire des termes candidats mono- et polylexicaux. Ne nécessitant aucune ressource morphologique ni morphosyntaxique extérieure¹, cet extracteur terminologique peut être appliqué sur un corpus de tout domaine de spécialité et

1. Même si la méthode peut fonctionner sur du texte brut, une étape de lemmatisation du corpus apporte une meilleure précision.

CONCLUSION ET PERSPECTIVES

toutes les ressources sont alors générées par le programme. Notre méthode combine des traitements statistiques et linguistiques, et permet entre autres d'extraire des termes mono et polylexicaux ayant des fréquences faibles dans le corpus. Cela permet d'inclure des nouveaux termes d'un domaine (des néologismes) dès leur apparition.

La méthode d'extraction terminologique s'appuie sur des n-grammes de caractères qui remplacent des listes de formants classiques et des étiquettes morphosyntaxiques. Le modèle de n-grammes de début et du milieu de mots peut être généré soit à partir des termes déjà présents dans l'ontologie source, soit à partir de termes candidats monolexicaux extraits par des approches statistiques. Ce modèle sert à l'identification des termes monolexicaux. La méthode se montre meilleure en matière de rappel et de F-mesure par rapport à TTC TermSuite 2.2 et est très proche des résultats de TTC TermSuite 1.6.

Les termes polylexicaux sont identifiés grâce aux patrons morphosyntaxiques endogènes qui combinent l'annotation par mots informatifs et vides de J. Vergne et les n-grammes de fin de mots informatifs. Les mots vides sont alors restreints à des mots de schéma les plus fréquents appris à partir du corpus et les étiquettes des mots informatifs sont remplacées par les n-grammes de fin de mot. Appliquée sur les corpus lemmatisés avec la liste de termes monolexicaux validés, la méthode dépasse les performances de TTC TermSuite 2.2.

De plus, notre extracteur terminologique peut fonctionner sur un corpus non lemmatisé. La perte de précision est alors d'environ 10 % pour les termes monolexicaux par rapport aux corpus lemmatisés et est quasi absente pour les termes polylexicaux. Cela permet d'exploiter cette méthode pour des langues peu dotées n'ayant pas de lemmatiseur, ni d'étiqueteur morphosyntaxique.

La deuxième contribution concerne l'exploitation des n-grammes de caractères pour identifier les traductions des termes extraits à partir de corpus parallèles et comparables. Cette méthode d'alignement bilingue des termes requiert seulement un petit nombre de traductions des termes d'un domaine en deux langues pour préparer les alignements des n-grammes avec l'outil Anymalign. L'application de l'aligneur statistique au niveau des n-grammes de caractères des termes et de leurs traductions, constitue un aspect original de notre méthode, car les aligneurs statistiques sont initialement conçus pour aligner des mots dans des corpus parallèles. L'avantage principal de notre méthode est sa capacité à identifier les traductions pour les termes composés monolexicaux, contenant des mots simples et les morphèmes, et polylexicaux, sans avoir à les segmenter en éléments constitutifs. Ce module prend en compte la position de n-grammes dans les mots et permet d'identifier les cognats, donc des termes proches graphiquement en langues source et cible, même si leur n-grammes ne sont pas connus dans le modèle.

Comparée avec les méthodes d'identification de cognats, notre approche d'alignement bilingue des termes a obtenu les meilleurs rappel (30,94 % en moyenne) et F-mesure (41,7 %) contre 19,08 % et 30 % respectivement, avec une légère différence en termes de précision : 63,13 % contre 70,30 %. Elle reste performante en termes de rappel et de F-mesure même si le modèle a été obtenu à partir d'un nombre limité de traductions, la précision se montre toutefois plus sensible à la taille du modèle.

Nous avons également comparé les résultats de cette deuxième contribution avec TTC TermSuite 2.2 sur les deux sous-corpus (B09C et C02F) et trois paires de langues (FR-EN, EN-DE et DE-FR) dans les deux sens, ce qui résulte en 12 échantillons. Notre approche

CONCLUSION ET PERSPECTIVES

identifie plus de traductions complètes et partielles (lorsque seulement une partie de terme est traduite) dans 9 échantillons sur 12 et plus de traductions dans l'ensemble de résultats. En même temps, notre méthode propose moins de fausses traductions, en leur préférant une absence de proposition.

Et enfin la troisième contribution consiste à organiser les termes candidats extraits lors des traitements précédents en familles morphologiques endogènes. Ces familles sont construites pour les termes de l'ontologie afin de permettre son élargissement, et pour les termes candidats, dans le cadre de l'enrichissement de l'ontologie. La méthode de construction des familles morphologiques endogènes a recours aux n -grammes de caractères appris de la ressource terminologique à enrichir (une ontologie du domaine, un dictionnaire ou des résultats de l'extraction terminologique) et du corpus.

Nous avons également mené une évaluation diachronique de ces familles en utilisant un corpus des années 1980 et celui des années 2000. La comparaison des familles morphologiques a démontré la capacité de l'approche à suivre l'évolution de la terminologie dans le temps en détectant les néologismes du domaine dès leur apparition dans des corpus.

Par rapport aux travaux existants sur l'enrichissement et l'élargissement des ontologies, notre méthode à base de familles morphologiques endogènes peut être classée dans la catégorie des approches mixtes, supervisées et outillées. L'utilisateur humain est en effet seul à pouvoir valider ou rejeter les suggestions du système qui pré-construit des familles pour faciliter son travail.

Ces trois contributions ont démontré que les n -grammes de caractères constituent un moyen accessible et polyvalent. Appris directement à partir des ressources du projet, ces n -grammes peuvent être réutilisés à différentes étapes du traitement ou générés à chaque étape, si l'on a besoin de varier la longueur n . Cette facilité de paramétrage permet également leur application à des langues différentes : si les mots d'une langue sont plus longs, il est possible d'augmenter cette longueur n . Une autre particularité de ces n -grammes, c'est la possibilité de les traiter différemment en fonction de leur position dans les mots. Pour les langues utilisées dans notre évaluation, les n -grammes grammaticaux (des suffixes ou des terminaisons) se trouvent à la fin des mots. Nous avons utilisé cette particularité pour donner plus de poids sémantique aux n -grammes du début et du milieu des mots.

Les résultats obtenus par notre système sont très encourageants. Avec peu de ressources et d'outils extérieurs, nous avons atteint des performances des systèmes plus complexes, comme TTC TermSuite. Un autre point fort est le temps d'exécution des traitements tout à fait adapté à l'exploitation du système en temps réel. En effet, l'indexation par n -grammes de caractères au cours des traitements limite significativement le temps de recherche des termes et de leurs traductions.

Bien évidemment, notre méthode a des limites : elle est capable de traiter un domaine de spécialité à la fois. En effet, l'hétérogénéité des données dans un corpus et une ontologie peut introduire du bruit et des résultats peu exploitables. Un autre point faible est l'incapacité de notre système de prendre en compte des termes très courts, comme *eau* ou *sol*. D'un côté, ces termes sont des fois étiquetés comme non-informatifs à l'étape du pré-traitement, car ils sont courts et fréquents. De l'autre côté, ils sont trop courts pour en apprendre des n -grammes de caractères. La prise en compte des termes courts peut être paramétrée dans les futures versions du système.

Perspectives

Les perspectives du projet concernent d'un côté l'optimisation des traitements et l'introduction d'autres ressources, et de l'autre côté, l'application du système à d'autres tâches.

Parmi les développements possibles, nous pouvons indiquer la prise en compte des skip-grammes (des n-grammes de caractères acceptant tout caractère à une position donnée) qui pourraient, par exemple, améliorer la détection des variations des caractères accentués. Nous pouvons également imaginer l'application combinée de notre méthode et des méthodes d'identification de cognats utilisés en tant que baseline à l'étape d'alignement bilingue des terminologies. Cela permettrait de gagner en rappel et en F-mesure, mais le blocage actuel se situe dans la différence significative de temps d'exécution entre les deux approches. Enfin, il est possible d'inclure des étiquettes morphosyntaxiques que nous avons volontairement omises dans nos traitements. Ces étiquettes peuvent améliorer la précision du système et constituer un moyen de détection des termes courts que nous avons évoqués précédemment.

Nous avons également prévu de continuer le travail sur les familles morphologiques endogènes, notamment en matière de l'évaluation d'un plus grand nombre de familles et sur davantage de critères. Nous aimerions également ajouter une étape d'alignement multilingue des familles comme nous l'avons illustré dans la discussion du dernier chapitre. Les familles multilingues permettront d'élargir et d'enrichir une ontologie en plusieurs langues en même temps.

Quant à la diversification d'applications du système, nous envisageons l'extension des fonctionnalités à la recherche des relations sémantiques au sein des familles morphologiques endogènes. Par exemple, il est tout à fait possible de paramétrer l'identification des relations hiérarchiques *est-un* ou encore des liens du type *relatif-à* que l'utilisateur pourrait affiner. De plus, les n-grammes de caractères situés au début des mots pourraient servir à la détection des antonymes au sein des familles. Une autre extension d'usage potentielle est l'utilisation de l'alignement bilingue des termes pour améliorer la traduction automatique statistique. Nous pouvons notamment injecter ces alignements dans le modèle avant l'étape de traduction, ou améliorer les traductions déjà effectuées.

Notre projet mérite un développement d'une plate-forme complète intégrant toutes les étapes et dotée d'une interface utilisateur. De plus, il faut prendre en compte les formats conventionnels des ontologies (RDF, OWL) aussi bien que des fichiers CSV, XML ou *.txt*. Cela permettrait son exploitation dans le cadre de projets de recherche et des activités industrielles. Notamment, le système peut être utilisé par des étudiants désirant de créer des terminologies multilingues à partir de corpus spécialisés comparables ou parallèles.

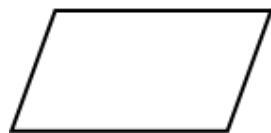
Notre méthodes d'extraction et d'alignement des termes à base des n-grammes de caractères pourra trouver d'autres applications, notamment dans le domaine de la bioinformatique. Étant donné que les mots sont des séquences de caractères, nous avons utilisé des méthodes de la bioinformatique servant à calculer les similarités au niveau des séquences de protéines. Il est intéressant de voir si nos traitements peuvent être utilisés pour ces tâches.

Enfin, notre méthode va l'encontre de la tendance actuelle d'utiliser toujours plus d'outils, de ressources et de données. Nous espérons avoir démontré qu'il est possible d'optimiser les ressources matérielles, temporelles et financières disponibles pour avoir des résultats tout à fait exploitables pour l'objectif fixé.

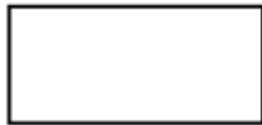
Annexes

Annexe A

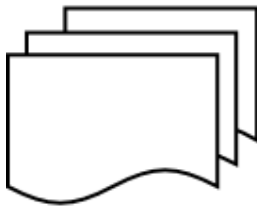
Guide de lecture



Entrée/sortie



Traitement



Corpus



Exemples
de données



Index

Annexe B

Catégories de la Classification Internationale de Brevets (CIB)

Domaine	CIB	Catégorie
Écologie	C02F	Traitement de l'eau, des eaux résiduaires, des eaux ou boues d'égout
Chimie	B09C	Régénération de sols pollués
	C01G	Composés contenant des métaux non couverts par les sous-classes
	C25B	Procédés électrolytiques ou électrophorétiques pour la production de composés organiques ou minéraux, ou de non-métaux ; appareillages à cet effet
	C25C	Procédés pour la production, la récupération ou l'affinage électrolytique des métaux ; appareillages à cet effet
	C25D	Procédés pour la production électrolytique ou électrophorétique de revêtements ; galvanoplastie
Énergie	C25F	Procédés pour le traitement d'objets par enlèvement électrolytique de matière ; appareillages à cet effet
	H01G	Condensateurs ; condensateurs, redresseurs, détecteurs, dispositifs de commutation, dispositifs photosensibles ou sensibles à la température, du type électrolytique
	H01M	Procédés ou moyens pour la conversion directe de l'énergie chimique en énergie électrique, p.ex. batteries
	H02J	Circuits ou systèmes pour l'alimentation ou la distribution d'énergie électrique ; systèmes pour l'accumulation d'énergie électrique
	H02M	Appareils pour la transformation de courant alternatif en courant alternatif, de courant alternatif en courant continu ou vice versa ou de courant continu en courant continu et employés avec les réseaux de distribution d'énergie ou des systèmes d'alimentation similaires ; transformation d'une puissance d'entrée en courant continu ou courant alternatif en une puissance de sortie de choc ; leur commande ou régulation

TABLE B.1 – Corpus PatTR : les catégories de la CIB sélectionnées

Annexe C

Statistiques du corpus PatTR

Domaine	CIB	Paire	Nombre de segments		
			Titres	Résumés	Revendications
Écologie	C02F	FR-DE	12 794	925	19 943
		EN-FR	16 969	26 548	46 395
		DE-EN	13 740	6 063	57 629
	B09C	FR-DE	1 065	82	1 820
		EN-FR	1 495	2 539	5 460
		DE-EN	1 138	652	6 604
	C01G	FR-DE	5 147	132	7 635
		EN-FR	6 058	6 161	20 189
		DE-EN	5 388	1 071	23 703
Chimie	C25B	FR-DE	2 814	145	4 355
		EN-FR	3 352	3 973	9 787
		DE-EN	2 987	744	12 185
	C25C	FR-DE	1 274	26	1 883
		EN-FR	1 620	2 252	5 020
		DE-EN	1 324	225	6 943
	C25D	FR-DE	4 684	445	7 036
		EN-FR	5 737	7 531	16 411
		DE-EN	5 090	2 119	20 327
	C25F	FR-DE	718	73	1 238
		EN-FR	867	1 246	2 518
		DE-EN	785	385	3 252
	H01G	FR-DE	4 009	192	6 496
		EN-FR	5 220	7 549	13 033
		DE-EN	4 236	1 045	15 818
Énergie	H01M	FR-DE	17 373	1 372	28 303
		EN-FR	24 018	42 912	51 978
		DE-EN	18 621	5 246	64 681
	H02J	FR-DE	7 184	616	11 551
		EN-FR	9 121	14 827	22 068
		DE-EN	7 761	3 027	25 976
	H02M	FR-DE	7 966	551	11 750
		EN-FR	9 832	14 050	22 826
		DE-EN	8 481	3 163	26 071

TABLE C.1 – Corpus PatTR : le nombre de segments parallèles

Annexe D

Échantillons des termes candidats

ÉCHANTILLONS DES TERMES CANDIDATS

Langue	Position dans la liste		
	Tête	Milieu	Fin
FR	photométrique (1)	abattoir (1)	monophasé (13)
	hydrométallurgiques (1)	polarographe (3)	syndiotactique (30)
	photo-électro- chimiquement (2)	perhalogénates (1)	di-n-propyléther(1)
	photo-électrochimique (1)	auto-portant (1)	atomiquement (1)
	photo-cathode (4)	échantillonner (7)	acidique (29)
EN	morphological (1)	dihydroxylated (2)	sand-blast (1)
	biological (7)	dihydroxyboron (2)	irradiation (5)
	physiologically (1)	chloroform (3)	polyetherketones(2)
	1-trifluoromethyl- 1,2,2-trifluoro- cyclobutane (2)	thermo-plastic (8)	mannose (2)
	1-trifluoromethyl- 1,2,2-trifluoro- 3-methylcyclobutane (2)	solid-walled (4)	intimately (16)
DE	wasserstoff- brennstoff- zeugungungs- elektrolyse- einrichtungen (10)	distanzelement- bereichs (1)	feinverteilen (1)
	probenwasser- strömungsdurchlass (1)	querschnitt (2)	wiederverwertung (2)
	wasserstoffspeicher- legierungspulver (9)	lithiumfluor- alkylphosphate (6)	durchflußelektrode (4)
	kohlenwasserstoff- brennstoff- reaktionsmittels (1)	induktionsspule (1)	querstrebe (1)
	kohlenwasserstoff- brennstoff- reaktionsmittel (1)	gasverteilermaterial (1)	aufeinanderfol-gende (1)

TABLE D.1 – Échantillons des termes candidats avec leurs fréquences (corpus 2000)

Bibliographie

- AKER, A., PARAMITA, M. et GAIZAUSKAS, R. (2013). Extracting bilingual terminologies from comparable corpora. *In ACL*, pages 402–411, Sofia, Bulgaria.
- AMINI, M. R. et GAUSSIER, E. (2013). *Recherche d'information : Applications, modèles et algorithmes - Fouille de données, décisionnel et big data*. Algorithmes. Eyrolles.
- ARAÚZ, P., REIMERINK, A. et FABER, P. (2011). Environmental knowledge in EcoLexicon. *In Computational Linguistics-Applications Conference*, numéro 14, pages 9–16.
- BALDWIN, T. et TANAKA, T. (2004). Translation by machine of complex nominals : Getting it right. *In Proceedings of the ACL04 Workshop on Multiword Expressions : Integrating Processing*, pages 24–31.
- BELTAGY, I., ERK, K. et MOONEY, R. (2014). Semantic Parsing using Distributional Semantics and Probabilistic Logic. *In Proceedings of ACL 2014 Workshop on Semantic Parsing (SP-2014)*, pages 7–11, Baltimore, MD.
- BENZ, D. et HOTH, A. (2014). Capturing emergent semantics from social tagging systems. *In LEHMANN, J. et VÖLKER, J., éditeurs : Perspectives on Ontology Learning*, pages 175–188.
- BERNHARD, D. (2006). *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales*. Thèse de doctorat, Université Joseph Fourier – Grenoble I.
- BERNIER-COLBORNE, G. (2012). Defining a Gold Standard for the Evaluation of Term Extractors. *ColabTKR 2012 - Terminology and Knowledge Representation Workshop*, pages 1–14.
- BERNIER-COLBORNE, G. et L'HOMME, M. C. (2015). Using a distributional neighbourhood graph to enrich semantic frames in the field of the environment. *In Proceedings of the conference Terminology and Artificial Intelligence (TIA)*, pages 9–16.
- BIBER, D., CONRAD, S. et CORTES, V. (2004). If you look at : Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405.
- BLOHM, S. (2010). *Large-scale pattern-based information extraction from the world wide web*. Thèse de doctorat, Karlsruher Institut für Technologie.

BIBLIOGRAPHIE

- BLUMENTHAL, P. (2007). Sciences de l'Homme vs sciences exactes : combinatoire des mots dans la vulgarisation scientifique. *Revue française de linguistique appliquée*, XII(2):15–28.
- BOJANOWSKI, P., GRAVE, E., JOULIN, A. et MIKOLOV, T. (2017). Enriching Word Vectors with Subword Information. *In Transactions of the Association for Computational Linguistics*, pages 135–146.
- BOLLEGALA, D., KONTONATSIOS, G. et ANANIADOU, S. (2015). A Cross-Lingual Similarity Measure for Detecting Biomedical Term Translations. *Plos One*, 10(6):1–28.
- BOSCA, A., CASU, M., DRAGONI, M. et REXHA, A. (2014). Modeling, Managing, Exposing, and Linking Ontologies with a Wiki-based Tool. *In Proceedings of LREC*, pages 1668–1675.
- BOUAMOR, D., SEMMAR, N. et ZWEIGENBAUM, P. (2013). Utilisation de la similarité sémantique pour l'extraction de lexiques bilingues à partir de corpus comparables. *In Proceedings of TALN*, pages 17–21, Les Sables d'Olonne, France.
- BOURIGAULT, D. (2002). Upéry : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN*, pages 75–84, Nancy.
- BOURIGAULT, D. et FABRE, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151.
- BOURIGAULT, D. et JACQUEMIN, C. (1999). Term extraction+ term clustering : An integrated platform for computer-aided terminology. *In Proceedings of EACL*, pages 15–22.
- BRAUNE, F. et FRASER, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. *In Proceedings of COLING*, pages 81–89.
- BROOKE, J., TSANG, V., HIRST, G. et SHEIN, F. (2014). Unsupervised Multiword Segmentation of Large Corpora using Prediction-Driven Decomposition of n -grams. *In Proceedings of COLING*, pages 753–761.
- BUITELAAR, P., CIMIANO, P. et MAGNINI, B. (2003). Ontology learning from text : An overview. *In Ontology learning from text*, pages 1–10. IOS Press.
- CABRÉ CASTELLVÍ, M. T. (2003). Theories of terminology : their description, prescription and explanation. *Terminology*, 9(2):163–199.
- CANAS, A. J. et NOVAK, J. D. (2009). What is a Concept ? ... from a Concept Mapping Perspective.
- CAP, F., FRASER, A., WELLER, M. et CAHILL, A. (2014). How to Produce Unseen Teddy Bears : Improved Morphological Processing of Compounds in SMT. *In Proceedings of EACL*, pages 579–587, Gothenburg, Sweden.

BIBLIOGRAPHIE

- CARLSON, A., BETTERIDGE, J., WANG, R. C. et MITCHELL, T. M. (2010). Coupled Semi-Supervised Learning for Information Extraction. *Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110.
- CASTRO, M. B. et FABER, P. (2016). EcoLexicon , une nouvelle version plus proche des besoins des traducteurs. *In Aspects of Specialised Translation*, pages 144–152. Tübingen.
- CASTRO, M. B., MARTÍNEZ, S. M. et FABER, P. (2014). Verb collocations and phraseology in EcoLexicon. *Yearbook of Phraseology*, 5(1):57–94.
- CHARLET, J., SZULMAN, S. et AUSSENAC-GILLES, N. (2010). DaFOE : une plateforme pour construire des ontologies à partir de textes et de thésaurus. *In 10ième Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances*, pages 1–3, Hammamet, Tunisie.
- CHARLET, J., SZULMAN, S., AUSSENAC-GILLES, N., NAZARENKO, A., HERNANDEZ, N., NADAH, N., DELAHOUSSE, J. et TEGUIAK, V. (2009). DaFOE : an Ontology Building Platform From Texts or Thesauri. *In Proceedings of International Conference on Knowledge Engineering and Ontology Development*, pages 1–4, Madeira, Portugal.
- CHIBOUT, K., MARIANI, J., MASSON, N., NÉEL, F. et GUILLOU, M. (2000). *Ressources et évaluation en ingénierie des langues*. Actualité scientifique. Louvain-la-Neuve : De Boeck université ; Bruxelles : Duculot ; Montréal ; Paris : AUPELF UREF.
- CHRIZMAN, N. et ITAI, A. (2014). How to Construct Multilingual Domain Ontologies. *In Proceedings of LREC*, pages 4345–4350.
- CIMIANO, P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Application*. Springer US.
- CIMIANO, P., HOTH, A. et STAAB, S. (2005). Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *In JAIR*, volume 24, pages 1–15.
- CIMIANO, P. et VÖLKER, J. (2005). Towards large-scale, open-domain and ontology-based named entity classification. *In Proceedings of RANLP*, pages 166–172.
- CIMIANO, P., VÖLKER, J. et STUDER, R. (2006). Ontologies on Demand ? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. *Information, Wissenschaft und Praxis*, 57:315–320.
- CLAVEAU, V. et KIJAK, E. (2011). Morphological analysis of biomedical terminology with analogy-based alignment. *International Conference Recent Advances in Natural Language Processing, RANLP*, (September):347–354.
- COPPOLA, B., GANGEMI, A., GLIOZZO, A., PICCA, D. et PRESUTTI, V. (2014). Learning Domain Ontologies by Corpus-Driven FrameNet Specialization. *In LEHMANN, J. et VÖLKER, J., éditeurs : Perspectives on Ontology Learning*, pages 115–133.

BIBLIOGRAPHIE

- CRAM, D. et DAILLE, B. (2016). TermSuite : Terminology Extraction with Term Variant Detection. *In Proceedings of ACL*, pages 13–18.
- CUADROS, M. et RIGAU, G. (2008). KnowNet : building a large net of knowledge from the web. *In Proceedings of COLING*, pages 161–168, Manchester, UK.
- DA SILVA, L. (2010). Extraction semi-automatique d’un vocabulaire savant de base pour l’indexation automatique. *In Actes de TALN*, volume 1, pages 19–23.
- DAILLE, B. (2003). Conceptual structuring through term variations. *In* BOND, F., KORHONEN, A., MACCARTHY, D. et VILLACICENCIO, A., éditeurs : *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.
- DAILLE, B., JACQUIN, C., MONCEAUX, L., MORIN, E. et ROCHETEAU, J. (2011). TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue. *In Actes de TALN*, Les Sables d’Olonne, France.
- DAILLE, B. et MORIN, E. (2012). Revising the compositional method for terminology acquisition from comparable corpora. *In Proceedings of COLING*, pages 1797–1810.
- DE CEA, G. A., GÓMEZ-PÉREZ, A., MONTIEL-PONSODA, E. et SUÁREZ-FIGUEROA, M. C. (2008). Natural language-based approach for helping in the reuse of ontology design patterns. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5268 LNAI:32–47.
- DE MARNEFFE, M.-C., MACCARTNEY, B. et MANNING, C. D. (2006). Generating typed dependency parses from phrase structure parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454.
- DEBILI, F. (1982). *Analyse syntactico-sémantique fondée sur une acquisition automatique des relations lexicales-sémantiques*. Thèse de doctorat, Université de Paris XI.
- DELPECH, E., CANDILLIER, L., LAPORTE, L. et PHAN, S. (2013). Identification de compatibilités entre descripteurs de lieux et apprentissage automatique. *In EGC*, pages 311—316, Toulouse, France.
- DELPECH, E., DAILLE, B., MORIN, E. et LEMAIRE, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. *In Proceedings of COLING*, pages 481–489.
- DRAGOS, V. et JAULENT, M.-c. (2010). Apprentissage de patrons lexico-syntaxiques à partir de textes. *In Proceedings of RNTI-E-19*, pages 615–620.
- DROUIN, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, XII:45–64.
- DROUIN, P. et LANGLAIS, P. (2006). Évaluation du potentiel terminologique de candidats termes. *In Actes de JADT*, pages 389–400.

BIBLIOGRAPHIE

- ENGUEHARD, C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle*, pages 373–384.
- ESTOPÀ, R., VIVALDI, J. et CABRÉ, M. T. (2000a). Use of Greek and Latin forms for term detection. *In Proceedings of LREC*, Athens, Greece.
- ESTOPÀ, R., VIVALDI, J. et CABRÉ, T. (2000b). Extraction of monolexical terminological units : requirement analysis. *In Workshop Proceedings of Second International Conference on Language Resources and Evaluation. Terminology Resources and Computation*, volume 56, pages 51–56, Athens, Greece.
- ETZIONI, O., CAFARELLA, M., DOWNEY, D., POPESCU, A. M., SHAKED, T., SODERLAND, S., WELD, D. S. et YATES, A. (2005). Unsupervised named-entity extraction from the Web : An experimental study. *Artificial Intelligence*, 165(1):91–134.
- FABER, P. (2011). The dynamics of specialized knowledge representation : Simulational reconstruction or the perception-action interface. *Terminology*, 17(1):9–29.
- FABER, P. et CASTRO, M. B. (2014). EcoLexicon. *Proceedings of the XVI EURALEX International Congress : The User in Focus*, pages 601–608.
- FABER, P., LEON, P. et PRIETO, J. A. (2009). Semantic Relations, Dynamicity, And Terminological Knowledge Bases. *Current Issues in Language Studies*, 1:1–23.
- FABER, P., LINARES, C. M. et EXPOSITO, M. V. (2005). Framing Terminology : A Process Oriented Approach. *Meta : journal des traducteurs*, 50(4):1492–1421.
- FABER, P., MONTERO MARTÍNEZ, S., CASTRO PRIETO, M. R., SENSO RUIZ, J., PRIETO VELASCO, J. A., LEÓN ARAUZ, P., MÁRQUEZ LINARES, C. et VEGA EXPÓSITO, M. (2006). Process-oriented terminology management in the domain of Coastal Engineering. *Terminology*, 12(2):189–213.
- FELLBAUM, C. (1998). *WordNet : An Electronic Lexical Database*. MA : MIT Press.
- FILLMORE, C. (1982). Frame semantics. *Linguistic Society of Korea (ed.) Linguistics in the Morning Calm*, pages 111–137.
- FILLMORE, C. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6:222–254.
- FILLMORE, C. J. et ATKINS, B. T. S. (1998). FrameNet and lexicographic relevance. *In Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain*.
- FRANTZI, K., ANANIADOU, S. et MIMA, H. (2000). Automatic Recognition of Multi-Word Terms : the C-value/NC-value Method. *International Journal on Digital Libraries*, 3(2):115–130.
- FRASER, A., WELLER, M., CAHILL, A. et CAP, F. (2012). Modeling inflection and word-formation in SMT. *In Proceedings of EACL*, pages 664–674.

BIBLIOGRAPHIE

- FRITZINGER, F. et FRASER, A. (2010). How to avoid burning ducks : combining linguistic analysis and corpus statistics for German compound processing. *In Proceedings of the 5th Workshop on Statistical Machine Translation*, pages 224–234.
- GARERA, N. et YAROWSKY, D. (2008). Translating Compounds by Learning Component Gloss Translation Models via Multiple Languages. *In Proceedings of IJCNLP*, pages 403–410.
- GETOOR, L., FRIEDMAN, N., KOLLER, D., PFEFFER, A. et TASKAR, B. (2007). Probabilistic relational models. *In Introduction to Statistical Relational Learning*, pages 129–174.
- GIUDICELLI, V. (2013). Ontologies : Langage OWL (IMGT-ONTOLOGY).
- GOUADEC, D. (1990). *Terminologie. Constitution des données*. AFNOR Gestion. Afnor édition.
- GRUBER, T. (1993). Toward principles for the design of ontologies used for knowledge sharing? *International Journal Human-Computer Studies*, 43(5-6):907–928.
- GUARINO, N., OBERLE, D. et STAAB, S. (2009). What Is an Ontology? *In STAAB, S. et STUDER, R., éditeurs : Handbook on Ontologies*, pages 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg.
- GUARINO, N. et WELTY, C. a. (2009). An overview of OntoClean. *Handbook on ontologies*, pages 1–20.
- HAASE, P., LEWEN, H., STUDER, R. et TRAN, D. (2008). The neon ontology engineering toolkit. *In Proceedings of WWW*, Beijing, China.
- HARASTANI, R., DAILLE, B. et MORIN, E. (2012). Neoclassical compound alignments from comparable corpora. *In Computational Linguistics and Intelligent Text Processing - 13th International Conference*, pages 72–82, New Delhi, India.
- HATHOUT, N. (2011). Morphonette : a morphological network of French. *Lingue e Linguaggio*, 2:243–262.
- HATIER, S. (2013). Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. *In Actes de RECITAL*, pages 17–21.
- HEARST, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In COLING '92 Proceedings of the 14th conference on Computational linguistics*, pages 539–545.
- HOFFART, J., LEWIS-KELHAM, E., SUCHANEK, F. M., MELO, G. D., BERBERICH, K. et WEIKUM, G. (2011). YAGO2 : Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. *Time*, 23(7):229–232.

BIBLIOGRAPHIE

- HRISTOVSKI, D., DINEVSKI, D., KASTRIN, A. et RINDFLESCH, T. C. (2015). Biomedical question answering using semantic relations. *BMC bioinformatics*, 16:6.
- JACQUEMIN, C., KLAVANS, J. et TZOUKERMANN, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. *In Proceedings of ACL*, pages 24–31, Madrid, Spain.
- JACQUEY, E., TUTIN, A., KISTER, L., JACQUES, M.-p., HATIER, S. et OLLINGER, S. (2013). Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. *In Terminologie et Intelligence Artificielle (TIA)*, pages 121–128, Villetaneuse, France.
- KERGOSIEN, E., KAMEL, M., SALLABERRY, C., BESSAGNET, M.-N., GILLES, N. A. et GAIO, M. (2009). Construction et enrichissement automatique d’ontologie à partir de ressources externes. *JFO’09 : 3es Journées Francophones sur les Ontologies*, pages 1–10.
- KISTER, L., JACQUEY, E. et GAIFFE, B. (2011). Du thesaurus à l’onto-terminologie : relations sémantiques vs relations ontologiques. *Corela*, 9(1):1–14.
- KORENCHUK, Y. (2014). Extraction terminologique : vers la minimisation de ressources. *In Actes de TALN-RECITAL 2014*, volume 2, pages 59–70.
- KRÖTZSCH, M., SIMANCIK, F. et HORROCKS, I. (2014). A description logic primer. *In LEHMANN, J. et VÖLKER, J., éditeurs : Perspectives on Ontology Learning*, numéro June, pages 1–17.
- LARDILLEUX, A. et LEPAGE, Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. *In Proceedings of AMTA*, pages 125–132.
- LARDILLEUX, A. et LEPAGE, Y. (2009a). Anymalign : un outil d’alignement sous-phrastique libre pour les êtres humains. *In Actes de TALN*, pages 24–26.
- LARDILLEUX, A. et LEPAGE, Y. (2009b). Sampling-based multilingual alignment. *In Proceedings of RANLP*, pages 214–218.
- LARDILLEUX, A., YVON, F. et LEPAGE, Y. (2013). Generalizing Sampling-Based Multilingual Alignment. *Machine Translation*, 27(1):1–23.
- LAROCHE, A. et LANGLAIS, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In Proceedings of COLING*, pages 617–625.
- ŁAWRYNOWICZ, A. et TRESP, V. (2014). Introducing Machine Learning. *In LEHMANN, J. et VÖLKER, J., éditeurs : Perspectives on Ontology Learning*, pages 36–50.
- LEHMANN, J. et FANIZZI, N. (2014). Concept Learning. *In Perspectives on Ontology Learning*, pages 71–91.

BIBLIOGRAPHIE

- L'HOMME, M. et POLGUÈRE, A. (2008). Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. In MANIEZ, F. et DURY, P., éditeurs : *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*, pages 191–206. Presses universitaires de Lyon.
- L'HOMME, M.-C. (2004). Sélection de termes dans un dictionnaire d'informatique : Comparaison de corpus et critères lexicosémantiques. In *Proceedings of Euralex*, pages 583–593, Lorient, France.
- L'HOMME, M.-C. (2005). Sur la notion de « terme ». *Meta : Journal des traducteurs*, 50(4):1112–1132.
- L'HOMME, M.-C. (2016). Terminologie de l'environnement et Sémantique des cadres. In *Actes du 5e Congrès Mondial de Linguistique Française, SHS Web of Conferences*, volume 27, page 05010.
- LIANG, A. C., LAUSER, B., SINI, M., KEIZER, J. et KATZ, S. (2006). From AGROVOC to the agricultural ontology service / concept server An OWL model for managing ontologies in the agricultural domain. *CEUR Workshop Proceedings*, 216.
- LIN, C. et HOVY, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.
- LING, W., LUIS, T., MARUJO, L., ASTUDILLO, R. F., AMIR, S., DYER, C., BLACK, A. W. et TRANCOSO, I. (2015). Finding Function in Form : Compositional Character Models for Open Vocabulary Word Representation. *Proceedings of EMNLP*, pages 1520–1530.
- LUONG, M.-T. et MANNING, C. D. (2016). Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. *Proceedings of ACL*, pages 1054–1063.
- MAHDISOLTANI, F., BIEGA, J. et SUCHANEK, F. M. (2015). YAGO3 : A Knowledge Base from Multilingual Wikipedias. In *Conference on Innovative Data Systems Research (CIDR)*.
- MAUREL, D. (2008). Prolexbase : A multilingual relational lexical database of proper names. In CALZOLARI, N., CHOUKRI, K., MAEGAARD, B., MARIANI, J., ODJIK, J., PIPERIDIS, S. et TAPIAS, D., éditeurs : *6th Language Resources and Evaluation Conference LREC 2008*, pages 334–338. European Language Resources Association (ELRA).
- MAYNARD, D. et BONTCHEVA, K. (2014). Natural Language Processing. In LEHMANN, J. et VÖLKER, J., éditeurs : *Perspectives on Ontology Learning*, pages 51–67.
- MAYNARD, D., FUNK, A. et PETERS, W. (2009). SPRAT : a tool for automatic semantic pattern-based ontology population. In *International Conference for Digital Libraries and the Semantic Web*, pages 1–15, Trento, Italy. Citeseer.

BIBLIOGRAPHIE

- MCNAMEE, P. (2008). N-gram Tokenization for Indian Language Text Retrieval. *In Working Notes of the Forum for Information Retrieval Evaluation*, pages 73 – 97.
- MCNAMEE, P., NICHOLAS, C. et MAYFIELD, J. (2009). Addressing morphological variation in alphabetic languages. *In 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 75–82.
- MIKOLOV, T., CHEN, K., CORRADO, G. et DEAN, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *In Advances in Neural Information Processing Systems*, pages 3111–3119.
- MILES, A., MATTHEWS, B., WILSON, M. et BRICKLEY, D. (2005). SKOS Core : Simple Knowledge Organisation for the Web. *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications : Vocabularies in Practice*, (5):1–9.
- MITCHELL, T., COHEN, W., HRUSCHKA, E., TALUKDAR, P., BETTERIDGE, J., CARLSON, A., DALVI, B., GARDNER, M., KISIEL, B., KRISHNAMURTHY, J., LAO, N., MAZAITIS, K., MOHAMMAD, T., NAKASHOLE, N., PLATANIOS, E., RITTER, A., SAMADI, M., SETTLES, B., WANG, R., WIJAYA, D., GUPTA, A., CHEN, X., SAPAROV, A., GREAVES, M. et WELLING, J. (2015). Never-Ending Learning. *In Proceedings of the Conference on Artificial Intelligence (AAAI)*.
- MORIN, E. et DAILLE, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. *In Actes de la Conférence conjointe JEP-TALN-RECITAL*, pages 141–154.
- NAKASHOLE, N., WEIKUM, G. et SUCHANEK, F. M. (2012). PATTY : A Taxonomy of Relational Patterns with Semantic Types. *In Proceedings of EMNLP-CoNLL*, pages 1135–1145.
- NAKOV, P. et TIEDEMANN, J. (2012). Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 301–305.
- NAVIGLI, R. (2013). BabelNet and Friends : A manifesto for multilingual semantic processing. *Intelligenza Artificiale*, 7:165–181.
- NAVIGLI, R. et VELARDI, P. (2008). From Glossaries to Ontologies : Extracting Semantic Structure from Textual Definitions. *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, (January):71–87.
- NAVLEA, M. (2014). *La traduction automatique statistique factorisée : une application à la paire de langues français-roumain*. Thèse de doctorat, Université de Strasbourg.
- NAVLEA, M. et TODIRASCU, A. (2011). Using Cognates in a French-Romanian Lexical Alignment System : A Comparative Study. *In Proceedings of RANLP*, pages 247–253.

BIBLIOGRAPHIE

- NOY, N. et MCGUINNESS, D. (2001a). Développement d'une ontologie 101 : Guide pour la création de votre première ontologie. *Traduction de l'anglais par Anila Angjeli, BnF, Bureau de normalisation documentaire*, pages 1–26.
- NOY, N. F. et MCGUINNESS, D. L. (2001b). Ontology Development 101 : A Guide to Creating Your First Ontology.
- OCH, F. J. et NEY, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.
- OMRANE, N., NAZARENKO, A. et SZULMAN, S. (2011a). Le poids des entités nommées dans le filtrage des termes d'un domaine. *In Proceedings of TIA*, pages 80–86.
- OMRANE, N., NAZARENKO, A. et SZULMAN, S. (2011b). Les entités nommées : des clés linguistiques pour la conceptualisation. *In Proceedings of IC*, pages 435–450, Chambéry, France.
- OOSTDIJK, N., VERBERNE, S. et KOSTER, C. (2010). Constructing a broad-coverage lexicon for text mining in the patent domain. *In Proceedings of LREC*, pages 2292–2299.
- OROBINSKA, O., LYON, E. et CHAUCHAT, J.-h. (2013). Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir d'un corpus spécialisé. *In Terminologie et Intelligence Artificielle (TIA)*, pages 129–136, Villetaneuse, France.
- PÉREZ, A., BAONZA, M. d. F. et VILLAZÓN, B. (2008). Neon methodology for building ontology networks : Ontology specification. Rapport technique February.
- PETASIS, G., KARKALETSIS, V., PALIOURAS, G., KRITHARA, A. et ZAVITSANOS, E. (2011). Ontology Population and Enrichment : State of the Art. *In PALIOURAS, G., SPYROPOULOS, C. D. et TSATSARONIS, G., éditeurs : Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, pages 134–166. Springer Berlin Heidelberg.
- POLGUÈRE, A. (2012). Propriétés sémantiques et combinatoires des quasi-prédicats sémantiques. *Scolia*, 26:131–152.
- POON, H. et DOMINGOS, P. (2010). Unsupervised ontology induction from text. *In Proceedings of ACL*, pages 296–305.
- RETTINGER, A., NICKLES, M. et TRESP, V. (2009). Statistical relational learning with formal ontologies. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5782 LNAI(PART 2):286–301.
- ROCHETEAU, J. et DAILLE, B. (2011). TTC TermSuite : A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. *5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 9–12.

BIBLIOGRAPHIE

- SANCHEZ-CISNEROS, D. et GALI, F. (2013). UEM-UC3M : An Ontology-based named entity recognition system for biomedical texts. *In Proceedings of SemEval*, pages 622–627.
- SAVARY, A., MANICKI, L. et BARON, M. (2013). Populating a multilingual ontology of proper names from open sources. *Journal of Language Modelling*, 1(2):189–225.
- SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester.
- SCHONE, P. et JURAFSKY, D. (2000). Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. *In Proceedings of COLING*, pages 67–72.
- SENNRICH, R., HADDOW, B. et BIRCH, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of ACL*, pages 1715–1725.
- SHVAIKO, P. et EUZENAT, J. (2013). Ontology matching : state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers (IEEE)*, 25(1):158–176.
- SIMPERL, E. (2014). Learning Ontologies via Games with a Purpose. *In Perspectives on Ontology Learning*, pages 225–246.
- STAAB, S. et STUDER, R. (2009). *Handbook on ontologies*. Springer Berlin Heidelberg, 2nd editio édition.
- SUÁREZ DE FIGUEROA BAONZA, M. C., GÓMEZ-PÉREZ, A. et FERNÁNDEZ-LÓPEZ, M. (2012). The NeOn Methodology for Ontology Engineering. *In SUÁREZ-FIGUEROA, M. C., GÓMEZ-PÉREZ, A., MOTTA, E. et GANGEMI, A., éditeurs : Ontology Engineering in a Networked World*, pages 9–34. Springer Berlin Heidelberg, Berlin, Heidelberg.
- SUÁREZ-FIGUEROA, M. (2010). *NeOn Methodology for building ontology networks : specification, scheduling and reuse*. Thèse de doctorat, Universidad Politécnica de Madrid.
- SUCHANEK, F. M. (2008). *Automated construction and growth of a large ontology*. Thèse de doctorat, Saarland University.
- SUCHANEK, F. M. (2014). Information Extraction for Ontology Learning. *In LEHMANN, J. et VÖLKER, J., éditeurs : Perspectives on Ontology Learning*.
- SYLLABS (2012). TTC project : Compilation of Monolingual Reference Term Lists (ID D3.2 – Chap3). Rapport technique.
- TESNIÈRE, L. (1959). *Éléments de syntaxe structurale*. Librairie C. Klincksieck, Paris.
- TIEDEMANN, J. (2012). Character-Based Pivot Translation for Under-Resourced Languages and Domains. *In Proceedings of EACL*, pages 141–151.

BIBLIOGRAPHIE

- TUTIN, A. (2007a). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, XII:5–14.
- TUTIN, A. (2007b). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. *In Actes de TALN*, pages 283–292.
- VAN CAMPENHOUDT, M. (2010). Le terme : condensation syntaxique et condensation des connaissances en langue spécialisée. *Les choses, les notions, les noms : le terme dans tous ses états, Romanica Wratislaviensia*, 57:29–46.
- VELARDI, P., FARALLI, S. et NAVIGLI, R. (2013). OntoLearn Reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.
- VELARDI, P., NAVIGLI, R., CUCCHIARELLI, A. et NERI, F. (2006). Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. *In* BUITELAAR, P., CIMIANO, P. et MAGNINI, B., éditeurs : *Ontology Learning from Text : Methods, Applications and Evaluation*, pages 1–15. IOS Press.
- VERGNE, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN*, 2:139–148.
- VERGNE, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. *In Actes de JADT*, pages 1157–1164, Louvain, Belgique.
- VERGNE, J. (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *In Actes de CIDE*, pages 155–168.
- VILARES, J., OAKES, M. et FERRO, M. V. (2007a). English-to-French CLIR : A Knowledge-Light Approach through Character N-Grams Alignment. *In Advances in Multilingual and Multimodal Information Retrieval*, pages 148–155.
- VILARES, J., OAKES, M. P. et VILARES, M. (2007b). Character N-Grams Translation in Cross-Language Information Retrieval. *In Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems : Natural Language Processing and Information Systems, NLDB '07*, pages 217–228, Berlin, Heidelberg. Springer-Verlag.
- VILARES, J., VILARES, M., ALONSO, M. a. et OAKES, M. P. (2016). On the feasibility of character n-grams pseudo-translation for Cross-Language Information Retrieval tasks. *Computer Speech & Language*, 36:136–164.
- WÄSCHLE, K. et RIEZLER, S. (2012a). Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27.
- WÄSCHLE, K. et RIEZLER, S. (2012b). Structural and Topical Dimensions in Multi-Task Patent Translation. *In Proceedings of EACL*, pages 818–828, Avignon, France.

BIBLIOGRAPHIE

- WELLER, M. et BLANCAFORT, H. (2011). Terminology extraction and term variation patterns : a study of French and German data. *In Jahrestagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)*.
- WELLER, M., FRASER, A. et HEID, U. (2014). Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT. *In Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 11–18.
- WELLER, M., FRASER, A. et SCHULTE IM WALDE, S. (2013). Using subcategorization knowledge to improve case prediction for translation to German. *In Proceedings of ACL*, volume 1, pages 593–603.
- WELLER, M., GOJUN, A. et HEID, U. (2011). Simple methods for dealing with term variation and term alignment. *In Proceedings of TIA*, pages 87–93.
- WELLER, M. et HEID, U. (2012). Analyzing and Aligning German compound nouns. *In Proceedings of LREC*, pages 2395–2400.
- WITSCHER, H. F. (2005). Terminology Extraction and Automatic Indexing. *In Terminology and Knowledge Engineering (TKE)*, pages 1–12.
- WU, F. et WELD, D. S. (2008). Automatically Refining the Wikipedia Infobox Ontology. *In Proceeding of the 17th international conference on World Wide Web WWW 08*, numéro January, pages 635–644.
- XU, Z., TRESP, V., YU, K. et KRIEGEL, H.-P. (2006). Infinite Hidden Relational Models. *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence*.

Liste de publications

Articles publiés dans des conférences internationales avec comité de lecture

Korenychuk, Y. (2015), Dealing with Large Corpora for Ontology Population. 11th International Conference on Terminology and Artificial Intelligence (TIA) "Multilingual Terminology, Ontologies and Knowledge Representation Models", Grenade, Espagne, les 4-6 novembre 2015, pp. 213-216.

Korenychuk, Y. (2015), Identification automatique de familles morphologiques endogènes et de néologismes. Les 10e journées internationales LTT (Lexicologie, Terminologie, Traduction), Strasbourg, France, le 4 septembre 2015, (en cours de publication).

Articles publiés dans des conférences nationales avec comité de lecture

Korenychuk, Y. (2014), Extraction terminologique : vers la minimisation de ressources. RECITAL - Traitement Automatique du Langage Naturel, Marseille, France, 1-4 juillet 2014, pp. 59-70.

Communications sans publication

Korenychuk, Y. (2017), Présentation des résultats de l'extraction terminologique sous forme de familles morphologiques. Atelier Acquisition terminologique et défi du multilinguisme, Congrès Mondial de Traductologie, Nanterre, France, le 12 avril 2017

Korenychuk, Y. (2015), Doctorat et entreprise : les clés du succès. Séminaire FDT, Strasbourg, France, le 29 mai 2015

Korenychuk, Y. (2015), Processus d'élargissement et d'enrichissement des ontologies. Soutenance à mi-parcours. Doctoriales de LiLPa, Strasbourg, France, le 12 juin 2015

Korenychuk, Y. (2014), Utilisation des ressources endogènes dans l'extraction terminologique multilingue. Atelier corpus multilingues des Journées internationales d'Analyse statistique des Données Textuelles (JADT), Paris, France, le 3 juin 2014

Korenychuk, Y. (2014), Formes émergentes dans l'extraction terminologique. Doctoriales de LiLPa, Strasbourg, France, le 16 mai 2014

Korenychuk, Y. (2014), Formes émergentes dans l'extraction terminologique. Journée doctorale "Formes émergentes", École doctorale des Humanités (ED520), Strasbourg, France, le 12 avril 2014

Méthode d'enrichissement et d'élargissement d'une ontologie à partir de corpus de spécialité multilingues

Résumé

Cette thèse propose une méthode pour alimenter une ontologie, une structure de concepts liés par des relations sémantiques, par des termes français, anglais et allemands à partir de corpus spécialisés comparables. Son apport principal est le développement des méthodes d'extraction utilisant des ressources endogènes apprises à partir de corpus et d'ontologie. Exploitant des n-grammes de caractères, elles sont disponibles et indépendantes vis-à-vis à la langue et au domaine. La première contribution porte sur l'utilisation des ressources morphologiques et morphosyntaxiques endogènes pour extraire des termes mono- et polylexicaux à partir de corpus. La deuxième contribution vise à exploiter des ressources endogènes pour identifier leurs traductions. La troisième contribution concerne la construction des familles morphologiques endogènes servant à alimenter l'ontologie.

Mots-clés : ontologies, extraction terminologique, alignement bilingue de termes, corpus comparables, ressources endogènes

Résumé en anglais

This thesis proposes a method of enrichment and population of an ontology, a structure of concepts linked by semantic relations, by terms in French, English and German from comparable domain-specific corpora. Our main contribution is the development of extraction methods based on endogenous resources, learned from the corpus and the ontology being analyzed. Using character n-grams, these resources are available and independent of a particular language or domain. The first contribution concerns the use of endogenous morphological and morphosyntactic resources for mono- and polylexical terms extraction from the corpus. The second contribution aims to use endogenous resources to identify translations for these terms. The third contribution concerns the construction of endogenous morphological families designed to enrich and populate the ontology.

Keywords: ontologies, terminology extraction, bilingual alignment of terms, comparable corpora, endogenous resources