Université de Montréal

Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue

par

Sahara Iveth Carreño Cruz

Département de linguistique et de traduction

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de Maître

en traduction

option recherche

Décembre 2004

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue

présenté par :
Sahara Iveth Carreño Cruz

a été évalué par un jury composé des personnes suivantes :

Gilles Bélanger
Président-rapporteur

Marie-Claude L'Homme
Directrice de recherche

Patrick Drouin
Membre du jury

Mémoire accepté le 8 février 2005

Résumé

Le présent travail a pour objectif la caractérisation de la variation terminologique en corpus parallèle spécialisé. L'étude est entreprise afin de fournir des pistes contribuant à la prise en compte de ce phénomène en extraction de termes bilingue.

Depuis quelques années, la variation terminologique a fait l'objet de plusieurs études et la plupart d'elles ont été réalisées sur des corpus unilingues. Il existe peu de travaux envisageant ce phénomène linguistique du point de vue d'un corpus bilingue. Pourtant, nous pensons qu'il convient de continuer les recherches dans cette direction, car la variation terminologique a une incidence importante sur l'identification et l'extraction automatiques de termes bilingue.

Pour mener à bien cette étude, nous avons constitué un corpus parallèle de 23 documents originaux anglais et de leur traduction espagnole. Les textes sélectionnés relèvent du domaine de l'environnement et portent sur les composés chimiques présentant des risques pour la santé et pour l'environnement. Le corpus, qui contient environ 350 000 mots par langue, a ensuite été aligné. À la suite d'une présélection et de l'application de certains critères, nous avons constitué une liste de 25 termes simples et de 25 termes complexes anglais. Enfin, nous avons procédé à l'analyse de la façon dont chacun des termes anglais a été rendu dans les textes traduits.

Nous avons observé divers cas de variation sémantique, syntaxique et morphosyntaxique pour les deux types de termes (simples et complexes). Le taux de variation estimé pour les termes simples est de 4,08 et de 4,44 pour les termes complexes. La fréquence, dans notre corpus, de certaines variations non relevées par les études unilingues montre que la variation de termes affecte la qualité des

extracteurs automatiques bilingues, surtout lors de la mise en correspondance des termes source et cible.

Mots-clés : variation terminologique, variante terminologique, terme de base, terme simple, terme complexe, extraction automatique de termes, extraction de termes bilingue, corpus parallèle, corpus aligné, terminologie de l'environnement

Abstract

This research is a study of term variation in specialized parallel corpora. The problem is characterized from the point of view of the handling of different terminological forms by bilingual term extractors.

During the last two decades, term variation has been the subject of several studies, and most of them have been conducted on monolingual corpora. There are very few studies approaching this linguistic phenomenon from a bilingual perspective. However, we believe that more research on this subject is needed, for term variation has an important impact on automatic bilingual recognition and extraction of terms.

For our own study, we constructed a parallel corpus containing 23 original documents in English and their corresponding translation in Spanish. The selected texts belong to the field of environment and deal with industrial chemicals that threaten human health and the environment. The corpus, which contains approximately 350 000 words per language, has then been aligned. Following a preselection process and the application of certain criteria, a list of 25 English simple terms and 25 English complex terms has been defined. We then proceeded to the observation of the ways each selected term had been represented in the translated texts.

We observed various cases of semantic, syntactic and morphosyntactic variations for both types of terms (simples and complex). The analysis yielded a variation rate of 4.08 for the simple terms, and of 4.44 for the complex terms. The frequency of certain types of variation in our corpus that were not detected in

monolingual studies shows that term variation affects the performance of bilingual automatic extraction systems, particularly during source-target term alignment.

Keywords : term variation, term variant, base term, simple term, complex term, automatic term extraction, bilingual term extraction, parallel corpus, aligned corpus, environmental terminology

Table des matières

Liste des tableaux

Liste des figures

À ma merveilleuse famille

Remerciements

Avec tout mon respect et mon admiration, je remercie sincèrement ma directrice de recherche, Marie-Claude L'Homme, pour son infinie patience, ses sages conseils, sa grande générosité et, surtout, pour m'avoir fait confiance à tous moments. J'ai beaucoup appris de toi.

Je remercie également le professeur Patrick Drouin pour sa constante disponibilité et son soutien technique dans la préparation de mon corpus d'étude.

Mes chaleureux remerciements vont à ma chère amie et mentor, Gertrudis Payàs, qui a toujours su m'encourager à entreprendre de nouvelles aventures.

À mes parents, mes frères, mes soeurs et toute ma famille mexicaine, qui m'ont donné leur grand soutien moral et financier et qui ont toujours cru en moi : ¡muchas gracias!

Je remercie infiniment mon ami Jean-Sébastien d'avoir consacré beaucoup de son temps à m'aider à améliorer mon français et de m'avoir redonné confiance dans les moments de doute.

La dernière mais non la moindre; je désire remercier Elizabeth Marshman, qui est toujours prête à aider et à encourager ses amis. Tu es un exemple à suivre.

Introduction

Aujourd'hui, le phénomène linguistique de la variation est un sujet d'étude important dans le domaine de la terminologie, plus particulièrement en terminologie textuelle bilingue et multilingue.

Les terminologues, traducteurs et autres spécialistes qui doivent réaliser de grands projets terminologiques multilingues – tels que glossaires, dictionnaires spécialisés et banques terminologiques – dans des délais très courts ont souvent besoin de ressources informatiques afin de faciliter au moins une partie de leur travail. Dans ce contexte, les extracteurs automatiques de termes bilingues, qui s'appuient très souvent sur des corpus de traductions, se révèlent des outils précieux.

Malheureusement, ces extracteurs automatiques sont encore loin de produire des résultats parfaits. La plupart d'entre eux restent limités en ce qui concerne l'identification d'unités terminologiques véritables et la mise en correspondance des termes source avec leurs équivalents dans d'autres langues.

Nous considérons que la variation terminologique joue un rôle important dans ces deux aspects de l'extraction automatique. En effet, lorsqu'un terme subit des transformations sur le plan formel, la précision de la reconnaissance des équivalents pour ce terme est considérablement affectée.

Malgré tout, au meilleur de notre connaissance, presqu'aucun extracteur de termes bilingue ne prend en considération les diverses variations que subissent les termes. En fait, la variation terminologique dans un contexte bilingue, et plus particulièrement dans un contexte de traduction, n'a pas été suffisamment étudiée. Nous pensons qu'il existe des cas de variation de termes que les études unilingues ne relèvent pas et qui peuvent être importants en extraction automatique bilingue.

À la lumière de ces observations, nous avons décidé de réaliser une analyse de la variation terminologique dans un corpus spécialisé composé de textes originaux anglais et de leur traduction en espagnol.

Notre objectif est donc de caractériser la variation terminologique en traduction et de mettre l'accent sur les conséquences de ce phénomène pour les extracteurs de termes bilingues.

Pour atteindre cet objectif, nous avons sélectionné un ensemble de textes anglais et espagnols du domaine de l'environnement et portant sur les composés chimiques industriels qui posent des risques pour la santé et l'environnement. Nous avons ainsi constitué un corpus parallèle d'environ 350 000 mots par langue.

À partir du corpus anglais, nous avons sélectionné 25 termes simples et 25 termes complexes qui font l'objet de notre analyse. Ils convient de mentionner que la plupart des études portant sur la variation terminologique ont porté sur les termes complexes de nature nominale. De même, les extracteurs automatiques bilingues visent principalement les termes complexes. Toutefois, nous croyons que tant les termes complexes que les termes simples, qu'ils soient des noms, des adjectifs ou des verbes, sont susceptibles de subir des variations et, en conséquence, il faut les prendre tous en considération.

Ensuite, nous avons observé la façon dont les 50 termes choisis sont représentés dans le corpus espagnol. Dans une base de données, conçue spécialement pour les fins de cette recherche, nous avons enregistré toutes les variations que chacun de nos termes ont subies.

Ainsi, en nous appuyant sur la classification des variantes proposée dans le premier chapitre du mémoire, nous décrivons tous les cas de variation sémantique, syntaxique et morphosyntaxique observés dans notre corpus parallèle. Bien sûr,

nous apportons également des informations sur la fréquence de chaque type de variation, la comparaison entre les types de variation pour les termes simples et pour les termes complexes, les cas particuliers de variation en traduction et, enfin, les cas de variation qui n'ont pas été considérés au début, dans la classification proposée.

La présente étude est organisée en trois chapitres.

Le premier chapitre présente les trois éléments thématiques qui constituent la base de notre recherche. Dans la première section, nous définissons la notion de « variation terminologique » et présentons une classification générale des variantes. Dans la deuxième section, nous décrivons cinq travaux portant sur la variation en terminologie. Les quatre premiers ont été réalisés sur une seule langue; le cinquième a envisagé le problème du point de vue d'un corpus bilingue. Dans la troisième section, nous discutons des principes généraux de l'extraction automatique et, plus particulièrement, de l'extraction automatique de termes basée sur des corpus parallèles.

Le deuxième chapitre présente la méthodologie adoptée pour notre recherche. D'abord, les critères de sélection des textes formant notre corpus parallèle sont décrits. Ensuite, nous exposons la méthode d'alignement des textes du corpus parallèle ainsi que les outils employés pour y parvenir. Dans la troisième section, nous décrivons les étapes de présélection et de sélection des 50 termes faisant l'objet de notre analyse et les critères que nous avons établis pour cette sélection. Ensuite, nous expliquons comment nous avons recueilli les termes et leurs diverses variantes à partir de l'observation du corpus parallèle. Nous concluons le deuxième chapitre par la présentation de la méthode d'analyse de la variation des termes sélectionnés.

Le troisième et dernier chapitre est consacré à la présentation détaillée des résultats de l'analyse des termes choisis. La section 3.1 contient les critères

appliqués pour sélectionner les termes de base espagnols à partir desquels les variantes ont été produites. La section 3.2 décrit les divers cas de variation sémantique, syntaxique et morphosyntaxique observés dans les 25 termes simples choisis. La section 3.3 correspond à la description des variantes observées dans les termes complexes sélectionnés. Dans la section 3.4, nous réalisons une comparaison entre les types de variantes et la fréquence de ces variantes observées dans les termes simples par rapport aux termes complexes. Enfin, dans la section 3.5, nous analysons quelques implications que le phénomène de la variation terminologique et les résultats obtenus par notre analyse ont pour les extracteurs automatiques bilingues existants.

Chapitre 1 : État de la question

La présente étude a pour objectif de caractériser la variation des termes en corpus bilingue spécialisé. Nous réalisons une analyse comparative des variations dans un corpus parallèle composé de textes anglais et espagnols. Nous jugeons donc pertinent de commencer par décrire le concept de « variation terminologique » sur lequel l'étude se fonde. Ainsi, dans la première section de ce chapitre, nous présentons une définition de « variation terminologique » et une classification des variations sur laquelle notre analyse s'aligne.

Dans la deuxième section, nous exposons quelques travaux importants dont l'objet d'étude est la variation en terminologie. D'abord, nous présentons les travaux portant sur une langue et ensuite un travail très récent qui analyse certains types de variation en traduction. Il s'agit donc d'une analyse bilingue.

Puisque notre étude s'inscrit dans le domaine de l'extraction automatique de termes, nous décrivons, dans la troisième section, les principes de l'extraction de terminologie et expliquons les principales méthodes d'extraction automatique de termes se basant sur des corpus parallèles. Nous concluons par une analyse des difficultés et des limites auxquelles les extracteurs automatiques bilingues font face à présent.

1.1 La variation terminologique

Depuis quelques années, le phénomène linguistique de la variation occupe une place importante en terminologie, en traitement automatique de la langue (TAL), en traduction automatique, en extraction automatique de termes et dans d'autres domaines reliés aux technologies de l'information. L'incidence de la

variation sur les résultats et les produits issus de ces domaines a été révélée par des études descriptives basées sur des corpus textuels.

Mais, qu'est ce que la variation ? Étant donné que la présente recherche s'inscrit dans le domaine de la terminologie, nous essayerons de décrire la variation telle qu'elle se présente dans ce domaine.

1.1.1 Définition de « variation terminologique »

En termes généraux, la variation terminologique est le phénomène selon lequel une même unité lexicale spécialisée (que nous appellerons *terme*) est représentée de différentes manières sur le plan formel. Une variante terminologique est ainsi un énoncé sémantiquement et conceptuellement relié au même terme d'origine (Daille *et al.* 1996).

Il convient de souligner trois éléments importants de cette définition. Le premier est associé au concept « terme d'origine ». Pour les fins de la présente recherche, *terme d'origine* ou *terme de base* est le terme à partir duquel se produisent les variantes. C'est l'unité qui fait l'objet du plus grand consensus dans un domaine. Il s'agit souvent de la forme apparaissant dans les dictionnaires spécialisés et les banques terminologiques. Par exemple, dans le domaine de l'environnement, le terme *elemento medioambiental* est le terme utilisé par les spécialistes pour faire référence à un secteur de l'environnement. Nous le considérons alors comme le *terme de base* ou *d'origine*. Toutefois, dans notre corpus espagnol, ce concept est représenté par d'autres formes lexicales : *medio ambiente*, *entorno ambiental*, *medio ambiental*. Il s'agit alors des variantes de notre terme d'origine.

Le deuxième élément est celui de l'équivalence conceptuelle. En principe, le terme d'origine et ses diverses variantes sont associés à un même concept. C'est la raison pour laquelle nous faisons référence aux différences sur le plan formel. Cependant, comme nous le verrons dans notre analyse, il existe certains types de variation qui peuvent entraîner une différence peu significative ou non significative sur le plan du sens.

Malgré tout, nous considérons que, même dans ce dernier cas, il existera toujours un lien sémantique fort entre un terme d'origine et ses variantes.

Le troisième élément est relié au concept « unité lexicale spécialisée ». Ce concept regroupe les termes simples (composés d'un seul mot) et les termes complexes (composés de deux ou de plusieurs mots). Ainsi, nous considérons que, bien que la plupart des études existantes sur la variation terminologique visent exclusivement les termes complexes, les termes simples peuvent également être l'objet d'une variation.

Dans la section qui suit nous présentons une classification assez générale de variantes terminologiques sur laquelle s'aligne notre étude.

1.1.2 Typologie de la variation terminologique

Il existe une grande variété de formes de variation terminologique mais aucune typologie n'a encore fait l'unanimité. En effet, chaque étude, chaque groupe de chercheurs détermine sa propre classification en fonction des objectifs poursuivis, des langues traitées, de l'application visée ainsi que les types de termes envisagés.

La classification que nous présentons se base principalement sur les travaux de Carl *et al.* (2004), Daille *et al.* (1996), Daille (1995), Freixa (2001; 2002), Jacquemin (1999) Hamon *et al.* (1998) et Ibekwe-Sanjuan (1998a).

Sur le plan de la forme, une variante d'un terme peut se produire au niveau morphologique et au niveau syntaxique. Toutefois, la variation peut aussi se produire sur le plan sémantique.

1.1.2.1 Variation morphologique

Une variation morphologique comporte une modification d'un ou de plusieurs morphèmes du terme. Les cas les plus connus de variation morphologique sont reliés à la flexion. Les variations en genre et en nombre (*residuo – residuos*) dans certaines langues en est un bon exemple. Les flexions donnant lieu à un gérondif (*reciclando*) ou à un participe (*reciclado*) à partir d'un verbe est un autre exemple de variation morphologique.

1.1.2.2 Variation graphique et orthographique

Une unité lexicale peut subir des changements au niveau graphique et orthographique. L'alternance entre la majuscule et la minuscule (*risk assessment – Risk Assessment*), l'utilisation d'un trait d'union ou d'un espace dans un terme complexe (*long-range transport – long range transport*), la présence d'éléments typographiques tels que les parenthèses et les guillemets au milieu d'un terme (*"point" source*) sont des exemples de variations graphiques et orthographiques.

1.1.2.3 Variation syntaxique

Ce type de variation vise surtout les termes complexes, car elle implique un changement dans leur structure. Elle comprend, par exemple, l'insertion d'autres mots, la substitution de certains éléments du terme, le changement de l'ordre des unités entrant dans la composition du terme.

Pour mieux définir la variation syntaxique, nous présentons une subdivision selon les types de modifications syntaxiques que peut subir un terme complexe de base.

a) Variation par coordination

Ce phénomène fait référence à la fusion de deux ou de plus de deux termes dans un même syntagme lorsqu'ils possèdent une structure syntaxique similaire. Dans la variation par coordination, les termes combinés peuvent partager la même tête (*point and diffuse source – point source + diffuse source*) ou les mêmes modificateurs (*risk assessment and communication - risk assessment + risk communication*). De cette façon une structure plus compacte est produite.

b) Variation par insertion

Le fait d'ajouter des mots non grammaticaux (par exemple des adverbes, des adjectifs, des noms) dans la structure d'un terme complexe de base produit une variation par insertion. Les insertions modifient la structure originale du terme et peuvent impliquer une « substitution » ou une « modification » du terme de base.

La substitution fait référence à la transformation du terme de base en un autre terme, normalement plus spécifique – c'est-à-dire, un hyponyme (Daille *et al.* 1996 : 229), par exemple, *fuente puntual + atmosférica = fuente atmosférica puntual*, *body burden + lead = body lead burden.*

La modification se produit quand le mot inséré est un adverbe ou un adjectif qui qualifie le terme de base (Daille *et al.* 1996 : 229). Par exemple, dans le terme *gestión racional* l'adverbe *ambientalmente* a été ajouté. De cette façon la variante par modification *gestión ambientalmente racional* est produite.

c) Variation par juxtaposition

La juxtaposition est l'union d'un terme de base et d'un autre mot ou d'un ensemble de mots, union parfois indiquée par un élément graphique comme le trait d'union. À la différence de la variation par insertion, la juxtaposition est une composition qui ne coupe pas la structure linéaire du terme de base. La juxtaposition peut aussi entraîner la modification d'un terme de base ou générer un nouveau terme. Par exemple, la juxtaposition du terme de base *municipal solid waste* et du nom *management* produit un nouveau terme : *municipal solid waste management*. Voici un autre exemple : la juxtaposition de l'adverbe *environmentally* au terme de base *sound management* produit la variante *environmentally sound management*.

d) Variation par omission

Quand certains éléments de la structure originale d'un terme complexe sont éliminés (par exemple, pour des raisons d'économie lorsque le terme est souvent répété dans le texte) et que cela n'entraîne aucun changement conceptuel, on parle d'une variation terminologique par omission. Par exemple, dans notre corpus

anglais, le terme de base *landfill site* apparaît plus souvent dans sa forme réduite : *landfill*. Les phénomènes d'abréviation et de siglaison sont aussi considérés comme étant des variations par omission. Par exemple, l'acronyme *POP* est souvent utilisé au lieu du terme de base *persistent organic pollutant*. *POP* est considéré comme une variante par omission.

e) Variation par permutation

La permutation fait référence au changement structural d'un terme causé par la présence d'un mot ou d'un ensemble de mots qui permutent autour d'un élément pivot du terme (Daille *et al.* 1996 : 213). Dans le cas de la langue anglaise, les prépositions sont assez souvent les éléments provoquant la variation par permutation. Par exemple, le syntagme *assessment of the risks* trouvé dans le corpus anglais est considéré comme étant une variante du terme *risk assessment*.

1.1.2.4 Variation morphosyntaxique

Ce type de variation fait référence plus particulièrement à un changement se situant sur le plan de la morphologie dérivationelle. La nominalisation des adjectifs, la nominalisation des verbes, l'adjectivation des noms, et tout autre changement de la catégorie grammaticale d'un terme simple ou complexe constitue une variation morphosyntaxique. L'utilisation du syntagme verbal *evaluar los riesgos* au lieu du terme complexe *evaluación de riesgos* (syntagme nominal) est un exemple de variation morphosyntaxique.

1.1.2.5 Variation sémantique

La variation sémantique implique parfois une modification sur le plan conceptuel. En terminologie, quelques phénomènes décrits au préalable, tels que la coordination, l'insertion et la juxtaposition, peuvent entraîner un changement de sens par rapport au terme de base. Ainsi, ces transformations produisent d'autres termes entretenant différentes relations sémantiques avec le terme de base, comme une relation de méronymie ou d'hyponymie. Par exemple, les variantes *plan regional de acción* et *plan nacional de acción* ont été produites lors d'insertions dans le terme de base *plan de acción*. Le terme de base et ses variantes entretiennent une relation d'hyponymie.

La forme de variation sémantique la plus fréquente est la synonymie qui consiste à utiliser une unité lexicale complètement différente du terme de base mais qui fait référence au même concept que celui-ci. Par exemple, dans notre corpus espagnol, le terme de base espagnol *liberación* est parfois remplacé par l'un de ses synonymes : *emanación*, *descarga*, *emisión*.

Évidemment, il n'est pas facile d'aborder la synonymie comme une forme de variation terminologique parce qu'il s'agit d'un phénomène beaucoup plus complexe qui relève assez souvent des discussions théoriques controversées quant à l'équivalence conceptuelle entre termes.

Malgré cet inconvénient, nous avons décidé d'inclure ce type de variation parce que nous supposons qu'il s'agit de l'une des variations importantes dans la présente recherche et qu'elle apportera des informations utiles.

1.1.3 Les variations considérées dans cette étude

Compte tenu des termes sélectionnés, des objectifs de la présente recherche et surtout du fait qu'elle se base sur l'analyse d'un corpus bilingue contenant des textes originaux et des traductions, nous avons décidé de nous concentrer sur les types de variation suivants : syntaxique, morphosyntaxique et sémantique. Nous excluons donc la variation morphologique (flexionnelle) et les variantes graphiques et orthographiques parce que, d'après les études sur le sujet, ce sont les formes de variation qui sont facilement caractérisées. Leur reconnaissance dans les textes ne représente plus un problème du point de vue du TAL.

Pour notre analyse comparative ainsi que pour la présentation des résultats, nous allons nous aligner en premier lieu sur la classification des variantes terminologiques que nous venons de présenter. Cependant il est important de souligner que cette classification se base sur des analyses portant sur des corpus unilingues et visant plutôt les termes complexes. Par conséquent, il est fort probable que d'autres types de variation puissent apparaître lors de l'observation de notre corpus parallèle. Nous supposons que le fait d'analyser un corpus de traductions relèvera des phénomènes de variation qui échappent à l'observation des corpus unilingues. L'inclusion des termes simples dans notre analyse peut également apporter un éclairage différent sur la variation terminologique.

1.2 Études de la variation terminologique basées sur corpus

Dans cette section nous présentons quelques études importantes sur la variation terminologique qui se basent sur l'analyse de textes spécialisés. La première sous-section décrit quatre recherches portant sur la variation en corpus unilingue et la deuxième sous-section présente une étude qui a envisagé la variation

dans un corpus bilingue (anglais-français). La plupart de ces études portent sur les termes complexes.

1.2.1 Études unilingues

1.2.1.1 Caractérisation de la variation des termes complexes anglais

L'une des premières analyses exhaustives de la variation en terminologie est celle réalisée par Daille *et al.* (1996).

Cette recherche vise à démontrer l'incidence aux niveaux qualitatif et quantitatif de la variation sur l'identification de termes. Basée sur une observation empirique de termes dans trois corpus techniques (médecine, communication et physique), l'étude présente en détail divers patrons de variation terminologique et propose une série de règles de formation de variantes qui peuvent être appliquées soit en génération de variantes candidates, soit en identification de ces unités dans les textes.

Les auteurs illustrent l'importance du rôle de la variation pour la reconnaissance de termes et suggèrent que la création d'une méthode d'extraction et de description de variantes est utile dans divers champs du TAL, tels que l'indexation automatique, la catégorisation de textes, la traduction automatique et, évidemment, l'extraction automatique de terminologie.

L'analyse se concentre sur les termes complexes anglais binaires – constitués de deux unités lexicales, par exemple, *epithelial cell* –, car il s'agit, selon les chercheurs, des termes les plus fréquents. De plus, les termes à trois unités lexicales ou plus se forment assez souvent à partir de termes binaires ou de termes simples.

De toute façon, lors de l'application des règles de formation de variantes, les termes ternaires sont inclus et étudiés comme des variantes. Les termes binaires sont alors les termes de base pour l'analyse.

Même si quatre différents types de variantes ont été observés dans les corpus techniques sélectionnés (variantes graphiques et orthographiques, par exemple : *packet mode/paquet-mode;* variantes flexionnelles, par exemple : *acoustic test/acoustic testing;* variantes syntaxiques, par exemple : la coordination de deux termes dans le syntagme *systolic and diastolic blood pressure;* et variantes morphosyntaxiques, par exemple : les dérivés morphologiques), les chercheurs ont décidé d'analyser exclusivement les variantes syntaxiques.

Le travail de recherche peut se diviser en deux étapes. Dans une première étape, les patrons de formation des termes binaires les plus fréquents en anglais ont été établis. Après avoir observé un ensemble de termes sélectionné à l'aide d'un programme d'extraction de syntagmes nominaux pour chaque corpus, les chercheurs ont déterminé que trois structures morphosyntaxiques sont les plus fréquentes dans les termes binaires : a) structure adjectif + nom (A N : *thermal control*), b) structure nom + nom (N_2 N_1 : *data transmission*) et c) structure nom + préposition + nom (N_1 P N_2 : *mass in orbit*).

La deuxième étape est celle de la description détaillée et de la caractérisation des variantes syntaxiques observées dans les corpus. La caractérisation des variantes – que les chercheurs ont appelée *règles de description* – a été réalisée sur la base d'un modèle « lexico-syntaxique à deux niveaux » qui représente tant la structure idiomatique que la structure syntagmatique des termes. Voici un exemple de cette description pour le terme *coronary artery* (Daille *et al.* 1996 : 214) :

Representation de la structure compositionnelle ou syntagmatique :
[N' [$_A$ *coronary*] [$_N$ *artery*]] [sem: [is-a: *artery*, quality: *coronary*]]

Representation de la structure idiomatique :
[N' [$_A$ *coronary*] [$_N$ *artery*]] [sem: [kind of: *artery*, property: *coronary*]]

Les chercheurs divisent les divers cas de variation en deux grands groupes : toutes les variations par coordination et par permutation d'un côté et tous les types de variation par insertion et par juxtaposition, de l'autre. Cette distinction se fait en fonction du type de règles qui est dérivé de chaque groupe. Ni les variantes par coordination ni les variantes par permutation n'entraînent, en général, la création ou l'apparition d'un nouveau terme. En conséquence, seules les règles définies pour ces deux types de variation sont considérées comme des « règles de variation ». Au contraire, divers cas de variation par insertion et de variation par juxtaposition impliquent souvent aussi la transformation du terme. Par conséquent, les règles définies pour les variations par insertion et par juxtaposition sont des « règles de production ».

Bien sûr, les chercheurs proposent aussi des règles pour les cas complexes de variation des termes de base. Dans les corpus, ils ont observé, par exemple, des cas de variation par coordination et par insertion (*cell function – cell* **membrane structure** and *function*) et des cas de variation par permutation combinée à des insertions (*membrane interaction – interaction* with **hemodialysis** *membranes*).

1.2.1.2 Un modèle de description syntagmatique et paradigmatique

Jacquemin (1999) présente un modèle de description des variations de termes complexes se basant sur des connaissances linguistiques. Sa recherche s'inscrit dans le champ de l'indexation automatique de terminologie en recherche d'information. Avec un modèle à deux axes – paradigmatique et syntagmatique – Jacquemin vise à

représenter les termes et leur variabilité dans les trois dimensions linguistiques : morphologie, syntaxe et sémantique.

L'auteur a effectué quatre expériences sur des corpus spécialisés anglais et français. Les expériences consistent à établir des liens paradigmatiques et syntagmatiques entre les termes et d'autres unités lexicales en utilisant divers outils et bases de données afin de générer des variantes. Pour cette étude, l'auteur se concentre sur les termes binaires en raison de leur fréquence élevée dans les corpus choisis.

Les termes de base et leurs variations sont représentés dans deux cadres parallèles. Le premier cadre, celui des termes, est composé d'une paire d'éléments : une structure syntaxique (axe syntagmatique) et l'ensemble des unités lexicales composant les termes reliées morphologiquement et sémantiquement à d'autres unités lexicales (axe paradigmatique) à partir de leurs lemmes. Voici un exemple de la représentation à deux axes du terme anglais *speed measurement* tiré de Jacquemin (1999 : 341) :

$$\left\{\begin{array}{l}\text{Syntagme : } \{N_0 \rightarrow N_2\ N_1\} \\ \text{Paradigme : } \left\{\begin{array}{l}(N_1 \text{ lemme}) = \textit{measurement} \\ (N_2 \text{ lemme}) = \textit{speed}\end{array}\right.\end{array}\right.$$

Le cadre correspondant aux variations a deux paires d'éléments de description, l'une correspond au terme de base et l'autre correspond à la variante. Par exemple, la variation qui relie le terme *speed measurement* à une forme verbale de sa tête et un synonyme de l'élément modificateur (*measuring maximal shortening velocity*) s'exprime de cette façon, selon Jacquemin (1999 : 342) :

Syntagme :

$$\left\{ \begin{array}{l} N_0 - N_2\ N_1 \ \Rightarrow \\ V_0 - V_1\ (\text{Prep}^?\ \text{Det}^?\ (A|N|\text{Part})^*)N'_2) \end{array} \right\}$$

$$\text{Paradigme} : \left\{ \begin{array}{l} (N_1\ \text{root}) = (V_1\ \text{root}) \\ (N_2\ \text{sem}) = (N'_2\ \text{sem}) \end{array} \right\}$$

Pour établir les liens morphologiques dans son modèle, Jacquemin a recours aux bases de données lexicales Multext[1] et Celex[2]. Pour les liens sémantiques, il a choisi le thésaurus intégré au programme de traitement de textes Microsoft Word97, qui établit des liens binaires entre unités lexicales en proposant une liste de synonymes, et le thésaurus Agrovoc[3] et la base WordNet[4] qui génèrent des classes sémantiques.

La première expérience a été réalisée sur un corpus français de l'agriculture et les outils utilisés pour l'établissement des liens sont le thésaurus de Word97 et la base Multext. La deuxième expérience a été effectuée sur un autre corpus français (de l'agriculture aussi) à l'aide du thésaurus Agrovoc et de Multext. La troisième expérience a été réalisée sur un corpus anglais de médecine en ayant recours à Word97 et à Celex. Enfin, la dernière expérience a porté sur un deuxième corpus anglais de médecine et elle a eu recours à WordNet et à Celex.

---

[1] En fait, cette base lexicale fait partie d'un ensemble de projets appelé Multilingual Text Tools and Corpora, Multext, visant à développer des corpus, des outils et d'autres ressources linguistiques pour plusieurs langues (http://www.lpl.univ-aix.fr/projects/multext/).

[2] CELEX comporte trois bases lexicales (anglaise, allemande et néerlandaise) créées par le Dutch Centre for Lexical Information (http://www.ldc.upenn.edu/readme_files/celex.readme.html).

[3] Agrovoc est un thésaurus spécialisé dans le domaine de l'agriculture créé par l'Organisation de Nations Unies pour l'Alimentation et l'Agriculture.

[4] Le système de référence lexicale WordNet pour la langue anglaise est développé par le Cognitive Science Laboratory de l'Université Princeton (http://www.cogsci.princeton.edu/~wn/index.shtml).

Les variations sont extraites des corpus annotés avec l'analyseur syntaxique FASTR.

Après avoir obtenu une liste de variations pour chaque expérience effectuée, l'auteur a évalué la précision de la génération de variantes avec les divers outils employés ainsi que le nombre de variantes terminologiques obtenues pour chaque expérience. Dans ce contexte, la précision fait référence au nombre de variantes générées et considérées comme étant correctes après une révision par des experts.

Selon l'évaluation des expériences faites sur les corpus français, le taux de précision de l'extraction de variantes syntaxiques, morphosyntaxiques et sémantiques est élevé (78 %). Cependant, lorsque certaines variations impliquent une transformation sémantique plus une modification syntaxique ou un lien morphologique (variations hybrides) la précision diminue considérablement (55 % - 29 %) à cause des décalages générés lors de ces combinaisons. Par exemple, la variante *former un réseau continu* a été générée par erreur à partir du terme *formation permanente*. Malgré tout, les variations hybrides n'ont pas une incidence importante sur le taux général de précision, car elles ne représentent qu'une fraction minimale des variations.

D'autres études, comme celle de Yoshikane *et al.* (1999) portant sur le japonais, montrent que ce modèle de description de la variation terminologique peut s'appliquer à d'autres langues que le français et l'anglais.

1.2.1.3 Analyse de la variation en catalan

La thèse de doctorat de Freixa (2002) porte aussi sur la variation terminologique en corpus spécialisé. Il s'agit d'une analyse contrastive profonde du phénomène de la variation dénominative (c'est-à-dire la variation au niveau formel) dans des textes catalans ayant des niveaux de spécialisation différents. Nous considérons que, étant donné l'envergure de l'étude de Freixa, il est impossible de la présenter au complet. Dans cette section nous nous limitons à en signaler les aspects les plus importants par rapport à notre étude.

Avec ce travail de recherche, Freixa veut vérifier une hypothèse générale voulant que le niveau de spécialisation d'un texte détermine la variation dénominative sur les plans quantitatif et qualitatif. Elle suppose que la variation dénominative est un phénomène plus fréquent dans les textes moins spécialisés que dans les textes plus spécialisés. Autrement dit, dans les textes moins spécialisés, les concepts sont représentés par plus d'une dénomination. Freixa suppose aussi que les types de variation dénominative diffèrent selon le niveau de spécialisation des textes.

Pour cette étude empirique, l'auteur a utilisé deux corpus composés de textes appartenant au domaine de l'environnement. Les textes sélectionnés portent sur les déchets et la contamination. Le premier corpus (nommé *ESPEC*) contient 13 textes ayant un niveau très élevé de spécialisation. Freixa explique que ce sont des textes dont l'émetteur (à savoir l'auteur ou les auteurs) et le récepteur (les lecteurs visés) sont des spécialistes du domaine. Le deuxième corpus (nommé *DIVUL*) est constitué de 7 textes de vulgarisation (brochures, livrets, etc.), beaucoup moins spécialisés, dont l'émetteur est un spécialiste et le récepteur est le public en général.

Freixa a procédé à l'extraction manuelle des unités terminologiques contenues dans chaque corpus. Les unités ont été placées dans deux bases terminologiques. Elle a sélectionné tous les termes simples et complexes appartenant au sous-domaine des déchets et de la contamination. La plupart des termes appartiennent à la catégorie des noms mais les bases terminologiques contiennent aussi certains verbes et adjectifs ainsi que des abréviations et des unités non linguistiques (par exemple, des formules chimiques). À l'aide de l'application *Multiterm* de *TRADOS,* les deux bases de données ont été construites; l'une pour les termes extraits du corpus *ESPEC*, l'autre pour ceux extraits du corpus *DIVUL*.

L'étape suivante a été celle de l'établissement de l'équivalence conceptuelle entre les termes sélectionnés. Pour ce faire, Freixa a comparé les définitions et les contextes définitoires des termes et a appliqué d'autres critères théoriques reliés au concept de la synonymie sur lequel cette étude s'appuie. Il s'agit de l'une des étapes les plus importantes, car elle conditionne l'analyse et l'interprétation des données et régit la structure des bases des donnés ainsi que l'organisation des unités terminologiques dans ces bases.

À la fin de cette étape, toutes les unités dénominatives associées à un même concept ont été regroupées dans une seule fiche terminologique. Ainsi, chaque fiche contient les diverses dénominations conceptuellement associées, accompagnées d'informations linguistiques (catégorie grammaticale, structure morphosyntaxique) et textuelles (contexte, page, document source, fréquence). La Figure 1 est un exemple de la fiche terminologique mise au point par Freixa.

Après la constitution des fiches terminologiques définitives, Freixa a calculé le nombre total de concepts et le nombre total de dénominations pour chaque corpus.

Ensuite, elle a estimé le taux de variation dénominative des concepts (nombre de dénominations / nombre de concepts). Le Tableau I reprend les chiffres de Freixa.

Tableau I. Taux de dénomination des concepts étudiés par Freixa (2002 : 222)

	DIVUL	ESPEC
Nombre de concepts	452	633
Nombre de dénominations	914	1004
Taux de dénomination	**2, 02**	**1, 58**

Fitxa núm. 21
anhídrid carbònic m [N+A]
Fonts i ocurrències: vola (3)
Context i font: *Fixeu-vos que els arbres han hagut de fabricar les seves fulles a partir d'elements simples -de l'aigua, l'anhídrid carbònic i les sals minerals- i utilitzen l'energia solar per a realitzar tot tipus de síntesis.* (vola, pàg. 30)
diòxid de carboni m [N+SP]
Fonts i ocurrències: tots (1)
Context i font: *La descomposició biològica d'un ser viu genera diòxid de carboni i deixa sals minerals diverses al medi.* (tots, pàg. 1)
carboni m [ML]
Fonts i ocurrències: tots (1)
Context i font: *La descomposició biològica d'un ser viu genera diòxid de carboni i deixa sals minerals diverses al medi. Estimacions recents avaluen que entre 2 i 3 mil milions de tones de carboni a l'any són lliurades a l'atmosfera procedents de l'activitat descomponedora a la superfície dels continents.*
(tots, pàg. 1)
CO2 m [FQ]
Fonts i ocurrències: tots (1)
Context i font: *És un dibuix.* (tots, pàg. 1)
gas carbònic m [N+A]
Fonts i ocurrències: vola (1)
Context i font: *Al voltant dels mil graus, la quasi totalitat d'aquests materials es crema i es transforma bàsicament en gas carbònic i vapor d'aigua.* (vola, pàg. 26)

Figure 1. Exemple de fiche terminologique élaborée par Freixa (2002)

La première ligne indique le numéro de fiche. Les diverses dénominations conceptuellement associées sont marquées en gras. La catégorie grammaticale et la représentation de la structure morphosyntaxique (entre accolades) sont placées devant chacune des dénominations. La ligne suivante indique les noms des documents où la dénomination apparaît (c'est-à-dire les sources) et son nombre d'occurrences. La fiche contient aussi un contexte (accompagné de la source et la page) pour chaque dénomination. Ils apparaissent en italiques.

Les résultats confirment l'hypothèse de l'auteur, hypothèse selon laquelle le taux de variation dénominative est plus élevé dans les textes moins spécialisés. Il est

aussi important de signaler que le corpus *ESPEC* présente une densité conceptuelle plus élevée, car le nombre de concepts différents détectés est plus élevé que celui du corpus *DIVUL*.

En ce qui concerne le nombre de dénominations observées par concept, les résultats montrent que, pour 60 % des concepts du corpus *ESPEC* et pour 43 % des concepts du corpus *DIVUL,* il n'y a qu'une seule unité dénominative. Pour 28 % et 32 % (*ESPEC* et *DIVULG*, respectivement) des concepts on trouve deux dénominations. Dans le corpus *ESPEC*, il y a peu de cas de concepts associés à plus de trois dénominations, ce qui représente 10 % du total. À l'inverse, dans le cas du corpus *DIVUL*, le pourcentage de concepts associés à plus des trois dénominations est plus élevé : 32 %. Le reste du pourcentage est constitué par les concepts correspondant à quatre, cinq ou plus de dénominations.

Quant à l'aspect morphosyntaxique des termes, les analyses montrent que les unités simples sont les plus fréquentes dans les deux corpus. En deuxième place, par ordre de fréquence, se trouvent les unités complexes constituées d'un nom et d'un syntagme prépositionnel [N+SP] (ex. *compactadora de pota de cabra)* ou d'un nom et d'un adjectif [N+A] (ex. a*cer inoxidable*). Les résultats de Freixa montrent également que les unités complexes ont un taux plus élevé de variation que les unités simples. D'une façon générale, tous ces résultats valident l'hypothèse de l'auteure.

La présentation d'une classification formelle des variations selon le type de changement linguistique observé constitue l'étape suivante de la recherche. Freixa propose cinq catégories de variation. Il faut remarquer que cette typologie diffère de celle que nous proposons dans la section 1.1.2.

1. Variation par changements graphiques – Cette catégorie regroupe les changements orthographiques (ex. *esprai/spray*), les abréviations (ex. *acer inoxidable/acer inox.*), les sigles (ex. *clorofluorocarboni/CFC*) et les représentations par une forme « artificielle » comme les symboles et les formules chimiques (*amoníac/NH_3*)

2. Variation par changements morphosyntaxiques – Cette catégorie comporte deux sous-catégories :

a) Changements qui ne modifient pas la structure du terme : absence ou présence du déterminant (ex. *gestió de residus/gestion dels residus*), changement de préposition (ex. *condicions del condensador/conditions en el condensador*), changement du genre (ex. *màxima absoluta/màxim absolut*) et changement du nombre (ex. *contaminació del aigua/contaminació de les aigües*).

b) changements entraînant une modification à la structure : substitution d'une unité simple par une unité complexe ou vice versa (ex. *producte ecològic/ecoproducte*) et transformations [N+A] / [N+SP] (ex. *residus miners/residus de la mineria*).

3. Variation par réduction – qui regroupent les réductions portant sur la tête (ex. *planta depuradora/depuradora*) ou sur des éléments modificateurs de la structure (ex. *temps de residència del gas/temps de residència*).

4. Variation par changements lexicaux – Dans le cas des unités simples, il s'agit de la substitution par un synonyme proprement dit (ex. *contaminació/pol-lució*). Dans les cas des unités complexes, il s'agit de la

substitution d'une unité lexicale, soit la tête (ex. *bé de consum/producte de consum*), soit le modificateur (ex. *dipòsit d'assentament/dipòsit de decantació*) de la structure.

5. Variation par changements complexes – Il s'agit d'une combinaison de différents types de changements. Par exemple, dans ces deux dénominations : *pesticida de síntesi* et *plaguicide químic*, un changement lexical (*pesticida – plaguicida*) ainsi qu'un changement morphosyntaxique entraînant une modification de la structure (*de síntesi – químic*) se sont produit.

En comparant les nombres et les types de variation observés dans chacun des deux corpus, Freixa obtient les résultats présentés dans le Tableau II.

Tableau II. Résultats de l'analyse des variantes de Freixa (2002 : 287)

Types de variation		DIVUL	ESPEC
I. Changements graphiques	1. Terme et forme artificielle	0.21 %	6.70 %
	2. Terme et abréviation	1.26 %	4.89 %
	3. Changements orthographiques	2.73 %	1.80 %
	Total	**4.20 %**	**13.40 %**
II. Changements morphosyntaxiques	1. Même structure	7.35 %	11.08 %
	2. Changements dans la structure	2.31 %	3.09 %
	Total	**9.66 %**	**14.17 %**
III. Réductions	1. Réduction du modificateur	19.32 %	24.74 %
	2. Réduction de la tête	3.99 %	6.44 %
	3. Autres réductions	2.52 %	5.67 %
	Total	**25.84 %**	**36.85 %**
IV. Changements lexicaux	1. Termes simples	14.07 %	4.89 %
	2. Termes complexes	28.99 %	20.10 %
	Total	**43.06 %**	**25 %**
V. Changements complexes	1. Similitude formelle	11.34 %	9.27 %
	2. Aucune similitude formelle	5.88 %	1.28 %
	Total	**17.22 %**	**10.56 %**

À partir de ces données, on peut attirer l'attention sur plusieurs faits reliés aux variations prédominantes en général et par corpus :

- dans les deux corpus, les variations graphiques et les réductions sont les variations les plus fréquentes;
- dans les deux corpus, la réduction des modificateurs des structures est plus fréquente que la réduction de tête;
- quant à la variation par changements lexicaux, les changements dans les termes complexes sont plus fréquents que dans les termes simples;
- dans la plupart des cas de variation morphosyntaxique, les changements ne modifient pas la structure ; et
- les changements graphiques sont plus fréquents dans le corpus *ESPEC* que dans le corpus *DIVUL*, surtout, comme l'observe Freixa, en raison de l'utilisation plus fréquente de symboles, de formules et d'abréviations dans les textes plus spécialisés.

La partie de l'étude de Freixa que nous venons de présenter apporte, sans doute, des informations importantes sur la variation terminologique qui s'avèrent très utiles pour le développement des outils de TAL, même si ceci n'est pas l'objectif principal de l'auteure.

1.2.1.4 Étude de la variation terminologique pour extraire des sujets de recherche

La recherche de Ibekwe-Sanjuan (1999) porte sur l'établissement de relations entre les termes et trois types de variation syntaxique afin de détecter des sujets de recherche dans un corpus spécialisé.

Ibekwe-Sanjuan affirme que les variations reflètent une évolution terminologique et, par conséquent, elles peuvent aussi impliquer une évolution du concept associé à ces variations.

Ce travail consiste à : 1) analyser un corpus anglais spécialisé, 2) en extraire un ensemble de termes, 3) identifier trois types de variantes syntaxiques et, enfin, 4) développer un cadre pour l'organisation des sujets à partir des relations conceptuelles établies par les variations. L'objectif principal n'est pas l'extraction des variantes comme telles, mais plutôt l'établissement des liens entre elles.

Le corpus spécialisé est constitué par des titres et des résumés scientifiques appartenant au domaine de la biotechnologie des plantes. L'étude se concentre sur les termes représentés par des syntagmes nominaux et sur les variations syntaxiques par permutation, par expansion et par substitution.

L'extraction des termes a été réalisée de façon manuelle et elle s'appuie sur trois critères fondamentaux : caractéristiques morphologiques, structure syntaxique et longueur des termes. Ainsi, les termes à extraire doivent être des noms ou des adjectifs; certaines prépositions sont aussi incluses. Les termes peuvent apparaître dans deux structures différentes : comme un composé (ex. *the specific alfalfa nodulation*) ou comme une structure syntagmatique (*the specific nodulation of alfalfa*). Le terme à extraire doit être constitué d'un maximum de sept mots et d'un minimum de 2 mots, sans prendre en considération ni les prépositions ni les déterminants. À cette étape, un total de 4463 termes candidats ont été extraits. Ensuite, un filtrage combinant des critères lexicaux et statistiques a été appliqué pour éliminer des termes candidats non pertinents. Du total des termes, seulement 70 % ont été retenus.

L'étape qui suit l'extraction de termes est celle de l'identification de variantes syntaxiques.

Selon l'analyse faite par Ibekwe-Sanjuan, peu de termes ont subi une transformation par permutation (c'est-à-dire, une transformation d'une structure syntagmatique en une structure composée). Voici un exemple donné par l'auteure (Ibekwe-Sanjuan 1998a : 661) :

> Structure syntagmatique : *accession of azolla-anabaena*
> Structure composée : *azolla-anabaena accession*

Une grande proportion des termes a subi une substitution (remplacement d'un mot du terme de base par un autre mot dans la variante). La variation par substitution peut se produire au niveau de la tête du terme (*infection thread development – infection thread formation*) ou au niveau d'un élément modificateur (*alfalfa root hair – curled root hair*).

Quant à la variation par expansion, l'auteure a observé que l'ajout de mots peut se produire à gauche, à droite ou au milieu (insertion) d'un terme. Voici quelques exemples présentés par Ibekwe-Sanjuan (1998a : 661):

> Expansion à gauche pour le terme *self-licking* : *refractory self-licking*
> Expansion à droite pour le terme *blue light* : *blue light induction*
> Insertion dans le terme *conserved domain* : *conserved protein domain*

Certains cas combinés d'expansion ont aussi été identifiés : 1014 termes ont subi une variation par expansion.

Au total, dans 82% des termes, les trois types de variation syntaxique ont été observés, ce qui met en évidence l'importance de ce phénomène en terminologie.

Afin d'établir une méthode d'extraction de sujets de recherche par le biais de relations termes-variantes, Ibekwe-Sanjuan a déterminé les propriétés conceptuelles des variantes syntaxiques observées. Trois relations conceptuelles ont été identifiées : relation « classe de », relation d'équivalence et relation générique/spécifique (ou relation hyperonyme-hyponyme).

La variation par substitution génère entre diverses variantes terminologiques une relation regroupant des « classes ». Par exemple, les variantes : *template dna*, *genomic dna* et *target dna*, où la substitution se produit au niveau du modificateur, peuvent être considérées comme des « classes » de *dna*.

La variation par permutation peut indiquer une équivalence conceptuelle entre terme et variante, tel est le cas de *dna fragment* et de *fragment of dna*.

Tous les sous-types de variation par expansion peuvent impliquer une relation générique/spécifique entre les termes. Autrement dit, l'expansion peut introduire une hiérarchisation entre les termes et les variantes, ce qui peut correspondre à des familles de concepts. Pour ce type de relation conceptuelle, Ibekwe-Sanjuan ne donne pas d'exemples. En fait, il semble que ce soit l'aspect le moins développé de cette recherche.

Dans la dernière partie de son travail, Ibekwe-Sanjuan développe une méthode pour la structuration des sujets de recherche en utilisant les informations obtenues à l'étape d'analyse des variantes terminologiques et des calculs mathématiques. La méthode consiste à organiser les termes extraits du corpus dans un réseau en s'appuyant sur les relations sémantiques entre termes et leurs variantes dérivées.

Elle divise les variantes en deux groupes : d'une part, les variantes produites par un changement de la tête (le groupe a été nommé *CLAS*) et, d'autre part, les variantes produites par un changement du modificateur (le groupe a été nommé *COMP*). Dans le groupe *CLAS,* se trouvent les variantes représentant des classes ou des types de sujets de recherche. Le groupe *COMP* relie et structure les variantes ayant la même tête (ensemble de variantes appartenant au même sujet de recherche), représentant ainsi les paradigmes dans le corpus. Ces paradigmes correspondent alors à des groupes de sujets isolés qui seront ensuite reliés par des liens transversaux du groupe *CLAS*. De cette façon, les associations entre les divers sujets du corpus sont mises en évidence. La Figure 2 présente le réseau des sujets de recherche créé à partir de la méthode proposée par Ibekwe-Sanjuan.

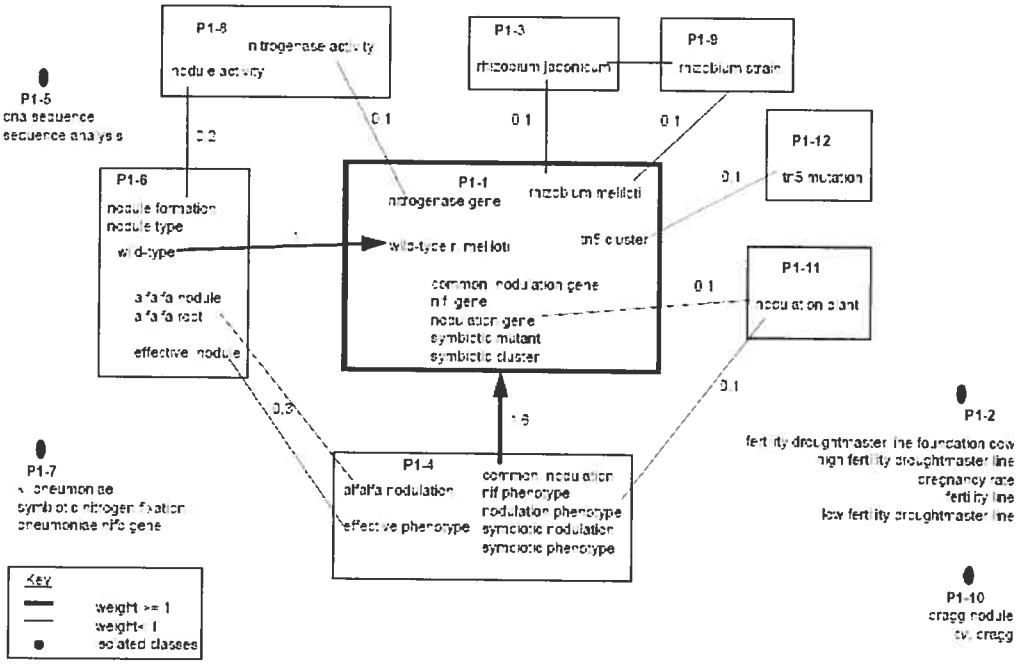

Figure 2. Réseau des sujets de recherche obtenu par l'établissement des relations entre termes et variantes (Ibekwe-Sanjuan 1998b : 174)

1.2.2 Étude de la variation dans un corpus bilingue

Carl *et al.* (2004) ont récemment publié une étude portant sur la variation terminologique observée dans des textes parallèles. C'est l'un des rares travaux réalisés sur des textes traduits.

Les auteurs de cette étude partent de la prémisse selon laquelle l'identification des termes et de leurs variantes joue un rôle important dans diverses applications du TAL : par exemple, la traduction automatique, ainsi que dans les processus de validation de l'emploi de termes et de mise à jour de ressources terminologiques. Leur objectif est donc de développer un système semi-automatique d'identification de termes et de leurs variantes dans des textes parallèles. Le système s'appuie sur une base de données contenant les termes de référence et leurs traductions ainsi que des modèles de variation terminologique.

La conception et l'application de cette base de données sont fondées sur le raisonnement par abduction. En termes généraux, l'abduction part de l'observation de phénomènes pour ensuite formuler des hypothèses qui peuvent expliquer ces phénomènes, mais ces hypothèses ne sont pas logiquement impliquées par les prémisses. Dans le contexte qui nous intéresse, le raisonnement par abduction s'applique comme suit : la supposition de base est que, dans le corpus parallèle visé, chaque terme source est traduit par un autre terme en langue cible. Cependant, si la traduction de ce terme n'est pas retrouvée (cette partie constituerait le phénomène observé), la base de données essayera de démontrer la présence d'une variante de la traduction du terme (cette partie correspondrait à la formulation d'un hypothèse). Dans ce qui suit nous expliquerons le processus d'abduction de variantes proposé par Carl *et al.*

Cette étude a été réalisée sur des textes anglais spécialisés alignés avec leur traduction correspondante en français et se concentre sur les termes complexes. Elle comporte trois étapes.

La première étape consiste à analyser des textes alignés pour en extraire un lexique bilingue contenant des termes de base et normalisés, des synonymes et un ensemble de patrons généraux de variation terminologique. Le lexique bilingue a été extrait manuellement. Les patrons de variation ont été développés en fonction des variantes observées : omission, insertion, permutation, coordination, synonymie, dérivation et variantes typographiques.

Dans la deuxième étape, les chercheurs ont fait une série d'expériences avec trois différentes versions d'une « base de données terminologiques abductive » (*Abductive Terminology Data Base*, *ATDB*) générée avec les ressources obtenues dans la première étape. La première version de la base de données, $ATDB_0$, contient le lexique bilingue; elle ne contient que les termes de base normalisés et leurs équivalents. La deuxième version, $ATDB_1$, a été enrichie; au lexique bilingue, s'ajoute un patron pour la génération de synonymes. La troisième base, $ATDB_2$, est la plus riche, car elle est composée des ressources se trouvant dans les deux premières bases plus les patrons conçus pour l'identification de variantes par insertion, par omission, par permutation et par coordination.

Les trois versions de la *ATDB* ont été testées sur deux paires de textes alignés. D'abord, les termes de base et leurs traductions ont été annotés manuellement sans spécifier si les termes traduits sont des termes de base ou des variantes. Ensuite, en appliquant la base de données aux textes, les termes et leurs variantes sont marqués de façon automatique.

L'évaluation de la performance des bases de données *ATDB* ainsi que la comparaison de leurs résultats avec ceux obtenus après l'annotation manuelle constituent la troisième étape de l'étude.

Après l'utilisation de chaque version de la base de données, les chercheurs ont calculé le nombre de termes traduits et annotés correctement, le bruit produit, le silence généré. Ensuite, ils ont estimé le taux de précision (traduction correcte / traduction correcte + bruit) et le taux de rappel (traduction correcte / traduction correcte + silence). Le Tableau III montre les résultats de ces évaluations.

Tableau III. Résultats des évaluations des bases des données abductives de Carl *et al.* (2004 : 116)

	Corpus SNIPER2			Corpus SNIPER3		
	$ATDB_0$	**$ATDB_1$**	**$ATDB_2$**	**$ATDB_0$**	**$ATDB_1$**	**$ATDB_2$**
précision	0,54	0,61	0,62	0,57	0,60	0,62
rappel	0,45	0,78	0,89	0,39	0,78	0,86
traductions correctes	467	802	921	338	683	754
bruit	402	508	559	250	446	468
silence	566	231	112	534	189	118

En comparant les résultats, on constate que la précision et le rappel augmentent au fur et à mesure que la base de données est enrichie mais, malheureusement, il en va de même pour le bruit. En général, dans les trois versions de la *ATDB*, le bruit est produit par des correspondances incorrectes établies entre les paires de termes des deux langues. Les auteurs considèrent que ce problème pourrait être minimisé en implémentant dans le système une analyse syntaxique plus profonde qui établit des connexions adéquates entre les termes. En plus, le silence a considérablement diminué avec la version de la *ATDB* la plus robuste – $ATDB_2$ –, ce qui démontre l'utilité de l'application des patrons de variation.

La troisième partie de l'étude présente aussi l'évaluation des résultats obtenus pour les trois *ATDB* par patron de variation. Les résultats montrent que la variation par omission est le type de variation le plus productif pour les termes anglais et français. Le nombre de cas de variation par permutation est considérable mais inférieur à celui des variations par omission. Très peu de cas d'insertion et de coordination ont été trouvés dans les deux paires de textes alignés. Dans les trois derniers cas de variation, le taux de bruit est plus élevé et la précision diminue à cause des connexions incorrectes entre les termes anglais et leurs équivalents (terme de base ou variante) en français. À ce propos, les chercheurs concluent qu'il faudra raffiner ces patrons de variation, inclure une analyse syntaxique plus profonde et tester le système sur d'autres corpus plus volumineux afin de mieux évaluer et caractériser ces types de variation.

De plus, ils suggèrent que, puisque les ambiguïtés d'alignement et de connexion entre termes et variantes risquent d'augmenter au fur et à mesure que le nombre de termes et les patrons de variations augmentent dans la base de données, il conviendrait de développer un système de structuration des patrons en fonction de leur fréquence dans les textes.

1.2.3 Synthèse des études portant sur la variation terminologique

Dans la section précédente nous avons décrit brièvement cinq études sur la variation terminologique. Toutes ces études s'inscrivent dans le contexte du TAL, et se basent sur l'observation de corpus textuels.

Parmi les cinq études, trois portent sur la variation en anglais et en français, une analyse porte sur la langue anglaise uniquement et la cinquième porte sur le catalan. Nous avons également des références d'autres travaux similaires en langue

allemande, espagnole et japonaise mais, jusqu'à présent, il semble que la langue anglaise a été la plus étudiée.

Une caractéristique commune à toutes ces recherches est l'approche générale adoptée. Même si les modèles de description des variations diffèrent, elles ont toutes pour but la caractérisation de la variation et l'identification des variantes au moyen de règles ou de patrons de transformation à partir de l'observation de termes dans un ensemble de textes spécialisés. Trois des travaux (Carl *et al.* 2004; Jacquemin 1999 et Ibekwe-Sanjuan 1998a) visent une application immédiate et bien précise – l'indexation automatique de termes et de leurs variantes. Par contre, le travail de Daille *et al.* (1996) et celui de Freixa (2002) se différencient un peu du reste parce qu'ils décrivent de façon exhaustive le phénomène de la variation et établissent une base théorique pour son étude formelle.

Un autre aspect intéressant de ces recherches est le fait qu'elles se concentrent sur la variation de type syntaxique et laissent de côté d'autres formes de variation, à l'exception de celle de Freixa (2002). En fait, certains types de variation (graphique, orthographique, flexionnelle) ne sont plus étudiés parce que ce sont des phénomènes facilement modélisés par les systèmes d'indexation/extraction terminologique existants; ces variations ne représentent plus un problème. Cependant, il semble que la variation morphosyntaxique et la variation au niveau sémantique restent encore peu étudiées.

En ce qui concerne le type d'unités terminologiques analysées, quatre des travaux présentés n'étudient que les unités complexes. Parfois, ce choix s'explique par le fait que les unités complexes sont les plus fréquentes dans les textes spécialisés, qu'elles varient plus que les unités simples, et qu'elles posent plus de problèmes lors de leur identification. Par contre, Freixa (2002) aborde la variation

dans les termes simples et elle montre que ceux-ci subissent aussi des transformations. Notre étude cherche aussi à apporter une réponse à cette question et c'est la raison pour laquelle nous analysons des termes complexes et des termes simples. De même, bien que la plupart des unités terminologiques étudiées soient de type nominal, les travaux exposés montrent qu'il y a une tendance à inclure des unités appartenant à une autre catégorie grammaticale (par exemple, verbes, adjectifs ou, encore, syntagmes verbaux ou adjectivaux).

D'ailleurs, et comme nous l'avons déjà mentionné, il existe à présent très peu de travaux sur la variation dans la perspective de l'extraction bilingue, des études en textes parallèles constitués assez souvent par des traductions. Le seul travail bilingue que nous avons présenté est celui de Carl *et al.* (2004).

Tout comme les études que nous avons présentées, notre étude a pour objet la caractérisation de la variation terminologique basée sur l'analyse d'un corpus textuel spécialisé. Elle s'inscrit aussi dans le contexte du TAL, plus précisément, elle cherche à proposer des pistes pour l'extraction automatique de termes bilingue. Par ailleurs, notre étude s'inspire des approches descriptives de Daille *et al.* (1996) et de Freixa (2002).

De la même façon que l'analyse de Freixa (2002), notre travail prend en considération les termes complexes ainsi que les termes simples et se base sur l'analyse d'un corpus constitué de textes du domaine de l'environnement. Une différence importante entre le travail de Freixa et le nôtre est le point de départ pour l'analyse. Freixa part d'un concept pour ensuite rechercher les diverses variantes dénominatives de ce concept ; nous, au contraire, prenons un terme de base espagnol pour ensuite rechercher les variantes produites à partir de ce terme.

Par ailleurs, à la différence des travaux présentés, nous allons étudier non seulement la variation syntaxique, mais aussi la variation morphosyntaxique et sémantique, plus particulièrement, la synonymie. De même, étant donné que les travaux portant sur la variation terminologique sont réalisés majoritairement sur les langues anglaise et française, nous observons ce phénomène linguistique dans une autre langue, à savoir l'espagnol.

Enfin, nous avons aussi pris comme référence le travail de Carl *et al.* (2004), car notre objectif est de réaliser une analyse comparative dans un corpus constitué de traductions. Comme *Carl et al.*, nous voulons mettre l'accent sur les cas de variation terminologique non relevés par les études unilingues, par exemple l'omission. Nous prenons comme langue de référence l'anglais et analysons les cas de variation en espagnol.

1.3 Extraction automatique de termes

Dans la dernière section du présent chapitre nous décrivons brièvement les principes relatifs à la conception et au fonctionnement des extracteurs automatiques de termes. Étant donné que notre étude a pour objectif de réfléchir à l'incidence de la variation terminologique en extraction terminologique bilingue, nous présenterons aussi les caractéristiques principales de l'extraction bilingue. Enfin, nous discuterons de l'efficacité et des limites des extracteurs de termes bilingues existants.

1.3.1 Principes généraux de l'extraction automatique de termes

La nécessité de réaliser de grands projets terminologiques (par exemple, élaboration de glossaires, dictionnaires spécialisés ou vocabulaires, alimentation et

mise à jour de banques terminologiques) dans des délais très courts a contribué au développement d'outils informatiques qui facilitent au moins une partie de ce travail. Une façon d'atteindre cet objectif est l'automatisation de certaines tâches, telle que l'identification de termes sur laquelle s'appuient les descriptions terminologiques. C'est dans ce contexte que les extracteurs automatiques de termes sont apparus.

La fonction générale des extracteurs automatiques consiste à « trouver dans un texte ou un ensemble de textes les mots graphiques ou les suites de mots graphiques susceptibles d'être des termes » (L'Homme 2004 : 167), à savoir, des *candidats-termes*. Ce travail constitue un défi immense puisqu'il fait appel à des connaissances linguistiques et extralinguistiques. Parmi les difficultés les plus importantes auxquelles les extracteurs automatiques font face se trouvent : 1) déterminer la longueur correcte des termes complexes, autrement dit, identifier où le syntagme terminologique commence et où il se termine, 2) distinguer, dans le corpus, les syntagmes terminologiques des séquences libres, 3) déterminer le statut terminologique des unités lexicales ; et 4) déterminer la pertinence des candidats-termes proposés pour les objectifs poursuivis par l'extraction (ex, pour le dictionnaire ou glossaire qu'on veut construire) (Cabré *et al.* 2001).

Par ailleurs, il est important de souligner que la plupart des extracteurs automatiques se concentrent sur les termes complexes de nature nominale, et cela détermine en grande mesure les approches à adopter et les techniques à utiliser dans le processus d'identification.

D'une manière générale, les extracteurs automatiques s'appuient sur des connaissances linguistiques et statistiques pour l'identification de termes.

a) L'approche statistique

L'approche statistique utilise des informations numériques qui peuvent être dégagées des corpus et est donc indépendante des langues. L'indice de base le plus utilisé est celui de la fréquence. Les concepteurs des extracteurs automatiques considèrent que l'un des indices primordiaux pouvant conférer à une forme un statut terminologique est sa récurrence dans un corpus spécialisé. Alors, en se basant sur un seuil de fréquence minimale, les extracteurs automatiques extraient les mots graphiques ou les chaînes de mots graphiques correspondant au paramètre établi; les mots dont la fréquence est inférieure au seuil ne seront pas retenus comme candidats-termes.

Lorsqu'il s'agit d'extraire des termes complexes, les extracteurs automatiques peuvent avoir recours à une technique reliée à la fréquence de suites de mots graphiques appelée *calcul de segments répétés*. La technique des segments répétés recherche « les séquences de mots (en fait de « formes graphiques ») non séparés par un caractère délimiteur de séquence (des signes de ponctuation), qui apparaissent plus d'une fois dans un corpus de textes » (Lebart et Salem 1988 : 24). Une fois repérés, ces segments répétés peuvent, par exemple, être présentés dans un « tableau des segments répétés » en les classant par ordre de longueur décroissante. La Figure 3 présente un exemple de segments répétés tiré de l'étude de Ladouceur et Drouin (1997).

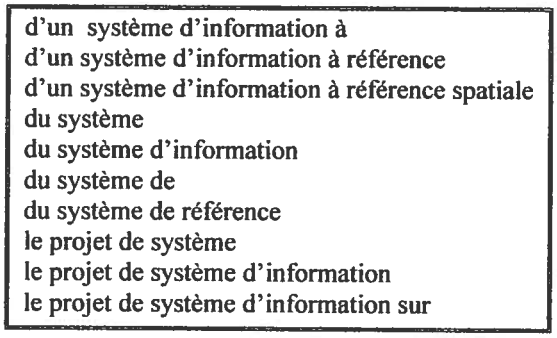

Figure 3. Liste partielle des résultats obtenus par l'application de la technique des segments répétés (Ladouceur et Drouin 1997 : 211)

Une autre façon de rechercher des suites de mots graphiques qui peuvent être des termes s'appuie sur le calcul du « degré d'association ». En se basant sur le principe selon lequel les unités constituant un terme complexe se caractérisent par une affinité non accidentelle, un extracteur automatique peut être paramétré pour extraire des couples de mots apparaissant souvent ensemble dans un corpus, car il y a une forte probabilité qu'il s'agisse d'un terme. Church et Hanks proposent une mesure du degré d'association des mots basée sur la théorie de « l'information mutuelle ». Cette mesure compare la probabilité de retrouver deux mots graphiques ensemble – (P *x, y*) – à la probabilité de les retrouver de façon indépendante dans le corpus – P (*x*) P (*y*). Si une association forte existe entre *x* et *y*, alors la probabilité de les trouver ensemble sera plus élevée que la probabilité de les retrouver de façon indépendante (Church et Hanks 1990). La Figure 4 présente la formule du calcul de l'information mutuelle proposée par Church et Hanks. Le système utilisant cette technique extraira ainsi les couples de mots ayant une association forte.

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)\,P(y)}$$

Figure 4. Calcul de l'information mutuelle proposée par Church et Hanks (1990 : 23)

b) L'approche linguistique

Les méthodes linguistiques d'identification de candidats-termes se fondent plutôt sur les caractéristiques morphosyntaxiques des termes. À la différence de l'approche statistique, l'approche linguistique exige un prétraitement des textes du corpus sur lequel l'extracteur va travailler. Chacun des mots graphiques du corpus doit porter une « étiquette » indiquant la catégorie grammaticale à laquelle il appartient. L'étiquetage des mots graphiques peut aussi comprendre d'autres indications grammaticales telles que le lemme. Le processus d'extraction dépendra alors de l'analyse portée sur le corpus étiqueté. Nous décrivons, ci-dessous, deux des techniques linguistiques d'extraction les plus connues :

- Identification de termes au moyen de « patrons syntaxiques typiques » : Comme les concepteurs d'extracteurs automatiques considèrent que la plupart des termes sont des syntagmes nominaux (très souvent constitués de deux unités lexicales), ils cherchent à caractériser la construction syntaxique de ces structures, autrement dit, à définir les patrons de formation de ces séquences régulières dans les textes spécialisés. Par exemple, en anglais, un patron syntaxique typique d'un terme complexe binaire est la combinaison d'un adjectif et d'un nom (*environmental fate*). Ces patrons de formation sont alors transformés en règles sur lesquelles l'extracteur automatique s'alignera pour générer une liste de candidats-termes. La Figure 5 montre des patrons syntaxiques typiques pour extraire des termes français.

N Adj : station brouilleuse

N1 N2 : diode tunnel, mémoire tampon

N1 à (Det) N2 : antenne à réflecteur, assignation à la demande

N1 de (Det) N2 : modulation de fréquence, synchronisation des paquets

N1 Prep N2 : multiplexage en fréquence

Figure 5. Exemple des patrons syntaxiques typiques et de termes correspondant à ces patrons (Daille 1995 : 105)

Étant donné que les patrons typiques de formation de termes varient d'une langue à l'autre, les extracteurs utilisant cette technique ne peuvent pas être appliqués à plusieurs langues sans avoir défini au préalable les patrons pour chaque langue désirée. Cela représente parfois une limite importante de ces extracteurs.

- Isolation des séquences terminologiques par repérage de frontières : Au lieu d'exploiter les patrons de formation de termes complexes, cette technique s'appuie sur des connaissances linguistiques « en négatif » concernant les catégories grammaticales qui ne donnent pas lieu à des termes complexes : conjonctions, verbes, préposition + adjectif possessif, etc. Comme l'explique son concepteur, la technique consiste à « découper le texte en repérant ces frontières potentielles entre lesquelles on isole des syntagmes nominaux susceptibles d'être des occurrences de termes » (Bourigault 1993). La Figure 6 présente un exemple de découpage d'une phrase et les groupes nominaux générés à partir du découpage.

Texte initial
le circuit d'aspersion de l'enceinte de confinement assure le maintien de sa température nominale de fonctionnement après une augmentation de pression.
Règles de découpage
verbe – coupe préposition + adj. possessif – coupe préposition + art. indéfini – coupe
Groupes nominaux maximaux
circuit d'aspersion de l'enceinte de confinement maintien température nominale de fonctionnement augmentation de pression

Figure 6. Exemple de découpage d'un texte en utilisant la technique de repérage de frontières (Bourigault 1993 : 108)

Cette technique, ne nécessitant pas une analyse lexico-syntaxique profonde, est moins lourde que celle des patrons de formation : le processus de repérage de frontières est moins long et moins complexe que celui des patrons typiques.

Nous venons de présenter quelques techniques linguistiques et statistiques utilisées par les extracteurs automatiques de termes, mais la question qui se pose est la suivante : quelles techniques sont plus efficaces ? Autrement dit, lesquelles produisent les meilleurs résultats ?

Normalement, la liste de candidats-termes produite à la suite d'une extraction automatique est évaluée en fonction du *bruit* généré et du *silence* obtenu. Le bruit fait référence aux candidats-termes qui, en réalité, ne constituent pas de vrais termes. Une mesure très courante du bruit est la *précision,* qui « estime la proportion de bons candidats extraits dans la liste de candidats-termes. Lorsque la précision est élevée, il y a peu de bruit » (L'Homme 2004 : 193). Le silence, au contraire, fait

référence aux termes qui n'ont pas été extraits du corpus. Le rappel est une façon de mesurer le silence : « il évalue la proportion de bons termes extraits parmi les possibilités dans le texte. Le rappel est élevé lorsqu'il y a peu de silence » (L'Homme 2004 : 193).

D'après les évaluations réalisées par les concepteurs des extracteurs de termes, aucune technique, appliquée toute seule, produit des listes de candidats-termes sans qu'il y ait des taux importants de bruit ou de silence. Par exemple, les techniques basées sur le critère de fréquence établissent très souvent un seuil minimal de 3; les termes ayant une fréquence inférieure ne seront pas extraits. En conséquence, le rappel risque d'être élevé. L'application du calcul de l'information mutuelle a aussi une incidence importante sur le taux de précision puisque la liste produite par l'extracteur comprendra des couples de mots graphiques présentant une association forte mais qui ne constituent pas des termes.

Les techniques basées sur des connaissances linguistiques posent aussi des problèmes et produisent du bruit et du silence. Par exemple, les patrons de formation de termes visent l'extraction de syntagmes (termes complexes), excluant ainsi la possibilité d'identifier des termes simples. D'ailleurs, il peut avoir des cas où des termes complexes dont les structures syntaxiques ne correspondent pas aux règles de formation appliquées. Certains termes complexes seront alors omis des listes de candidats-termes. De plus, cette même technique peut repérer des syntagmes s'alignant sur les patrons de formation mais qui n'ont pas un caractère terminologique.

Compte tenu de tous ces problèmes et de la nécessité d'améliorer la qualité des résultats d'extraction, les chercheurs ont développé des extracteurs de termes qui combinent différentes techniques linguistiques et statistiques. Il y a des extracteurs

de termes qui, par exemple, appliquent la technique linguistique de patrons typiques de formation et ensuite filtrent la liste de candidats-termes en s'appuyant sur des mesures statistiques. Dans leurs évaluations, les concepteurs d'extracteurs de termes commentent que le fait d'adopter des approches hybrides permet, en effet, d'améliorer les taux de précision et de rappel. Toutefois, il reste encore plusieurs autres difficultés à résoudre.

Jusqu'ici, nous avons décrit le fonctionnement des extracteurs automatiques de termes conçus pour traiter une seule langue à la fois. Cependant, il existe aussi quelques extracteurs conçus pour dépouiller des corpus constitués de textes originaux et de leur traduction – que nous appelons *corpus parallèles* – ce qui implique une extraction simultanée de termes dans deux langues différentes. Dans la section suivante nous présentons ces systèmes d'extraction bilingue.

1.3.2 Les extracteurs automatiques de termes bilingues

Dans certains systèmes de traduction automatique ainsi que dans les systèmes de traduction assistée par ordinateur, l'extracteur de termes constitue souvent l'une des composantes fondamentales. De plus, la création et la gestion de banques terminologiques multilingues pourraient tirer des possibilités offertes par un extracteur automatique.

Dans ces contextes, un extracteur automatique de termes doit non seulement identifier et proposer une liste de candidats-termes pour chacune des langues visées, mais aussi mettre en correspondance les termes et leurs équivalents. Normalement, ces extracteurs travaillent par paires de langues et, la plupart du temps, ils s'appuient sur des corpus parallèles. Un corpus parallèle est un corpus constitué de deux sous-ensembles de textes de traduction, le premier contenant les textes source et le

deuxième les textes cible[5]. Les deux sous-ensembles sont mis en correspondance au moyen du processus d'alignement.

Toutefois, il existe aussi quelques études s'appuyant sur des corpus comparables. Un corpus comparable est un ensemble de textes en deux ou en plusieurs langues ayant des caractéristiques communes, surtout aux niveaux thématique, de la structuration du texte, de la date de parution et du niveau de spécialisation, s'il s'agit des textes spécialisés (Bowker et Pearson 2002). Dans le contexte de l'extraction automatique de termes, Déjean et Gaussier affirment que « deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1 » (2002 : 2).

L'alignement est le processus mettant en correspondance les textes originaux et leurs traductions afin de faciliter le traitement linguistique d'un corpus parallèle. Isabelle et Simard donnent une définition graphique de l'alignement :

> « Considérons un texte *S* et sa traduction *T* comme deux ensembles de segments successifs : $S = \{s1, s2, \ldots, sn\}$ et $T = \{t1, t2, \ldots, tm\}$. Un alignement est tout simplement un sous-ensemble du produit cartésien *S* X *T*. Par exemple, si $S = \{s1, s2, s3, s4, s5\}$ et $T = \{t1, t2, t3, t4, t5\}$, alors l'alignement A = {*s1-t1, s2-t2, s2-t3, s3-t4, s4-t5, s5-t5*} associe le segment *s1* au segment *t1* ; le segment *s2* aux segments *t2* et *t3* ; le segment *s3* au segment *t4* ; et les segments *s4* et *s5* au même segment *t5*. »
> (Isabelle et Simard 1996 : 1)

---

[5] D'autres auteurs (Harris 1988, Isabelle 1992) utilisent plutôt le terme *bi-texte* pour faire référence au corpus aligné constitué de traductions.

La mise en correspondance s'appuie sur les caractéristiques communes aux textes originaux et aux traductions, par exemple, le nombre de paragraphes, le nombre de phrases et le nombre de correspondances lexicales entre les deux ensembles de textes (Bowker et Pearson 2002). Ainsi, les alignements les plus fréquents sont réalisés au niveau des paragraphes, des phrases et des mots.

Puisque les systèmes que nous décrirons s'appuient sur des textes traduits, nous allons faire référence à des corpus parallèles alignés et, donc, à la correspondance entre terme source et terme cible.

Selon Gaussier (2001), il existe trois modèles d'extraction automatique bilingue : a) double extraction de candidats-termes – alignement de termes source et cible, b) analyse de la langue source – identification de séquences de traduction contenant le terme cible, et c) analyse parallèle – identification simultanée de termes source et cible. Nous décrirons brièvement ces modèles ci-dessous.

a) Double extraction de candidats-termes – alignement de termes source et cible

D'une façon générale, ce modèle identifie, dans une première étape, les candidats-termes par langue; une liste de candidats-termes est extraite de chaque corpus. Dans une deuxième étape, les candidats-termes source et cible sont mis en correspondance au moyen du processus d'alignement au niveau des mots. Les extracteurs utilisant ce modèle se concentrent souvent sur les termes complexes binaires de nature nominale.

Au cours de l'étape d'extraction, les extracteurs se basant sur ce modèle font appel aux techniques statistiques et linguistiques comme celles que nous avons

décrites dans la section 1.3.1. La plupart d'entre eux (Blank 2000, Daille *et al.* 1994 et Gaussier et Langé 1995; 1997) travaillent sur des corpus alignés au niveau des phrases, étiquetés et lemmatisés pour pouvoir ensuite appliquer des règles associées aux patrons syntaxiques typiques. Afin de filtrer les listes de candidats-termes extraits, qui contiennent souvent une grande quantité de syntagmes libres (du bruit), les extracteurs font appel à des calculs statistiques, par exemple, à la fréquence, à l'information mutuelle, au rapport de vraisemblance et à d'autres mesures d'association.

Par la suite, pour aligner les termes source et cible, ce modèle se base très souvent sur des calculs statistiques. Parmi les calculs les plus utilisés on trouve les calculs du degré d'association, qui peuvent être appliqués aux mots simples constituant un candidat terme complexe ou bien aux termes en tant qu'entités à part entière (Gaussier et Langé 1995).

Certains chercheurs ont ajouté aux calculs statistiques des calculs basés sur des informations linguistiques, tel que le critère d'affinité des patrons, pour améliorer leurs résultats. Le critère d'affinité des patrons consiste à établir des correspondances entre les patrons syntaxiques des termes source et les patrons des termes cible à l'aide de règles de transformation de patrons. Par exemple, les termes français dont le patron de formation est NdeN sont souvent traduits en anglais par un terme suivant le patron NN.

Un autre élément utilisé pour améliorer les résultats de l'alignement de candidats-termes est celui relié à l'information positionnelle. L'information positionnelle permet, par exemple, de distinguer deux mots identiques dans une même phrase et de chercher la traduction de deux mots source contigus dans la même fenêtre de la phrase cible (Gaussier 2001).

Un exemple d'extraction bilingue basée sur le modèle que nous venons de décrire est celui de Daille *et al.* (1994). Leur système, qui fait appel à des techniques linguistiques et statistiques, identifie des syntagmes nominaux anglais et français susceptibles d'être des termes. L'extraction est réalisée à partir d'un corpus parallèle aligné au niveau des phrases et ensuite étiqueté et lemmatisé.

Dans une première étape, les chercheurs établissent les patrons syntaxiques typiques, pour les syntagmes anglais et français constitués de deux unités lexicales. Ils créent aussi des règles modélisant trois variations que ces syntagmes binaires peuvent subir dans le corpus : surcomposition (ex. régénération de + lobes latéraux = *régénération des lobes latéraux*), modification (ex. station terrienne + brouilleuse = *station terrienne brouilleuse*) et coordination (ex. assemblage de paquets + désassemblage de paquets = *assemblage et désassemblage de paquets*). Les chercheurs procèdent ainsi à l'extraction en chaque langue des syntagmes suivant les patrons syntaxiques définis. Une liste contenant des paires composées des deux lemmes (ex. *interference*, *level* ; *circuit*, *numérique*, etc.) est créée.

Pour filtrer cette liste de paires, qui constituent les candidats-termes par langue, cet extracteur a recours à la mesure de la fréquence, au rapport de vraisemblance et à l'information mutuelle.

La deuxième étape consiste à mettre en correspondance les candidats-termes source et cible en utilisant les associations bilingues de chaque unité lexicale constituant les candidats-termes. Ainsi, *worden1* et *worden2* représentent les deux unités constituant un candidat-terme anglais et *wordfr1* et *wordfr2* représentent un candidat-terme français. À chaque fois qu'un candidat-terme source et un candidat-terme cible apparaissent dans une phrase alignée, une valeur d'association leur est

attribuée. Cette technique suppose que, si un candidat-terme est la traduction d'un autre, cela implique que *wordfr1* et *wordfr2* peuvent être les traductions de *worden1* et de *worden2*. Par exemple, les unités *station* et *terrienne* du candidat-terme français seront individuellement associées aux unités anglaises *earth* et *station*.

Les taux de rappel et de précision de la plupart des extracteurs utilisant ce premier modèle d'extraction bilingue sont relativement satisfaisants mais ils font face encore à bon nombre de difficultés importantes tant au niveau de repérage de candidats-termes qu'au niveau de leur alignement. Comme nous l'avons déjà dit, l'utilisation de ces calculs et techniques s'applique surtout aux termes complexes binaires ayant une fréquence élevée, ce qui exclut les termes simples, les termes peu fréquents et les termes ayant une longueur supérieure à deux.

De même, le processus d'alignement des textes à un niveau plus fin que les phrases – dans ce cas, mot à mot – est une tâche complexe en raison des incompatibilités entre les structures des deux langues. Cette tâche se complique encore plus par le fait que les termes source ne sont pas toujours traduits de la même façon; un seul terme source peut avoir différents équivalents dans le texte cible.

Compte tenu de ces difficultés, les chercheurs proposent d'autres modèles d'extracteurs de termes, comme ceux que nous présentons dans ce qui suit.

b) Analyse de la langue source – identification de séquences de traduction contenant le terme cible

Dans ce deuxième modèle, l'extraction de candidats-termes est faite dans la langue source seulement. La deuxième étape consiste à identifier en langue cible, lors de l'alignement, une séquence susceptible de contenir la traduction du terme

source. Normalement, cette étape se base sur des modèles probabilistes de traduction. Nous décrirons deux extracteurs différents qui utilisent ce modèle d'extraction bilingue : *Termight*, système linguistique conçu par Dagan et Church (1997) et *TRINITY*, système statistique conçu par Hull (2001).

Pour l'extraction de candidats-termes source, le système *Termight* s'appuie sur le modèle linguistique des patrons syntaxiques proposé par Justeson et Katz (1995). Ce modèle extrait tous les syntagmes constitués de noms ou d'une combinaison d'adjectifs et de noms, même ceux séparés par des prépositions. L'avantage de ce modèle simple d'extraction est qu'il permet de retenir tous les syntagmes nominaux indépendamment de leur longueur et de leur fréquence dans le corpus.

Quant à la deuxième étape, le système *Termight* part de l'alignement au niveau des mots. Le système réalise une mise en correspondance partielle entre les mots du corpus parallèle, en omettant les mots qui ne peuvent pas être alignés à un niveau satisfaisant (Dagan et Church 1997 : 99). La Figure 7 est un exemple d'alignement des mots. Même si l'alignement partiel est imparfait, le système peut proposer les équivalents des termes source dans les textes cible, car il extrait la séquence de mots se trouvant entre la première et la dernière positions alignées avec l'un des mots du terme source. Ainsi, pour le candidat-terme *Optional Parameter box*, la séquence *zone Paramètres optionnels* puisque *zone* et *optionnels* constituent le premier et le dernier mots et ils sont alignés avec les mots du terme anglais (Dagan et Church 1997 : 100).

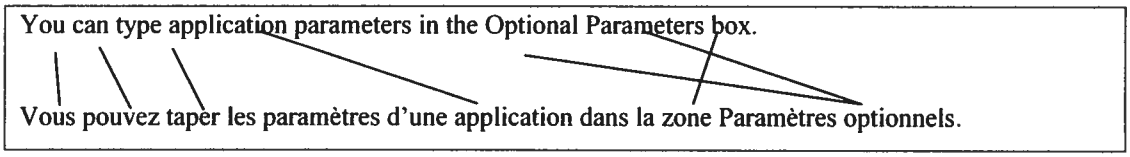

Figure 7. Exemple d'alignement partiel de mots par le système *Termight* (Dagan et Church 1997 : 100)

Hull (2001), au contraire, utilise une approche statistique pour l'identification et l'extraction de candidats-termes en langue source. Malheureusement, dans cette étude, l'auteur ne spécifie pas les techniques utilisées, il se concentre sur l'explication de la méthode d'alignement de termes.

Hull s'appuie sur le model d'alignement de Hiemstra. Ce modèle représente chaque phrase dans un tableau à double entrée, les lignes correspondant aux mots des textes source et les colonnes à ceux des textes cible. Les cellules du tableau définissent la fréquence estimée à laquelle s'alignent les mots de la ligne et de la colonne correspondantes (Hull 2001 : 229). Ensuite, un algorithme de flux des coûts minimum est appliqué entre les cellules pour repérer la traduction des termes source.

Étant donné que l'alignement des mots grammaticaux affecte le taux d'erreur, Hull traite seulement les mots pleins (noms, adjectifs, verbes et adverbes) dans ce processus. Cela signifie que, pour fonctionner, le modèle ne requiert que l'étiquetage grammatical des textes, en ce qui concerne l'information linguistique.

Comme nous l'avons vu dans les deux exemples précédents, le fait que ce deuxième modèle d'extraction bilingue de termes ne repose ni sur une analyse syntaxique profonde dans les deux langues, ni sur un alignement rigoureux des mots, procure deux avantages importants. D'abord, les termes simples peuvent être considérés et traités de la même façon que les termes complexes. De même, la

longueur des termes complexes peut varier sans que cela affecte leur identification. Enfin, le modèle permet de retenir plusieurs équivalents apparaissant dans le corpus pour un même terme source.

c) Analyse parallèle – identification simultanée de termes source et cible

Le troisième modèle consiste à aligner le corpus parallèle au niveau des mots. Ensuite, une analyse syntaxique des phrases source et phrases cible est réalisée. Enfin, une identification simultanée du terme source et du terme cible est faite. Une caractéristique importante de ce modèle est que, à la différence des autres, il fait appel à des connaissances linguistiques non seulement pour l'extraction des termes mais aussi pour l'alignement proprement dit.

Il y a moins d'extracteurs utilisant ce modèle d'extraction bilingue. Donc, dans ce qui suit nous en décrirons un seul, celui proposé par Ozdowska et Bourigault (2004).

D'abord, *SYNTEX*, un analyseur syntaxique se basant sur un corpus étiqueté et aligné au niveau des paragraphes, est utilisé pour repérer les sujets et les objets des verbes dans les deux langues. Une version du système *SYNTEX* est utilisée pour l'anglais et une autre, pour le français. L'analyse est effectuée dans les deux langues mais de façon indépendante. Après l'analyse, *SYNTEX* extrait un ensemble de mots et de syntagmes. Il faut remarquer que ce système extrait des candidats-termes simples et complexes appartenant à diverses catégories grammaticales (noms, verbes, adjectifs, adverbes).

L'étape suivante est celle de l'alignement qui, elle aussi, se divise en deux parties. D'abord, les candidats-termes source et cible sont alignés en fonction de leur

fréquence d'apparition dans les paragraphes. Par conséquent, les termes ayant une fréquence élevée sont d'abord mis en correspondance. Cela est fait en calculant la fréquence de cooccurrence. La deuxième étape consiste à aligner les candidats-termes moins fréquents. Ces termes son mis en correspondance phrase à phrase à partir des relations de dépendance syntaxique identifiées par *SYNTEX*.

Le deuxième processus d'alignement se fonde sur l'idée selon laquelle « les relations de dépendance syntaxique sont susceptibles, d'une part, de confirmer ou d'infirmer des liens d'appariement et, d'autre part, de créer de nouveaux liens » (Ozdowska et Bourigault 2004 : 2). Les auteurs expliquent ce raisonnement de la façon suivante :

> « Si deux mots Ts_j et Tc_q sont appariés et s'il existe une relation de dépendance syntaxique entre Ts_j et Ts_i, d'une part, et entre Tc_q et Tc_p, d'autre part, alors Ts_i et Tc_p peuvent être appariés. »
> (Ozdowska et Bourigault 2004 : 2)

Un avantage de ce mécanisme d'alignement est que, puisque les candidats-termes sont alignés sans prendre en considération leur catégorie grammaticale, il est possible de mettre en correspondance des paires des termes ayant une longueur différente ou n'appartenant pas à la même catégorie grammaticale.

Selon les évaluations faites par les chercheurs, ce modèle d'extraction bilingue a un taux de précision considérablement élevé (91,7 %) ; le mécanisme d'alignement par le moyen des propagations de liens d'appariement le long de relations syntaxiques s'avère prometteur.

1.3.3 Synthèse des méthodes d'extraction de termes bilingue

Dans les sous-sections précédentes, nous avons décrit quelques critères linguistiques et statistiques sur lesquels plusieurs extracteurs de termes se basent. À la fin de la section 1.3.1, nous avons affirmé que la combinaison de diverses techniques produit des meilleurs résultats que leur application isolée.

D'ailleurs, même si une application entièrement automatique est visée, les évaluations des niveaux de précision et de rappel obtenus par les extracteurs montrent que la tâche d'extraction de termes est encore loin de pouvoir être automatisée entièrement. La liste de termes générée par un extracteur n'est qu'une liste de candidats que le spécialiste humain devra épurer et compléter avec ses propres recherches.

Une caractéristique importante de plusieurs extracteurs est le fait qu'ils sont configurés pour identifier et pour extraire seulement certains types de termes, normalement les termes complexes ayant une longueur précise et appartenant à la catégorie grammaticale des noms. Même si quelques-uns font appel à des techniques pouvant repérer d'autres types de termes, les extracteurs restent encore limités dans ce sens.

En ce qui concerne les extracteurs de termes bilingues, nous avons présenté trois modèles de fonctionnement différents.

Le premier modèle – celui de l'extraction de candidats-termes dans les deux langues – est considéré comme le modèle de base. Il combine les techniques linguistiques et statistiques pour l'identification de termes mais s'appuie en grande partie sur des méthodes statistiques pour la mise en correspondance des termes

source et cible. L'un des inconvénients de l'alignement des termes de ce modèle est qu'il ne permet qu'un équivalent par terme source et souvent cet équivalent doit plus ou moins correspondre aux caractéristiques morphosyntaxiques du terme source.

Compte tenu des limites imposées par le premier modèle, deux modèles alternatifs sont apparus. Le modèle analysant seulement la langue source pour ensuite identifier une séquence de traduction contenant le terme cible permet de traiter tant les termes complexes que les termes simples. Il inclut aussi des syntagmes de longueurs diverses. De plus, il permet de repérer plusieurs équivalents pour un même terme source, ce qui signifie que certaines variantes terminologiques sont déjà considérées.

Le modèle qui part de l'alignement mot à mot du corpus bilingue et d'une analyse syntaxique parallèle pour ensuite extraire des candidats-termes dans les deux langues semble venir à bout des limites du premier modèle. À la différence des autres modèles, ce dernier exige l'utilisation de connaissances linguistiques tout le long du processus. Outre la possibilité de repérer des termes appartenant à différentes catégories grammaticales, ce modèle ouvre la possibilité de considérer certains phénomènes linguistiques intervenant dans la traduction des termes source. Cela implique d'étudier davantage les types de variation entre les termes source et cible pour pouvoir modéliser ce phénomène linguistique et, de cette façon, améliorer l'efficacité des extracteurs automatiques bilingues.

Jusqu'à présent, peu d'extracteurs bilingues considèrent les connaissances linguistiques, surtout celles reliées à la variation, comme moyen d'amélioration des taux de précision et de rappel. Nous jugeons alors pertinent de réaliser plus d'études sur ce sujet qui puissent apporter des connaissances applicables au développement des systèmes d'extraction.

Enfin, et puisqu'il s'agit de connaissances dépendantes de la langue et que la majorité des extracteurs existants portent sur l'anglais et le français, nous proposons une étude de la variation terminologique anglais-espagnol.

Chapitre 2 : Méthodologie

Dans ce chapitre nous présentons les détails de la sélection des textes en anglais et en espagnol qui font partie du corpus, la construction du corpus parallèle et la méthode utilisée pour sélectionner l'ensemble de termes qui font l'objet de l'analyse de la variation terminologique. De même, nous présentons la méthode adoptée pour recueillir et décrire les variantes terminologiques des termes sélectionnés.

2.1 Sélection des textes formant le corpus parallèle

Nous avons décidé de réaliser l'analyse comparative de la variation terminologique sur un corpus parallèle portant sur le domaine de l'environnement pour trois raisons que nous décrirons dans cette section.

Premièrement, c'est un domaine d'intérêt général qui a acquis plus d'importance au cours des dernières années. Aujourd'hui, le respect de l'environnement est un élément essentiel des politiques commerciales et économiques de plusieurs pays.

Deuxièmement, l'environnement constitue un domaine de recherche très large et il a des liens étroits avec une grande variété d'autres domaines : la chimie, la biologie, la médicine, la santé publique, l'écologie et la législation nationale et internationale, pour n'en mentionner que quelques-uns. Par conséquent, il existe une grande quantité et une variété d'études portant sur ce sujet. En outre, ce matériel est assez souvent du domaine public, en raison de sa nature. Il est donc très facile d'y avoir accès.

Dernièrement, en tant que traductrice professionnelle, l'auteure du présent travail possède une expérience importante dans le domaine de l'environnement et nous considérons que nos connaissances dans ce domaine constituent certainement un avantage pour l'analyse et l'interprétation des résultats de la recherche.

Évidemment, il serait presque impossible de couvrir tout le domaine de l'environnement dans la présente recherche. C'est la raison pour laquelle nous nous concentrons sur un seul sous-domaine : les composés chimiques industriels qui présentent des risques pour la santé humaine et l'environnement.

Nous avons exploré trois sites Web d'organismes régionaux et internationaux de l'environnement : la Commission de coopération environnementale de l'Amérique du Nord (CCE), la Convention sur la Diversité Biologique (CDB) et le Programme des Nations Unies pour l'environnement (PNUE), d'où nous avons extrait un total de 23 documents que nous avons jugés représentatifs du sous-domaine choisi.

Tous les textes sélectionnés, bien sûr en format électronique, ont été rédigés à l'origine en anglais et ensuite traduits, entre autres langues, vers l'espagnol.

D'une façon générale, on peut diviser les textes sélectionnés en deux types : d'une part, les conventions, protocoles et plans d'action et, d'autre part, les études scientifiques.

1. *Les conventions, protocoles ou plans d'action.* Le processus d'élaboration et de publication du premier type de textes est très particulier ; il peut être facilement caractérisé. Les textes sont le résultat de discussions tenues régulièrement sur un sujet donné dont les acteurs sont les représentants

gouvernementaux et non gouvernementaux de diverses nations poursuivant un intérêt commun. Normalement, un groupe de spécialistes en la matière est responsable de la rédaction et de la mise au point du texte final. Compte tenu de son caractère officiel et multinational, le texte est écrit dans un langage formel et normalisé et il doit être traduit et publié dans les langues des pays impliqués. On pourrait affirmer que le lectorat cible est d'abord l'ensemble des représentants et des nations qui signent le document et, en dernière instance, le public en général. Par conséquent, ces textes – originaux et traductions – ont toujours une structure bien définie et assez homogène au niveau de leur présentation en format électronique, les sections comprises, la mise en garde, les paragraphes, etc. Les documents sélectionnés pour l'analyse comparative ont une taille respectable, entre 8 000 et 22 000 mots chacun et ils ont été publiés entre les années 1989 et 2000.

2. *Les études scientifiques.* Nous avons choisi les études et les évaluations scientifiques sur certains composés chimiques qui font l'objet des conventions, protocoles et plans d'action mentionnés au préalable, ou encore, qui ont un rapport direct avec eux. Ce sont des documents rédigés par des spécialistes visant à décrire les caractéristiques, l'usage, les sources et les risques que les substances présentent dans diverses régions du monde. Les études sont ensuite publiées par l'instance – normalement un organisme international – qui a sollicité leur élaboration. Étant donné qu'il s'agit de rapports qui constituent la base d'une prise de décision, les lecteurs visés sont les instances qui ont commandé l'étude ainsi que les pays ou les parties et régions impliqués. Compte tenu des thèmes, des objectifs et des auteurs de tels documents, nous considérons que leur niveau de spécialisation n'est pas trop élevé puisqu'ils s'adressent à des lecteurs moins spécialisés. Les textes

du deuxième groupe sont volumineux – entre 10 000 et 62 000 mots chacun – et ils ont été publiés entre 1990 et 2003.

Le Tableau IV présente la liste des 23 documents anglais qui forment la première partie du corpus parallèle et qui contient un total de 364 890 mots. Bien sûr, nous avons aussi téléchargé la traduction correspondante en espagnol de chaque document original, et cette deuxième partie du corpus contient 418 490 mots.

Il faut remarquer que les documents téléchargés se trouvaient en format Word ou PDF. Donc, nous avons dû les convertir en format texte seul pour faciliter leur traitement.

Tableau IV. Liste des textes anglais formant le corpus sur les composés chimiques dangereux

	Titre du document	Nom du fichier	Nombre de mots	Date de publica-tion	Source
1	Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal	baselconvention-en.txt	15 423	1989	PNUE
2	Convention on Biological Diversity	cbd-en.txt	9 258	1992	CDB
3	Decision Document on Lead under the Process for Identifying Candidate Substances for Regional Action under the Sound Management of Chemicals Initiative	lead-public-consult-en.txt	26 286	2003	CCE
4	Decision Document on Lindane under the Process for Identifying Candidate Substances for Regional Action under the Sound Management of Chemicals Initiative	linddd-en.txt	13 204	2000	CCE
5	Dioxin and Furan Inventories	dioxin-furan-inventory-en.txt	34 294	1999	PNUE
6	Global Mercury Assessment. Excerpts of the full report	global_mercury-en.txt	16 378	2002	PNUE
7	Kyoto Protocol to the United Nations Framework Convention on Climate Change	kyotoprotocol-en.txt	8 634	1992	CDB

8	Nomination Dossier on Lead. Submission by The United States to the Working Group of the Sound Management of Chemicals	Nom-lead-en.txt	4 827	1998	CCE
9	North American Regional Action Plan on Chlordane	narap-chlordane-en.txt	2 763	1997	CCE
10	North American Regional Action Plan on DDT	narap-ddt-en.txt	6 182	1997	CCE
11	North American Regional Action Plan on Mercury. Phase II	hgnarap-en.txt	8 119	2000	CCE
12	North American Regional Action Plan on PCB	narap-pcb-en.txt	12 883	1996	CCE
13	PCB Transformers and Capacitors: From management to reclassification	PCBtranscap-en.txt	14 570	2002	PNUE
14	Regionally Based Assessment of Persistent Toxic Substances. Central America and the Caribbean	pts-centralamerica-en.txt	60 251	2002	PNUE
15	Regionally Based Assessment of Persistent Toxic Substances. Eastern and Western South America	pts-southamerica-en.txt	46 598	2002	PNUE
16	Rotterdam Convention on the Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade	rotterdam-convention-en.txt	8 150		PNUE
17	Stockholm Convention on Persistent Organic Pollutants	stockholmconvention-en.txt	13 400		CDB
18	The Montreal Protocol on Substances that Deplete the Ozone Layer	MontrealProtocol-en.txt	12 207	2000	CDB
19	The Status of Mercury in Canada. Report two	Hgcan-en.txt	18 932	2000	CCE
20	The Status of Mercury in Mexico. First Draft	Hgmex.en.txt	5 492	2000	CCE
21	The Vienna Convention for the Protection of the Ozone Layer	viennaconvention-en.txt	6 772	2001	CDB
22	United Nations Framework Convention on Climate Change	UNFCC-en.txt	8 403	1992	PNUE
23	US Status Report on Mercury Activities	Hgus-en.txt	12 235	2000	CCE
TOTAL			**23 FICHIERS - 364 891 MOTS**		

2.2 Alignement du corpus parallèle

Pour l'alignement du corpus parallèle, nous utilisons l'une des applications du logiciel *Logiterm,* système de traduction assistée par ordinateur (TAO) créé par l'entreprise canadienne Terminotix Inc.

Nous avons choisi le générateur de textes parallèles de *Logiterm* – conçu au Laboratoire de Recherche appliquée en linguistique informatique (RALI) de l'Université de Montréal –, en raison de sa grande efficacité d'alignement. Ainsi, la construction du corpus parallèle avec les 23 textes anglais et les 23 textes espagnols – tous en format texte seul – a été réalisée rapidement et a produit des résultats satisfaisants.

L'aligneur de *Logiterm* génère les textes alignés en format HTML, ce qui facilite leur consultation et leur manipulation hors de l'environnement de ce système de TAO. La Figure 8 montre une portion de l'un des textes alignés de notre corpus en format HTML.

Ensuite, les textes alignés ont été regroupés et placés dans le même répertoire afin de pouvoir procéder à l'interrogation du corpus, toujours dans l'environnement de *Logiterm*.

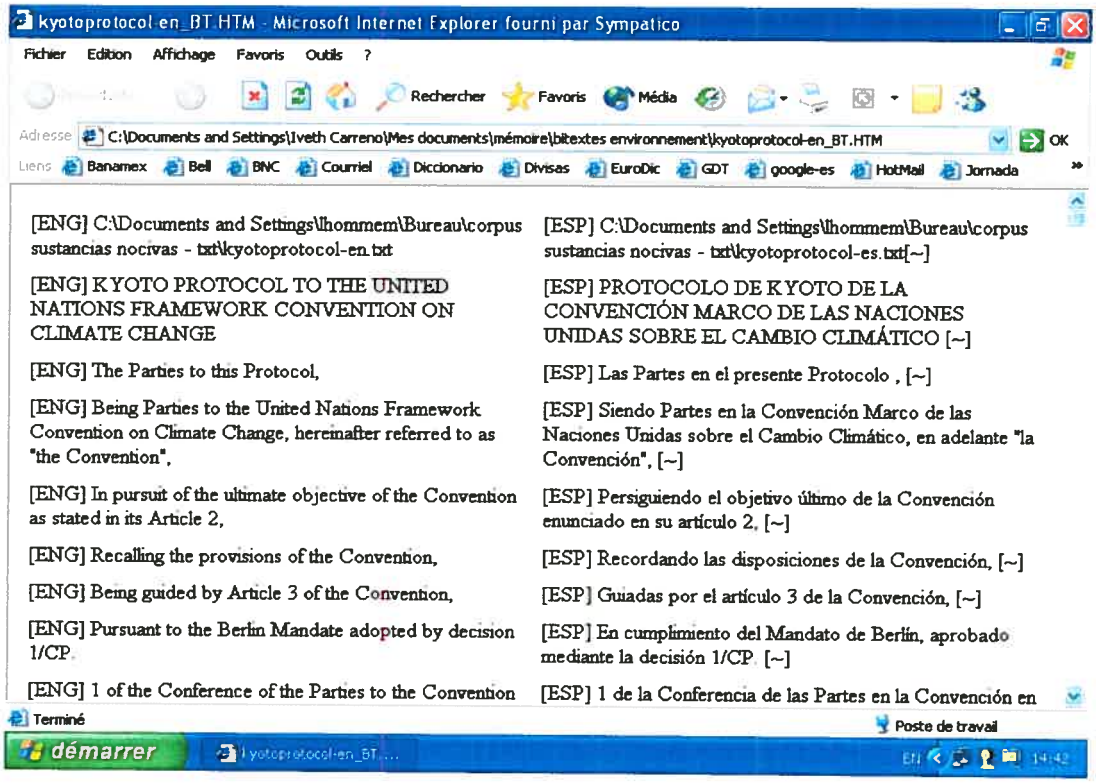

Figure 8. Texte aligné généré par *Logiterm* en format HTML à partir du document original du Protocole de Kyoto et de sa traduction en espagnol

2.3 Sélection des termes à analyser

Une fois le corpus parallèle aligné, nous avons procédé à la sélection des termes qui font l'objet de l'analyse comparative.

Tout d'abord, il fallait déterminer la quantité la plus appropriée de termes à retenir à partir d'un corpus de plus de 300 000 mots. Prenant en considération les objectifs, les caractéristiques et les limites de la présente recherche, nous avons jugé que 50 termes était une quantité convenable.

La sélection des termes a été réalisée en deux étapes : présélection de termes candidats par des méthodes automatisées et sélection finale des termes basée sur trois critères. La langue de référence pour l'extraction terminologique est la langue des textes originaux, soit l'anglais.

2.3.1 Présélection de termes

À cette première étape, nous avons réalisé une extraction automatique des termes candidats simples et complexes avec la version Web du logiciel *TermoStat,* conçu par Patrick Drouin (2002). *TermoStat* est un système qui a recours à des méthodes statistiques et linguistiques (méthode hybride, voir la section 1.3.1 du présent travail). En outre, il identifie des pivots lexicaux spécialisés (PLS) à la suite de la comparaison du lexique d'un corpus technique (corpus d'analyse) à celui d'un corpus non technique (corpus de référence). Avec les résultats de cette comparaison et des PLS identifiés, le système génère automatiquement une liste de candidats-termes (Drouin 2002).

À la différence d'autres extracteurs automatiques, *TermoStat* propose non seulement des termes complexes mais aussi des termes simples dans la même liste de résultats, et cette particularité du système convient à nos besoins de sélection terminologique.

L'interface Internet en anglais de *TermoStat* nous a permis d'appliquer le système à chacun des fichiers du corpus anglais. L'extracteur automatique génère une table de trois colonnes : la première contient le nombre d'occurrences du terme candidat, la deuxième présente les termes candidats et la troisième montre la tête de chaque terme candidat. De plus, chaque terme proposé constitue en même temps un lien à ses contextes. La table est triée en ordre décroissant selon la fréquence du

terme. La Figure 9 montre un exemple des résultats générés par *TermoStat* appliqué à l'un des textes anglais de notre corpus.

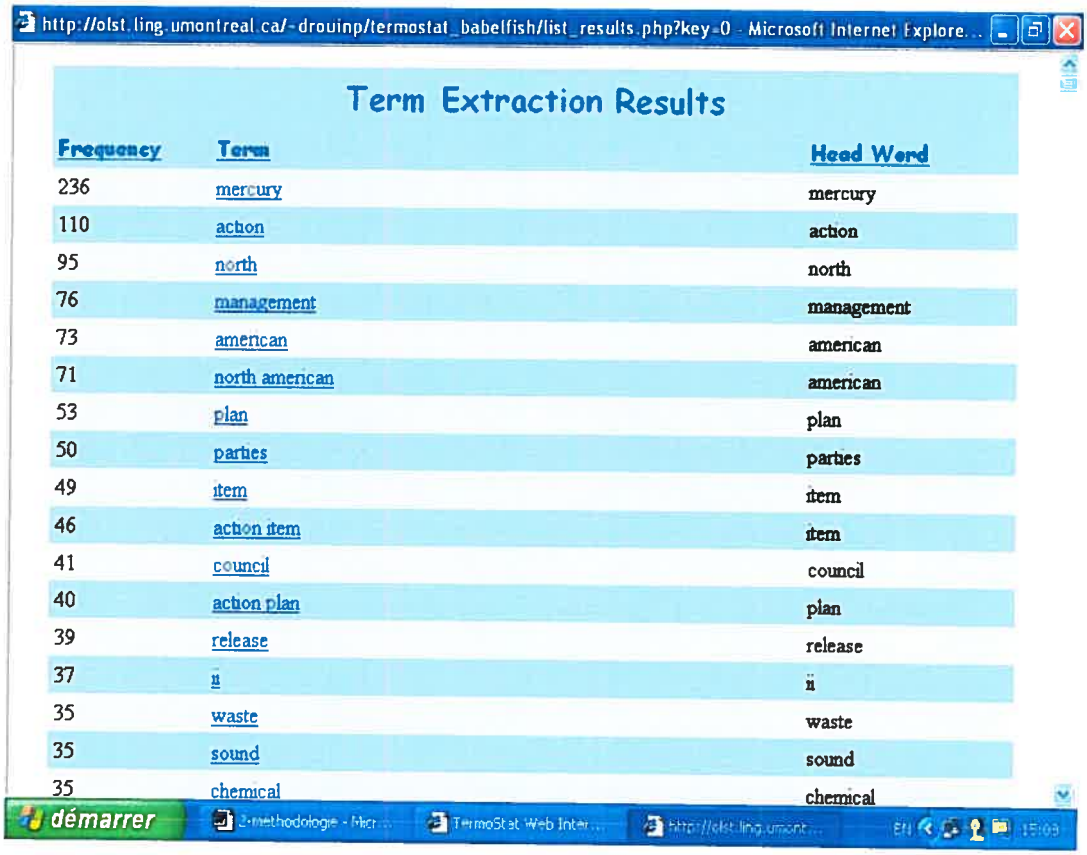

Figure 9. Liste des termes candidats générée par l'extracteur automatique *TermoStat* appliqué à l'un des textes anglais du corpus sur les substances dangereuses

Nous avons obtenu 23 listes de termes candidats, que nous avons regardées attentivement afin de déterminer les caractéristiques des termes ainsi que leur utilité et leur adéquation pour la recherche.

Nous avons conclu que, effectivement, les listes générées automatiquement seraient utiles car, selon les critères établis pour la sélection, certains termes proposés étaient appropriés.

2.3.2 Sélection finale des termes

Pour l'établissement de la liste définitive des 50 termes à analyser, nous nous sommes appuyée sur trois critères fondamentaux : la pertinence, la fréquence et la répartition.

a) Pertinence

La première considération importante est celle de la pertinence des termes candidats. Nous avons retenu les termes reliés particulièrement aux composés chimiques industriels en tant que substances dangereuses pour l'environnement et la santé humaine. Nous avons ainsi écarté une grande partie des termes proposés dans chaque liste de termes candidats. De chaque liste nous n'avons retenu qu'entre 8 et 30 termes candidats. Dans le but de vérifier leur validité, nous nous sommes assurée que les termes retenus se trouvaient dans des glossaires spécialisés et banques terminologiques sur Internet, telles que *Eurodicautom* et *Termium*, et qu'ils étaient classifiés dans des rubriques ou des sous-domaines identiques à celui choisi pour la présente recherche.

b) Fréquence

Après avoir appliqué le premier filtre, soit le critère de pertinence, nous avons évalué la fréquence des termes retenus pour chaque texte original. Il faut remarquer que, à ce moment-là, nous avions 23 listes épurées contenant entre 8 et 30 termes simples et complexes.

Ensuite, en appliquant le critère de fréquence, nous avons jugé convenable de conserver seulement les termes dont le nombre d'occurrences dans le texte était

élevé. Le nombre d'occurrences minimal pour la rétention d'un terme a été fixé à 5. De cette façon, les listes de termes ont été encore une fois raffinées.

c) Répartition

Jusqu'à cette étape, nous avions des listes contenant des termes pertinents ayant une fréquence élevée, mais ces listes étaient encore très hétérogènes, car certains termes n'apparaissaient que dans une seule liste, c'est-à-dire, dans un seul document du corpus. Étant donné que l'un des objectifs particuliers de cette étude est d'observer l'usage d'un terme dans divers textes, nous avons décidé d'appliquer un troisième critère, celui de la répartition dans tout le corpus. En conséquence, seuls les termes qui figuraient dans plus de deux documents ont été retenus.

Comme dernière considération, et dans le but d'assurer un équilibre dans la sélection, nous avons choisi le même nombre de termes simples et de termes complexes. C'est ainsi que nous avons constitué les listes de 25 termes simples et 25 termes complexes pour l'analyse comparative. Quant à la liste de termes simples, il faut signaler que nous avons choisi des termes appartenant à diverses catégories grammaticales. Bien que la plupart de ces termes soient des noms, nous avons aussi inclus des verbes et des adjectifs ainsi que le sigle d'un terme complexe.

En ce qui concerne les termes complexes, tous sont des noms. Cependant, comme on peut le constater dans le Tableau VI, leurs structures syntagmatiques sont variées.

Les Tableaux V et VI présentent respectivement les 25 termes simples et les 25 termes complexes qui feront l'objet de l'analyse comparative de la variation.

Tableau V. Liste des termes simples sélectionnés pour l'analyse de la variation

	Terme	Catégorie grammaticale
1	abatement	nom
2	anthropogenic	adjectif
3	bioaccumulate	verbe
4	bioconcentrate	verbe
5	carcinogen	nom
6	chlor-alkali	nom
7	coal-fired	adjectif
8	compartment	nom
9	deposition	nom
10	disposal	nom
11	emission	nom
12	foundry	nom
13	incinerator	nom
14	landfill	nom
15	leakage	nom
16	monitoring	nom
17	pathway	nom
18	PCB	abréviation
19	pesticide	nom
20	recycling	nom
21	release	nom
22	remediation	nom
23	reservoir	nom
24	sink	nom
25	trace	nom

Tableau VI. Liste des termes complexes sélectionnés pour l'analyse de la variation

	Terme	Catégorie grammaticale	Structure syntagmatique
1	action plan	nom	n + n
2	adverse effect	nom	adj + n
3	ambient air concentration	nom	n + n + n
4	body burden	nom	n + n
5	daily intake	nom	adv + n
6	dangerous goods	nom	adj + n
7	environmental fate	nom	adj + n
8	environmental media	nom	adj + n
9	flue gas	nom	adj + n
10	hazardous air pollutant	nom	adj + n + n
11	hazardous waste	nom	adj + n
12	industrial chemical	nom	adj + n
13	life cycle	nom	n + n
14	long-range transport	nom	(adj + n) + n
15	municipal solid waste	nom	adj + adj + n
16	persistent organic pollutant	nom	adj + adj + n
17	persistent toxic substance	nom	adj + adj + n
18	point source	nom	adj + n
19	risk assessment	nom	n + n
20	sewage sludge	nom	n + n
21	sinter plant	nom	n + n
22	sound management	nom	adj + n
23	transboundary movement	nom	adj + n
24	vector control	nom	n + n
25	waste management	nom	n + n

2.4 Analyse des termes sélectionnés

Pour l'interrogation du corpus bilingue ou, autrement dit, pour l'observation en contexte de chaque terme sélectionné, nous avons utilisé le concordancier bilingue du logiciel *Logiterm*.

L'interface des résultats de la requête est constituée de deux grandes colonnes verticales qui alignent les contextes en anglais et en espagnol. La colonne de gauche présente les contextes anglais dans lesquels se trouve le terme recherché. Chaque occurrence du terme est surlignée en jaune pour son identification rapide. La colonne de droite présente les contextes correspondants en espagnol. En cliquant sur le côté gauche d'un contexte anglais, on accède aux fichiers complets. Au-dessous des deux colonnes verticales se trouve une barre d'état qui indique le nombre total d'occurrences du terme recherché ainsi que le nombre de fichiers dans lesquels celui-ci apparaît. La Figure 10 est un exemple de la présentation de résultats du concordancier bilingue de *Logiterm*.

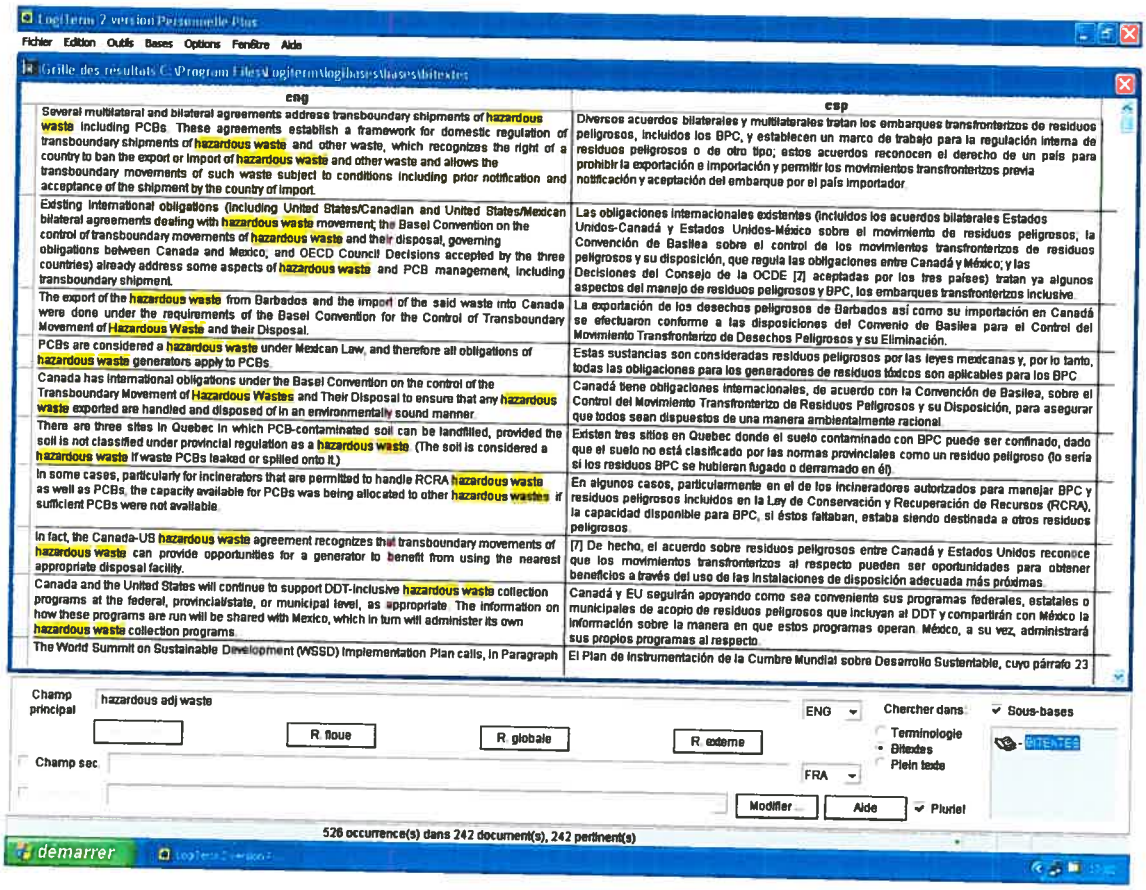

Figure 10. Présentation des résultats de la requête « hazardous waste » appliquée à notre corpus et générée par le concordancier bilingue de *Logiterm*

Pour l'enregistrement systématique des informations dérivées de l'observation en contexte des termes anglais ainsi que de leurs équivalents et variantes espagnols, nous avons créé une base de données à l'aide du logiciel *Access*.

Dans un premier tableau à cinq champs nous avons organisé les données reliées aux termes anglais. Le premier champ correspond aux termes, le deuxième indique sa catégorie grammaticale, le troisième contient le nombre total d'occurrences de chaque terme dans le corpus, le quatrième champ présente le nombre total de fichiers dans lesquels le terme apparaît et, finalement, dans le cinquième champ, nous spécifions s'il s'agit d'un terme simple ou d'un terme complexe. La Figure 11 montre un extrait de ce tableau.

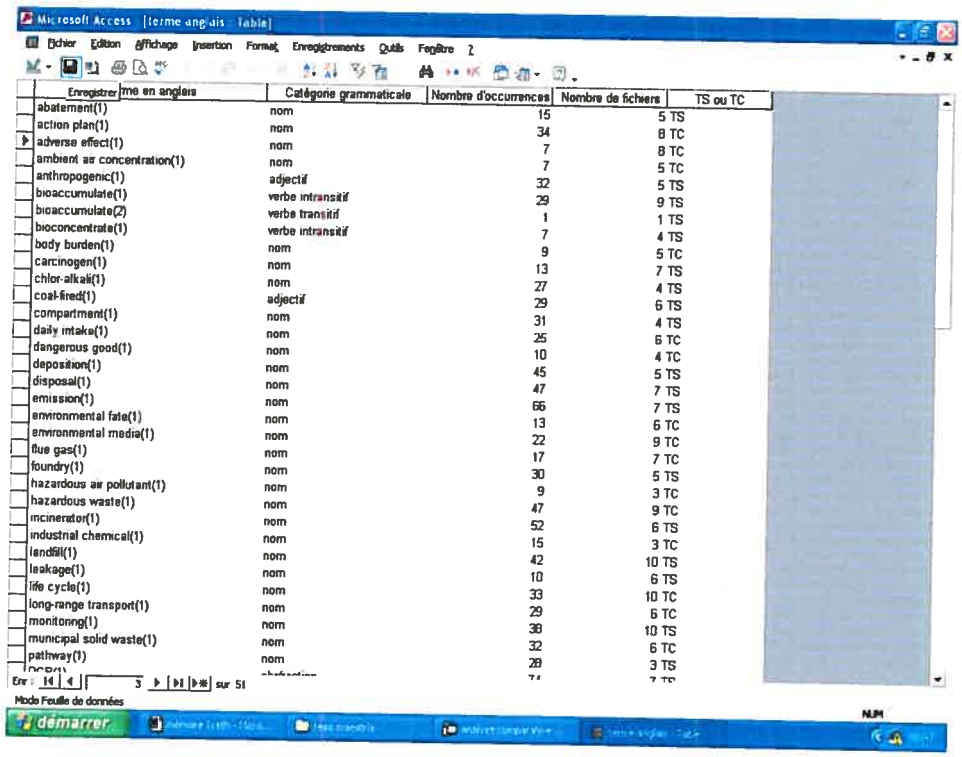

Figure 11. Extraits du tableau décrivant les termes anglais

Les données concernant les équivalents et les diverses variantes espagnoles ont été enregistrées dans deux tableaux différents, un tableau pour les termes simples et un autre pour les termes complexes. Cependant, tous les deux ont la même structure. Le premier champ est celui du terme anglais, le deuxième contient son équivalent ou sa variante, le troisième décrit la catégorie grammaticale de l'équivalent/variante. Les quatrième et cinquième champs présentent, respectivement, le nombre d'occurrences et le nombre de fichiers pour chaque équivalent/variante. Dans le sixième champ, nous indiquons si l'équivalent/variante est un terme simple ou complexe. Le dernier champ nous permet d'ajouter toutes sortes d'observations ou de commentaires additionnels qui peuvent être pertinents lors de l'interprétation des résultats.

Il faut préciser que, pour faciliter l'analyse des données, nous avons enregistré chaque équivalent et variante sur une ligne différente, de telle sorte que pour un même terme anglais il y aura autant de lignes que d'équivalents et variantes trouvés en espagnol. De plus, dans le champ du terme anglais, le nom du terme est accompagné d'un numéro entre parenthèses. Le numéro 1 indique qu'il s'agit toujours de la même unité lexicale qui revêt le même sens. Par contre, le numéro 2, ou 3, indique que la forme lexicale est la même mais que le sens diffère de celui indiqué par le numéro 1. La Figure 12 présente la structure du tableau des équivalents et des variantes en espagnol.

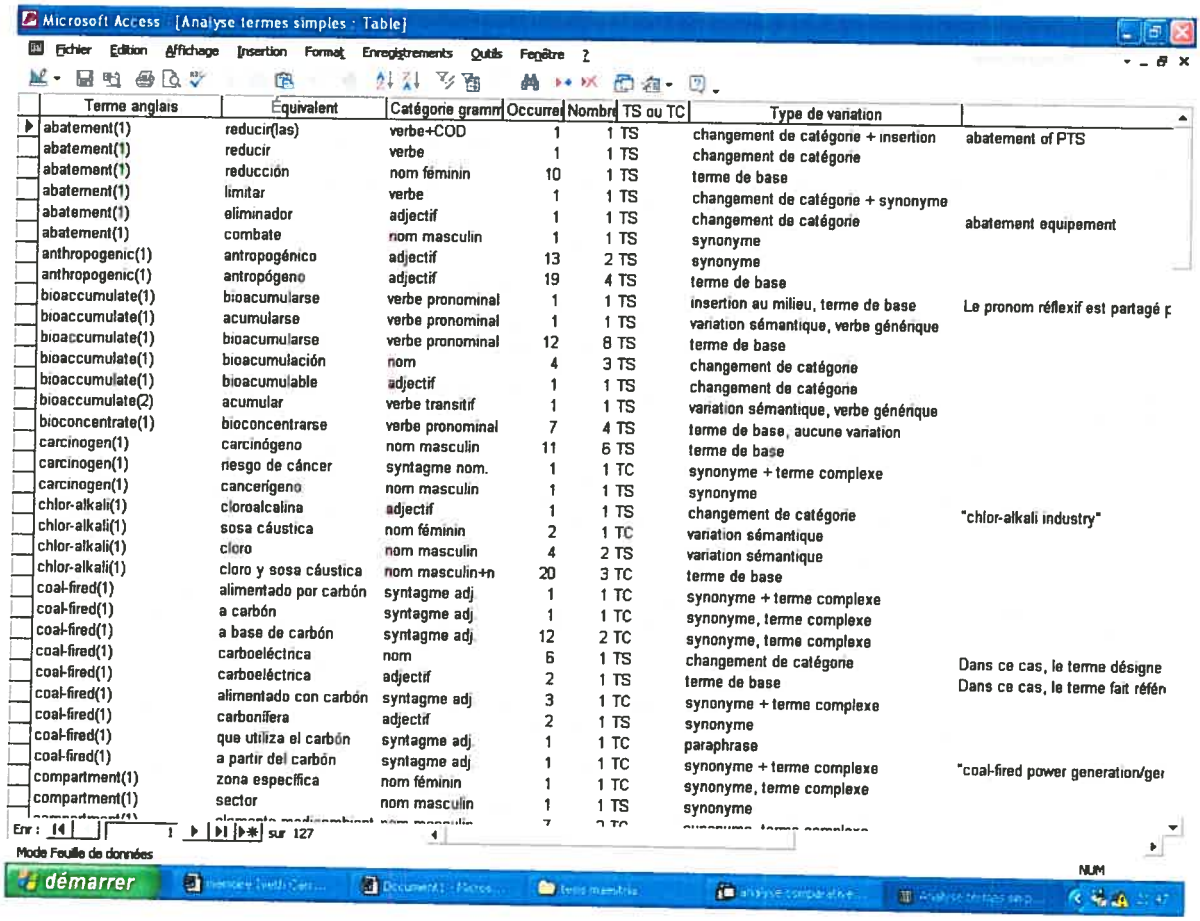

Figure 12. Extrait du tableau contenant les données sur les équivalents et les variantes en espagnol

2.4.1 Fréquence et répartition des termes simples

Un total de 25 termes simples a été analysé, dont 19 sont des noms, 3 sont des adjectifs, 2 sont des verbes et 1 est un sigle.

Le Tableau VII présente une version résumée de la base de données des termes simples anglais pour montrer le nombre d'occurrences de chaque terme ainsi que le nombre de fichiers dans lesquels ils apparaissent.

Tableau VII. Fréquence et répartition dans le corpus des termes simples en anglais

	Terme	Catégorie grammaticale	Nombre d'occurrences	Nombre de fichiers
1	abatement(1)	nom	15	5
2	anthropogenic(1)	adjectif	32	5
3a	bioaccumulate(1)	verbe tr.	29	9
3b	bioaccumulate(2)	verbe intr.	1	1
4	bioconcentrate(1)	verbe	7	4
5	carcinogen(1)	nom	13	7
6	chlor-alkali(1)	nom	27	4
7	coal-fired(1)	adjectif	29	6
8	compartment(1)	nom	31	4
9	deposition(1)	nom	45	5
10	disposal(1)	nom	47	7
11	emission(1)	nom	66	7
12	foundry(1)	nom	30	5
13	incinerator(1)	nom	52	6
14	landfill(1)	nom	42	10
15	leakage(1)	nom	10	6
16	monitoring(1)	nom	38	10
17	pathway(1)	nom	28	3
18	PCB(1)	abréviation	74	7
19	pesticide(1)	nom	53	6
20	recycling(1)	nom	37	9
21	release(1)	nom	71	10
22	remediation(1)	nom	36	8
23	reservoir(1)	nom	31	9
24	sink(1)	nom	30	6
25	trace(1)	adjectif	21	8

Nous observons un seul cas de changement de sens pour une même forme lexicale en anglais : le verbe *bioaccumulate*, qui, dans le corpus, est utilisé soit comme un verbe transitif, soit comme un verbe intransitif. Comme nous l'avons expliqué à la section précédente, nous distinguons les deux sens en ajoutant un numéro entre parenthèses à la fin de chaque terme.

Le terme ayant le nombre d'occurrences le plus élevé est *PCB*, avec 74 occurrences réparties dans 7 fichiers. Le terme ayant le moins grand nombre d'occurrences est *bioconcentrate*, avec 7 occurrences réparties dans 4 fichiers. Le terme ayant la répartition la plus élevée est *landfill*, avec 42 occurrences réparties dans 10 fichiers. Le terme ayant la répartition la plus basse est *pathway*, avec 28 occurrences dans seulement 3 fichiers.

2.4.2 Fréquence et répartition des termes complexes

Comme nous l'avons spécifié auparavant, nous avons analysé 25 termes complexes. Tous appartiennent à la même catégorie grammaticale, à savoir celle du nom. Cependant, comme on peut le constater au Tableau VIII, nous avons relevé une variété de structures syntagmatiques. Ce tableau est aussi une version résumée de la base de données des termes complexes et montre le nombre d'occurrences de chaque terme ainsi que le nombre de fichiers dans lesquels ils apparaissent.

Tableau VIII. Fréquence et répartition dans le corpus des termes complexes en anglais

	Terme	Catégorie grammaticale	Structure syntagmatique	Nombre d'occur-rences	Nombre de fichiers
1	action plan(1)	nom	n + n	34	8
2	adverse effect(1)	nom	adj + n	7	8
3	ambient air concentration(1)	nom	n + n + n	7	5
4	body burden(1)	nom	n + n	9	5
5	daily intake(1)	nom	adv + n	25	6
6	dangerous goods(1)	nom	adj + n	10	4
7	environmental fate(1)	nom	adj + n	13	6
8	hazardous waste(1)	nom	adj + n	47	9
9	industrial chemical(1)	nom	adj + n	15	3
10	life cycle(1)	nom	adj + n + n	33	10
11	municipal solid waste(1)	nom	adj + n	32	6
12	persistent organic pollutant(1)	nom	adj + n	27	3
13	point source(1)	nom	n + n	25	9
14	risk assessment(1)	nom	(adj + n) + n	30	11
15	sewage sludge(1)	nom	adj + adj + n	33	7
16	sound management(1)	nom	adj + adj + n	31	11
17	transboundary movement(1)	nom	adj + adj + n	31	6
18	waste management(1)	nom	adj + n	34	10
19	long-range transport(1)	nom	n + n	29	6
20	environmental media(1)	nom	n + n	22	9
21	flue gas(1)	nom	n + n	17	7
22	hazardous air pollutant(1)	nom	adj + n	9	3
23	sinter plant(1)	nom	adj + n	29	3
24	vector control(1)	nom	n + n	21	4
25	persistent toxic substance(1)	nom	n + n	27	6

Le terme ayant le nombre d'occurrences le plus élevé est *hazardous waste*, avec 47 occurrences réparties dans 9 fichiers. Le terme ayant le moins grand nombre d'occurrences est *ambient air concentration*, avec 7 occurrences réparties dans 5 fichiers. Le terme ayant la répartition la plus élevée est *sound management*, avec 31 occurrences réparties dans 11 fichiers. Le terme ayant la répartition la plus basse est *hazardous air pollutant*, avec 9 occurrences dans seulement 3 fichiers.

Chapitre 3 : Résultats de l'analyse

Ce chapitre présente les données générées à partir de l'observation en contexte des termes simples et complexes qui font l'objet de cette étude ainsi qu'une analyse de ces résultats. La première section présente la méthode de sélection de nos termes de base espagnols correspondant aux termes simples et aux termes complexes anglais. La deuxième section décrit les résultats obtenus pour les termes simples ; la troisième, les résultats obtenus pour les termes complexes. Dans la quatrième section, nous réalisons une comparaison entre les résultats des termes simples et les résultats des termes complexes. Enfin, dans la cinquième section, nous discutons des possibilités de repérage des variantes terminologiques observées ainsi que leur incidence sur l'extraction de termes bilingue.

3.1 Sélection des termes de base espagnols

Avant d'entreprendre l'observation en contexte des termes, il a été nécessaire de définir une liste d'équivalents terminologiques anglais-espagnol. Autrement dit, il nous a fallu déterminer le terme de base espagnol qui correspondait à chaque terme simple anglais et à partir duquel les variantes sont générées.

D'abord, nous avons établi une liste d'équivalents candidats en nous appuyant sur nos connaissances de la terminologie du domaine de l'environnement. À cette étape, il y avait des termes anglais pour lesquels nous avions plus d'un terme équivalent espagnol. Pour valider notre liste, nous l'avons comparée aux équivalents proposés dans les sources suivantes :

- Banque terminologique TERMIUM (2004) ;
- Glossaire multilingue en ligne de l'Agence Européenne pour l'Environnement, AEE (2004) ;
- Glosario electrónico de la Conferencia de las Naciones Unidas sobre el Medio Ambiente y el Desarrollo, CNUMAD (1998) ;
- Glosario electrónico de la Organización Mundial del Comercio, OMC (1998).

Nous avons décidé de retenir les équivalents espagnols faisant le plus grand consensus parmi les sources consultées. Toutefois, puisque notre étude s'appuie sur l'analyse d'un corpus parallèle, il fallait aussi prendre en considération les équivalents y apparaissant. C'est la raison pour laquelle, parmi deux ou plusieurs équivalents candidats, nous avons donné priorité à ceux qui apparaissaient dans notre corpus[6].

Bien sûr, il y a eu certains cas où tous les équivalents candidats pour un terme anglais étaient valables et se trouvaient dans le corpus (par exemple, le terme anglais *remediation* et ses équivalents espagnols *saneamiento*, *rehabilitación*, *correctivo*, *recuperar*). Dans ces cas, nous avons fait appel à deux critères : a) catégorie grammaticale et b) fréquence. Le terme de base espagnol devait appartenir à la même catégorie grammaticale que le terme anglais (l'équivalent *recuperar* a été ainsi écarté). Puis, nous avons choisi l'équivalent candidat ayant le plu grand nombre d'occurrences dans les textes espagnols.

---

[6] Toutefois, nous avons trouvé un cas problématique. L'équivalent espagnol du verbe anglais *bioaccumulate*(2), à savoir *bioacumular*, n'apparaît pas dans le corpus. Malgré cela, nous l'avons retenu comme terme de base parce qu'il s'agit, sans aucun doute, de l'équivalent correct.

3.2 Analyse des termes simples

Une fois notre liste des termes de base espagnols constituée, nous avons procédé à l'analyse comparative. À l'aide du concordancier bilingue de *Logiterm* nous avons observé la façon dont chacun de nos termes simples anglais a été rendu dans les textes traduits. Le Tableau IX montre, pour chaque terme simple anglais, le terme de base espagnol et les différentes variantes observées. Dans la colonne « variantes » les insertions sont placées entre parenthèses tandis que les cas d'omission du terme dans la traduction sont représentés par des traits continus.

Tableau IX. Variantes terminologiques observées dans les termes simples

Terme anglais	Terme de base espagnol	Variantes	Nombre total de variantes
abatement(1)	reducción	reducir(las) combate eliminador limitar reducir	5
anthropogenic(1)	antropógeno	antropogénico	1
bioaccumulate(1)	bioacumularse	bioacumulable acumularse bioacumulación se (bioconcentra y) bioacumula	4
bioaccumulate(2)	bioacumular	acumular	1
bioconcentrate(1)	bioconcentrarse		0
carcinogen(1)	carcinógeno	riesgo de cáncer cancerígeno	2
chlor-alkali(1)	cloro y sosa cáustica	cloroalcalina sosa cáustica cloro	3
coal-fired(1)	carboeléctrica	a carbón a base de carbón carboeléctrica alimentado con carbón alimentado por	8

		carbón carbonífera a partir del carbón que utiliza el carbón	
compartment(1)	compartimento	compartimiento segmento elemento medioambiental sector zona específica	5
deposition(1)	deposición	sedimentación eliminación	2
disposal(1)	vertido	disposición eliminación --- respectivas	4
emission(1)	emisión	las de emisor	2
foundry(1)	fundidora	foundry fundería siderúrgica fundición empresa fundidora ---	6
incinerator(1)	incinerador	éstos incineradora --- IDMS de ellos	5
landfill(1)	relleno sanitario	Landfill terraplén de ellos dichos rellenos vertedero de desechos --- relleno confinamiento vertedero	9
leakage(1)	filtración	fuga lixiviado	2
monitoring(1)	monitoreo	existente Monitoring vigilancia	3
pathway(1)	trayectoria	vía ---	4

		forma vía de paso	
PCB(1)	BPC	estos compuestos --- estos residuos PCB	4
pesticide(1)	plaguicida	pesticide estos compuestos --- pesticida	4
recycling(1)	reciclado	--- se reciclan Recycling (se sigue) reciclando reciclar(los) reciclaje	6
release(1)	liberación	emisión descarga emanación Release las que	5
remediation(1)	saneamiento	Remediation rehabilitación --- recuperar correctivo recuperación	6
reservoir(1)	embalse	estanque depósito ---	3
sink(1)	depósito	sumidero colector	2
trace(1)	traza	residual mínima cantidades mínimas de oligo- de trazas en trazas	6

Total de variantes : 102

Comme le montre le Tableau IX, le terme présentant la variation la plus élevée est *relleno sanitario* (neuf variantes), tandis que *antropógeno* et *bioacumular*(2) sont les termes qui présentent une variation minimale (une seule variante).

À noter également qu'un seul terme reste invariable: *bioconcentrarse* ; le terme de base a toujours été utilisé dans les quatre fichiers dans lesquels il apparaît.

Un autre phénomène intéressant est le fait que, dans le corpus espagnol, il n'y a aucune occurrence du terme de base *bioacumular*, qui correspond au terme anglais *bioaccumulate*(2), une seule variante a été observée.

L'analyse comparative de termes a produit un total de 102 cas différents de variation terminologique qui peuvent être organisés dans six catégories : variation sémantique, changement de catégorie grammaticale, omission du terme dans la traduction, substitution du terme par une anaphore, occurrence du terme anglais dans la traduction et substitution du terme par une paraphrase.

3.2.1 Variation sémantique

Cette catégorie comprend deux sous-types de variation : a) l'utilisation d'un synonyme et b) l'utilisation d'un autre terme qui entraîne un changement de sens.

3.2.1.1 Utilisation d'un synonyme

Nous avons observé 47 cas d'utilisation d'un synonyme au lieu du terme de base en espagnol. 21 termes présentent cette variation par synonymie et nous avons

constaté la présence de deux, trois ou plus de synonymes pour le même terme de base. Le Tableau X contient des exemples de synonymes observés pour les termes simples.

Tableau X. Quelques exemples du premier cas de variation sémantique : utilisation d'un synonyme

Terme anglais	Terme de base espagnol	Synonymes observés
abatement(1)	reducción	combate
anthropogenic(1)	antropógeno	antropogénico
carcinogen(1)	carcinógeno	cancerígeno
compartment(1)	compartimento	compartimiento segmento sector zona específica
landfill(1)	relleno sanitario	terraplén vertedero de desechos relleno confinamiento vertedero
release(1)	liberación	emisión descarga emanación
remediation(1)	saneamiento	rehabilitación correctivo recuperación

3.2.1.2 Utilisation d'un autre terme qui entraîne un changement de sens

Huit des variantes sémantiques observées dans le corpus espagnol ne sont pas incluses dans la catégorie de synonymes parce qu'il y a une différence conceptuelle entre ces variantes et les termes de base correspondants. Le Tableau XI contient la liste des variantes sémantiques entraînant un changement de sens.

Tableau XI. Variantes sémantiques des termes simples entraînant un changement de sens

Terme anglais	Terme de base	Variante sémantique	Commentaires
bioaccumulate(1)	bioacumular	acumular	*Acumular* est un terme plus général que *bioacumular*.
bioaccumulate(2) (verbe transitif)	bioacumular	acumular	*Acumular* est un terme plus général que *bioacumular*.
chlor-alkali(1)	cloro y sosa cáustica	- cloro - sosa cáustica	Les deux variantes excluent l'une de deux substances chimiques auxquelles le terme de base et le terme anglais font référence.
deposition(1)	deposición	eliminación	La *deposición* (*dépôt*) d'un contaminant n'implique pas nécessairement son élimination.
disposal(1)	vertido	disposición	Il n'y a aucun rapport entre le sens du terme *disposición* (qui, dans le sous-domaine de la législation de l'environnement, fait référence à une aliénation, par exemple) et le terme *vertido* (qui fait référence à l'évacuation des déchets).
leakage(1)	filtración	lixiviado	Le terme *lixiviado* fait référence à un procédé chimique tandis que *filtración* designe plûtot un événement accidentel, non contrôlé.
sink(1)	depósito	sumidero	*Sumidero* est un hyponyme de *depósito*, car le premier fait référence seulement à certains types de réservoirs (*depósitos*) : ceux qui contiennent du carbone.

3.2.2 Variation syntaxique

3.2.2.1 Omission du terme dans la traduction

Le phénomène d'omission a été observé dans 10 termes. Dans les 12 différents cas trouvés, le terme anglais apparaît deux ou plusieurs fois dans le même paragraphe, voire dans la même phrase du texte original. Il semble que le traducteur

utilise ce recours stylistique pour éviter les répétitions non nécessaires et « alléger » la phrase. L'exemple (1) montre un cas où le terme *recycling* figure deux fois dans une phrase anglaise et une seule fois dans la phrase correspondante en espagnol.

(1) [EN] A CEC effort could determine whether the same concerns are applicable to **recycling** of products containing lead or **recycling** of lead.

[ES] La CCA podrá emprender una acción para determinar si las mismas consideraciones son aplicables al **reciclado** de productos con contenido de plomo o del propio plomo.

3.2.2.2 Substitution du terme par une anaphore

Dans 8 termes, nous avons observé 11 cas de substitution par une anaphore pronominale ou par la combinaison d'une anaphore et le terme abrégé ou un terme générique. Parmi les anaphores observées se trouvent des pronoms personnels, ainsi que des pronoms et adjectifs démonstratifs. Les phrases (2) et (3) sont des exemples de l'utilisation d'éléments anaphoriques. Les termes et les anaphores sont marqués en gras.

(2) [EN] The virtually universal distribution of **PCBs** throughout the world, including the arctic and other remote areas, suggests that **PCBs** are transported via the atmosphere and ocean currents (Ballschmiter and Wittlinger, 1991).

[ESP] La distribución prácticamente universal de los **PCB**, incluso en el ártico y otras regiones remotas, sugiere que **estos compuestos** se transportan por la atmósfera y las corrientes oceánicas (Ballschmiter y Wittlinger, 1991).

(3) [EN] While the use of **incinerators** in Region X for disposing of municipal waste is the exception rather than the rule, **incinerators** are widely used for the disposal of hospital wastes.

[ES] Si bien en la Región X el uso de **incineradores** para eliminación de desechos municipales constituye la excepción más que la regla, **éstos** se usan generalmente para eliminar desechos de hospitales.

3.2.2.3 Occurrence du terme anglais dans la traduction

Quant à ce type de variation, nous avons observé deux cas différents. Dans le premier cas, présent dans 7 termes, le terme simple fait partie d'un nom propre, du titre d'un norme, d'une loi, d'un programme ou organisme en particulier. Le traducteur traduit normalement le titre, mais il place après lui et entre parenthèses le titre original, en anglais. Il semble qu'il s'agisse d'un procédé auquel ont recours les traducteurs des textes formant le corpus d'étude. La phrase (4) constitue un exemple de cette variation.

(4) [EN] EPA has established a Regional Environmental **Monitoring** and Assessment Program (R-EMAP) project to assess trace elements on precipitation and aerosol samples.

[ES] El EPA ha establecido un proyecto de Programa regional de evaluación y **vigilancia** ambiental (Regional Environmental **Monitoring** Assessment Program, R-EMAP) para estimar los oligoelementos en muestras de aerosoles y precipitaciones.

Dans le deuxième cas, observé seulement dans le terme *BPC*, il s'agit d'une préférence pour l'utilisation du terme original dans les textes espagnols. En voici un exemple :

(5) [EN] Electrical transformers and capacitors are one such major source of **PCBs**.

[ES] Los transformadores y condensadores eléctricos son una de las fuentes más importantes de **PCB**.

3.2.3 Variation morphosyntaxique

3.2.3.1 Changement de catégorie grammaticale

Sur les 24 termes pour lesquels nous avons observé des variations, 8 présentent au moins un cas de changement de catégorie grammaticale par rapport au terme de base. Dans 7 cas, il s'agit d'une transformation d'un nom en verbe ; dans 3 cas, le nom est remplacé par un adjectif. Nous avons aussi repéré un cas de transformation verbe-nom, un cas de transformation nom-adjectif, un cas de transformation verbe-adjectif et un autre dans lequel le nom a été transformé en préfixe accompagnant un nom. Nous avons donc observé un total de 14 cas de changement de la catégorie grammaticale. En voici trois exemples marqués en gras :

Transformation nom - verbe

(6) [EN] R7 Recovery of components used for pollution **abatement**

[ES] R7 Recuperación de componentes utilizados para **reducir** la contaminación

Transformation nom - adjectif

(7) [EN] Industrial processes are considered the major source of lead **emissions** to the atmosphere, …

[ES] Se considera que los procesos industriales constituyen el principal foco **emisor** de plomo a la atmósfera, …

Transformation nom - préfixe

(8) [EN] Likewise, any mercury-containing compounds, together with any other **trace** elements that are left over during the refining stages of gold and other precious metals, …

[ES] Asimismo, todo compuesto que contenga mercurio, junto con otros **oligo**elementos que hayan quedado de las etapas de refinación del oro y otros metales preciosos, …

3.2.3.2 Transformation du terme original en une paraphrase

Nous avons observé 2 cas où le terme, en tant qu'unité lexicale figée, n'est pas conservé dans la traduction; à sa place, on trouve une phrase explicative plus longue, qui nous considérons comme étant une paraphrase. De toute évidence, il s'agit d'un phénomène de variation très particulier des textes traduits. Voici les deux cas des termes transformés en paraphrases trouvés dans le corpus :

(9) [EN] Additionally, the ICR will gather information on stack emissions of mercury for a segment of the **coal-fired** electric utility industry.

[ES] Además, la ICR reunirá información sobre las emisiones de chimenea de mercurio de un sector de la industria eléctrica **que utiliza el carbón**.

(10) [EN] (in which the lime raw material often contains **trace** mercury)

[ES] (en el que la cal, como materia prima, suele contener **cantidades mínimas de** mercurio)

En plus des types de variation que nous venons de décrire, nous avons observé deux autres phénomènes : a) la combinaison de deux types de variation et b) la transformation du terme simple en terme complexe.

L'exemple (11) présente une combinaison de variation par changement de catégorie grammaticale du verbe anglais *recycling* avec l'utilisation d'une anaphore pour le terme *metal*.

(11) [EN] Amongst these may be cited the complete dismantling of the transformer, with separation into similar **metals**, decontamination of these **metals**, and **recycling** as appropriate.

[ES] Se puede por ejemplo desguazar el transformador, separar sus partes por **metales** similares, descontaminarlos y **reciclarlos** como corresponda.

Par ailleurs, douze des variantes catégorisées comme des synonymes sont des transformations d'un terme simple en terme complexe. Il s'agit du deuxième phénomène accompagnant une variante. Voici quelques exemples de synonymes qui sont devenus des termes complexes :

Terme anglais	**Terme de base**	**Synonyme qui est un terme complexe**
carcinogen(1)	carcinógeno	riesgo de cáncer
coal-fired(1)	carbonífera	a base de carbón
		alimentado por carbón
		a partir del carbón
compartment(1)	compartimento	elemento medioambiental
pathway(1)	trayectoria	vía de paso

Nous considérons que ces deux phénomènes jouent un rôle important lors de l'identification ou l'extraction automatique de termes et, par conséquent, il faut toujours les prendre en considération.

Le Tableau XII est une récapitulation des variations observées dans les termes simples. La première colonne du tableau regroupe toutes les variations dans trois catégories : sémantique, syntaxique et morphosyntaxique. La deuxième colonne présente chacune des variantes observées. La troisième colonne indique le nombre de termes ayant subi chaque type de variation. La quatrième colonne donne, à titre illustratif, le pourcentage des termes ayant subi chaque type de variation, proportion estimée sur la base du nombre total de termes analysés (25). Par exemple, nous avons observé la variation par synonymie dans 21 termes : ainsi, 84 % des 25 termes sélectionnés ont subi cette variation. La cinquième colonne indique le nombre total de cas par type de variation observé dans notre corpus parallèle. Enfin, la sixième colonne contient la représentation en pourcentage du nombre de cas par

variation par rapport au nombre total de variations observées dans le corpus parallèle. Ainsi, les 47 cas de variation par synonymie représentent 46 % du total des 102 variantes observées dans les termes simples. Il faut remarquer que les valeurs de la troisième et de la quatrième colonnes différeront des celles de la cinquième et sixième colonnes étant donné qu'un même terme présente souvent divers types de variation et qu'un type de variation donné peut se présenter plus d'une fois dans le corpus.

Comme nous le verrons dans ce tableau, presque tous les termes simples ont subi au moins un cas de variation par synonymie. Presque la moitié des cas de variation sont reliés à l'utilisation des synonymes.

Les chiffres présentées dans le tableau montrent également que l'omission est un cas de variation ayant une fréquence élevée dans les termes simples, ce qui tend à montrer que le fait d'éviter des répétitions nécessaires est un recours important en traduction. Toutefois, les omissions peuvent causer des erreurs lors de la mise en correspondance de termes source et équivalents.

Il faut souligner aussi que plusieurs types de variation (substitution par une anaphore, occurrence du terme anglais dans la traduction et changement de la catégorie grammaticale) ont la même fréquence chez les termes simples. De plus, leur pourcentage reflète qu'il s'agit de phénomènes linguistiques importants en traduction.

Enfin, les chiffres correspondant aux variantes morphosyntaxiques démontrent que, dans le domaine de l'extraction bilingue, et plus particulièrement celle basée sur des corpus de traductions, il convient de prendre en considération ce

type de variation. Sinon, on risque de perdre bon nombre de variantes/termes équivalents.

Tableau XII. Types de variation observés dans les termes simples

Type de variation		Termes	% (25)	Nombre de cas	% (102)
1. Variation sémantique	a) synonyme	21	84	47	46
	b) changement de sens	6	24	8	7,8
2. Variation syntaxique	a) omission du terme	10	40	12	11,7
	b) substitution par une anaphore	8	32	11	10,7
	c) occurrence de terme anglais dans la traduction	8	32	8	7,8
3. Variation morphosyntaxique	a) changement de la catégorie grammaticale	8	32	14	13,7
	b) substitution par une paraphrase	2	8	2	1,9

3.3 Analyse des termes complexes

Après avoir défini la liste des termes de base en espagnol qui correspondent à chacun des termes complexes anglais (en fonction des critères décrits à la section 3.1), nous avons réalisé l'analyse comparative de ces termes et nous avons ainsi obtenu une longue liste de variantes terminologiques.

Le Tableau XIII montre les 25 termes complexes anglais sélectionnés, leur terme de base en espagnol, la liste de différentes variantes observées et le nombre total de variantes par terme.

Tableau XIII. Variantes terminologiques observées dans les termes complexes

Terme anglais	Terme de base espagnol	Variantes	Nombre de variantes
action plan(1)	plan de acción	------ - plan (regional) de acción - PAR	3
adverse effect(1)	efecto adverso	efecto nocivo	1
ambient air concentration(1)	concentración en el aire ambiente	- concentración en el aire ambiental - concentración (común) en el aire - concentración (promedio del plomo) en el aire - concentración (de plomo) en el aire ambiente - concentración (de plomo) en el aire	5
body burden(1)	carga corporal	- carga (de plomo) en el cuerpo humano - cantidad (de plomo) en el cuerpo	2
daily intake(1)	ingesta diaria	- ingesta - la de - ingestión diaria (de plomo) - ingestión diaria (de dioxina) - ingesta diaria (de pesticidas)	5
dangerous goods(1)	material peligroso	- bien peligroso - sustancia peligrosa - mercadería peligrosa - mercancía peligrosa	4
environmental fate(1)	destino ambiental	- destino en el medio ambiente - destino (y transporte) ambiental - (distribución ambiental y) destino - (transporte,) destino final (y transferencia) en el medio ambiente - (distribución y) destino ambiental	5
environmental media(1)	elemento medioambiental	- medio ambiental - entorno ambiental - entorno - medio ambiente - medio	5
flue gas(1)	gas de escape	- gas de combustión ---------- - gas de chimenea - gas de salida	4

hazardous air pollutant(1)	contaminante atmosférico peligroso	- Hazardous Air Pollutant - contaminante peligroso del aire	2
hazardous waste(1)	desechos peligrosos	- residous peligrosos --- - residuos (no) peligrosos - residuos (industriales no) peligrosos - Hazardous Waste - residuo peligroso - residuos tóxicos - al respecto	8
industrial chemical(1)	producto químico industrial	- productos (y subproductos) químicos industriales - sustancias químicas industriales - químicos industriales	3
life cycle(1)	ciclo de vida	- dicho ciclo - ciclo vital	2
long-range transport(1)	transporte a larga distancia	- transporte de grandes distancias - transporte a grandes distancias - transportarse a grandes distancias - transporte de largo alcance - transporte (del mercurio) a largas distancias	5
municipal solid waste(1)	desechos sólidos municipales	- desechos sólidos urbanos - residuos sólidos municipales - desechos municipales sólidos - (incineradores) municipales (de) desechos sólidos - desechos urbanos sólidos	5
persistent organic pollutant(1)	contaminante orgánico persistente		1
persistent toxic substance(1)	sustancia tóxica persistente	Persistent Toxic Substance	1
point source(1)	fuente puntual	- fuente fija - fuentes fijas (y móviles) - fuente (atmosférica) puntual - fuente puntual (y difusa) - fuente puntual (o de área) - fuente (no) puntual - fuente potencial - foco (emisor de plomo) - fuente - Non Point Source - fuente puntual (o no puntual) - fuente (emisora de plomo)	12
risk assessment(1)	evaluación de riesgos	- evaluación de riesgo - evaluación - evaluación (y comunicación) de riesgos	9

		- indicadores de riesgo - evaluar (mejor) los riesgos - evaluacion del riesgo - evaluación (y manejo) de riesgos - estudio de riesgo - evaluar el riesgo	
sewage sludge(1)	lodos de depuración	- lodos de alcantarilla - fangos cloacales - fango cloacal	3
sinter plant(1)	planta de sinterización	- estas instalaciones - planta sinterizadora - instalación de sinterización - planta	4
sound management(1)	manejo racional	- manejo (ambientalmente) racional - manejo (ambiental) racional - manejo (eficiente y ambientalmente) racional - manejo (ambiental y económicamente) racional - manejo adecuado - manejo - gestión racional - gestión (ecologicamente) racional - gestión (ambientalmente) racional	9
transboundary movement(1)	movimiento transfronterizo	envío transfronterizo	1
vector control(1)	control de vectores	- control del vector (del paludismo) - control de los vectores - control (más integral de la enfermedad y sus) vectores - control (biológico) de los vectores - luchar contra los vectores	5
waste management(1)	gestión de desechos	- gestión (sin riesgo) de desechos - gestión (ambientalmente racional) de desechos - manejo (general) de residuos (peligrosos) - manejo de desechos - manejo de los desechos - manejo de los residuos - manejo de residuos - Waste Management	8

Total de variantes : 112

Comme le Tableau XIII le montre, tous les termes complexes donnent lieu à des variantes.

Les termes de base espagnols *efecto advers*, *contaminante orgánico persistente, sustancia tóxica persistente* et *movimiento transfronterizo* présentent une variation minimale (seulement une variante pour chacun), suivis des termes *carga corporal*, *contaminante atmosférico peligroso* et *ciclo de vida*, qui présentent deux variantes.

Le terme *fuente puntual* possède la variation la plus élevée (douze variantes), suivi des termes *evaluación de riesgos* et *manejo racional*, avec neuf variantes chacun.

L'analyse comparative a produit un total de 112 différents cas de variation sémantique, syntaxique et morphosyntaxique que nous présentons dans ce qui suit.

3.3.1 Variation sémantique

Nous avons détecté trois types de variantes sémantiques dans les termes complexes sélectionnés : a) des synonymes, b) des quasi-synonymes qui introduisent une certaine ambiguïté, et c) des termes qui entraînent un changement de sens.

3.3.1.1 Synonymie

Nous avons observé 41 cas d'utilisation de synonymes pour 18 termes ; certains termes étant remplacés par plus d'un synonyme. Le tableau XIV présente des exemples de synonymes relevés dans le corpus.

Tableau XIV. Exemples de variation par synonymie

Terme anglais	Terme de base	Synonymes
dangerous goods	material peligroso	mercadería peligrosa bien peligroso mercancía peligrosa
environmental fate	destino ambiental	destino en el medio ambiente
long-range transport	transporte a larga distancia	transporte de largo alcance transporte de grandes distancias
municipal solid waste	desechos sólidos municipales	residuos sólidos municipales desechos municipales sólidos desechos sólidos urbanos desechos urbanos sólidos
sewage sludge	lodos de depuración	fangos cloacales fango cloacal

Comme on peut l'observer dans le tableau précédent, les synonymes sont assez souvent des syntagmes très similaires qui se distinguent seulement par l'utilisation d'une certaine préposition ou d'un adjectif ou adverbe simple au lieu d'un syntagme adjectival ou prépositionnel. Ce sont des cas de variation des éléments modificateurs du terme. Toutefois, il y a aussi des cas de synonymes complexes dont la tête varie. Enfin, dans d'autres cas, il s'agit tout simplement d'un changement dans la structure du syntagme.

3.3.1.2 Quasi-synonymes introduisant des ambiguïtés

Dans l'analyse de 3 termes complexes, nous avons trouvé 4 cas où l'utilisation d'un quasi-synonyme entraîne une ambiguïté. Nous présentons dans le Tableau XV la liste des cas ambigus. Les composantes syntagmatiques causant l'ambiguïté sont marquées en gras.

Tableau XV. Quasi-synonymes introduisant des ambiguïtés

Terme anglais	Terme de base	Quasi-synonymes ambigus	Cause de l'ambiguïté
environmental media	elemento medioambiental	**medio ambiente**	*Medio ambiente* peut faire référence à l'environnement en général ou à un segment de l'environnement (air, eau, etc.)
life cycle	ciclo de vida	ciclo **vital**	L'adjectif *vital* peut être interprété comme « faisant référence à la vie » ou bien comme « très important »
risk assessment	evaluación de riesgos	evaluación **de riesgo** estudio **de riesgo**	***De riesgo*** peut être interprété comme « une évaluation qui implique des risques/risquée » et non comme « une évaluation des risques », comme c'est le cas ici.

3.3.1.3 Utilisation d'un terme entraînant un changement de sens

Dans les termes complexes nous avons observé 3 cas seulement de variation avec changement de sens par rapport aux termes anglais et termes espagnols de base. Le Tableau XVI présente les cas relevés.

Tableau XVI. Variantes sémantiques entraînant un changement de sens

Terme anglais	Terme de base	Variante comportant un changement de sens	Explication
point source	fuente puntual	fuente fija	Une *fuente puntual* est une source de pollution bien identifiée/précise, tandis que *fuente fija* fait référence à une source non mobile.
point source	fuente puntual	fuente potencial	Une *fuente potencial* es une source susceptible de contaminer, mais pas une source bien identifiée.
hazardous waste	desechos peligrosos	residuos tóxicos	Les *residuos peligrosos* ne sont pas forcement toxiques

3.3.2 Variation syntaxique

Nous avons observé différents cas de variation syntaxique : par insertion, par coordination, par omission, par abréviation et, enfin, par occurrence du terme anglais dans le corpus espagnol.

3.3.2.1 Variation par insertion

Au cours de l'analyse comparative, nous avons constaté que la variation par insertion est un type de variation très fréquent dans les termes complexes. Nous avons observé un total de 41 cas de variation par insertion pour 14 termes complexes. De plus, cette variante prend des formes divers : nous avons trouvé toutes sortes de mots insérés à gauche, à droite ou au milieu des termes.

Dans 16 cas, les insertions donnent lieu à la création d'un autre terme plus spécifique, par exemple un hyponyme. Il s'agit de l'insertion causant une substitution. Le Tableau XVII montre quelques exemples de ce type d'insertion. Dans tous les exemples de cette section, les insertions au terme espagnol sont placées entre parenthèses et les mots anglais qui produisent cette variation sont marqués en gras.

Tableau XVII. Exemples de variation par insertion générant un terme plus spécifique

Terme anglais	Variante par insertion – terme plus spécifique	Insertions aux termes anglais
action plan	plan (regional) de acción	**regional** action plan
daily intake	ingestión diaria (de plomo) ingesta diaria (de pesticidas)	daily **dietary lead** intake daily **pesticide** intake
point source	fuente (atmosférica) puntual	**air** point source
vector control	control del vector (del paludismo) control (biológico) de los vectores	**malaria** vector control **biological** vector control

Dans 8 cas, les mots insérés sont des adjectifs et des adverbes qui modifient le terme en question. Il s'agit alors d'une insertion causant une « modification ». Le Tableau XVIII contient quelques exemples de ce type d'insertion.

Tableau XVIII. Insertion d'adjectifs ou d'adverbes modifiant le terme complexe

Terme anglais	Variante par insertion – adjectifs, adverbes	Occurrence du terme en anglais
ambient air concentration	concentración (común) en el aire	**typical** ambient air concentration
sound management	manejo (ambiental y económicamente) racional gestión (ecológicamente) racional	**economically and environmentally** sound management **environmentally** sound management

3.3.2.2 Variation par coordination

Dans 11 cas, les insertions au terme complexe sont directement reliées à la combinaison de deux termes, lorsque la tête ou le modificateur sont partagés par ceux-ci. Le Tableau XIX en donne trois exemples.

Tableau XIX. Insertion due à la combinaison de deux ou de plusieurs termes anglais

Terme anglais	Occurrence du terme en anglais	Variante par insertion – combinaison de termes
environmental fate(1)	environmental fate **and transport** environmental transport, fate and transportation	destino (y transporte) ambiental (transporte), destino final (y transferencia) en el medio ambiente
point source(1)	point **or area** source	fuente puntual (o de área)

De même, la variation par coordination peut relier le terme complexe et son antonyme. Dans notre corpus, nous avons observé 4 cas de ce type de coordination et ils figurent dans le Tableau XX.

Tableau XX. Insertion qui produit la combinaison du terme et son antonyme

Terme anglais	Variante par insertion – terme et son antonyme	Occurrence du terme en anglais
point source	fuente **(no)** puntual fuente puntual **(o no puntual)**	**non** point source point **or non** point source
hazardous waste	residuos **(no)** peligrosos residuos (industriales **no**) peligrosos	**non**-hazardous waste **non**-hazardous industrial waste

Bien sûr, il y a des cas où l'on trouve deux types différents d'insertion pour un même terme complexe (par exemple, l'insertion d'un adjectif au milieu et une insertion dérivée de la combinaison de deux termes).

De plus, les divers types d'insertion peuvent s'effectuer sur le terme de base ou sur une autre variante du terme, comme un synonyme, qui est le cas le plus fréquent dans notre corpus.

Il est important de signaler que, dans la plupart des cas, les insertions apparaissent aussi dans les termes anglais. En anglais, les éléments insérés apparaissent normalement à gauche des termes alors qu'en espagnol ces éléments apparaîtront très souvent au milieu des termes. Par conséquent, lors de l'observation des termes complexes anglais (dans l'environnement du concordancier de *Logiterm*) ces éléments sont difficiles à localiser. Toutefois, dans l'analyse des termes espagnols, il serait impossible de ne pas les prendre en considération.

De même, la variation par coordination peut relier le terme complexe et son antonyme. Dans notre corpus, nous avons observé 4 cas de ce type de coordination et ils figurent dans le Tableau XX.

Tableau XX. Insertion qui produit la combinaison du terme et son antonyme

Terme anglais	Variante par insertion – terme et son antonyme	Occurrence du terme en anglais
point source	fuente **(no)** puntual fuente puntual **(o no puntual)**	**non** point source point **or non** point source
hazardous waste	residuos **(no)** peligrosos residuos (industriales **no**) peligrosos	**non**-hazardous waste **non**-hazardous industrial waste

Bien sûr, il y a des cas où l'on trouve deux types différents d'insertion pour un même terme complexe (par exemple, l'insertion d'un adjectif au milieu et une insertion dérivée de la combinaison de deux termes).

De plus, les divers types d'insertion peuvent s'effectuer sur le terme de base ou sur une autre variante du terme, comme un synonyme, qui est le cas le plus fréquent dans notre corpus.

Il est important de signaler que, dans la plupart des cas, les insertions apparaissent aussi dans les termes anglais. En anglais, les éléments insérés apparaissent normalement à gauche des termes alors qu'en espagnol ces éléments apparaîtront très souvent au milieu des termes. Par conséquent, lors de l'observation des termes complexes anglais (dans l'environnement du concordancier de *Logiterm*) ces éléments sont difficiles à localiser. Toutefois, dans l'analyse des termes espagnols, il serait impossible de ne pas les prendre en considération.

3.3.2.3 Variation par omission

Nous avons observé trois types d'omission dans notre corpus.

a) Omission totale du terme dans le corpus espagnol

Comme nous l'avons observé dans les termes simples, si le terme complexe apparaît deux ou plusieurs fois dans le même paragraphe, voire dans la même phrase du texte original, il peut être omis dans la traduction à partir de sa deuxième ou de sa troisième occurrence. Nous supposons que le traducteur utilise ce recours stylistique pour éviter les répétitions et « alléger » la phrase. Seulement 4 termes complexes ont présenté la variation par omission totale. La phrase (12) en est un exemple:

(12) [EN] (c) Ensure that persons involved in the management of **hazardous wastes** or other wastes within it take such steps as are necessary to prevent pollution due to **hazardous wastes** and other wastes arising from such management…

[ES] c) Velar porque las personas que participan en el manejo de los **desechos peligrosos** y otros desechos dentro de ella se adopten las medidas necesarias para impedir que ese manejo dé lugar a una contaminación…

b) Occurrence du terme abrégé

Comme dans le cas précédent, lorsqu'un terme apparaît plus d'une fois dans la même phrase ou le même paragraphe, il est possible que le terme en question soit présent dans une forme abrégée à partir de sa deuxième occurrence. Dans notre corpus, 5 termes ont présenté cette variation. Ils sont reproduits dans le Tableau XXI.

Tableau XXI. Variation par abréviation du terme complexe

Terme anglais	Terme de base	Terme abrégé
daily intake	ingesta diaria	ingesta
point source	fuente puntual	fuente
risk assessment	evaluación de riesgos	evaluación
sinter plant	planta de sinterización	planta
sound management	manejo racional	manejo

c) Transformation du terme en acronyme

Nous avons observé un cas spécial de variation dans les termes complexes : le terme *action plan* est inclus dans un acronyme (PAR, par « Plan de Acción Regional »). Ce cas de variation a une fréquence très basse dans notre corpus (2 occurrences), mais il s'agit peut-être d'une variation plus importante dans d'autres corpus spécialisés ou pour d'autres termes complexes non étudiés.

3.3.2.4 Occurrence du terme en anglais dans la traduction

Nous avons observé 5 cas dans 5 termes complexes où le terme anglais figure dans le corpus espagnol. Dans tous les cas, le terme fait partie d'un nom propre, du titre d'une norme, d'une loi, d'un programme ou d'un organisme en particulier. Le traducteur traduit normalement le titre, mais il place entre parenthèses le titre original, en anglais. Il semble qu'il s'agisse d'un procédé courant utilisé par les traducteurs des textes formant le corpus d'étude. La phrase (13) est un exemple de cette variation.

(13) [EN] **Hazardous Waste** Combustion Facilities Rule: In September 1999, EPA promulgated air emission standards for a number of chemicals…

[ES] Normativa para plantas de combustión de **desechos peligrosos (Hazardous Waste** Combustion Facilities Rule): En septiembre de 1999 el EPA promulgó normas de emisiones atmosféricas para varias sustancias químicas…

3.3.2.5 Substitution du terme par une anaphore

Contrairement à ce que nous supposions au début de l'analyse, nous avons observé seulement 2 cas de remplacement du terme par une anaphore et trois cas de combinaison d'une anaphore et du terme abrégé. Le Tableau XXII présente la liste des cas d'utilisation d'une anaphore.

Tableau XXII. Substitution du terme complexe par une anaphore

Terme anglais	Terme de base	Anaphore ou combinaison anaphore + terme abrégé
daily intake	ingesta diaria	la de
hazardous waste	desechos peligrosos	- al respecto - tales desechos
life cycle	ciclo de vida	dicho ciclo
sinter plant	planta de sinterización	estas instalaciones

3.3.3 Variation morphosyntaxique

3.3.3.1 Transformation du terme en syntagme verbal

Nous avons trouvé 4 cas où le terme disparaît dans la traduction et on trouve à sa place un syntagme verbal qui véhicule le sens spécialisé du terme complexe. Apparemment, c'est un autre recours stylistique que les traducteurs utilisent lorsqu'ils décident de ne pas reproduire la structure de la phrase originale qui contient le terme complexe en question. Les phrases (14) et (15) illustrent cette transformation du terme en syntagme verbal.

(14) [EN] Also, countries with a longer tradition of environmental management of mercury have expressed the need to continue to expand their knowledge base on mercury to improve **risk assessment** and ensure effective risk management.

[ES] Además, los países que tienen más tradición en la gestión ambiental del mercurio han manifestado la necesidad de seguir ampliando su base de información sobre el mismo para poder **evaluar** mejor **los riesgos** y garantizar la gestión efectiva de los mismos.

(15) [EN] It is capable of undergoing **long-range transport** due to its relative volatility (VPL = 4.76 Pa; H = 52 Pa m 3 /mol).

[ES] Debido a su volatilidad (VP[GP4]L = 4.76 Pa, H = 52 Pa m 3 /mol), el mirex puede **transportarse a grandes distancias**.

Tous les types de variantes observées dans les termes complexes sont présentés dans le Tableau XXIII. De la même façon que dans le Tableau XII sur les termes simples, la première colonne du tableau XXIII regroupe toutes les variations dans trois catégories : sémantique, syntaxique et morphosyntaxique. La deuxième colonne présente chacune des variantes observées. La troisième colonne indique le nombre de termes ayant subi chaque type de variation. La quatrième colonne donne, à titre illustratif, le pourcentage des termes ayant subi chaque type de variation, proportion estimée sur la base du nombre total de termes analysés (25). Par exemple, nous avons observé la variation par synonymie dans 18 termes : ainsi, 72 % des 25 termes sélectionnés ont subi cette variation. La cinquième colonne indique le nombre total de cas par type de variation observé dans notre corpus parallèle. Enfin, la sixième colonne contient la représentation en pourcentage du nombre de cas par variation par rapport au nombre total de variations observées dans le corpus parallèle. Ainsi, les 41 cas de variation par synonymie représentent 36,6 % du total des 112 variantes observées dans les termes complexes. Il faut remarquer que les valeurs de la troisième et de la quatrième colonnes différeront des celles de la cinquième et sixième colonnes étant donné qu'un même terme présente souvent

divers types de variation et qu'un type de variation donné peut se présenter plus d'une fois dans le corpus.

Tableau XXIII. Types de variation observés dans les termes complexes

Type de variation		Termes	% (25)	Nombre de cas	% (112)
1. Variation sémantique	a) synonyme	18	72	41	36,6
	b) quasi-synonyme	3	12	4	3,5
	b) changement de sens	2	8	3	1,7
2. Variation syntaxique	a) insertion	14	56	41	36,6
	b) coordination	7	28	11	9,8
	c) omission	8	32	10	8,9
	d) occurrence du terme anglais	5	20	5	4,4
	e) substitution par une anaphore	4	16	5	4,4
3. Variation morphosyntaxique	a) transformation en syntagme verbal	3	12	4	3,5

Comme nous pouvons le voir dans le tableau précédent, la substitution par un synonyme est le type de variation le plus fréquent dans les termes complexes. Un total de 18 termes a subi cette variation. De plus, pour un même terme complexe, nous avons trouvé jusqu'à quatre synonymes. Les variations sémantiques impliquant un changement de sens sont beaucoup moins fréquentes mais elles nous incitent à réfléchir davantage sur la variation sémantique en général et sur son incidence sur l'identification de termes équivalents en traduction.

La variation par insertion est aussi importante, puisque plus de la moitié des termes ont fait l'objet d'une insertion.

D'ailleurs, même si leurs proportions sont moins élevées, les variantes par omission, par substitution par une anaphore, par occurrence du terme anglais et par

transformation en syntagme verbal sont des phénomènes importants dans le domaine de la traduction et, par conséquent, en extraction automatique bilingue. D'après ce que nous avons vu dans le chapitre 1, ce sont de types de variation non repérés par les analyses unilingues.

3.4 Comparaison des résultats des termes simples et des termes complexes

Dans les sections précédentes du présent chapitre, nous avons décrit tous les cas de variation observés dans notre corpus parallèle anglais-espagnol pour 25 termes simples et 25 termes complexes. Les résultats de l'analyse montrent que tous les types de termes (qu'il s'agisse de termes simples, de termes complexes de différentes longueurs, des noms, d'adjectifs ou de verbes) sont susceptibles de subir une variation.

Le Tableau XXIV montre que le taux de variation des termes simples est très similaire à celui des termes complexes. Cela montre qu'en principe, il n'y a pas de raison pour écarter les termes simples dans les études sur la variation terminologique, qu'elles soient unilingues, bilingues ou multilingues.

Tableau XXIV. Taux de variation des termes simples et complexes

	Termes simples	Termes complexes
Nombre de termes ayant subi de la variation	**25**	**25**
Nombre total de variantes	**102**	**112**
Taux de variation	**4,08**	**4,44**

Comparés aux taux de variation produits par l'étude de Freixa (2002) (entre 1,58 et 2,02, selon le niveau de spécialisation des textes), et présentés à la section 1.2.1.3, les taux de variation de notre étude s'avèrent nettement plus élevés.

Il est difficile de déterminer avec précision les raisons de cette différence, mais nous considérons que trois facteurs ont joué un rôle important. Premièrement, l'étude de Freixa a été réalisée dans un corpus unilingue; le nôtre s'est basé sur un corpus bilingue. Deuxièmement, Freixa prend comme base de son analyse un ensemble de concepts; nous, au contraire, avons établi d'abord une liste de termes anglais et une liste de termes de base espagnols. Cela affecte le nombre de variantes produites. Dernièrement, le nombre de concepts analysés par Freixa est beaucoup plus élevé que le nombre de termes que nous avons sélectionnés. Il conviendrait peut-être de rapprocher ces deux études afin d'évaluer l'évolution des résultats.

En ce qui concerne les types de variation, nous avons observé que, d'une façon générale, les termes simples et les termes complexes peuvent être sujets aux mêmes types de variation sémantique (utilisation d'un synonyme ou d'un quasi-synonyme et utilisation d'un terme entraînant un changement de sens) et syntaxique (substitution par une anaphore, omission et occurrence du terme anglais dans la traduction). Bien sûr, il existe certains cas de variation syntaxique qui peuvent se produire seulement à partir des termes complexes, tels que l'insertion et la coordination. En ce qui concerne la variation morphosyntaxique, dans les termes simples, nous avons observé des cas de changement de la catégorie grammaticale et de substitution par une paraphrase, tandis que, dans les termes complexes, nous avons observé seulement des cas où certains noms ont été transformés en syntagmes verbaux.

Quant à la fréquence par type de variation, nous remarquons que dans les termes simples et les termes complexes, la variation par synonymie est le cas le plus fréquent ; pour 84 % des termes simples et 72 % des termes complexes, nous avons relevé au moins un synonyme. Dans le cas des termes simples, la variation par changement de catégorie grammaticale (représentant 13,7 % du total des variantes) est aussi très fréquente. En troisième position, par ordre de fréquence, nous observons les cas d'omission et de substitution par une anaphore (10,7 % pour chacun). Les cas d'occurrence du terme anglais dans la traduction ont aussi une fréquence importante ; un tiers des termes ont subi cette variation.

Dans les termes complexes, la variation par insertion a une fréquence élevée (36,6 % du total de variantes) ce qui confirme le caractère dynamique des termes. Les autres types de variation syntaxique ont une fréquence variée mais, d'après nous, leurs pourcentages sont aussi importants (variation par coordination : 9,8 % du total des variantes ; variation par omission : 8,9 %; substitution par une anaphore : 4,4 % et occurrence du terme anglais dans la traduction : 4,4 %).

Grâce à l'analyse comparative de la variation, nous avons repéré certains types de variation qui n'ont pas été considérés dans la typologie présentée à la section 1.1.2. En effet, les variations syntaxiques par substitution par une anaphore et par occurrence du terme anglais dans la traduction, ainsi que les variations morphosyntaxiques par changement de catégorie grammaticale et par transformation en syntagme verbal ne figurent pas dans cette typologie. La raison est simple : il est très difficile de relever ces variantes terminologiques dans les études de corpus unilingues.

Ainsi, les taux de variation pour les termes simples et les termes complexes (4,08 et 4,44, respectivement) mettent en évidence le fait que la variation joue un

rôle important dans l'extraction de termes bilingue. La variation peut rendre difficile la mise en correspondance entre les termes source et les termes cible. Si un terme source risque d'être représenté sous plusieurs formes en langue cible, est-il possible de l'identifier ? Dans la section qui suit, nous analysons quelles sont les variantes qui peuvent être reconnues et lesquelles posent des difficultés pour les extracteurs automatiques bilingues. Cette analyse se basera, bien sûr, sur les résultats de notre étude.

3.5 Incidence des résultats de l'analyse sur l'extraction de termes bilingue

D'après ce que nous avons présenté à la section 1.3.2 portant sur les extracteurs bilingues, très peu de modèles prennent en considération le fait que l'équivalent d'un terme source est susceptible de subir des variations dans la langue cible, c'est-à-dire, qu'il puisse être représenté de différentes façons. Le phénomène linguistique de la variation pose alors des problèmes au niveau de l'identification d'équivalents en langue cible ainsi qu'au niveau de la mise en correspondance entre un terme source et un terme cible. Analysons les implications de la variation terminologique par modèle d'extraction bilingue.

Il est important de souligner que cette section présente une réflexion qui s'appuie sur nos connaissances des méthodes d'extraction de termes bilingue, connaissances acquises par des lectures portant sur le sujet. Les analyses présentées ci-dessous n'ont pas fait l'objet d'une expérimentation.

Tout d'abord, il faut souligner que la variation sémantique comme telle n'est pas considérée par les extracteurs bilingues. Jusqu'à présent, aucun extracteur ne réalise une analyse conceptuelle pour vérifier que l'équivalent proposé pour un

terme source n'entraîne pas un changement de sens. Ce sera le spécialiste humain qui accomplira cette tâche. Cependant, nous considérons que la variation par synonymie ne représente pas vraiment un problème du point de vue de l'identification des candidats-termes et de la mise en correspondance des termes et leurs équivalents, à condition que le synonyme ait la même structure syntaxique et qu'il appartienne à la même catégorie grammaticale que le terme source. Nous ferons, donc, référence à la variation par synonymie seulement si cette condition n'est pas respectée.

Les extracteurs bilingues utilisant le modèle qui consiste à extraire, de façon indépendante, une liste de candidats-termes par langue pour ensuite mettre en correspondance les termes de deux listes (point a) de la section 1.3.2) visent généralement les termes complexes binaires de nature nominale. Donc, lors de l'identification des candidats-termes, ces systèmes écartent :

1) les termes simples;
2) les termes source appartenant à une catégorie grammaticale autre que le nom;
3) les termes complexes composés de plus de deux unités lexicales.

D'ailleurs, les techniques statistiques d'alignement utilisées par ce type d'extracteurs réussissent difficilement à mettre en correspondance les termes source ayant subi ces variations en langue cible :

a) Les insertions et les coordinations. Ces variantes impliquent un changement de la longueur de la structure des termes complexes, ce qui peut causer un décalage lors de l'alignement mot à mot.

b) Les changements de catégorie grammaticale. Lorsqu'un changement de catégorie grammaticale se produit, la position du terme cible dans la phrase peut aussi changer. Il sera donc difficile pour l'extracteur de déterminer où se trouve l'équivalent en langue cible.

c) Les transformations des termes source de nature nominale en paraphrases ou en syntagmes verbaux. La longueur des structures des termes source sera différente de celle des équivalents en langue cible. De plus, la position des équivalents peut varier. Par conséquent, l'extracteur sera incapable de localiser le segment en langue cible qui correspond au terme source.

d) Les variations par omission (ce qui inclut les abréviations) et par substitution par une anaphore. En raison de l'absence d'un terme équivalent en langue cible, il est fort probable qu'un alignement incorrect soit produit. Le terme source sera aligné aux unités lexicales se trouvant à la place du terme cible.

e) Les synonymes qui n'ont pas la même structure syntaxique que le terme source. Comme dans les cas d'insertion et de coordination, la différence de longueur entre le terme source et le terme cible pose des difficultés d'alignement mot à mot.

En considérant la fréquence de chacune de ces variantes dans notre analyse, et surtout le nombre de termes subissant chaque type de variation (voir les Tableaux XXII et XXIII), il est évident que le modèle d'extraction visant les termes binaires de nature nominale rate une proportion très élevée de termes ainsi que presque toutes leurs variantes.

Pour pallier cette limite, certains extracteurs se basant sur ce modèle d'extraction font appel à la technique des patrons syntaxiques typiques pour modéliser certaines variations syntaxiques des termes complexes binaires. Grâce à cette technique linguistique, les variantes par insertion, par coordination et par permutation correspondant aux patrons définis, seront alors identifiées et mises en correspondance avec les termes source.

Un système extrayant d'abord une liste de candidats-termes dans la langue source et qui, ensuite, retrouve la séquence de traduction contenant les équivalents (point b) de la section 1.3.2) vise très souvent les termes complexes de longueurs diverses et de nature nominale. Cependant, grâce aux méthodes probabilistes de traduction appliquées par ces extracteurs, les termes simples peuvent aussi être considérés. Alors, ces systèmes écartent seulement :

1) les termes source appartenant à une catégorie grammaticale autre que le nom.

Les extracteurs s'appuyant sur les modèles probabilistes de traduction (qui n'exigent pas des alignements exacts, mot à mot) pour mettre en correspondance les termes source et les termes pourraient ainsi identifier les variantes :

- ✓ Tous les synonymes, même si la longueur de leurs structures syntaxiques diffère de celles des termes sources.

- ✓ Certaines insertions. Si un seul élément a été inséré dans le terme source, il est probable que la variante dérivée soit incluse dans la séquence de traduction correspondante.

- ✓ Les omissions, les abréviations et les anaphores. Puisque, ni les techniques linguistiques, ni les techniques statistiques utilisées dans ce modèle d'extraction n'exigent un alignement parfait des mots, ces cas de variation deviendront évidents; ils apparaîtront dans la séquence de traduction correspondante.

Toutefois, plusieurs autres variantes poseront encore des difficultés pour ce deuxième type d'extracteurs :

a) Les coordinations. Les syntagmes terminologiques dérivés d'une coordination des termes risquent d'être mal découpés, surtout si le phénomène de coordination n'a pas été considéré lors de l'extraction des candidats-termes en langue source.

b) L'occurrence du terme anglais dans la traduction. S'il s'agit du cas récurrent que nous avons observé dans notre corpus parallèle (le terme anglais apparaissant à coté, souvent entre parenthèses, de son équivalent espagnol dans la traduction), il est possible que l'extracteur ne tienne pas compte de cette variante lors de la mise en correspondance des termes. En raison de sa position dans la phrase, le terme anglais ne fera pas partie de la séquence de traduction correspondant au terme source.

c) Les changements de catégorie grammaticale. Comme nous l'avons déjà mentionné, lorsqu'un changement de catégorie grammaticale se produit, la position du terme cible dans la phrase peut aussi changer. Il sera donc difficile pour l'extracteur de déterminer où se trouve l'équivalent en langue cible.

d) Les transformations des termes source de nature nominale en une paraphrase ou en syntagme verbal. La longueur des structures des termes source sera très différente de celle des équivalents en langue cible. De plus, la position des équivalents peut aussi varier. Par conséquent, l'extracteur sera incapable de déterminer la séquence de traduction dans laquelle l'équivalent du terme source apparaît.

Enfin, les extracteurs de termes qui partent de l'alignement mot à mot du corpus parallèle pour ensuite extraire simultanément les candidats-termes source et cible (point c) de la section 1.3.2) et qui, en plus, s'appuient plutôt sur des techniques linguistiques peuvent extraire, en principe, tous les types de termes : simples, complexes, de longueur variée et appartenant à diverses catégories grammaticales.

De même, les techniques linguistiques comme celle des relations de dépendance syntaxique permettent d'identifier plusieurs cas de variation :

- ✓ Les synonymes, même si les structures syntaxiques des termes source et cible diffèrent;
- ✓ Les insertions et les coordinations;
- ✓ Les changements de catégorie grammaticale;
- ✓ Les abréviations.

Par contre, d'autres cas de variation entraîneront des difficultés lors de la mise en correspondance des termes source et cible :

a) Les anaphores;
b) Les omissions;
c) L'occurrence du terme anglais dans la traduction;

d) La transformation des termes en paraphrases ou en syntagmes de nature non nominale, par exemple, des syntagmes verbaux.

Dans les Tableaux XXV et XXVI, nous résumons les types de termes visés et les variations considérées par chaque modèle d'extraction bilingue. Les résultats de l'évaluation générale de l'incidence de la variation terminologique que nous venons de présenter mettent en évidence le fait que les extracteurs bilingues ne peuvent pas reconnaître l'ensemble des termes et de leurs variantes.

Tableau XXV. Types de termes visés par modèle d'extraction de termes bilingue

Type de termes	EB-A	EB-A, PST	EU-T	A-EB
Simples			●	●
Complexes binaires	●	●	●	●
Complexes de plus de deux unités lexicales		●	●	●
Noms	●	●	●	●
Autres catégories grammaticales				●

EB-A = extraction de termes dans les deux langues – alignement des termes
EB-A = modèle EB-A faisant appel à la technique des patrons syntaxiques typiques
EU-T = extraction dans la langue source – identification des séquences de traduction contenant le terme cible
A-EB = alignement mot à mot – extraction simultanée de termes source et cible

Tableau XXVI. Types de variations considérés par modèle d'extraction de termes bilingue

Type de variation	EB-A	EB-A, PST	EU-T	A-EB
Synonymie (même structure syntaxique)	●	●	●	●
Synonymie (différente structure syntaxique)		●	●	●
Omission			●	
Anaphore			●	
Terme anglais				
Insertion		●	●	●
Coordination		●		●
Transformation en paraphrase				
Transformation en syntagme verbal				
Changement de catégorie grammaticale				●

EB-A = extraction de termes dans les deux langues – alignement des termes
EB-A, PST = modèle EB-A faisant appel à la technique des patrons syntaxiques typiques
EU-T = extraction dans la langue source – identification des séquences de traduction contenant le terme cible
A-EB = alignement mot à mot – extraction simultanée de termes source et cible

Les données présentées au Tableau XXV montrent que les termes simples et les termes ayant une catégorie grammaticale autre que le nom sont moins souvent pris en considération par les extracteurs automatiques bilingues, peu importe la méthode mise en œuvre.

Selon les données du Tableau XXVI, il y a encore divers cas de variation syntaxique qui ne sont pas considérés lors de l'extraction de termes bilingue : l'occurrence du terme anglais, l'omission et la substitution par des éléments anaphoriques, même si ce sont des cas assez récurrents en corpus de traductions. De plus, il est vrai que quelques extracteurs prennent en considération les variations syntaxiques par insertion et par coordination, mais ils se limitent à certains cas très

précis. La différence entre la longueur des termes source et cible est un autre aspect encore difficile à modéliser.

Par ailleurs, nous estimons que d'autres types de variation syntaxique, à savoir la substitution par une anaphore, l'omission et l'occurrence du terme anglais dans la traduction, qui ne sont pas souvent considérés, ne représentent pas un réel problème pour l'extraction de termes bilingue. Si les reprises anaphoriques et les omissions sont utilisées pour éviter des répétitions du terme source (ce qui démontre que le terme source a une fréquence élevée), cela implique que l'équivalent en langue cible apparaîtra forcement ailleurs dans le texte. De plus, peut-être que, pour les fins de l'élaboration de glossaires et de dictionnaires spécialisés ou d'autres projets terminographiques, ces trois types de variation n'apportent pas d'informations additionnelles pertinentes. Prévoir ces variations contribue, cependant, à mieux préciser où l'équivalent pour un terme source peut se trouver dans les textes.

De même, nous observons que les diverses variantes morphosyntaxiques (transformations en paraphrases, en syntagmes verbaux, changement de la catégorie grammaticale) ne sont presque jamais prises en considération, bien qu'il s'agisse aussi de cas de variation ayant une fréquence importante en traduction.

Un autre fait important que ces deux tableaux mettent en évidence est que les techniques linguistiques, plus particulièrement celle des patrons syntaxiques typiques et celle des relations de dépendance syntaxique, semblent potentiellement plus adéquates pour l'identification de variantes. Malheureusement, un inconvénient important de ces techniques est qu'elles sont dépendantes de la langue, ce qui implique que l'extracteur devra ajuster ou modifier certaines de ses composantes en fonction de chaque langue visée.

En conclusion, nous considérons que le fait de négliger les diverses variations que l'équivalent d'un terme source peut subir en langue cible affecte les taux de précision et de rappel des extracteurs de termes bilingues. La modélisation des variantes syntaxiques et morphosyntaxiques aura donc une incidence positive sur le repérage de candidats-termes ainsi que sur leur alignement.

Il reste donc à explorer davantage les techniques d'extraction existantes pour déterminer lesquelles sont plus efficaces pour l'identification de variantes ou, le cas échéant, pour en développer d'autres qui s'attaquent mieux à ce phénomène linguistique important.

Conclusion

La présente étude avait pour objectif de caractériser la variation terminologique en traduction et d'analyser les conséquences de ce phénomène linguistique sur l'extraction de termes bilingue.

Pour réaliser cette caractérisation, nous avons analysé un ensemble de termes anglais et leurs équivalents espagnols dans un corpus parallèle spécialisé. Nous avons sélectionné 23 textes anglais et leur traduction correspondante en espagnol. Les textes portaient sur les composés chimiques industriels qui posent des risques pour la santé et l'environnement. Nous avons ainsi constitué un corpus textuel d'environ 350 000 mots par langue, que nous avons ensuite aligné.

La deuxième étape de notre méthodologie a consisté à sélectionner les termes à analyser. Après une présélection faite à l'aide d'un extracteur automatique unilingue et l'application de trois critères de raffinement des listes de candidats-termes, nous avons constitué une liste de 25 termes simples et de 25 termes complexes anglais. Notre liste de termes contient non seulement des noms mais aussi quelques adjectifs et verbes. Nous voulions voir si les termes de différentes natures pouvaient varier.

Ensuite, à l'aide d'un concordancier bilingue, nous avons observé la façon dont chacun des termes choisis étaient représentés dans le corpus espagnol. Tous les équivalents espagnols (termes de base et variantes) retrouvés ont été enregistrés dans une base de données. Nous y avons également noté le nombre total d'occurrences de chaque équivalent dans le corpus ainsi que le nombre de fichiers où il apparaissait.

Nous avons organisé les données générées à l'étape d'analyse de termes en nous basant sur la classification proposée dans la section 1.1.2 du premier chapitre,

classification s'inspirant de divers travaux. Ainsi, nous avons procédé à la présentation des résultats sur les variantes observées pour nos 50 termes.

En ce qui concerne les termes simples, nous avons observé qu' un seul terme reste invariable; les 24 autres ont subi diverses variations. Au total, 102 variantes ont été retrouvées, ce qui donne un taux de variation de 4,08 pour les termes simples. Les variantes sémantiques observées sont : substitution par un synonyme et utilisation d'un terme entraînant un changement de sens. Nous avons trouvé aussi trois types de variation syntaxique : omission, substitution par une anaphore et occurrence du terme anglais dans la traduction. Le changement de catégorie grammaticale et la substitution par une paraphrase sont les deux types de variation morphosyntaxique observés. La synonymie est la forme de variation la plus fréquente dans les termes simples (elle constitue 46 % du total de variantes). Les variations par changement de catégorie grammaticale, par substitution par une anaphore et par omission ont également une fréquence significative dans le corpus (13, 7 %, 10,7 %, et 10,7 %, respectivement).

En ce qui concerne les termes complexes, tous ont fait l'objet d'au moins une variation. Nous avons retrouvé un total de 112 variantes, ce qui donne un taux de variation de 4,4. Ces termes présentent trois types de variation sémantique : par synonymie, utilisation d'un quasi-synonyme et utilisation d'un terme entraînant un changement de sens. Nous avons aussi observé cinq types de variation syntaxique : par insertion, par omission, par coordination, par substitution par une anaphore et par occurrence du terme anglais dans la traduction. Une seule forme de variation morphosyntaxique a été observée : la transformation d'un terme nominal en syntagme verbal. Comme dans les termes simples, la synonymie est la variation la plus fréquente dans les termes complexes (représentant 36,6 % du total des variantes). Les variations par insertion et par coordination ont aussi une fréquence élevée (36 % et 9,8 %, respectivement).

À partir de ces résultats nous avons formulé plusieurs conclusions. Premièrement, comme nous le supposions, tant les termes complexes que les termes simples, qu'ils soient des noms, des adjectifs ou des verbes, sont susceptibles de subir des variations. La plupart des extracteurs traitent seulement les termes complexes – se limitant parfois à ceux constitués de deux unités lexicales – de nature nominale. Toutefois, nous jugeons pertinent de prendre tous les termes en considération lors du développement d'extracteurs de termes.

À la classification proposée au début de notre étude, certains types de variation se sont ajoutés : occurrence du terme anglais dans la traduction, changement de la catégorie grammaticale, substitution par une anaphore et substitution par une paraphrase. Nous concluons ainsi que ces cas de variation sont plus fréquents en traduction. De plus, il est difficile de les traiter dans des études unilingues. D'ailleurs, les variations par omission, par synonymie et par transformation en syntagme, même si elles sont considérées dans la classification originale, se relèvent plus importantes dans un corpus de traductions que dans les études unilingues, car leur fréquence est élevée.

Or, les résultats de notre analyse confirment l'importance des variantes syntaxiques par insertion et par coordination, qui ont déjà été considérées dans les travaux unilingues portant sur la variation et caractérisées dans certains extracteurs de termes bilingues.

Compte tenu des taux de variations estimés, la diversité de variantes observées, ainsi que les limites de l'extraction bilingue, nous constatons que la variation terminologique a une incidence importante sur l'efficacité des extracteurs automatiques de termes bilingues. Malgré cela, et prenant en considération les modèles d'extraction bilingue décrits à la section 1.3, nous réalisons que peu de variantes sont considérées lors de la conception de ces extracteurs.

Par exemple, si on considère la variation sémantique, seule la synonymie est facilement prise en compte par les extracteurs bilingues. Par contre, nous sommes conscientes du fait que l'analyse sémantique des termes proposés comme équivalents n'est pas envisageable à court terme dans le contexte de l'extraction automatique de termes. Il revient au spécialiste humain d'accomplir cette tâche, puisqu'elle fait appel à des connaissances extralinguistiques.

Comme nous l'avons mentionné au préalable, les variations syntaxiques par insertion et par coordination sont déjà considérées par certains modèles d'extraction, surtout par ceux qui ont recours à des techniques linguistiques telles que les patrons syntaxiques typiques et les relations de dépendance syntaxique. Les variations morphosyntaxiques ainsi que d'autres variations syntaxiques (substitution par une anaphore, omission, occurrence du terme anglais dans la traduction) ne sont pas encore modélisées.

Comme nous l'exposons à la section 3.5 de notre étude, le fait de ne pas considérer certaines variantes entraîne des lacunes dans l'identification d'équivalents pour les termes source, mais d'autres variantes affectent également la précision de l'alignement réalisé par les extracteurs bilingues. Par exemple, il sera difficile de repérer l'équivalent d'un terme source nominal qui a été transformé en syntagme verbal dans la traduction. Il est évident que le problème se situe sur le plan de l'identification. Cependant, cela implique aussi que la phrase cible sera plus longue que la phrase cible et que les structures syntaxiques sont, évidemment, différentes. Dans ce cas, il est fort probable que l'extracteur produise des erreurs d'alignement. Déterminer avec précision les types de difficultés que posent les variantes peut contribuer à mieux définir lesquelles doivent forcement être considérées et comment elles peuvent être modélisées par le système d'extraction bilingue.

À la lumière de ces conclusions, nous pensons qu'il convient d'analyser quelles techniques et quels modèles d'extraction de termes bilingues existants se relèvent utiles pour la considération et la modélisation de variantes terminologiques. Peut-être sera-t-il nécessaire de développer de nouveaux modèles d'extraction de termes bilingues qui prennent en considération le phénomène de la variation terminologique.

De même, nous jugeons important d'étendre la présente recherche (augmenter la quantité des termes simples et complexes, inclure plus de termes appartenant à d'autres catégories grammaticales, augmenter la taille du corpus d'étude ou analyser d'autres corpus spécialisés, etc.) afin de vérifier si les résultats changent considérablement ou si les tendances que nous avons observées se maintiennent.

Bibliographie

BLANK, Ingeborg (2000) : « Terminology extraction from parallel technical texts », dans VÉRONIS, Jean (ed.) *Parallel Text Processing*, Dordrecht, Boston : Kluwer Academic Publishers, pp. 237-252.

BOURIGAULT, Didier (1993) : « Analyse syntaxique locale pour le repérage de termes complexes dans un texte », *Traitement automatique des langues,* vol. 34, n° 2, pp. 105-117.

BOURIGAULT, Didier et FABRE, Cécile (2000). « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaire*, n° 25, pp. 131-151.

BOURIGAULT, Didier et JACQUEMIN, Christian (1999) : « Term extraction and term clustering: An integrated platform for computer-aided terminology », *Proceedings of Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, pp. 15-22.

BOWKER, Lynne et PEARSON, Jennifer (2002) : *Working with specialized language. A practical guide to using corpora*, Londres, New York : Routledge, 242 p.

BROWN, Peter, LAI, Jennifer et MERCER, Robert (1991) : « Aligning sentences in parallel corpora », *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Californie, pp. 169-176.

CABRÉ, M. Teresa, ESTOPÀ Rosa et VIVALDI, Jordi (2001) : « Automatic term detection: A review of current systems », dans BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M.-C. (ed.), *Recent advances in Computational Terminology*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 53-87

CARL, Michael, RASCU, Ecaterina, HALLER, Johann et LANGLAIS, Philippe (2004) : « Abducing term variant translations in aligned texts », *Terminology*, vol. 10, n° 1, pp. 103-133.

CHURCH, Kenneth et HANKS, Patrick (1990) : « Word association norms, mutual information, and lexicography », *Computational Linguistics*, vol. 16, n° 1, pp. 22-29.

DAGAN, Ido et CHURCH, Ken (1997) : « *Termight* : Coordinating humans and machines in bilingual terminology acquisition », *Machine Translation*, vol. 12, n° 1, pp. 89-107.

DAILLE, Béatrice (1995) : « Repérage et extraction de terminologie par une approche mixte statistique et linguistique », *Traitement Automatique des Langues,* vol. 36, n° 1-2, pp. 101-118.

DAILLE, Béatrice, GAUSSIER, Éric et LANGÉ Jean-Marc (1994) : « Towards automatic extraction of monolingual and bilingual terminology », *Proceedings of the 15th International Conference on Computational Linguistics* (*COLING 94*), Kyoto, pp. 515-521.

DAILLE, Béatrice, HABERT, Benoît, JACQUEMIN, Christian et ROYAUTÉ, Jean (1996) : « Empirical observation of term variations and principles for their description", *Terminology*, vol. 3, n° 2, pp. 197-257.

DÉJEAN, H. et GAUSSIER, E. (2002) : « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », *Lexicometrica, Alignement lexical dans les corpus multilingues*, numéro spécial, pp. 1-21.

DROUIN, Patrick (2002) : *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*, thèse de doctorat présentée à l'Université de Montréal, 274 p.

EURODICAUTOM (2004), Base de donnée terminologique de la Commission Européenne. http://europa.eu.int/eurodicautom/Controller.

FREIXA, Judit (2001) : « Reconocimiento de unidades denominativas : incidencia de la variación en el reconocimiento de las unidades terminológicas », dans CABRÉ, M. Teresa et FELIU, Judit (ed.), *La terminología, científico técnica: Reconocimiento, análisis y extracción de información formal y semántica*, Barcelone : IULATERM, pp. 57-65.

----- (2002) : *Anàlisi de la variació denomitiva en textos de different grau d'especialització de l'àrea de medi ambient*, thèse de doctorat présentée à l'Université de Barcelone, 397 p.

GALE, William et CHURCH, Kenneth (1991) : « A program for aligning sentences in bilingual corpora », *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Californie, pp. 177-183.

GAUSSIER, Éric (2001) : « General considerations on bilingual terminology extraction », dans BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M.-C. (ed.), *Recent advances in Computational Terminolgy*, Amsterdam, Philadelphia : John Benjamins Publishing Company, pp. 167-182.

GAUSSIER, Éric et LANGÉ, Jean-Marc (1995) : « Modèles statistiques pour l'extraction de lexiques bilingues », *Traitement automatique des langues*, vol. 36, n° 1-2, pp. 133-155.

----- (1997) : « Some methods for the extraction of bilingual terminology », dans Jones, Daniel et Somers Harold (ed.), *New methods in language processing*, Londres : UCL Press, pp. 145-153.

GAUSSIER, Éric, HULL, David et Aït-Mokhtar, Salah (2000) : « Term alignment in use », dans VÉRONIS, Jean, *Parallel Text Processing*, Dordrecht, Boston : Kluwer Academic Publishers, pp. 253-274.

Glosario electrónico de la Conferencia de las Naciones Unidas sobre el Medio Ambiente y el Desarrollo, CNUMAD (1998).

Glosario electrónico de la Organización Mundial del Comercio, OMC (1998).

GLOSSAIRE MULTILINGUE DE L'ENVIRONNEMENT (2004), Agence européenne pour l'environnement. http://glossary.eea.eu.int/EEAGlossary/.

HAMON, Thierry, NAZARENKO, Adeline et GROS, Cécile (1998) : « A step towards the detection of semantic variants of terms in technical documents », *Proceedings of the 17th International Conference on Computational Linguistics*, Montréal, pp. 498-504.

HARRIS, Brian (1988) : « Bi-text: A new concept in translation theory », *Language Monthly*, n° 54, pp. 8-10.

HULL, David (1997) : « Automating the construction of bilingual terminology lexicons », *Terminology*, vol. 4, n° 1, pp. 225-244.

HULL, David (2001) : « Software tools to support the construction of bilingual terminology lexicons », dans BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M.-C. (ed.), *Recent advances in Computational Terminology*, Amsterdam, Philadelphia : John Benjamins Publishing Company, pp. 225-244.

IBEKWE-SANJUAN, Fidelia (1998a) : « Terminological variation, a means of identifying research topics from texts », *Joint International Conference on Computational Linguistics* (ACL-COLING'98), Montréal, pp. 654-570.

----- (1998b) : « A linguistic and mathematical method for mapping thematic trends from texts », *13th European Conference on Artificial Intelligence (ECAI 98)*, Brighton, pp. 170-174.

ISABELLE, Pierre (1992) : « La bi-textualité : Vers une nouvelle génération d'aides à la traduction et la terminologie », *META*, vol. 37, n° 4, pp. 721-737.

ISABELLE, Pierre et SIMARD, Michel (1996) « Propositions pour la représentation et l'évaluation des alignements et des textes parallèles », *Rapport technique du CITI*. Laval, Canada. http://www-rali.iro.umontreal.ca/arc-a2/PropEval

JACQUEMIN, Christian (1999) : « Syntagmatic and paradigmatic representations of term variation », *Proceedings of the 37th Meeting of the Association for Computational Linguistics*, Maryland, pp. 341-348.

LADOUCEUR, Jacques et DROUIN, Patrick (1997) : « Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité », *META*, vol. 42, n° 1, pp. 207-218.

LEBART, Ludovic et SALEM, André (1988) : Analyse statistique des données textuelles, Paris : Dunod, 209p.

L'HOMME, Marie-Claude (2004) : *La terminologie : principes et techniques*, Montréal, Les Presses de l'Université de Montréal, 278 p.

MARTÍN, Antonio et SANTAMARÍA, Jesús Miguel (2000) : *Diccionario Terminológico de contaminación ambiental*, Navarra : Ediciones Universidad de Navarra, 312 p.

OZDOWSKA, Sylwia et BOURIGAULT, Didier (2004) : « Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus », *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, Toulouse. www.laas.fr/rfia2004/actes/ARTICLES/326.pdf

TERMIUM (2004) Base de données terminologique et linguistique du gouvernement du Canada. http://www.termium.gc.ca/.

TERMOSTAT (2003) http://olst.ling.umontreal.ca/~drouinp/termostat_web.

VÉRONIS, Jean (2000) : « From the Rosetta stone to the information society. A survey of parallel text processing », dans VÉRONIS, Jean (ed.), *Parallel Text Processing*, Dordrecht, Boston : Kluwer Academic Publishers, pp. 1-24.

VÉRONIS, Jean et LANGLAIS, Philippe (2000) : « Evaluation of parallel text alignment systems: The Arcade project », dans VÉRONIS, Jean (ed.), *Parallel Text Processing*, Dordrecht, Boston : Kluwer Academic Publishers, pp. 369-388.

YOSHIKANE, Fuyuki, TSUJI Keita, KAGEURA, Kyo et JACQUEMIN Christian (1999) : « Detecting Japanese Term Variation in Textual Corpus », *4th International Workshop on Information Retrieval with Asian Languages*, Taipei, pp. 97-108.