Une tentative d'exploitation bi-directionnelle d'un corpus bilingue

Jennifer Pearson*

*Cet article cherche à établir si on peut utiliser un corpus bilingue pour faire des travaux terminologiques, et plus particulièrement pour repérer des éléments définitoires. La première partie de l'article traite de la construction et traitement d'un corpus bilingue (anglais-français) de textes spécialisés. Dans la deuxième partie, des marqueurs linguistiques signes de la présence d'éléments définitoires (*called *en anglais,* c'est-à-dire *en français) seront examinés, l'un repéré dans les textes-source et l'autre dans les textes-cible afin d'établir comment chacun de ces marqueurs est rendu dans l'autre langue. L'intention est d'établir si les marqueurs linguistiques que l'on peut repérer dans une langue réapparaissent dans l'autre et s'ils sont rendus de façon systématique. A plus long terme, le but est d'établir si l'on peut utiliser des marqueurs linguistiques dans une langue pour accéder à des éléments définitoires dans une autre langue, des éléments qui seraient peut-être autrement difficilement repérables.*

*This paper attempts to establish whether it is possible to use a bilingual corpus for terminological purposes and especially for retrieving definitional information in two languages. The first section of the article describes the compilation and preliminary processing of the bilingual (English-French) corpus of specialised texts used in the study. In the second section, two linguistic markers (*called *in English,* c'est-à-dire *in French) are analysed with a view to ascertaining whether and how they are translated. When markers are translated, we examine whether they are translated systematically. When they are not translated, we try to establish whether we can use markers in one language to retrieve definitional information in the other language, even when it is not explicitly marked.*

* SALIS, Dublin City University, Dublin, Irlande.

1. Introduction

Les traductologues, c'est-à-dire des chercheurs qui s'intéressent à ce qui se passe pendant le processus de la traduction, ont commencé à s'intéresser à des corpus bilingues, et plus particulièrement à des corpus dits parallèles (des corpus contenant des textes et les traductions de ces textes) au cours des années 1990 (Baker 1999, Laviosa 1998, Malmkjær 1998, Kenny 1998)[1]. Un des buts principaux du travail des traductologues utilisant des corpus parallèles est d'établir si les textes qui ont été traduits (en français par exemple) diffèrent de façon essentielle des textes non traduits, c'est-à-dire des textes rédigés en français. Ils cherchent à savoir si les mots et les structures utilisés par les traducteurs sont représentatifs des mots et des structures utilisés dans le même genre de texte non traduit et, s'ils ne sont pas représentatifs, ils cherchent à évaluer les différences. Certaines études (Laviosa 1998 par exemple) montrent que les traducteurs ont tendance à utiliser, quand ils traduisent, un vocabulaire plus réduit et une gamme de structures grammaticales plus étroite qu'ils n'utiliseraient en situation classique de rédaction de texte.

Jusqu'à présent, les études en traductologie basées sur des corpus parallèles portaient principalement sur des textes de langue générale[2]. Les corpus parallèles avec lesquels les traductologues travaillent contiennent souvent des textes narratifs de fiction (tels que des romans ou des nouvelles) ou des articles de journaux. Les traductologues ne se sont pas encore penchés sur des textes spécialisés, c'est-à-dire sur des textes non-narratifs traitant de domaines spécialisés tels que la science, l'économie, etc. C'est pour cette raison que nous avons créé le corpus parallèle de textes spécialisés sur lequel cet article va s'appuyer.

Le travail présenté ici[3] s'inscrit dans le débat concernant le processus de traduction et ce qui arrive au contenu sémantique d'un texte quand on le transfère d'une langue à une autre. Il constitue un premier pas dans notre quête d'établir quels travaux terminologiques liés à la traduction on peut envisager de faire avec un corpus parallèle. Nous allons regarder ce qui arrive à des éléments définitoires au cours de la traduction. Nous proposons d'accéder aux éléments définitoires terminologiques par l'intermédiaire de marqueurs linguistiques. Nous voulons voir comment, et si, les marqueurs sont rendus dans l'autre langue. Si les marqueurs linguistiques sont rendus dans l'autre langue de façon systématique, nous pensons que cela permettrait

1 Les études de Daille (1994) et de Bourigault, Chodiewicz et Humbley (1999) sont, à notre connaissance, les seules études *terminologiques* à avoir été faites sur des corpus bilingues et parallèles.

2 D'autres types de corpus parallèles sont parfois utilisés par des chercheurs en *data driven language learning* ou dans des travaux sur l'alignement.

3 Je remercie le relecteur anonyme à l'Université de Toulouse-Le Mirail pour ses commentaires qui ont permis de notablement améliorer cet article.

de récupérer simultanément dans les deux langues les éléments définitoires associés avec ces marqueurs. Si les marqueurs ne sont pas traduits ou ne sont pas traduits de façon systématique, nous pensons que cela ne nous empêchera pas de récupérer les éléments définitoires dans les deux langues parce que nous pourrons utiliser les marqueurs dans une langue pour accéder dans l'autre langue aux éléments définitoires non marqués. Nous voulons aussi aller au-delà d'un travail purement traductologique parce que nous voulons non seulement travailler à partir de marqueurs repérés dans les textes-source mais aussi à partir de marqueurs repérés dans les textes-cible. A plus long terme, le but est d'établir si l'on peut utiliser des marqueurs linguistiques dans une langue pour accéder à des éléments définitoires dans une autre langue, éléments qui seraient peut-être autrement difficilement repérables.

Dans la première partie de cet article le corpus qui a été créé sera décrit. On parlera aussi de certaines caractéristiques du corpus qui risquent de poser des problèmes pour l'alignement. Dans la deuxième partie, deux marqueurs linguistiques (*called* en anglais, *c'est-à-dire* en français) seront examinés pour voir comment, et si, ils sont rendus dans l'autre langue.

2. Construction et description du corpus

Quand nous avons commencé à chercher des textes candidats pour le corpus, quelques critères importants sont entrés en jeu. Il fallait que les textes soient spécialisés, qu'ils soient publiés, qu'ils soient traduits, et il fallait que le rapport auteur-lecteur soit un rapport de communication entre des experts et des gens qui n'ont aucune ou peu de formation dans le domaine en question, mais qui s'y intéressent pour des raisons professionnelles ou privées (cf. Pearson 1999 pour une discussion plus détaillée de cette dernière question). Nous voulions donc trouver des textes qui étaient susceptibles de contenir des éléments définitoires.

Puisque nous avions une préférence pour des textes scientifiques nous avons pensé que des articles parus dans *Scientific American* pourraient convenir. La revue *Scientific American* répondait à nos critères de typologie et les articles sont traduits et publiés en français dans *Pour la Science*[4]. Ayant pris contact avec les éditeurs de ces deux revues, nous avons pu obtenir 63 articles et leurs traductions en français. Les articles traduits nous ont été envoyés par courrier électronique et les articles d'origine ont été obtenus sur le site web de *Scientific American* ou scannés. La collecte des textes était une tâche non négligeable, ce qui fait que la création et la correction du corpus ont pris beaucoup plus de temps que prévu.

[4] Je remercie les éditeurs de *Pour la Science,* et Philippe Pajot en particulier, d'avoir bien voulu me fournir les articles par courrier électronique et d'avoir répondu à toutes mes questions.

2.1. Quelques chiffres indicatifs

Une fois que les deux collections de textes furent disponibles sous forme électronique, quelques données statistiques (résumées dans le tableau 1) ont été relevées[5].

Tableau 1 : Statistiques pour les 63 articles parus dans *Scientific American* (SA) et leurs traductions parues dans *Pour la Science* (PLS)

Pub	Occurrences	Types	Rapport type/occurrence	Rapport type/occurrence normalisé[6]	Nombre de phrases
SA	187.159	15.613	8,34	45,27	6.737
PLS	168.518	17.478	10,37	44,10	4.899

On peut voir qu'il y a 187.159 *tokens* ou *occurrences* (mots) dans les articles de *Scientific American et* 168.518 *occurrences* dans les traductions. Le corpus est donc assez petit. L'intention à moyen terme est de l'agrandir avec d'autres types de textes.

On peut aussi constater qu'il y a moins d'*occurrences* dans les traductions que dans les textes d'origine, ce qui, à première vue, est contraire à la norme pour les traductions d'anglais vers le français. Cependant, quand on regarde les textes d'un peu plus près on constate que si les traductions sont en effet beaucoup plus courtes que les textes d'origine, l'explication ne vient pas de ce que les traductions sont plus précises ou plus concises . La raison réside plutôt dans le fait qu'il y a de nombreux segments (titres, phrases et même des paragraphes entiers) dans les articles d'origine qui ne sont pas du tout reproduits dans les traductions, ce que l'on aurait peut-être pu deviner en comparant le nombre de phrases dans les textes d'origine (6.737) et le nombre de phrases dans les traductions (4.899 seulement).

Bien que les traductions soient plus courtes que les textes d'origine, le nombre de *types* (mots différents) est beaucoup plus élevé : 17.478 dans les traductions par rapport à seulement 15.613 dans les textes d'origine. Ceci est probablement dû en partie au plus grand nombre de *types* dits de 'classe

[5] Les analyses statistiques ont été faites avec WordSmith Tools, un logiciel conçu par Mike Scott de l'Université de Liverpool et commercialisé par Oxford University Press.

[6] Puisque le rapport type/occurrence variera suivant la longueur d'un texte ou la taille d'un corpus, il n'est pas très intéressant en lui-même. C'est pour cette raison que nous produisons un rapport type/occurrence normalisé pour l'ensemble du corpus. Le rapport type/occurrence normalisé est calculé tous les *n* mots pendant que le logiciel traite chaque fichier (*n* = 1.000). Le rapport est calculé pour les 1.000 premiers mots, ensuite pour les 1.000 mots suivants et ainsi de suite jusqu'au bout du corpus. La moyenne est calculée, donnant un rapport type/occurrence moyen basé sur des morceaux consécutifs de 1.000 mots. Il en est le même de la normalisation des longueurs de phrases.

fermée' en français et aussi au fait que la morphologie française est plus riche que la morphologie anglaise.

3. Alignement des textes

Quand on veut faire des études systématiques sur des collections de textes parallèles bilingues, on a besoin d'un certain nombre d'outils qui facilitent la tâche. Le travail décrit dans cet article nécessite des outils de balisage, d'alignement et de production de concordances bilingues.

Un corpus aligné est essentiel pour tous ceux qui veulent faire des études contrastives. Le but de l'alignement consiste à repérer et à créer des liens entre les textes d'origine et leurs traductions pour permettre l'utilisation d'un concordancier bilingue. L'alignement peut se faire au niveau des paragraphes, au niveau des phrases et quelquefois même au niveau des mots. Pour les travaux décrits ici, nous avons utilisé un outil, nommé MinMark (conçu par David Woolls à Birmingham en Angleterre) qui balise les textes en identifiant la fin des phrases et des paragraphes et les aligne au fur et à mesure pendant la production de concordances bilingues.

En règle générale les algorithmes d'alignement sont basés sur un certain nombre de suppositions concernant les textes à aligner (cf. Gale et Church 1993 ; Johansson, Ebeling et Hofland 1996). Malheureusement certaines des suppositions ne se confirment pas toujours en réalité. Par exemple, les algorithmes d'alignement nécessitent généralement une correspondance séquentielle entre les textes d'origine et les traductions, au moins au niveau des paragraphes et de préférence au niveau des phrases. Cela veut dire que le premier segment (un segment peut être une phrase, plusieurs phrases ou même un paragraphe entier) du texte d'origine est censé correspondre au premier segment de la traduction, le deuxième au deuxième et ainsi de suite jusqu'à ce que des liens aient été établis entre chaque paire de segments.

Malheureusement, la réalité est souvent bien différente parce que rien n'empêche les traducteurs de restructurer des textes en changeant l'ordre des phrases à l'intérieur des paragraphes, en deplaçant des phrases d'un paragraphe à un autre, en inversant l'ordre des paragraphes dans un texte, et c'est précisément ce qui s'est passé dans notre corpus.

Bien qu'il soit vrai qu'il existe maintenant des outils d'édition d'alignement (tel WinAlign de Trados) qui permettent à l'utilisateur de corriger ce genre de problème, il n'empêche que cela demande un travail manuel assez important. Puisque nous n'avions pas accès à une version professionnelle de WinAlign et que nous ne voulions pas changer l'ordre des segments (cela risquait de nuire plus tard à la possibilité de faire des études contrastives sur la structuration d'information dans les textes), nous n'avons pas cherché à avoir une correspondance parfaite en ce qui concernait l'ordre des paragraphes.

Une deuxième supposition sur laquelle reposent les outils d'alignement est que chaque segment du texte d'origine va être reproduit dans la traduction, c'est-à-dire que tout le contenu sémantique du texte d'origine va se retrouver dans la traduction.

Dans notre corpus, par exemple, cela n'est très souvent pas le cas. Il y a de nombreux segments dans les textes d'origine qui ne sont pas traduits du tout ; nous avons compté cinquante deux paragraphes, trente titres ou sous-titres manquants et il y avait un nombre non calculé de phrases non traduites. Les segments qui ont été omis contiennent, entre autres, des précisions sur la recherche menée aux Etats Unis.

Quand on constate qu'il manque des segments, on peut insérer des marqueurs du genre 'paragraphe manquant' 'titre manquant' suivi d'un retour chariot pour rétablir un parallèle exact. C'est ce que nous avons fait pour tout paragraphe et titre manquant mais nous n'avons pas cherché à le faire au niveau des phrases, tout simplement parce que cela aurait été pratiquement impossible.

Une troisième supposition est que l'on ne trouvera pas dans les traductions des informations qui n'ont pas été fournies dans les textes d'origine. Une fois de plus, la réalité est bien différente. Dans les traductions, il y a six paragraphes, vingt-huit titres ou sous-titres et de nombreuses phrases supplémentaires sans équivalent dans les textes d'origine. En règle générale, l'information ajoutée consiste en des précisions concernant la recherche en France ou en Europe.

Là aussi, chaque fois qu'il manquait un paragraphe ou un titre, nous avons inséré un marqueur suivi d'un retour chariot pour avoir un parallèle exact. Dans les figures 1 et 2, on trouvera un extrait d'un des articles du corpus avant, et après, balisage. Dans la figure 1, on peut voir qu'au début du texte-cible, il y a un paragraphe qui n'a pas d'équivalent dans le texte-source, et qu'il y a un titre dans le texte-source qui n'a pas d'équivalent dans le texte-cible. Il y a aussi des différences entre le texte d'origine et la traduction au niveau de la segmentation des phrases. On peut comprendre pourquoi il est important d'insérer des titres et des paragraphes fantômes. Sans cela, le concordancier que nous utilisons qui aligne les textes au fur et à mesure qu'il produit les concordances risquerait de donner de mauvais résultats.

Figure 1 : Extrait d'un article de notre corpus avant balisage et avant correction

Cosmic Rays at the Energy Frontier Roughly once a second, a subatomic particle enters the earth's atmosphere carrying as much energy as a well-thrown rock. Somewhere in the universe, that fact implies, there are forces that can impart to a single proton 100 million times the energy achievable by the most powerful earthbound accelerators. Where and how? Giants Those questions have occupied physicists since cosmic rays were first discovered in 1912 (although the entities in question are now known to be particles, the name « ray » persists). The interstellar medium contains atomic nuclei of every element in the periodic table, all moving under the influence of electrical and magnetic fields. Without the screening effect of the earth's atmosphere, cosmic rays would pose a significant health threat.	Les rayons cosmiques Ces noyaux atomiques provenant de l'espace ont une énergie supérieure à celle que pourraient communiquer les plus gros accélérateurs de particules. Les astrophysiciens recherchent leur origine. Une fois par seconde, un noyau atomique pénètre dans l'atmosphère terrestre avec autant d'énergie qu'une pierre lancée vigoureusement. Appliquée à un proton, cette même énergie est 100 millions de fois supérieure à celle qu'imprimeraient les accélérateurs terrestres les plus puissants. Quels mécanismes accélèrent-ils des noyaux atomiques et où se produisent-ils ? Cette question préoccupe les physiciens depuis la découverte des rayons cosmiques, en 1912. À l'époque, on ignorait la nature particulaire des rayons cosmiques, d'où leur nom impropre, qui a subsisté. Le milieu interstellaire contient des noyaux atomiques de tous les éléments, qui se déplacent dans des champs électriques et magnétiques. Sans la protection de l'atmosphère terrestre, beaucoup plus de rayons cosmiques arriveraient jusqu'au sol.

Figure 2 : Extrait d'un article de notre corpus après balisage et après correction

<body> <p><s>Cosmic Rays at the Energy Frontier <p><s>Paragraph missing <p><s>Roughly once a second, a subatomic particle enters the earth's atmosphere carrying as much energy as a well-thrown rock. <s>Somewhere in the universe, that fact implies, there are forces that can impart to a single proton 100 million times the energy achievable by the most powerful earthbound accelerators. <s>Where and how?	<body> <p><s>Les rayons cosmiques <p><s>Ces noyaux atomiques provenant de l'espace ont une énergie supérieure à celle que pourraient communiquer les plus gros accélérateurs de particules. <s>Les astrophysiciens recherchent leur origine. <p><s>Une fois par seconde, un noyau atomique pénètre dans l'atmosphère terrestre avec autant d'énergie qu'une pierre lancée vigoureusement.

<p><s>Giants <p><s>Those questions have occupied physicists since cosmic rays were first discovered in 1912 (although the entities in question are now known to be particles, the name « ray » persists). <s>The interstellar medium contains atomic nuclei of every element in the periodic table, all moving under the influence of electrical and magnetic fields. <s>Without the screening effect of the earth's atmosphere, cosmic rays would pose a significant health threat.	<s>Appliquée à un proton, cette même énergie est 100 millions de fois supérieure à celle qu'imprimeraient les accélérateurs terrestres les plus puissants. <s>Quels mécanismes accélèrent-ils des noyaux atomiques et où se produisent-ils? <p><s>No heading <p><s>Cette question préoccupe les physiciens depuis la découverte des rayons cosmiques, en 1912. <s>À l'époque, on ignorait la nature particulaire des rayons cosmiques, d'où leur nom impropre, qui a subsisté. <s>Le milieu interstellaire contient des noyaux atomiques de tous les éléments, qui se déplacent dans des champs électriques et magnétiques. <s>Sans la protection de l'atmosphère terrestre, beaucoup plus de rayons cosmiques arriveraient jusqu'au sol.

Etant donné tous les problèmes potentiels posés par l'organisation des textes dans notre corpus, on pourrait penser que les résultats de concordances bilingues laisseraient beaucoup à désirer. Curieusement, et à notre grande surprise, les résultats pour le travail que nous faisons ont généralement été assez bons. A titre d'exemple, nous reproduisons dans la figure 3 un extrait de la concordance bilingue pour le type *asbestos.* Cette concordance bilingue a été produite avec l'outil MultiConcord de David Woolls à Birmingham. L'outil aligne les textes au fur et à mesure qu'il produit les lignes de concordance.

Figure 3 : Extrait de la concordance bilingue pour *asbestos*

The future for asbestos appears downright grim.	L'avenir industriel de l'amiante est compromis.
Efficient asbestos air filters were used in hospital ventilators, cigarette tips and military gas masks.	Des filtres à air en amiante étaient utilisés avec succès dans les ventilateurs d'hôpitaux, dans les cigarettes et dans les masques à gaz militaires.
Over the next 1,000 years, asbestos continued to attract the attention of kings and chemists from western Europe to China.	Paragraphe manquant
Somewhere along the line, though	Check Paragraph View

the fact that asbestos was a stone seems to have been forgotten.	
Marco Polo serendipitously brought asbestos back to the realm of science.	D'autres pensent qu'elle provient d'écailles de lézard ou de plumes d'oiseau.

Il y a 101 occurrences d'*asbestos* dans le corpus ; sur ces 101 occurrences, il y en a 86 où l'outil d'alignement et de concordance trouve automatiquement une phrase qu'il estime être équivalente. Il y en a 5 où il récupère un équivalent fantôme ; c'est le cas du troisième exemple dans la figure 3 où il n'y pas de paragraphe correspondant dans le texte-cible. Il y a 10 occurrences où le concordancier ne trouve pas automatiquement de phrase équivalente dans le texte-cible. Cela est indiqué par l'instruction *check paragraph view* (voir le quatrième exemple dans la figure 3). Quand *check paragraph view* s'affiche, cela veut dire que le logiciel a du mal à aligner le paragraphe en question. *Check paragraph view* permet à l'utilisateur de regarder en amont et en aval de la ligne de concordance affichée. Quand on fait cela, on découvre généralement qu'il y a moins de phrases dans le paragraphe cible que dans le paragraphe source, et que c'est pour cette raison que le processus d'alignement et de production de concordances n'ont pas pu continuer pour le paragraphe en question. *Check paragraph view* permet aussi à l'utilisateur d'identifier et de récupérer une phrase équivalente si elle existe. Sur les 101 lignes de concordance bilingue, il y en a seulement cinq qui sont incorrectes (voir le dernier exemple dans notre figure). Elles sont incorrectes parce que l'outil d'alignement a mal fonctionné. Etant donné la nature du corpus, on aurait pu s'attendre à des résultats bien pires que ceux décrits ici.

Nous concluons donc que même si certaines études ne sont pas possibles et que d'autres (telles que des études lexicales quantitatives) ne donneront pas toujours des résultats utiles, beaucoup d'autres types d'études restent possibles. Par exemple, le fait d'ajouter des marqueurs 'paragraphe ou titre manquant' permet de repérer les endroits où de l'information a été supprimée ou ajoutée. On peut donc envisager de mener des études sur ces omissions ou ces additions pour comprendre ce qui s'est passé. Au niveau lexical ou syntaxique, on peut examiner comment certaines unités lexicales ou certaines structures grammaticales sont rendues en traduction. Le corpus a, par exemple, déjà été utilisé pour étudier comment sont traduits les adverbes, et des mots 'ordinaires' comme *researcher* et *scientist,* avec des résultats parfois étonnants. Même s'il est vrai qu'il aurait été préférable de travailler avec un corpus qui se laisse parfaitement aligner, le fait que cela ne soit pas le cas ne nous a pas trop handicapés.

4. Traduction d'éléments définitoires

Le rapport auteur-lecteur qui existe entre les auteurs des articles pour *Scientific American* et les lecteurs n'est pas un rapport d'égalité, dans le sens où les auteurs sont susceptibles d'avoir plus de connaissances que leurs lecteurs. On peut donc s'attendre à ce que les auteurs fournissent des explications quand ils introduisent des notions qu'ils pensent ne pas être connues par leurs lecteurs. C'est en effet ce qui se passe dans notre corpus. Les auteurs fournissent des explications de façons différentes : ils peuvent, entre autres, utiliser une analogie, donner une définition, préciser un synonyme. Quand ils donnent une définition, ils peuvent le faire en fournissant une définition classique qui suit la formule X = Y + caractéristique distinctive (assez rare) ou en donnant des éléments définitoires (beaucoup plus fréquent). Il y a un certain nombre de patrons syntactiques et de marqueurs linguistiques qui semblent signaler la présence d'éléments définitoires dans les textes. Malheureusement, puisque nous travaillons actuellement avec un corpus non étiqueté, nous ne pouvons pas encore accéder automatiquement à des éléments définitoires par le biais de patrons syntactiques. Par contre, nous pouvons accéder automatiquement à des éléments définitoires associés avec des marqueurs linguistiques en utilisant un concordancier bilingue, et nous n'avons pas besoin d'étiqueter notre corpus pour le faire. C'est pour cette raison que nous avons choisi de focaliser d'abord sur les marqueurs linguistiques.

Nous proposons d'examiner deux marqueurs linguistiques, signes de présence d'éléments définitoires, *called* en anglais et *c'est-à-dire* en français. Nous voulons établir quelles sortes d'informations sont asssociées avec ces marqueurs. Nous savons déjà que, dans certains types de textes, les informations associées avec les marqueurs linguistiques sont souvent des éléments définitoires qui peuvent être utilisés ensuite dans la construction de définitions terminologiques (cf. Condamines et Rebeyrolle, à paraître, Pearson 1998). Nous voulons voir comment, et si, les marqueurs et les éléments définitoires sont rendus dans l'autre langue.

Voici les questions que nous nous posons. Quand un marqueur linguistique dans une langue est rendu de façon systématique dans l'autre langue, est-ce qu'on peut utiliser le marqueur pour récupérer les éléments définitoires associés avec le marqueur dans les deux langues simultanément? Si cela était possible, un travail terminographique bilingue deviendrait plus facile. Quand un marqueur constaté dans une langue est remplacé par quelque chose de moins facilement repérable dans l'autre langue (quand il est remplacé par une virgule, par exemple), ou quand un marqueur n'est pas rendu du tout dans l'autre langue est-ce que, étant donné que nous utilisons un concordancier bilingue, on peut accéder à l'information dans le texte-cible par le biais du marqueur dans le texte-source, et vice versa ? Si cela s'avèrait être possible, cela voudrait dire qu'on pourrait envisager d'utiliser des marqueurs linguistiques d'une langue pour accéder à des éléments

définitoires dans une autre langue qui ne seraient peut-être pas autrement repérables.

4.1. Traduction du marqueur linguistique *called*

Les informations associées avec le marqueur linguistique *called* peuvent être classées de la façon suivante (Pearson 1998) :

- Mot ou phrase de la langue générale à gauche du marqueur, terme correct à droite. Exemple de notre corpus :
  A better technique is to encase the gene in a fatty bubble ***called*** *a liposome.*
  Une meilleure technique consiste à insérer d'abord le gène dans une sphère, ***nommée*** *liposome.*

- Hypéronyme à gauche du marqueur, hyponyme à droite. Exemple de notre corpus :
  We found that in starfish, cells ***called*** *coelomocytes (the equivalent of macrophages) produce IL-1.*
  Nous avons montré que des cellules ***nommées*** *cœlomocytes (l'équivalent des macrophages) produisent de l'interleukine 1.*

- Terme à gauche du marqueur, synonyme à droite. Aucun exemple de ce patron dans notre corpus.

Sur les 130 occurrences du mot *called* dans le corpus, 77 occurrences correspondaient au simple marqueur *called,* les autres étaient précédées du verbe *être* et ont donc été éliminées de cette discussion. Sur les 77 occurrences examinées, la présence du marqueur en anglais indiquait ou une relation d'équivalence ou une relation d'hypéronymie-hyponymie mais jamais une relation de synonymie, contrairement à ce que nous avions trouvé dans d'autres corpus (Pearson 1998).

Quand on regarde comment le marqueur *called* a été traduit, on peut constater qu'il a été traduit relativement souvent par le marqueur *nommé** (42 fois), qu'il a été remplacé assez souvent (19 fois) par une virgule, que quelquefois, le marqueur n'étais pas rendu du tout (15 fois) et que, assez rarement, un des éléments définitoires (ou a droite ou a gauche du marqueur) n'a pas été traduit.

Chaque fois que le marqueur *called* est rendu par *nommé** en français dans notre corpus, le terme défini est à droite du marqueur et l'équivalent ou l'hypéronyme est à gauche. Cela veut dire qu'en ce qui concerne l'utilisation du marqueur *nommé** pour traduire le marqueur *called,* les structures anglaises et françaises sont identiques. Elles devraient donc être faciles à repérer et à extraire à des fins terminologiques.

Quand le marqueur *called* est rendu par une virgule en français, le terme défini se retrouve, comme précédemment avec *nommé*,* à droite du

marqueur et l'équivalent ou l'hypéronyme est à gauche. Quelques exemples sont fournis ici à titre indicatif.

The soliders of acquired immunity are the specialized white blood cells ***called*** *lymphocytes*
Les « soldats » de l'immunité acquise sont des globules blancs spécialisés, les lymphocytes.

a positively charged organic polymer ***called*** *DEAE-dextran.*
un polymère organique chargé positivement, le DEAE-dextrane.

the most vulnerable cells in the brain are the many neurons that respond to an extremely powerful neurotransmitter ***called*** *glutamate.*
les cellules cérébrales les plus vulnérables sont les neurones sensibles à un neuromédiateur excitateur, le glutamate.

that HIV can infect and persist for years in another class of CD4-carrying immune cells ***called*** *macrophages.*
que le VIH infecte une autre classe de cellules immunitaires présentant des molécules CD4, les macrophages,

Puisque la virgule en français a, en plus de son rôle de marqueur de la présence d'un élément définitoire, de nombreuses autres fonctions très différentes, il est très probable que si le marqueur linguistique anglais n'avait pu être utilisé pour accéder aux éléments définitoires séparés par une virgule en français, ces éléments n'auraient pas été repérés en français. Cela confirmerait que l'on peut envisager d'utiliser des marqueurs dans une langue pour accéder à des éléments définitoires dans une autre langue même quand les éléments dans la deuxième langue ne sont pas très fortement marqués.

A part les cas où le marqueur est rendu par une virgule ou par le marqueur français *nommé**, il existe aussi des cas où le marqueur n'est pas rendu du tout dans la traduction. Voici quelques exemples.

a protein ***called*** *bcl-2, which suppresses cell suicide.*
un gène qui code la protéine bcl-2, laquelle bloque le suicide cellulaire.

a bacterium, ***called*** *Thermus aquaticus, that would later make*
il s'agissait de la bactérie Thermus aquaticus, dont l'utilisation fut

This property, ***called*** *giant magnetoresistance, will have a considerable impact on magnetic.*
cette magnétorésistance géante devrait être utilisée pour le stockage des données.

Comme on peut voir dans les deux premiers exemples, l'hypéronyme est précisé en anglais et en français mais il n'est pas marqué de façon

explicite en français. Dans des cas comme ceux-ci, on pourrait envisager d'utiliser le marqueur anglais pour accéder à des éléments définitoires en français que l'on n'aurait peut-être pas décelés autrement. Dans le troisième exemple (et il y en a eu quelques-uns comme celui-ci) l'hypéronyme est précisé en anglais mais non pas en français. On pourrait donc envisager d'induire l'hypéronyme français à partir de l'hypéronyme anglais, et ainsi le terme *magnétorésistance* acquérrait l'hypéronyme *property.*

4.2. Origines du marqueur linguistique *c'est-à-dire*

Comme nous l'avons indiqué dans l'introduction, nous voulions aller au-delà d'un travail purement traductologique. Nous voulions non seulement travailler dans le sens de la traduction mais aussi dans le sens inverse. Ayant examiné ce qui arrive à un marqueur linguistique anglais quand il est traduit en français, nous voulions ensuite constater quelles étaient les origines dans les textes-source d'un marqueur repéré dans les textes-cible. En faisant cela, nous espérions constater qu'il serait utile d'envisager, à plus long terme, une exploitation bidirectionnelle du corpus. Nous avons choisi d'examiner les origines du marqueur *c'est-à-dire.* Si l'on regarde les occurrences de ce marqueur dans notre corpus, on peut classer les informations qui sont associées avec ce marqueur de la façon suivante :

- Terme à gauche du marqueur linguistique, hypéronyme et phrase définitoire à droite.
  Exemple de notre corpus :
  *un interféromètre, **c'est-à-dire** un appareil composé de deux miroirs et de deux lames séparatrices.*
  an interferometer - a device consisting of two mirrors and two beam splitters.

- Terme à gauche du marqueur linguistique, phrase équivalente à droite.
  Exemple de notre corpus :
  *la « fixation » naturelle de l'azote (**c'est-à-dire** la dissociation de la molécule de diazote et l'incorporation ultérieure des deux atomes d'azote dans l'ammoniac) est l'œuvre de bactéries.*
  most natural nitrogen « fixation » (the splitting of paired nitrogen molecules and subsequent incorporation of the element into the chemically reactive compound ammonia) is done by certain bacteria.

- Terme à gauche du marqueur linguistique, explication sans hypéronyme à droite.
  Exemple de notre corpus :
  *relatives aux cellules somatiques, **c'est-à-dire** ne concernant ni les spermatozoïdes ni les ovocytes.*
  affecting somatic cells, the kinds that are neither sperm nor egg.

- Phrase terminologique à gauche du marqueur linguistique, phrase équivalente à droite.
  Exemple de notre corpus :
  *comment réduire l'immunogénicité de la protéine (**c'est-à-dire** diminuer sa capacité à activer le système immunitaire), pour de potentielles applications pharmacologiques.*
  also examining chemical modifications to reduce immunogenicity of alpha-hemolysin - its tendency to provoke an attack by the immune system-for biotherapeutic applications.

Sur les 37 occurrences du marqueur *c'est-à-dire* dans notre corpus, il y avait seulement une occurrence avec une origine linguistique dans le texte-source. C'était l'exemple suivant, où l'origine de *c'est-à-dire* était *that is* (l'équivalent de *i.e.* en anglais) :

> *forment des métastases, **c'est-à-dire** qu'elles migrent vers des sites éloignés et y prolifèrent.*
> *to metastasize that is, to migrate to distant sites and to grow in unfamiliar territory.*

Quand nous avons cherché l'origine, dans les textes-source, de toutes les occurrences, nous avons pu constater qu'elle reposait souvent dans un des signes de ponctuation suivants : un trait (-) (14 fois), deux points (3 fois), une virgule (7 fois).Sinon, à la place du marqueur, ce qui aurait dû apparaître à droite du marqueur apparaît entre des parenthèses (4 fois), comme dans l'exemple suivant :

> *la « fixation » naturelle de l'azote (**c'est-à-dire** la dissociation de la molécule de diazote et l'incorporation ultérieure des deux atomes d'azote dans l'ammoniac) est l'œuvre de bactéries.*

> *most natural nitrogen « fixation » (the splitting of paired nitrogen molecules and subsequent incorporation of the element into the chemically reactive compound ammonia) is done by certain bacteria.*

Ce qui est intéressant avec le marqueur *c'est-à-dire* est d'abord le fait qu'il n'ait pas d'origine textuelle dans les textes-source. Alors que le traductologue chercherait peut-être à comprendre pourquoi le traducteur a choisi d'insérer des marqueurs à des endroits où il n'y en avait pas auparavant, le terminologue va se réjouir parce que cela luit permet de déceler des éléments définitoires dans les textes-source qu'il n'aurait peut-être pas repérés autrement.

5. Conclusion et perspectives

Dans cet article nous avons voulu établir si l'on pouvait exploiter un corpus bilingue à des fins terminologiques. Nous espérions aussi découvrir si un travail bi-directionnel était possible. Nous avons utilisé un corpus de textes spécialisés rédigés en anglais et traduits en français. Nous avons constaté qu'il y avait des différences importantes entre les textes-source et les textes-cible. Ces différences risquaient de poser des problèmes pour un travail bilingue. Nous avons observé notamment dans les textes-source des phrases, des titres voire même des paragraphes entiers qui n'ont pas été traduits ; dans les traductions, nous avons observés des segments (phrases, titres et paragraphes) qui n'avaient pas de correspondance dans les textes-source. Bien qu'il existe des algorithmes d'alignement – qui utilisent des 'mots ancres', par exemple (Johansson et alia 1996) - qui auraient peut-être réglé une partie des difficultés, il n'empêche qu'il restera difficile de concevoir un algorithme qui puisse détecter certaines des caractéristiques de notre corpus, telles que des changements dans les séquences des phrases ou des paragraphes, des omissions ou même l'ajout d'éléments nouveaux. Cependant, à notre grand étonnement, ces différences ont eu des conséquences moins lourdes pour notre travail que l'on ne pouvait penser. Si l'outil d'alignement et de production de concordances bilingues que nous utilisions n'a pas toujours trouver la bonne solution, les résultats ont été suffisamment encourageants pour nous permettre de continuer.

Puisque nous n'avions pas la possibilité d'étiqueter notre corpus, nous avons choisi de faire un travail qui ne nécessitait pas d'étiquetage et nous avons analysé deux marqueurs linguistiques. L'analyse du marqueur *called* nous a montré plusieurs choses. Quand le marqueur est traduit, il est traduit par *nommé** ; ces cas-là sont les plus faciles à récupérer dans un travail terminologique bilingue. Quand le marqueur n'est pas traduit, il peut être remplacé par une virgule ou par rien du tout. Ces derniers sont très intéressants parce que le marqueur anglais est notre seul moyen d'accéder à ces éléments définitoires non marqués en français. Nous avons constaté aussi que les éléments définitoires (de chaque côté de *called*) dans les articles d'origine ne sont pas toujours traduits. Cela veut dire qu'on peut envisager d'utiliser l'information dans le texte d'origine pour compléter l'information dans la traduction (indication de relation hypéronyme-hyponyme par exemple). L'analyse du marqueur *c'est-à-dire* a été très révélatrice. Nous avons pu constater qu'à une exception près ce marqueur n'a pas d'origine linguistique dans les textes sources et que ses origines résident plutôt dans des signes de ponctuation. Cela veut dire que nous pouvons repérer des éléments définitoires dans les textes anglais en utilisant un marqueur français confirmant qu'un travail bi-directionnel est possible.Nous avons donc pu confirmer qu'il est possible d'utiliser des marqueurs linguistiques dans une langue pour accéder à des éléments définitoires dans l'autre langue même si le marqueur lui-même n'a pas été rendu. Pour nous, cela met en relief le

bénéfice que le travail sur corpus bilingue pourrait nous apporter par rapport au travail sur corpus unilingue. Bien que le travail présenté ici ne représente qu'un début très timide, nous trouvons que les résultats sont suffisamment prometteurs pour nous encourager à poursuivre un travail plus approfondi. Nous avons donc l'intention de poursuivre ce travail en traitant d'autres marqueurs de la même façon.

En ce qui concerne les éléments définitoires que l'on pourrait récupérer en précisant des patrons syntactiques, nous avons l'intention d'étiqueter les textes-source et les textes-cible afin de pouvoir faire un travail plus approfondi sur l'identification de ces éléments et sur la traduction de ces éléments.

Les résultats obtenus jusqu'à présent nous indiquent qu'il est utile d'exploiter des corpus parallèles de textes spécialisés pour des travaux terminologiques et plus particulièrement pour repérer simultanément des éléments définitoires dans deux langues. Un autre bénéfice inattendu est l'idée que le travail dans un sens va donner des résultats différents d'un travail fait dans l'autre. On peut donc envisager une approche bi-directionnelle qui enrichirait les résultats d'un travail uniquement uni-directionnel.

Références bibliographiques

Baker, M. (1999), « The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators », in *International Journal of Corpus Linguistics* 4. 2, pp. 281-298.

Bourigault, D., Chodiewicz, C. & Humbley, J. (1999), « Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné », in *Terminologies Nouvelles* 19, décembre 1998-juin 1999, pp. 70-77.

Bowker, L. & Pearson, J. (à paraître 2001), *Working with Specialised Language : a practical guide to using corpora,* London, Routledge.

Condamines, A. & Rebeyrolle, J. (à paraître), « Searching for and identifying conceptual relationships via a Corpus-based approach to a Terminological Knowledge Base (CTKB) : Method and Results » in M.-C. L'homme, C. Jacquemin, D. Bourigault (éds), *Recent Advances in Computational Terminology,* Amsterdam/Philadelphia.

Daille, B. (1994), *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat en informatique fondamentale, Université Paris VII.

Gale, W. A. & Church, K.W. (1993), « A Program for Aligning Sentences in Bilingual Corpora »,in *Computational Linguistics*, 19.1, pp 75-90.

Johansson, S., Ebeling, J. & Hofland, K. (1996), « Coding and aligning the English-Norwegian Parallel Corpus », in K. Aijmer, B. Altenberg (eds),

Languages in Contrast : Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994. Lund, Lund University Press, pp. 87-112.

Kenny, D. (1998), « Creatures of Habit? What Translators Usually Do with Words », in *Meta* 43.4, pp. 515-523.

Kenny, D. (1999), *Norms and Creativity : Lexis in Translated Text.* PhD thesis, Manchester, Centre for Translation Studies, Department of Language Engineering, UMIST.

Laviosa, S. (1998), « Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose »,in *Meta* 43.4, pp. 557-570.

Malmkjær, K. (1998), « Love thy Neighbour : Will Parallel Corpora Endear Linguists to Translators? », in *Meta* 43.4, pp. 534-541.

Monachini, M., Peters, C., & Picchi, E. (1993), « The PISA Tools : A Survey of Computational Tools for Corpus-based Lexicon Building », DELIS Working Paper for TR01/1-2.

Pearson, J. (1998), *Terms in Context,* Amsterdam, John Benjamins Publishing Company.

Pearson, J. (1999), « Comment accéder aux éléments définitoires dans les textes spécialisés ? », in *Terminologies Nouvelles,* 19, décembre 1998-juin 1999, pp. 21-28.