Université de Montréal

Étude sur l'équivalence de termes extraits automatiquement d'un corpus parallèle : contribution à l'extraction terminologique bilingue

par

Annaïch Le Serrec

Département de linguistique et 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

Mai 2008

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Étude sur l'équivalence de termes extraits automatiquement d'un corpus parallèle : contribution à l'extraction terminologique bilingue

présenté par :
Annaïch Le Serrec

a été évalué par un jury composé des personnes suivantes :

Gilles Bélanger, président-rapporteur
Marie-Claude L'Homme, directeur de recherche
Patrick Drouin, co-directeur
Olivier Kraif, membre du jury

Résumé

L'étude entreprise dans le cadre de ce mémoire se veut une contribution à l'extraction bilingue de termes en vue notamment de construire des dictionnaires spécialisés et des lexiques. Plus précisément, notre but est d'examiner en corpus des candidats termes (CT) français et leurs équivalents anglais afin d'analyser ces derniers et de catégoriser les CT selon les types d'équivalents qu'ils possèdent.

À l'heure actuelle, les techniques d'alignement de textes et d'extraction de termes sont suffisamment au point pour donner des résultats satisfaisants. Par contre, l'extraction bilingue de termes connaît encore de nombreuses difficultés et les données empiriques sur les problèmes linguistiques que pose ce genre d'approche sont peu nombreuses. Par ailleurs, nous avons constaté que les travaux portant sur l'acquisition de termes s'intéressaient principalement aux substantifs et aux termes complexes. Dans ce mémoire, pour nous démarquer de ces travaux, nous étudions les termes simples appartenant aux parties du discours du nom, de l'adjectif, du verbe et de l'adverbe.

Pour mener à bien cette étude, nous avons constitué un corpus parallèle relevant du domaine du changement climatique comptant plus de 500 000 mots par langue. Une fois le corpus aligné, nous avons procédé à l'extraction automatique des termes français et anglais. Nous avons ensuite prélevé les 50 premiers CT de la liste d'extraction française afin d'identifier en corpus leurs équivalents anglais. Puis nous avons classé chacun des CT en fonction des types d'équivalents qui lui sont associés. Nous avons repéré les équivalents dans la liste d'extraction anglaise et calculé les écarts entre les CT français et leurs équivalents. Enfin, nous avons comparé la liste d'équivalents compilée manuellement avec une liste d'extraction du lexique générée par le module d'extraction lexicale d'un système d'alignement de textes bilingue.

L'identification des équivalents nous a permis de constater que l'établissement de l'équivalence pose des problèmes à peu près identiques pour les termes simples que pour les termes complexes. Nous avons observé que lorsqu'un CT possède plusieurs équivalents, le nombre d'occurrences de l'équivalent privilégié est presque toujours beaucoup plus élevé que le nombre des autres équivalents. Nous avons également remarqué que les équivalents anglais étaient tous présents, sauf deux, dans la liste d'extraction anglaise et que les écarts entre les CT français et leur équivalent privilégié étaient généralement peu élevés. Enfin, nous avons constaté que la liste d'extraction du lexique générée par le système d'alignement proposait 69 % des équivalents compilés manuellement.

Mots-clés : terminologie bilingue, corpus spécialisé, corpus parallèle, extraction semi-automatique de termes, équivalence, changement climatique.

Abstract

This work is a contribution to bilingual term extraction as a means to build specialized dictionaries, lexicons, etc. More precisely, our aim is to identify in a corpus French candidate terms (CTs) and their English equivalents in order to study the latter, and classify the CTs according to the types of their equivalents.

At the present time, text alignment and term extraction techniques are sufficiently developed and offer good results. On the other hand, bilingual term extraction still faces a number of difficulties, and empirical data concerning the linguistic problems this approach poses are scarce. Furthermore, we have observed that term extraction studies were mostly interested in nouns and multi-word terms. In this research, in order to distance ourselves from previous work, we study single-word terms belonging to the following parts of speech: noun, adjective, verb and adverb.

To conduct this study, we built a parallel corpus pertaining to climate change, and containing more than 500,000 words per language. Once the corpus was aligned, we have proceeded to automatically extract English and French terms. We then took the 50 first CTs from the French list of terms to identify in our corpus their English equivalents. After that, we have classified each of the French CTs according to the types of equivalents. We have then located the equivalents in the English list of terms and we have calculated the distances between the French CTs and their equivalents. Finally, we have compared the list of equivalents compiled manually with a lexical extraction list proposed by an aligner.

The identification of the equivalents has allowed us to see that the establishment of equivalence poses similar difficulties for single-word terms and for multi-word terms. We have observed that, when a CT possesses several equivalents, the number of occurrences of the privileged equivalent is practically always far more important than the number of the other equivalents. We have also noted that, except for two, the English equivalents were all

present in the English list of terms, and that the position between the French CTs and their privileged equivalent was generally quite similar. Finally, we have observed that in the lexical extraction, the aligner proposes 69% of the equivalents compiled manually.

Keywords: bilingual terminology, specialized corpora, aligned corpora, semi automatic extraction, equivalence, climate change.

Table des matières

Liste des tableaux

Liste des figures

Vers la fin du XII^e siècle, Pierre de Blois écrivait à peu près ceci :

« Nous nous croyons des géants, mais nous ne sommes que des nains juchés sur les épaules de nos prédécesseurs, et si nous voyons plus loin qu'eux, c'est parce que nous prenons appui sur eux. »

Remerciements

En premier lieu, je tiens à remercier vivement le Conseil de recherches en sciences humaines du Canada (CRSH) et le Fonds québécois de la recherche sur la société et la culture (FQRSC). Je leur suis redevable du soutien financier qu'ils m'ont accordé pour réaliser mon projet de recherche. Je suis également redevable à l'Université de Montréal pour les bourses qu'elle m'a octroyées durant ma scolarité.

Je souhaite exprimer toute ma reconnaissance et ma gratitude à ma directrice de recherche, Marie-Claude L'Homme, pour m'avoir invitée à suivre la voie de la recherche, pour la confiance qu'elle m'a accordée d'emblée, pour l'intérêt qu'elle a porté à mon travail et pour sa constante disponibilité. Par ses nombreuses lectures et relectures de mon mémoire, j'ai pu profiter de sa très précieuse expérience. Enfin, je la remercie tout particulièrement d'avoir toujours su m'aiguiller, me conseiller avec générosité et bienveillance.

Je tiens également à offrir mes sincères remerciements à Patrick Drouin, mon codirecteur de recherche. Ses lectures attentives, ses remarques judicieuses et ses explications techniques m'ont été d'un grand secours dans la rédaction de ce mémoire. Je lui sais gré d'avoir paramétré TermoStat pour une extraction terminologique « personnalisée ».

Je voudrais qu'Olivier Kraif, mon superviseur de stage à Grenoble et concepteur d'Alinea, trouve ici le témoignage de mon estime. Je le remercie pour ses qualités scientifiques et humaines et son gratin savoyard.

J'adresse mes remerciements à ma collègue, Iveth Carreño, pour son écoute attentive et ses commentaires. Je n'oublie pas non plus ceux et celles qui ont suivi mon parcours d'un peu plus loin (Françoise, Stéphane, etc.).

Qu'il me soit enfin permis de présenter publiquement mes plus sincères remerciements à Georges et à nos enfants, Yan, Gaël et Christelle. Ensemble, nous avons vécu ces années d'études comme une aventure passionnante et enrichissante. Je leur dédie ce mémoire.

Introduction

L'utilisation généralisée des outils de bureautique, l'avènement d'Internet et la mondialisation ont donné lieu à une production massive et sans cesse grandissante de documents électroniques de tout genre. Or, parmi tous ces documents, il existe de nombreux « textes parallèles[1] » en deux langues, voire plusieurs langues. Dans le domaine du traitement automatique du langage naturel (TALN), on s'est rapidement aperçu des nombreuses possibilités que présentent ces textes. L'étude d'un corpus bilingue permet, entre autres, de faire ressortir des faits de langue moins visibles dans un corpus unilingue. Les corpus bilingues peuvent notamment faire l'objet d'une multitude d'applications dans divers domaines : la traduction, la lexicologie, l'ingénierie des langues, les études comparées en littérature, l'enseignement et la terminologie. Dans la présente étude, nous nous intéressons à la dernière application et plus particulièrement à la problématique de l'extraction bilingue de termes à partir de corpus parallèles[2,3].

La constitution de recueils terminologiques spécialisés de tout genre bénéficie d'une forte demande de la part des traducteurs, des entreprises et des institutions (Bourigault et Slodzian 1999). Pour répondre à cette demande, le recours aux outils informatiques et à ses techniques est devenu une nécessité. Dès la fin des années 1980, on a été en mesure d'aligner les textes de façon satisfaisante au niveau de la phrase (Kay et Röscheisen 1988). Vers la même époque, les extracteurs de termes (David et Plante 1990) ont vu le jour, suivis peu après par les premières tentatives d'extraction de termes bilingues (van der Eijk 1993). Depuis ce temps, des progrès considérables ont été réalisés sur le plan de l'alignement des textes et de l'extraction des termes. Par contre, les recherches sur l'extraction bilingue de

---

[1] Deux textes sont dits parallèles lorsque l'un d'eux est la traduction de l'autre.

[2] Un corpus parallèle est « un ensemble composé de textes sources et de textes cibles dont les composantes formelles ([les mots], les phrases ou les paragraphes) ont été alignées afin d'en faciliter la consultation » (L'Homme 2005b).

[3] De nombreux auteurs (Harris B. 1988; Kraif 2001; L'Homme 2005b) utilisent le terme *bi-texte, bitexte* ou encore *textes alignés*.

termes n'ont pas produit les résultats escomptés. Il reste encore beaucoup de difficultés à résoudre et les données empiriques sur les problèmes linguistiques que pose ce genre d'approche sont peu nombreuses.

Ne possédant pas la connaissance du monde, les outils informatiques éprouvent de la difficulté à apparier les termes automatiquement. Ils se heurtent à des obstacles qui leur sont propres. Parmi les problèmes fréquemment évoqués par les auteurs (L'Homme 2004; Gaussier 2001; Kraif 2001; etc.), on peut citer la complexité de distinguer les unités terminologiques des mots de la langue générale, d'effectuer le découpage des termes adéquatement et d'établir les équivalences.

Outre les difficultés que nous venons d'évoquer, les auteurs qui se sont penchés sur l'extraction de termes, que ce soit en extraction unilingue ou plurilingue, se sont principalement intéressés aux termes complexes. Au-delà du fait que ces derniers sont plus facilement repérables par les systèmes d'extraction, de nombreux chercheurs affirment que les termes spécifiques à un domaine sont principalement des syntagmes nominaux (Daille *et al.* 1994; Gaussier 2001). Cette idée ne fait pourtant pas l'unanimité comme en témoignent les écrits de L'Homme (2005a : 1126) ou d'Estopà (2001 : 217-237). Voici ce que dit cette dernière à ce propos :

> Des tests réalisés dans le domaine de la biomédecine signalent que l'ensemble des unités monolexicales spécialisées d'un texte thématiquement spécialisé ne peut être ignoré, car cela correspond à 35 % et à 45 % des unités avec un signifié spécialisé (Estopà 2001 : 220).

Par ailleurs, ces deux auteurs ont également relevé que l'on s'intéresse principalement aux termes nominaux, les autres parties du discours étant largement négligées (L'Homme 2005a : 1119). Comme de juste, les observations dont nous venons de rendre compte semblent se vérifier, puisque aucune étude n'a porté uniquement sur les termes simples appartenant à plusieurs parties du discours. Enfin, une autre idée à laquelle plusieurs chercheurs en extraction bilingue de termes souscrivent est que les termes n'ont qu'un seul équivalent dans les textes spécialisés parallèles (Fung 1998; Gurrutxaga *et al.* 2006). Ici

encore, cette notion est remise en cause dans plusieurs travaux (Jacquemin 1997; Carreño 2004).

Dans le présent mémoire, en réponse à ces observations et en vue de contribuer à l'extraction terminologique bilingue, il nous a semblé opportun d'étudier des termes simples appartenant aux parties du discours du nom, de l'adjectif, du verbe et de l'adverbe extraits automatiquement d'un corpus parallèle. Plus précisément, notre but est d'examiner, du point de vue de l'extraction automatique, des candidats termes (CT)[4] français et leurs équivalents anglais afin d'analyser ces derniers et de catégoriser les CT selon leurs équivalents.

Pour ce faire, nous avons constitué un corpus parallèle (anglais – français) dont les textes se rapportent au domaine du changement climatique. Au terme de la cueillette, nous avons ainsi obtenu un corpus constitué de 31 paires de textes comptant plus de 500 000 mille mots pour l'anglais et plus de 600 000 mots pour le français.

À l'aide de l'aligneur Alinea (Kraif 2001), nous avons tout d'abord aligné le corpus parallèle au niveau des phrases. Puis, nous avons extrait de ce corpus deux listes de CT simples (anglais – français) appartenant aux quatre parties du discours citées plus haut. Pour effectuer cette tâche, nous avons utilisé l'extracteur de termes TermoStat (Drouin 2002), un des rares systèmes à extraire des termes simples. Il convient de préciser ici que les CT peuvent être des termes simples, des têtes de syntagme, des modificateurs ou encore des non-termes. Cela s'explique, d'une part, par le fait que les logiciels dans leur ensemble ne peuvent pas déterminer avec exactitude le statut terminologique d'une unité lexicale et que, d'autre part, les logiciels d'extraction qui ne reposent pas sur une analyse complète de la phrase ne sont pas en mesure de faire la différence entre la tête ou le

---

[4] Les candidats termes extraits par le logiciel d'acquisition de termes ne sont pas des termes tant qu'ils n'ont pas été validés par l'humain.

modificateur d'un syntagme. Nous avons ensuite nettoyé les listes d'extraction en éliminant uniquement les CT n'appartenant pas à l'une des quatre parties du discours étudiées dans ce mémoire[5].

Notre méthode d'analyse des CT français et de leurs équivalents anglais s'est déroulée en quatre étapes : 1) prélèvement des 50 premiers CT de la liste d'extraction française nettoyée; 2) sélection dans les 31 paires de textes d'au plus 310 paires de contextes (français – anglais) par CT; 3) description des équivalents de chacun des CT; 4) classification des CT français selon les types d'équivalents anglais. Nous aimerions souligner ici que notre démarche est descriptive et non pas normative, c'est-à-dire qu'elle vise à décrire les équivalences en contexte, sans porter de jugement de valeur sur la qualité de ces équivalents. Notre étude se situe donc en amont de la validation du terminologue. Une fois la classification des CT effectuée, nous avons décrit dans le détail leurs équivalents.

Nous avons procédé ensuite au repérage de la position des équivalents dans la liste d'extraction anglaise de TermoStat et au calcul des écarts entre les CT français et leurs équivalents. Ces deux dernières opérations avaient pour but de vérifier si les équivalents étaient présents dans la liste d'extraction anglaise et de voir le rang qu'ils occupaient par rapport aux CT français.

Pour terminer, en nous guidant sur les résultats de notre étude pour paramétrer le module d'extraction du lexique d'Alinea[6], nous avons comparé la liste d'équivalents

---

[5] Même si le logiciel d'extraction n'a été paramétré que pour extraire les CT spécifiques au corpus appartenant à la partie du discours du nom, de l'adjectif, du verbe et de l'adverbe, il produit quand même une certaine quantité de bruit.

[6] Une fois l'alignement effectué au niveau des mots, Alinea offre la possibilité d'extraire le lexique du corpus.

compilée manuellement à partir de nos fiches avec la liste des équivalents offerts par ce logiciel.

Notre mémoire s'organise selon quatre chapitres :

Le premier chapitre se divise en quatre sections et présente la partie théorique constituant la base de notre étude. Dans la première section, nous y abordons la définition du terme telle qu'appréhendée par deux courants théoriques opposés. Dans la deuxième section, nous examinons les travaux de plusieurs auteurs sur les principes théoriques de l'équivalence en terminologie. Dans la troisième section, nous nous attardons sur la description des difficultés liées à l'établissement de l'équivalence du point de vue de l'extraction automatique. Enfin, dans la quatrième section, nous dégageons les enseignements de chacun des points abordés au cours de ce chapitre et nous présentons les bases pratiques et théoriques sur lesquelles s'appuie notre mémoire.

Dans le deuxième chapitre, qui comporte quatre sections, nous nous penchons sur les méthodes et les techniques qui permettent l'extraction bilingue de termes. La première section présente le fonctionnement des outils automatiques d'alignement de textes. La deuxième section donne les principales caractéristiques des techniques et des systèmes d'extraction de termes. La troisième section décrit dans le détail les systèmes d'extraction bilingue de termes à partir de corpus parallèles. Enfin, pour clore le chapitre, dans la quatrième section, nous faisons la synthèse des travaux effectués en extraction bilingue de termes.

Le troisième chapitre est consacré à la présentation de la méthodologie adoptée dans ce mémoire. Dans la première section, nous passons en revue toutes les étapes de la préparation du corpus et de son alignement. Dans une deuxième section, nous y décrivons le logiciel d'acquisition automatique de termes et les listes de candidats termes (CT) produites par ce dernier. Dans la troisième et dernière section, nous expliquons la démarche suivie pour analyser 50 CT français et leurs équivalents anglais.

Le quatrième chapitre est dévolu à la présentation des résultats et à leur analyse. La première section sert à décrire les équivalents des 50 CT étudiés et à classifier ces derniers selon leurs équivalents. La deuxième section donne les résultats du repérage des équivalents dans la liste d'extraction anglaise et les résultats du calcul des écarts entre les CT français et leurs équivalents. Dans la troisième section, la liste d'équivalents produite manuellement et la liste d'extraction lexicale générée par Alinea sont comparées. Pour terminer, la quatrième section, commente les résultats obtenus.

Chapitre 1 : L'équivalence en terminologie

Cette mémoire se donne pour objectif d'étudier les possibilités et les difficultés que présente la constitution d'une nomenclature bilingue à partir de listes de candidats termes extraits automatiquement. En terminologie multilingue, l'équivalence est au cœur des préoccupations. Dans le chapitre qui suit, afin de mieux comprendre ce qu'est l'équivalence, nous nous penchons sur la façon dont deux courants théoriques opposés définissent le terme (section 1.1). Nous examinons les principes théoriques de l'équivalence en terminologie (section 1.2). Nous nous attardons sur les difficultés liées à l'établissement de l'équivalence du point de vue de l'extraction automatique (section 1.3). Enfin, après un résumé du Chapitre 1, nous posons les bases pratiques et théoriques sur lesquelles s'appuie cette étude (section 1.4).

1.1 Qu'est-ce qu'un terme?

Suivant les courants théoriques, le terme s'envisage selon des points de vues différents. Pour illustrer notre propos, nous limitons notre examen à deux courants souvent présentés à l'opposé l'un de l'autre : l'optique conceptuelle et la terminologie basée sur corpus.

Aussi appelée *Théorie générale de la terminologie* ou *théorie traditionnelle de la terminologie*, l'optique conceptuelle, apparue vers les années 1930, est associée à l'ingénieur Eugen Wüster. Dans le but de résoudre les problèmes de communication entre experts, Wüster élabore une théorie de la terminologie dans laquelle la normalisation des termes occupe une place privilégiée. La définition du terme selon cette théorie nous vient de Felber (1987 : 3), disciple de Wüster : « Un terme est un symbole conventionnel représentant une notion définie dans un certain domaine du savoir ». Dans cette définition, le mot *notion* est d'une importance capitale. En effet, du point de vue de l'optique

conceptuelle, le concept[7] ou *notion* représente la pierre angulaire de la terminologie, car il existe avant le terme. Ce dernier est une étiquette servant à dénommer le concept. La démarche qui consiste à partir du concept pour aller vers la forme est dite *onomasiologique*. Cette démarche, qui peut déboucher sur la normalisation (le choix d'un seul terme pour dénommer un concept), postule que pour un concept donné il n'y a en principe qu'une seule étiquette par langue. Par ailleurs, étant donné que l'on nomme les concepts, les noms sont privilégiés. Dernier point important de l'optique conceptuelle : ce ne sont que les étiquettes qui changent d'une langue à l'autre, les concepts ne varient pas puisqu'ils sont indépendants des langues.

En ce qui concerne la terminologie basée sur corpus, nous prenons comme modèle *La terminologie textuelle* (Bourigault et Slodzian 1999). Cette approche est née de la constatation de l'évolution importante de la pratique terminologique : accroissement et diversification de la production de vocabulaires spécialisés, masse des données imposant l'emploi d'outils informatiques, différentes visées des applications, etc. Les auteurs proposent une définition du terme revisitée et réadaptée aux réalités de la terminologie contemporaine :

> Le terme est un construit. Il est le produit d'un travail d'analyse, mené par le linguiste terminologue, dont les choix sont guidés par une double contrainte de pertinence :
>
> - Pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques et stables [...]
> - Pertinence vis-à-vis de l'application. Les unités finalement retenues doivent l'être en fonction de leur utilité dans l'application visée, qui s'exprime en termes d'économie et d'efficacité [...] (Bourigault et Slodzian 1999 : 31).

---

[7] En terminologie, certains auteurs utilisent également le synonyme *concept*. En linguistique, ces deux termes sont le plus souvent remplacés par le mot *signifié*. Dans cette étude, nous adoptons l'appellation *concept*, sauf lorsque le terme *notion* est utilisé dans les définitions des auteurs que nous citons.

D'après les auteurs, le texte constitue le point de départ de la collecte de ressources terminologiques. Ce faisant, ils adoptent une démarche sémasiologique – démarche plus souvent associée à la lexicographie. Ils rejettent la doctrine wüstérienne qui attribue au concept une préexistence et une priorité sur les termes, car un domaine n'est pas « comme un fragment de connaissances bien structurées, permanentes et clairement circonscrites » (Bourigault et Slodzian 1999 : 31). En outre, selon les auteurs, d'un point de vue pratique, cette optique permet plus facilement de retenir des termes appartenant à des parties du discours différentes du nom, étant donné qu'ils « ne sont pas les seules unités lexicales à décrire » (*ibid.*). Enfin, l'approche textuelle se veut plus descriptive que normative.

1.2 Principes théoriques de l'équivalence en terminologie

Pour commencer, nous proposons deux définitions de l'équivalence en terminologie. La première est empruntée à Termium (2007) : « Relation entre deux termes de langues différentes qui désignent une même notion ». La deuxième est tirée de L'Homme (2004 : 115) : « Des termes sont équivalents lorsqu'ils ont les mêmes composantes sémantiques ». On remarquera que la première définition reflète nettement la vision conceptuelle de la terminologie classique. La deuxième définition, quant à elle, s'inspire de l'approche dite *lexico-sémantique* (L'Homme 2005a). Cette approche, qui part nécessairement des textes, considère les termes comme des unités lexicales et les décrit dans leur fonctionnement linguistique (*ibid.* : 1123). Il est à noter également que l'approche lexico-sémantique s'intéresse aux termes appartenant aux parties du discours du nom, de l'adjectif, du verbe et de l'adverbe et envisage surtout les termes simples (seuls les termes complexes non compositionnels sont retenus).

Les deux définitions que nous venons de présenter décrivent l'équivalence sous sa forme parfaite. Toutefois, comme précisé dans plusieurs ouvrages de terminologie (Dubuc 2002 : 73-74; L'Homme 2004 : 115; Rondeau 1981 : 33; Van Campenhoudt 2001 : 181-209), l'équivalence exacte n'est pas toujours possible; on parlera alors d'*équivalence*

partielle ou même d'*absence d'équivalence*. Ainsi, l'équivalence peut être divisée selon trois cas de figure : l'équivalence exacte, l'équivalence partielle et l'absence d'équivalence.

La méthode classique pour déterminer à quel niveau d'équivalence se situent deux termes susceptibles d'être équivalents consiste à comparer le type de rapport qu'entretient une dénomination (D) d'une langue L_1 et la notion (N) qu'elle recouvre avec une dénomination (D) d'une langue L_2 et la notion (N) qu'elle recouvre (Rondeau 1981 : 33). L'approche lexico-sémantique, de son côté, établit le niveau d'équivalence entre deux termes par l'analyse et la comparaison de chacune de leurs composantes sémantiques.

1.2.1 Équivalence exacte

La relation d'équivalence est dite *exacte* quand deux termes, T_1 et T_2, de langues différentes, L_1 et L_2, affichent un rapport identique entre la notion (N) et la dénomination (D). La formule employée par Rondeau s'exprime comme suit :

$$T_{1\,(L_1)} = T_{2\,(L_2)} = \left(\frac{D}{N}\right)_{L1} = \left(\frac{D}{N}\right)_{L2} \text{(Rondeau 1981 : 33)}$$

Pour illustrer la formule, prenons un exemple classique, *ordinateur* et *computer* (Rondeau 1981 : 33; L'Homme 2004 : 115). Ces termes, en informatique, entretiennent un rapport d'équivalence parfaite et peuvent être appariés puisqu'ils recouvrent la même notion[8] : « Équipement informatique comprenant les organes nécessaires à son fonctionnement autonome, qui assure, en exécutant les instructions d'un ensemble structuré de programmes, le traitement rapide de données codées sous forme numérique qui peuvent être conservées et transmises » (Académie Française 1997). En appliquant la formule de Rondeau, nous obtenons le résultat présenté à la Figure 1.1.

---

[8] Ou encore, du point de vue de la lexico-sémantique, parce qu'ils possèdent les mêmes composantes sémantiques.

$$\text{computer} = \text{ordinateur} = \left(\frac{\text{computer}}{\text{« Équipement informatique… »}\ \text{(Académie 1997)}} \right)_{L1} = \left(\frac{\text{ordinateur}}{\text{« Équipement informatique… »}\ \text{(Académie 1997)}} \right)_{L2}$$

Figure 1.1 : Exemple d'équivalence parfaite

1.2.2 Équivalence partielle

Toutefois, il n'est pas toujours possible d'apparier les termes aussi facilement que dans la section précédente, car comme le précise Van Campenhoudt :

> La linguistique a depuis longtemps montré que toutes les langues n'approchent pas la réalité de la même manière et que de nombreux problèmes se posent lors de l'établissement d'équivalences (Van Campenhoudt 1996b : section 7.1).

Par exemple, un terme d'une langue L_1 ne sera pas tout à fait équivalent à un terme d'une langue L_2 si le terme de L_1 ne recouvre pas exactement la même réalité que le terme de L_2. Dans ce cas de figure, nous avons affaire à de l'équivalence partielle.

Selon Rondeau, l'équivalence partielle a pour origine deux causes principales :

1) dans un premier cas, l'équivalence est partielle « […] parce qu'il faut deux termes en langue L_1 pour correspondre à un terme en langue L_2 :

$$\left(\left(\frac{D}{N} \right)_X + \left(\frac{D}{N} \right)_Y \right)_{L1} = \left(\frac{D}{N} \right)_{L2} \text{ » (Rondeau 1981 : 34).}$$

Rondeau donne l'exemple suivant : en anglais, *coffee mill* et *coffee grinder* veulent respectivement dire « moulin à café à couteaux » et « moulin à café à meule », alors qu'en français il n'existe qu'un terme plus générique *moulin à café* (Figure 1.2).

$$\left(\left(\frac{\text{coffee grinder}}{\text{« moulin à café à couteaux »}} \right)_X + \left(\frac{\text{coffee mill}}{\text{« moulin à café à meule »}} \right)_Y \right)_{L1} = \left(\frac{\text{moulin à café}}{\text{« moulin à café à couteaux ou à meule »}} \right)_{L2}$$

Figure 1.2 : Exemple d'équivalence partielle

2) dans un deuxième cas, « [...] la dénomination en langue L_1 ne recouvre que partiellement la notion exprimée en langue L_2 :

$$\left(\frac{D}{N_X} + \frac{}{N_Y}\right)_{L1} = \left(\frac{D}{N}\right)_{L2}$$ » (Rondeau 1981 : 34).

Rondeau ne donnant pas d'exemple, nous avons emprunté le suivant à Kraif (2001 : 117). Le terme anglais *nut*, employé comme hyperonyme, recouvre les hyponymes *almond, peanut, walnut, pecan nut, cashew nut, hazelnut* (Figure 1.3) :

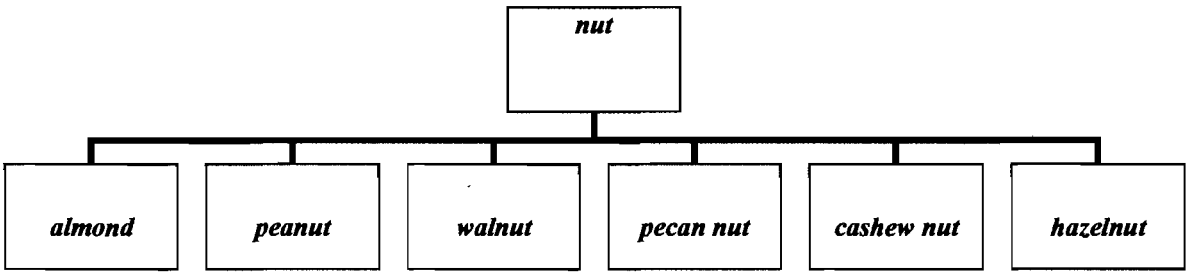


Figure 1.3 : Le terme anglais *nut* et ses hyponymes (d'après Kraif 2001 : 117)

En français, nous n'avons pas d'hyperonyme qui remplisse le même rôle qu'en anglais, par exemple, *noix* ne couvre qu'une partie des hyponymes : *noix, noix de pécan, noix de cajou*, alors qu'*amande, cacahuète* et *noisette* n'en font pas partie (Figure 1.4) :

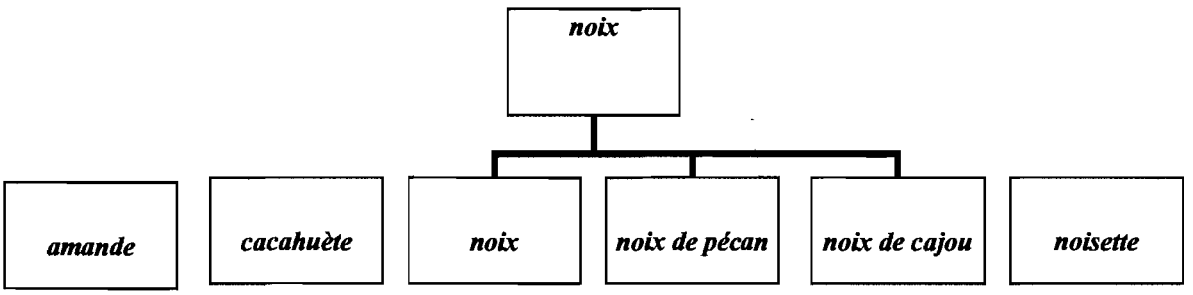


Figure 1.4 : Le terme français *noix* et ses hyponymes (d'après Kraif 2001 : 117)

Moins par préoccupation théorique que par souci de la traduction exacte, Dubuc (2002 : 74) s'intéresse également au fait que deux termes ne s'équivalent pas d'une langue à l'autre. Il distingue la *disparité de sens* et la *disparité d'usage*.

La disparité de sens se manifeste selon deux scénarios : 1) la relation de générique à spécifique; et 2) la polysémie. Dans le premier cas, un terme en langue L_1 est plus générique qu'un terme en langue L_2. Ainsi, si l'on reprend l'exemple de Dubuc (2002 : 74), le terme français *table de salon*, dans le domaine du mobilier, englobe les termes anglais *coffee table, end table, lamp table* (Tableau 1.1). On observera que cet exemple présente des points communs avec celui de la Figure 1.2 de Rondeau (1981 : 34).

Tableau 1.1 : Représentation du terme *table de salon* et ses équivalents

Langue	**Terme**		
français	*table de salon*		
anglais	*coffee table*	*end table*	*lamp table*

Dans le deuxième cas, dans un domaine donné, un terme en langue L_1 est polysémique alors qu'en langue L_2 chacun des sens du terme de la langue L_1 est exprimé par un terme différent (Tableau 1.2). L'exemple dont nous nous servons pour illustrer ce cas de disparité de sens nous vient de Van Campenhoudt (1996a : 284). Dans le domaine de la marine, le terme anglais *watch* exprime : 1) l'action de veiller; 2) une période pendant laquelle une partie de l'équipage est en service; et 3) la partie de l'équipage en service. Le français, pour exprimer ces trois sens, utilise trois termes différents, *veille, quart, bordée.*

Tableau 1.2 : Représentation du terme *watch* et ses équivalents

Sens	**Anglais**	**Français**
Action de veiller.	*watch* 1	*veille*
Période pendant laquelle une partie de l'équipage est en service.	*watch* 2	*quart*
Partie de l'équipage en service.	*watch* 3	*bordée*

La disparité d'usage se rencontre lorsqu'un concept dans une langue L_1 est désigné par plusieurs termes de niveaux de langue différents alors qu'en langue L_2, il n'existe qu'un terme. Dubuc (2002 : 75) nous donne l'exemple du terme anglais *zoom* qui, dans le domaine de la photographie, représente un « dispositif optique qui sert à effectuer des plans rapprochés ou éloignés sans avoir à se déplacer ». En français, ce même dispositif est désigné sous le nom d'*objectif à focale variable* en langue technico-scientifique, et de *zoom* dans le jargon des studios.

1.2.3 Absence d'équivalence

Enfin, à l'autre extrémité du spectre de l'équivalence, nous trouvons l'absence d'équivalence. Cette absence survient : 1) soit parce qu'un concept dans une langue L_1 est inconnu dans une langue L_2; 2) soit parce qu'un nouveau concept nommé dans une langue L_1 n'a pas encore reçu de dénomination dans une langue L_2 (Rondeau 1981 : 34).

En ce qui concerne le premier cas de figure, certains domaines sont particulièrement marqués par ce genre de difficulté. Le domaine juridique est l'exemple par excellence parce qu'il est fortement influencé par les systèmes particuliers à chaque société. Par conséquent, de nombreux concepts présents dans une langue ne trouvent pas d'équivalents dans d'autres langues. Ainsi pour les termes anglais *common law* et *Equity*, Gémar (2002 : 170-171) nous dit qu'on ne peut leur trouver d'équivalent en français et que même l'emprunt ne résout pas le problème.

Le deuxième cas de figure est très fréquent en terminologie. Constamment de nouvelles techniques, des inventions, des découvertes voient le jour et reçoivent des dénominations dans la langue de la communauté où elles sont apparues. Dans les communautés de langue différente, le concept est connu, mais il n'a pas encore été nommé. Afin de pallier le problème, on aura recours à la néologie pour nommer les nouveaux concepts qui entrent dans les habitudes d'un groupe ne partageant pas la même langue.

1.3 Problèmes d'équivalence du point de vue de l'extraction

Depuis que l'on dispose d'ordinateurs suffisamment puissants pour étudier en corpus le comportement des termes, la pratique terminologique s'est considérablement modifiée. Aujourd'hui, une quantité formidable de documents électroniques peuvent être facilement interrogés. Ainsi, comme le soulignent Bourigault et Slodzian (1999 : 29), « ce changement d'échelle met en évidence des phénomènes largement sous-estimés jusqu'ici »,

nous songeons à titre d'exemple au problème des variantes terminologiques dont il est question au point 1.3.1.

Un autre aspect intéressant du développement de l'informatique est l'apparition de nombreux outils qui facilitent, accélèrent et systématisent la collecte de ressources terminologiques : extracteurs de termes, concordanciers, aligneurs de corpus, etc. Toutefois, si ces outils constituent une aide appréciable, il n'en demeure pas moins que leur portée est limitée par certains obstacles inhérents à leur nature. Par exemple, au chapitre qui traite de l'extraction bilingue, L'Homme (2004 : 208) nous présente quelques cas de figure dans lesquels les extracteurs bilingues de termes éprouvent des difficultés pour établir des équivalences terminologiques :

1. Un terme dans une langue A a plusieurs équivalents dans le corpus de la langue B. Par exemple, *intelligent terminal* peut se rendre par *terminal intelligent, terminal programmable* et *terminal avec mémoire* dans un corpus français.

2. Les termes complexes ont, dans chacune des langues, des structures ou des longueurs différentes. Par exemple, *portable life support system,* qui comporte quatre noms, se traduit par *équipement de survie* composé de deux noms et d'une préposition.

3. Un terme complexe dans une langue équivaut à un terme simple dans l'autre langue. *Computer-assisted terminography* se rend en français par *terminotique.* De même, *base de données* est rendu par *database* en anglais.

4. Un syntagme nominal s'exprime dans l'autre langue par un syntagme d'une autre nature. Par exemple, *law suit* peut se traduire par *poursuite en justice.* Mais dans d'autres phrases, il donnera lieu à *poursuivre en justice.*

5. Il arrive qu'une phrase d'un texte source comporte une mention explicite à un terme complexe, mais que la phrase du texte cible utilise plutôt une anaphore. Par exemple, *...the disk drive is identified...* peut se rendre par *...ce dispositif est identifié...* ou encore *...celui-ci est identifié...* (L'Homme 2004 : 208).

Dans les sections qui suivent, 1.3.1 à 1.3.5, nous reprenons les points énumérés ci-dessus et les examinons à tour de rôle d'un peu plus près.

1.3.1 Un terme en L_1 possède plusieurs équivalents en L_2

Lorsqu'un terme en L_1 possède plusieurs équivalents en L_2, cela peut être dû à la polysémie du terme en L_1 ou à la synonymie (ou variation) observable en L_2. Ces phénomènes sont connus et les dictionnaires en rendent compte. En ce qui concerne la variation, elle a été l'objet de plusieurs études, notamment celle de Carreño (2004).

> 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 (Carreño 2004 : 6).

En extraction bilingue de termes comme en extraction monolingue, la question des variantes pose des problèmes. Les travaux de Carreño montrent que la variation en corpus parallèle anglais et espagnol touche la presque totalité des termes étudiés et se manifeste sous différentes formes. Sur les 50 termes étudiés par l'auteure, un seul restait invariable dans le corpus, soit le terme *bioconcentrate* qui ne donne qu'un équivalent, *bioconcentrarse*. Tous les autres termes du texte source, simples ou complexes, ont pour équivalent au moins un terme et une variante dans le texte cible. Par exemple, *landfill* donne l'équivalent attendu, *relleno sanitario*, et neuf variantes *landfill, terraplén, de ellos, dichos rellenos, vertedero de desechos, ø, relleno, confinamiento, vertedero* (Carreño 2004 : 81). En conclusion, l'auteure souligne l'importance que joue la variation terminologique sur la performance des extracteurs bilingues de termes et constate que ces outils informatiques ne prennent pas suffisamment en compte ce phénomène (Carreño 2004 : 122-123).

1.3.2 Différence de structure ou de longueur entre termes de L_1 et de L_2

En ce qui concerne le cas de figure touchant à la différence de structure ou à la longueur des termes complexes, nous avons trouvé deux études en extraction bilingue de termes dans lesquelles on a catégorisé les règles de transformation des termes afin de rendre

compte du problème de la différence de structure entre équivalents. Ces règles peuvent, en quelque sorte, être associées à ce que nous appelons *typologie de l'équivalence formelle*. On remarquera cependant que les travaux que nous présentons se sont essentiellement intéressés à la structure des termes plutôt qu'à leur longueur. Les termes complexes étudiés sont composés exclusivement de syntagmes nominaux formés de deux unités lexicales, c'est-à-dire que les termes complexes de plus de deux unités lexicales ne sont pas considérés.

Dans une première étude portant sur l'extraction de termes bilingues en biologie moléculaire à partir de corpus comparables, Tran *et al.* (2003) ont observé que la règle de transformation la plus fréquente pour les termes complexes du français vers l'anglais est la transposition syntaxique. Cette dernière est elle-même sous-divisée selon trois types de transposition syntaxique : inversion syntaxique isocatégorielle, inversion syntaxique transcatégorielle, inversion syntaxique en supprimant la préposition et/ou le déterminant.

a) Inversion syntaxique isocatégorielle

L'inversion isocatégorielle se caractérise par le fait que les catégories grammaticales sont conservées lorsqu'un terme français est traduit en anglais, mais que chacun des constituants du terme permute :

N + ADJ	→	ADJ + N
chaleur sensible	→	*sensible heat*

N + Abréviation/Symbole	→	Abréviation/Symbole + N
plantes c_3	→	c_3 *plants*

b) Inversion syntaxique transcatégorielle

L'inversion syntaxique transcatégorielle se traduit par la transformation de la catégorie grammaticale du modificateur du terme français par une autre catégorie

grammaticale du modificateur du terme anglais. Cette transformation est accompagnée par une inversion des constituants du terme :

N + ADJ	→	N2 + N1
changement climatique	→	*climate change*

Dans le terme complexe *changement climatique,* on observe que l'adjectif *climatique* du modificateur du terme français se transforme en nom *climate* dans le terme anglais *climate change*.

c) Inversion syntaxique en supprimant la préposition et/ou le déterminant

Dans ce type d'inversion syntaxique, la préposition et/ou le déterminant du terme complexe français sont supprimés au moment de la transformation vers l'anglais. Cette transformation est également accompagnée par une permutation des constituants du terme. Les transformations suivantes illustrent différentes possibilités de l'inversion syntaxique en supprimant la préposition et/ou le déterminant :

N1 + prép/det + N2	→	N2 + N1
absorption de la chaleur	→	*heat absorption*
cycle du carbone	→	*carbon cycle*

N1 + prép + N2	→	N2 + N1
surcharge en fer[9]	→	*iron overload*
couche d'ozone	→	*ozone layer*

N + de + éponyme	→	Éponyme + N
force de Coriolis	→	*Coriolis force*

N + de + Abréviation	→	Abréviation + N
synthèse de TNFa[10]	→	*TNFa synthesis*

[9] Ce terme n'appartient pas au domaine du changement climatique, il est emprunté aux auteurs Tran *et al.*

[10] Voir note précédente.

Les auteurs de l'étude dont nous venons de décrire les règles de transposition ont ensuite effectué un test d'automatisation de traduction à l'aide de ces règles. Le pourcentage de termes traduits correctement s'élevait à 69,9 %. Ce résultat nous montre que dans un peu plus de 30 % des cas, les équivalents ne suivent pas les règles de transposition attendues. Notons au passage que la traduction automatique n'a touché que certains types de termes et s'est effectuée du français vers l'anglais.

La deuxième étude qui a attiré notre attention nous vient de Gaussier (2001 : 167-183). Dans cette étude, le chercheur explore les règles de transformation des termes complexes comportant deux unités lexicales, en anglais et en français. Il se limite à l'examen de cette sorte de termes, car, explique-t-il, ils sont les plus fréquents et jouent un rôle de premier plan en terminologie puisque, très souvent, ils servent à construire des termes plus longs. Dans un tableau, qui présente les types de transformations formelles que subissent les termes lorsqu'ils passent d'une langue à l'autre, il donne le nombre d'occurrences observées à partir d'une petite portion du corpus sur lequel il a travaillé (Tableau 1.3).

Tableau 1.3 : Patrons de correspondances (d'après Gaussier 2001 : 173)

		ANGLAIS				
		N N	Adj N	N of N	N's N	N
FRANÇAIS	N de N	122	15	2	1	8
	N prép N	28	9	—	—	2
	N Adj	23	63	—	—	1
	N N	11	—	—	—	—
	N	1	1	—	—	—

Après avoir présenté les patrons de correspondance des termes anglais et français, Gaussier explique cependant que l'on ne peut se fier uniquement à ces patrons pour extraire des paires d'équivalents, car, pour un même patron, il peut exister plusieurs types de règles de transformation, par exemple, nous voyons dans le Tableau 1.3 que le patron français N de N peut être traduit en anglais selon les patrons suivants : N N, adj N, N of N, N's N, et N.

1.3.3 Un terme complexe en L_1 s'exprime par terme simple en L_2

Les études portant sur la transposition d'un terme complexe dans une langue en un terme simple dans une autre semblent rares. Dans le Tableau 1.3, Gaussier (2001 : 173-174) en rend compte partiellement, car son étude s'intéresse uniquement aux termes complexes. Ce phénomène semble pourtant assez courant comme en témoignent les exemples suivants : *relleno sanitario/landfill* (Carreño 2004 : 81); *terre humide/wetland*; *ice sheet/inlandsis*; *mot de passe/password*; *bande passante/bandwith, etc.*

On remarquera dans les exemples de termes simples qui précèdent qu'il existe des noms composés (compounds). Ces derniers sont constitués d'éléments qui peuvent être utilisés de façon autonome, par exemple dans le nom composé *wetland,* nous avons les mots *wet* et *land.* Les éléments des noms composés peuvent être accolés (*groundwater*), reliés par un trait d'union (*human-induced*) ou séparés par un blanc (*sea rise*) (Ahronian 2005). Il n'est pas rare de voir un nom composé dont les éléments sont accolés en L_1 avoir pour équivalent en L_2 un nom composé dont les éléments sont séparés par un blanc ou reliés par un trait d'union, par exemple le terme *wetland* est rendu par *terre humide* dans notre corpus. Du point de vue de l'extraction automatique, lorsqu'il y a des différences de structures entre des noms composés en L_1 et en L_2, cela pose des problèmes d'établissement d'équivalence.

1.3.4 Un terme dont la partie du discours en L_2 est différente de celle de la L_1

Comme expliqué au début de la section 1.3, dans le quatrième énoncé de L'Homme (2004 : 2008), il arrive qu'une unité lexicale d'une certaine nature grammaticale dans le texte de départ soit exprimée par une unité lexicale d'une autre nature dans le texte d'arrivée. À titre d'exemple, dans le corpus de Carreño (2004 : 88), le nom anglais (en caractères gras) *pollution* ***abatment*** a été traduit par un verbe espagnol (en caractère gras) *para* ***reducir*** *la contaminación.* Dans son étude portant sur la variation terminologique, Carreño catégorise ce type de phénomène comme une variation morphosyntaxique. Sur les 50 termes simples et complexes étudiés, 11 changent de nature grammaticale au moins une fois[11] lorsqu'ils sont traduits de l'anglais vers l'espagnol. Ce qui représente un pourcentage non négligeable de 22 % sur les termes étudiés (Carreño 2004).

1.3.5 Un terme en L_1 est traduit par une anaphore en L_2

La substitution de l'équivalent par une anaphore ou même par l'omission est également fréquente en traduction, surtout pour les langues qui privilégient ces techniques afin d'éviter les répétitions. Carreño (2004) considère ces deux cas comme de la variation syntaxique et les a inclus dans sa typologie de la variation. Sur les 50 termes simples et complexes étudiés, 30 termes subissent, au moins une fois[12] l'anaphore et/ou l'omission lors du passage d'une langue à l'autre, ce qui correspond à 60 % des termes étudiés (Carreño 2004).

---

[11] Le changement n'est pas systématique, il s'opère aléatoirement, selon les choix effectués par le traducteur au cours de la reformulation de la phrase en langue cible.

[12] Voir note précédente.

1.4 Conclusion

Dans ce chapitre, à la section 1.1, nous avons vu qu'il n'existe pas de consensus sur la définition du terme. Selon les optiques, les terminologues envisagent le terme sous différents angles. Par exemple, pour les uns, les termes sont principalement de nature nominale et pour les autres, les termes peuvent facilement appartenir à plusieurs catégories grammaticales.

Dans la section 1.2, nous avons abordé très succinctement l'optique conceptuelle de la terminologie et l'approche lexico-sémantique et nous avons décrit trois niveaux d'équivalence : l'équivalence exacte, l'équivalence partielle et l'absence d'équivalence. Ainsi, nous avons pu constater que, même en terminologie, l'équivalence pose des difficultés.

Dans la section 1.3, nous avons examiné de plus près les problèmes d'équivalence du point de vue de l'extraction automatique à partir des énoncés de L'Homme (2004), à savoir que dans les textes de la langue cible, les termes de la langue source ne sont pas toujours traduits par les équivalents attendus. En langue cible, le traducteur est libre d'utiliser des variantes (synonymes) ou un terme de nature grammaticale différente de celle du terme de la langue source, d'employer une anaphore ou même d'omettre le terme. Nous avons vu aussi que la structure et la longueur des termes sont rarement les mêmes d'une langue à l'autre. À la suite de cet examen, nous avons constaté que s'il est possible de dégager des règles générales, elles ne peuvent être appliquées systématiquement.

À la lumière de ces observations, il nous semble important et nécessaire dans le cadre de cette étude d'énoncer les deux points essentiels suivants :

Premièrement, compte tenu du fait que nous travaillons à partir de corpus, nous prenons pour modèles l'approche de la *terminologie textuelle* (Bourigault et Slodzian 1999) et l'optique lexico-sémantique (L'Homme 2004 : 52-82) pour définir le terme :

1. le terme est une construction linguistique qui fait référence à une signification particulière dans un domaine donné;
2. le terme ne se limite pas à l'unité lexicale nominale, il peut également appartenir à la partie du discours de l'adjectif, du verbe et de l'adverbe;
3. le terme complexe n'est retenu que s'il est non compositionnel[13];
4. le terme est choisi selon un objectif précis[14].

Deuxièmement, si pour nous les principes d'équivalence en terminologie évoqués à la section 1.2 sont d'un point de vue théorique instructifs et nécessaires, en extraction automatique de termes, il nous faut inévitablement nous orienter vers des méthodes adaptées aux outils informatiques qui tiennent compte des problèmes sous une perspective plus pratique, plus formelle. Ainsi, afin de mieux comprendre les difficultés liées à l'établissement de l'équivalence, nous nous inspireront des problèmes d'établissement d'équivalence automatique examinés à la section 1.3. Pour cela, nous recombinerons la plupart des éléments qui s'y trouvent et les intégrerons à notre classification des candidats termes selon leurs équivalents. Comme notre travail se situe en amont de la validation, nous entendons par « équivalence », l'équivalence telle qu'observée en contexte et non pas l'équivalence que l'on pourrait trouver dans les dictionnaires.

Par conséquent, dans cette étude, afin d'étudier l'équivalence entre candidats termes (CT) anglais et français, nous nous proposons de constituer un corpus parallèle dans les deux langues et appartenant au domaine du changement climatique. Ensuite, nous alignerons ce corpus au niveau des phrases avec Alinea. Nous extrairons, dans les deux langues, des candidats termes simples appartenant aux parties du discours du nom, de

---

[13] Termes complexes n'ayant pas un sens compositionnel : *traitement de texte, système d'exploitation* (d'après L'Homme 2005 : 1126)

[14] Autrement dit, c'est le type d'ouvrage, de projet qui va guider le terminologue à déterminer le type de terme à inclure dans sa nomenclature.

l'adjectif, du verbe et de l'adverbe à l'aide de l'outil d'acquisition de termes TermoStat. Toutefois, avant de continuer, nous aimerions préciser que ces candidats termes sont également susceptibles d'être des têtes ou des modificateurs de syntagme ou encore des non-termes.

En résumé, ce mémoire se propose de répondre aux points énumérés ci-après :

- vérifier ce qui se dégage de notre analyse par rapport aux résultats de la liste des problèmes d'établissement des équivalences de L'Homme (2004 : 208) à l'aide de la classification des candidats termes et de leurs équivalents établie par nous;
- comparer un échantillon de la liste d'extraction de termes français à la liste d'extraction de termes anglais de TermoStat pour vérifier si tous les équivalents anglais sont présents dans la liste d'extraction et pour observer le rang qu'ils occupent par rapport aux CT français;
- comparer la liste des équivalents compilés manuellement avec une extraction lexicale d'Alinea afin d'analyser le bruit et le silence généré par ce dernier.

Nous espérons ainsi que notre analyse manuelle mette en évidence certaines des limites de l'appariement de CT qu'une évaluation directe des résultats n'aurait pas permis de révéler.

Chapitre 2 : Corpus parallèles, alignement et extraction de termes

Pour constituer des ressources terminologiques bilingues à partir de corpus parallèles, des outils d'acquisition automatique de termes sont généralement combinés à des aligneurs de textes. Dans le présent chapitre, avant de nous pencher sur les systèmes d'extraction bilingue, nous avons jugé opportun de décrire dans les grandes lignes, à la section 2.1, les différentes techniques d'alignement au niveau des phrases et au niveau des mots. Puis de poursuivre, à la section 2.2, par un rapide tour d'horizon des méthodes d'extraction de termes. Finalement, à la section 2.3, partie centrale de ce chapitre, nous explorons en détail les méthodes d'extraction bilingue à partir de corpus parallèles. Pour terminer, à la section 2.4, nous récapitulons brièvement le Chapitre 2.

2.1 Alignement automatique de textes parallèles

De nombreux textes parallèles ont jalonné l'histoire : traités, contrats, textes sacrés ou littéraires. Mais, le plus connu est sans conteste la pierre de Rosette. Ce fragment de stèle en granite noir sur lequel figuraient trois systèmes d'écriture, les hiéroglyphes, le démotique et le grec, a permis à Champollion de déchiffrer l'écriture hiéroglyphique. De nos jours, grâce à l'informatique, les textes parallèles peuvent être exploités plus facilement et à plus grande échelle. Selon Véronis (2000 : 152), dès les années 1950, on avait déjà pensé les utiliser en traduction automatique, mais à cette époque les ordinateurs n'étaient pas encore suffisamment puissants. Ce n'est qu'au cours des années 1980 que la première méthode automatique d'alignement de textes parallèles a été développée par Kay et Röscheisen (1988). Depuis, de nombreux chercheurs se sont intéressés à l'alignement de textes.

2.1.1 Alignement au niveau des phrases

Plusieurs techniques sont employées pour aligner les textes au niveau des phrases. Les premières sont celles mises au point par Kay et Röscheisen (1988), elles se

caractérisent par l'utilisation exclusive d'informations internes comme la distribution lexicale :

> We present an algorithm for aligning texts with their translations that is based only on internal evidence. The relaxation process rests on a notion of which word in one text corresponds to which word in the other text that is essentially based on the similarity of their distributions (Kay et Röscheisen 1993 : 121)[15].

Les auteurs montrent qu'il est possible d'aligner les textes au niveau des phrases en exploitant la similarité de distribution de certains mots (noms propres, dates, etc.) à l'intérieur de zones[16] qui se correspondent dans les deux textes. Et comme le note Kraif, ces techniques « ont le grand mérite [...] de montrer qu'il est possible d'aligner sans passer par le sens, en se basant sur des propriétés purement formelles » (Kraif 2001 : 227). À peu près à la même époque, Gale et Church (1991) et Brown *et al.* (1991) présentent une autre méthode reposant également sur des informations internes, mais à la différence de la première, celle-ci repose sur l'observation de la longueur des phrases. Gale et Church se basent sur le nombre de caractères contenus dans une phrase : « [...] a method and a program (align) for aligning sentences based on a simple statistical model of character lengths » (Gale et Church 1991 : 75); alors que Brown *et al.* (1991) s'appuient sur le nombre de mots dans la phrase. À la suite de ces travaux, d'autres chercheurs, pour améliorer la technique, ont introduit de nouvelles notions telles que les cognats[17] (Simard *et al.* 1992; Church 1993; Langlais 1997; Kraif 1999).

Si les techniques d'alignement dont on vient de parler donnent de très bons résultats pour les langues comme l'anglais et le français, il n'en va pas ainsi pour d'autres langues.

---

[15] Le texte de 1988 qui décrit la méthode de Kay et Röscheisen n'étant pas accessible, nous avons eu recours au texte de 1993.

[16] Par exemple, le début et la fin des textes sont des zones, qui en principe, se correspondent.

[17] Cognats : « (de l'anglais *cognate*), des mots qui se traduisent l'un par l'autre et qui présentent une ressemblance graphique » (Kraif 1999 : 205).

Certaines langues indo-européennes comme l'allemand sont plus difficiles à aligner (Chiao *et al.* 2006). Les langues asiatiques, quant à elles, posent des problèmes encore plus complexes. Par exemple, « le système d'écriture du japonais ne dispose pas de séparateur graphique indiquant les frontières entre les mots » (Nakamura-Delloye 2005). Par conséquent, les chercheurs ont dû recourir à des ressources externes en intégrant des analyseurs morphologiques et des dictionnaires bilingues dans leurs systèmes d'alignement. Enfin, il va sans dire que toutes les méthodes dont nous venons de parler ne donnent pas de bons résultats avec des textes parallèles présentant des différences structurelles : paragraphe manquant, ajout, inversion, etc.

2.1.2 Alignement au niveau des mots

L'alignement au niveau des mots, qui consiste à « repérer les mots et expressions du texte source et du texte cible, puis de les mettre en correspondance » (Véronis 2000 : 163), est encore plus difficile à réaliser que l'alignement au niveau des phrases à cause de nombreux facteurs (différence de structure des syntagmes, particularités grammaticales propres à chaque langue, locution, phraséologismes, etc.). À la Figure 2.1, nous illustrons quelques-uns des problèmes reliés à l'alignement au niveau des mots dans un couple de phrases tiré de notre corpus, *Changement climatique*. On y voit, entre autres, que le terme composé anglais *wetland* se traduit à l'aide de deux unités lexicales en français *terres humides* et que les mots grammaticaux français *la, des, des, l', du* n'ont pas de contrepartie en anglais.

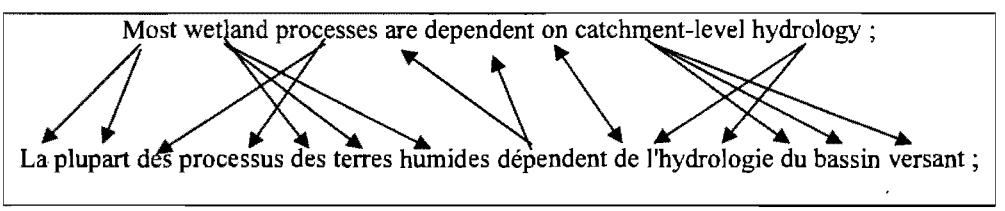


Figure 2.1 : Illustration de quelques problèmes reliés à l'alignement au niveau des mots

Les stratégies adoptées pour l'alignement au niveau des mots, des expressions[18] et des syntagmes nominaux dépendent du type de tâche que l'on veut accomplir : traduction automatique, conception de lexiques multilingues, extraction terminologique, etc. Pour les deux premières tâches, par exemple, on s'intéresse généralement à apparier le plus de mots et d'expressions possible dans un texte alors que pour la troisième tâche, comme nous le verrons à la section 2.3, on cherche à isoler les termes spécifiques à un domaine donné, tout particulièrement les syntagmes nominaux.

Selon Ozdowska et Claveau (2005) et Tiedemann (2003), il existe globalement deux approches pour aligner les mots : celle « à dominante statistique qui s'appuie notamment sur les modèles IBM (Brown *et al.* 1993) » (Ozdowska et Claveau 2005 : 2) et celle qui met en œuvre des heuristiques statistiques et des ressources linguistiques (Figure 2.2). Dans le cadre de cette étude, nous ne décrirons pas les méthodes d'alignement des approches à dominante statistique, car elles touchent principalement la traduction automatique (exemple, traduction automatique fondée sur l'exemple). Qu'il suffise de dire que les travaux fondateurs de Brown *et al.* (1993) ont été une source d'inspiration pour de nombreux chercheurs, tant dans la branche utilisant l'approche à dominante statistique (Och 1999; Nevado *et al.* 2003; Moore 2005) que dans la branche ayant recours aux heuristiques statistiques et aux ressources linguistiques (Dagan *et al.* 1993; Melamed 1997). Cette dernière branche, est celle qui nous intéresse, car elle s'est surtout employée à extraire des lexiques (Church *et al.* 1991; Fung et Church 1994; Tiedemann 1997; Kraif 2002).

---

[18] Segments de phrases, par exemple, *les mains dans les poches* et *with his hands in his pokets*.

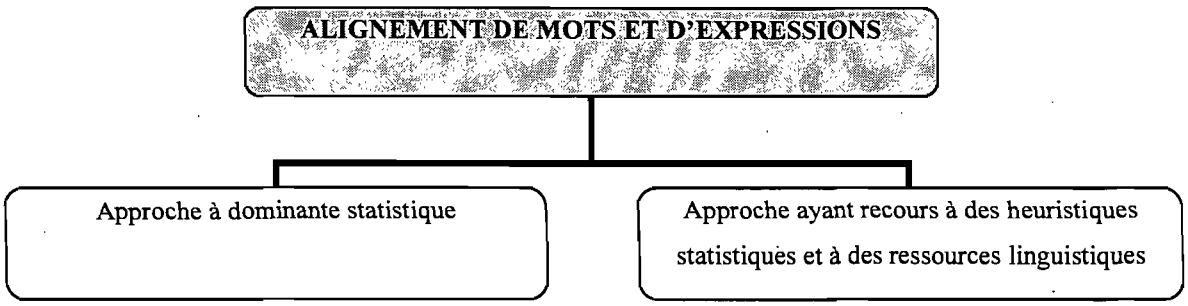


Figure 2.2 : Deux approches pour aligner au niveau des mots

À la section 2.3, nous décrivons plus en détail les travaux spécifiquement dédiés à l'extraction de termes, toutefois il nous semble opportun de présenter auparavant certaines des stratégies employées par de nombreux chercheurs dans la branche ayant recours à des heuristiques statistiques et à des ressources linguistiques. Ces stratégies, également utilisées dans les travaux d'extraction bilingue de termes, font appel à toute une batterie d'indices d'association telles la cooccurrence parallèle[19], la cognation, la position des mots dans la phrase et les parties du discours.

2.1.2.1 Modèles de cooccurrence parallèle

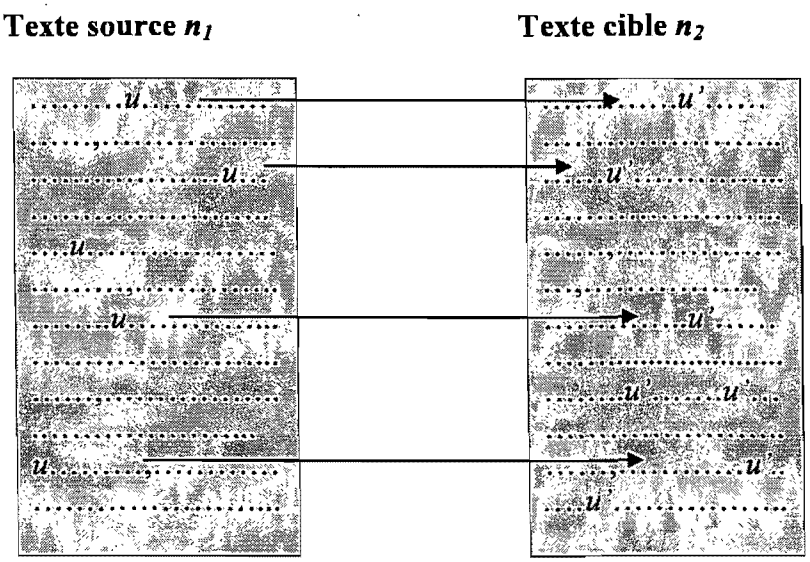
Les modèles de cooccurrence parallèle sont parmi les approches les plus exploitées pour extraire des correspondances lexicales au sein d'un corpus parallèle (Fung et Church 1994; Gaussier et Lange 1995; Melamed 1997; Kraif 2000). À partir de l'observation de la distribution des occurrences et des cooccurrences parallèles des unités à l'intérieur d'un corpus parallèle, divers calculs peuvent être appliqués : l'information mutuelle (Church et Hanks 1990), le *t-score* (Fung et Church 1994); le rapport de vraisemblance (*log-likelihood*

[19] L'expression *cooccurrence parallèle* est prise dans Kraif (2001 : 400) : « Afin qu'il n'y ait pas de confusion avec les cooccurrences unilingues, à l'intérieur de chaque moitié du bi-texte, nous désignons par cooccurrence parallèle les cooccurrences entre deux parties alignées du bi-texte. »

ratio) (Dunning 1993) et la probabilité de l'hypothèse nulle (Kraif 2000), parmi d'autres. Toutes ces mesures se basent sur le même principe : « si le nombre de cooccurrences observé est très supérieur au nombre de cooccurrences estimé dans le cas d'une distribution aléatoire, c'est qu'il y a sans doute un lien entre les deux unités » (Kraif 2001 : 408).

Nous avons choisi de présenter l'indice décrit dans les travaux de Kraif (2001) et de Kraif et Chen (2004) puisque c'est celui qui est implémenté dans Alinea, l'aligneur utilisé pour cette étude[20]. Pour extraire les correspondances lexicales, Kraif se base donc sur l'indice qu'il nomme *la probabilité de l'hypothèse nulle (P0)*. Dans un texte source (n_1), le logiciel cherche toutes les occurrences d'une chaîne de caractères (u); puis il cherche dans le texte cible (n_2) toutes les occurrences d'une autre chaîne de caractères (u'), ensuite il calcule le nombre de fois, dans le texte source et dans le texte cible (n_{12}), où les occurrences u et u' sont en correspondance dans un couple de phrases alignées, c'est-à-dire en cooccurrence parallèle. Sur la base de ce nombre, il fait l'hypothèse que si les cooccurrences de u et de u' dépassent le nombre de cooccurrences attendues par le simple jeu du hasard (P0), alors u et u' sont fort probablement des équivalences traductionnelles. Dans la Figure 2.3, nous illustrons ce qui précède.

[20] Le système d'alignement Alinea est décrit plus en détail au Chapitre 3, Section 3.1.7.



$n_1 = 5, n_2 = 7, n_{12} = 4$

Figure 2.3 : Occurrences et cooccurrences de deux unités (Kraif 2001 : 261)

Afin d'extraire les correspondances lexicales, Alinea doit calculer P0 à partir d'un corpus parallèle de très grande taille (au moins 1 million de mots) d'où il extrait des statistiques d'occurrences et de cooccurrences.

2.1.2.2 La cognation

La cognation, autre indice d'association très utilisé, est fondée sur la ressemblance graphique entre mots d'un texte source et d'un texte cible. Les cognats, comme on les appelle, sont particulièrement nombreux entre langues apparentées comme le français et l'anglais. Ces ressemblances ne sont pas que formelles, car il a été observé que très souvent les cognats sont également comparables sémantiquement, donc susceptibles d'être des équivalents traductionnels.

Les cognats peuvent se présenter sous diverses formes. Simard *et al.* (1992) ont établi les catégories suivantes :

1. les couples de tokens[21] alphanumériques identiques contenant au moins un chiffre (CO_2, 2007);
2. les couples de tokens dont, au moins, les quatre premières lettres sont identiques (*terrestre, terrestrial*);
3. les couples de ponctuation semblables (= { } ( ) * $ £ ! ?).

Ainsi, de nombreux cognats peuvent être identifiés si, à l'exemple de la catégorie 2, au moins les quatre premiers caractères (4-grammes) d'un couple de tokens sont identiques, comme dans *solaire* et *solar*. Cependant, cette méthode, bien que simple et efficace dans de nombreux cas, ne permet pas de découvrir les cognats tels que *précipitation* et *precipitation*[22]. Pour parvenir à identifier les cognats de ce type, les indices d'association ont été affinés grâce, en particulier, à l'indice de la sous-chaîne maximale (SCM) (Kraif 1999) ou à une variante du coefficient de Dice (Brew et McKelvie 1996). En ce qui concerne la méthode de calcul avec la SCM on considère comme cognat tout couple de mots présentant une sous-chaîne commune d'une longueur au moins égale à un certain pourcentage[23] de la longueur du mot le plus long. Par exemple entre *précipitation* et *precipitation*, la plus longue sous-chaîne commune est ***p-r-c-i-p-i-t-a-t-i-o-n***, soit 12 caractères sur 13 ou 92 % de *precipitation*. Maintenant, en réutilisant le même couple de mots afin d'y appliquer le coefficient de Dice. Les sous-chaînes communes sont tout d'abord décomposées en bigrammes[24] : ***pr-ci-ip-pi-it-ta-at-ti-io-on.*** Ces bigrammes sont ensuite utilisés pour effectuer le calcul montré à la Figure 2.4.

---

[21] Un *token* est une chaîne de caractères délimitée par des espaces.

[22] La différence se situe au niveau de l'accentuation du mot français.

[23] Alinea considère, par défaut, tout couple de mots comme des cognats s'il « peut en extraire une sous-chaîne commune d'une longueur au moins égale à 66 % du mot le plus long » (Kraif 2007).

[24] Bigramme : groupe de deux mots, de deux syllabes ou de deux caractères.

$$\text{Dice} = \frac{2 * (pr\text{-}ci\text{-}ip\text{-}pi\text{-}it\text{-}ta\text{-}at\text{-}ti\text{-}io\text{-}on)}{13 + 13} = \frac{20}{26} = 0{,}77$$

Figure 2.4 : Calcul avec la formule modifiée de Dice (d'après Tiedemann 2003 : 17)

Avant de terminer, il est important de rappeler que pour limiter le bruit certaines précautions sont à prendre. Premièrement, en règle générale, on ne considère que les mots de 4 lettres et plus. Deuxièmement, il faut établir un seuil en deçà duquel les résultats ne sont pas considérés.

2.1.2.3 Position des mots dans la phrase

Pour améliorer l'alignement au niveau des mots, divers indices d'association peuvent être combinés. Par exemple, Ahrenberg *et al.* (1998), à partir des règles qui régissent la position des mots dans la phrase[25], attribuent un poids à chacun des couples de mots relevés à partir d'une paire de phrases alignées. Un couple dont la position est relativement identique par rapport aux mots qui les entourent de part et d'autre de l'alignement reçoit un score plus élevé qu'un couple présentant un écart de position important. L'indice de position des mots dans la phrase est considéré comme suffisamment fiable pour être utilisé dans de nombreux travaux (Kraif et Chen 2004; Tiedemann 2003).

2.1.2.4 Parties du discours

Pareillement, les parties du discours peuvent servir d'indices complémentaires pour améliorer l'alignement au niveau des mots. Étant donné que les équivalents traductionnels appartiennent souvent à la même partie du discours, Melamed (1995) explique qu'en se basant sur cette observation, il est possible, par exemple, d'implémenter un algorithme

[25] Pour certaines langues, comme l'anglais et le français, la position des mots n'est pas toujours libre dans la phrase, il existe un certain ordre entre eux.

permettant de déceler les couples possédant la même partie du discours et les couples appartenant à des catégories différentes.

Notre description des indices d'association prend fin ici, même s'il en existe plusieurs autres. Nous pensons par exemple à la distribution des mots dans le texte, aux dictionnaires bilingues accessibles par ordinateur (*machine readable bilingual dictionary*) ou aux informations syntaxiques.

2.2 Acquisition automatique de termes

Afin de faciliter la tâche des terminologues, les outils d'acquisition automatique de termes repèrent dans les corpus des chaînes de caractères ou des suites de chaînes de caractères susceptibles d'être des termes. Le progiciel TERMINO (David et Plante 1990) est le premier outil d'acquisition automatique de termes à avoir vu le jour. Parallèlement, à la même période ou même un peu avant, des chercheurs (Choueka 1988; Lebart et Salem 1988; Church et Hanks 1990) ont mis au point différents systèmes statistiques de repérage de collocations dans les corpus. Ces techniques, bien qu'elles ne soient pas spécifiquement dédiées à l'extraction de termes, ont inspiré de nombreux travaux terminologiques. Par exemple, le calcul des segments répétés au sein d'un corpus brut (Lebart et Salem 1988) a servi d'amorce à plusieurs travaux portant sur la terminologie.

Plusieurs auteurs (Cabré *et al.* 2001; Drouin 2002; Carreño *et al.* 2006) classent les méthodes d'acquisition de termes selon trois grandes catégories : linguistique, statistique, hybride (linguistique et statistique). Pour simplifier les choses, nous utilisons la même classification. Toutefois, il est à noter que ce découpage est à nuancer, car la plupart des outils font appel à plus d'une méthode.

2.2.1 Méthode linguistique

Parmi les méthodes linguistiques utilisées pour extraire les termes, nous en présentons deux qui nous semblent les plus représentatives.

David et Plante (1990), créateurs de TERMINO (maintenant connu sous le nom de Nomino), font appel, entre autres, à des règles de désambiguïsation lexico-syntaxiques pour extraire des syntagmes nominaux. Après avoir repéré les noms et leur expansion[26], le module procède à une désambiguïsation, puis il propose une liste de candidats jugés valides. Bourigault (1994), développeur de LEXTER, adopte quant à lui une stratégie différente. Une fois l'analyse syntaxique de surface effectuée, les termes sont isolés par des frontières, c'est-à-dire des éléments ne pouvant pas faire partie d'un terme : verbe conjugué, adverbe, conjonction, pronom, etc.

2.2.2 Méthode statistique

Afin d'identifier les termes, les approches à dominante statistique mettent en oeuvre diverses stratégies. L'outil ANA (Enguehard 1992), par exemple, détecte à partir de deux termes connus les associations récurrentes, ou encore, repère la cooccurrence d'un mot avec un terme connu. À l'aide de ce qu'il appelle le *co-efficient of weirdness,* Ahmad (1996) mesure la fréquence relative d'une chaîne de caractères dans un corpus spécialisé par rapport à un corpus de langue générale. Avec l'indice *C-value,* Frantzi (1998), quant à elle, mesure la fréquence d'apparition d'un candidat terme (CT) dans le corpus, la fréquence à laquelle il apparaît au sein de candidats termes plus longs, le nombre de ces candidats termes plus longs et la longueur, en terme de mots, du candidat terme.

---

[26] Toute chaîne de caractères située après un nom et avant une frontière syntaxiquement marquée.

2.2.3 Méthode hybride

Rapidement, les chercheurs se sont orientés vers des méthodes d'extraction hybrides. Comme son nom l'indique, cette méthode combine les approches statistique et linguistique. Avec XTRACT[27], Smadja (1993) commence par les étapes statistiques avant de passer à l'étape linguistique : extraction de couples de mots qui se rencontrent fréquemment, extension de ces couples par identification d'enchaînements plus longs et enfin étiquetage des collocations au sein desquelles on peut trouver des termes. À l'inverse de Smadja, Daille (1994), créatrice du logiciel ACABIT, commence par identifier des candidats termes à l'aide de patrons morphosyntaxiques[28] sur un corpus préalablement étiqueté, la liste de CT obtenue est ensuite soumise à une série de filtres statistiques destinés à observer différentes propriétés : fréquence, critères d'association, diversité et mesures de distance. Pour extraire les candidats termes, Frantzi et Ananiadiou (1999), utilisent aussi la méthode des patrons morphosyntaxiques, mais elles se servent en plus d'une liste de suffixes spécifiques au domaine étudié. La valeur terminologique des CT est ensuite calculée selon la *C-value* évoquée en 2.2.2. TERMOSTAT, le logiciel conçu par Drouin (2002) que nous décrirons plus amplement dans la section 3.2.1, extrait tout d'abord des CT au moyen de patrons morphosyntaxiques. Après quoi, en s'appuyant sur une approche contrastive (mise en opposition de corpus spécialisés et non spécialisés) et à l'aide du calcul des spécificités, le système propose des candidats termes propres au corpus étudié. Enfin, le système TERMINATOR mis au point par Patry et Langlais (2005) se

---

[27] Bien que conçu pour l'extraction des collocations, XTract peut être utilisé pour l'acquisition de termes.

[28] Pour identifier les termes complexes, certains logiciels repèrent des combinaisons de mots qui correspondent à des patrons morphosyntaxiques fondés sur la formation typique des syntagmes nominaux. Ainsi, en français, les termes complexes se composent le plus souvent d'un nom suivi d'un adjectif (*changement climatique*), d'un nom suivi d'un nom (*gaz traceur*), d'un nom suivi d'une préposition et d'un nom (*carotte de glace*), etc.

distingue des extracteurs précédents par le recours à un corpus d'apprentissage et par l'exploitation d'un algorithme d'apprentissage automatique. L'extraction s'effectue en quatre phases : 1) entraînement sur corpus d'apprentissage; 2) étiquetage du corpus d'analyse; 3) extraction des candidats termes; 4) attribution d'un score et classification des CT.

Les méthodes d'extraction qui viennent d'être évoquées très succinctement nous permettent déjà de comprendre que les approches à dominante linguistique ou statistique n'ayant pas toujours été suffisantes à elles seules nous sommes passés aux méthodes hybrides. Les systèmes linguistiques sont dépendants de la langue et, comme ils sont généralement conçus pour une langue en particulier, il est difficile de les adapter aux besoins d'un autre code linguistique. Les systèmes statistiques, quant à eux, en tentant de rester indépendants des langues, ne prennent pas en compte la nature linguistique des phénomènes observés. Enfin, si les systèmes hybrides ont grandement amélioré les performances des extracteurs, ils ne sont pas encore parfaits, ils sont également dépendants de la langue et le bruit et le silence restent encore importants.

2.3 Extraction bilingue de termes à partir de corpus parallèles

L'extraction de termes bilingue fait appel aux applications décrites dans les sections 2.1 et 2.2, soit l'alignement au niveau des phrases et au niveau des mots et l'acquisition automatique de termes. Les méthodes d'extraction bilingue de termes se distinguent des autres types d'alignement au niveau des mots du fait que l'on cherche à isoler les termes spécifiques au corpus étudié. Sous certains aspects, c'est un avantage, car, comme on l'a vu dans la Figure 2.1, tous les mots d'une phrase ne peuvent être alignés. Toutefois, sous d'autres aspects, c'est un inconvénient. En effet, l'identification des termes et leur découpage posent des difficultés.

Selon Gaussier (2001) et Ozdowska et Bourigault (2004), les méthodes d'extraction bilingue de termes se divisent elles-mêmes selon trois approches. Pour les illustrer, nous avons repris la Figure 2.2, et à partir de la branche ayant recours à des heuristiques statistiques et à des ressources linguistiques, nous avons ajouté la branche *Alignement de termes*[29] et ses sous-divisions (Figure 2.5). Dans la première approche (I), la plus classique, il s'agit, dans un premier temps, de repérer les termes dans le texte source et dans le texte cible, puis, dans un deuxième temps, de les apparier (Daille *et al.* 1994). Dans la deuxième approche (II), l'extraction des termes ne se fait que dans une langue, les termes de l'autre langue sont quant à eux identifiés au moment de l'alignement (Gaussier 1998). Enfin, dans la troisième approche (III), on commence par effectuer une analyse syntaxique parallèle des textes source et cible, puis on procède à ce que les auteurs nomment *identification simultanée des termes* (Wu 2000; Ozdowska et Bourigault 2004). Ici, nous aimerions faire remarquer que la division simplifiée que nous proposons des approches (Figure 2.5), tant pour l'alignement de mots et d'expressions que pour l'alignement de termes, est purement conventionnelle, dans la réalité leur découpage est beaucoup moins net.

[29] Expression empruntée à Bourigault et Jacquemin (2000 : 230).

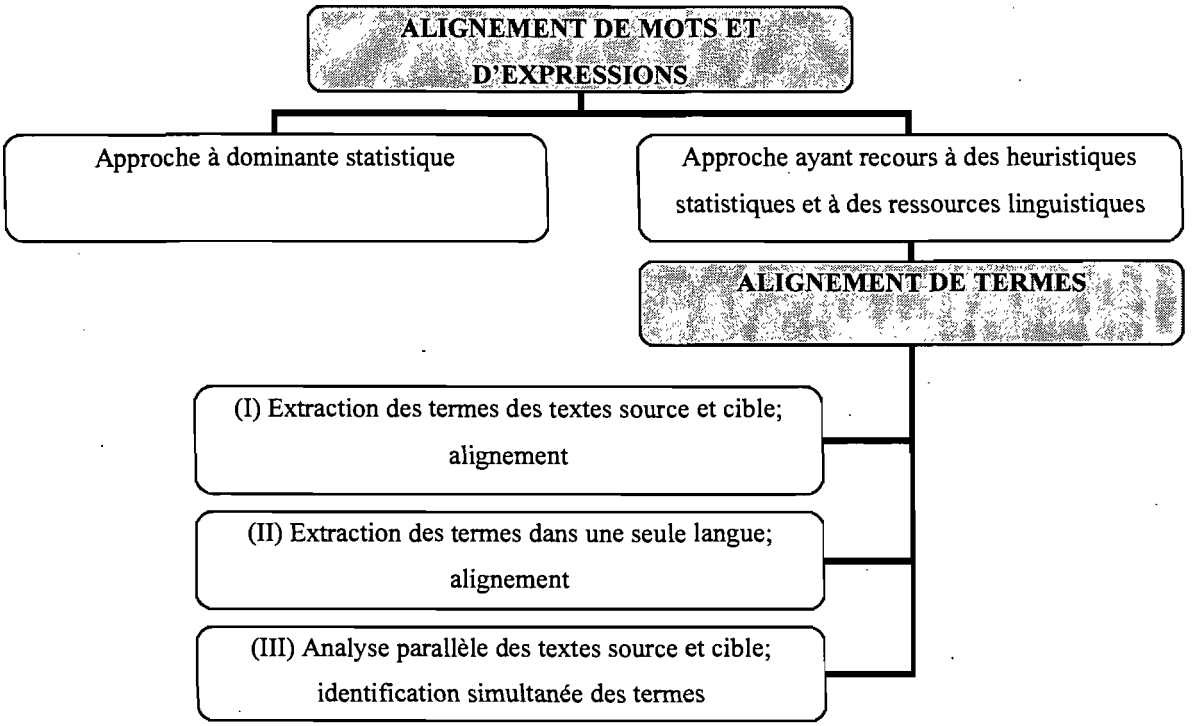


Figure 2.5 : Graphique des types d'alignement de mots et de termes

Dans les sections qui suivent, nous présentons par ordre chronologique des travaux sur l'extraction de termes bilingues en ne citant que ceux qui exploitent les corpus parallèles.

2.3.1 Van der Eijk (1993)

Dès l'introduction, van der Eijk (1993) annonce clairement l'objet de son étude : automatiser au maximum l'acquisition de listes bilingues de termes à partir de corpus parallèles. La méthode employée, qui se conforme à l'approche I, se déroule en deux étapes : prétraitement linguistique des textes sources et cibles afin d'extraire des termes, sélection statistique des syntagmes nominaux identifiés au prétraitement.

Avant d'effectuer l'extraction de termes, le corpus anglais et néerlandais de 25 000 mots est soumis à un prétraitement. Il est tout d'abord analysé lexicalement et

aligné, au niveau des phrases, suivant la méthode de Gale et Church (1991). Puis, un étiquetage et une analyse syntaxique de surface sont effectués. Les candidats termes recherchés sont des syntagmes nominaux dont les patrons sont basés sur l'anglais. Un algorithme de filtrage (*pattern matching algorithm*) considère comme terme potentiel toute séquence de zéro ou plus adjectif suivi par un nom ou plus (ainsi, des termes simples et complexes peuvent être identifiés). Van der Eijk a pris le parti de ne pas tenir compte des compléments et des modificateurs qui suivent un nom.

Afin d'extraire les syntagmes nominaux identifiés à l'étape précédente, des statistiques de cooccurrences sont mises en oeuvre (section 2.1.2.1). Pour atténuer le bruit, un seuil minimum et une mesure de la position des mots dans la phrase sont fixés (section 2.1.2.3).

Dans son analyse des résultats, l'auteur fait état du rappel[30] très faible qu'il attribue principalement à la différence de structure des deux langues. Étant donné que le niveau d'agglutination des composés est très différent entre l'anglais et le néerlandais, il est difficile d'apparier des CT tels que *programme management* et *programmabeheer* ou encore *high speed data processing capability* et *snelle gevensverwerkingscapaciteit.* Plusieurs variantes de la méthode ont été testées. L'alignement des syntagmes a donné de meilleurs résultats que l'appariement des termes simples.

2.3.2 Dagan et Church (1994)

Dagan et Church (1994) présentent TERMIGHT comme un outil semi-automatique utile aux traducteurs et aux terminologues pour identifier des candidats termes et leur

---

[30] Le rappel est égal au nombre d'alignements corrects par rapport au nombre d'alignements de référence.

traduction. La méthode proposée s'apparente à l'approche II puisque seuls les termes source sont extraits à la première étape.

L'outil d'extraction repose sur l'étiquetage pour identifier les termes simples et complexes. Les termes complexes, de forme nominale, sont extraits à partir de patrons catégoriels qui peuvent être modifiés par l'utilisateur. En ce qui concerne l'extraction des termes simples, tous les mots ne faisant pas partie d'une liste d'exclusion sont considérés. Les candidats termes ainsi recueillis sont regroupés sous une même tête de syntagmes. Puis chacun des groupes est trié par ordre de fréquence décroissant. Les concepteurs, qui concèdent que cette façon d'opérer génère beaucoup de bruit, ont prévu une interface dans laquelle l'utilisateur peut procéder à une validation.

En ce qui concerne l'extraction des équivalents, les auteurs émettent l'hypothèse qu'un alignement du corpus au niveau des mots est mieux adapté. La composante bilingue de TERMIGHT utilise *word_align* (Dagan *et al.* 1993). Ce système aligne les textes au niveau des mots en s'appuyant sur les données de sortie de *char_align*[31] (Church 1993) sur lesquelles il applique un algorithme inspiré du modèle 2 de Brown *et al.* (1993). En se basant sur l'alignement des mots, TERMIGHT identifie les traductions candidates de chaque terme source (un terme source peut avoir plusieurs traductions candidates). Ensuite, il présente les traductions candidates de chaque terme source par ordre décroissant de fréquence. À l'aide du concordancier bilingue de TERMIGHT, l'utilisateur a le choix de valider les candidats termes ou de chercher les traductions manquantes dans les concordances.

---

[31] *Char_align,* système d'alignement au niveau des phrases, se base sur l'observation de la longueur des phrases au niveau des caractères et les ressemblances superficielles (cognats).

Une évaluation du système a donné les résultats suivants : la bonne réponse était située à 40 % dans le premier choix des traductions candidates et à 7 % dans le deuxième. Les 53 % de bonnes réponses manquantes, se trouvaient quant à eux dans les concordances.

2.3.3 Daille, Gaussier et Langé (1994)

Selon Daille *et al.* (1994), dans le domaine technique, les termes les plus courants sont des syntagmes nominaux qui se traduisent par un nombre limité de patrons syntaxiques. Sur le principe de l'approche I, ils proposent un système qui fait appel à des connaissances linguistiques et à des modèles statistiques pour extraire une terminologie bilingue.

À partir d'un corpus étiqueté, les candidats termes anglais et français sont extraits à l'aide de patrons syntaxiques de base (deux unités lexicales). Les patrons retenus en français sont les suivants : nom + adjectif (*orbite géostatique*), nom1 + nom2 (*diode tunnel*), nom1 + de (dét.) + nom2 (*bande de fréquence*), nom1 + préposition (dét.) + nom2 (*assignation à la demande*). En anglais, les patrons sont : adjectif + nom (*multiple access*) et nom1 + nom2 (*data transmission*). À ces patrons de base, les concepteurs autorisent des variantes qui permettent de rendre compte de la surcomposition, de la modification et de la coordination. Les listes obtenues à partir des patrons morphosyntaxiques ne sont toutefois pas exemptes d'erreurs. Pour atténuer le bruit, plusieurs filtres ont été testés, les meilleurs résultats ont été obtenus avec la mesure de fréquence.

Les CT sont appariés sur la base de techniques d'association des correspondances combinées à des méthodes d'identification des correspondances de patrons linguistiques.

Les auteurs annoncent un taux de précision[32] d'environ 70 % sur une liste de plus de 1 000 candidats termes.

2.3.4 Gaussier (1998)

La démarche de Gaussier (1998) repose sur l'approche II, les candidats termes ne sont extraits qu'à partir d'une seule langue. Dans son article, l'auteur rappelle que l'approche I s'appuie sur l'idée que les termes complexes se correspondent d'une langue à l'autre, ce qui ne se vérifie malheureusement pas toujours[33]. Cette constatation expliquerait les faiblesses de l'approche I. Ensuite, il émet l'hypothèse que les candidats termes sont plus faciles à repérer en anglais (hypothèse principalement basée sur le fait qu'en français on utilise plus souvent qu'en anglais des mots grammaticaux pour construire des syntagmes nominaux).

Pour extraire les CT anglais, un corpus parallèle de 1 000 paires de phrases a été étiqueté et lemmatisé. Les candidats termes anglais sont repérés à l'aide de patrons morphosyntaxiques.

Les termes français sont ensuite « devinés » au cours de l'alignement. Pour y parvenir, Gaussier utilise une technique basée sur les graphes qu'il appelle *alignment flow networks*. Cette méthode, qui lui permet d'intégrer différents types de contraintes et de paramètres, consiste à découvrir le meilleur alignement des mots à l'intérieur d'une phrase. La méthode, se distingue des autres par les trois points suivants : 1) elle ne se base pas sur des patrons pour identifier les équivalents; 2) elle considère toute la phrase pour permettre une certaine désambiguïsation; et 3) les probabilités d'association sont calculées à partir de l'ensemble du syntagme ou de chacun de ses constituants.

---

[32] La précision est égale au nombre d'alignements corrects par rapport au nombre d'alignements proposés.

[33] Voir Tableau 1.3 : Patrons de correspondances et de non-correspondances (d'après Gaussier 2001 : 173)

2.3.5 Hull (2001)

À l'instar de Gaussier, Hull (2001) rejette l'approche I pour extraire les termes en évoquant les mêmes raisons que lui, mais aussi parce que, selon l'auteur, les outils d'analyse et d'extraction de termes ne sont pas de qualité identique en anglais et en français.

TRINITY, un système qui fonctionnait suivant l'approche I, a donc été modifié pour être adapté sur le modèle de l'approche II. Le système procède en quatre étapes : extraction de candidats termes sur le texte source anglais, validation et normalisation des CT extraits, repérage des équivalents dans le texte cible français, validation et normalisation des CT français.

Pour l'extraction des candidats termes en langue source, Hull utilise une approche statistique qui exploite la reconnaissance des segments répétés. Le résultat est validé manuellement. Les variantes sont regroupées et remplacées par une forme canonique.

Le repérage des candidats termes français se fait à partir de l'identification dans le texte cible d'une séquence de mots susceptibles de contenir la traduction du terme source. L'algorithme d'alignement de TRINITY peut s'appliquer à tous les mots. Toutefois, pour l'acquisition de termes, il ne se concentre que sur les mots pleins (nom, adjectif, verbe et adverbe) et laisse de côté les mots grammaticaux. Pour améliorer la performance de l'alignement, les mots des textes sont lemmatisés afin de diminuer le nombre de formes à analyser.

L'algorithme d'alignement employé repose sur trois suppositions. Premièrement, les termes sources ne peuvent être modifiés puisqu'ils ont déjà été validés par des terminologues. Deuxièmement, pour faciliter la compréhension au moment de la validation, la séquence de mots français ne doit pas être scindée. Troisièmement, les séquences de mots trop longues sont préférées aux séquences trop courtes, car il est plus facile de

nettoyer une séquence trop longue que de revenir dans le texte chercher les parties manquantes d'une séquence trop courte.

2.3.6 Névéol et Ozdowska (2005)

Afin de procéder à une extraction bilingue de termes médicaux, Névéol et Ozdowska (2005) ont recours à une méthode identique à celle décrite dans Ozdowska et Bourigault (2004) et Ozdowska (2004). Cette méthode fait appel à l'approche III, car elle est fondée sur une analyse syntaxique parallèle des phrases sources et sur une identification simultanée des candidats termes. Les auteures sont ainsi parvenues à enrichir la terminologie du CISMeF[34] de 133 synonymes appartenant au domaine médical.

L'étude, qui met en oeuvre une méthode « d'appariement par propagation syntaxique » (Névéol et Ozdowska 2005), est fondée à partir de l'hypothèse selon laquelle « les liaisons paradigmatiques peuvent aider à déterminer les relations syntagmatiques, et inversement » (Debili et Zribi 1996). Ainsi, comme l'explique Ozdowska :

> Si deux mots $T1_i$ (*protective*, dans l'exemple) et $T2_p$ (*protecteurs*) sont appariés et s'il existe une relation de dépendance syntaxique entre $T1_i$ (*protective*) et $T1_j$ (*clothing*), d'une part, et entre $T2_p$ (*protecteurs*) et $T2_q$ (*vêtements*), d'autre part, alors $T1_j$ (*clothing*) et $T2_q$ (*vêtements*) peuvent être appariés (d'après Ozdowska 2004)[35].

La Figure 2.6, d'après celle qui est présentée dans l'article des auteures, illustre l'explication qui précède.

---

[34] Le CISMeF (Catalogue et Index des Sites Médicaux Francophones) est un catalogue qui indexe des ressources francophones spécialisées dans le domaine de la santé.

[35] Pour une meilleure compréhension, les exemples qui figuraient à l'origine dans l'explication ont été remplacés par ceux de la Figure 2.6.

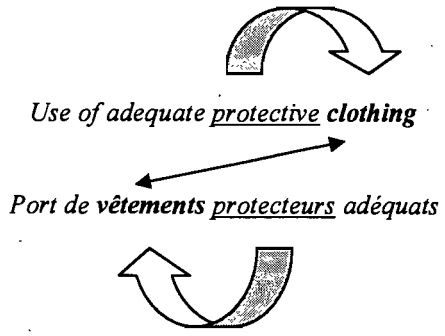


Figure 2.6 : Propagation des liens d'appariement (d'après Névéol et Ozdowska 2005)

L'acquisition des candidats termes s'effectue selon le scénario suivant : 1) analyse du corpus avec SYNTEX[36]; 2) identification de couples amorces; et 3) appariement local.

1) Les corpus anglais–français du domaine médical (CISMeF/Hansard[37], RCP[38]), qui comptent à eux deux environ 970 000 mots, sont préalablement alignés automatiquement au niveau des phrases et étiquetés. Ensuite, les analyseurs anglais et français SYNTEX effectuent l'analyse en dépendance syntaxique de chaque phrase (sujet, objet direct et indirect, modificateur, etc.).

2) L'étape de l'identification de couples amorces consiste à trouver des couples susceptibles de servir de point de départ à la propagation. Un couple amorce est constitué de deux mots en relation d'équivalence. Les couples amorces peuvent provenir de

---

[36] « L'analyseur syntaxique de corpus SYNTEX (Bourigault et Fabre, 1999) permet d'extraire d'un corpus une liste de noms et syntagmes nominaux, structurée par des relations de dépendance syntaxique. La fonction de cet analyseur est d'identifier des relations de dépendances entre mots et d'extraire d'un corpus des syntagmes (verbaux, nominaux, adjectivaux) ». ERSS, *Syntex,* [en ligne]. http://www.irit.fr/RFIEC/syntex/ (page consultée le 23 janvier 2008).

[37] Le Hansard (anglais–français) est un corpus juridique aligné composé de couples de phrases provenant des transcriptions des débats du parlement Canadien. Les textes retenus pour cette étude traitent du droit de la santé.

[38] Corpus constitué à partir de résumés des caractéristiques du produit (RCP) dans le cadre du projet PERTOMed (domaine principalement étudié : pharmacovigilance).

ressources lexicales externes, du repérage des cognats (section 2.1.2.2) ou de l'observation de la distribution des occurrences et des cooccurrences des unités à l'intérieur d'un corpus parallèle (section 2.1.2.1).

3) Les couples amorces ainsi obtenus servent à initier le processus d'appariement local (phrase à phrase) afin de découvrir les termes médicaux. L'appariement repose sur des règles de propagation[39] préalablement établies. Dans cette étude, les règles sont fondées sur des mots simples. Par exemple, à partir du couple amorce *protective/protecteurs*, le lien d'équivalence est propagé vers *clothing* et *vêtements*.

L'appariement par propagation présente un taux de précision de 70 % pour un rappel de 57 %. Les auteures expliquent que ces résultats sont liés au taux de fréquence des termes analysés. Plus le taux de fréquence est élevé, plus la précision chute; et plus le taux de fréquence est bas, plus le rappel est faible.

2.3.7 Gurrutxaga, Saralegi et Ugartetxea (2006)

Les auteurs de cet article présentent ELEXBI, un outil d'extraction terminologique bilingue pour l'espagnol et le basque. Ils partent du principe que l'acquisition automatique de termes est un cas particulier d'extraction lexicale. Pour cette première expérience, ils

---

[39] Dans les travaux d'Ozdowska et Bourigault (2004), d'Ozdowska (2004) et de Névéol et Ozdowska (2005), les règles de propagation sont établies manuellement, ce qui nécessite une expertise humaine et du temps. Pour contourner ce problème, Ozdowska et Claveau (2005) appliquent à la méthode « d'appariement par propagation syntaxique » une technique d'apprentissage artificiel (programmation logique inductive (PLI)) pour inférer automatiquement des règles de propagation. Cette dernière méthode permet, entre autres, de faciliter l'identification des cas d'isomorphismes et de non-isomorphismes syntaxique entre les deux langues analysées. La PLI est une technique d'apprentissage automatique qui permet, dans le cas de l'étude d'Ozdowska et Claveau (2005) d'inférer des règles de propagation générales à partir d'exemples de propagation valides au sein de deux phrases alignées. Les règles inférées sont ensuite recherchées à travers l'ensemble des règles possibles afin de découvrir celles qui maximisent un score.

disposent d'une mémoire de traduction[40], c'est-à-dire que les alignements phrase à phrase sont en principe sûrs à 100 %. Dans de futurs travaux, ils envisagent de travailler sur des corpus parallèles. Adoptant l'approche I, ils se proposent d'extraire des termes simples (noms) et des syntagmes nominaux.

Dans une première étape, ils procèdent à l'extraction automatique des candidats termes basques et espagnols. Les CT basques sont extraits à l'aide d'ERAUZTERM (Alegria *et al.* 2004). Cet outil, qui utilise des techniques linguistiques et statistiques, extrait les syntagmes nominaux, leur attribue une étiquette et les présente dans leur forme canonique. Pour les termes espagnols, l'extraction a été confiée à FREELING 2.1 (Carreras *et al.* 2004). L'extracteur repère les syntagmes nominaux candidats à partir d'une analyse de surface.

La deuxième étape consiste à regrouper en paires de candidats termes (CT) les termes provenant d'un même segment de traduction. Les paires sont ensuite confiées à une base de données relationnelle. Le processus d'appariement est itératif et commence par l'algorithme qui repère des chaînes de caractères identiques, par exemple, *extracción de terminología* et *terminologia-erauzketa*. Les paires de CT partageant cette caractéristique sont alors retirées de la base de données. Un deuxième algorithme prend la relève pour identifier les cognats dont la sous-chaîne maximale (section 2.1.2.2) est supérieure à 0,8. Comme dans l'étape précédente, les paires de CT qui correspondent à cette valeur sont enlevées de la base de données. Enfin, les paires de CT restant sont appariées grâce à des mesures d'association, comme l'information mutuelle (MI), le rapport de vraisemblance ou *log-likelihood ration* (LR), le coefficient de Dice, etc.

---

[40] « Une mémoire de traduction est une base de données contenant des segments de texte ainsi que l'équivalent de ces segments dans une autre langue. » Wikipedia. *L'encyclopédie libre,* [en ligne]. http://fr.wikipedia.org/wiki/M%C3%A9moire_de_traduction (page consultée le 12 janvier 2008).

Les auteurs estiment que la précision est acceptable (jusqu'à 80 % avec l'indice LR), mais qu'elle pourrait être améliorée si, dans la détection des termes, les extracteurs de termes monolingues basques et espagnols étaient plus performants et s'ils se montraient de valeur égale. Dans de futurs travaux, ils chercheront à implémenter un algorithme qui soit également capable d'extraire plus d'un équivalent pour chaque terme.

2.4 Synthèse et conclusion

En guise de synthèse, le Tableau 2.1 récapitule les travaux d'acquisition bilingue de termes décrits dans la section 2.3.

Tableau 2.1 : Récapitulatif des travaux sur l'extraction de termes

Année	Nom	Approche	Méthode	Type de CT	
1993	Van der Eijk	I	Analyse sémantique de surface et extraction – alignement sur la base de statistique de cooccurrences et de mesures positionnelles au niveau de la phrase.	Syntagmes nominaux	Termes simples (nom)
1994	Dagan et Church	II	Extraction de termes source par patrons catégoriels et par liste d'exclusion et validation – alignement avec *word-align*.	Syntagmes nominaux	Termes simples (nom, verbe, adjectif, adverbe)
1994	Daille *et al.*	I	Extraction par patrons catégoriels de CT anglais et français – alignement basé sur la cooccurrence et les patrons morphosyntaxiques.	Syntagmes nominaux	
1998	Gaussier	II	Extraction par patrons catégoriels – alignement : recherche de flots maxima dans un réseau de flots.	Syntagmes nominaux	
1998	Hull	II	Extraction statistique basée sur la reconnaissance des segments répétés – repérage des équivalents sur la base de la cooccurrence	syntagmes nominaux	Termes simples (nom, verbe, adjectif, adverbe)
2005	Névéol et Ozdowska	III	Analyse syntaxique parallèle du corpus – alignement par propagation syntaxique	Syntagmes nominaux	Termes simples (nom)
2006	Gurrutxaga *et al.*	I	Extraction par patrons catégoriels – alignement par cognation et cooccurrence	Syntagmes nominaux	Termes simples (nom)

Pour conclure, dans le Chapitre 2, nous avons voulu montrer que l'acquisition automatique bilingue de termes à partir de corpus parallèles s'appuie principalement sur deux techniques, l'extraction monolingue de termes et l'alignement des textes parallèles. Or, comme nous l'avons vu, ces deux techniques peuvent être mise en oeuvre au moyen de diverses méthodes. Ainsi, l'originalité de chacun des travaux sur l'acquisition bilingue de termes présentés dans la section 2.3 repose sur la combinaison des méthodes choisies par les concepteurs.

Depuis le début des années 1990, les progrès en matière d'acquisition bilingue de termes ont été remarquables, néanmoins, il reste toujours place à l'amélioration. Le bruit et le silence sont encore élevés, sans oublier que l'humain doit toujours intervenir de façon importante.

De plus, sous l'influence d'une pratique terminographique traditionnelle (donnant lieu à des dictionnaires contenant beaucoup de syntagmes nominaux), les chercheurs s'intéressent principalement à l'extraction des syntagmes nominaux, car ceux-ci sont plus immédiatement accessibles par les systèmes d'extraction. En outre, dans ces études, on tient pour acquis que les syntagmes nominaux sont plus nombreux que les termes simples ou que les autres types de termes (adjectif, verbe, adverbe). Par exemple, sur les 30 extracteurs recensés par Carreño et L'Homme (2007) et Carreño *et al.* (2007), sept systèmes seulement pouvaient extraire des termes simples.

Au passage, nous aimerions signaler qu'il existe des travaux d'acquisition de termes à partir de corpus comparables[41] (Dejean et Gaussier 2002; Morin *et al.* 2004). Ces corpus sont constitués de textes traitant du même sujet, mais écrits dans différentes langues. En

---

[41] « 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 » (Dejean et Gaussier 2002).

conséquence, on peut présumer que les textes qui les composent possèdent des termes dont les usages sont comparables entre les langues. Pour exploiter ces corpus, on utilise des méthodes statistiques qui se basent sur la similarité des contextes immédiats des termes.

Pour terminer, comme nous l'avons signalé plus haut, en extraction bilingue de termes, on s'est penché soit sur les termes complexes seulement, soit sur les termes complexes et les termes simples. À notre connaissance, il n'y a pas eu d'étude effectuée uniquement sur des termes simples, ce qui rend notre travail d'autant plus intéressant et pertinent.

Chapitre 3 : Méthodologie

Dans le présent chapitre, le cadre méthodologique adopté pour cette recherche est présenté en trois sections. La section 3.1 reprend étape par étape l'élaboration du corpus parallèle et son alignement. La section 3.2 s'emploie à décrire le logiciel d'acquisition automatique de termes TermoStat et les listes de candidats termes (CT) obtenues. Enfin, la section 3.3 décrit la méthode suivie pour l'analyse de 50 CT français et de leurs équivalents anglais.

3.1 Constitution du corpus

L'avènement d'Internet, en facilitant la constitution de corpus sous format numérique, a profondément modifié la pratique de la terminologie. En effet, le corpus, « ensemble de textes représentatifs du domaine [en vue d'en] décrire la terminologie » (L'Homme 2004 : 123), constitue de nos jours un matériau d'étude privilégié. Toutefois, pour être utile, selon les conseils de Bowker et Pearson, il doit être élaboré suivant les exigences du projet choisi.

> [...] corpora are not merely random collections of texts but, rather, they are collections that have been put together according to specific criteria. These criteria are determined by your needs and by the goal of your project (Bowker et Pearson 2002 : 45).

Dans les sections qui suivent, nous développons dans le détail les étapes qui ont mené à la mise en place du corpus parallèle *Changement climatique*.

3.1.1 Critères de sélection des textes

Notre corpus parallèle (anglais–français) se compose de textes se rapportant à un domaine qui prend de plus en plus le devant de la scène : le changement climatique. C'est pourquoi nous pensons que la constitution d'un corpus traitant de ce sujet peut être utile, non seulement pour ce travail de recherche, mais également pour des travaux futurs. Avant de commencer la cueillette des textes, nous avons, à l'instar de nombreux chercheurs travaillant sur des corpus, établi une liste de critères de sélection. La nôtre, inspirée des

ouvrages de Bowker et Pearson (2002 : 12-13, 50-51) et de L'Homme (2004 : 126-127), se présente de la façon suivante :

1. les auteurs des textes sources doivent être des spécialistes ou du moins des personnes connaissant très bien le domaine;
2. les textes doivent provenir de sites Internet ou de publications reconnus dans le domaine;
3. les textes doivent être diversifiés au niveau de la spécialisation (technique, scientifique, vulgarisé, etc.);
4. les textes doivent provenir de différents types de documents (manuels, périodiques, sites Internet, etc.);
5. l'équilibre entre la taille et le nombre de documents doit être respecté.

Une fois ces critères de base établis, nous avons voulu que nos textes puissent être classés selon leur lieu d'origine, c'est-à-dire que nous désirions réunir suffisamment de textes provenant du Canada, de l'Union Européenne et d'organismes internationaux officiels comme l'UNESCO.

Nous avons aussi tenu compte du fait que la collecte de textes parallèles est plus difficile à réaliser que celle de textes unilingues (Véronis 2000; Pearson 2000), car les sources électroniques multilingues où l'on peut recueillir des données parallèles fiables sont beaucoup moins nombreuses que les sources unilingues. À titre d'exemple, sur le site français de l'Agence européenne pour l'environnement[42], on ne peut trouver que les résumés des documents sur l'environnement. Pour obtenir les textes intégraux, on nous redirige vers le site principal de l'Agence[43] qui est en anglais. Malheureusement, sur ce site, la très grande majorité des textes sont distribués dans cette langue seulement.

---

[42] Agence européenne pour l'environnement : http://local.fr.eea.europa.eu/about_us

[43] European Environment Agency : http://www.eea.europa.eu/

Après avoir consulté des travaux antérieurs (Carreño 2004; Lemay 2003), il nous semblait que, pour obtenir des listes de candidats termes intéressantes, le corpus anglais–français devrait comporter au moins 350 000 mots par langue.

3.1.2 Description des textes retenus

Pour créer le corpus *Changement climatique*, nous avons orienté principalement nos recherches vers les sites officiels bilingues ou multilingues. À l'occasion, ces sites indiquent des liens externes à partir desquels nous avons extrait quelques documents qui présentaient de l'intérêt. En ce qui concerne les ouvrages papier, nous n'en avons retenu qu'un, que nous décrirons plus loin.

Les textes canadiens sont tous tirés du site du gouvernement fédéral, *Environnement Canada,* et s'adressent au grand public. Étant donné que le français et l'anglais sont les langues officielles de ce pays, les sites canadiens offrent l'avantage de passer d'une langue à l'autre d'un simple clic et de présenter des textes bilingues structurés sur le même modèle. Par ailleurs, d'autres chercheurs ont utilisé les sites du gouvernement du Canada parce qu'ils se prêtent bien à la constitution des corpus parallèles (Névéol et Ozdowska 2005).

Les documents de source européenne sont de provenances beaucoup plus diversifiées. Tout d'abord, nous avons trouvé des textes sur les sites officiels de l'Agence européenne pour l'environnement et d'EUROPA. Les informations contenues sur ces pages Web s'adressent au grand public. Ensuite, nous avons obtenu un document en anglais et en français sur le site de Greenfacts, organisation sans but lucratif. Ce site s'est donné pour mission de vulgariser de façon objective l'information scientifique sur des questions d'environnement et de santé. Puis, sur *L'Encyclopédie de l'Environnement Atmosphérique*, site bilingue conçu pour les utilisateurs de tout âge, nous avons prélevé les chapitres *Changement climatique* et *Réchauffement de la planète* et nous avons laissé de côté les

parties qui n'étaient par directement liées à notre sujet. Enfin, pour le dernier couple de textes de ce groupe, nous avons numérisé l'ouvrage *Le changement climatique* et sa traduction *Climate change* (Jacques et Le Treut 2004, 2005 (Annexe A)). Les auteurs, directeurs de recherche au CNRS, s'investissent depuis plusieurs années dans la communication scientifique en publiant de nombreux ouvrages. Dans celui-ci, ils présentent, dans un langage accessible, la complexité de fonctionnement de la machine climatique.

Les documents téléchargés des sites internationaux officiels proviennent de deux sources, mais ils sont de loin les plus nombreux. L'UNESCO propose sur son site Internet des magazines et des bulletins dans lesquels nous avons prélevé les articles se rapportant au domaine visé. Le site de l'IPCC (Intergovernmental Panel on Climate Change) ou GIEC[44] (Groupe d'experts intergouvernemental sur l'évolution du climat) offre des rapports à l'intention des décideurs et des scientifiques.

Finalement, en plus des textes appartenant aux trois groupes énumérés ci-dessus, nous avons ajouté un document des États-Unis : le rapport « An Abrupt Climate Change Scenario and Its Implications for United States National » commandé par le Pentagone et sa traduction française.

Pour conclure, bien qu'il ne soit pas toujours facile de connaître la langue source d'un texte (celle dans laquelle l'original est rédigé), nous pouvons quand même affirmer avec certitude que la très grande majorité de ceux que nous avons retenus a d'abord été écrite en anglais puis traduite en français. *Le changement climatique* (Jacques et Le Treut 2004 (Annexe A)) est, à notre connaissance, le seul document dont la langue d'origine soit le français.

---

[44] À partir de ce point, chaque fois que nous parlerons de cet organisme nous lui donnerons son appellation française.

3.1.3 Récapitulation des critères de sélection de base

Une fois les documents du corpus réunis, nous avons jugé bon de revenir sur les critères de sélection afin d'évaluer dans quelle mesure nous les avions respectés. Selon nous, tous les critères de base ont été satisfaits (Tableau 3.1). Bien entendu, on pourrait se demander s'il était justifié d'incorporer un si grand nombre de documents provenant du GIEC, soit 14 sur 31. Toutefois, étant donné que ces rapports sont devenus une référence incontournable dans le domaine et qu'ils sont cités à l'échelle internationale (presses, chercheurs, gouvernements, organismes, etc.), nous avons décidé d'en inclure un grand nombre dans le corpus. Par ailleurs, nous considérons que la terminologie contenue dans les textes du GIEC est la plus représentative de l'usage actuel de ce domaine.

Tableau 3.1 : Récapitulatif des critères de sélection de base

N°	Critères	Évaluation
1	Les auteurs des textes sources doivent être des spécialistes ou du moins des personnes connaissant très bien le domaine.	✓
2	Les textes doivent provenir de sites Internet ou de publications reconnus dans le domaine.	✓
3	Les textes doivent être diversifiés au niveau de la spécialisation (technique, scientifique, vulgarisé, etc.).	✓
4	Les textes doivent provenir de différents types de documents (manuels, périodiques, sites Internet, etc.).	✓
5	L'équilibre entre la taille et le nombre de documents doit être respecté.	✓

3.1.4 Référencement des textes

Dès qu'un texte est sélectionné, avant même qu'il ne soit prétraité, il doit être nommé et référencé.

Puisque nous voulions confier au logiciel Alinea l'alignement du corpus (section 3.1.7), nous avons suivi les indications de l'aide en ligne de cet outil pour nommer les fichiers. Ainsi, pour faciliter la reconnaissance du format, de la langue et des fichiers qui se correspondent, il est important de nommer les textes de la façon suivante : NOM.L1.FFF et NOM.L2.FFF. Par exemple, dans le fichier *chang_3europa.fr.txt* :

- *chang_* : indique que ce texte appartient au projet *Changement climatique*;
- *3europa* : est un préfixe quelconque commun aux deux fichiers (anglais–français);
- *.fr* : est un des codes langues sur deux caractères (norme ISO 639-1);
- *.txt* : est une extension indiquant le format.

Afin de nous conformer au désir d'uniformisation du processus de gestion des corpus du groupe ÉCLECTIK[45], nous nous sommes alignée sur le modèle suggéré par Marshman (2003). Nous avons donc construit, dans une base de données Access, une table comportant 17 champs servant à consigner les informations pertinentes de chaque document (exemple de formulaire à la Figure 3.1).

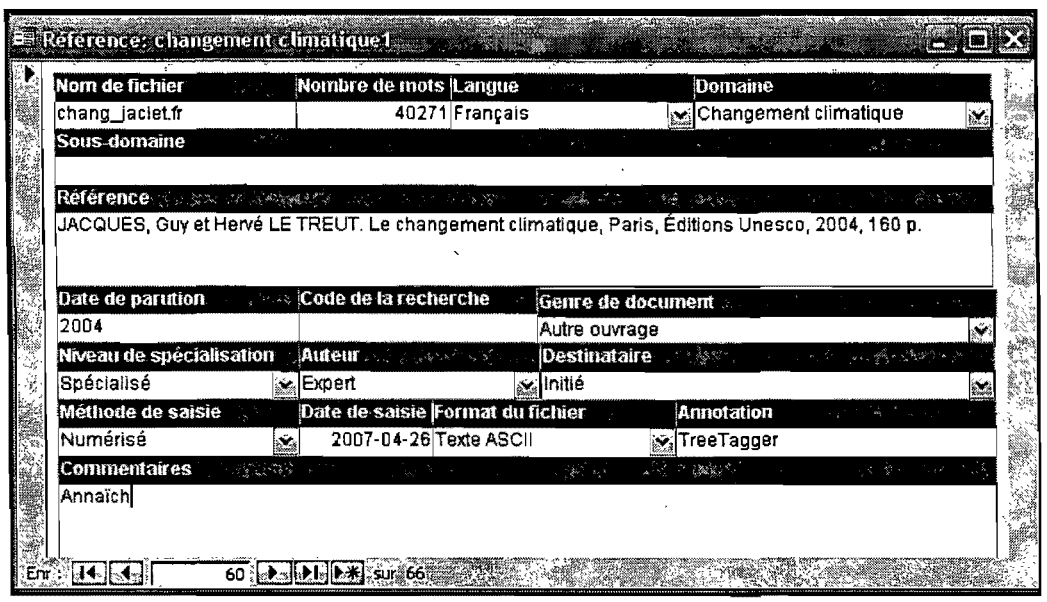


Figure 3.1 : Exemple de formulaire de la base de données des documents

[45] ÉCLECTIK est la composante « terminologie » de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal.

Une bibliographie complète des textes français composant le corpus *Changement climatique* est présentée à l'Annexe A.

3.1.5 Prétraitement

Lors de l'élaboration d'un corpus, il faut également prendre en considération le format sous lequel un document est disponible (PDF, HTML, papier). En effet, pour la constitution d'un corpus, la conversion en texte brut d'ouvrages papier et de fichiers PDF exige un prétraitement beaucoup plus important que pour celle des pages HTML (Bowker et Pearson 2002 : 6; L'Homme 2000 : 179-182). Ainsi, les documents provenant du groupe canadien apparaissent généralement sous format HTML, rarement en PDF. Un seul texte du groupe européen provient d'un ouvrage papier, les autres sont accessibles sous format HTML ou PDF. Les textes issus des sites officiels internationaux et du rapport du Pentagone sont tous des fichiers PDF.

À cause de l'ouvrage papier et du grand nombre de documents PDF retenus, nous avons dû procéder à un prétraitement assez important. Toutefois, il est à souligner que si nous nous étions limitée aux textes HTML, nous aurions eu de la difficulté à réunir un corpus parallèle d'une taille suffisante, de plus il aurait fallu passer à côté de textes présentant un intérêt de premier plan.

Le prétraitement des textes s'est effectué en trois étapes :

1. élimination de la presque totalité des péritextes[46];
2. déplacement des éléments qui empêchent un bon alignement (ou élimination lorsque ce n'était pas possible de les déplacer);
3. corrections dues aux erreurs de conversions de format.

---

[46] Ensemble des documents, notes, préface accompagnant un texte sans en faire partie (titre, préface, table des matières, bibliographie et listes des publications).

3.1.5.1 Élimination de la presque totalité des péritextes

Premièrement, afin de ne pas gonfler inutilement le corpus, nous avons éliminé des textes les parties du péritexte ne présentant pas d'intérêt pour cette étude. Nous avons cependant conservé les quelques annexes se présentant sous forme de texte. Dans les rapports du GIEC, le péritexte est particulièrement développé, principalement les préfaces qui sont consacrées aux remerciements adressés aux auteurs du document. À titre d'exemple, dans *L'aviation et l'atmosphère planétaire*, avant élimination du péritexte, on compte 13 307 mots et après : 8 569. Ce qui représente une différence non négligeable de 4 738 mots.

3.1.5.2 Déplacement des éléments qui empêchent un bon alignement (ou élimination lorsque ce n'était pas possible de les déplacer)

Deuxièmement, nous avons soit éliminé, soit déplacé certains éléments qui empêchent un bon alignement. Dans ce qui suit, nous présentons les opérations effectuées et nous en donnons les raisons :

- Nous avons ôté les glossaires et les listes des acronymes et abréviations, car l'ordre alphabétique des termes empêche un alignement parfait (l'ordre des termes anglais ne suit pas celui des termes français). Nous avons toutefois placé ces éléments dans un répertoire afin de les incorporer à une base de données dans un travail futur.
- Nous avons évité les encadrés, les graphiques, les figures ou les tableaux à cause du « décalage » des mises en page entre les textes anglais et français (les encadrés risquent de ne pas se trouver entre les mêmes paragraphes correspondants entre les deux langues). Par contre, les légendes ou textes compris dans ces données ont été copiés et collés en fin de texte dans le même ordre d'apparition pour les deux documents (anglais et français).

- Pour les notes de bas de page, nous avons procédé de la même manière que précédemment. Très souvent, les notes des textes français ne figurent pas sur les pages correspondantes des textes anglais.
- Nous avons pris soin d'enlever les numéros de page. L'aligneur Alinea utilise, dans sa première phase d'alignement, les chaînes numériques (entre autres) pour créer des points d'ancrage[47]. Il est donc important de ne pas l'induire en erreur avec des données qui pourraient fausser l'alignement. Dans l'ouvrage *Changement climatique*, par exemple, le chapitre 3 commence à la page 41 alors que, dans sa traduction, il commence à la page 47.
- Nous avons supprimé les en-têtes et les pieds de page, car, lors du transfert en .txt, ils s'insèrent à l'intérieur des phrases qui chevauchent deux pages. Dans l'exemple qui suit, on reconnaît en caractère gras l'en-tête d'un rapport intercalé à l'intérieur d'une phrase : « *Les options de politique pour* ***L'aviation et l'atmosphère planétaire*** *réduire davantage les émissions comprennent...* » (GIEC 1999) (Figure 3.2).

---

[47] Point d'ancrage : couple d'unités dont l'appariement est considéré comme fiable, entre lesquels les zones sont présumées alignées (Kraif 2001 : DCXLIII)

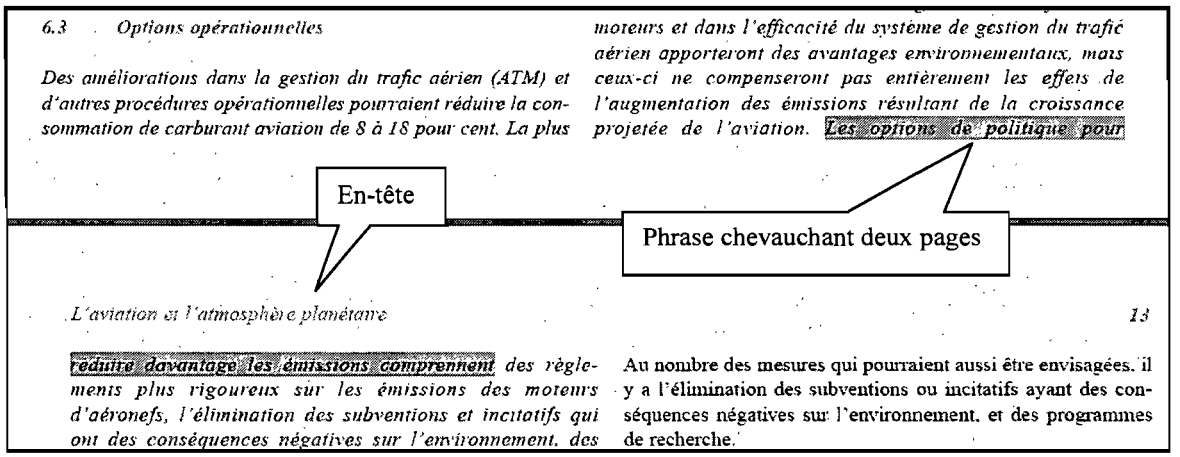


Figure 3.2 : Exemple de phrase chevauchant deux pages dans laquelle l'en-tête s'insère au moment du transfert en .txt.

- Enfin, étant donné que les PDF ont la particularité de terminer chacune de leur ligne par une marque de paragraphe (retour à la ligne), nous avons éliminé ceux qui nuisent à un bon alignement (Alinea utilise, par défaut, la marque de paragraphe pour la segmentation des phrases)[48]. Pour illustrer cette caractéristique, nous avons aligné un paragraphe tiré d'un PDF avant d'enlever les retours de chariot inutiles et après (Tableaux 3.2 et 3.3). Dans le Tableau 3.2, les phrases alignées sont incomplètes, alors que dans le Tableau 3.3, les mêmes phrases sont segmentées convenablement.

[48] Il est possible de modifier les fichiers de segmentation et de faire en sorte qu'Alinea n'utilise pas la marque de paragraphe comme fin de phrases. Toutefois, cette modification entraîne des inconvénients pour les textes contenant de nombreux titres et listes.

Tableau 3.2 : Paragraphe PDF aligné avec Alinea avant d'ôter les retours chariot

[s1] 2.	[s1] 2.
[s2] How Do Aircraft Affect Climate and Ozone ?	[s2] Quels sont les effets des aéronefs sur le climat et
[s3] Aircraft emit gases and particles directly into the upper	[s3 s4] l'ozone ? ¤ Les aéronefs émettent des gaz et des particules directement
[s4] troposphere and lower stratosphere where they have an impact	[s5] dans la haute troposphère et dans la basse stratosphère , où ils
[s5] on atmospheric composition .	[s6] ont un impact sur la composition de l'atmosphère .
[s6] These gases and particles alter	[s7] Ces gaz et
[s7] the concentration of atmospheric greenhouse gases , including	[s8] particules modifient la concentration des gaz à effet de serre
[s8] carbon dioxide (CO2) , ozone (O3) , and methane (CH4) ,	[s9] dans l'atmosphère , notamment le dioxyde de carbone (CO2) ,
[s9 s10] trigger ¤ formation of condensation trails (contrails) ;	[s10] l'ozone (O3) et le méthane (CH4) , déclenchent la formation de
[s11 s12] and may increase ¤ cirrus cloudiness—all of which contribute to climate change	[s11 s12] traînées de condensation et pourraient augmenter la nébulosité ¤ en cirrus , tout cela contribuant à des changements climatiques
[s13] (see Box on page 4) .	[s13] (voir encadré page 4) .

Tableau 3.3 : Paragraphe PDF aligné avec Alinea après avoir ôté les retours de chariot

[s1] 2.	[s1] 2 .
[s2] How Do Aircraft Affect Climate and Ozone ?	[s2] Quels sont les effets des aéronefs sur le climat et l'ozone ?
[s3] Aircraft emit gases and particles directly into the upper troposphere and lower stratosphere where they have an impact on atmospheric composition .	[s3] Les aéronefs émettent des gaz et des particules directement dans la haute troposphère et dans la basse stratosphère , où ils ont un impact sur la composition de l'atmosphère .
[s4 s5 s6] These gases and particles alter the concentration of atmospheric greenhouse gases , including carbon dioxide (CO2) , ozone (O3) , and methane (CH4) , ¤ trigger formation of condensation trails (contrails) , ¤ and may increase cirrus cloudiness—all of which contribute to climate change (see Box on page 4) .	[s4] Ces gaz et particules modifient la concentration des gaz à effet de serre dans l'atmosphère , notamment le dioxyde de carbone (CO2) , l'ozone (O3) et le méthane (CH4) , déclenchent la formation de traînées de condensation et pourraient augmenter la nébulosité en cirrus , tout cela contribuant à des changements climatiques voir encadré page 4) .

3.1.5.3 Corrections dues aux erreurs de conversion de format

Troisièmement, nous avons corrigé les erreurs créées lors de la conversion des formats. Si les documents HTML ne présentent pas plus de difficultés que la plupart de celles énumérées dans les points 3.1.5.1 et 3.1.5.2, il n'en va pas ainsi avec les PDF et les ouvrages papier.

En ce qui concerne la documentation papier, la qualité de la reconnaissance optique de caractères (ROC) varie selon les appareils ou selon les caractéristiques du texte à copier (type de caractères, qualité du papier, etc.). Il est par conséquent nécessaire de corriger les

coquilles produites lors de cette opération. Parfois, elles sont si nombreuses qu'il est préférable d'écarter le document.

Les documents PDF peuvent également présenter des problèmes de conversion[49], par exemple :

- il arrive qu'après les lettres *th*, *ff*, *fl*, *fi* le reste du mot soit séparé par une espace : *th e, eff ect, fl ows* ;
- les erreurs de conversion sont fréquentes : *likeiy!* au lieu de *likely7*;
- des mots sont parfois réunis : *usedfor, records.Instruments, sociétévont*;
- dans certains cas, des mots sont éclatés (toutes les lettres sont séparées par une espace) : *c o l l e c t i v e p i c t u r e o f*, au lieu de *collective picture of*[50].

Lorsque nous avons corrigé ces erreurs à l'aide du correcteur orthographique de Word, nous avons pris la précaution de ne pas surcorriger. Nous voulions seulement éliminer les erreurs commises lors de la conversion. Par exemple, dans les textes d'origine, il arrive que les mots apparaissent sous des variantes orthographiques différentes de la langue sélectionnée par le correcteur : *dewpoint, nighttime*. Des coquilles se glissent parfois dans les textes originaux : *disasasters, l'nfluence*. Nous croyons que, pour un terminologue, ces variantes et ces coquilles constituent des indices précieux (origine et qualité des textes par exemple).

Ce premier prétraitement terminé, il s'est révélé important de vérifier que les textes source et leur traduction soient organisés de la même façon afin de corriger les erreurs de copie (une page HTML peut par accident être rajoutée ou oubliée). Cette vérification sert en outre à détecter les différences structurelles entre les textes parallèles. Il peut arriver,

---

[49] Les documents PDF ont été convertis en .txt à l'aide du gratuiciel Adobe Reader 8.

[50] Le correcteur orthographique ne voit pas toutes les erreurs.

comme l'explique Pearson (2000 : 56), que des phrases, voire des paragraphes, soient modifiés, déplacés ou carrément enlevés. Il va sans dire qu'un alignement trop dissemblable ne donne pas de bons résultats (Véronis 2000; Kraif 2007). En ce qui nous concerne, nous avons décidé de ne pas nous attarder sur le parallélisme au niveau des phrases[51], mais de le faire au niveau des paragraphes.

Au cours du prétraitement, nous avons dû prendre de nombreuses décisions afin de permettre un bon alignement du corpus, la plupart de ces choix cependant n'affectent pas l'extracteur. Toutefois, si les nombreuses chaînes numériques et alphanumériques qui jalonnent les textes scientifiques constituent des points d'ancrage idéaux pour Alinea, ils peuvent être un désavantage pour l'extraction. Cet inconvénient se traduit par du bruit. En présence d'une chaîne de caractères inconnue, l'étiqueteur morphosyntaxique attribut souvent la mauvaise étiquette. Par exemple, dans les corpus français et anglais, le couple de dates « 1975–1995 » est étiqueté comme un nom commun.

À la fin du prétraitement, nous avons fait le décompte des mots dans Word, puis nous avons converti les fichiers dans le format nécessaire à chacun des logiciels utilisés.

3.1.6 Taille du corpus

Une fois finalisé, le corpus se compose donc de 31 paires de documents (anglais–français) dont la taille varie entre 779 mots et 49 922 mots. Le nombre total de mots pour le sous-corpus anglais s'élève à 509 955 alors que le sous-corpus français en compte 604 787 (Tableau 3.4). Les dates de parution des textes se situent entre 1997 et 2007, 6 documents ont été publiés avant 2000 et 25 à partir de cette date.

---

[51] Faire une vérification à un niveau de granularité aussi fin sur un corpus de cette taille, prendrait beaucoup trop de temps (dans ces conditions, cela reviendrait à pratiquement aligner le corpus manuellement).

Tableau 3.4 : Récapitulatif du nombre de documents et de la taille du corpus

Provenance	Nbre de paires de documents	Nbre de mots (sous-corpus anglais)	Nbre de mots (sous-corpus français)
Canada	6	89 002	102 738
Europe	8	99 413	111 448
Organismes internationaux	16	314 006	381 342
Pentagone	1	7 534	9 259
Total	**31**	**509 955**	**604 787**

3.1.7 Alignement

Avant de passer à l'extraction des candidats termes, nous avons procédé à l'alignement du corpus afin d'en vérifier la qualité[52]. La section 3.1.7.1 décrit l'aligneur Alinea. La section 3.1.7.2 explique la façon dont le corpus a été aligné et examine les résultats obtenus.

3.1.7.1 Description du logiciel Alinea

« Alinea est un programme dédié à la constitution et à l'édition de corpus bilingues alignés » (Kraif 2007). Ce logiciel, qui s'adresse surtout aux chercheurs, est distribué gratuitement, mais n'est pas un logiciel libre. Son concepteur le présente comme « un produit de laboratoire, élaboré au fil des années en fonction d'expériences ponctuelles » (Kraif 2007). Alinea est par conséquent un logiciel en constante évolution (Tableau 3.5). Nous utilisons la version 3.53, soit la dernière version disponible au moment de cette étude.

[52] Cette vérification supplémentaire, nous a permis par exemple de découvrir que, au cours de la collecte de textes parallèles sur Internet, nous avions oublié de copier les pages HTML correspondantes de certains textes.

Tableau 3.5 : Fiche technique du logiciel Alinea

Fiche technique	
Nom du logiciel	Alinea
Date de publication	En développement depuis 2000
Concepteur	Olivier Kraif (maître de conférences au département d'informatique pédagogique de l'Université Stendhal de Grenoble)
Distribution	Gratuiciel
Type	Programme dédié à la constitution et à l'édition de corpus alignés
Utilisateurs	Chercheurs, traducteurs, terminologues, linguistes, etc.
Langue d'affichage	Français (certaines boîtes de dialogue sont en anglais)
Support et document d'accompagnement	Aide en ligne (à terminer)
Environnement informatique	Windows 32 (2000, NT et XP)
Adresse Internet	http://w3.u-grenoble3.fr/kraif/

Le logiciel Alinea cumule plusieurs fonctions : comparaison d'alignements et évaluation (précision et rappel), édition manuelle d'alignement, recherche complexe et concordance avec des critères bilingues, extraction de lexiques bilingues, alignement au niveau des phrases et au niveau des mots. Dans notre étude, nous traiterons ces trois dernières fonctions.

Pour l'alignement au niveau des phrases, Alinea utilise des méthodes reposant principalement sur des informations internes qui combinent diverses stratégies : chaînes de caractères identiques (transfuges)[53], ressemblances superficielles (cognats), rapport des longueurs de phrases, calcul du meilleur chemin (algorithmes d'alignement). En utilisant les informations les plus fiables en premier et par un processus récursif (itératif), le programme converge vers un alignement de plus en plus précis.

[53] « Par transfuges, on désigne toutes les chaînes de caractères invariantes dans le passage à la traduction : les noms propres, les données numériques, certains sigles, les numéros de chapitre, etc. » (Kraif 2001 : 252).

L'alignement au niveau des mots est appelé *extraction des correspondances lexicales* par le concepteur (Kraif 2001 : 362), car il rejete le terme « alignement lexical » puisque la traduction unité à unité est impossible à ce niveau. Grefenstette (2004) va dans le même sens et parle « d'appariement de mots ». Pour extraire les correspondances lexicales phrase à phrase, Alinea se base sur la probabilité de l'hypothèse nulle (P0) ou *Null Hypothesis Approach* en anglais (Kraif et Chen 2004) décrite au Chapitre 2, section 2.1.2.1. L'extraction des correspondances lexicales réalisée, il est ensuite possible d'extraire le lexique[54] du corpus analysé à l'aide du module d'extraction du lexique. Ce module, conçu avant tout pour extraire tout le lexique d'un corpus, peut être paramétré pour filtrer les unités selon la partie du discours, le nombre d'occurrences, le pourcentage des occurrences, etc.

Alinea accepte les formats texte brut (.txt) et textes segmentés (.txs, .ces, .xip, CESAlign). Il exporte les fichiers d'alignement en format CESAlign, mais aussi en texte brut et en HTML. Alinea possède également une interface permettant de lancer l'étiqueteur morphosyntaxique TreeTagger (Schmid 1994), sans passer par la ligne de commande (à condition que TreeTagger soit installé sur l'ordinateur). L'aligneur peut ainsi lire directement les sorties étiquetées par TreeTagger (format .ttg).

3.1.7.2 Alignement du corpus et résultats

Nous avons procédé à deux alignements du corpus *Changement climatique*. Un premier projet a été créé sous format .txt et un deuxième sous format .ttg (étiqueté avec TreeTagger). Chacun de ces alignements présente des avantages, le premier se caractérise

---

[54] Ici, on entend l'ensemble des signifiants d'une langue L_1 et les équivalents d'une langue L_2 contenu dans le corpus.

par une segmentation de granularité plus fine[55], le deuxième permet d'effectuer une extraction des correspondances lexicales moins bruitée, car il fait appel à un corpus étiqueté.

L'alignement au niveau des phrases en .txt a généré 23 909 couples d'alignement et 15 855 points d'ancrage. L'opération s'est effectuée rapidement : lecture des fichiers, extraction des points d'ancrage, alignement au niveau des phrases n'ont pris, à eux tous, qu'à peu près une demi-heure avec un ordinateur de 2.2 GHz, 1 GHz de RAM. Toutefois, le corpus en texte brut n'étant pas étiqueté, nous n'avons pas procédé à l'extraction des correspondances pour les raisons expliquées au paragraphe précédent. Comme les fichiers portent tous un même nom de projet, *chang_* (section 3.1.4), nous avons construit le projet en définissant les fichiers sources et cibles à l'aide du caractère joker (*) de la façon suivante : *chang*.fr.txt* et *chang*.en.txt*. L'intérêt de cette opération réside dans le fait que chaque segment d'alignement est identifié selon le nom du fichier auquel il appartient. Ainsi, le segment 2061 qui porte l'identifiant [chang_11ip-s2061] appartient au texte *11ipccbilan2001* du projet *Changement climatique* (Figure 3.3). La Figure 3.3, montre le Navigateur bitextuel d'Alinea permettant de naviguer dans le corpus, d'y effectuer des recherches et de l'éditer.

[55] Les règles de segmentation appliquées sur format .txt sont basées à partir de paramètres de segmentation propre à Alinea. Dans le cas du format .ttg, la segmentation est prise en charge par TreeTagger qui met en œuvre des règles de segmentation moins fines que celles d'Alinea.

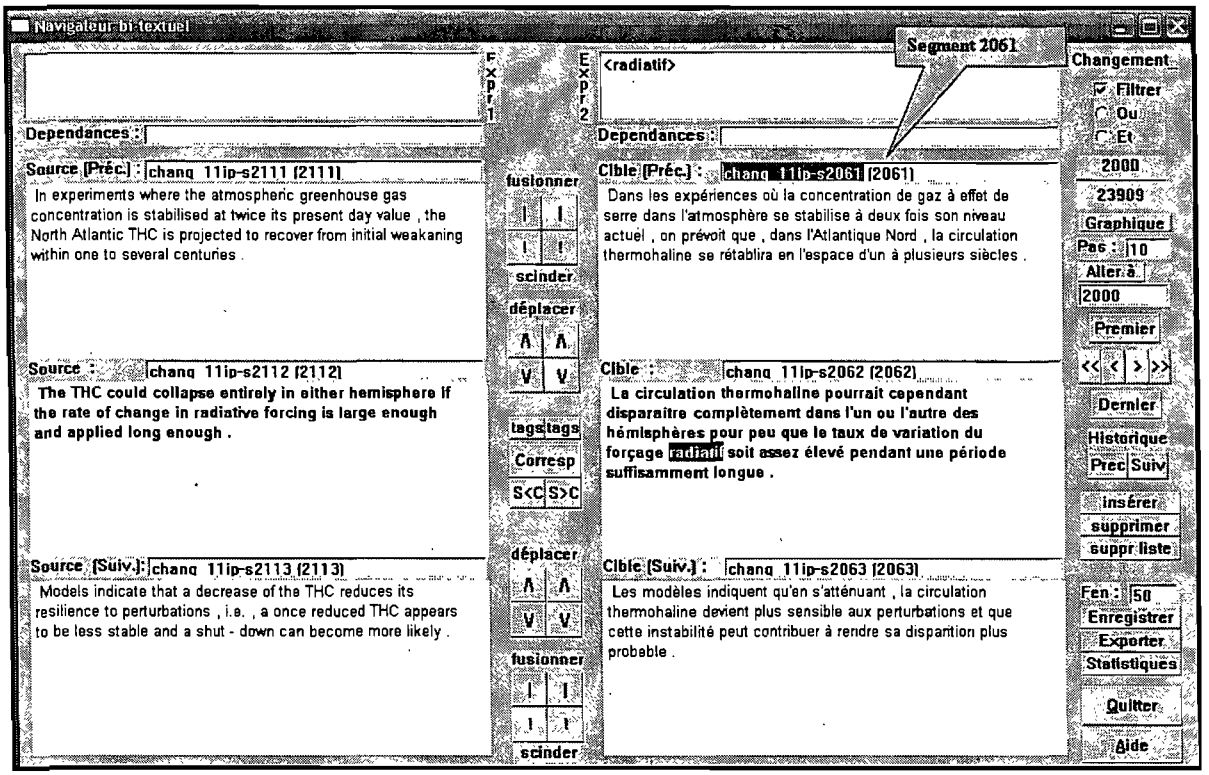


Figure 3.3 : Navigateur bitextuel d'Alinea

Étant donné que TreeTagger applique ses propres règles de segmentation, le projet sous format .ttg n'a généré que 18 396 couples d'alignement et 11 868 points d'ancrage, soit 5 513 alignements de moins qu'avec le projet en .txt. Ce nombre représente une différence de 23 %. À titre de comparaison, les Tableaux 3.6 et 3.7 présentent respectivement un échantillon du corpus parallèle en .txt et en .ttg. Comme on peut l'observer, la sortie .txt présente quatre couples d'alignement alors que celle en .ttg n'en produit que deux. Le calcul des points d'ancrage, l'alignement et l'extraction lexicale, à eux trois, n'ont pas pris beaucoup plus de temps que pour le projet précédent. Cependant, le temps de lecture des fichiers est de beaucoup augmenté (plus ou moins cinq heures dans notre cas).

Tableau 3.6 : Échantillon du corpus parallèle sous format .txt

[chang_3ipc-s10295] Vulnerabilities have been documented for a variety of coastal settings , initially by using a common methodology developed in the early 1990 s .	[chang_3ipc-s10092] La vulnérabilité d'une variété de zones côtières a été analysée , en utilisant au départ une méthodologie commune élaborée au début des années 90 .
[chang_3ipc-s10296] These and subsequent studies have confirmed the spatial and temporal variability of coastal vulnerability at national and regional levels .	[chang_3ipc-s10093] Ces études , et celles qui ont suivi , ont confirmé la variabilité spatiale et temporelle de la vulnérabilité côtière au sein des nations et des régions .
[chang_3ipc-s10297] Within the common methodology , three coastal adaptation strategies have been identified :	[chang_3ipc-s10094] Trois stratégies d'adaptation ont été définies :
[chang_3ipc-s10298] protect , accommodate , and retreat .	[chang_3ipc-s10095] protéger , composer , se retirer .

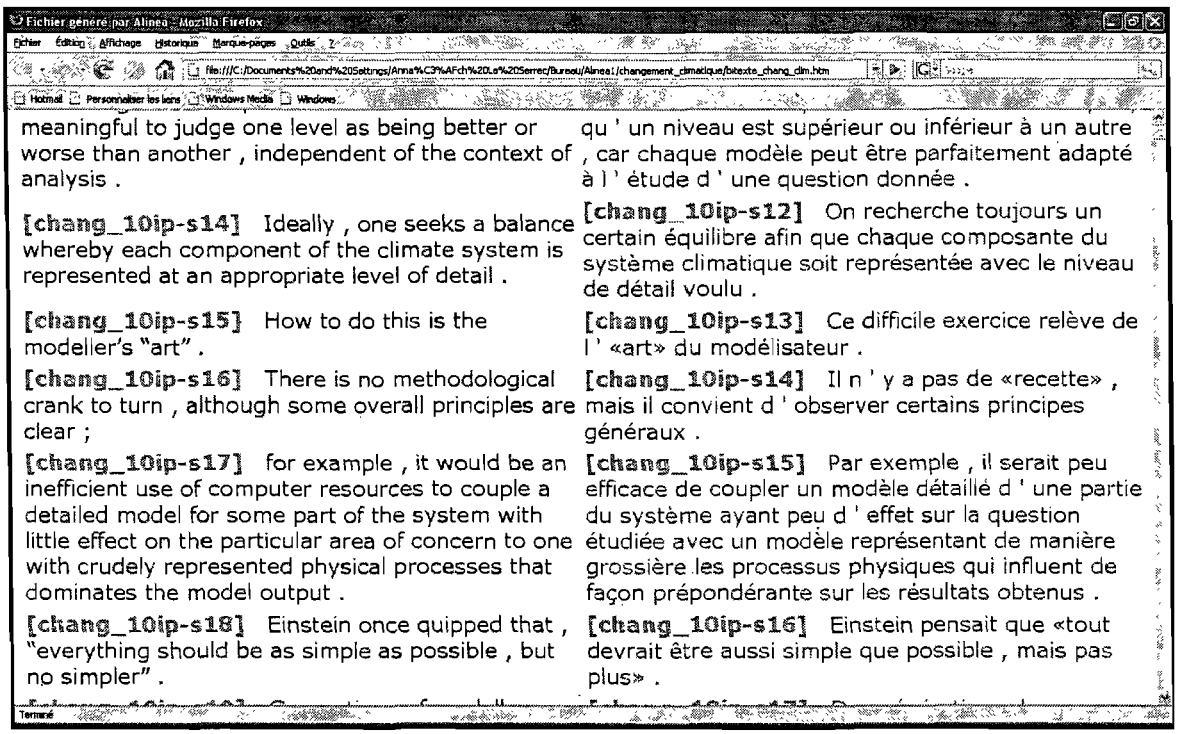
Tableau 3.7 : Échantillon du corpus parallèle sous format .ttg

[s4344 s4345] Vulnerabilities have been documented for a variety of coastal settings , initially by using a common methodology developed in the early 1990s . ¤ These and subsequent studies have confirmed the spatial and temporal variability of coastal vulnerability at national and regional levels .	[s4315] La vulnérabilité d'une variété de zones côtières a été analysée , en utilisant au départ une méthodologie commune élaborée au début des années 90. Ces études , et celles qui ont suivi , ont confirmé la variabilité spatiale et temporelle de la vulnérabilité côtière au sein des nations et des régions .
[s4346] Within the common methodology , three coastal adaptation strategies have been identified : protect , accommodate , and retreat .	[s4316] Trois stratégies d'adaptation ont été définies : protéger , composer , se retirer .

La qualité des alignements au niveau des phrases se révèle très satisfaisante[56] (Figure 3.4 et Tableau 3.13, colonne 7), à condition bien sûr que les textes soient parfaitement parallèles, ce qui n'est, bien entendu, pas toujours le cas. Tout bien considéré, nous croyons que, si dès le départ, on prend soin de soumettre à un prétraitement soigné des textes, sans toutefois y mettre trop de temps, il faut quand même accepter une certaine

[56] Au terme d'une évaluation du logiciel Alinea, nous en sommes arrivée à la conclusion qu'Alinéa est aussi efficace que les systèmes les plus performants en ce qui concerne l'alignement au niveau des phrases.

marge d'erreur. L'extraction des correspondances lexicales phrase à phrase, quant à elle, présente plus de bruit, de toute façon elle ne nous est pas utile pour le type de travail que nous désirons effectuer[57]. Par contre, l'extraction du lexique présente de l'intérêt et une de ses sorties sera comparée aux résultats de notre analyse manuelle.



meaningful to judge one level as being better or worse than another , independent of the context of analysis .	qu ' un niveau est supérieur ou inférieur à un autre , car chaque modèle peut être parfaitement adapté à l ' étude d ' une question donnée .
[chang_10ip-s14] Ideally , one seeks a balance whereby each component of the climate system is represented at an appropriate level of detail .	**[chang_10ip-s12]** On recherche toujours un certain équilibre afin que chaque composante du système climatique soit représentée avec le niveau de détail voulu .
[chang_10ip-s15] How to do this is the modeller's "art" .	**[chang_10ip-s13]** Ce difficile exercice relève de l ' «art» du modélisateur .
[chang_10ip-s16] There is no methodological crank to turn , although some overall principles are clear ;	**[chang_10ip-s14]** Il n ' y a pas de «recette» , mais il convient d ' observer certains principes généraux .
[chang_10ip-s17] for example , it would be an inefficient use of computer resources to couple a detailed model for some part of the system with little effect on the particular area of concern to one with crudely represented physical processes that dominates the model output .	**[chang_10ip-s15]** Par exemple , il serait peu efficace de coupler un modèle détaillé d ' une partie du système ayant peu d ' effet sur la question étudiée avec un modèle représentant de manière grossière les processus physiques qui influent de façon prépondérante sur les résultats obtenus .
[chang_10ip-s18] Einstein once quipped that , "everything should be as simple as possible , but no simpler" .	**[chang_10ip-s16]** Einstein pensait que «tout devrait être aussi simple que possible , mais pas plus» .

Figure 3.4 : Échantillon d'alignement sous format HTML

3.2 Extraction des candidats termes

L'extraction, qui porte sur les candidats termes simples anglais et français (nom, adjectif, verbe, adverbe), a été confiée au logiciel d'acquisition automatique de termes TermoStat (Drouin 2002). Dans le chapitre qui suit, les listes française et anglaise de

[57] Dans l'évaluation du logiciel Alinea évoquée dans la note précédente, nous avons également évalué l'extraction des correspondances lexicales.

candidats termes générées par ce logiciel sont examinées et nettoyées. Mais auparavant, le système d'extraction est brièvement présenté.

3.2.1 Description de l'extracteur TermoStat

TermoStat (Tableau 3.8) s'appuie sur une approche contrastive, c'est-à-dire qu'il exploite une méthode de mise en opposition de corpus non spécialisés et spécialisés. La méthode part du principe que la comparaison d'un corpus de référence de grande taille (langue générale) à un corpus d'analyse (langue de spécialité) permet de faire ressortir les spécificités lexicales de ce dernier. La notion de « spécificité » renvoie à l'idée que dans un corpus spécialisé certains termes lui sont particuliers et apparaissent statistiquement plus souvent que dans un corpus de référence.

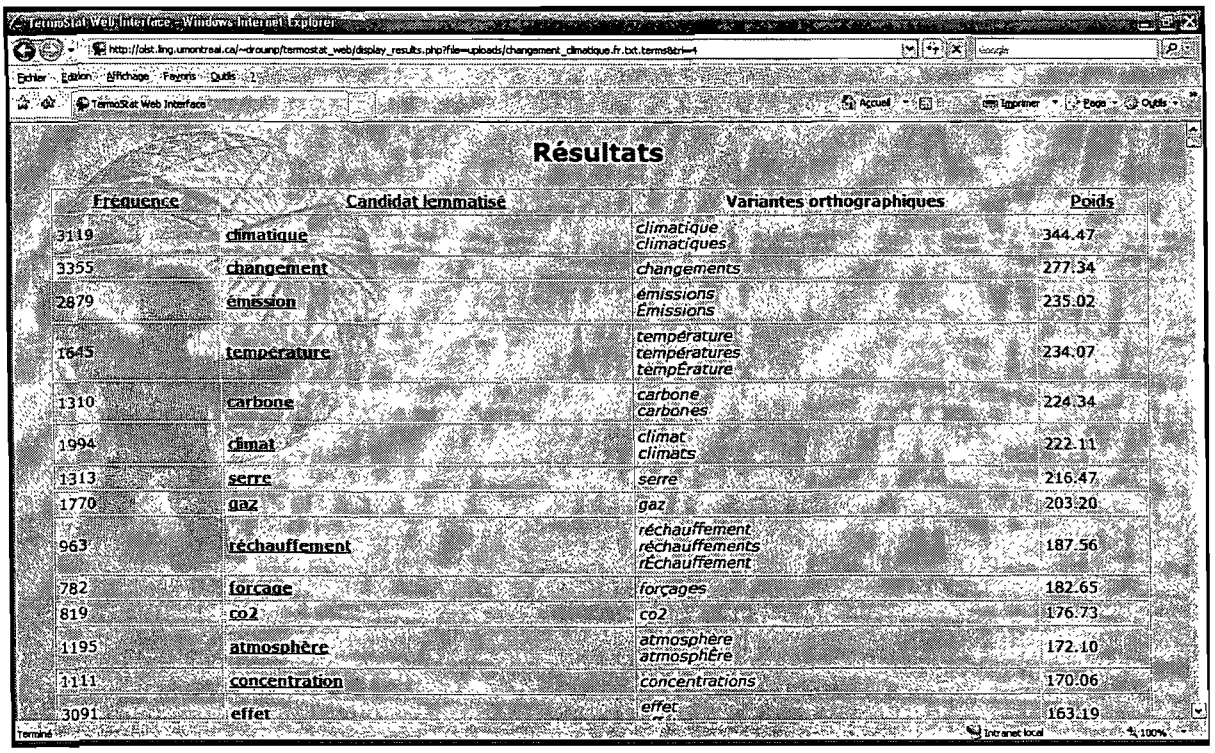
Après avoir étiqueté le corpus à l'aide de TreeTagger, TermoStat peut extraire des candidats termes simples et complexes en se basant sur des patrons morphosyntaxiques. Ensuite, à partir d'une technique qui exploite ce que l'on nomme *calcul des spécificités*, TermoStat attribue à chacun des CT un indice de spécificité. Ainsi, grâce à ce calcul, TermoStat propose des candidats termes fortement susceptibles d'être étroitement liés à la terminologie du corpus analysé. Ce système, qui a déjà fait l'objet d'évaluations (Lemay 2003; L'Homme 2004), a notamment été mis à contributions dans plusieurs travaux de recherche (Carreño 2004; Carrière 2006).

Tableau 3.8 : Fiche technique du logiciel TermoStat

Fiche technique	
Nom du logiciel	TermoStat
Date de publication	En développement depuis 2002
Concepteur	Patrick Drouin (professeur au Département de linguistique et de traduction de l'Université de Montréal)
Distribution	Accessible gratuitement par Internet (français, anglais, espagnol, italien)
Type	Logiciel d'acquisition automatique de termes
Utilisateurs	Chercheurs, traducteurs, terminologues, linguistes, etc.
Langue d'affichage	Français, anglais
Support et document d'accompagnement	s/o
Environnement informatique	Web
Adresse Internet	http://olst.ling.umontreal.ca/~drouinp/termostat_web/

Le présent travail de recherche utilise deux corpus de référence non spécialisés comme point de comparaison pour l'acquisition des termes dans le corpus *Changement climatique*. Le premier a été compilé à partir d'articles parus en 1989 dans le quotidien *The gazette* et comporte environ 7 millions d'occurrences. Le deuxième est constitué par des articles parus dans le journal *Le Monde* en 2002 et compte à peu près 30 millions de mots.

En entrée, il faut prendre soin de soumettre à TermoStat le corpus sous format texte brut. Le système confie alors ce dernier à TreeTagger pour étiquetage et lemmatisation. Puis il procède à l'extraction proprement dite. Une fois l'opération terminée, l'interface Web du logiciel affiche le résultat sur une page HTML (Figure 3.5).



Résultats

Fréquence	Candidat lemmatisé	Variantes orthographiques	Poids
3119	climatique	climatique climatiques	344.47
3355	changement	changements	277.34
2879	émission	émissions Émissions	235.02
1645	température	température températures tempÉrature	234.07
1310	carbone	carbone carbones	224.34
1994	climat	climat climats	222.11
1313	serre	serre	216.47
1770	gaz	gaz	203.20
963	réchauffement	réchauffement réchauffements réChauffement	187.56
782	forçage	forçages	182.65
819	co2	co2	176.73
1195	atmosphère	atmosphere atmosphère	172.10
1111	concentration	concentrations	170.06
3091	effet	effet	163.19

Figure 3.5 : Échantillon de l'acquisition de CT du corpus *Changement climatique*

Le logiciel donne accès aux informations suivantes :

- fréquence des candidats termes (CT) dans le corpus d'analyse;
- CT lemmatisé relié à ses contextes par hyperlien;
- variantes flexionnelles du CT présentes dans le corpus;
- poids de chaque CT (plus la valeur est élevée, plus le CT est spécifique au corpus).

Enfin pour terminer, il est à noter que, sur l'interface offerte gratuitement sur Internet, TermoStat génère des listes contenant à la fois des candidats termes simples et complexes. Il est cependant possible en contactant le concepteur du système, d'adapter les requêtes en fonction des besoins d'un projet donné. À titre d'exemple, pour notre étude,

TermoStat a été paramétré pour extraire seulement les noms, les adjectifs, les verbes et les adverbes.

3.2.2 Examen et nettoyage des listes de CT (anglais–français)

Dans les sections qui suivent, nous examinons les listes françaises et anglaises de CT simples afin de nous familiariser avec leur contenu et les nettoyer. Pour notre étude, une liste nettoyée est une liste de CT brute à laquelle nous n'avons enlevé que les CT n'appartenant pas à une des parties du discours définies dans les paramètres de l'extraction (nom, adjectif, verbe, adverbe). Ici, nous aimerions rappeler que cette étude n'est pas à caractère normatif, mais descriptif. Étant donné que nous nous situons en amont de la validation, les candidats termes contenus dans les listes d'extraction n'ont pas fait l'objet d'une analyse terminologique.

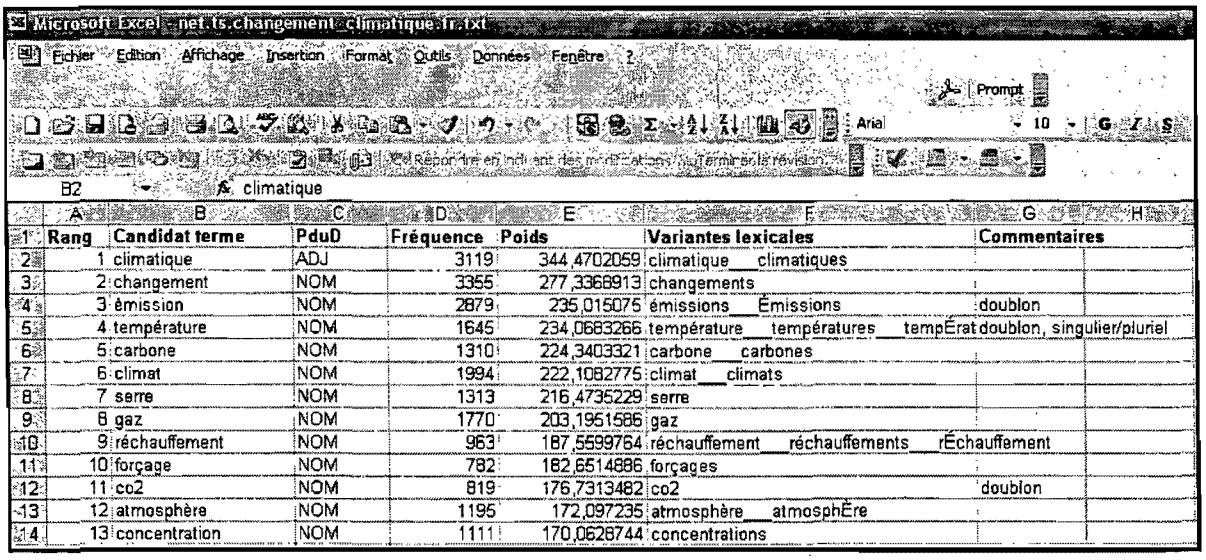
Les erreurs qui causent du bruit sont classées de la façon suivante :

1. erreurs contenues dans les textes avant prétraitement (coquilles, passages non traduits, etc.);
2. erreurs de prétraitement (reconnaissance imparfaite de caractères, CT avec appel de note, etc.);
3. erreurs commises par les systèmes (mauvaise partie du discours, problème de segmentation, double lemmatisation en raison d'une ambiguïté lexicale, CT tronqué).

Pour examiner, ordonner et nettoyer les listes, TermoStat n'étant pas conçu pour ces tâches, les résultats de l'extraction ont été importés dans des fichiers Excel. En plus des quatre colonnes contenant les données énumérées au point 3.2.1, le fichier Excel en présente trois de plus (Figure 3.6) :

- la colonne A indique le rang (le rang 1 correspond au plus haut score obtenu par un CT);

- la colonne C est utilisée pour y inscrire la partie du discours à laquelle appartient le CT;
- la colonne G reçoit les commentaires éventuels.



Microsoft Excel - net.ts.changement_climatique.fr.txt

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

B2 climatique

	A	B	C	D	E	F	G	H
1	Rang	Candidat terme	PduD	Fréquence	Poids	Variantes lexicales	Commentaires	
2	1	climatique	ADJ	3119	344,4702059	climatique__climatiques		
3	2	changement	NOM	3355	277,3368913	changements		
4	3	émission	NOM	2879	235,015075	émissions__Émissions	doublon	
5	4	température	NOM	1645	234,0683266	température__températures__tempÉrat	doublon, singulier/pluriel	
6	5	carbone	NOM	1310	224,3403321	carbone__carbones		
7	6	climat	NOM	1994	222,1082775	climat__climats		
8	7	serre	NOM	1313	216,4735229	serre		
9	8	gaz	NOM	1770	203,1951586	gaz		
10	9	réchauffement	NOM	963	187,5599764	réchauffement__réchauffements__rÉchauffement		
11	10	forçage	NOM	782	182,6514886	forçages		
12	11	co2	NOM	819	176,7313482	co2	doublon	
13	12	atmosphère	NOM	1195	172,097235	atmosphère__atmosphÈre		
14	13	concentration	NOM	1111	170,0628744	concentrations		

Figure 3.6 : Exemple de la liste de CT français importée dans Excel

3.2.2.1 Exemples d'erreurs retirées des deux listes

Les erreurs que nous avons retirées sont pratiquement du même type dans les deux listes. En outre, elles sont, à peu de chose près, semblables à celles qui ont été enlevées de la liste de candidats adjectifs étudiée par Carrière (2006 : 60-61). Les candidats que nous avons éliminés ont été classés selon les catégories suivantes :

1. coquille : *programmmes, cetaines, combusion, fasisant, efffet,* etc.;
2. mots non traduits (ex., un paragraphe n'a pas été traduit en français) : *contributing, induce, soils,* etc.;
3. CT avec appel de note accolé : *température5, établi6, globe9, likely20,* etc.;
4. erreur de conversion de fichier : *twodimentional, i'altération, està,* etc.;
5. problème de segmentation : *world's, earth's, puisqu'il, puisqu'un,* etc.;

6. double lemmatisation en raison d'une ambiguïté lexicale : *convenir|convier, accroire|accroître*, etc.;
7. Chaîne numérique et alphanumérique, mots vides : *mais, g, h, e, ~1150, $*, etc.;
8. chiffre romain : *ix, xviiie, v*, etc.;
9. partie de titre anglais : *climate, proposals, the*, etc.;
10. partie d'adresse Internet : *http, www, grida, unetonne*, etc.;
11. nom propre : *aristotle, atlantic, hadleycentre, jésus*, etc.;
12. nom étranger/latin/grec : *marinus, klima, elaphus*, etc.;
13. CT tronqué : *thermo, Î, pre, sub*, etc.

3.2.2.2 Résultats du nettoyage de la liste de CT français

La liste française contient 3 703 CT dont le poids est égal ou supérieur à 3,09. TermoStat est généralement configuré pour n'extraire que des CT dont le poids est supérieur à 3,09 afin, comme l'explique Drouin :

> [...] de ne retenir que les formes intéressantes et très significatives, nous nous attardons sur celles dont les valeurs-tests sont supérieures à 3,09, comme le suggèrent Lebart et Salem (1994 : 183). L'adoption de ce seuil nous permet d'assurer qu'il n'y a que 1 chance sur 1 000 que la fréquence observée dans le texte CA, soit due au hasard (Drouin 2002 : 148).

Sur 3 703 CT, il a été écarté 915 candidats n'appartenant pas à l'une des parties du discours désirées. Le Tableau 3.9 donne un aperçu du nombre des erreurs relevées dans la liste de CT français. L'extraction française compte 989 hapax[58], sur ce nombre, 599 ont été éliminés (tous ces hapax sont situés en fin de liste, c'est-à-dire que ce sont des CT présentant les scores les moins élevés). Nous avons constaté que TermoStat est très sensible

[58] Mot, forme, emploi dont on ne peut relever qu'un exemple dans un corpus défini (Termium 2007).

aux occurrences peu fréquentes. Par exemple, dans le corpus français, un paragraphe anglais de 127 mots n'avait pas été traduit, TermoStat en a extrait 31 candidats.

Tableau 3.9 : Catégories et nombre d'erreurs relevées dans la liste de CT française

Catégorie de CT		Nombre
Extraction brute de CT français		**3 703**
Erreur dans les textes avant prétraitement		
Coquille	57	
Mot non traduit	31	88
Erreur de prétraitement		
CT avec appel de note	186	
Erreur de copie	27	213
Erreur des systèmes : TreeTagger (TTG), TermoStat (TS)		
Problème de segmentation	18	
Double lemmatisation en raison d'une ambiguïté lexicale	17	
Chaîne numérique et alphanumérique, mots vides	391	
Chiffre romain	18	
Partie de titre anglais	9	
Partie d'adresse Internet	7	
Nom propre	81	
Nom étranger/latin/grec	22	
Autre	9	
CT tronqué	42	614
Liste nettoyée		**2 788**

3.2.2.3 Résultat du nettoyage de la liste de CT anglais

La liste anglaise contient 2 906 CT dont le poids est égal ou supérieur à 3,09. Les candidats éliminés sont au nombre de 290. Par ailleurs, l'extraction en anglais n'a présenté que des candidats possédant un minimum de deux occurrences (contrairement à la liste française, il n'y a pas d'hapax). Le Tableau 3.10 donne un aperçu du nombre des erreurs relevées dans la liste de CT anglais.

Tableau 3.10 : Catégories et nombre d'erreurs relevées dans la liste de CT anglaise

Type de CT		Nombre
Extraction brute de CT anglais		**2 906**
Erreur dans les textes avant prétraitement		
Coquille	8	
Mot non traduit	0	8
Erreur de prétraitement		
CT avec appel de note	19	
Erreur de copie	6	25
Erreur des systèmes : TreeTagger (TTG), TermoStat (TS)		
Problème de segmentation	15	
Double lemmatisation en raison d'une ambiguïté lexicale)	0	
Chaîne numérique et alphanumérique, mots vides	184	
Chiffre romain	3	
Partie de titre français	0	
Partie d'adresse Internet	2	
Nom propre	29	
Nom étranger/latin/grec	1	
Autre	7	
CT tronqué	16	257
Liste nettoyée		**2 616**

3.2.2.4 Comparaison entre les deux listes

Lorsqu'on compare les deux listes, une première constatation s'impose : la liste anglaise compte moins de CT que la française. Ensuite, la liste anglaise ne comporte pas d'hapax.

Deuxième observation : il a été enlevé beaucoup moins de CT de la liste anglaise que de la liste française, 290 contre 915. Le fait qu'il n'y ait pas d'hapax dans la première explique en grande partie le résultat de la seconde. Comme nous l'avons vu à la section 3.2.2.2, 599 hapax ont été éliminés de la liste française. Ainsi, en anglais, les coquilles, les erreurs de copie et les CT avec appel de note sont beaucoup moins nombreux puisqu'en général ils ne produisent qu'une fois.

Nous avons également remarqué que dans les deux extractions (anglais – français), les erreurs les plus fréquentes sont attribuables à l'étiqueteur, ce qui est tout à fait normal compte tenu du fait que ce système sert de point de départ lors du lancement de l'extraction.

Dernière constatation, il existe dans les deux listes de nombreux doublons. Environ 124 paires en français et 80 en anglais. Nous avons identifié deux types de doublons. Le premier est causé par des erreurs d'étiquetage. Par exemple, dans la liste, il y a deux CT *beaucoup,* un est étiqueté *adverbe* et l'autre *nom.* Le deuxième type se produit au moment de la lemmatisation. Lorsque le module de lemmatisation de TreeTagger rencontre des termes qui lui sont inconnus au pluriel, il ne les ramène pas sous leur forme canonique et les laisse tels quels, par exemple, nous avons les CT *écozone* et *écozones.*

Enfin pour terminer, toute liste de candidats termes comporte inévitablement une certaine quantité de bruit. S'il n'est pas encore envisageable d'éviter certaines erreurs, d'autres pourraient l'être. À la suite de l'analyse de nos résultats, nous pensons qu'un prétraitement automatique du corpus pourrait corriger un certain nombre de ces problèmes. Nous pourrions, par exemple, séparer des mots les appels de notes, les symboles (*$, %, ±, §, ~*), etc. Il serait également possible de procéder à un post-traitement du corpus étiqueté pour corriger les erreurs d'étiquetage les plus fréquentes.

3.3 Analyse des listes de CT (anglais – français)

Dans cette section, nous présentons la méthode suivie pour analyser les listes d'extraction des candidats termes français et anglais. Pour cela, nous avons procédé en trois grandes étapes : 1) identification manuelle et analyse des équivalents anglais de 50 CT français; 2) repérage de la position des équivalents dans la liste d'extraction anglaise et calcul de l'écart entre la position des CT français et leurs équivalents; 3) recours au module

d'extraction du lexique d'Alinea pour extraire les 50 CT et leur équivalents et comparaison de la liste produite avec celle recueillie manuellement.

Avant de continuer, nous voudrions préciser deux points importants. Premièrement, nous avons travaillé du français vers l'anglais et les équivalents anglais identifiés sont ceux qui apparaissent dans le corpus. Deuxièmement, étant donné qu'ils n'ont pas reçu de statut terminologique définitif, les 50 CT français sélectionnés sont susceptibles d'être des termes simples, des têtes de syntagme, des modificateurs, ou encore des non-termes.

3.3.1 Identification des équivalents anglais

Dans le but d'identifier, de classer et de décrire les équivalents anglais des 50 CT français choisis, nous avons suivi les cinq étapes décrites dans les sections 3.3.1.1 à 3.3.1.5.

3.3.1.1 Sélection des candidats termes français

Comme il n'est pas possible dans le cadre de ce projet d'analyser tous les candidats termes français, nous avons choisi 50 CT à partir du début de la liste d'extraction, c'est-à-dire les CT dont le poids est le plus élevé. Comme déjà souligné dans la section 3.2.1, plus la valeur du poids du CT est élevée, plus le CT est spécifique au corpus, ainsi les CT étudiés sont ceux qui sont les plus susceptibles d'être des termes propres au changement climatique. Dans le Tableau 3.11, nous présentons les 50 CT sélectionnés. Les colonnes indiquent, dans l'ordre, le rang du CT[59], son lemme, sa fréquence, son poids, ses variantes lexicales et enfin la partie du discours à laquelle il appartient.

---

[59] On remarquera que les CT sont numérotés jusqu'à 53. Cela s'explique par le fait qu'au nettoyage nous avons éliminé trois CT, ceux occupant le 24e rang « ° » (symbole de degré), le 40e rang « ci » (celle-ci, ci-devant, etc.) et le 51e rang « ppmv » (parties par million en volume), puisque ces CT ne sont pas des unités lexicales.

Tableau 3.11 : CT sélectionnés pour l'analyse des listes d'extraction

Rang	Candidat lemmatisé	Fréquence	Poids	Variantes	Partie du Discours
1	climatique	3119	34 4,4702059	climatique climatiques	adjectif
2	changement	3355	277,3368913	changements[60]	nom
3	émission	2879	235,015075	émissions Émissions	nom
4	température	1645	234,0683266	température températures tempÉrature[61]	nom
5	carbone	1310	224,3403321	carbone carbones	nom
6	climat	1994	222,1082775	climat climats	nom
7	serre	1313	216,4735229	serre	nom
8	gaz	1770	203,1951586	gaz	nom
9	réchauffement	963	187,5599764	réchauffement réchauffements rÉchauffement	nom
10	forçage	782	182,6514886	forçages	nom
11	co_2	819	176,7313482	co_2	nom
12	atmosphère	1195	172,097235	atmosphère atmosphÈre	nom
13	concentration	1111	170,0628744	concentrations	nom
14	effet	3091	163,19259	effet effets	nom
15	aérosol	624	159,6186728	aérosols	nom
16	écosystème	642	155,6772825	écosystèmes Écosystèmes ÉcosystÈmes	nom
17	atténuation	580	155,4898995	atténuation	nom
18	scénario	1373	152,6789183	scénarios scÉnarios	nom
19	incidence	627	145,3996366	incidences	nom
20	océan	810	144,2312509	océans	nom
21	modèle	1679	139,0221151	modèles	nom

[60] Nous avons observé que le logiciel commet des erreurs dans son fichier de sortie. La plupart du temps, lorsqu'il ne propose que la variante au pluriel, la variante au singulier devrait également être présente (ce qui se vérifie en consultant les contextes proposés par TermoStat).

[61] Les variantes avec des majuscules sont probablement attribuables à une incompatibilité entre les différents jeux de caractères utilisés dans les fichiers.

22	élévation	535	139,0083139	élévation élévations Élévation	nom
23	radiatif	437	136,5790928	radiatif radiatifs radiatives	adjectif
25	atmosphérique	490	131,0083631	atmosphériques	adjectif
26	variation	704	129,5251035	variations	nom
27	précipitation	543	126,4611336	précipitations	nom
28	variabilité	362	122,6647497	variabilité variabilités	nom
29	dioxyde	394	121,0837139	dioxyde	nom
30	surface	838	120,9979664	surface surfaces	nom
31	ozone	394	119,2815577	ozone	nom
32	échelle	690	115,4716948	échelles	nom
33	stabilisation	437	115,39972	stabilisation stabilisations	nom
34	eau	1439	112,151936	eau eaux	nom
35	anthropique	289	109,7520395	anthropiques	adjectif
36	augmentation	1050	109,7254122	augmentation augmentations	nom
37	océanique	311	105,1022137	océaniques	adjectif
38	coût	1017	101,1446987	coût coûts	nom
39	fossile	303	99,65303493	fossiles	adjectif
41	adaptation	611	97,35966773	adaptation adaptations	nom
42	latitude	330	96,9300139	latitude latitudes	nom
43	naturel	806	96,74779221	naturelle naturels naturelles	adjectif
44	combustible	325	96,58539107	combustibles	nom
45	évaluation	588	95,15224033	évaluation évaluations Évaluation Évaluations	nom
46	côtier	294	93,10451226	côtières côtiers cÔtiÈres	adjectif
47	niveau	1312	91,50472368	niveaux	nom
48	réduction	793	90,18969156	réduction réductions	nom
49	piégeage	193	90,01027058	piégeage	nom
50	météorologique	276	89,64142076	météorologiques	adjectif
52	glaciaire	226	88,60768818	Glaciaires	adjectif

53	mer	952	87,83363682	mer mers	nom

À la lecture du Tableau 3.11, on notera que nous avons conservé le CT CO_2. Cette abréviation étant employée majoritairement dans notre corpus comme un nom commun (avec un déterminant), nous avons jugé bon de la garder. Parmi les 50 CT sélectionnés, nous comptons 40 noms, 10 adjectifs, 0 verbe et 0 adverbe[62].

3.3.1.2 Sélection des paires de contextes (anglais–français)

Pour identifier les équivalents anglais des CT français, depuis Alinea, nous avons exporté, pour chacun des CT étudiés, toutes les paires de contextes dans un fichier Excel. Comme ces contextes sont très nombreux, nous avons ensuite sélectionné un maximum de 310 paires de contextes par CT. Le chiffre 310 s'explique par le fait que le corpus compte 31 paires de textes et que nous avons pris, au hasard (au début, au milieu ou la fin de chaque texte), un bloc contenant un maximum de 10 paires de contextes par paire de textes. Concrètement, pour le CT *climatique,* dont la fréquence est de 3119 et qui apparaît dans 31 paires de textes, nous avons prélevé, dans chacun de ces textes, 10 paires de contextes, pour un total de 310. Le CT *forçage*, dont la fréquence est de 782, quant à lui, apparaît dans 14 paires de textes, nous avons donc pris, dans ces 14 textes jusqu'à 10 paires de contextes et nous avons obtenu 98 contextes. Si dans ce dernier exemple, nous n'avons pas rassemblé les 140 paires de contextes attendues, c'est parce que certains documents présentent moins de 10 paires de contextes. Le Tableau 3.13, présente, pour chacun des CT, dans la 2^{e} colonne, le nombre de paires de textes dans lequel il est présent; et, dans la 3^{e} colonne, le nombre de paires de contextes sélectionnées. Le nombre total de contextes analysés s'élève à 9 283, ce qui donne une moyenne de 186 contextes par CT.

---

[62] Les verbes est les adverbes apparaissent plus loin dans la liste d'extraction.

3.3.1.3 Sélection des occurrences des candidats termes

Lorsque les paires de contextes ont été choisies, avec collage spécial (texte sans mise en forme), nous les avons exportées vers Word, puis converties en tableau de deux colonnes. Nous avons comptabilisé toutes les occurrences du CT mis en gras (parfois, un contexte contient plus d'une occurrence du CT étudié). La 4e colonne du Tableau 3.13 indique le nombre d'occurrences pour chaque CT avant analyse. Au cours de l'analyse, nous avons rejeté toutes les occurrences apparaissant dans des contextes mal alignées. Parfois, c'était toute la paire de contextes qui était mal alignée; d'autres fois, ce n'était qu'une partie de la paire. Dans l'exemple du Tableau 3.12, la partie en gris du contexte français n'est pas rendue dans le contexte anglais.

Tableau 3.12 : Exemple d'alignement partiel

chang_jacl-s24445 A vast number of socio - economic models were used by the IPCC to define **scenarios** for the evolution of temperature without attaching to any one of them a greater probability of occurrence .	chang_jacl-s23886 chang_jacl-s23887 Un vaste ensemble de modèles socio - économiques a été utilisé par le GIEC pour décrire les **scénarios** d ' évolution de la température sans attacher à aucun d ' eux une plus grande probabilité d ' occurrence . À chacun de ces **scénarios** est associée une estimation des émissions de gaz à effet de serre et d aérosols ;

Une fois l'analyse terminée, nous avons comptabilisé les occurrences analysées, les résultats figurent dans la 5e colonne du Tableau 3.13. Ainsi, à la fin de l'analyse du CT *climatique*, nous avons obtenu des résultats pour 322 occurrences sur 326 et pour le CT *forçage*, 111 occurrences sur 118. Dans la 6e colonne, nous indiquons le nombre d'occurrences mal alignées et, dans la 7e colonne, le pourcentage d'occurrences bien alignées. La moyenne des occurrences bien alignées avec leur équivalent s'élève à 98 %.

Tableau 3.13 : Récapitulatif de la sélection des contextes et de l'analyse des occurrences

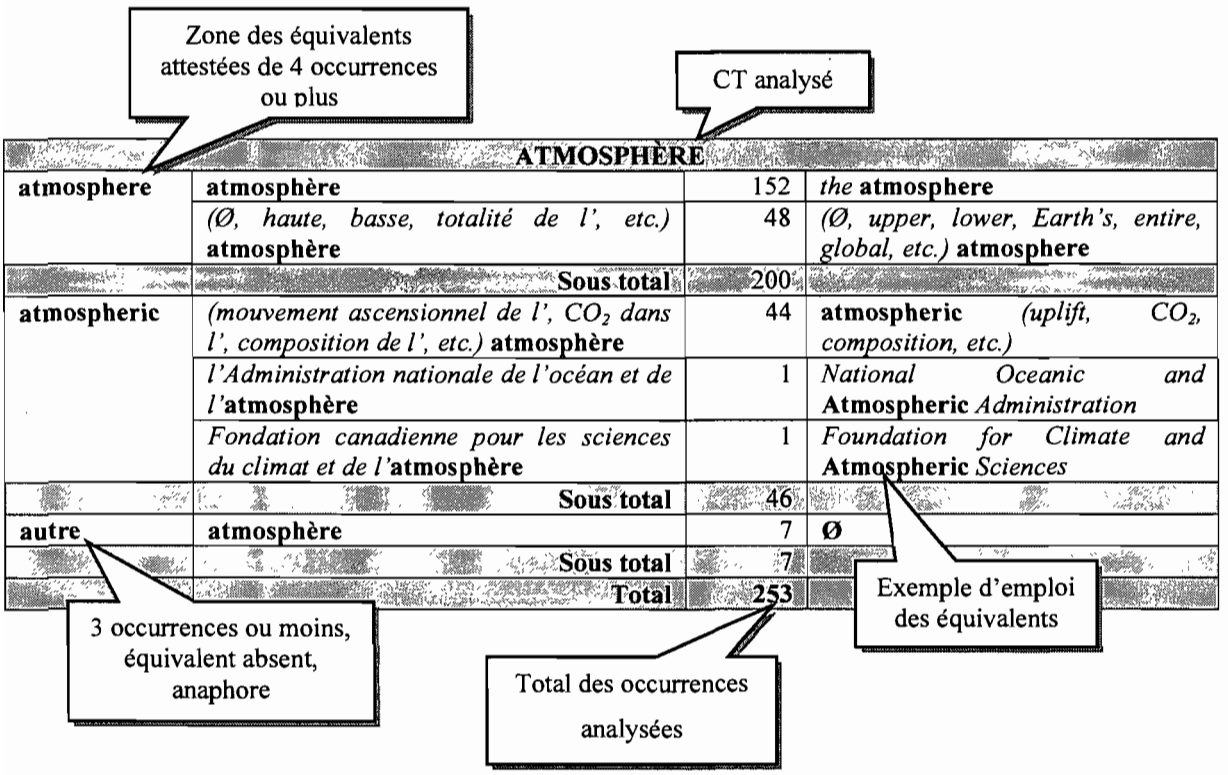
CT	Nombre de documents	Nombre de contextes	Occurrences avant analyse du CT	Occurrences après analyse du CT	Différence	Pourcentage d'occurrences bien alignées
climatique	31	310	326	323	3	99 %
changement	31	307	330	316	14	96 %
émission	31	255	280	268	12	96 %
température	23	204	226	225	1	100 %
carbone	27	205	231	225	6	97 %
climat	30	254	273	261	12	96 %
serre	31	269	284	279	5	98 %
gaz	31	267	282	275	7	98 %
réchauffement	26	202	223	216	7	97 %
forçage	14	97	118	114	4	97 %
co_2	28	195	219	216	3	99 %
atmosphère	30	245	256	253	3	99 %
concentration	29	214	233	228	5	98 %
effet	31	297	323	317	6	98 %
aérosol	23	153	170	165	5	97 %
écosystème	27	182	195	193	2	99 %
atténuation	19	93	98	96	2	98 %
scénario	25	193	232	227	5	98 %
incidence	20	139	141	137	4	97 %
océan	25	203	226	223	3	99 %
modèle	30	237	266	257	9	97 %
élévation	27	170	187	182	5	97 %
radiatif	22	104	110	107	3	97 %
atmosphérique	26	189	193	190	3	98 %
variation	28	192	207	202	5	98 %
précipitation	21	164	176	172	4	98 %
variabilité	23	164	169	167	2	99 %
dioxyde	28	165	172	165	7	96 %
surface	26	197	214	210	4	98 %
ozone	19	129	149	142	7	95 %
échelle	29	223	236	234	2	99 %
stabilisation	20	85	97	97	0	100 %
eau	27	213	260	253	7	97 %
anthropique	21	134	136	136	0	100 %
augmentation	30	244	261	255	6	98 %
océanique	22	152	160	154	6	96 %
coût	22	147	157	155	2	99 %
fossile	27	158	168	162	6	96 %
adaptation	23	126	135	134	1	99 %
latitude	20	145	155	153	2	99 %
naturel	29	233	242	238	4	98 %
combustible	24	149	158	153	5	97 %
évaluation	25	194	202	194	8	96 %

côtier	21	114	123	117	6	95 %
niveau	31	265	285	280	5	98 %
réduction	31	221	236	234	2	99 %
piégeage	13	42	43	42	1	98 %
météorologique	23	126	128	127	1	99 %
glaciaire	18	105	112	109	3	97 %
mer	26	212	229	223	6	97 %
Total des contextes analysés		**9 283**	**Moyenne des occurrences bien alignées**			**98 %**

3.3.1.4 Identification des équivalents anglais

Afin de répertorier les équivalents anglais, pour chaque CT français, nous avons pris en note toutes les formes sous lesquelles le CT français était rendu dans les contextes anglais correspondants et nous les avons comptabilisés. Les observations les plus pertinentes étaient consignées dans une fiche d'analyse attribuée à chaque CT. Le tableau 3.14 montre en exemple la fiche du CT *atmosphère*. On notera que le total des occurrences analysées est de 253, alors qu'au départ (Tableau 3.13) elles étaient au nombre de 256, cela s'explique par le fait que 3 occurrences pour ce CT étaient mal alignées (voir explications section 3.3.1.3). Au cours de l'analyse, chaque fois que nous avons compté 4 occurrences ou plus d'un même équivalent anglais, nous l'avons placé dans son propre enregistrement, ce qui faisait de lui un équivalents attesté. Autrement dit, les équivalents présents 3 fois ou moins n'ont pas été acceptés comme équivalents attestés et ont été placés dans la catégorie *autre* du tableau d'analyse avec les cas d'anaphore, d'abréviation d'une lettre et d'omission (Tableau 3.14). Toutefois, nous tenons à souligner que nous ne prétendons pas que les équivalents placés dans la section *autre* ne sont pas valables. Par ailleurs, dans cette étude, nous appelons *équivalent privilégié* (placé en tête de la première colonne) l'équivalent d'un CT qui possède le plus d'occurrences dans la fiche d'analyse. Dans le Tableau 3.14, l'équivalent *atmosphere* est donc l'équivalent privilégié, car il présente le plus d'occurrences.

Tableau 3.14 : Fiche d'analyse du CT *atmosphère*



ATMOSPHÈRE			
atmosphere	**atmosphère**	152	*the* **atmosphere**
	(Ø, haute, basse, totalité de l', etc.) **atmosphère**	48	*(Ø, upper, lower, Earth's, entire, global, etc.)* **atmosphere**
	Sous total	200	
atmospheric	*(mouvement ascensionnel de l', CO_2 dans l', composition de l', etc.)* **atmosphère**	44	**atmospheric** *(uplift, CO_2, composition, etc.)*
	*l'Administration nationale de l'océan et de l'***atmosphère**	1	*National Oceanic and* **Atmospheric** *Administration*
	*Fondation canadienne pour les sciences du climat et de l'***atmosphère**	1	*Foundation for Climate and* **Atmospheric** *Sciences*
	Sous total	46	
autre	**atmosphère**	7	**Ø**
	Sous total	7	
	Total	253	

3.3.1.5 Classification des CT par type d'équivalent

Afin de faciliter la description des CT français et leurs équivalents anglais, nous avons choisi de diviser les CT en deux grandes catégories : 1) les CT avec un seul équivalent; et 2) les CT avec plusieurs équivalents. Ces deux catégories sont elles-mêmes sous-divisées en fonction de la partie du discours à laquelle appartient l'équivalent ou de la forme (terme simple/terme complexe) qu'il prend. Pour construire notre système de classification, nous nous sommes basée en grande partie sur les problèmes d'établissement d'équivalence énumérés à la section 1.3 du Chapitre 1 (L'équivalence en terminologie). Naturellement, le point 2 (les termes complexes ont, dans chacune des langues, des structures ou des longueurs différentes) n'a pas été pris en considération puisqu'il ne touche que les termes complexes.

1. **CT avec un seul équivalent[63] :**
   a. CT avec un équivalent terme simple appartenant à la même partie du discours (*combustible/fuel*);
   b. CT avec un équivalent terme simple appartenant à une partie du discours différente;
   c. CT avec un équivalent terme complexe (*inlandsis/ice sheet).*

2. **CT avec plusieurs équivalents :**
   a. CT avec des équivalents termes simples appartenant à différentes parties du discours, mais morphologiquement apparentés entre eux (*atmosphère/atmosphere, atmospheric*);
   b. CT avec au moins deux équivalents termes simples morphologiquement différents l'un par rapport à l'autre (*précipitation/precipitation, rainfall*) (certains peuvent également avoir les caractéristiques de 2a, ex. *atténuation/mitigation, mitigate (to), moderate (to)*);
   c. CT avec au moins un équivalent terme complexe et/ou un équivalent faisant partie d'une des composantes d'un nom composé dont les éléments sont accolés[64] (*anthropique/anthropogenic, **anthropogenically induced**, human, **human activities**, human-induced*; *eau/water, **seawater**, sea*) (certains peuvent également avoir les caractéristiques de 2a et/ou 2b).

---

[63] Le fait d'avoir fixé le seuil à 4 occurrences et plus pour attester un terme, permet d'affirmer sans trop de risque que nous avons bien affaire à un CT n'ayant qu'un seul équivalent dans ce corpus.

[64] Par exemple, l'équivalent *water* est inclus dans le terme simple *seawater*. Le terme *seawater* est considéré comme un nom composé parce qu'il est formé à partir des mots *sea* et *water*.

3.3.2 Repérage de la position des équivalents dans la liste d'extraction et calcul des écarts entre les CT français et leurs équivalents

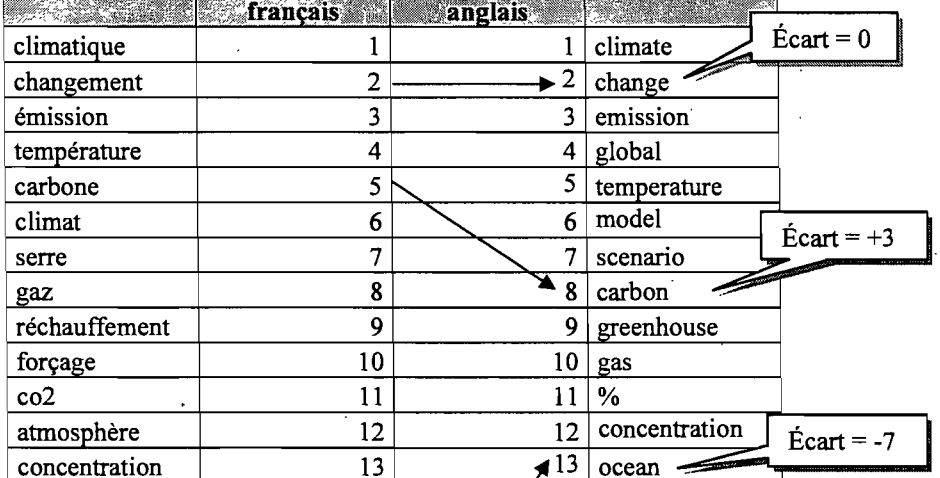
La deuxième étape de l'analyse des listes d'extraction (anglais – français) consiste à repérer, dans la liste d'extraction anglaise, la position de chacun des équivalents trouvés. Le repérage de la position permet de calculer les écarts entre les 50 CT français analysés et leurs équivalents respectifs.

Le calcul des écarts est présenté dans le Tableau 3.15. Le CT *changement* et son équivalent *change*, occupent tous deux le rang 2, par conséquent l'écart entre l'équivalent *change* et le CT *changement* est de 0. Le CT *carbone* et son équivalent *carbon* occupent respectivement le rang 5 et le rang 8; dans ces conditions, l'écart entre l'équivalent *carbon* et le CT *carbone* est de +3. À l'inverse, le CT *océan* et son équivalent *ocean* occupent respectivement le rang 20 et le rang 13, ce qui fait que l'écart entre l'équivalent *ocean* et le CT *océan* est de -7[65].

[65] Les signes + et – ont été choisis arbitrairement. Le signe + signifie que l'équivalent est situé plus bas dans la liste d'extraction que le CT français, alors que le signe – signifie que l'équivalent est situé plus haut que le CT français dans la liste d'extraction.

Tableau 3.15 : Exemples d'écarts entre équivalents anglais et CT français

CT français	Rang des CT français	Rang des CT anglais	CT anglais
climatique	1	1	climate
changement	2	2	change
émission	3	3	emission
température	4	4	global
carbone	5	5	temperature
climat	6	6	model
serre	7	7	scenario
gaz	8	8	carbon
réchauffement	9	9	greenhouse
forçage	10	10	gas
co2	11	11	%
atmosphère	12	12	concentration
concentration	13	13	ocean
effet	14	14	impact
aérosol	15	15	atmosphere
écosystème	16	16	warming
atténuation	17	17	sea
scénario	18	18	ecosystem
incidence	19	19	energy
océan	20	20	atmospheric



3.3.3 Recours au module d'extraction du lexique d'Alinea

Pour la troisième et dernière étape de l'analyse, nous avons utilisé le module d'extraction du lexique d'Alinea[66] pour extraire les équivalents des 50 CT français. À l'aide d'une expression régulière, nous avons soumis au module les 50 CT. En retour, le module génère la liste des 50 CT et leurs équivalents. La Figure 3.7 montre un échantillon de l'extraction lexicale, on y voit que le CT *anthropique* a été associé à 4 équivalents, le

[66] Voir Description du logiciel Alinea, section 3.1.7.1.

chiffre entre parenthèses indique le nombre d'occurrences des candidats équivalents trouvés dans le corpus, les hyperliens mènent aux contextes.

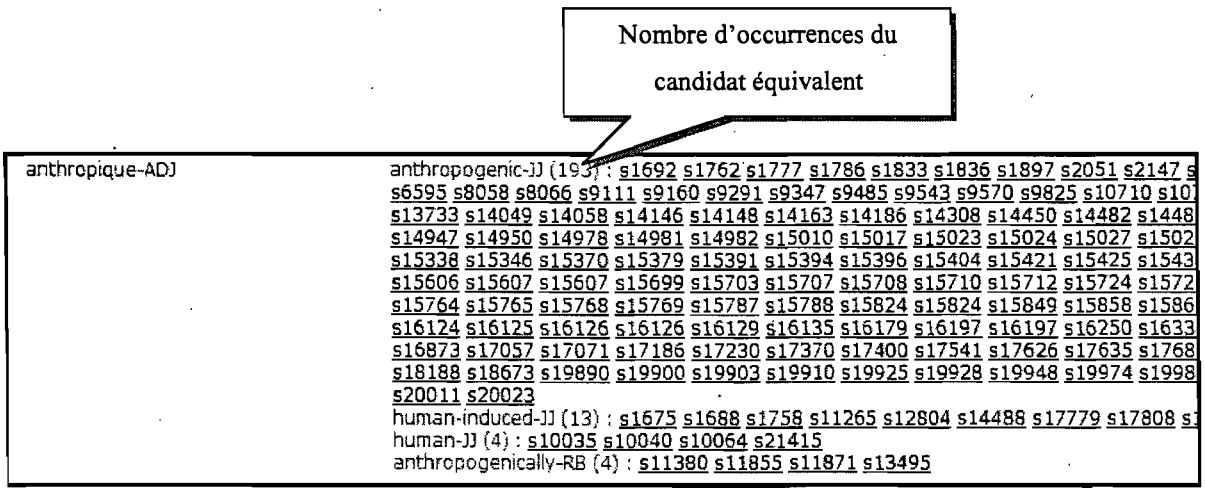


Figure 3.7 : Échantillon au format HTML de l'extraction du lexique

À l'aide d'un tableau, nous avons ensuite comparé la liste produite par Alinea avec les équivalents compilés manuellement afin d'étudier le bruit et le silence produit par le logiciel.

Chapitre 4 : Résultats de l'analyse

Dans ce chapitre, nous présentons les résultats de l'analyse des 50 CT français et de leurs équivalents. Dans la section 4.1, nous donnons les résultats de l'identification des équivalents anglais. Dans la section 4.2, nous décrivons les résultats du repérage des équivalents dans la liste d'extraction anglaise et les résultats du calcul des écarts entre les CT français et leurs équivalents. Enfin, dans la section 4.3, nous comparons l'analyse manuelle à l'extraction lexicale effectuée par Alinea.

4.1 Résultats de l'identification des équivalents anglais

Comme expliqué à la section 3.3.1.4, nous avons premièrement analysé tous les CT français afin d'identifier les équivalents. Les fiches d'analyse des 50 CT français sont placées à l'Annexe B en fonction du rang occupé par les CT à la sortie de l'extraction par TermoStat. Ensuite, afin de décrire les CT français et leurs équivalents anglais, nous avons classé les CT français selon les catégories décrites à la section 3.3.1.5. Dans la section 4.1.1, nous présentons un tableau général dans lequel les CT français sont répartis selon la classe d'équivalent à laquelle le CT est relié. Des sections 4.1.2 à 4.1.6, nous poursuivons par une description plus détaillée de chacune des classes d'équivalent. Finalement, à la section 4.1.7, nous terminons par des observations générales.

4.1.1 Classification générale

Le Tableau 4.1 montre la répartition générale des CT français par type d'équivalent qu'ils possèdent selon la méthode de classification et les critères établis à la section 3.3.1.5.

Tableau 4.1 : Classification générale des CT français par type d'équivalent

Catégorie principale	Sous catégorie	Nombre de CT	CT
1 Un seul équivalent	a	15	aérosol, changement, climat, combustible, concentration, co_2, côtier, coût, écosystème, émission, forçage, fossile, scénario, température, variabilité
	b	0	
	c	0	
2 Plusieurs équivalents	a	6	atmosphère, climatique, naturel, océanique, radiatif, stabilisation
	b	20	adaptation, atmosphérique, atténuation, augmentation, effet, élévation, évaluation, glaciaire, incidence, météorologique, modèle, niveau, océan, ozone, piégeage, précipitation, réduction, surface, variation,
	c	9	anthropique, carbone, dioxyde, eau, échelle, gaz, latitude, mer, réchauffement, serre
	Total	**50**	

Dans ce tableau, nous observons que la catégorie 1 (un seul équivalent) contient 15 CT, et que la catégorie 2 (plusieurs équivalents), en contient 35, soit le plus grand nombre.

4.1.2 CT appartenant à la catégorie 1a

Les CT appartenant à la catégorie 1a (CT avec équivalent terme simple appartenant à la même partie du discours) figurent dans le Tableau 4.2. La première colonne présente le CT français et la deuxième, l'équivalent anglais. Dans la troisième colonne, on trouve le nombre d'occurrences de l'équivalent anglais sur le nombre d'occurrences du CT français. La quatrième colonne indique le nombre de cas *autre*[67].

[67] Ces chiffres nous permettent d'apprécier la différence entre les CT attestés et les cas *autre*. Il est à remarquer que dans cette étude, comme nous n'avons travaillé que sur 50 CT, intentionnellement, nous ne donnons pas de pourcentages, car ils peuvent facilement conduire à des conclusions générales trop hâtives.

Tableau 4.2 : Liste des CT de la catégorie 1a

Terme français	Équivalent anglais	Équivalents angl. / occurrences fr.	autre
aérosol	aerosol	157/165	8
changement	change	293/316	23
climat	climate	258/261	3
combustible	fuel	145/153	8
concentration	concentration	202/228	26
co_2	co_2	211/216	5
côtier	coastal	116/117	1
coût	cost	143/155	12
écosystème	ecosystem	187/193	6
émission	emission	258/268	10
forçage	forcing	111/114	3
fossile	fossil	162/162	0
scénario	scenario	210/227	17
température	temperature	220/225	5
variabilité	variability	162/167	5

Comme le montre le Tableau 4.2, parmi les 15 CT français, nous avons 2 adjectifs : *côtier* et *fossile*. Dans les Tableaux 4.3 à 4.5, nous présentons en contextes quelques CT appartenant à la catégorie 1a. Les contextes que nous prenons en exemple sont tirés des paires de contextes sélectionnées pour effectuer notre analyse des 50 CT (section 3.3.3.2).

Tableau 4.3 : Illustration des CT *concentration* et *émission*

A given **concentration** target may be achieved through more than one **emission** pathway	Un même objectif en matière de **concentration** peut être atteint en faisant passer les **émissions** par plusieurs itinéraires
Humans are altering the **concentration** of greenhouse gases and aerosols, both of which influence, and are influenced by, climate.	Les activités humaines font varier la **concentration** des gaz à effet de serre et des aérosols, lesquels influencent et sont influencés par le climat.
In addition, carbon monoxide (CO) **emissions** have recently been identified as a cause of increasing CH4 **concentration**.	De plus, les **émissions** de monoxyde de carbone (CO) ont récemment été identifiées comme l'une des causes de l'augmentation de la **concentration** de CH4.

Tableau 4.4 : Illustration des CT *forçage* et *variabilité*

The response to anthropogenic changes in climate **forcing** occurs against a backdrop of natural internal and externally forced climate **variability**.	La réaction aux variations anthropiques du **forçage** climatique s'inscrit dans le contexte d'une **variabilité** naturelle propre au système climatique et d'une **variabilité** du climat due à des forçages externes.
The spatial patterns of some radiative **forcing** agents, especially aerosols, are very heterogeneous and so add further to the spatial **variability** of climate change.	La configuration dans l'espace de certains agents de **forçage** radiatif, notamment les aérosols, est très hétérogène et accroît la **variabilité** dans l'espace des changements climatiques.
Mathematical modelling is a powerful tool to explore the Earth's complex system and to study how the system responds to both external radiative **forcing** and to internal feedbacks and **variability** of the climate system.	La modélisation mathématique est un outil puissant, qui permet d'approfondir le système complexe de la Terre et d'étudier de quelle façon il réagit à la fois au **forçage** radiatif externe et à la **variabilité** et aux rétroactions internes du système climatique.

Tableau 4.5 : Illustration des CT *fossile, scénario* et *émission*

Global **fossil** CO2 **emissions** (GtC/y) for the IS92a **scenario**...	**Emissions** mondiales de CO2 d'origine **fossile** (GtC/an) selon le **scénario** IS92a...
Fossil CO2 **emissions** are those arising from **fossil** fuel combustion (including gas flaring) and cement production.	Les **émissions** de CO2 d'origine **fossile** sont celles découlant de l'utilisation des combustibles **fossiles** (y compris le torchage) et de la production de ciment.
In particular, for **scenarios** with higher **fossil** fuel use (hence, higher carbon dioxide **emissions**, e.g., A2), the SO2 **emissions** are also higher.	En particulier, pour les **scénarios** qui prévoient un usage intensif de combustibles **fossiles** (et par conséquent de fortes **émissions** de dioxyde de carbone, comme le **scénario** A2), les **émissions** de SO2 sont également plus élevées.
For example, deferring mitigation for a couple of decades would allow global **fossil** fuel **emissions** to increase significantly (e.g., IS92a and several other **scenarios**).	En retardant l'atténuation de quelques dizaines d'années, par exemple, on provoquerait à l'échelle mondiale une augmentation sensible des **émissions** de combustibles **fossiles** (voir par exemple, le **scénario** IS92a et plusieurs autres **scénarios**).

Le CT *fossile* est le seul qui n'a pas d'équivalent classé dans la section *autre*. Les CT ayant moins de 10 occurrences dans la section *autre* sont au nombre de 10. Par exemple, dans 6 cas, *aérosol* est rendu par la lettre *A* en anglais. Les CT *changement, concentration, écosystème* et *émission* possèdent 10 occurrences ou plus dans la section *autre*. Par exemple, l'équivalent du CT *concentration* est absent 16 fois du texte anglais.

Après examen des CT du Tableau 4.2, nous n'avons pas pu dégager de règle qui permet de prédire à l'avance qu'un certain type de CT possède un seul équivalent dans un corpus.

4.1.3 CT appartenant à la catégorie 1b et 1c

Parmi les 50 CT français sélectionnés, aucun ne remplissait les conditions de la catégorie 1b (CT avec équivalent terme simple appartenant à une partie du discours différente). En ce qui concerne cette catégorie, cela est dû au fait que nous avons établi l'attestation d'un équivalent à 4 occurrences ou plus (section 3.3.1.4). Toutefois, si nous l'avions fixée plus haut, il nous aurait été possible d'y mettre, par exemple, le CT *climatique*. En effet, cet adjectif français est pratiquement toujours traduit par le nom anglais *climate,* au lieu de l'adjectif *climatic.* Dans la fiche d'analyse du CT *climatique,* pour 323 occurrences analysées, nous avons 300 fois *climate,* 10 fois *climatic* et 13 fois *autre.* L'établissement du nombre d'occurrences pour faire en sorte qu'un équivalent soit attesté est une question de limite fixée arbitrairement pour les besoins de l'étude, du nombre de contextes étudiés, etc. Une fois ce seuil établi, il doit être appliqué à tous les CT.

Nous n'avons pas non plus trouvé de CT qui entre dans la catégorie 1c (CT avec équivalent terme complexe). Il faut dire que nous n'avons analysé que les 50 premiers CT de la liste d'extraction. Si nous avions exploré la liste plus avant, nous aurions trouvé, à la 183^{e} position, le CT *inlandsis* qui remplit les conditions de la catégorie 1c, car il n'a qu'un seul équivalent dans ce corpus : *ice sheet* (Tableau 4.6).

Tableau 4.6 : Illustration du CT *inlandsis*

Furthermore, the energy involved in melting Antarctic or Greenland **ice sheets** and albedo effects due to changes in their area, are small compared to the forcings.	En outre, l'énergie en jeu dans la fonte des **inlandsis** antarctique et groenlandais et les effets albédo attribuables à la réduction de leur superficie sont minimes par rapport aux forçages.

4.1.4 CT appartenant à la catégorie 2a

La catégorie 2a (CT avec des équivalents termes simples appartenant à différentes parties du discours, mais morphologiquement apparentés entre eux), compte 6 CT. Le Tableau 4.7 contient ces CT et se présente de la même façon que le Tableau 4.2, à la différence près que dans la 3^{e} colonne, nous indiquons pour chaque équivalent le nombre d'occurrences trouvées dans les contextes analysés et que sur la ligne *Total,* nous y exprimons le nombre des occurrences des équivalents sur le nombre des occurrences du CT français.

Tableau 4.7 : Liste des CT de la catégorie 2a

Terme français	Équivalent anglais	Équivalents angl. / occurrences fr.	autre
atmosphère	atmosphere	200	7
	atmospheric	46	
	Total	246/253	
climatique	climate	300	13
	climatic	10	
	Total	310/323	
naturel	natural	210	23
	naturally	5	
	Total	215/238	
océanique	ocean	122	5
	oceanic	27	
	Total	149/154	
radiatif	radiative	85	7
	radiation	15	
	Total	100/107	
stabilisation	stabilization	71	4
	stabilisation	13	
	stabilize (to)	4	
	stabilise (to)	5	
	Total	93/97	

Dans cette catégorie, parmi les 6 CT, se trouvent 4 adjectifs : *climatique, naturel, océanique* et *radiatif.* Il est intéressant de noter que dans le cas des adjectifs *climatique* et

océanique les équivalents privilégiés[68] sont des noms. Dans les tableaux 4.8 à 4.10, nous présentons en contextes quelques CT appartenant à la catégorie 2a.

Tableau 4.8 : illustration du CT *atmosphère*

Human Perturbations to the Composition of the **Atmosphere**	Perturbations anthropiques de la composition de l'**atmosphère**
(the latter two of which tend to reduce the **atmospheric** CO2 concentration)	(ces deux derniers tendant à réduire la concentration de CO2 dans l'**atmosphère**)

Tableau 4.9 : Illustration du CT *climatique*

This Technical Paper is intended as a primer on the **climate** system and SCMs, and has two objectives :	Le présent document technique d'introduction au système **climatique** et aux modèles **climatiques** simples vise deux objectifs:
Some coastal, high-latitude, and high-altitude ecosystems have also been affected by changes in regional **climatic** factors.	Certains écosystèmes côtiers, à latitudes et altitudes élevées, ont également subi les effets des variations des paramètres **climatiques** régionaux.

Tableau 4.10 : Illustration du CT *radiatif*

The **radiative** impact of a given change in cloud properties, cloud amount, or cloud height depends on the location and time of year and day when the changes occur.	L'effet **radiatif** d'une fluctuation donnée des propriétés, de la quantité ou de la hauteur des nuages dépend du lieu, de la période de l'année et du moment de la journée où elle se produit.
The importance of NOx in the **radiation** budget is because increases in NOx concentrations perturb several greenhouse gases; for example, decreases in methane and the HFCs and increases in tropospheric ozone.	L'importance des NOx dans le bilan **radiatif** tient au fait que l'accroissement de leur concentration a des répercussions sur plusieurs gaz à effet de serre et peut par exemple entraîner une diminution du méthane et des HFC et une augmentation de l'ozone troposphérique.

On remarquera également que, pour un CT donné, le nombre d'occurrences des équivalents privilégiés par rapport à celui des autres équivalents est beaucoup plus élevé,

[68] Nous rappelons que pour cette étude l'*équivalent privilégié* d'un CT est celui qui possède le plus d'occurrences dans la fiche d'analyse.

par exemple l'équivalent privilégié *climate* de *climatique* compte 300 occurrences alors que l'autre équivalent *climatic* n'en compte que 10.

Sauf pour le CT *naturel,* nous avons un nom parmi les équivalents anglais de chaque CT français (celui du CT *stabilisation* se présente sous deux graphies *stabilization* et *stabilisation*) : *atmosphere, climate, ocean, radiation, stabilization.* Les adjectifs sont aussi très présents : *atmospheric, climatic, natural, oceanic, radiative.* Nous n'avons qu'un seul équivalent adverbial : *naturally*; et un seul équivalent verbal (avec deux graphies) : *stabilize (to)* et *stabilise (to).*

Tous les CT ont au moins 2 occurrences dans la section *autre*, ceux qui en ont le plus sont les CT *climatique* (13) et *naturel* (23), mais le nombre ne dépasse pas ce que nous avons vu dans la catégorie 1a.

Dans la catégorie 2a, nous n'avons pas dégagé de règles permettant de prédire à l'avance qu'un certain type de CT présente la caractéristique de posséder des équivalents termes simples appartenant à différentes parties du discours, mais qui soient morphologiquement apparentés entre eux.

4.1.5 CT appartenant à la catégorie 2b

La catégorie 2b (CT avec au moins deux équivalents termes simples morphologiquement différents l'un par rapport à l'autre (certains peuvent également avoir les caractéristiques de 2a)) contient 20 CT, soit le plus grand nombre de CT français analysés. Le Tableau 4.11 est disposé de la même manière que le précédent.

Tableau 4.11 : Liste des CT de la catégorie 2b

Terme français	Équivalent anglais	Équivalent angl. / occurrence fr.	autre
adaptation	adaptation	91	
	adaptive	10	
	adapt (to)	19	10
	adjustment	4	
	Total	124/134	
atmosphérique	atmospheric	158	
	atmosphere	11	
	air	13	4
	weather	4	
	Total	186/190	
atténuation	mitigation	77	
	mitigate (to)	8	7
	moderate (to)	4	
	Total	89/96	
augmentation	increase	154	
	increase (to)	15	
	increased (adj.)	21	
	increasing (adj.)	8	
	rise	7	
	rising	7	22
	growth	7	
	higher	5	
	enhancement	5	
	enhanced	4	
	Total	233/255	
échelle	scale	126	
	level	16	88
	wide	4	
	Total	146/234	
effet	effect	107	
	impact	10	196
	affect (to)	4	
	Total	121/317	
élévation	rise	98	
	rising	15	
	rise (to)	7	27
	increase	31	
	higher	4	
	Total	155/182	
évaluation	assessment	134	
	assess (to)	5	
	estimate	15	22
	estimate (to)	4	
	evaluation	14	
	Total	172/194	

glaciaire	ice	92	3
	glacial	14	
	Total	106/109	
incidence	impact	81	12
	implication	14	
	effect	13	
	affect (to)	11	
	incidence	6	
	Total	125/137	
météorologique	weather	88	17
	meteorological	22	
	Total	110/127	
modèle	model	235	9
	modelling	5	
	pattern	8	
	Total	248/257	
niveau	level	232	42
	target	6	
	Total	238/280	
océan	ocean	203	2
	oceanic	12	
	sea	6	
	Total	221/223	
ozone	ozone	132	2
	o_3	8	
	Total	140/142	
piégeage	sequestration	23	6
	capture	13	
	Total	36/42	
précipitation	precipitation	141	2
	rainfall	31	
	Total	172/174	
réduction	reduction	133	25
	reduced	28	
	reduce (to)	35	
	decrease	8	
	cut	5	
	Total	209/234	
surface	surface	185	15
	area	10	
	Total	195/210	
variation	change	94	27
	variation	77	
	shift	4	
	Total	175/202	

Dans la catégorie 2b, nous avons 3 adjectifs : *atmosphérique, glaciaire* et *météorologique.* Pour appartenir à cette catégorie, tous les termes doivent posséder au moins deux équivalents morphologiquement différents, comme les équivalents du CT *surface* : *surface* et *area*; ou encore de celui du CT *variation* : *change, variation* et *shift.* Cette catégorie peut accueillir les CT ayant en plus les caractéristiques de la catégorie 1a, c'est-à-dire les CT possédant des équivalents morphologiquement apparentés entre eux, ce qui est le cas pour 9 CT : *adaptation, atmosphérique, atténuation, augmentation, élévation, évaluation, modèle, océan, réduction.* Dans les Tableaux 4.12 à 4.14, nous présentons en contextes quelques CT appartenant à la catégorie 2b.

Tableau 4.12 : Illustration du CT *adaptation*

Transfer of technology for **adaptation** to climate change is also an important element of reducing vulnerability to climate change.	Le transfert de technologie en vue de l'**adaptation** aux changements climatiques est aussi un élément important dans la réduction de la vulnérabilité aux changements climatiques.
Evaluation and **adjustment** to local conditions, and replication2 are other important stages.	L'évaluation des conditions locales et l'**adaptation** à ces conditions, ainsi que la reproduction2) sont d'autres étapes importantes.

Tableau 4.13 : Illustration du CT *atmosphérique*

Human influences will continue to change **atmospheric** composition throughout the 21st century.	L'influence des activités humaines continuera à modifier la composition **atmosphérique** tout au long du XXIe siècle.
Effects of climate change on other **air** pollutants are less well established.	Les répercussions des changements climatiques sur les autres polluants **atmosphériques** sont moins bien connus.

Tableau 4.14 : Illustration du CT *effet*

Because greenhouses retain heat in somewhat the same way, this phenomenon has been called the greenhouse **effect**, and the absorbing gases that cause it, greenhouse gases.	Comme les serres conservent la chaleur de façon un peu semblable, le phénomène a été appelé **effet** de serre, et les gaz absorbants qui le causent sont dits gaz à **effet** de serre.
Some of these **impacts** are already irreversible.	Certains de ces **effets** sont déjà irréversibles.

L'écart entre le nombre d'occurrences de l'équivalent privilégié et les autres équivalents d'un même CT est également important dans la catégorie 2b, à l'exception du CT *variation* et dans une moindre mesure du CT *piégeage* (Tableau 4.11).

Parmi tous les équivalents, des 4 parties du discours étudiées, les adverbes n'ont aucun représentant. Les noms sont les plus nombreux (46); suivent les adjectifs (14) et les verbes (10).

Dans cette catégorie, le nombre d'équivalents par CT varie beaucoup, 7 CT ont 2 équivalents, les autres en ont 3 et plus. Le CT *augmentation* est celui qui en possède le plus (10) : *increase, increase (to), increased*[69], *increasing, rise, rising, growth, higher, enhancement, enhanced.* Nous remarquons aussi que dans cette catégorie beaucoup d'équivalents ont moins de 10 occurrences, surtout dans le cas du CT *augmentation*[70].

Dans la section *autre* de cette catégorie, les occurrences varient de 2 à 193. Cinq CT ont plus de 25 occurrences dans *autre* : *échelle* (88), *effet* (193), *élévation* (27), *niveau* (42) et *variation* (27). Comme ces cas sont intéressants, nous allons les examiner d'un peu plus près ci-après.

Sur 234 occurrences analysées (voir la fiche *échelle* à l'Annexe B), le CT *échelle* est utilisé dans la préposition *à l'échelle* 63 fois, alors qu'en anglais, comme le montre l'exemple suivant (Tableau 4.15), la même idée est souvent rendue par un adverbe.

---

[69] Il s'agit de la forme adjectivale.

[70] Comme évoquée à la section 4.1.3, la limite fixée empiriquement à 4 occurrences pour attester un terme doit être rigoureusement respectée dans tous les cas.

Tableau 4.15 : Illustration de la préposition *à l'échelle*

It has since provided over 40,000 jobs **nationally**, through some 300 programmes, most of them for the poorest of the poor, and is one of the most successful examples anywhere that environmental change can be reversible.	Ce programme a permis la création de 40 000 emplois **à l'échelle nationale**, la plupart destinés aux plus déshérités. Son succès est un des exemples les plus parlants de la réversibilité du changement environnemental.

Sur 304 occurrences analysées (voir la fiche *effet* à l'Annexe B), le CT *effet* entre dans le syntagme *gaz à effet de serre* 137 fois; dans le syntagme anglais *greenhouse gases* (GHG), il n'est pas rendu. Ce CT entre également 21 fois dans la locution adverbiale *en effet,* dans ce cas-ci, il existe en anglais une foule de façon de rendre cette expression : *as the, it is, indeed, in fact, may, that is because, in some respect, in effect,* Ø, etc.

L'omission est ce qui fait monter le nombre d'occurrences à 27 dans la catégorie *autre* pour les CT *élévation* et *variation* (voir les fiches *élévation* et *variation* à l'Annexe B). *Élévation* n'est pas rendu 11 fois et *variation* 9 fois.

Sur 280 occurrences analysées (voir la fiche *niveau* à l'Annexe B), le CT *niveau* entre 19 fois dans le syntagme *au niveau de*, en anglais cette expression s'exprime de différentes façons : *in, of, with respect, on a basis, relative to, on a scale, in terms of.* En outre, le CT *niveau* n'est pas rendu en anglais 13 fois.

Finalement, il ne nous a pas été possible de dégager de règle permettant de prédire qu'un CT possède les caractéristiques de la catégorie 2b.

4.1.6 CT appartenant à la catégorie 2c

Dans la catégorie 2c (CT avec au moins un équivalent terme complexe et/ou un équivalent faisant partie d'une des composantes d'un nom composé, dont les éléments sont

accolés[71] (certains peuvent également avoir les caractéristiques de 2a et/ou 2b)), nous trouvons 9 CT. Le Tableau 4.16 est organisé comme les deux précédents. Il est à souligner que, d'une catégorie à l'autre, le nombre de critères va en augmentant et en se complexifiant.

Tableau 4.16 : Liste des CT de la catégorie 2c

Terme français	Équivalent anglais	Équivalents angl. / occurrences fr.	autre
anthropique	anthropogenic	75	12
	anthropogenically induced	4	
	human	28	
	human activities	7	
	human-induced	10	
	Total	124/136	
carbone	carbon	216	4
	co_2	5	
	Total	221/225	
dioxyde	dioxide	133	7
	co_2	25	
	Total	158/165	
eau	water	190	26
	ocean	7	
	sea	6	
	seawater	4	
	groundwater	7	
	freshwater	8	
	streamflow	5	
	Total	227/253	
gaz	gas	253	7
	GHG	15	
	Total	268/275	
latitude	latitude	143	5
	midlatitude	5	
	Total	148/153	
mer	sea	199	12
	seawater	8	
	offshore	4	
	Total	211/223	
réchauffement	warming	189	9

[71] Par exemple, l'équivalent *water* est inclus dans le terme simple *seawater*. Le terme *seawater* est considéré comme un nom composé parce qu'il est formé à partir des mots *sea* et *water*.

	warmer	8	
	warm (to)	4	
	GWP	6	
	Total	207/216	
serre	greenhouse	262	2
	GHG	16	
	Total	278/279	

Dans la catégorie 2c, il ne se trouve qu'un adjectif : *anthropique*. Par ailleurs, ce CT est le seul qui est rendu par des termes complexes : *anthropogenically induced, human-induced* et *human activities* (Tableau 4.17).

Tableau 4.17 : Illustration du CT *anthropique*

There are also some GHGs that are almost entirely due to **anthropogenic** sources.	D'autres GES proviennent presque entièrement de sources **anthropiques.**
The carbon cycle is an integral part of the climate system, and governs the build-up of atmospheric CO2 in response to **human** emissions.	Le cycle du carbone fait partie intégrante du système climatique et régule l'accumulation de dioxyde de carbone dans l'atmosphère en réponse aux émissions **anthropiques.**
...whether due to natural processes (El Nino episodes) or **human activities;**...	...qu'elle soit naturelle (épisodes El Nino) ou d'origine **anthropique;**...

Les 8 autres CT ont des équivalents qui font partie d'une des composantes d'un nom composé dont les éléments sont accolés. Par exemple, l'équivalent *water* du CT *eau* est compris dans les noms composés *seawater* (Tableau 4.18), *groundwater, freshwater*. Pour l'équivalent *streamflow* du même CT, l'idée d'eau est quant à elle rendue dans *stream* (Tableau 4.18).

Tableau 4.18 : Illustration du CT *eau*

(ii) the quantity and quality of forests, grazing lands, soils, fisheries, and **water** resources;	ii) la quantité et la qualité des forêts, des prairies, des sols, des ressources halieutiques et des ressources en **eau**;
Seawater in the high latitudes readily sinks, forming deep-water currents.	L'**eau** de mer dans les latitudes élevées coule facilement, formant les courants d'eau profonde.
Streamflow during seasonal low flow periods would decrease in many areas due to greater evaporation...	Pendant les périodes de basses eaux, le débit des cours d'**eau** devrait diminuer dans de nombreuses régions en raison d'une évaporation accrue...

L'équivalent du CT *dioxyde* se trouve inclus dans le symbole chimique CO_2 sous la forme de O_2 (dioxide). À l'inverse, l'équivalent du CT *carbone* est représenté dans le même symbole par *C* (carbon) (Tableau 4.19). De la même manière, l'équivalent *greenhouse* du CT *serre* peut être représenté par les lettres *GH* dans le sigle *GHG*, alors que l'équivalent *gas* du CT *gaz* est représenté dans le même sigle par la lettre *G* (Tableau 4.20).

Tableau 4.19 : Illustration des CT *dioxyde, carbone*

...are due primarily to the lower projected sulphur **dioxide** emissions in the SRES scenarios relative to the IS92 scenarios.	...est avant tout due aux prévisions d'un abaissement des émissions de **dioxyde** de soufre dans les scénarios SRES par rapport aux scénarios IS92.
The carbon cycle is an integral part of the climate system, and governs the build - up of atmospheric **CO_2** in response to human emissions.	Le cycle du carbone fait partie intégrante du système climatique et régule l'accumulation de **dioxyde** de **carbone** dans l'atmosphère en réponse aux émissions anthropiques.
The one-dimensional upwelling-diffusion model can be used as the oceanic part of the **carbon** cycle...	On peut utiliser le modèle unidimensionnel remontée-diffusion pour la partie océanique du cycle du **carbone**...

Tableau 4.20 : Illustration des CT *gaz* et *serre*

For ppm, this can be visualised as 1 cubic centimetre (cm^3) of **gas** per cubic metre of air.	Une ppm représente un centimètre cube (cm^3) de **gaz** par mètre cube d'air.
Similar approaches could be used for fluxes of non-CO_2 **greenhouse gases**.	On pourrait appliquer des méthodes analogues aux flux des **gaz** à effet de **serre** autres que le dioxyde de carbone.
What are the main driving forces of the **GHG** emissions in the scenarios?	Quelles sont les principales forces motrices des émissions de **gaz** à effet de **serre** dans les scénarios?
The change could be the result of an entirely natural process, such as an increase in solar radiation, or it could be a consequence of human actions, most notably the enhancement of the **greenhouse** effect.	le changement peut découler d'un processus entièrement naturel, comme une augmentation du rayonnement solaire, ou des activités humaines, en particulier du renforcement de l'effet de **serre**.

Parmi tous les équivalents de la catégorie 2c, des 4 parties du discours étudiées, nous obtenons 1 adverbe, *offshore* et 1 syntagme dans lequel se trouve un adverbe, *anthropogenically induced*. Les noms, au nombre de 23, sont les plus nombreux. Il n'y a que 3 adjectifs et 1 verbe. Pour cette catégorie, à l'exception du CT *anthropique*, l'écart entre l'équivalent privilégié et les autres équivalents d'un CT analysé est très marqué.

Dans la section *autre,* le nombre d'occurrences est de moins de 25, sauf pour le CT *eau* qui en a 27. toutefois, nous n'avons relevé rien de particulier à propos de ces occurrences, excepté que le CT *eau* n'est pas rendu 8 fois dans le texte anglais.

Pour la catégorie 2c, nous remarquons que les CT faisant partie d'un symbole chimique ou appartenant à des expressions fréquemment employées dans un texte scientifique, comme *gaz à effet de serre*, sont susceptibles d'appartenir à cette catégorie, car ces CT sont souvent rendus par une abréviation, un sigle ou un acronyme.

4.1.7 Observations générales

De cette analyse, il ressort que sur 50 CT analysés 15 trouvent place dans la catégorie 1 et 35 dans la catégorie 2.

À partir de 50 CT français étudiés, nous avons obtenu 128 équivalents anglais en comptant les équivalents appartenant à plusieurs CT (ex. l'équivalent *sea* est présent dans les analyses des CT *eau*, *mer* et *océan*), et 106 sans les compter.

Tous les CT, à l'exception de *fossile*, sont rendus au moins une fois par des anaphores, des abréviations avec une lettre, des omissions ou des reformulations. Par exemple : le terme *atténuation* de la phrase française « *...les mesures d'**atténuation** incluent...* » peut être rendu en anglais par une anaphore « ...these activities include... »; il arrive que le terme *atmosphère* soit omis dans le texte anglais; le terme *aérosol* est parfois rendu par *A* dans le corpus anglais; etc.

Nous avons également observé que pour un même CT les équivalents et leurs variantes morphologiquement apparentées sont souvent présents dans le corpus. De l'anglais vers le français, le même phénomène se produirait probablement, par exemple, pour le CT anglais *mitigate* nous aurions les équivalents français *atténuation, atténué* (adj.), *atténuer*.

Il est intéressant de noter que, dans les textes scientifiques, de nombreux équivalents sont des abréviations soit pour des éléments chimiques (*ozone* → O_3), soit pour des syntagmes revenant fréquemment dans les textes (*global warming potential* → *GWP*).

Nous avons remarqué tout au long de l'analyse que l'écart entre l'équivalent privilégié et les autres équivalents d'un même CT est souvent très marqué. Dans la section 4.2.2, nous exploiterons cet aspect.

En se basant à peu près sur le même modèle que l'énumération des problèmes d'établissement d'équivalence soulevés à la section 1.3, nous pouvons classer les résultats obtenus de la façon suivante :

1. Un CT simple dans une langue A peut avoir un seul équivalent dans une langue B (ex. *aérosol* se traduit toujours dans le corpus *Changement climatique* par *aerosol*) [72];
2. Un CT simple dans la langue A peut posséder plusieurs équivalents dans la langue B (ex. *précipitation* peut se rendre par *precipitation* ou par *rainfall*);
3. Un CT simple peut équivaloir à un CT complexe dans l'autre langue (ex. *inlandsis* se traduit par *ice sheet* dans le corpus anglais);
4. Un CT simple peut être exprimé par un équivalent appartenant à une autre catégorie grammaticale (ex. *...le système* ***climatique***... se rends par *...the climate* ***system***...);
5. Un CT simple dans une langue A peut être rendu par une anaphore, être omis, etc. (ex. *...les mesures d'atténuation incluent*...peut être traduit par *...these activities include*...).

[72] Du point de vue de l'extraction automatique, ce premier point n'est pas considéré comme un problème.

4.2 Position des équivalents dans la liste d'extraction des CT anglais

Dans cette section, nous donnons, dans un premier temps, les résultats du repérage des équivalents dans la liste d'extraction anglaise. Dans un deuxième temps, nous présentons les résultats du calcul des écarts entre les 50 CT français analysés et leurs équivalents respectifs.

4.2.1 Résultats du repérage des équivalents dans la liste d'extraction

En vue de vérifier si tous les équivalents sont présents dans la liste d'extraction anglaise et de calculer les écarts entre les 50 CT français et leurs équivalents respectifs, nous avons, pour chacun des équivalents, relevé dans la liste d'extraction anglaise le rang qu'il occupe. Le Tableau 4.21 présente ces résultats et par la même occasion montre des informations complémentaires. Dans la première colonne, nous rappelons la catégorie d'équivalent à laquelle le CT français appartient. Les deuxième, troisième et quatrième colonnes indiquent respectivement le CT français, son rang et le nombre d'occurrences. Les cinquième, sixième et septième colonnes indiquent, quant à elles, par ordre d'apparition, l'équivalent anglais, son rang et le nombre d'occurrences.

Tableau 4.21 : Position des équivalents anglais dans la liste d'extraction

RÉSULTATS GLOBAUX						
Catégorie d'équivalent	**CT français**	**Rang**	**Nombre d'occurrences**	**CT anglais**	**Rang**	**Nombre d'occurrences**
2a	climatique	1	3119	climate	1	5020
				climatic	52	312
1a	changement	2	3355	change	2	4556
1a	émission	3	2879	emission	3	3049
1a	température	4	1645	temperature	5	1733
2c	carbone	5	1310	carbon	8	1268

				co_2[73]	27	516
				co_2	130	152
1a	climat	6	1994	climate	1	5020
2c	serre	7	1313	greenhouse	9	1299
				GHG[74]	413	45
				GHGs	438	42
2c	gaz	8	1770	gas	10	1654
				GHG	413	45
				GHGs	438	42
2c	réchauffement	9	963	warming	16	853
				warm (to)	161	201
				warmer	1745	5
				GWPs	474	37
				GWP	776	17
1a	forçage	10	782	force[75]	47	645
				forcings	217	90
				forcing	426	44
1a	co_2	11	819	co_2	27	516
				co_2	130	152
2a	atmosphère	12	1195	atmosphere	15	1017
				atmospheric	20	658
1a	concentration	13	1111	concentration	12	1088
2b	effet	14	3091	effect	21	1362
				impact	14	1215
				affect (to)	73	590
1a	aérosol	15	624	aerosol	24	617
1a	écosystème	16	642	ecosystem	18	660
2b	atténuation	17	580	mitigation	22	606
				mitigate (to)	471	45
				moderate (to)	858	25
1a	scénario	18	1373	scenario	7	1228
2b	incidence	19	627	impact	14	1215
				implication	266	141
				effect	21	1362
				affect (to)	513	590
				incidence	884	27
2b	océan	20	810	ocean	13	1041
				oceanic	279	79
				sea	17	939

[73] TreeTagger attribue à CO_2 soit la partie du discours nom commun, soit la partie du discours nom propre, par conséquent il présente deux CT.

[74] Pour un mot qu'il ne reconnaît pas, TreeTagger ne réunit pas sous un même CT les singuliers et les pluriels.

[75] TreeTagger éprouve de la difficulté à désambiguïser forcing, force, forces, etc.

2b	modèle	21	1679	model	6	1669
				modelling	507	33
				pattern	119	320
2b	élévation	22	535	rise	72	442
				rising et rise (to)[76]	764	241
				increase	31	1100
				higher		absent[77]
2a	radiatif	23	437	radiative	33	427
				radiation	157	230
2b	atmosphérique	25	490	atmospheric	20	658
				atmosphere	15	1017
				air	309	426
				weather	134	384
2b	variation	26	704	change	2	4556
				variation	90	273
				shift	517	137
2b	précipitation	27	543	precipitation	30	480
				rainfall	89	218
1a	variabilité	28	362	variability	43	362
2c	dioxyde	29	394	dioxide	53	375
				co_2	27	516
				CO_2	130	152
2b	surface	30	838	surface	25	806
				area	175	779
2b	ozone	31	394	ozonophere[78]	60	385
				o_3	1091	10
2b	échelle	32	690	scale	55	478
				level	26	1363
				wide	1126	132
2a	stabilisation	33	437	stabilization et stabilisation[79]	41	394
				stabilize (to) et stabilise (to)	237	132
2c	eau	34	1439	water	34	1246
				ocean	13	1041
				sea	17	939
				seawater	663	28
				groundwater	530	32
				freshwater	321	67

[76] TreeTagger ne fait pas la distinction entre l'adjectif et le verbe.

[77] CT absent de la liste d'extraction anglaise.

[78] TreeTagger a lemmatisé *ozone* (anglais) par *ozonosphere*.

[79] TreeTagger lemmatise avec la graphie *stabilization* et *stabilize* les CT anglais *stabilisation et stabilise*.

				streamflow	611	25
2c	anthropique	35	289	anthropogenic	68	248
				human	70	710
				anthropogenetically induced	5	partiellement[80]
				human activities	?	partiellement
				human-induced	?	partiellement
2b	augmentation	36	1050	increase	31	1100
				increase (to)	23	1344
				increased	118	308
				increasing	1902	5
				rise	72	442
				rising	764	241
				growth	323	284
				higher		absent
				enhancement	423	53
				enhanced	264	160
2a	océanique	37	311	ocean	13	1041
				oceanic	279	79
1a	coût	38	1017	cost	32	1213
1a	fossile	39	303	fossil	57	324
2b	adaptation	41	611	adaptation	36	460
				adaptive	191	185
				adapt (to)	140	148
				adjustment	877	61
2c	latitude	42	330	latitude	50	318
				midlatitudes	1828	4
				midlatidude	2694	2
2a	naturel	43	806	natural	37	739
				naturally	1350	52
1a	combustible	44	325	fuel	102	452
2b	évaluation	45	588	assessment	79	330
				assess (to)	195	193
				estimate	92	370
				estimate (to)	302	284
				evaluation	1055	39
1a	côtier	46	294	coastal	54	341
2b	niveau	47	1312	level	26	1363
				target	347	193
2b	réduction	48	793	reduction	39	672
				reduced	375	108
				reduce (to)	64	813
				decrease	143	199
				cut		absent

[80] Pour cette extraction, TermoStat a été paramétré pour n'extraire que des CT simples.

2b	piégeage	49	193	sequestration	434	43
				capture	145	159
2b	météorologique	50	276	weather	134	384
				meteorological	487	43
2b	glaciaire	52	226	ice	86	709
				glacial	448	46
2c	mer	53	952	sea	17	939
				seawater	633	28
				offshore	1922	14

Une première observation importante en ce qui concerne les équivalents est qu'ils sont presque tous présents dans la liste d'extraction anglaise. *Higher* et *cut* sont les seuls équivalents qui manquent. Cela peut s'expliquer par le fait que ces unités lexicales sont très courantes dans le corpus de référence anglais utilisé par TermoStat. Par ailleurs, ces équivalents ne sont pas des équivalents privilégiés et leur nombre d'occurrences est très peu élevé dans l'analyse des CT *élévation, réduction et augmentation* (9 pour *higher*, 5 pour *cut*). Comme, dans ce projet, TermoStat a été paramétré pour n'extraire que des CT simples, il manque aussi les équivalents complexes, *anthropogenetically induced, human activities et human-induced.* Par contre, chacun des constituants de ces termes complexes fait partie de la liste d'extraction (Tableau 4.22) (*human* est présent dans le Tableau 4.21, puisque c'est également un équivalent par lui-même).

Tableau 4.22 : Constituants de CT complexes présents dans la liste d'extraction

CT	Rang	Occurrence
activity	76	603
induced	254	131
anthropogenetically	1593	5
human	70	710

On remarquera aussi que, pour diverses raisons, expliquées dans les notes de bas de page du Tableau 4.21, certains équivalents sont soit dédoublés (ex., CO_2), soit réunis (ex., *stabilization* et *stabilisation*).

4.2.2 Calcul des écarts entre les 50 CT français et leurs équivalents

Dans cette section, nous présentons en premier le Tableau 4.23 qui indique les écarts entre les 50 CT français et leurs équivalents, puis nous illustrons les résultats dans des graphiques.

Tableau 4.23 : Calcul des écarts entre les 50 CT français et leurs équivalents

CT français	Rang des CT français	Écart	Rang des CT anglais	CT anglais
climatique	1	0	1	climate
		+51	52	climatic
changement	2	0	2	change
émission	3	0	3	emission
température	4	+1	5	temperature
carbone	5	+3	8	carbon
		+22	27	co_2
		+125	130	co_2
climat	6	-5	1	climate
serre	7	+2	9	greenhouse
		+406	413	GHG
		+431	438	GHGs
gaz	8	+2	10	gas
		+405	413	GHG
		+430	438	GHGs
réchauffement	9	+7	16	warming
		+152	161	warm (to)
		+1736	1745	warmer
		+465	474	GWPs
		+767	776	GWP
forçage	10	+37	47	force
		+207	217	forcings
		+416	426	forcing
co_2	11	+16	27	co_2
		+119	130	co_2
atmosphère	12	+3	15	atmosphere
		+8	20	atmospheric
concentration	13	-1	12	concentration
effet	14	+7	21	effect
		0	14	impact
		+59	73	affect (to)
aérosol	15	+9	24	aerosol
écosystème	16	+2	18	ecosystem

atténuation	17	+5	22	mitigation
		+454	471	mitigate (to)
		+841	858	moderate (to)
scénario	18	-11	7	scenario
incidence	19	-5	14	impact
		+247	266	implication
		+2	21	effect
		+494	513	affect (to)
		+865	884	incidence
océan	20	-7	13	ocean
		+259	279	oceanic
		-3	17	sea
modèle	21	-15	6	model
		+486	507	modelling
		+98	119	pattern
élévation	22	+50	72	rise
		+742	764	rising et rise (to)
		+9	31	increase
radiatif	23	+10	33	radiative
		+134	157	radiation
atmosphérique	25	-5	20	atmospheric
		-10	15	atmosphere
		+284	309	air
		+109	134	weather
variation	26	-24	2	change
		+64	90	variation
		+491	517	shift
précipitation	27	+3	30	precipitation
		+62	89	rainfall
variabilité	28	+15	43	variability
dioxyde	29	+24	53	dioxide
		-2	27	co_2
		+101	130	co_2
surface	30	-5	25	surface
		+145	175	area
ozone	31	+29	60	ozone
		+1060	1091	o_3
échelle	32	+23	55	scale
		-6	26	level
		+1094	1126	wide
stabilisation	33	+8	41	stabilization et stabilisation
		+204	237	stabilize (to) et stabilise (to)
eau	34	0	34	water
		-21	13	ocean
		-17	17	sea
		+629	663	seawater
		+496	530	groundwater

		+287	321	freshwater
		+577	611	streamflow
anthropique	35	+33	68	anthropogenic
		+35	70	human
augmentation	36	-5	31	increase
		-13	23	increase (to)
		+82	118	increased
		+1866	1902	increasing
		+36	72	rise
		+728	764	rising
		+287	323	growth
		+387	423	enhancement
		+228	264	enhanced
océanique	37	-24	13	ocean
		+242	279	oceanic
coût	38	-6	32	cost
fossile	39	+18	57	fossil
adaptation	41	-5	36	adaptation
		+150	191	adaptive
		+99	140	adapt (to)
		+836	877	adjustment
latitude	42	+8	50	latitude
		+1786	1828	midlatitudes
		+2652	2694	midlatidude
naturel	43	-6	37	natural
		+1307	1350	naturally
combustible	44	+58	102	fuel
évaluation	45	+34	79	assessment
		+150	195	assess (to)
		+47	92	estimate
		+257	302	estimate (to)
		+1010	1055	evaluation
côtier	46	+8	54	coastal
niveau	47	-21	26	level
		+300	347	target
réduction	48	-9	39	reduction
		+327	375	reduced
		+16	64	reduce (to)
		+95	143	decrease
piégeage	49	+385	434	sequestration
		+96	145	capture
météorologique	50	+84	134	weather
		+437	487	meteorological
glaciaire	52	+34	86	ice
		+396	448	glacial
mer	53	-36	17	sea
		+580	633	seawater
		+1869	1922	offshore

Dans la Figure 4.1, nous illustrons à l'aide d'un graphique l'ensemble des écarts des équivalents et leur CT respectif (dans le graphique, les CT sont écrits au long à tous les deux CT). Toutefois, à l'analyse du graphique, nous constatons que la présentation de l'ensemble des écarts ne permet pas de dégager des observations suffisamment significatives.

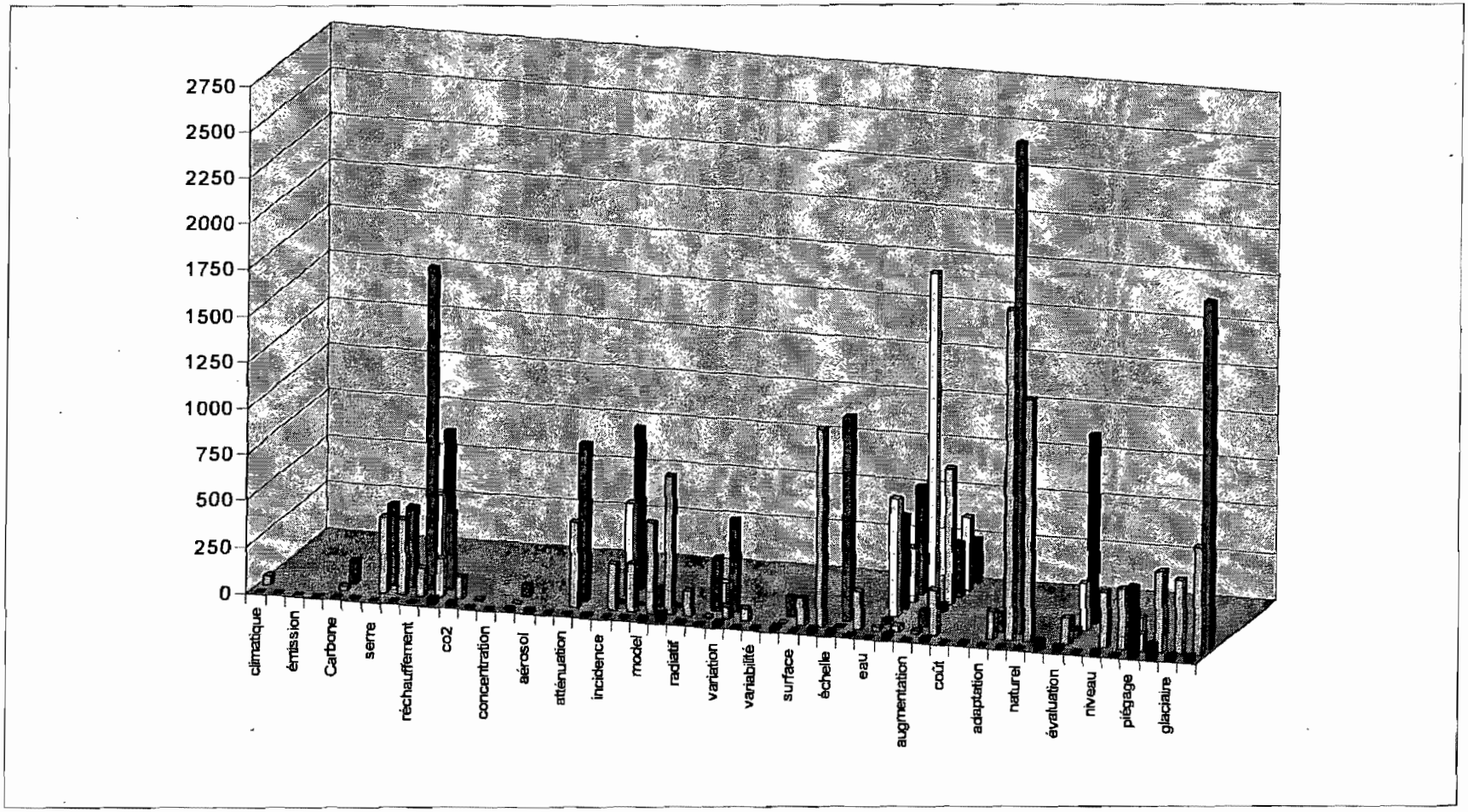


Figure 4.1 : Ensemble des écarts des équivalents et leur CT respectif

Maintenant, si nous considérons seulement les écarts entre les équivalents privilégiés et leur CT respectif, nous obtenons un graphique beaucoup plus intéressant (Figure 4.2), dans lequel 47 équivalents montrent un écart avec leur CT respectif ne dépassant pas 50 (Tableau 4.24). Trois équivalents seulement ont un écart supérieur à 50, le CT *piégeage* et son équivalent *sequestration* présentent le plus grand écart, +385, suivi du CT *météorologique* et de son équivalent *weather* (+84) et du CT *combustible* et de son équivalent *fuel* (+58).

Tableau 4.24 : Nombre d'équivalents classés par tranche d'écart

Nombre d'équivalents par tranche d'écart	Valeur des tranches d'écart
4	0
19	0 – 5
30	0 – 10
40	0 – 25
47	0 – 50

Tout en gardant à l'esprit que nous n'avons analysé que 50 CT sur plus de 3 000 et que ces CT se trouvent en début de liste, nous constatons que les équivalents privilégiés sont peu distants des CT français, 40 d'entre eux ont un écart de 25 ou moins avec leur CT respectif. Par contre, par rapport aux autres équivalents pour un même CT, les équivalents privilégiés ne présentent pas toujours l'écart le plus faible, par exemple, l'équivalent privilégié *effect* (+7) du CT *effet* possède un écart plus élevé que l'équivalent *impact* (0) qui vient en deuxième position (Tableau 4.23). Nous avons constaté le même phénomène pour les CT *incidence, océan, élévation, dioxyde, échelle* et *piégeage*. Ce qui veut dire que 7 CT sur 50 n'ont pas pour équivalent privilégié celui dont l'écart est le plus faible. Par conséquent, même si la valeur de l'écart[81] peut se révéler utile, employée seule, elle ne suffit pas à attester un équivalent privilégié.

[81] Dans le cas de notre étude, le pourcentage de CT dont la valeur de l'écart est la plus élevée pour l'équivalent privilégié est de 86 %.

Compte tenu de ce qui précède, nous nous sommes demandé si la cognation, qui joue un rôle important dans l'alignement des textes et l'extraction des termes, pourrait servir à l'identification des équivalents privilégiés et se combiner à la valeur de l'écart pour contribuer à l'amélioration de l'attestation des équivalents privilégiés.

Ainsi, parmi les 127 équivalents des 50 CT français analysés, nous avons 48 cognats et 79 non-cognats. Le Tableau 4.25 présente les cognats relevés parmi les équivalents anglais (en gras figurent les équivalents privilégiés cognats). La première colonne indique le type de cognat; la deuxième, le nombre de cognats; la troisième, l'équivalent anglais. Les cognats représentent 38 % du total des équivalents. Sur les 50 équivalents privilégiés, nous avons 28 cognats, soit 56 %. Ce dernier chiffre, nous indique que la cognation est un indice moins performant que celui du calcul des écarts pour la portion de CT que nous avons étudiée. Par exemple, l'équivalent cognats *incidence* du CT français *incidence* n'occupe que le 5^{e} rang des équivalents de ce CT, le premier étant *impact*. La même observation s'applique aux cognats qui ne sont pas en gras dans le Tableau 4.25.

Tableau 4.25 : CT et équivalents cognats

Type de cognat	Nombre de cognats	Équivalent anglais cognat de CT français
Cognats complètement identiques	9	**adaptation, concentration, co_2,** incidence, **latitude, ozone,** stabilisation, **surface,** variation
Cognats identiques, mais avec accentuation en français	9	**aerosol, atmosphere, emission,** evaluation, **ocean, precipitation, reduction, scenario, temperature**
Cognats avec des lettres en plus, en moins ou différentes	30	adaptive, affect, affect (to), **anthropogenic, atmospheric, carbon, change, climate,** climatic, cost, **dioxide, ecosystem, effect, fossil,** glacial, meteorological, midlatitude, **model,** modelling, **natural,** naturally, oceanic, radiation, **radiative,** reduced, reduce (to), stabilise (to), stabilize (to), **stabilization, variability**
Total	**48**	

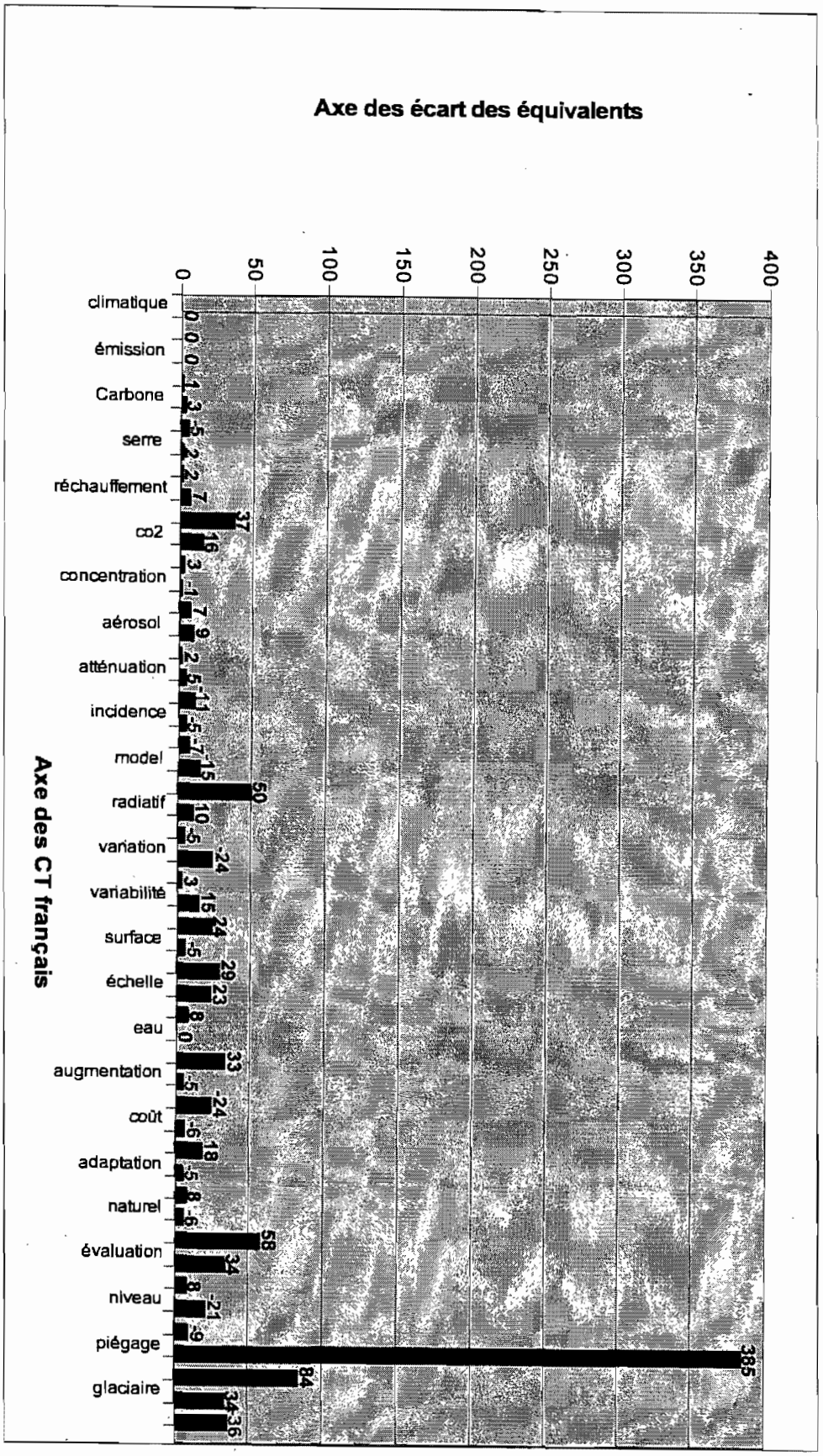


Figure 4.2 : Écarts entre les équivalents privilégiés et leur CT respectif

4.3 Comparaisons de l'analyse manuelle à l'extraction lexicale d'Alinea

Dans cette section, nous comparons notre liste d'équivalents compilés manuellement aux équivalents proposés par l'extraction lexicale d'Alinea afin d'analyser le bruit et le silence généré par ce dernier. Nous espérons ainsi que notre analyse manuelle mette en évidence certaines des limites de l'appariement de CT qu'une évaluation directe des résultats n'aurait pas permis de révéler.

Comme indiqué à la section 3.1.7.1, il est possible de paramétrer l'extraction lexicale. En conservant le paramétrage par défaut, nous avons obtenu presque uniquement les équivalents privilégiés, ce résultat nous semble beaucoup trop silencieux. En réglant les paramètres à zéro, nous avons recueilli trop de bruit. Pour obtenir des résultats intéressants, nous nous sommes guidé sur les résultats obtenus dans notre analyse des équivalents pour paramétrer le module d'extraction lexicale de la façon suivante : nombre maximum d'occurrences 4, pourcentage d'occurrences 0,013[82].

Les résultats de notre paramétrage sont représentés dans le Tableau 4.26, la première colonne contient les CT français, la deuxième représente la compilation manuelle des équivalents anglais, la troisième colonne signale la présence de l'équivalent dans l'extraction lexicale d'Alinea et la quatrième colonne indique les propositions d'Alinea absentes de la compilation manuelle.

[82] 0,013 représente le pourcentage minimum d'occurrences.

Tableau 4.26 : Comparaison des équivalents compilés manuellement aux équivalents produits par Alinea

CT français	Équivalents anglais	Présent (✓) dans l'extraction lexicale d'Alinea	Autre proposition d'Alinea
climatique	climate	✓	
	climatic	✓	
changement	change	✓	change (to) climate[82]
émission	emission	✓	
température	temperature	✓	
carbone	carbon	✓	carbon-free
	co_2		
climat	climate	✓	climatic
serre	greenhouse	✓	
	GHG	✓	
gaz	gas	✓	
	GHG		
réchauffement	warming	✓	heating
	warm (to)	✓	
	warmer	✓	
	GWP		
forçage	forcing	✓	
co_2	co_2	✓	non-CO_2
atmosphère	atmosphere	✓	
	atmospheric	✓	
concentration	concentration	✓	
effet	effect	✓	greenhouse
	impact	✓	
	affect (to)		
aérosol	aerosol	✓	sulphate
écosystème	ecosystem	✓	
atténuation	mitigation	✓	mitigative
	mitigate (to)		
	moderate (to)		
scénario	scenario	✓	emission SRES
incidence	impact	✓	
	implication	✓	
	effect		

[82] Les équivalents surlignés en gris constituent du bruit, en ce sens qu'ils n'ont pas été relevés dans notre étude.

	affect (to)		
	incidence	✓	
océan	ocean	✓	
	oceanic	✓	
	sea		
modèle	model	✓	
	modelling		
	pattern		
élévation	rise	✓	sea-level
	rising	✓	sea
	rise (to)	✓	low-lying
	increase		
	higher		
radiatif	radiative	✓	
	radiation	✓	
atmosphérique	atmospheric	✓	
	atmosphere	✓	
	air	✓	
	weather		
variation	change	✓	
	variation	✓	
	shift		
précipitation	precipitation	✓	
	rainfall	✓	
variabilité	variability	✓	
dioxyde	dioxide	✓	
	co_2		
surface	surface	✓	
	area		
ozone	ozone	✓	ODSs
	o_3	✓	ozone-depleting
			depletion
échelle	scale	✓	large-scale
	level	✓	globally
	wide		time-scale
			global
			timescale
			larger-scale
stabilisation	stabilization	✓	
	stabilisation	✓	
	stabilize (to)	✓	
	stabilise (to)	✓	
eau	water	✓	
	ocean		
	sea		
	seawater		
	groundwater		
	freshwater		
	streamflow		

anthropique	anthropogenic	✓	human-induced anthropogenically
	human	✓	
	anthropogenetically induced	non considéré[83]	
	human activities	non considéré	
	human-induced	non considéré	
augmentation	increase	✓	
	increase (to)	✓	
	increased	✓	
	increasing	✓	
	rise		
	rising		
	growth		
	higher		
	enhancement		
	enhanced		
océanique	ocean	✓	
	oceanic	✓	
coût	cost	✓	
fossile	fossil	✓	fossil-fuel non-fossil
adaptation	adaptation	✓	
	adaptive	✓	
	adapt (to)	✓	
	adjustment		
latitude	latitude	✓	mid-latitude high-latitude latitudinal
	midlatitude		
naturel	natural	✓	
	naturally	✓	
combustible	fuel	✓	
évaluation	assessment	✓	SAR valuation
	assess (to)		
	estimate		
	estimate (to)		
	evaluation	✓	
côtier	coastal	✓	
niveau	level	✓	
	target		
réduction	reduction	✓	abatement
	reduced	✓	
	reduce (to)	✓	
	decrease		
	cut		
piégeage	sequestration	✓	capture (to)

[83] Alinea n'étant pas en mesure d'extraire des termes complexes, nous n'avons pas considéré ces CT.

	capture	✓	trap (to)
météorologique	weather	✓	weather-related
	meteorological	✓	event
			instrumental
glaciaire	ice	✓	sheet
	glacial	✓	
mer	sea	✓	sea-level
	seawater	✓	
	offshore		

Sur 124[85] équivalents compilés manuellement, Alinea en propose 86, soit 69 %. Même s'il n'est pas très élevé, ce résultat est néanmoins très intéressant. Il est à remarquer qu'Alinea propose tous les équivalents privilégiés. Par ailleurs, il les présente toujours en première position, sauf l'équivalent privilégié *sequestration*[86]. Par contre, il éprouve de la difficulté à extraire des équivalents dont le nombre d'occurrences est peu élevé.

En ce qui concerne les silences (31 %), on se souviendra que, pour aller chercher le maximum de CT n'ayant qu'un seul équivalent dans le corpus, nous avons fixé pour attester un équivalent le nombre d'occurrences à 4. Si dans le corpus il n'existe qu'un nombre d'occurrences peu élevé pour un équivalent, Alinea risque de ne pas le proposer si l'alignement au niveau des mots présente trop de bruit. Par exemple, dans la paire de phrase du Tableau 4.27, l'équivalent *sea* (en gras) a été aligné avec le mot *élévation* (souligné) au

[85] À l'origine nous avions 127 équivalents compilés manuellement, mais, comme Alinea n'est pas programmé pour proposer des termes complexes, nous n'avons pas considéré *anthropogenetically induced, human activities* et *human-induced.*

[86] L'équivalent *sequestration* apparaît dans plusieurs textes du corpus, mais il est moins fréquent que l'équivalent *capture* qui se trouve principalement dans un texte du corpus. Lorsque nous avons composé notre échantillon de corpus pour procéder à l'analyse des 50 CT, nous n'avons prélevé qu'un maximum de 10 contextes par texte. Ainsi, dans notre échantillon, l'équivalent *sequestration* est devenu l'équivalent privilégié et l'équivalent *capture* a été placé en deuxième position. Comme Alinea extrait les équivalents de tout le corpus, *capture* est l'équivalent privilégié puisqu'il apparaît en plus grand nombre que *sequestration.*

lieu du CT *océans* (en gras). Ce cas de figure se présente suffisamment souvent pour les équivalents de faible occurrence.

Tableau 4.27 : Exemple d'équivalent mal apparié

The overall impact of **sea** surface temperature increase and elevated CO2 concentrations...	Les effets conjugués de l'élévation thermique à la surface des **océans** et de l'augmentation des concentrations de CO2...

Le silence est aussi imputable aux CT qui ont des équivalents compris dans : 1) des noms composés dont les éléments sont accolés (*eau/seawater*); 2) des symboles (*dioxyde/*CO_2) ou 3) des sigles (*gaz/GHG*). Dans ces cas de figure, le logiciel ne peut pas convenablement apparier les CT à leur équivalent en raison de la différence de structure qui existe entre eux.

Par contre, les propositions d'Alinea contiennent peu de bruit (13 CT, soit 10 %) et la plupart du temps elles s'expliquent par le fait que le logiciel éprouve de la difficulté à choisir le bon cooccurrent, par exemple il propose parfois *climate* au lieu de *change* et *greenhouse* au lieu de *effect*[87].

Étant donné qu'Alinea a extrait les équivalents de tout le corpus, il présente des équivalents que nous n'avons pas répertoriés dans notre liste d'équivalents compilée manuellement à partir d'un échantillon du corpus, soit parce que leur nombre était inférieur à 4 dans l'échantillon (ex. *heating, valuation, etc.*), soit parce qu'ils y étaient absents (*mitigative, instrumental*).

Chose intéressante, Alinea présente des équivalents avec des traits d'union (15 CT). Avec Alinea, nous avons fait appel à la version Windows de TreeTagger. Cette version ne

[87] Dans le tableau, nous avons surligné en gris les équivalents présentant cette caractéristique.

segmente pas les CT avec un trait d'union. Tandis que la version Linux utilisée par TermoStat segmente aux traits d'union.

Enfin, on observera que, tout comme dans notre analyse, Alinea propose des équivalents dont la partie du discours est différente de celle du candidat terme (ex. pour le CT *climatique* on obtient l'équivalent *climate*).

Conclusion

L'objectif du présent mémoire visait à observer l'équivalence des termes en corpus afin d'étudier les possibilités et les difficultés que présente la constitution d'une nomenclature bilingue à partir de listes de candidats termes extraits automatiquement. Dans les paragraphes qui suivent, nous passons en revue les étapes de notre étude et nous apportons nos commentaires en ce qui concerne les résultats obtenus, les limites de l'étude et les perspectives de travail envisagées.

Dans la première étape de notre méthodologie pour étudier l'équivalence en corpus, nous avons mis en forme, à l'aide de 31 paires de textes, un corpus parallèle spécialisé de plus 500 000 mots pour l'anglais et de plus de 600 000 mots pour le français. Le domaine spécialisé choisi pour cette étude portait sur le changement climatique. Pour permettre un alignement plus précis du corpus, ce dernier a été prétraité. Avec Alinea (Kraif 2001), nous avons procédé à deux alignements, un premier au format texte au niveau des phrases et un deuxième au format .ttg jusqu'au niveau des mots. Nous avons poursuivi par l'extraction de termes français et anglais avec le logiciel d'acquisition de termes TermoStat (Drouin 2002). Nous avons ainsi obtenu une liste de 3 703 CT français et une liste de 2 906 CT anglais. Ces listes ont fait l'objet d'une analyse et d'un nettoyage.

La deuxième partie de notre méthodologie a consisté à décrire les trois étapes de l'analyse. Le nombre des CT proposés par TermoStat étant beaucoup trop élevé pour être étudié dans ce mémoire, nous avons choisi d'analyser les 50 premiers CT français apparaissant en début de liste. Les CT placés en début de liste sont les plus susceptibles d'être des termes. Afin d'obtenir un échantillonnage représentatif du corpus, pour ces 50 CT, nous avons sélectionné dans les 31 paires de textes jusqu'à 310 paires de contextes par CT. Pour chacun des CT, nous avons relevé les équivalents observés en corpus. Chaque CT a reçu sa propre fiche dans laquelle nous avons consigné les résultats. Dans le but de faciliter la description des CT français et leurs équivalents anglais, nous avons établi une classification des CT par rapport aux équivalents trouvés en corpus. Cette classification est basée sur le nombre d'équivalents possédé par chaque CT et sur le type des équivalents.

Nous avons par la suite repéré la position des équivalents dans la liste d'extraction anglaise et calculé les écarts entre les CT français et leurs équivalents. Ces deux dernières étapes consistaient à vérifier si tous les équivalents se trouvaient dans la liste d'extraction anglaise et à quel niveau par rapport aux CT français. Enfin, nous avons comparé la compilation manuelle des équivalents avec l'extraction lexicale d'Alinea.

En ce qui concerne les résultats de l'identification des équivalents anglais, nous avons constaté que 15 CT possédaient un seul équivalent et que 35 CT en avaient plusieurs. Parmi les CT n'ayant qu'un seul équivalent, aucun ne présentait un équivalent appartenant à une autre partie du discours et aucun n'avait un équivalent complexe. Par contre, parmi les CT possédant plusieurs équivalents, nous avons dénombré dans une première sous-catégorie six CT dont les équivalents appartenaient à une partie du discours différente, mais morphologiquement apparentés. Dans une deuxième sous-catégorie, nous avons identifié 20 CT (soit la majorité) dont les équivalents étaient morphologiquement différents et présentaient la caractéristique de relever de plusieurs parties du discours. Dans une troisième sous-catégorie, nous avons placé 9 CT possédant au moins un équivalent terme complexe et/ou nom composé. De cette analyse, il ressort que l'équivalence est un problème complexe même en terminologie. Nous n'avons pas pu dégager de règles permettant de prédire ce qui prédispose un terme à n'avoir qu'un équivalent et un autre à en avoir plusieurs. Nous avons observé que les problèmes d'établissement de l'équivalence de L'Homme (2004) sont identiques les termes simples et pour les termes complexes, excepté le point 2 évidemment. Nous avons observé que le nombre d'occurrences d'un équivalent privilégié se démarque presque toujours de façon importante par rapport aux autres équivalents. Il est à noter que parmi les 50 premiers CT, les verbes et les adverbes n'étaient pas présents.

Nous avons ensuite présenté les résultats du repérage de la position des équivalents dans la liste d'extraction anglaise et le calcul des écarts entre les CT français et leurs équivalents. Nous avons observé que seuls deux équivalents n'étaient pas présents dans la

liste d'extraction anglaise, soit *higher* et *cut*. Par contre, ces deux CT ne sont pas des équivalents privilégiés. Ce résultat est d'autant plus intéressant que le corpus anglais d'analyse est de nature très différente du corpus d'analyse français : 1) sur le plan de la date de parution des textes qui les composent, le premier date de 1987 et le deuxième de 2002; 2) sur le nombre de mots qu'ils contiennent, le corpus anglais comprend 7 millions de mots alors que le corpus français en compte 30 millions. À l'égard du calcul des écarts, nous avons constaté que les équivalents privilégiés étaient relativement peu éloignés des CT français. Cette dernière constatation doit être cependant prise avec réserve, car nous n'avons analysé que les 50 premiers CT, c'est-à-dire ceux dont le poids est le plus élevé. Nous avons également constaté que la cognation ne permet pas toujours de déceler les équivalents privilégiés.

Enfin, nous avons donné les résultats de la comparaison de la liste compilée manuellement avec les équivalents obtenus à l'aide de l'extraction lexicale d'Alinea. Sur 124 équivalents compilés manuellement, Alinea a proposé 86 équivalents, ce qui représente un pourcentage de 69 %. Il a été constaté qu'Alinea produisait toujours les équivalents privilégiés, mais qu'il éprouvait de la difficulté à extraire les équivalents dont le nombre d'occurrences est faible (Alinea a extrait de tout le corpus). La comparaison a également permis de voir qu'Alinea présentait des équivalents valides qui étaient absents ou trop peu nombreux dans l'échantillon sur lequel nous avons travaillé et qu'il produisait peu de bruit.

À partir de ces résultats, il nous est maintenant possible de faire part de nos observations et de proposer des perspectives de travail. Avant de passer aux observations générales, nous nous attardons sur les étapes de l'analyse.

Relativement à l'identification des équivalents anglais, même si nous n'avons travaillé que sur un petit nombre de CT, elle nous a permis de montrer que même en terminologie l'équivalence est un problème complexe. Bien sûr, pour obtenir un portrait plus juste, il faudrait travailler sur un plus grand nombre de CT. Comme les verbes et les adverbes étaient absents des 50 CT analysés, une autre étude pourrait se pencher sur

l'identification de leurs équivalents. Enfin, il serait intéressant de voir parmi les équivalents la place occupée par la synonymie et la polysémie.

En ce qui concerne les résultats du repérage de la position des équivalents dans la liste d'extraction anglaise, il serait intéressant d'étendre l'analyse à tous les CT afin de calculer le pourcentage d'équivalents anglais présents dans la liste anglaise. Il faudrait aussi comparer les écarts entre un plus grand nombre de CT et leur équivalent privilégié pour examiner le comportement de l'écart au fur et à mesure que le poids des CT diminue. Par ailleurs, nous pensons qu'en améliorant l'étiquetage du corpus et en utilisant un corpus de référence parallèle (ex. Hansard), nous obtiendrions des listes de termes anglais et français possédant un contenu terminologique plus équivalent, plus « parallèle ».

À l'égard des résultats de la comparaison de la liste compilée manuellement avec les équivalents obtenus à l'aide de l'extraction lexicale d'Alinea, nous trouvons les résultats très intéressants. Il aurait été toutefois intéressant de pouvoir comparer des listes étiquetées avec la même version de TreeTagger. Tel quel, le module d'extraction lexicale peut déjà rendre de grands services aux terminologues. À notre avis, le silence observé n'est pas a priori un inconvénient. Souvenons-nous que dans notre étude, nous avions fixé à 4 occurrences le nombre d'équivalents pour l'attester. Dans un travail terminologique appliqué, il est probable que ce nombre serait augmenté. Afin de rendre ce module beaucoup plus convivial pour le terminologue, il serait utile toutefois d'y apporter quelques modifications. Nous suggérons notamment la création dans la boîte de dialogue *Extraction de lexique bilingue* d'un champ permettant l'importation de listes de termes simples préalablement extraites à l'aide de logiciels d'acquisition de termes afin d'obtenir les équivalents de ces derniers.

Dans de futurs travaux, on pourrait mettre en place une méthode plus conviviale d'extraction bilingue de termes simples. Pour valider des termes, les besoins des terminologues sont multiples. Il serait par exemple utile de mettre au point une interface offrant la possibilité de consulter toutes les étapes du processus d'extraction de termes

unilingue ou bilingue, de la constitution du corpus à la validation des termes. En effet, il est important pour un terminologue de pouvoir revenir dans les textes originaux, de vérifier l'étiquetage, de trier les termes, d'extraire les contextes, de facilement produire la bibliographie des contextes servant à décrire les termes, etc.

Par notre expérience en extraction de termes complexes et en extraction de termes simples, nous trouvons qu'il est plus facile de travailler avec des termes simples. Les listes d'extraction de ces derniers comptent moins de CT que celles des termes complexes. Elles sont à notre avis beaucoup plus gérables : au lieu d'effectuer un travail de « débroussaillage », nous réalisons un travail de construction. D'ailleurs, en partant de cette dernière idée et en guise de perspective pour de futurs travaux, il serait intéressant de se pencher sur le problème de l'extraction de termes complexes à partir de l'extraction de termes simples. Ces derniers combinés à des patrons morphosyntaxiques pourraient servir d'amorces à l'identification de termes complexes.

Bibliographie

ACADÉMIE FRANÇAISE (1992). *Dictionnaire de l'Académie française,* 9e éd., Revue, [en ligne]. http://atilf.atilf.fr/academie9.htm (page consultée le 15 octobre 2007).

AHMAD, K. (1996). *Language Engineering and the Processing of Specialist Terminology,* [en ligne]. http://www.computing.surrey.ac.uk/ai/pointer/paris.html (page consultée le 16 juin 2007).

AHRENBERG, L., M. ANDERSSON et M. MERKEL (1998). « A simple hybrid aligner for generating lexical correspondences in parallel texts », dans *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, Canada, p. 29–35.

AHRONIAN, C. (2005). *Les noms composés anglais, français et espagnol du domaine d'Internet,* Thèse de doctorat en Lexicologie et Terminologie Multilingue – Traduction, Université Lumière Lyon 2.

ALEGRIA I., A. GURRUTXAGA, P. LIZASO, X. SARALEGI, S. UGARTETXEA et R. URIZAR (2004). « A Xml-Based Term Extraction Tool for Basque », dans *4 th International Conference On Language Resources And Evaluation (LREC'04),* Lisbonne, Portugal.

BOURIGAULT, D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*, Thèse en Mathématiques, Informatique Appliquée aux Sciences de l'Homme, Paris, École des Hautes Études en Sciences Sociales.

BOURIGAULT, D. et C. JACQUEMIN (2000). « Construction de ressources terminologiques », dans J.-M. PIERREL (dir.), *Ingénierie des langues*, Hermès, p. 215–233.

BOURIGAULT, D. et M. SLODZIAN (1999). « Pour une terminologie textuelle », Terminologies Nouvelles Spécial TIA, 19, p. 29–32.

BOWKER, L. et J. PEARSON (2002). *Working with Specialized Language: A Practical Guide to Using Corpora,* London, New York, Routledge.

BREW, C et D. MCKELVIE (1996). « Word-pair extraction for lexicography », dans *Proceedings of International Conference on New Methods in Natural Language Processing*, Bilkent, Turquie, p. 45–55.

BROWN, P., J. LAI, R. MERCER (1991). « Aligning Sentences in Parallel Corpora », dans *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, ACL-91, Morristown, NJ, p. 169–176.

BROWN, P. F., S. A. D. PIETRA, V. J. D. PIETRA et L. R. MERCER (1993). « The mathematics of statistical machine translation: parameter estimation », *Computational Linguistics*, 19(2), p. 263–311.

CABRÉ, M. T., R. ESTOPÀ et J. VIVALDI (2001). « Automatic Term Detection: A review of current systems », dans BOURIGAULT, D., C. JACQUEMIN et M.-C. L'HOMME (dir.). *Recent advances in Computational Terminology*, Amsterdam/Philadelphia, John Benjamins, p. 53–87.

CARREÑO, I. (2004). *Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingues*, Mémoire de maîtrise, Département de linguistique et de traduction, Université de Montréal.

CARREÑO, I., P. DROUIN et M.-C. L'HOMME (2006). *Current Methods for Automatic Term Extraction and A Review of Existing Term-Extraction Systems for English*, Montreal, Observatoire de linguistique Sens-Texte (OLST).

CARREÑO, I. et M.-C. L'HOMME (2007). *Detailed Description and Evaluation of five Automatic Term-Extraction Systems for English.* Montreal, Observatoire de linguistique Sens-Texte (OLST).

CARREÑO, I., A. LE SERREC et M. BOUDREAU (2007). À paraître. « Evaluating automatic terminology extraction methods for the construction of an ontology ».

CARRERAS, X., I. CHAO, L. PADRÓ et M. PADRÓ (2004). « FreeLing : An open-source suite of language analyzers », dans *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04),* Lisbonne, Portugal.

CARRIÈRE, I. (2006). *Adjectifs dérivés de noms : analyse en corpus médical et modèle d'encodage terminologique*, Mémoire de maîtrise, Département de linguistique et de traduction, Université de Montréal.

CHIAO, Y.C., O. KRAIF, D. LAURENT, T. M. H. NGUYÊN, N. SEMMAR, F. STUCK, J. VÉRONIS et W. ZAGHOUANI (2006). « Evaluation of multilingual text alignment systems: the ARCADE II project », dans *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC'06)*, Gène, Italie, p. 1975–1978.

CHOUEKA, Y. (1988). « Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases », dans *Proccedings of the conférence « User-Oriented Context Based Text And Image Handling »*, (RIAO'88), Cambridge, p. 609–623.

CHURCH, K. (1993). « Char align : A program for Aligning Parallel Texts at the Character Level », dans *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-93,* Columbus, Ohio, p. 1–8.

CHURCH, K. et P. HANKS (1990). « Word association norms, mutual information, and lexicography », *Computational Linguistics* 16(1), p. 22–29.

CHURCH, K., W. GALE, P. HANKS et D. HINDLE (1991). « Using statistics in lexical analysis », dans U. Zernik (dir.), *Lexical Acquisition*, Englewood Cliff, NJ, Erlbaum, 115–164.

DAGAN I. et K. CHURCH (1994). « Termight : Identifying and translating technical terminology », dans *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, Stuttgart, p. 34–40.

DAGAN, I., K. CHURCH et W. GALE (1993). « Robust Bilingual Word Alignment for Machine Aided Translation », dans *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, p. 1–8.

DAILLE, B. (1994). *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*, Thèse de doctorat, Paris, Université Paris-7.

DAILLE, B., É. GAUSSIER et J.-M. LANGÉ (1994). « Towards automatic extraction of monolingual and bilingualterminology », dans *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japon, p. 515–521.

DAVID, S. et P. PLANTE (1990). « De la nécessité d'une approche morpho-syntaxique en analyse du texte », *Intelligence artificielle et sciences cognitives au Québec,* 2(3), p. 140–155.

DEBILI, F., et A. ZRIBI (1996). « Les dépendances syntaxiques au service de l'appariement des Mots », dans *Actes du 10e Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, Rennes, France, p. 81–90.

DEJEAN, H. et É. GAUSSIER (2002). « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », *Lexicometrica*, N° spécial 2002, p. 1–21.

DROUIN, P. (2002). *Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés*, Thèse de doctorat, Département de linguistique et de traduction, Université de Montréal.

DUBUC, R. (2002). *Manuel pratique de terminologie,* 4e édition, Montréal, Linguatech.

DUNNING, T. (1993). « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics*, 19(1), p. 61–74.

ENGUEHARD, C. (1992). *Acquisition Naturelle Automatique d'un réseau sémantique*, Thèse en Contrôle des Systèmes, Université de Technologie de Compiègne.

ESTOPÀ, R. (2001). « Les unités de signification spécialisées : élargissant l'objet du travail en terminologie » dans Terminology, 7(2), p. 217–237.

FELBER, H. (1987). *Manuel de terminologie*, Paris, Organisation des Nations Unies pour l'éducation, la science et la culture.

FRANTZI, K. T. (1998). *Automatic Recognition of Multi-Word Terms*, Thèse de doctorat, Manchester Metropolitan University dep. Of Computing and Mathematics.

FRANTZI, K. T. et S. ANANIADOU (1999). « The C-value / NC-value domain independent method for multi-word term extraction », *Journal of Natural Language Processing*, 6(3), p. 145–179.

FUNG, P. (1998). « A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora », dans *Lecture Notes in Artificial Intelligence AMTA 98*, Springer Publisher, p. 1–17.

FUNG, P. et K. CHURCH (1994). « K-vec : A new approach for aligning parallel texts », dans *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, p. 1096–1102.

GALE, W. et K. W. CHURCH (1991). « A program for aligning sentences in bilingual corpora », dans *Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, California, p. 177–184.

GAUSSIER, É. (1998). « Flow network models for word alignment and terminology extraction from bilingual corpora », dans Boitet, C. (dir.), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguistics (COLING)*, Montréal, Canada, p. 444–450.

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

GAUSSIER, E. et J.-M. LANGE (1995). « Modèles statistiques pour l'extraction de lexiques bilingues », *Traitement Automatique des Langues*, 36(1-2), p. 133–155.

GÉMAR, J.-C. (2002). « Le *plus* et le *moins-disant* culturel du texte juridique. Langue, culture et équivalence », *Meta*, (47)2, p. 163–176.

GREFENSTETTE, G. (2004). « Corpus Bilingues Alignés », Laboratoire Ingénierie de la Connaissance Multimédia Multilingue (LIC2M), [en ligne]. http://w3.u-grenoble3.fr/lebarbe/elc/supports/grefenstette.pdf (page consultée le 31 mars 2007).

GURRUTXAGA, A., X. SARALEGI et S. UGARTETXEA (2006). « ELeXBI, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora », dans *Actes du*

12e Congrès international de lexicographie (EURALEX,06), Torino, Italie, p. 159–165.

HULL, D. (2001). « Software tools to support the construction of bilingual terminology lexicons », dans BOURIGAULT, D., C. JACQUEMIN et M.-C. L'HOMME (dir.), *Recent advances in Computational Terminolgy*, Amsterdam, Philadelphia, John Benjamins Publishing Company, p. 167–182.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, France.

KAY, M. et M. RÖSCHEISEN (1988). « Text-Translation alignment », Technical Report, Xerox Palo Alto Research Center.

KAY, M. et M. RÖSCHEISEN (1993). « Text-Translation alignment », *Computational Linguistics*, 19(1), p. 121–142.

KRAIF, O. (1999). « Identification des cognats et alignement bi-textuel : une étude empirique », dans *Actes de la 6e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN'99)*, Cargèse, France, p. 205–214, [en ligne]. http://www.atala.org/doc/actes_taln/AC_0008.pdf (page consultée le 31 mars 2007).

KRAIF, O. (2000). « Extraction automatique de correspondances lexicales évaluation d'indices et d'algorithmes », dans *Actes de la 7e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN'00)*, Lausanne, Suisse, p. 225–236.

KRAIF, O. (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction*, Thèse de doctorat, sous la dir. de H. ZINGLÉ, Université de Nice Sophia Antipolis.

KRAIF, O. (2002). « Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné », dans Véronis J. (dir.), *Alignement lexical dans les corpus multilingues*, *Lexicometrica*.

KRAIF, O. (2007). *Page personnelle*, *Alinea, aide d'Alinea*, [en ligne]. http://w3.u-grenoble3.fr/kraif/index.php (page consultée le 19 avril 2007).

KRAIF, O. et B. CHEN (2004). « Combining clues for lexical level aligning using the Null hypothesis approach », dans *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Suisse, p. 1261–1264.

L'HOMME, M.-C. (2000). *Initiation à la traductique*, Brossard, Linguatech.

L'HOMME, M.-C. (2004). *La terminologie : principes et techniques,* Montréal, Les Presses de l'Université de Montréal.

L'HOMME, M.-C. (2005a). « Sur la notion de "terme" », *Meta*, 50(4), p. 1112–1132.

L'HOMME, M.-C. (2005b). « Glossaire des termes de traductique », TRA2000A, Montréal, Université de Montréal, .pdf.

LANGLAIS, P. (1997). « Alignement de corpus bilingues : intérêt, algorithmes et évaluation », dans *Bulletin de Linguistique Appliquée et Générale*, numéro Hors Série, Université de Franche-Comté, France, p. 245–254.

LEBART, L. et A. SALEM (1988). *Analyse statistique des données textuelles*, Dunod, Bordas.

LEMAY, C. (2003). *Identification automatique du vocabulaire caractéristique de l'informatique fondée sur la comparaison de corpus,* Mémoire de maîtrise, Département de linguistique et de traduction, Université de Montréal.

MARSHMAN, E. (2003). « Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie », [en ligne]. http://www.olst.umontreal.ca/pdf/terminotique/corpusentermino.pdf (page consultée le 23 juillet 2006).

MELAMED, D. (1995). « Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons », dans *Proceedings of the 3th Workshop on Very Large Corpora*, Boston, MA, p. 184–198.

MELAMED, D. (1997). « A word-to-word model of translation equivalence », dans *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, Espagne, p. 490–497.

MOORE, R. (2005). « Improving IBMWord-Alignment Model 1 », dans *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelone, Espagne, p. 519–526.

MORIN, E., S. DUFOUR-KOWALSKI et B. DAILLE (2004). « Extraction de terminologies bilingues à partir de corpus comparables », dans *Actes de la 11e Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN'04)*, Fès, Maroc, p. 309–318.

NAKAMURA-DELLOYE, Y. (2005). « Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais », dans *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, Dourdan, France, p. 285–294.

NEVADO, F., F. CASACUBERTA et E. VIDAL (2003). « Parallel corpora segmentation using anchor words », dans *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resources and tools for building MT*, Budapest, Hongrie, p. 33–40.

NÉVÉOL, A. et S. OZDOWSKA (2005). « Terminologie médicale bilingue anglais–français : usages clinique et législatif », *Glottopol,* 8.

OCH, F. J., C. TILLMANN et H. NEY (1999). « Improved alignment models for statistical machine translation », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP'99)*, University of Maryland, College Park, MD, p. 20–28.

OZDOWSKA, S. (2004). « Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés », dans *Actes de la 8e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'04)*, Fès.

OZDOWSKA, S. et D. BOURIGAULT (2004). « Détection de relations bilingues entre termes à partir d'une analyse syntaxique de corpus », dans *Actes du 14e Congrès*

Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA'04), Toulouse.

OZDOWSKA, S. et V. CLAVEAU (2005). « Alignement de mots par apprentissage artificiel de règles de propagation syntaxique en corpus », dans *Actes de la conférence Traitement automatique des langues naturelles, (TALN'05)*, Dourdan, France, p. 243–252.

PATRY, A. et P. LANGLAIS (2005). « Corpus-based terminology extraction », dans *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (TKE'05)*, Copenhague, Danemark, p. 313–321.

PEARSON, J. (2000). « Une tentative d'exploitation bi-directionnelle d'un corpus bilingue », dans *Cahiers de grammaire*, 25, « *Sémantique et Corpus* », p. 53–69.

RONDEAU, G. (1981). *Introduction à la terminologie*, Montréal, Centre éducatif et culturel.

SCHMID, H. (1994). « Probabilistic part-of-speech tagging using decision trees », dans *Proceedings of International Conference on New Methods on Language Proceeding*, Manchester, Royaume-Uni, p. 44–49.

SIMARD, M., FOSTER G. et P. ISABELLE (1992). « Using cognates to align sentences in bilingual corpora », dans *Proceedings of the 4th International Conference on Thoretical and Methodological Issues in Machine Translation (TMI'92)*, Montréal, Canada, p. 67–81.

SMADJA, F. (1993). « Retrieving collocations from text: Xtract », dans *Computational Linguistics*, 19(1), p. 143–177.

TERMIUM PLUS. *La base de données terminologiques et linguistiques du gouvernement du Canada*, [en ligne]. http://www.termiumplus.com/ (pages consultées le 2 septembre 2007).

TIEDEMANN, J. (1997). *Automatical Lexicon Extraction From Aligned Bilingual Corpora*, Diploma thesis, Otto-von-Guericke-University, Magdeburg, Department of Computer Science.

TIEDEMANN, J. (2003). *Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Thèse de doctorat, Université d'Uppsala, Uppsala, [en ligne]. http://publications.uu.se/theses/abstract.xsql?dbid=3791 (page consultée le 24 juillet 2007).

TRAN, T. D., A. BURGUN et N. GARCELON (2003). « Acquisition semi-automatique de terminologie bilingue en biologie moléculaire à partir des corpus comparables », dans *Terminologie et Intelligence Artificielle (TIA '03)*, Strasbourg, p. 166–175.

VAN CAMPENHOUDT, M. (1996a). « Réseau notionnel, intelligence artificielle et équivalence en terminologie multilingue : essai de modélisation », dans Clas A., P. Thoiron et H. Béjoint (dir.). *Lexicomatique et dictionnairiques, IV^es journées scientifiques du réseau thématique Lexicologie, terminologie, traduction, Université Lumière (Lyon II), 28-30 septembre 1995*, Montréal, AUPELF-UREF et Beyrouth, F.M.A., p. 281–306.

VAN CAMPENHOUDT, M. (1996b). *Abrégé de terminologie multilingue*, [en ligne]. http://www.termisti.refer.org/theoweb1.htm#table (page consultée le 15 octobre 2007).

VAN CAMPENHOUDT, M. (2001). « Pour une approche sémantique du terme et de ses équivalents », dans *International Journal of Lexicography*, 14(3), p. 181–209.

VAN DER EIJK, P. (1993). « Automating the acquisition of bilingual terminology », dans *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL '93)*, Utrecht, p. 113–119.

VÉRONIS, J. (2000). « Alignement de corpus multilingues », dans *Ingénierie des langues*, J.-M. PIERREL (dir.), Paris, Éditions Hermès, p. 151–172.

WU, D. (2000). « Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars », dans Véronis, J. (dir.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht, Kluwer Academic Publishers, p. 139–167.

Annexe A : Liste des textes du corpus et bibliographie

NOM DU FICHIER	Nbres MOTS	RÉFÉRENCE
chang_10ipccmodele.fr	22052	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (1997). « Introduction aux modèles climatiques simples employés dans le Deuxième Rapport », [en ligne]. http://www.ipcc.ch/pub/IPCCTP.II(F).pdf (page consultée le 6 janvier 2007).
chang_11ipccbilan2001.fr	43408	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2001). « Bilan 2001 des changements climatiques : Rapport de synthèse », [en ligne]. http://www.grida.no/climate/ipcc_tar/vol4/french/pdf/wg1sum.pdf (page consultée le 7 octobre 2006).
chang_12ipccsremissions.fr	5635	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2000). « Scénario d'émission », [en ligne]. http://www.grida.no/climate/ipcc/spmpdf/sres-f.pdf (page consultée le 7 octobre 2006).
chang_13ipccstabilisationdesgaz.fr	30264	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (1997). « Stabilisation des gaz atmosphériques à effet de serre : conséquences physiques biologiques et socio-économiques », [en ligne]. http://www.ipcc.ch/pub/IPCCTP.III(F).pdf (page consultée le 7 octobre 2006).
chang_14ipccsrl.fr	14829	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2000). « L'utilisation des terres, le changement d'affectation des terres et la foresterie », [en ligne]. http://www.grida.no/climate/ipcc/spmpdf/srl-f.pdf (page consultée le 07 octobre 2006).
chang_1canadaccd.fr	13133	ENVIRONNEMENT CANADA (2002). *Projections du climat du Canada.* CCD 00-01, [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/publications/reports_papers/index_f.html (page consultée le 15 janvier 2007).
chang_1europaenv.fr	6261	EUROPA (2005). « Un environnement de qualité : Le rôle de l'UE », [en ligne]. http://ec.europa.eu/publications/booklets/move/55/fr.doc (page consultée le 6 janvier 2007).
chang_1ipccaviation.fr	8569	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (1999). « L'aviation et l'atmosphère planétaire », [en ligne]. http://www.grida.no/climate/ipcc/spmpdf/av-f.pdf (page consultée le 7 octobre 2006).
chang_2canadaccd9801.fr	11381	ENVIRONNEMENT CANADA (2002). « Phénomènes météorologiques extrêmes et changement climatique », [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/understanding/ccd/ccd_9801/CCD_9801_f.pdf (page consultée le 20 janvier 2007).
chang_2europachange.fr	5666	EUROPA. *Change*, [on line]. http://ec.europa.eu/environment/climat/campaign/index_fr.htm (page consultée le 8 janvier 2007).

chang_2ipccbiodiversité.fr	40660	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2002). « Les changements climatiques et la biodiversité », [en ligne]. http://www.ipcc.ch/pub/tpbiodiv_f.pdf (page consultée le 7 octobre 2006).
chang_3canadaenviron.fr	24777	ENVIRONNEMENT CANADA. *La science du changement climatique*, [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/index_f.html (page consultée le 15 janvier 2007).
chang_3europa.fr	1052	EUROPA (2002). Opter pour un avenir plus vert, [en ligne]. http://ec.europa.eu/publications/booklets/move/32/txt_fr.pdf (page consultée le 6 janvier 2007).
chang_3ipccconsequence.fr	49665	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2001). « Bilan 2001 des changements climatiques : Conséquences, adaptation et vulnérabilité », [en ligne]. http://www.grida.no/climate/ipcc_tar/vol4/french/pdf/wg2sum.pdf (page consultée le 7 octobre 20.
chang_4canadafacc.fr	6212	ENVIRONNEMENT CANADA (1998). « Fonds d'action pour le changement climatique : fiches d'information », [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/publications/reports_papers/index_f.html (page consultée le 20 janvier 2007).
chang_4ipccsrocspmts.fr	37892	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2005). « Préservation de la couche d'ozone et du système climatique planétaire: Questions relatives aux hydrofluorocarbures et aux hydrocarbures perfluorés », [en ligne]. http://www.ipcc.ch/activity/spe (page consultée le 15 janvier 2007).
chang_5canadaicc.fr	29199	ENVIRONNEMENT CANADA (2005). *Une introduction au changement climatique – Une perspective canadienne*, [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/publications/reports_papers/index_f.html (page consultée le 15 janvier 2007)
chang_5ipccdyoxyde.fr	27378	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2005). « Piégeage et stockage du dioxyde de carbone », [en ligne]. http://www.ipcc.ch/activity/srccs/IPCC%20F.pdf (page consultée le 07 octobre 2006).
chang_6canadaqfp.fr	18123	ENVIRONNEMENT CANADA (2002). *FAQ - Foires aux questions - SCC*, [en ligne]. http://www.msc-smc.ec.gc.ca/education/scienceofclimatechange/publications/reports_papers/index_f.html (page consultée le 15 janvier 2007).
chang_6ipccemissions.fr	15592	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (1997). « Incidences des propositions de limitation des émissions de CO_2 », [en ligne]. http://www.ipcc.ch/pub/IPCCTP.IV(F).pdf (page consultée le 7 octobre 2006).
chang_7ipccregion.fr	17125	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (1997). « Evaluation de la vulnérabilité », [en ligne]. http://www.grida.no/climate/ipcc/spmpdf/region-f.pdf (page consultée le 7 octobre 2006).

chang_8ipccattenuation.fr	49923	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2001). « Bilan 2001 des changements climatiques : mesures d'atténuations », [en ligne]. http://www.grida.no/climate/ipcc_tar/vol4/french/pdf/wg3sum.pdf (page consultée le 7 octobre 2006).
chang_9ipccsrtt.fr	6286	GROUPE D'EXPERTS INTERGOUVERNEMENTAL SUR L'ÉVOLUTION DU CLIMAT (2000). « Questions méthodologiques et technologiques dans le transfert de technologie », [en ligne]. http://www.grida.no/climate/ipcc/spmpdf/srtt-f.pdf (page consultée le 07 octobre 2006).
chang_aworldofscience.fr	8266	UNESCO. Planète science, Bulletin trimestriel d'information sur les sciences exactes et naturelles, vol. 4, no 1 mars 2006, [en ligne]. http://unesdoc.unesco.org/images/0014/001432/143225f.pdf (page consultée le 7 octobre 2006).
chang_climatechangeyouth.fr	4761	EUROPA (2006). « Le changement climatique : qu'est-ce que c'est? » [en ligne]. http://ec.europa.eu/environment/climat/campaign/pdf/climate_change_youth_fr.pdf (page consultée le 8 janvier 2007).
chang_ency.fr	28779	MÜLLER, S. et J. BUCHDAHL (2000). *L'Encyclopédie de l'Environnement Atmosphérique*, [en ligne]. http://www.ace.mmu.ac.uk/eae/french/french.html (page consultée le 7 janvier 2007).
chang_greenfacts.fr	23080	GREENFACTS. *Faits sur la Santé et l'Environnement : changement climatique et le réchauffement de la planète*, [en ligne]. http://www.greenfacts.org/fr/dossiers/changement-climatique/index.htm (page consultée le 11 janvier 2007).
chang_jaclet.fr	40278	JACQUES, G. et H. LE TREUT (2004). *Le changement climatique*, Paris, Éditions Unesco.
chang_newcourier2005.fr	3844	UNESCO. *Le nouveau courrier*, Paris, Organisation des Nations unies pour l'éducation, la science et la culture, décembre 2005, [en ligne]. http://unesdoc.unesco.org/images/0014/001420/142021f.pdf#142037 (page consultée le 6 janvier 2007).
chang_rapportpentagone.fr	9259	SCHWARTZ, P. et D. RANDALL (2003). « An Abrupt Climate Change Scenario and Its Implications for United States National », [on line]. http://www.gbn.com/GBNDocumentDisplayServlet.srv?aid=26231&url=%2FUploadDocumentDisplayServlet.srv%3Fid%3D28566 (page consultée le 15 janvier 2007).
chang_vulnerability.fr	1720	AGENCE EUROPÉENNE POUR L'ENVIRONNEMENT (2005). « Changements climatiques et inondations liées aux rivières et fleuves en Europe », [en ligne]. http://reports.eea.europa.eu/index_table?lang=French (page consultée le 11 janvier 2007).

Annexe B : Fiches d'analyse des CT

CLIMATIQUE			
climate	*changement* **climatique**	178	**climate** *change*
	système **climatique**	51	**climate** *system*
	modèle **climatique**	8	**climate** *model*
	(condition, politique, variable, processus, etc.) **climatique**	63	**climate** *(condition, policy, variable, process, etc.)*
	Sous total	300	
climatic	*condition* **climatique**	3	**climatic** *condition*
	effet **climatique**	2	**climatic** *effect*
	changement **climatique**	2	**climatic** *change*
	(impact, contrainte, paramètre) **climatique**	3	**climatic** *(impact, constraint, factor)*
		10	
autre	*modèles* **climatiques** *simples*	1	*SCMs* (simple **climate** models)
	aspects scientifiques du changement **climatique** – *Shine,* et al	1	*IPCC Scientific Assessment — Shine, et al.* (Intergovernmental Panel on **Climate** Change)
	(condition, origine, etc.) **climatique**	3	*weather (condition, related, etc.)*
	climatique	8	**Ø** *(par exemple,* **réchauffement climatique** *est rendu par* **global warming***)*
	Sous total	13	
	total	**323**	

CHANGEMENT			
change	**changement** *climatique*	157	*climate* **change**
	changement (employé sans collocation particulière)	107	**change**
	changement *climatique*	5	**change** *in climate*
	changement *(planétaire, de température, d'affectation des terres, etc.)*	24	*(global, temperature, land-use, etc.)* **change**
	Sous total	293	
autre	**changement**	1	**changing** *(adj.)*
	dans les **changements**	1	*in* **changing**
	changement *climatique*	1	*by* **changing** *climate*
	Convention - cadre des Nations Unies sur les **changements** *climatiques (CCNUCC)*	1	*UNFCCC (United Nations Framework Convention on Climate Change)*
	changements *d'affectation des terres et de la foresterie*	6	*LULUCF (Land Use, Land-Use* **Change,** *and Forestry)*
	changement *de température*	1	**changed** *temperature (adj.)*
	changement *climatique*	1	*global warming*
		1	*climate system*
	changement	8	**Ø**
		1	*which*
		1	*Altered*
	Sous total	23	
	Total	**316**	

ÉMISSION			
emission	**émission** (employé sans collocation particulière)	207	**emission**
	émission *de gaz à effet de serre*	32	*greenhouse gas* **emission**
		6	*GHG* **emission**
	émission *des gaz à effet de serre*	5	*greenhouse gas* **emission**
		3	**emission** *of greenhouse gases*
	émission *de gaz à effet de serre*	5	**emission** *of greenhouse gases*
	Sous total	**258**	
autre	**émission** *de CH4*	1	**generation** *of CH4*
	avec **émissions** *de 1,6 ±1*	1	*and* **emitting** *1.6 ±1*
	émission	3	**release**
		1	*those*
		4	**Ø**
	Sous total	10	
	total	**268**	

TEMPÉRATURE			
temperature	**température**	220	**temperature**
	Sous total	220	
autre	**température**	1	*above 1990 levels*
		1	*those*
	L'élévation des **températures**	1	*this*
	températures *minimales nocturnes*	1	**Ø** *nighttime lows*
	températures *maximales diurnes*	1	**Ø** *daytime highs*
	Sous total	5	
	Total	**225**	

CARBONE			
carbon	*dioxyde de* **carbone**	72	**carbon** *dioxide*
	cycle du **carbone**	13	**carbon** *cycle*
	carbone (employé sans collocation particulière)	131	**carbon**
	Sous total	216	
co_2	*dioxyde de* **carbone**	5	**co_2**
	Sous total	5	
autre	**carbone**	3	**Ø**
	industries à forte intensité de **carbone**	1	*energy-intensive industries*
	Sous total	4	
	Total	**225**	

CLIMAT			
climate	**climat** (employé sans collocation particulière)	202	**climate**
	évolution du **climat**	10	**climate** *change*
	Groupe d'experts intergouvernemental sur l'évolution du **climat**	32	*Intergovernmental Panel on* **Climate** *Change*
	modèle du **climat**	5	**climate** *model*
	climat *mondial*	7	*Global* **climate**
		1	*Earth's* **climate**
	pressions liées au **climat**	1	**climate**-*related stresses*
	Sous total	258	
autre	*déplacements liés au* **climat**	1	**climatically** *associated shifts*
	climat	2	**Ø**
	Sous total	3	
	Total	**261**	

SERRE			
greenhouse	*gaz à effet de* **serre**	229	**greenhouse** *gas*
	serre	2	**greenhouse**
	effet de **serre**	22	**greenhouse** *effect*
		1	**greenhouse** *gas*
		1	**greenhouse** *gas effect*
	effet naturel de la **serre** *de la Terre*	2	*Earth's natural* **greenhouse** *effect*
	réchauffement par effet de **serre**	5	**greenhouse** *warming*
	Sous total	262	
GHG	*gaz à effet de* **serre**	16	**GHG** (greenhouse gas)
	Sous total	16	
autre	**serre**	1	**Ø**
	Sous total	1	
	Total	**279**	

GAZ			
gas	**gaz** *à effet de serre*	199	*greenhouse* **gazes**
	gaz *naturel*	6	*natural* **gaz**
	gaz *(naturel, organique, fluoré, etc.)*	8	**gas** *(reservoir, organic, fluorinated, etc.)*
	gaz (employé sans collocation particulière)	38	**gas**
	gaz *à effet de serre*	2	*greenhouse effect* **gases**
	Sous total	253	
GHG	**gaz** *à effet de serre*	15	*GHG (greenhouse gases)*
	Sous total	15	
autre	**gaz** *carbonique*	2	*carbon dioxide*
		1	*carbon credits*
		1	*carbon exchange*
	gaz *à effet de serre*	1	*greenhouse concentrations*
	gaz	1	CO_2
		1	*both species*
	Sous total	7	
	Total	**275**	

RÉCHAUFFEMENT			
warming	**réchauffement** (employé sans collocation particulière)	131	**warming**
	réchauffement *(climatique, planétaire, global, mondial, etc.)*	58	*(global, climate)* **warming**
	Sous total	189	
warmer	**réchauffement** *de la planète*	1	**warmer** *temperatures*
	réchauffement *de notre planète*	1	*world* **warmer**
	réchauffement *climatique*	1	**warmer** *climate*
	réchauffement *des températures*	2	**warmer** *temperatures*
	réchauffement	1	**warmer** *climate*
	réchauffement *du climat*	2	**warmer** *climate*
		8	
warm (to)	**réchauffement** *du climat*	1	**warm** *the atmosphere*
	réchauffement	1	**warm** *faster*
	réchauffement *du climat*	1	*climate* **warms**
	réchauffement *climatique*	1	*climate* **warms**
	Sous total	4	
GWP	**réchauffement** *global*	4	*GWP (Global Warming Potential)*
	réchauffement	2	*GWP*
	Sous total	6	
autre	**réchauffement** *(radiatif, solaire, Ø)*	3	*(radiative, solar, Ø)* **heating**
	réchauffement *mondial*	1	**Ø**
	réchauffement *de l'eau*	1	**increasing** *water* **temperature**
	réchauffement	2	**temperature increase**
		1	**temperature rise**
		1	**increase in temperature**
	Sous total	9	
	Total	**216**	

FORÇAGE			
Forcing	**forçage** *radiatif*	48	*radiative* **forcing**
	forçage (employé sans collocation particulière)	43	**forcing**
	forçage *(négatif, positif, climatique, solaire, etc.)*	20	*(negative, positive, climate, solar, etc.)* **forcing**
	Sous Total	111	
autre	**forçage**	1	**Ø**
		1	**which**
		1	**that**
	total	114	

CO_2			
co_2	**co_2**	211	**co_2**
	Sous total	211	
autre	**co_2**	4	**Ø**
		1	**it**
	Sous total	5	
	Total	216	

ATMOSPHÈRE			
atmosphere	**atmosphère**	152	*the* **atmosphere**
	(Ø, haute, basse, totalité de, etc.) **atmosphère**	48	*(Ø, upper, lower, Earth's, entire, global, etc.)* **atmosphere**
	Sous total	200	
atmospheric	*(mouvement ascensionnel de l', CO_2 dans l', composition de l', etc.)* **atmosphère**	44	**atmospheric** *(uplift, CO_2, composition, etc.)*
	*l'Administration nationale de l'océan et de l'***atmosphère**	1	*National Oceanic and* **Atmospheric** *Administration*
	*Fondation canadienne pour les sciences du climat et de l'***atmosphère**	1	*Foundation for Climate and* **Atmospheric** *Sciences*
	Sous total	46	
autre	**atmosphère**	7	**Ø**
	Sous total	7	
	Total	253	

CONCENTRATION			
concentration	**concentration** *(de co_2, atmosphérique, de gaz à effet de serre, d'ozone, etc.)*	122	*(co_2, atmospheric, greenhouse gas, ozone, etc.)* **concentration**
	concentration *(de co_2, des gaz à effet de serre)*	49	**concentration** *(of co_2, of greenhouse gases)*
	concentrations *(préindustrielle, atmosphériques, etc.) de (gaz à effet de serre, co_2, méthane, etc.)*	18	*(pre-industrial, atmospheric, etc.)* **concentrations** *of (greenhouse gases, co_2, methane, etc.)*
	concentration	13	**concentration**
	Sous total	202	
autre	**concentration**	2	**build-up**
	concentration *de chlore*	1	*chlorine* **level**
	concentration *de co_2*	1	*co_2* **level**
	concentration	19	**Ø**
	concentrations *de capitaux*	1	*capital* **pool**
	concentration *démographique*	1	*population* **pressure**
	concentration	1	*warm* ***pool***
	Sous total	26	
	Total	228	

EFFET			
effect	**effet** *(radiatif, de rétroaction, etc.)*	41	*(radiative, feedback, etc.)+* **effect**
	effet	49	**effect**
	(cet, ces) **effet**	5	*(this, these)* **effect**
	effet *de serre*	1	*greenhouse gas* **effect**
		9	*greenhouse* **effect**
	gaz à **effet** *de serre*	2	*greenhouse* **effect** *gases*
	Sous total	107	
impact	**effet**	10	**impact**
	Sous total	10	
affect (to)	*à comprendre les* **effets** *du réchauffement*	1	*understand how warming will* **affect**
	ont déjà subi des **effets** *des changements climatiques*	1	*have already been* **affected** *by changes in climate*
	aident à comprendre les **effets** *du réchauffement sur la société canadienne et sur l'environnement*	1	*help us understand how warming will* **affect** *Canadian society and the environment*
	ont des **effets** *sur la productivité*	1	**affect** *the productivity*
	Sous total	4	
autre	*gaz à* **effet** *de serre*	121	**Ø** *(greenhouse gas)*
	gaz à **effet** *de serre*	16	**Ø** *(GHG)*
	réchauffement par **effet** *de serre*	5	**Ø** *(greenhouse warming)*
	effet *de serre*	3	*greenhouse gas*
	en **effet**	21	*as the, it is, indeed, in fact, may, that is because, in some respect, in effect, Ø, etc.*
	effets *des changements climatiques d'origine anthropique*	1	**influence** *of anthropogenic climate change*
	effet	1	**consequence**
	effet *initial*	1	*initial* **change**
	effet	2	**response**
	ce qui a pour **effet**	1	*in so doing*
	effets *positifs dans le domaine environnemental*	1	*environmental* **benefits**
	un **effet** *additionnel*	1	*additionality*
	effets *bénéfiques*	1	*benefits*
	*due à l'***effet** *combiné*	1	*due to a combination*
	aurait un **effet** *positif*	1	**enhance**
	effet	1	**sign**
	effet	2	**loop**
	correspondant aux **effets** prévus	1	*in the manner expected*
	effet	10	**Ø**
	effets *cumulés*	1	**synergy**
	effet	1	**feedback**
	effet négatif	1	*negative* **factor**
	*sous l'***effet** *du soleil*	1	*from solar* **heating**
	*sous l'***effet** *de l'accroissement démographique*	1	*as the world's population increases*
	Sous total	196	
	Total	317	

AÉROSOL			
aerosol	**aérosol** *sulfaté*	18	*sulphate* **aerosol**
		4	*sulfate* **aerosol**
	GES + **aérosol**	2	*GHG* + **aerosol**
	aérosol	104	**aerosol**
	aérosol *(troposphérique, stratosphérique, volcanique, anthropogénique, etc.)*	24	*(tropospheric, stratosphériques, volcanic, anthropogenic, etc.)* **aerosol**
	forçage des **aérosols**	2	**aerosol** forcing
	forçage imputable aux **aérosols**	1	**aerosol** forcing
	forçage dû aux **aérosols**	2	**aerosol** forcing
	Sous total	157	
autre	*GES* + **aérosol**	6	*GHG* + **A**
	pour les **aérosols** *carbonés organiques*	1	*for fossil fuel organic carbon*
	propulseurs d'**aérosols**	1	*spray can propellants*
	Sous total	**8**	
	Total	**165**	

ÉCOSYSTÈME			
ecosystem	**écosystème**	185	**ecosystem**
	les **écosystèmes** terrestres, marins et autres **écosystèmes** aquatiques	2	terrestrial , marine , and other aquatic **ecosystems**
	Sous total	187	
autre	*incidences plus profondes sur les* **écosystèmes** *et les sociétés humaines*	1	*impacts on* **natural** *and human* **systems**
	écosystème	1	**natural systems**
	*capacité de l'***écosystème** *de la Terre*	1	*capacity of the Earth's* **environment**
	dépassent les seuils de tolérance des humains et des **écosystèmes**	1	*will exceed human and* **ecological** *tolerance thresholds*
	écosystème	1	*while* **one**
	l'absorption nette de carbone par les **écosystèmes** *terrestres*	1	**Ø** *(the net terrestrial carbon uptake)*
	Sous total	6	
	Total	**193**	

ATTÉNUATION			
mitigation	**atténuation**	48	**mitigation**
	atténuation *des changements*	12	*climate change* **mitigation**
		1	**mitigation** *of climate change*
	atténuation *(des émissions de gaz à effet de serre, des gaz à effet de serre)*	4	*greenhouse gas* **mitigation**
	atténuation *des effets des émissions de gaz à effet de serre*	1	**mitigation** *of greenhouse gas emissions*
	atténuation	1	*greenhouse gas* **mitigation**
	atténuation *des émissions*	1	*emissions* **mitigation**
	*technologies d'***atténuation**	3	**mitigation** *technologies*
	*scénario d'***atténuation**	6	**mitigation** *scenario*
	Sous total	77	
mitigate (to)	**atténuation**	4	**mitigate**
	atténuation	4	*(for, in)* **mitigating** *climate change*
	Sous total	8	
moderate (to)	**atténuation**	4	**moderating**
	Sous total	4	
autre	*des mesures d'***atténuation**	1	*of the* **mitigating** *measures*
	atténuation *du changement climatique*	1	**attenuate** *climate change*
	*futurs coûts d'***atténuation**	1	*future* **abatement** *costs*
	*comme l'***atténuation** *ou l'arrêt de la circulation thermohaline*	1	*such as* **slowing** *or shutdown of thermohaline circulation*
	*les mesures d'***atténuation** *incluent*	1	**these activities** *include*
	mais il y a pu avoir **atténuation** *de l'appauvrissement de la diversité*	1	*but diversity* **losses** *were ameliorated*
	*Les mesures d'***atténuation** *des gaz à effet de serre doivent être replacées dans le contexte des nombreux biens et services fournis par les écosystèmes.*	1	*The production of greenhouse gas* **offsets** *should be placed in the context of the many goods and services that ecosystems produce.*
	Sous total	7	
	Total	96	

SCÉNARIO			
scenario	**scénario**	108	**scenario**
	scénario *d'émission*	20	*emission* **scenario**
	scénario *IS92*	23	*IS92* **scenario**
	scénario *du SRES*	9	*SRES* **scenario**
	scénario *(A2, A1, B1, B2)*	5	**scenario** *(A2, A1, B1, B2)*
	scénario *(du GIEC, de définitions, de référence, etc.)*	43	*(IPCC, definitional, reference, etc.)* **scenario**
	Sous total	210	
autre	**scénarios** *du changement climatique à venir*	2	*future climate change* **projections**
	scénario *(IS92a, IS92c, IS92e)*	13	**Ø** *(IS92a, IS92c, IS92e, etc.)*
	scénario	1	**model**
		1	**storyline**
	Sous total	17	
	Total	227	

INCIDENCE			
impact	**incidence**	18	**impact**
	incidence *(sur l'écosystème, de divers scénarios d'émissions, etc.)*	28	**impact** ***(on ecosystem, of*** *diverse emissions scenarios, etc.)*
	incidence *(climatique, environnemental, majeure, sur l'environnement, etc.)*	27	*(climatic, environmental, major, etc.)* **impact**
	(évaluation des, évaluation, etc.) **incidences**	5	*(assessment of, global mean, etc.)* **impact**
	(forte, peu d', réduire l', etc.) **incidence**	3	*(significant, little, lessen the, etc.)* **impact**
	Sous total	81	
implication	**incidence**	14	**implication**
	Sous total	14	
effect	**incidence**	3	**effect**
	incidence *(du climat, sur l'élévation du niveau de la mer)*	4	**effect** *(of the climate,* on *sea level rise)*
	*calculer l'***incidence**	1	*calculate the* **effect**
	incidence *(nette, sensible, directe, etc.)*	5	*(net, significant, direct, etc.)* **effect**
	Sous total	13	
affect (to)	**incidence**	11	**affect** (to)
	Sous total	11	
incidence	**incidence** *du régime des perturbations*	1	**incidence** *of disturbance regimes*
	incidence *de l'assèchement saisonnier*	1	**incidence** *of seasonal flow*
	incidence *accrue*	2	*increased* **incidence**
	*augmentant l'***incidence**	1	*increasing the* **incidence**
	incidence *des parasites d'arbres*	1	**incidence** *of tree pests*
	Sous total	6	
autre	*avoir des* **incidences** *sur la dynamique*	1	**to interact** *with the dynamics*
	incidence	6	**Ø**
		1	**futur change**
		1	**affected**
		1	**influence**
		1	**consequence**
		1	**be important**
	Sous total	12	
	Total	137	

OCÉAN			
ocean	**océan**	203	**ocean**
	Sous total	203	
oceanic	*température sous la surface des* **océans**	1	*sub-surface* **oceanic** *temperature*
	la rétroaction des **océans**	2	**oceanic** *feedback*
	océans	1	**oceanic** *uptake*
	stockage du carbone dans les **océans**	2	**oceanic** *carbon storage*
	processus potentiellement essentiel relatif aux **océans**	3	*potentially critical* **oceanic** *process*
	Des relevés à plus long terme sur des carottes de glace et dans les **océans**	1	*Longer ice core and* **oceanic** *records*
	*pompes biologique et physico-chimique de l'***océan**	1	**oceanic** *biological and physico-chemical pumps*
	*Administration nationale de l'***océan** *et de l'atmosphère*	1	*National* **oceanic** *and Atmospheric Administration (NOAA)*
	Sous total	12	
sea	**océan**	6	**sea**
	Sous total	6	
autre	*on prévoit un accroissement de la productivité écologique des* **océans**	1	**marine** *ecological productivity should rise*
	cet **océan**	1	**Pacific**
	Sous total	2	
	Total	223	

MODÈLE			
model	**modèle**	235	**model**
	Sous total	235	
modelling	*Les* **modèles** *montrent également que les températures devraient continuer à augmenter*	1	**Modelling** *also shows that temperatures should continue to rise*
	modèle	2	**modelling**
	modèle *intégré*	1	*integrated* **modelling**
	simulations par **modèles**	1	**modelling** *studies*
	Sous total	5	
pattern	**modèle**	6	**pattern**
	modèle *de climat*	2	*weather* **pattern**
	Sous total	8	
autre	*selon les* **modèles**	1	*are estimated by* **modelling**
	modèle	2	**they**
		1	**these**
		1	**performer**
		4	**Ø**
	Sous total	9	
	Total	257	

ÉLÉVATION			
rise	**élévation** *du niveau de la mer*	68	*sea level* **rise**
	élévation *du niveau marin*	7	*sea level* **rise**
	élévation	15	**rise**
	élévation *du niveau de la mer*	7	**rise** *in sea level*
		1	*sea levels to* **rise**
	Sous total	98	
rising	**élévation** *du niveau de la mer*	5	**rising** *sea level*
	élévation *du niveau des mers*	5	**rising** *sea level*
	élévation *du niveau marin*	1	**rising** *sea level*
	élévation *du niveau de l'océan*	1	**rising** *ocean levels*
	élévation *de la température de l'eau*		**rising** *water temperatures*
	élévation	3	**rising**
	Sous total	15	
rise (to)	**élévation**	7	**rises** **rise** *(to)* **rising**
	Sous total	7	
increase	**élévation**	31	**increase**
	Sous total	31	
higher	**élévation** *(de la température, de la vitesse des vents, etc.)*	4	**higher** *(temperature, wind speeds, etc.)*
	Sous total	4	
autre	**élévation** *de la température*	3	**increasing** *temperature*
	élévation	11	**Ø**
	deltas de faible **élévation**	3	*low-lying deltas*
	élévation *des températures*	1	*warmer climates*
	*risques d'***élévation** *des coûts*	1	*risks of* **high** *costs*
	*entraîner l'***élévation** *du coût total*	1	**raise** *aggregate costs*
	élévation	1	**growth**
	élévation *du niveau de la mer*	1	**elevation** *in sea level*
	élévation *des températures*	1	**increased** *temperature*
	élévation	3	**change**
		1	**build-up**
	Sous total	27	
	Total	182	

RADIATIF			
radiative	*forçage* **radiatif**	73	**radiative** *forcing*
	(processus, effet, etc.) **radiatif**	12	**radiative** *(process, effect, etc.)*
	Sous Total	85	
radiation	*(bilan, équilibre, etc.)* **radiatif**	15	**radiation** *(budget, balance of, etc.)*
	Sous Total	15	
autre	**radiatif**	7	**Ø** *(exemple:* **forçage radiatif** *par* **forcing** *seulement)*
	Sous Total	7	
	Total	107	

ATMOSPHÉRIQUE			
atmospheric	*(gaz, méthane, concentration, CO_2, aérosol, composition, etc.)* **atmosphérique**	158	*(gas, CH_4, concentration, CO_2, aerosol, composition, etc.)* **atmospheric**
	Sous total	158	
atmosphère	**atmosphérique**	11	*(in the, of the)* **atmosphere**
	Sous total	11	
air	*pollution* **atmosphérique**	13	**air** *pollutant* ou *pollution*
	Sous total	13	
weather	*(condition, extrême)* **atmosphérique**	4	**weather** *(condition, extreme)*
	Sous total	4	
autre	**atmosphérique**	2	**Ø**
	pollution **atmosphérique**	1	**gas** *pollution*
	conditions **atmosphériques**	1	**climate**
	Sous total	4	
	Total	**190**	

VARIATION			
change	**variation** *(de concentration, de la température, etc.)*	94	**change** *(in concentration, in temperature, etc.)*
	Sous total	94	
variation	**variation**	77	**variation**
	Sous total	77	
shift	**variation**	4	**shift**
	Sous total	4	
autre	**variation**	1	**impact**
		1	**influence**
		9	**Ø**
		2	**fluctuation**
		1	**vary**
		1	**glide**
		1	**rise**
		1	**trend**
		1	**changing**
		1	**variable**
		1	**variance**
		1	**varied**
		1	**changing**
		1	**varying**
		1	**deviation**
		1	**different**
		1	**difference**
		1	**variability**
	Sous total	27	
	Total	**202**	

PRÉCIPITATION			
precipitation	**précipitation**	141	**precipitation**
	Sous total	141	
rainfall	**précipitation**	31	**rainfall**
	Sous total	31	
autre	**précipitations** *de pluie ou de neige*	1	*rain or snowfall*
	précipitation	1	**Ø**
	Sous total	2	
	Total	174	

VARIABILITÉ			
variability	**variabilité**	162	**variability**
	Sous total	162	
autre	**variabilité** *naturelle*	1	**vary** naturally
	grande **variabilité**	1	*high* **variance**
	variabilité *temporelle*	1	*temporal* **variation**
	variabilité	2	**Ø**
	Sous total	5	
	Total	167	

DIOXYDE			
dioxide	**dioxyde** *de carbone*	125	*carbon* **dioxide**
	dioxyde *de soufre*	7	*sulphur* **dioxide**
		1	*sulfur* **dioxide**
	Sous total	133	
co_2	**dioxyde** *de carbone*	25	**co_2**
	Sous total	25	
autre	**dioxyde**	4	**Ø**
		2	**carbon**
		1	*ccs (carbon dioxide capture and storage)*
	Sous total	7	
	Total	165	

SURFACE			
surface	*(réflexion en, température de, etc.)* **surface**	185	**surface** *(reflectivity, temperature, etc.)*
	Sous total	185	
area	**surface**	6	**area**
	surface *boisée*	1	*forest* **area**
	surface *brûlée*	1	**area** *burnt*
	par unité de **surface**	1	*per unit* **area**
	surface	1	*land* **area**
	Sous total	10	
autre	**surface**	7	**Ø**
	matériaux de **surface**	1	**facing** *material*
	fait **surface**	1	*have* emerged
	la **surface** *de glace*	2	*sea-ice* **extent**
	surface	1	**ground**
	température de la mer en **surface**	3	*SST (sea surface temperature)*
	Sous total	15	
	Total	210	

OZONE			
ozone	*(appauvrissement de la couche d', raréfaction, couche d', etc.)* **ozone**	132	**ozone** *(depletion, layer, etc.)*
	ozone *(troposphérique, stratosphérique, etc.)*		*(tropospheric, stratospheric, etc.)* **ozone**
	Sous total	132	
o_3	**ozone**	8	
	Sous total	8	
autre	**ozone**	2	**O***DS (ozone depletion substance)*
	Sous total	2	
	Total	142	

ÉCHELLE			
scale	**échelle** *(mondiale, régionale, etc.)*	95	*(global, regional, etc.)* **scale**
	petite **échelle**	2	*small*-**scale**
	grande **échelle**	14	*large*-**scale**
	échelle *(temporelle, de temps, séculaire, régionale, planétaire, etc.)*	15	*(time-, regional-, global, etc.)* **scale**
	Sous total	126	
level	**échelle** *(du pays, nationale, régionale, mondial, etc.)*	16	*(country, national, regional, global, etc.)* **level**
	Sous total	16	
wide	*à l'***échelle** *(mondiale, de l'UE, du secteur)*	4	*(world-, Eu-, sector-)* **wide**
	Sous total	4	
autre	*à l'***échelle** *(du globe, mondial, de la planète, planétaire)*	19	**globally** **internationally** **nationally** **locally** **worldwide**
	*à l'***échelle** *(du globe, mondial, planétaire, régionale, nationale, internationale)*	44	**global** **regional** **national** **international** **local**
	échelle *de temps*	3	**timescale**
	échelle	8	**Ø**
		13	**extensive, large, metrics, massive, great extent, shorter term, within countries, widely, is scaled, scaling method, downscaling**
	transfert rapide et à grande **échelle**	1	*rapid and widespread transfer*
	Sous total	25	
	Total	234	

STABILISATION			
stabilization	**stabilisation**	71	**stabilization**
	Sous total	71	
stabilisation	**stabilisation**	13	**stabilisation**
	Sous total	13	
stabilize (to)	**stabilisation**	4	**stabilized, stabilizing**
	Sous total	4	
stabilise (to)	**stabilisation**	5	**stabilised**
	Sous total	5	
autre	**stabilisation**	3	**Ø**
		1	STAB
	Sous total	4	
	Total	97	

EAU			
water	**eau**	178	**water**
	(thermopompes à, condenseur à) **eau**	4	**water-** *(heating, cooled)*
	eau *(de pluie, profonde, potable, souterraine)*	8	*(rain, deep, drinking, ground)* **-water**
	Sous total	190	
ocean	**eau** *océanique*	3	**ocean**
	eau	4	**ocean**
	Sous total	7	
sea	**eau**	5	**sea**
	montée des **eaux**	1	**sea-***level rise*
	Sous total	6	
seawater	**eau** *de mer*	4	**seawater**
	Sous total	4	
groundwater	**eau** *souterraine,*	7	**groundwater**
	Sous total	7	
freshwater	**eau** *douce*	8	**freshwater**
	Sous total	8	
streamflow	*débit d'***eau**	5	**streamflow**
	Sous total	5	
autre	*débit des cours d'***eau**	1	*river flow*
	eaux *profondes*	1	*regions at depth*
	basses **eaux**	2	*low flow*
	*cours d'***eau**	2	**river** *stream*, **stream**
	*remontée d'***eau**	2	**upwelling**
	eau *de fonte*	1	**meltwater**
	eau *usée*	1	**wastewater**
	*collecteurs d'***eaux** *pluviales*	1	*storm drains*
	eau	8	**Ø**
		1	its
		1	**sewage**
		2	**moisture**
		3	**irrigation, drainage**
	Sous total	26	
	Total	253	

ANTHROPIQUE			
anthropogenic	**anthropique**	75	**anthropogenic**
	Sous total	75	
anthropogenically induced	**anthropique**	4	**anthropogenically induced**
	Sous total	4	
human	*émissions* **anthropiques**	15	**human** *emissions*
	(impact, échelle du temps, source, etc.) **anthropique**	13	**human** *(impact, timescale, source, etc.)*
	Sous total	28	
human activities	**anthropique**	7	**human activities**
	Sous total	7	
human-induced	**anthropique**	10	**human-induced**
	Sous total	10	
autre	**anthropique**	3	**Ø**
		1	**man-made**
		1	**manmade**
	anthropique	2	**human-made**
	anthropique	1	**human-related**
	effet de serre **anthropique**	2	*enhanced greenhouse gas effect*
	facteurs **anthropiques**	2	**human** *influences*
	Sous total	12	
	Total	**136**	

AUGMENTATION			
increase	**augmentation**	154	**increase**
	Sous total	154	
increase (to)	**augmentation**	15	*(will, to, etc.)* **increase,** *(has, have, being, etc.)* **increased**
	Sous total	15	
increased	**augmentation** *(des rendements des cultures, des émissions, etc.)*	21	**increased** *(crop yields, emissions, etc.)*
	Sous total	21	
increasing	**augmentation** *(de la concentration,*	8	**increasing** *(concentration*
	Sous total	8	
rise	**augmentation** *de la température*	2	**rise** *in temperature*
	augmentation *(du niveau de la mer, net, etc.)*	5	*(sea level, net, etc.)* **rise**
	Sous total	7	
rising	**augmentation** *(des émissions, des températures, etc.)*	7	**rising** *(emissions, baselines, etc.)*
	Sous total	7	
growth	**augmentation**	7	**growth**
	Sous total	7	
higher	**augmentation**	5	**higher**
	Sous total	5	
enhancement	**augmentation**	5	**enhancement**
	Sous total	5	
enhanced	**augmentation**	4	**enhanced**
	Sous total	4	
autre	**augmentation** *de la masse*	1	**gain** *mass*
		1	**greater**
		6	**Ø**
		1	*are* **rising**
		13	**adding, increment, elevation, more, capabilities, growing, enhancing**
	Sous total	22	
	Total	255	

OCÉANIQUE			
ocean	*modèle* **océanique**	12	**ocean** *model*
	(courrant, circulation, etc.) **océanique**	94	**ocean** *(current, circulation, etc.)*
	océanique	16	**ocean**
	Sous total	122	
oceanic	*(GCM, composante, etc.)* **océanique**	27	**oceanic** *(MCG, component, etc.)*
	Sous total	27	
autre	**océanique**	1	**Ø**
	fonds **océaniques**	3	**seafloor**
	système **océanique**	1	**sea level** *system*
	Sous total	5	
	Total	154	

COÛT			
cost	**coût**	143	**cost**
	Sous total	143	
autre	**coût**	2	**Ø**
		1	**costly**
		1	**price**
		1	**subsidized**
		2	**expense**
		1	**cheaper**
		1	**cheapest**
		1	**pay**
		1	**low price**
		1	**cost (to)**
	Sous total	12	
	Total	155	

FOSSILE			
fossil	*combustible* **fossile**	114	**fossil** *fuel*
	(carburant, matière) **fossile**	9	**fossil** *fuel*
	combustible **fossile**	9	**fossil-***fuel*
	(émissions, ressources) **fossiles**	2	**fossil** *fuel (emissions, resources)*
	co_2 d'origine **fossile**	7	**fossil** *co_2*
	(origine, pollen, dune, émission, etc.) **fossile**	16	**fossil** *(origin, pollen, dune, emission, etc.)*
	combustible **fossile**	5	**fossil** *(resources, intensive,*
	Total	162	

ADAPTATION			
adaptation	*(Ø, mesure d', politiques d', options d', etc.)* **adaptation**	91	**adaptation** *(Ø, measures, policies, options, etc.)*
	Sous total	91	
adaptive	**adaptation**	10	**adaptive**
	Sous total	10	
adapt (to)	**adaptation**	19	**adapt**
	Sous total	19	
adjustment	**adaptation**	4	**adjustment**
	Sous total	4	
autre	**adaptation**	3	**retrofitting**
		1	**retrofit**
		2	**Ø**
		2	**maladaptation**
		1	**mismatch**
		1	**adoption**
	Sous total	10	
	Total	134	

LATITUDE			
latitude	**latitude**	143	**latitude**
	Sous total	143	
midlatitude	**latitude** *(moyenne, intermédiaire)*	5	**midlatitude**
	Sous total	5	
autre	**latitude**	2	**Ø**
	latitude	2	**latitudinal**
	imposent des limites à la **latitude** *maximale*	1	*set limits to the poleward range*
	Sous total	5	
	Total	**153**	

NATUREL			
natural	*(variabilité, effet de serre, aérosol, émission, cycle du carbone, etc.)* **naturel**	210	**natural** *(variability, greenhouse effect, aerosol, emission, carbon cycle, etc.)*
	Sous total	210	
naturally	**naturel**	5	**naturally**
	Sous total	5	
autre	*non* **naturel**	1	**unnaturally**
	prairie **naturelle**	1	**rangeland**
	prairie **naturelle**	1	**native** *meadow*
	pâturage **naturel**	1	**native** *grassland*
	réserve **naturelle**	2	**nature** *reserve*
	les milieux **naturels**	1	*the ecology*
	catastrophes **naturelles**	1	*environmental disasters*
	gaz **naturel**	8	**Ø** *gas*
	naturel	5	**Ø**
		1	**reliable**
		1	**Non-anthropogenic**
	Sous total	23	
	Total	**238**	

COMBUSTIBLE			
fuel	**combustible** *fossile*	116	*fossil* **fuel**
	combustible	28	**fuel**
	Sous total	145	
autre	**combustible** *fossile*	5	*fossil* **Ø**
		1	**carbon**
		1	**conventional oil and gas**
	combustible	1	**combustible**
	Sous total	8	
	Total	**153**	

ÉVALUATION			
assessment	*(modèle d', rapport d, etc.)* **évaluation** *(des incidences, du changement climatique, etc.)*	134	*(climate change, report, etc.)* **assessment** *(model, impact, etc.)*
	Sous Total	134	
asses (to)		5	**assess (to)**
	Sous Total	5	
estimate	**évaluation**	15	**estimate**
	Sous Total	15	
estimate (to)	**évaluation**	4	**estimating, estimated**
	Sous Total	4	
evaluation	**évaluation**	14	**evaluation**
	Sous Total	14	
autre	*deuxième rapport d'***évaluation**	11	*SAR, TAR (second* **assessment** *report, third* **assessment** *report)*
	évaluation	2	**Ø**
		2	**valuation**
		1	**quantification**
		1	**quantified**
		1	**estimation**
		1	**reviewing**
		1	**rating**
		1	**monitoring**
		1	**test**
	Sous Total	22	
	Total	194	

CÔTIER			
coastal	**côtier**	116	**coastal**
	Sous total	116	
autre	**côtier**	1	**shoreline**
	Sous total	1	
	Total	117	

NIVEAU			
level	**niveau** *de la mer*	93	*sea* **level**
	niveau *(marin, des mers, etc.)*	23	*sea* **level**
	niveau *de stabilisation*	17	*stabilization* **level**
	niveau	99	**level**
	Sous total	232	
target	**niveau** *(à atteindre, fixé, etc.)*	6	**target**
	Sous total	6	
autre	*au* **niveau**	11	**in**
		8	**of, with respect, on a basis, relative to, on a scale, in terms of**
	au **niveau** *du globe*	1	**globally**
	haut **niveau**	1	**higher** *altitudes*
	bas **niveau**	2	**lower**
	différents **niveaux**	1	*different* **concentration**
	niveau	1	**values**
		1	**layer**
		1	**rate**
		2	**degree**
		13	**Ø**
	Sous total	42	
	Total	280	

RÉDUCTION			
reduction	réduction	133	reduction
	Sous total	133	
reduced	réduction	28	reduced
	Sous total	28	
reduce (to)		35	reducing
	Sous total	35	
decrease	réduction	8	decrease
	Sous total	8	
cut	réduction	5	cut
	Sous total	5	
autre	réduction	3	Ø
		3	loss
		2	decline
		2	mitigation
		2	diminishing
		2	decreased
		1	limitation
		1	narrowing
		1	drop
		1	fewer
		1	less
		1	low
		1	lower
		1	downscaling
		1	disappearance
		1	abatement
		1	negatively
	Sous total	25	
	Total	234	

PIÉGEAGE			
sequestration	piégeage *du carbone*	10	*carbon* sequestration
		2	sequestration *of carbon*
	piégeage	11	sequestration
	Sous total	23	
capture	piégeage	13	capture
	Sous total	13	
autre	piégeage	1	sequester (to)
		2	trapping
		2	absorption
		1	removal
	Sous total	6	
	Total	42	

MÉTÉOROLOGIQUE			
weather	**météorologique**	88	**weather**
	Sous total	**88**	
meteorological	**météorologique**	22	**meteorological**
	Sous total	**22**	
autre	**météorologique**	13	**Ø**
		1	*WMO (World Meteorological Organization)*
		2	**climate**
		1	**meteorology**
	Sous total	**17**	
	Total	**127**	

GLACIAIRE			
ice	*calotte* **glaciaire**	26	**ice** *cap*
		11	**ice** *sheet*
	(carotte, écoulement, période, etc.) **glaciaire**	38	**ice** *(core, flow, age, etc.)*
	nappe **glaciaire**	17	**ice** *sheet*
	Sous total	**92**	
glacial	*(masse, ère, etc.)* **glaciaire**	14	**glacial** *(mass, period, etc.)*
	Sous total	**14**	
autre	**glaciaire**	1	**glaciation**
		2	**glacier**
	Sous total	**3**	
	Total	**109**	

MER			
sea	*(niveau de la, glace de, etc.)* **mer**	174	**sea** *(level, ice, etc.)*
	(niveau de la, glace de, etc.) **mer**	25	**sea-** *(level, ice, etc.)*
	Sous total	**199**	
seawater	*eau de* **mer**	8	**seawater**
	Sous total	**8**	
offshore	*en* **mer**	4	**offshore**
	Sous total	**4**	
autre	**mer**	3	**Ø**
		2	**ocean**
	bordure de **mer**	2	**coastal** *(zone, etc.)*
	au-delà des **mers**	2	**overseas**
	fond de la **mer**	1	**seabed**
	haute **mer**	1	*expenses of ocean*
	d'une plate-forme en **mer**	1	*from a stationary floating vessel*
	Sous total	**12**	
	Total	**223**	