Évaluation de la fonction de recherche dans les outils d'exploitation de corpus parallèles : proposition d'une méthode d'évaluation appliquée aux outils MultiTrans Prism, ParaConc et myCAT

Rappazzo, Giovanna

Giovanna RAPPAZZO

Évaluation de la fonction de recherche dans les outils d'exploitation de corpus parallèles.

Proposition d'une méthode d'évaluation appliquée aux outils *MultiTrans Prism*, *ParaConc* et *myCAT*.

Directrice du mémoire :
Prof. Aurélie PICTON

Jurée :
Mme Donatella PULITANO

Mémoire présenté à la Faculté de traduction et d'interprétation (Département de traitement informatique multilingue) pour l'obtention de la Maîtrise universitaire en traduction, mention technologies de la traduction.

Université de Genève
Juin 2014

« E quindi uscimmo a riveder le stelle.»
DANTE ALIGHIERI (1265-1321)
La Divina Commedia - Inferno: C. XXXIV, v. 139

Remerciements

Ce mémoire n'aurait pas été possible sans l'intervention de plusieurs personnes que je tiens à remercier une à la fois.

En premier lieu, je souhaite remercier ma directrice de mémoire Mme Aurélie Picton pour son aide constante, pour sa patience et pour ses innombrables relectures et conseils, mais surtout pour la motivation qu'elle a su me transmettre.

Je tiens à remercier vivement Mme Donatella Pulitano pour ses précieux conseils et sa grande disponibilité, ainsi que Mme Marianne Starlander et Mme Lucia Morado Vazquez pour les astuces techniques concernant les mémoires de traduction.

Merci à Mme Pierrette Bouillon et à M. Jean-Pierre Sossauer pour avoir mis à disposition le matériel informatique de l'université pour l'installation de *myCAT*.

J'adresse mes remerciements à la Fondation Olanto pour m'avoir transmis une copie du logiciel *myCAT* et, en particulier, à M. Karim Benzineb pour son immense disponibilité et pour avoir pris le temps de répondre à mes questions.

Un remerciement particulier va à l'agence de traduction Transpose SA, et à son directeur, M. Tolis Cléopas, qui m'a autorisée à utiliser les corpus d'entreprise afin de tester les outils, objets de ce mémoire. Mais l'équipe de Transpose a joué un rôle bien plus important pour moi! Je ne remercierai jamais assez tous ses composants pour le support moral qui va bien au-delà d'un simple rapport professionnel ! Merci à M. Jean-François Roch pour sa flexibilité et pour m'avoir permis de disposer du bureau et du matériel informatique de l'agence à n'importe quelle heure du jour et de la nuit. Merci à M. Guillaume Ginet pour l'assistance technique et morale, à Mme Marie Dabbadie pour les astuces techniques, pour l'assistance linguistique, mais surtout pour la relecture finale ! Merci également Mme Clara Zoldan, une grande « motivatrice », mais surtout une amie qui m'a accompagnée tout au long de ce parcours.

Enfin un grand merci à ma famille pour avoir cru en moi et m'avoir permis de terminer mes études et à mes amis, proches et lointains, pour le soutien, pour les encouragements et pour m'avoir accompagnée tout au long de cette aventure.

Table des matières

Table des figures et des tableaux

Liste des abréviations

BNC : British National Corpus

KWIC : key-word in context

LISA : Localisation Industry Standard

LORIA : Laboratoire Lorrain de Recherche en Informatique et ses Applications

MT : mémoire de traduction

ND : non disponible

OSCAR : Open Standards for Container/Content Allowing Reuse

RegExp : regular expression

SRX : Segmentation Rules eXchange

TA : traduction automatique

TAO : traduction assistée par ordinateur

TMX : Translation Memory eXchange

TPC : true points of correspondance

UT : unité de traduction

Introduction

Ce mémoire est issu d'un stage que nous avons effectué au sein de l'agence de traduction Transpose SA. Au cours de cette année à l'agence, la première tâche qui nous a été confiée consistait à récupérer de l'archive et à classer tous les travaux de 2004 à 2012 afin de créer des corpus parallèles à exploiter avec des outils d'aide à la traduction, notamment une mémoire de traduction (MT) ainsi que le concordancier parallèle *myCAT*. En effet, à cette époque, Transpose venait de débuter une collaboration avec la Fondation Olanto afin d'installer le logiciel *myCAT* et de l'adapter aux besoins de l'agence.

Quelques mois plus tard, lorsqu'une bonne partie des documents avait enfin été classée et chargée sur *myCAT*, nous avons commencé à l'utiliser pendant les traductions. Bien que ce logiciel nous ait paru un excellent instrument pour le traducteur, nous nous sommes demandée s'il était possible d'apporter quelques modifications.
Par la suite, nous avons eu l'occasion de discuter à ce sujet avec M. Karim Benzineb, le concepteur de *myCAT*, qui a pris le temps de nous expliquer comment avait été conçu ce logiciel et quelle était sa politique de base : un outil intuitif permettant d'obtenir un grand nombre de données en très peu de temps.

Suite à cette rencontre, en comparant ce que nous venions d'apprendre aux enseignements de la FTI, nous nous sommes demandé si cette idée d'obtenir un grand nombre de données en peu de temps correspondait réellement aux attentes du traducteur ou si le traducteur avait besoin d'un outil un peu différent, avec d'autres fonctions et peut-être personnalisable. En même temps, nous nous sommes demandé ce que pouvait apporter concrètement un concordancier parallèle au processus de traduction et quels étaient ses atouts par rapport à d'autres outils classiques auxquels le traducteur a normalement recours, tels que les dictionnaires, les bases terminologiques, les documents écrits et les ressources en ligne.

Dans ce mémoire, nous cherchons à apporter des éléments de réponse à ces questions. En suivant la logique de la comparaison d'outils, nous confrontons deux concordanciers parallèles (*ParaConc* et *myCAT*), avec un autre outil d'exploitation de corpus parallèles : la

mémoire de traduction de *MultiTrans Prism*. Ce choix n'est pas un hasard. En effet, tout comme les concordanciers, les mémoires de traduction permettent d'analyser les corpus parallèles, mais avec une grande différence dans l'usage. Les concordanciers sont des outils de recherche dans les textes. Ils servent donc uniquement à la consultation, raison pour laquelle on s'attend à avoir des modalités de recherche avancées. Par contre, les mémoires de traduction sont des outils d'assistance à la traduction, qui permettent donc de récupérer automatiquement les segments des traductions précédentes pendant la phase de traduction. Étant donné qu'elles contiennent des bitextes et que parfois la mémoire de traduction n'est pas en mesure de récupérer automatiquement les segments si le taux de correspondance avec les segments dans la mémoire est trop bas, les MT mettent à disposition une interface de recherche qui est souvent très basique. Le but de cette comparaison est donc de voir dans la pratique comment ces modalités de recherche différentes influencent la tâche de recherche et d'affichage des résultats.

Cette étude se divise en deux parties, dont la première, plus théorique, représente un état des lieux et la deuxième, pratique, commence par un aperçu des outils d'exploitation de corpus disponibles sur le marché pour arriver ensuite à une évaluation.

Dans le chapitre 1, nous définissons et présentons les principaux types de corpus, puis nous analysons leurs apports dans le processus de traduction en focalisant notre attention sur les corpus parallèles qui sont au centre de ce mémoire.

Dans le chapitre 2, nous présentons les outils classiques d'exploitation des corpus parallèles, notamment les mémoires de traduction, les aligneurs et les concordanciers parallèles. Nous expliquons donc le fonctionnement de base des mémoires de traduction en approfondissant l'aspect de la recherche. Nous passons ensuite aux aligneurs en présentant les méthodes d'alignement classiques qui sont à la base de la plupart des outils d'exploitation des corpus disponibles sur le marché. Enfin, nous parlons des concordanciers, dont nous présentons les principales modalités de recherche et d'affichage des données.

La deuxième partie commence avec le chapitre 3, dans lequel nous faisons un tour d'horizon des outils disponibles sur le marché, puis nous choisissons les trois logiciels dont nous nous servons pour notre évaluation, c'est-à-dire la mémoire de traduction *MultiTrans Prism* et

deux concordanciers parallèles : *ParaConc* et *myCAT.* Nous présentons ensuite les fonctions principales de ces trois logiciels.

Dans le chapitre 4, nous développons notre méthode d'évaluation de la fonction de recherche dans les outils d'exploitation de corpus parallèles. Nous présentons brièvement les études qui sont à la base de notre travail, notamment le projet EAGLES (1996), la norme ISO 9126 : 1991 (1991), la théorie du modèle de tâche issue du domaine du génie du logiciel et l'étude de Höge (2002). Nous définissons ensuite nos critères d'évaluation que nous rassemblons en deux grilles d'évaluation : la grille d'évaluation des tâches qui nous permet d'évaluer les tâches et la grille d'évaluation des outils qui est utilisée pour l'évaluation des caractéristiques des outils d'exploitation de corpus qui ne peuvent pas être relevées au moment de l'accomplissement de la tâche, car elles sont orientées vers l'outil même. Nous concluons le chapitre en présentant les deux corpus dont nous nous servons pour l'évaluation des outils : le corpus CHOCOLAT et le corpus FORMATION.

Le chapitre 5 est dédié à l'exécution de l'évaluation en appliquant les critères tels que décrits dans le chapitre 4. Afin de pouvoir tester nos trois outils, nous définissons d'abord les trois tâches de recherche. Nous passons ensuite au déroulement des tâches que nous évaluons à partir des deux grilles d'évaluation définies dans le chapitre 4, c'est-à-dire la grille d'évaluation des tâches de recherche et la grille d'évaluation des outils. Pour finir, nous discutons les résultats en mettant en évidence les différences entre ces trois logiciels.

Partie I : Présentation et atout des corpus

Chapitre 1 : Les corpus dans la traduction

Dans ce chapitre, nous définissons la notion de corpus dans une perspective de traduction (section 1). Nous examinons ensuite les différents types de corpus (section 2) et nous montrons de quelle manière ils peuvent intervenir dans le processus de traduction, en nous concentrant sur les corpus parallèles qui sont au centre de ce mémoire (section 3).

1. Définition

Plusieurs auteurs ont défini la notion de corpus. Étant donné que le présent mémoire est issu du domaine de la traduction, nous nous basons sur la définition que Bowker et Pearson (2002 : 9) proposent dans le cadre de leurs études sur les langues de spécialité :

> « A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. »

En d'autres termes, un corpus est un recueil de textes authentiques, sous format électronique et sélectionnés selon des critères spécifiques.
Suivant ces auteures, nous passons brièvement en revue les quatre mots-clés de cette définition qui permettent de distinguer un corpus d'une récolte de textes : « *authentic* », « *electronic* », « *large* » et « *specific criteria* ».

1.1. Authenticité des textes

Les textes doivent être authentiques, c'est-à-dire qu'ils doivent être le résultat d'un échange d'information que les individus exercent dans la vie réelle et qu'ils ne doivent donc pas être écrits dans le seul but d'être introduits dans un corpus afin de soutenir une thèse (Bowker et Pearson, 2002 : 9). Comme l'affirme L'Homme, ils « doivent apparaître dans un environnement 'naturel' » (L'Homme, 2004 : 123). Ce critère est important car il distingue les contextes d'un terme que l'on peut trouver dans un corpus qui sont authentiques, par rapport à ceux que l'on peut trouver dans un dictionnaire qui sont créés dans le but de montrer le fonctionnement du terme.

1.2. Format

L'intérêt de disposer d'un corpus sous format électronique est que les textes peuvent être manipulés à l'aide de logiciels. Bowker et Pearson soulignent un énorme avantage des corpus électroniques, à savoir celui d'avoir un accès immédiat aux informations essentielles. En effet, lorsque l'on consulte un livre ou un article, il faut le lire du début jusqu'à la fin pour trouver les informations pertinentes (Bowker et Pearson, 2002 : 9). Avec les corpus, il est possible, par exemple, de focaliser l'attention sur ce que Meyer appelle le *knowledge-rich contexts* (en français les *contextes riches en connaissance*) et d'ignorer le reste (Meyer, 2001 repris par L'Homme 2004 : 155).

1.3. Taille

La taille d'un corpus est un concept relatif. Comme l'expliquent Bowker et Pearson (2002: 45) :

> « […] there are no hard and fast rules that can be followed to determine the ideal size of a corpus. Instead, you will have to make this decision based on factors such as the needs of your project, the availability of data and the amount of time that you have. It is very important, however, not to assume that big is always better. »

La taille d'un corpus est donc souvent un compromis entre le but de l'étude et le temps que l'on a à disposition. Un corpus ne doit pas forcément être énorme, ce qui compte c'est qu'il soit performant, voilà pourquoi il faut surtout que les textes soient diversifiés et soigneusement classés (Habert, Fabre et Issac, 1998 : 35). De plus, si l'on pense à ce que représentait le concept de taille des corpus dans les années quatre-vingt-dix, lorsque les supports informatiques étaient moins performants, moins de textes étaient disponibles sous format électronique et des techniques telles que l'océrisation n'étaient pas encore largement diffusées, on comprend pourquoi le corpus de référence *British National Corpus*[1] (BNC), qui contient 100 millions de mots, apparaissait aux chercheurs de l'époque d'une grandeur extraordinaire, alors qu'actuellement les corpus comptent souvent plusieurs centaines de millions de mots.

[1] Subventionné par le Gouvernement britannique, le BNC a été développé entre 1991 et 1994 par un consortium composé par la presse universitaire d'Oxford, les éditeurs du dictionnaire Longman et Chambers, les centres de recherche des universités d'Oxford et Lancaster et la British Library. Son but était de représenter toute la production de l'anglais britannique de la fin du XXème siècle.
http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro - Consulté le 15/01/2014

1.4. Critères de sélection

Il est important de définir attentivement les critères de sélection des textes à introduire dans le corpus. À ce propos, Bowker et Pearson (2002 : 45) expliquent :

> « [...] 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. »

Les textes doivent donc être sélectionnés en suivant des critères spécifiques qui dépendent de l'objet d'étude. Par exemple, si le but du corpus est celui d'examiner la terminologie d'une entreprise, il faut tout d'abord que tous les textes soient produits par cette dernière et, afin d'affiner la sélection, il faut établir des critères permettant d'inclure dans le corpus uniquement des textes représentatifs. Ces critères doivent ainsi permettre de collecter des « textes représentatifs du domaine dont il compte décrire la terminologie » (L'Homme, 2004 : 123).

2. Types de corpus

Dans cette section, nous présentons les principaux types de corpus, en les classant selon plusieurs critères. Précisons que ces critères peuvent être combinés entre eux et que cette subdivision est effectuée uniquement dans un but descriptif. Soulignons également qu'il existe d'autres types de corpus dont nous ne parlons pas dans ce travail. Pour approfondir le sujet, nous renvoyons, par exemple, à Sinclair (1996) ou à Habert (2001).

2.1. Corpus spécialisés et corpus de référence

La première distinction que nous pouvons faire est celle entre les corpus spécialisés et les corpus de référence.

Les corpus de référence ont pour but de représenter un échantillon de l'ensemble d'une langue. D'après Sinclair (1996 : 17) :

> « A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary […]. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text type that are selected. »

Bowker et Pearson (2002 : 12) ajoutent que les corpus de référence contiennent généralement des documents écrits ou la transcription de documents oraux ainsi qu'une grande variété de textes qui vont des débats aux articles de journaux, aux émissions radio, diversité essentielle pour assurer la « représentativité » du corpus de référence.

En ce qui concerne les corpus spécialisés, Habert, Fabre et Issac (1998 : 37) écrivent :

> « Les corpus spécialisés réunissent des données linguistiques relatives à une dimension particulière : un domaine, un thème, une situation de communication. »

En d'autres termes, les textes sont choisis selon des critères bien précis, afin de représenter une variété spécifique de la langue ou un domaine spécifique.
Ils permettent donc de cerner un sujet bien précis, comme, par exemple, le droit de succession ou la crise économique.

2.2. Corpus écrits, oraux et mixtes

Selon l'origine des textes, les corpus peuvent être subdivisés en écrits, oraux ou mixtes.
Les corpus écrits contiennent des textes originaux, produits directement sous forme manuscrite, imprimée ou électronique. C'est le cas pour la grande majorité des corpus.
Les corpus oraux, au contraire, sont issus de la langue parlée. Ils contiennent des transcriptions de production orale, tels que les conversations formelles ou informelles, des émissions radio, etc. Ils servent à étudier, par exemple, un groupe de personnes sélectionné par rapport à des paramètres spécifiques tels que le lieu de provenance, l'âge ou le niveau d'instruction.
Enfin, les corpus mixtes sont composés de transcriptions de l'oral et de textes écrits. C'est le cas par exemple de corpus de référence tels que le BNC (section 1.3).

2.3. Corpus fermés et corpus ouverts

Une autre distinction importante est celle entre les corpus fermés et les corpus ouverts.
Baker, Hardie et McEnery (2006 : 152) appellent les corpus fermés *static corpus* et les définissent ainsi :

> « A **sample text corpus** that intended to be of a particular **size**- once that target is reached, no more texts are included in it. More **corpora** are static, providing a 'snapshot' of a particular language variety at a given time. »

Les corpus fermés sont donc conçus de façon à ne pas permettre d'ajouter de nouveaux textes une fois que le corpus est terminé (Bowker et Pearson, 2002 : 13). Or, comme le souligne Sinclair (1996 : 18), le fait de limiter l'ajout de documents est une restriction qui n'est pas nécessaire. En effet, il n'existe actuellement aucune limite technique qui impose qu'un corpus doit être fermé. Au contraire, dans l'optique de créer un corpus durable, il est bien d'en permettre une mise à jour.

En partant de cette idée sont nés les corpus ouverts. Un corpus ouvert (ou *dynamic corpus* ou *monitor corpus*) est, tel que le définit Habert (2001 : 13), un corpus qui « ne cesse de croître », dans lequel il est possible d'ajouter sans cesse de nouveaux documents. Comme l'explique Sinclair, cela donne une nouvelle dimension aux corpus, c'est-à-dire la dimension diachronique (Sinclair, 1996 : 18), et permet donc d'étudier « l'évolution de certains phénomènes langagiers » (Habert, 2001 : 13).

2.4. Corpus monolingues et corpus multilingues

En fonction du nombre de langues traitées, les corpus peuvent être qualifiés de monolingues ou multilingues.

Les premiers contiennent des textes en une seule langue, tandis que les corpus multilingues contiennent des textes en plusieurs langues sélectionnés selon les mêmes critères.

2.4.1. Corpus comparables

Les corpus comparables représentent ainsi une sous-catégorie de corpus multilingues. Bowker et Pearson (2002 : 93) les définissent ainsi :

> « [...] comparable corpora consist of sets of text in different languages that are not translations of each other. We use the word 'comparable' to indicate that the text in different languages have been selected because they have some characteristics or features in common; the one and only features such as that distinguishes one set from another in a comparable corpus is the languages in which the text are written. »

En d'autres termes, il s'agit de deux ensembles de textes sélectionnés selon les mêmes critères, qui diffèrent, dans la plupart des cas, uniquement par la langue[2], mais qui ne sont pas des traductions mutuelles.

[2] Les corpus comparables sont également utilisés pour analyser la dimension diachronique dans une langue. Dans ce cas, les deux corpus diffèrent sur la base de laps de temps.

Pour que deux corpus soient comparables, il faut qu'ils soient « équivalents », c'est-à-dire qu'ils soient composés des mêmes types de textes dans la même proportion (Kilgarriff, 2010 : 1). Ils doivent de fait contenir approximativement le même nombre de textes et d'occurrences.

2.4.2. Corpus parallèles

Les corpus parallèles se composent de textes mis en correspondance avec leur traduction dans une ou plusieurs langues. Voici la définition de Sinclair (1996 : 19):

> « A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other. »

Si l'on considère le rapport de traduction entre les textes du premier et du second corpus, une autre distinction peut être faite. À ce propos, Aijmer (2008 : 276) prend en exemple un corpus parallèle anglais-suédois et explique que, si ce dernier contient des textes originaux en anglais et leur traduction en suédois, il s'agit d'un corpus unidirectionnel ; si par contre il contient aussi des traductions vers l'anglais, le corpus est bidirectionnel. McEnery et Xiao (2008 : 20) ajoutent que les corpus parallèles multilingues peuvent être aussi multidirectionnels dans le cas où le même texte est disponible en plusieurs langues, par exemple en anglais, français et allemand, comme dans le cas de textes officiels de l'UE rédigés simultanément.

Comme l'expliquent McEnery et Wilson (1996 : 58), pour qu'un corpus parallèle soit exploitable, il faut que chaque phrase (ou même chaque mot) du texte source soit mise en correspondance avec sa traduction dans le texte cible. Il faut donc que ces textes soient alignés. On parle alors de corpus alignés (L'Homme, 2004 : 131) ou de bitextes. Cet alignement peut se faire à plusieurs niveaux (principalement au niveau des mots ou des phrases). Nous en parlons dans le chapitre 2.

3. Apport des corpus parallèles à la traduction

Dans cette section, nous abordons la question de l'apport des corpus dans le processus de traduction, avec un regard particulier sur les corpus parallèles qui sont au centre de ce mémoire.

Pour commencer, nous traitons dans les grandes lignes la question de la fiabilité des corpus parallèles (section 3.1) qui est devenue incontournable en raison du grand nombre de critiques à ce sujet. Nous examinons ensuite l'apport des corpus parallèles dans le processus traduction (section 3.2) en montrant qu'ils peuvent intervenir aussi bien dans la pré-traduction (section 3.2.1) que dans la traduction (section 3.2.2) et dans la révision (section 3.2.3).

3.1. Fiabilité des corpus parallèles

Comme nous l'avons vu dans la section précédente, il existe deux types de corpus bilingues : les corpus parallèles et les corpus comparables. Dans les corpus parallèles, les textes sont mis en relation et constituent des traductions mutuelles, tandis que dans les corpus comparables tous les textes sont des créations authentiques dans les deux langues.
À ce sujet, plusieurs auteurs (souvent dans le domaine de l'extraction terminologique ou de la linguistique), parmi lesquels Déjean et Gaussier (2002), Sinclair (1996 : 20), Dipper, Seiss et Zinsmeister (2012) affirment que les corpus comparables sont plus fiables que les corpus parallèles. La plupart de ces critiques se basent sur la qualité de la langue des traductions.

D'après Dipper, Seiss et Zinsmeister (2012 : 139), le processus de traduction en lui-même a un impact sur les textes traduits. En particulier, le texte cible peut subir l'influence du texte source si la traduction est trop fidèle ou alors le traducteur peut rendre explicite ce qui était implicite dans le texte source, dans le but de rendre sa traduction plus compréhensible.
Déjean et Gaussier (2002 : 3), soulignent eux aussi que « le vocabulaire de la langue source influence le choix du traducteur ». Zweigenbaum et Habert (2006 : 22) sont du même avis et soulignent les risques de calques ou de « biais de traduction », en prenant l'exemple de « consistent » (de l'anglais, « cohérent ») traduit en français par « consistant ». Tout cela conduirait à une distorsion inévitable des traductions, comme l'affirme Sinclair (1996 : 20).

Un autre avis est celui de Baker (2000 : 33), qui affirme que les chercheurs du domaine de la traduction voient les textes traduits comme des représentations authentiques de communication. Elle continue en affirmant qu'ils sont sûrement différents par rapport aux autres textes, mais ces différences concernent le fait que le contexte du processus de production et de réception est différent. Les traductions sont liées à des contraintes sociales, culturelles, idéologiques et cognitives (Baker, 2000 : 35-36).

De son côté, Le Serrec (2012 : 56) remarque que, même si l'on veut exclure toutes traductions des corpus, il est parfois impossible de comprendre si un texte est vraiment original et que, dans certains domaines, « certaines traductions ont autant de poids sinon plus, du point de vue terminologique, que certains originaux ». En d'autres termes, certains ouvrages de référence sont des traductions et sont souvent incontournables. Elle observe également que les textes originaux ne sont pas à l'abri de fautes, car, souvent, les auteurs de textes spécialisés citent des textes écrits dans une autre langue et qu'ils risquent eux-mêmes « d'introduire dans la langue de nouvelles caractéristiques linguistiques », tout comme les traducteurs, car les autres auteurs qui ne sont pas de langue maternelle pourraient utiliser « un vocabulaire ou des tournures inhabituelles » (Le Serrec, 2012 : 57).

À travers l'analyse des équivalences terminologiques entre les corpus parallèles et les corpus comparables, l'auteure montre que la terminologie extraite du corpus comparable et celle extraite du corpus parallèle sont très similaires. Elle ajoute que, d'après l'analyse de la partie en français des corpus comparables et parallèles, on peut déduire que ce sont les auteurs des textes originaux qui introduisent le plus d'emprunts fautifs, probablement parce qu'ils lisent de nombreux ouvrages et participent à des colloques en anglais (Le Serrec, 2012 : 231). Elle conclut que les traductions sont tout aussi valables que les originaux pour étudier la terminologie d'un domaine, à condition que les critères de sélection des textes soient rigoureux (Le Serrec, 2012 : 131).

Pour notre part, nous considérons, comme le fait Baker (2000 : 33), que les traductions sont des représentations authentiques de communication et, comme l'affirme Le Serrec (2012 : 131), qu'en terminologie et en traduction, ils sont tout aussi valables que les originaux. Il ne faut pas négliger le fait que les traducteurs sont des professionnels ayant suivi une formation qui les amène à être très méticuleux dans le choix des termes et à savoir éviter les pièges tels que les calques. De plus, avec l'expérience, les traducteurs acquièrent des connaissances linguistiques et terminologiques dans leurs domaines de compétences et cela donne une valeur aux traductions.

Nous soulignons à notre tour que les critères de sélection des textes sont fondamentaux pour garantir des corpus fiables, car nous sommes conscients que de mauvaises traductions engendrent de mauvais corpus.

3.2. Apport des corpus dans le processus de traduction

Dans cette partie, nous présentons la manière dont les corpus parallèles peuvent s'intégrer dans le processus de traduction et ce qu'ils peuvent apporter de plus par rapport aux outils traditionnels dans les trois phases de la traduction. Ceci ne signifie pas pour autant renoncer aux outils traditionnels, mais les compléter.

Dans ce mémoire, nous proposons de nous concentrer sur les corpus parallèles. Cependant, il est important de souligner que plusieurs types de corpus peuvent être utilisés pour atteindre les mêmes résultats. Souvent, on les utilise de manière complémentaire. On peut par exemple faire des recherches terminologiques sur un corpus comparable et vérifier les correspondances dans un corpus parallèle et vice-versa ou alors comparer les résultats obtenus sur un corpus spécialisé avec ceux d'un corpus de référence.

3.2.1. Phase de pré-traduction

La phase de pré-traduction rassemble toutes les actions qui précèdent la traduction proprement dite. Elle concerne donc la lecture du texte source et, le cas échéant, la recherche d'informations sur le sujet et la recherche terminologique. Nous montrons ici l'apport des corpus parallèles au niveau conceptuel (3.2.1.1), terminologique (3.2.1.2) et stylistique (3.2.1.3).

3.2.1.1. Niveau conceptuel

Avant de commencer une traduction, surtout s'il s'agit d'un domaine spécialisé, il est nécessaire de se documenter. La méthode classique impose de consulter des livres, des articles ou des documents sur internet et de les lire quasi entièrement avant de trouver l'information recherchée. Cela peut prendre du temps, sans compter que parfois les documents sont peu fiables ou difficiles d'accès. Comme le souligne Aston (2000 : 22), les corpus peuvent aider au niveau conceptuel pour se familiariser avec le sujet et avec les concepts récurrents et pour avoir des schémas de référence plus vastes et en accord avec le domaine en question. Ils permettent entre autres de vérifier jusqu'à quel point ces concepts ou schémas de référence sont communs aux deux cultures.

À ce propos, Munday (1998 : 7) explique comment les outils d'exploitation de corpus, même très basiques (comme la liste de mots dont nous parlons dans le chapitre 2, section 3.4.1), peuvent donner une idée sur le contenu du texte et fournir des points d'accès aux concepts

clés que l'on peut ensuite examiner avec les concordances et voir dans leur contexte d'apparition. Ces contextes peuvent, par ailleurs, contenir des définitions, des explications ou des descriptions d'un concept (Bowker et Pearson 2002 : 38).

3.2.1.2. Niveau terminologique

Les corpus permettent d'obtenir des informations sur les caractéristiques des langues de spécialité, notamment en ce qui concerne la terminologie.

Il existe plusieurs manières de repérer les termes importants. Par exemple, Aston (2000 : 23) propose d'extraire des listes d'unités que l'on peut ensuite analyser et comparer avec d'autres listes (par exemple avec celles issues des corpus de référence) afin de voir quels éléments sont typiques d'un corpus. Aston (ibidem) continue en expliquant que, en partant de ces listes et en approfondissant leur analyse, il est possible d'obtenir des informations concernant la distribution et les variantes des termes récurrents, ainsi que de voir les termes qui leur sont associés.

De plus, les corpus permettent d'extraire la terminologie propre à une entreprise afin de respecter les besoins du client. En effet, dans certains secteurs, comme par exemple le secteur bancaire, il arrive de devoir traduire un document du même type (par exemple un contrat type de gestion de fortune) pour des banques différentes. Bien que les deux textes puissent présenter des parties identiques, les deux clients pourraient utiliser une terminologie différente. De ce fait, une base terminologique générique est inutile et il est nécessaire de faire référence aux documents du client pour extraire la terminologie correcte. Une analyse sur corpus de ces documents facilite la tâche de recherche terminologique.

3.2.1.3. Niveau stylistique

Dans les langues de spécialité, il est impératif de respecter le style et le registre propres à la langue et au domaine traités. En effet, dans chaque langue, il peut y avoir des constructions grammaticales ou des collocations qui n'existent pas dans l'autre ou qui diffèrent totalement.

Aston (2000 : 22) explique que les corpus permettent d'améliorer les connaissances stylistiques, lexicales et grammaticales des textes d'un domaine spécifique et dans les deux cultures, notamment en ce qui concerne la structure du texte et les formes lexico-

grammaticales. De plus, les corpus bilingues permettent une comparaison entre plusieurs langues et plusieurs cultures (Aston 2000 : 23).

Bowker et Pearson (2002 : 20) ajoutent que les corpus peuvent aider à vérifier et améliorer le style d'un texte. Bowker et Pearson (2002 : 36) prennent l'exemple des textes juridiques. Le style est certainement différent par rapport à celui de la langue générale, car il y a des constructions syntaxiques spécifiques à ce genre de textes (par exemple en anglais l'usage du passif, de phrases longues, de conditionnels ou de négations), c'est pourquoi il est utile de vérifier certains éléments dans un corpus afin de choisir des solutions qui respectent les caractéristiques du domaine et de la langue.

3.2.2. Phase de traduction

Une fois que le traducteur dispose de toutes les informations concernant le contenu, la terminologie et le style du texte, il passe à la traduction. Dans cette phase, il peut avoir recours à des outils, parmi lesquels les concordanciers parallèles. Nous présentons, dans cette partie, la manière dont on peut les utiliser.

3.2.2.1. Comparaison du texte avec les traductions précédentes

Lorsque l'on traduit souvent pour le même client ou qu'il faut mettre à jour une traduction précédente, il est possible de consulter les bitextes pour vérifier si le texte ou une partie du texte a déjà été traduite précédemment. Cela permet de gagner du temps et de garantir la cohérence avec les traductions précédentes.

3.2.2.2. Complément aux outils de référence

Lorsque le traducteur se trouve face à un concept ou un terme difficile à rendre dans la langue cible, il a recours à des outils de référence classiques tels que des dictionnaires monolingues et bilingues, des encyclopédies, des bases terminologiques en ligne ou privées, des sites internet. Or, comme l'affirme Isabelle (1992 : 726), ces outils demeurent incomplets, car parfois ils ne contiennent pas la solution recherchée. La raison est très simple : les dictionnaires spécialisés valident les synonymes, mais, comme l'explique Le Serrec (2012 : 227) :

> « énumérer tous les équivalents non typiques d'un terme serait une entreprise trop difficile à réaliser, car il faudrait connaître d'avance tous les contextes possibles et imaginables dans lesquels ces notions peuvent être utilisées. »

C'est justement dans ce sens que les corpus peuvent intervenir, en montrant le terme dans son contexte. Isabelle (1992 : 726) remarque que:

> « La masse de traduction produite chaque année contient infiniment plus de solutions à plus de problèmes que les outils de référence existants et imaginables ! »

Künzli (2001 : 501) partage cet avis et, en se référant aux bitextes, il argumente ainsi :

> « Dans la pratique, les textes parallèles comme source de documentation sont souvent plus précieux que n'importe quel dictionnaire, car ils offrent des équivalents en langue cible dépassant le niveau du mot et permettent la réalisation de traductions adaptées aux besoins du client. »

À ce propos, Bowker et Pearson (2002 : 194) font l'exemple du terme « *plaignant* » dans le contexte de la traduction d'un amendement d'une convention collective de travail au Canada. Le dictionnaire propose trois équivalents anglais : « *litigant* », « *plaintiff »* et « *complainant* » en indiquant que les trois sont utilisés dans le domaine juridique, mais sans en donner le contexte. En consultant un corpus parallèle anglais-français qui réunit des conventions collectives de travail pour les employés d'une institution bilingue du Canada, on peut voir qu'aucun des trois termes n'est utilisé dans ce cas, car on utilise « grievor ».

Aston (2000 : 25) affirme que l'on peut consulter les corpus parallèles afin de découvrir si et comment une certaine expression du texte de départ a été traduite précédemment et lequel parmi les équivalents possibles a été effectivement utilisé, c'est-à-dire quelles sont les variantes admises et quelle est la variante homologuée par l'usage (Zweigenbaum et Habert, 2006 : 22). De plus, ils permettent de trouver des correspondances pour les néologismes. Ils peuvent donc aider le traducteur à accomplir sa tâche avec plus de précision, dans le respect de la terminologie et de la phraséologie (McEnrey et Xiao, 2008 : 26).

3.2.2.3. Collocations

Afin de trouver les collocations d'un terme, on peut avoir recours aux dictionnaires, toutefois, comme nous l'expliquons à la section précédente, parfois les contextes d'exemples ne sont pas pertinents avec l'usage du terme dans le texte source. Le traducteur peut alors consulter un corpus parallèle.

Pour clarifier l'importance de la recherche de collocation, nous reprenons l'exemple de Bowker et Pearson (2002 : 194-195), qui analysent le terme français « grief » et son

correspondant en anglais « *grievance* » dans le contexte des contrats collectifs de travail. Pour découvrir les collocations de ces deux termes et leurs emplois dans la phrase, elles consultent un corpus juridique anglais-français. Elles observent, ainsi, qu'en anglais « *grievance »* est associé principalement à « *initiated »* (cinq occurrences) et parfois à « *filed* » (une occurrence), alors qu'en français on retrouve les verbes « amorcer » (une occurrence), « soumission » (une occurrence), « présenter » (trois occurrences) et « déposer » (une occurrence). Cet exemple montre donc que « grief » et « *grievance* » ne se construisent pas de la même manière dans les deux langues.

3.2.2.4. Équivalences d'expressions idiomatiques

Comme l'explique Isabelle (1992 : 728), les corpus parallèles aident le traducteur lorsqu'il se trouve face à une expression idiomatique qui n'est pas toujours répertoriée dans les dictionnaires bilingues en lui fournissant très rapidement des solutions satisfaisantes et auxquelles il n'avait pas pensé.

3.2.2.5. Analyse des choix de traduction

Laviosa (2003 : 107) explique comment un certain type d'analyse des corpus permet de comprendre le processus de décision du traducteur et les normes de traduction qu'il utilise. Dans ce sens, cette analyse peut offrir une stratégie de traduction systématique des structures linguistiques qui n'ont pas d'équivalent direct dans la langue cible (McEnrey et Xiao, 2008 : 26).

De manière générale, on peut utiliser les corpus parallèles pour vérifier comment un mot ou une phrase ont été traduits et utilisés dans une autre langue (Bowker et Pearson, 2002 : 94), et pour observer les relations d'équivalences lexicales ou des structures grammaticales dans les textes sources et cibles (Hansen-Schirra et Teich, 2009 : 1161).

Bowker et Pearson (2002 : 94) expliquent qu'ils permettent d'observer ce qu'il se passe dans la traduction, comment une information est transférée, ou si des informations ont été perdues, adaptées ou altérées dans le processus de traduction.

3.2.3. Phase de révision

Les corpus peuvent intervenir également dans la phase de révision, c'est-à-dire pendant la relecture et la correction de la traduction. Aston (2000 : 26-27) suggère de les utiliser afin de

vérifier la structure interne et la cohérence de la traduction. Une fois la traduction terminée, il conseille d'ajouter les deux textes alignés au corpus parallèle personnel du traducteur afin de pouvoir exploiter les outils d'analyse des corpus et vérifier, par exemple, si toutes les expressions du texte source correspondent à une même traduction dans le texte cible.

Sans vouloir forcément aller si loin, les corpus peuvent être utilisés pour la simple consultation afin de fournir des réponses à des doutes en phase de révision. Ils permettent ainsi de vérifier l'harmonisation du texte par rapport aux traductions précédentes, c'est-à-dire de vérifier si le traducteur a utilisé la terminologie officielle requise par le client et éventuellement de voir s'il existe des variantes terminologiques homologuées par l'usage et si le style a été respecté.

Habert, Nazarenko et Salem (1997 : 137) affirment que les textes alignés fournissent un « appui critique à la traduction » qui peut consister à vérifier qu'il n'y a pas d'omissions dans la traduction ou que le traducteur n'a pas utilisé de faux-amis. Il permet également de vérifier si une solution de traduction est adéquate.

Chapitre 2 : Outils classiques d'exploitation des corpus

Dans la première partie de ce mémoire, nous avons défini la notion de corpus et la manière dont ils peuvent aider le traducteur dans les différentes tâches de traduction.

Nous présentons maintenant les outils qui permettent d'exploiter les corpus, et plus spécifiquement les corpus parallèles : les aligneurs, les mémoires de traduction et les concordanciers parallèles.

Afin de faciliter la description de leurs modes de fonctionnement, ces trois outils font partie de sections distinctes dans ce chapitre. Il est toutefois utile de rappeler que, bien qu'il existe des outils qui servent exclusivement à l'alignement, les aligneurs sont souvent intégrés à d'autres outils d'aide à la traduction, comme certains concordanciers ou les mémoires de traduction.

Pour chaque type d'outils, nous examinons son fonctionnement de base et décrivons ses fonctionnalités principales.

1. Mémoires de traduction

Les mémoires de traduction (MT) représentent notre premier type d'outils pour exploiter les corpus parallèles. Ces outils sont devenus presque indispensables pour les traducteurs et le sont encore plus dans les agences de traduction, car ils permettent de gérer en peu de temps de gros projets avec plusieurs traducteurs en assurant une cohérence stylistique et terminologique (Webb, 1992 : 19). Comme l'expliquent Astermühl (2001 : 139), Quah (2006 : 94), Bowker (2002 : 112-114) et L'Homme (2008 :189), les MT sont également utiles lorsque les textes contiennent de nombreuses répétitions internes de mots, de phrases ou même de paragraphes entiers, ainsi que pour les textes des mêmes clients ou du même domaine, pour les traductions dans les domaines spécialisés tels que les textes juridiques, les documents techniques et les manuels. De plus, les MT sont très utiles lorsqu'il faut mettre à jour d'anciennes traductions, car elles permettent de récupérer tous les passages identiques, ce qui garantit la cohérence avec l'ancienne version, ainsi qu'un énorme gain de temps pour le traducteur.

Les MT sont généralement intégrées dans des outils de traduction assistée par ordinateur (TAO) qui se composent des parties suivantes :

- un aligneur qui permet de segmenter et d'aligner les textes à introduire dans la base de données ;
- un outil pour la création et la gestion de bases terminologiques ;
- un éditeur pour le texte cible qui peut être intégré dans la MT ou dans un éditeur de texte (par exemple dans Word) ;
- des outils de gestion de projets permettant de gérer plusieurs traducteurs en même temps et de planifier des deadlines ;
- d'autres outils tels que les dictionnaires, les correcteurs d'orthographe, etc.

Comme tous les outils de TAO, les MT sont toujours en évolution. On parle actuellement de seconde génération de MT qui se distinguent des mémoires de traduction classiques par des techniques différentes d'archivage des textes (comme dans le cas de *MultiTrans Prism* qui archive des textes en entier) ou par de nouvelles techniques de récupération des segments (comme par exemple *Similis*[3] qui utilise la technologie des *chunks*, permettant de découper les phrases en groupes terminologiques intelligents et donc de récupérer des sous-segments). Dans cette partie, nous nous limiterons à décrire le fonctionnement de base des mémoires de traduction qui est très semblable dans tous les outils, tout en sachant qu'il existe des techniques différentes qui dépassent le cadre de notre travail.

1.1. Définition et fonctionnement de base

Le principe de base d'une mémoire de traduction est celui de « recycler » les traductions précédentes pour ne plus devoir traduire deux fois le même segment. Dans ce but, les anciennes traductions sont stockées dans une base de données et récupérées au moment de la traduction.

Voici la définition de mémoire de traduction que l'on trouve dans le rapport final du projet EAGLES (1996):

> « a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions. »

[3] Similis : http://similis.fr/ - Consulté le 27/04/2014

Il s'agit donc d'une base de données constituée à l'aide des traductions précédentes qui sont alignées avec les textes sources correspondants, subdivisés en segments. Chacun de ces segments est appelé unité de traduction (Astermühl, 2001 : 135). Comme le précise Bowker (2002 : 92), la structure de cette base de données peut être considérée comme un corpus parallèle ou un bitexte.

Lorsqu'on introduit un nouveau texte à traduire, ce dernier est tout d'abord segmenté en phrases. Puis, à l'aide d'un algorithme, le système cherche les correspondances entre les nouvelles phrases à traduire et celles contenues dans la base de données. S'il y a des correspondances parfaites ou partielles, celles-ci sont affichées et le traducteur peut décider de les utiliser ou de les rejeter. De plus, s'il y a des répétitions au fil du texte, il suffit de traduire la phrase une première fois pour qu'elle s'affiche ensuite automatiquement dans le reste du texte. C'est ce que l'on appelle la propagation (L'Homme, 2004 : 180). De cette manière, le traducteur travaille uniquement sur les nouvelles phrases. Une fois la traduction terminée, les nouveaux segments sont ajoutés à la base de données existante. Ceci signifie donc que la MT s'agrandit au fur et à mesure que l'on traduit.

1.2. Segmentation et alignement

La segmentation est une phase très importante pour le bon fonctionnement des MT car elle détermine les unités de traduction (UT). Plus cette segmentation est précise, plus la MT sera performante.

Quah (2006 : 100) définit ainsi la segmentation :

> « Segmentation is the process of breaking a text up into units consisting of a word or a string of word that is linguistically acceptable. Segmentation is needed in order for a translation memory to perform the matching (perfect and fuzzy) process. A pair of old source and target-language text is usually segmented into individual pairs of sentences. »

En d'autres termes, la segmentation est la subdivision du texte en unités de traduction. Ces unités de traduction sont généralement constituées par des phrases (Quah, ibidem), mais, comme nous le rappellent entre autres Quah (ibid.), Austermühl (2001 : 135) et Bowker (2002 : 94), il existe des UT qui ne sont pas des phrases, comme par exemple les cellules d'un tableau ou les listes d'éléments. Le système identifie les UT grâce aux retours à la ligne ou à la ponctuation. Toutefois, ces critères ne sont parfois pas suffisants. Bowker (2002 : 92)

explique que le point est un marqueur fort de fin de phrase, mais qu'il est également utilisé dans les abréviations. Cela peut créer des ambiguïtés dans la segmentation. Voilà pourquoi la plupart des MT proposent une liste des abréviations qui contient par défaut les abréviations les plus courantes dans chaque langue et que l'utilisateur peut personnaliser.

L'alignement a lieu grâce à un aligneur qui est généralement intégré à la MT. Nous analyserons son fonctionnement dans la section 2.

1.3. Stockage des segments

Dans les MT de première génération, les UT sont stockées de manière indépendante, c'est-à-dire que chaque segment est séparé de son contexte d'origine. Il est donc impossible de voir le texte dans son format d'origine.

Dans certaines MT de seconde génération, les UT sont stockés dans leur contexte, c'est-à-dire que la mémoire contient des bitextes alignés. De cette manière, il est possible de remonter aux textes de départ.

1.4. Types de correspondances

Comme nous l'expliquons dans la section 1.1, pendant la phase de traduction, le système récupère dans la base de données les unités de traduction correspondantes. Ces correspondances peuvent être parfaites ou partielles.

On parle de correspondances parfaites (*exact matches*) lorsque le segment récupéré correspond à 100 % avec le segment à traduire, aussi bien du point de vue linguistique que pour son formatage (Bowker, 2002 : 96). Aucun changement n'est nécessaire avant d'introduire ce segment dans la nouvelle traduction.

Lorsque deux segments ne se distinguent que par un seul élément que Bowker (2002 : 98) et Austermühl (2001 : 137) appellent « variable » (par exemple le format de la date, l'heure, une mesure ou un nombre) ou pour la police, ils ne sont plus considérés comme correspondants à 100 %. Austermühl et Bowker les appellent *full matches,* d'autres auteurs (Somers, 2003 : 37-39 et Quah, 2006 : 96) les considèrent comme des correspondances partielles.

Comme l'explique Bowker (2002 : 99), on a des correspondances partielles (*fuzzy matches*), lorsque les segments sources récupérés sont semblables mais non identiques au nouveau segment à traduire. Afin de déterminer le degré de correspondance des segments sources trouvés dans la base de données par rapport au segment source du nouveau texte à traduire, un pourcentage est fixé grâce à un algorithme qui se base sur la comparaison des chaînes de caractères dans les deux phrases (Quah, 2006 : 96 et L'Homme, 2004 : 177). Les segments avec le pourcentage le plus élevé s'affichent en premier. Pour faciliter la tâche du traducteur, la plupart des logiciels mettent en évidence les parties qui diffèrent à travers un jeu de couleurs.

Les correspondances partielles vont de 1 à 99 %, mais généralement l'utilisateur fixe un taux entre 60 % et 70 % (Bowker, 2002 : 100) afin d'avoir un équilibre entre bruit et silence[4].

1.5. Modalités de recherche

La MT dispose d'une interface de recherche pour la consultation en tant que bitexte (L'Homme, 2004 : 179) à travers des requêtes ponctuelles. L'utilisateur peut donc lancer une recherche et la MT affiche les segments contenant la séquence recherchée ainsi que les segments traduits correspondants.

C'est de cet aspect des mémoires de traduction que nous nous occupons dans la partie II de ce mémoire. En partant des mêmes corpus, nous montrons en quoi diffère la recherche dans un concordancier parallèle par rapport à celle effectuée dans une mémoire de traduction.

Nous décrivons ici les principaux types de recherche disponibles dans les MT en les subdivisant par recherche simple, recherche avancée et recherche par fuzzy match.

1.5.1. Recherche simple

La recherche de base pour toutes les MT est celle par mot simple. Ce type de recherche permet de retrouver tous les segments contenant le mot exact. Certains logiciels admettent des jokers[5] qui se limitent souvent à l'astérisque.

[4] La notion de bruit indique, dans ce cas, les segments non pertinents mais récupérés par le logiciel, tandis que, la notion de silence indique les segments pertinents mais omis par le logiciel.

[5] Les jokers sont des caractères spéciaux permettant de remplacer un ou plusieurs caractères dans une séquence de recherche. L'astérisque remplace généralement un nombre quelconque de caractère et peut se trouver partout dans la chaîne de caractères.

1.5.2. Recherche avancée

Certaines mémoires de traduction proposent des fonctions de recherche avancées. Il y en a principalement deux : la recherche par troncature et la recherche par filtres. La première permet de retrouver des mots à travers leurs radicaux ou à travers une chaîne de caractères. Cette fonction est utile lorsqu'il faut chercher les formes fléchies d'un mot ou des mots appartenant à la même famille sémantique.

La recherche par filtre permet de trier les résultats à l'aide de critères qui diffèrent d'une MT à l'autre. Par exemple, on peut trier les résultats par rapport au domaine, au client, à la date de création, à la date de la dernière modification, à l'utilisateur qui a créé le segment ou qui l'a modifié, au nombre de balise. On peut également choisir de lancer la recherche sur les segments sources ou cibles.

1.5.3. Recherche par « fuzzy match »

En choisissant cette option, il est possible de chercher un groupe de mots. On peut choisir d'afficher uniquement les segments contenants tous les mots dans le même ordre, dans ce cas le logiciel récupère uniquement les *exact matches*, ou d'afficher tous les segments contenant les mots dans n'importe quel ordre, dans ce cas la MT affiche les *fuzzy matches*. Enfin, il est possible de récupérer les segments qui contiennent au moins un certain nombre de mots (par exemple 2).

1.6. Importation et exportation de MT

Vers la fin des années 1990, plusieurs mémoires de traduction et outils d'aide à la traduction ont été développés. Ces outils n'étaient pas compatibles entre eux à cause du fait que chaque développeur utilisait un format différent pour stocker les fichiers et les bases de données dans les mémoires de traduction. Par conséquent, les mémoires de traduction ne pouvaient pas être importées et exportées sur tous les outils.

Comme l'explique Esselink (1998 : 136), en 1997, la Localisation Industry Standard Association[6] (LISA) a créé un groupe formé par des fournisseurs de services informatiques et d'outils, des vendeurs et des clients appelé OSCAR (Open Standards for Container/Content

[6] La société LISA a fermé ses portes en février 2011. À sa place, l'European Telecommunications Standard Institute (ETSI) a fondé la Localization Industry Standards (LIS) Industry Specification Group (ISG) afin de développer des spécifications basées sur XML pour l'échange d'informations dans le domaine de la TAO, y compris pour TMX (Translation Memory eXchange) et pour SRX (Segmentation Rules eXchange).

Allowing Reuse) qui a mis au point le format Translation Memory eXchange (TMX). Le but était d'uniformiser le format des mémoires de traduction afin de permettre d'importer et d'exporter les bases de données sur tous les outils de TAO et de faciliter l'échange de MT. Cela a permis également aux traducteurs de réutiliser leurs propres mémoires de traduction en cas de changement d'outil ou d'entreprise.

Quah (2006 : 120) explique la manière dont le format TMX facilite l'échange de données entre deux bases de données qui supportent des formats de fichiers différents (figure 2.1). Un texte stocké en format Word dans l'outil A est exporté à l'aide du format TMX et importé dans l'outil B qui supporte le format HTML.

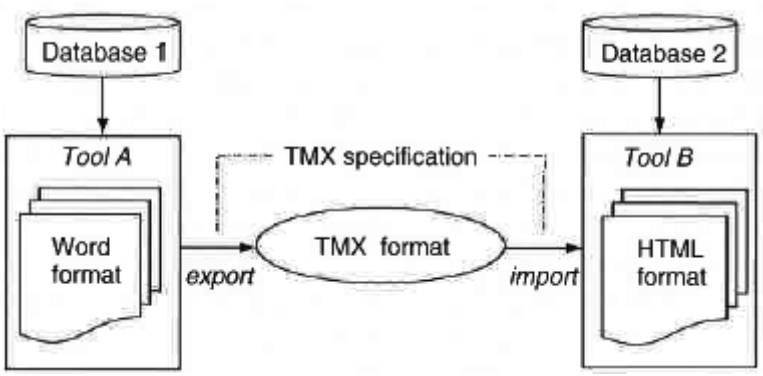


Figure 2.1 : Exemple d'échange de données avec TMX (« Exemple of the TMX data-sharing » : Quah, 2006 : 120)

Pour de plus amples informations concernant les spécifications techniques du format TMX, il est possible de consulter la dernière version du document disponible sur le lien: http://www.gala-global.org/oscarStandards/tmx/tmx14b.html#refLISA [7].

D'autres formats sont également utilisés pour l'importation et l'exportation des MT, parmi lesquels XLIFF, RTF, TDA et XLS.

2. Aligneurs

Les aligneurs sont des outils fondamentaux pour les corpus parallèles, car ils permettent de segmenter les textes et de mettre en relation les segments des textes sources avec leurs correspondants dans les textes cibles. Cette phase est essentielle, car elle prépare les corpus pour l'exploitation avec d'autres outils.

[7] Consulté le 23/05/2014

Les aligneurs se distinguent entre eux par les méthodes qu'ils utilisent afin d'accomplir leur tâche. Comme nous l'expliquons à la section 2.2, il existe différentes techniques d'alignement basées sur des calculs purement statistiques et d'autres qui intègrent des facteurs linguistiques.

Les nouvelles techniques cherchent à surmonter les problèmes communs aux méthodes classiques. La tendance est celle de mélanger les techniques et d'intégrer des éléments linguistiques aux méthodes statistiques ou alors de développer un algorithme capable de mélanger la statistique à la traduction automatique.

L'alignement n'étant pas le sujet principal de ce mémoire, il est impossible de décrire ici toutes les techniques existantes. C'est pourquoi nous nous limitons aux méthodes classiques qui sont à la base de nombreuses nouvelles recherches et que nous retrouvons dans la plupart des aligneurs disponibles sur le marché.

2.1. Définition

Kraif (2002 : 273) définit ainsi l'alignement:

> « Aligning consists in finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents »

Il s'agit donc de mettre en correspondance les segments des textes sources avec ceux des textes cibles afin que le n-ième segment du texte source et le n-ième segment du texte cible soient une traduction mutuelle l'un de l'autre (Kraif, 2002 : 275).

Afin de s'assurer que ces conditions soient effectivement respectées, Bowker et Pearson (2002 : 96) conseillent de prétraiter les documents pour vérifier si les textes sources et cibles contiennent le même nombre de paragraphes et de phrases.

Une fois que cette tâche est accomplie, il est possible de charger les documents dans le logiciel afin de procéder à l'alignement des textes.

2.2. Méthodes d'alignement automatique

Il existe principalement deux méthodes d'alignement des textes parallèles : l'alignement par phrase et celui par mot. L'alignement par phrase reste pour l'instant la méthode la plus exploitée, car elle permet d'obtenir des résultats satisfaisants. En effet, déjà en 2000, dans le

cadre du projet d'évaluation des systèmes d'alignement de textes parallèles ARCADE, Véronis et Langlais (2000 : 369) estimaient que les systèmes pouvaient atteindre un taux de réussite de 98.5 % sur les « textes normaux », c'est-à-dire des textes qui ne contiennent pas d'adjonctions et d'omissions, et dans lesquels l'ordre des paragraphes est respecté. Toutefois, la grande limite de cette étude était représentée par le fait qu'un seul couple de langue avait été testé (anglais-français). Afin d'approfondir ces résultats, un nouveau projet a été conduit : le projet ARCADE II de 2006. Ce dernier a en partie confirmé les résultats de l'étude précédente. ARCADE II a pris en compte 10 langues subdivisées en langues d'Europe de l'Ouest (anglais, allemand, italien et espagnol) et langues distantes du français (arabe, chinois, grec, japonais, persan et russe). Chaque couple a comme langue pivot le français (Chiao & all, 2006 : 1975). Pour le premier groupe, les résultats d'alignement sont confirmés, avec 99 %[8] sur les corpus segmentés (ibidem : 1977). Par contre, les résultats pour les langues distantes du français sont plus modestes : ils atteignent en moyenne 87 %, avec 97 % pour le grec mais seulement 42 % pour les langues avec écriture non latine.

En tenant compte de ces résultats, on comprend bien pourquoi les recherches de ces dernières années se concentrent surtout sur l'alignement de langues mineures, comme par exemple le corpus anglais-croate CorAl (Seljan & all, 2010) ou de langues distantes entre elles, comme par exemple le corpus multilingues MultiUN (Chen et Eisele, 2012) qui propose l'accès à des documents alignés dans cinq des langues officielles des Nations unies (anglais, français espagnol, russe et chinois).

Nous présentons dans la section suivante les méthodes principales d'alignement par phrase, en particulier l'ancrage lexical (section 2.2.1.2), l'alignement par longueur de phrases (section 2.2.1.3) et les cognats (section 2.2.1.4). Nous passons ensuite aux méthodes d'alignement par mot, notamment aux cooccurrences parallèles (section 2.2.2.1) et aux catégories grammaticales (section 2.2.2.2).

[8] L'unité de mesure utilisée dans cette étude est en réalité la F-mesure. Les résultats sont compris entre 0 et 1, les meilleurs résultats devant se rapprocher de 1. Afin de faciliter la compréhension des résultats, nous avons converti les valeurs en pourcentage. Pour la définition de F-mesure, nous renvoyons le lecteur à Veronis et Langlais (2000 : 376).

2.2.1. Alignement par phrase

Comme nous l'expliquons dans la section précédente, il existe de nombreuses méthodes d'alignement par phrase. Cependant, deux d'entre elles sont considérées à la base des techniques d'alignement développées par la suite : d'une part nous avons les études de Kay et Röscheisen (1988, 1993), fondées sur l'ancrage lexical, de l'autre les études de Brown, Lai et Mercer (1991) et de Gale et Church (1993), basées sur la longueur des phrases.

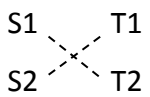
2.2.1.1. Prérequis des systèmes d'alignement par phrase

Comme l'explique Véronis (2000 : 160), la plupart des méthodes d'alignement fonctionnent uniquement si certaines conditions sont respectées, c'est-à-dire si :

> « - l'ordre des phrases dans les deux textes est identique ou très proche ;
> - les textes contiennent peu de suppressions ou d'adjonctions ;
> - les alignements 1:1 sont très largement prépondérants et que les rares alignements m:n sont limités à de petites valeurs de m et n (typiquement 2). »

Une correspondance 1:1 a lieu lorsqu'un segment du texte source correspond à un segment dans le texte cible. Lorsqu'on parle d'alignement m:n, on indique avec *m* un nombre de segments du texte source et avec *n* un nombre de segments du texte cible. Par exemple, dans le cas d'une correspondance 1:0, le segment de la langue source ne correspond à aucun segment dans la langue cible, donc il n'a pas été traduit ; la correspondance 2:1 signifie que deux segments de la langue source ont été traduits par un seul segment dans la langue cible. Ceci est le cas lorsque le traducteur lie deux phrases. En retournant à la définition de Véronis, les alignements admis sont donc 0:1, 1:0, 1:1, 1:2, 2:1 et 2:2.

Isabelle et Warwick-Armstrong (1993 : 294) ajoutent que les correspondances croisées sont refusées. Les correspondances croisées mettent en relation par exemple le segment 1 de la langue source avec le segment 2 de la langue cible ainsi que le segment 2 de la de la langue source avec le segment 1 de la langue cible c'est-à-dire, pour reprendre l'exemple de Warwick-Armstrong (ibidem) :



Cependant, le fait de devoir respecter l'ordre des segments et donc de refuser les correspondances croisées représente une grosse limite des systèmes d'alignement dans le

cas, par exemple, de listes de mots par ordre alphabétique. En passant d'une langue à l'autre, l'ordre des termes change. Comme les correspondances croisées ne sont pas admises, l'alignement automatique échoue et il faut corriger les erreurs manuellement.
C'est pour surmonter ce problème que l'on est à la recherche d'autres techniques orientées, par exemple, sur le développement d'algorithmes capables de mélanger la statistique à la traduction automatique. Nous en reparlons au chapitre 3.
En partant de ces observations, nous allons maintenant voir les différentes méthodes d'alignement par phrase.

2.2.1.2. Ancrage lexical

L'ancrage lexical représente la première méthode d'alignement automatique. Elle a été mise au point à partir de 1988 par Kay et Röscheisen (Kraif, 2006 : 3). Il s'agit d'une technique basée exclusivement sur la distribution des mots dans le texte, donc sur des informations internes. Kay et Röscheisen (1993 : 122) observent qu'un couple de phrases contenant un couple de mots alignés sont elles aussi alignées. Par conséquent, un alignement partiel au niveau des mots peut conduire à un alignement très complet au niveau des phrases (Kay et Röscheisen, 1993 : 129).

L'alignement se déroule en deux phases. Pour commencer, les phrases des textes sources et cibles sont placées sur un diagramme cartésien. Le système fait donc un premier alignement en mettant en correspondance les phrases par rapport à la distance dans le texte. Le système exclut d'abord les phrases qui ont une position trop éloignée et qui ont donc peu de probabilités d'être alignées et procède, par exclusion, à un premier alignement partiel. Sur la base de ce premier alignement, le système crée une liste de mots dans les deux textes, en se basant sur la similarité et sur la fréquence dans les textes respectifs, et les aligne en comparant leur distribution dans le texte. La distribution d'un couple de mots est semblable si la plupart des phrases dans lesquelles le premier terme apparaît peuvent être alignées avec des phrases dans lesquelles le deuxième apparaît et vice-versa. Ces couples de mots représentent donc des points d'ancrage, c'est-à-dire des points de repère que le système utilise pour réaligner les phrases.

2.2.1.3. Longueur des phrases

L'alignement basé sur la longueur des phrases a été mis en place par Gale et Church (1993). Il s'agit d'une méthode purement statistique. L'idée est que les phrases longues dans le texte source ont tendance à être traduites par des phrases longues dans le texte cible et que les phrases courtes seront traduites par des phrases courtes. Ils ajoutent qu'il existe une relation constante entre la longueur des phrases en nombre de caractères d'une langue à l'autre. À chaque correspondance proposée, on assigne un score de probabilité basé sur le rapport de longueur de la phrase en caractères et la variation de ce rapport (Gale & Church, 1993 : 79). Ces scores sont ensuite utilisés dans un algorithme qui permet de mettre en correspondance les segments sources et cibles qui ont la probabilité la plus élevée de correspondre, en tenant compte des correspondances possibles décrites à la section 2.2.1.1.

L'étude de Brown, Lai et Mercer (1991), contemporaine à celle de Gale et Church, utilise la même approche, mais se fonde sur le nombre de mots plutôt que sur le nombre de caractères.

2.2.1.4. Croisement des deux systèmes : les cognats

Le concept de « cognat » a été introduit en 1992 par Simard, Foster et Isabelle et repris par Church (1993), Johansson (1996) et McEnery (1995).

Simard, Foster et Isabelle (1992), en reprenant le modèle de Gale et Church, développent un algorithme qui combine les critères de longueur des phrases à celui d'ancrage lexical, en utilisant les cognats.

Simard, Foster et Isabelle partent d'une critique faite au modèle de Gale et Church. D'après eux, dès que deux paragraphes ne sont pas composés par le même nombre de phrases ou qu'ils contiennent un grand nombre de suppressions ou d'adjonctions, le système risque d'échouer l'alignement et n'est pas en mesure de rattraper l'erreur, ce qui fausse tout le reste de l'alignement. Comme solution à ce problème, les auteurs proposent d'ajouter des cognats qu'ils définissent ainsi :

> « Informally speaking, cognates are pairs of tokens of different languages which share "obvious" phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pairs *generation/génération* and *error/erreur* constitute typical examples for English and French. One might want to extend the notion so as to include such things as proper nouns (*Paris*; *London* and

Londres), numerical expressions and even punctuation (question marks, parentheses, etc.). » (Simard, Foster & Isabelle 1992 : 4)

Les auteurs précisent ensuite que les cognats peuvent être de trois types : un couple de caractères alphanumériques identiques (par exemple une année), un couple de mots dont au moins les quatre premiers caractères sont identiques ou une ponctuation (ibidem : 5).

Cette méthode se base donc encore une fois sur un calcul statistique et non pas linguistique. Simard, Foster et Isabelle considèrent que certaines suites de caractères sont similaires dans les deux textes et peuvent donc être considérées comme des points d'ancrage afin de perfectionner l'alignement lorsque le critère de la longueur des phrases n'est pas suffisant. Comme le soulignent Isabelle et Warwick-Armstrong (1993 : 297), cette technique marche bien uniquement entre des langues similaires.

2.2.2. Alignement par mot

L'alignement par mot reste encore un défi à surmonter. Comme l'explique Véronis (2000 : 161), afin d'obtenir un bon alignement mot à mot, il faut tenir compte des « unités complexes » telles que « mots composés, locutions, phraséologie ». Cet alignement se complique avec des couples de langues distantes entre elles ou avec des langues comme l'allemand qui compte un grand nombre de mots composés. Les mots grammaticaux ne sont pas forcément traduits en passant d'une langue à l'autre, mais ne peuvent être ignorés, car parfois ils font partie de la locution à repérer (Véronis, 2000 : 161). Prenons par exemple le mot composé allemand *Notausgang* et sa traduction en français *sortie de secours :* un mot lexical en allemand correspond à deux mots lexicaux liés par un mot grammatical en français.

En l'état actuel, aucune méthode statistique ne semble être en mesure d'effectuer de manière satisfaisante l'alignement d'éléments complexes (Véronis, 2000 : 161). C'est pourquoi les chercheurs exploitent des solutions linguistiques ou ajoutent des éléments linguistiques aux méthodes statistiques. Nous en décrivons deux : les cooccurrences parallèles et les catégories grammaticales.

2.2.2.1. Cooccurrences parallèles et alignement géométrique

Les modèles de cooccurrences parallèles se fondent surtout sur des modèles d'alignement géométrique. Sans vouloir entrer trop dans les détails, Melamed (1996 : 2) explique que ce système place le texte source et le texte cible sur un axe cartésien de manière à former un rectangle où l'unité de mesure est le caractère. L'origine représente le début des deux textes (*origin*), l'angle supérieur à droite représente la fin (*terminus*). Ces deux points sont reliés par la diagonale principale (*main diagonal*) et l'inclinaison de la diagonale représente l'inclinaison du texte (figure 2.2). Dans cet espace, il y a toute une série de *true points of correspondance* (TPC que Kraif (2006 : 5) appelle *transitions*), c'est-à-dire les coordonnées (p, q) formées par un terme du texte 1 et un terme du texte 2 qui sont l'un la traduction de l'autre. L'union de ces points représente ce que Kraif (2006 : 5) définit le *chemin d'alignement*.

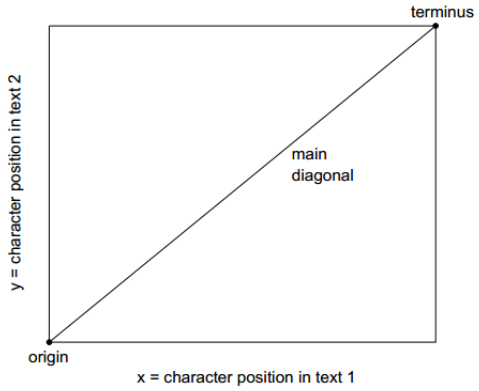


Figure 2.2 : Espace d'un bitexte (« *A bitext space* » : Melamed, 1996 : 2)

En se basant sur l'alignement géométrique, Melamed (2001 : 57) explique que la plupart des modèles de cooccurrences parallèles partent de l'idée que, dans un bitexte, si deux mots sont la traduction mutuelle l'un de l'autre, il y a beaucoup de chance qu'ils apparaissent dans des régions de bitextes correspondantes. Il définit ainsi ce modèle :

> « A *model of co-occurrence* is a boolean predicate that indicates whether a given pair of word *tokens* co-occur in corresponding regions of bitext space. Co-occurrence is a precondition for the possibility that two tokens might be mutual translations. » (ibidem)

Kraif (2006 : 9) précise qu'afin de mesurer ce degré d'association entre deux unités source et cible, différents indices statistiques peuvent être utilisés, tels que l'information mutuelle, le t-score, le rapport de vraisemblance, ou la probabilité de l'hypothèse nulle.

2.2.2.2. Catégories grammaticales

Une autre possibilité pour l'alignement par mot est celle d'utiliser les catégories grammaticales (ou patrons) sur des corpus annotés. Ces systèmes combinent généralement une technique statistique à des éléments linguistiques. Comme notre travail se fonde sur des corpus non annotés[9], nous nous limitons à voir très brièvement de quoi il s'agit en prenant comme exemple l'étude de Piperidis, Boutsis et Papageorgiou (2000) qui a pour but celui d'aligner un corpus parallèle au niveau des mots afin de pouvoir extraire une terminologie bilingue. Pour arriver à l'alignement par mot, cette méthode passe par trois phases. Premièrement, les textes sont alignés au niveau des phrases et des éléments (tels que les limites des mots et des phrases ainsi que les dates ou les abréviations) sont marqués. Dans un deuxième temps, les textes sont annotés et chaque terme est rapporté à sa forme canonique (lemmatisation). Ensuite, le système recherche les noms à travers des patrons et extrait les termes à travers une expression régulière. Nous traitons ces deux derniers concepts dans la section 3.

3. Concordanciers parallèles

Le troisième type d'outils que nous examinons est le concordancier parallèle. Dans le domaine de la traduction, le concordancier est utilisé, entre autres, pour la recherche ponctuelle de termes et d'équivalents dans une ou plusieurs langues.

Nous donnons tout d'abord une définition de concordancier parallèle (section 3.1), puis nous décrivons les différentes modalités de recherche (section 3.2) et enfin nous passons en revue les divers affichages de données (section 3.3) et nous présentons les affichages d'informations spécifiques (section 3.4).

[9] D'après la définition de Laporte (2000 : 27-28), l'annotation ou étiquetage des corpus regroupe « l'ensemble des techniques qui concourent à passer d'un texte brut, exempt d'information linguistique, à une séquence de mots étiquetés par des informations linguistiques, au premier rang desquelles les informations morphologiques et grammaticales. »

Précisons que les modalités de recherche et les types d'affichage que nous décrivons sont présents également sur les concordanciers monolingues, à la seule exception de la recherche parallèle que nous traitons à la section 3.2.2.3.

3.1. Définition

Le concordancier est un outil d'analyse des corpus. L'Homme (2004 : 143) le définit ainsi :

> « [Le concordancier] est conçu pour retrouver les ***occurrences*** d'une ou plusieurs ***chaînes de caractères*** dans un ou plusieurs textes électroniques. Ces chaînes sont des ***mots graphiques*** (par exemple, *informatique*), des parties de mots (comme *inform-)* ou une combinaison de mots (*systèmes informatiques*). »

Dans notre cas, nous allons nous concentrer sur les concordanciers parallèles, outils qui servent à analyser les corpus parallèles, et plus particulièrement sur leurs modalités de recherche. Ces outils permettent de retrouver des occurrences dans deux (ou plusieurs) langues simultanément et d'afficher les résultats en parallèle.

3.2. Modalités de recherche

Nous présentons ici les principales modalités de recherche disponibles pour les corpus parallèles non annotés, en les subdivisant en recherche simple, recherche avancée et recherche par expressions régulières.

3.2.1. Recherche simple

3.2.1.1. Recherche par mot

Le premier type de recherche disponible est celui par mot. Cette recherche permet de retrouver toutes les occurrences d'une chaîne de caractères exacte. Il s'agit de la modalité de recherche la plus rapide, mais, en même temps, c'est la moins efficace, car elle ne tient pas compte des formes fléchies. Elle peut tenir compte de la casse.

3.2.1.2. Recherche par troncature

La recherche par troncature permet de lancer une requête en saisissant une séquence de caractères contenue dans le terme. Si l'outil accepte les mots simples, la troncature est généralement indiquée par un astérisque qui peut se trouver n'importe où dans la chaîne de caractères.

Cette modalité de recherche est très utile pour retrouver les formes fléchies d'un verbe (L'Homme, 2008 : 156), des séquences caractérisées par le même suffixe ou préfixe ou encore des mots appartenant à la même famille morphologique.

3.2.2. Recherche avancée

3.2.2.1. Opérateur OU

L'opérateur OU[10] permet de chercher deux séquences qui apparaissent dans des contextes différents. En d'autres termes, cela équivaut à lancer deux requêtes en une seule fois. L'Homme (2008 : 156) explique que cette modalité de recherche est utile pour trouver des synonymes, des quasi-synonymes ou des variantes graphiques (par exemple clé ou clef).

3.2.2.2. Recherche par cooccurrents

La recherche par cooccurrents repère tous les contextes dans lesquels deux ou plusieurs mots apparaissent ensemble. De cette manière, la recherche est ciblée afin d'afficher uniquement les résultats pertinents.

Ces mots peuvent être contigus ou non contigus. Dans le second cas, certains concordanciers disposent d'un paramétrage pour définir le nombre de mots entre les deux occurrences.

Cette fonction permet de retrouver les mots complexes même lorsqu'ils sont interrompus par un autre mot.

3.2.2.3. Recherche parallèle

Les concordanciers parallèles offrent la possibilité de lancer des requêtes dans les deux langues simultanément. Il est donc possible de chercher un mot et une traduction spécifique. Par exemple, nous pouvons chercher dans notre corpus CHOCOLAT[11] toutes les occurrences dans lesquelles le terme italien « *scatola* » est traduit par « boîte ».

On peut également lancer une requête et exclure une traduction donnée. Ceci permet de mieux cibler la recherche lorsqu'une requête simple donne trop d'occurrences. Faisant suite à l'exemple précédent, nous pouvons chercher les occurrences dans lesquelles « *scatola* »

[10] Il existe trois opérateurs booléens : ET, OU et SAUF. Cependant, nous considérons que l'opérateur ET, lorsqu'il se réfère à une recherche dans la même langue, correspond à la recherche par cooccurrents. Nous retenons que les opérateurs ET et SAUF sont plutôt à considérer dans la recherche parallèle (voir la section 3.2.2.3).

[11] Le corpus CHOCOLAT est un des deux corpus que nous utilisons dans le chapitre 5 pour l'évaluation des outils. Il est composé de textes dont les sujets principaux sont le processus d'élaboration du chocolat et la présentation des produits finis. Nous en parlons dans les détails à la section 5.3.2.1 du chapitre 4.

n'est pas traduit par « boîte ». Nous découvrons ainsi que parfois il est traduit par « coffret ».

3.2.3. Recherche par expressions régulières

Une expression régulière (ou RegExp) est une chaîne de caractères alphanumériques pouvant contenir des caractères spéciaux que l'on utilise pour chercher, modifier ou remplacer des mots dans un texte. La chaîne de caractères recherchée est appelée aussi motif ou patron.

On retrouve les RegExp dans la plupart des langages de programmation, mais aussi dans des éditeurs de textes tels que *Microsoft Word*.

Dans un concordancier, les expressions régulières permettent d'effectuer des recherches plus ciblées à l'aide de métacaractères. Elles permettent entre autres de retrouver, en une seule fois, toutes les formes fléchies d'une séquence, par exemple, si l'on cherche *\bprésente?(s|nt|ons|z)?\b* [12] on retrouve toutes les formes du présent de l'indicatif du verbe *présenter*. On peut aussi retrouver des chiffres ou des dates, en utilisant une fourchette de caractères. Par exemple, le patron *[0-9]{1,}.?[0-9]{0,2} CHF* retrouve tous les prix en francs suisse.

Les expressions régulières permettent également d'effectuer des recherches avec des patrons interrompus : par exemple, si on cherche *plus (\w+?)+?que* on obtiendra tous les comparatifs de supériorité contenus dans un corpus français.

Enfin, lors de la préparation de corpus, les RegExp facilitent la tâche de nettoyage des textes. Par exemple, dans *Microsoft Word*, il est possible d'utiliser les expressions régulières dans la fenêtre « Rechercher et remplacer ». Lorsqu'un document est océrisé ou lorsqu'on copie un texte à partir d'un PDF, souvent, les phrases sont coupées par des retours à la ligne qu'il faut supprimer avant d'ajouter un texte au corpus. Avec les expressions régulières, il est possible de les effacer d'un seul coup.

3.3. Affichage des données : les concordances

Une concordance est un ensemble de toutes les occurrences contenant la chaîne de caractères recherchée insérée dans son contexte immédiat. Dans un corpus parallèle, ce contexte est mis en relation avec le segment traduit correspondant.

[12] Les exemples d'expressions régulières que nous présentons ici se réfèrent au langage utilisé par le concordancier parallèle ParaConc dont nous parlons dans le chapitre 3 (section 2.2).

Il s'agit d'un affichage qui permet au traducteur de voir une séquence dans plusieurs contextes simultanément afin de détecter des modèles linguistiques et conceptuels souvent difficiles à repérer dans d'autres ressources, comme par exemple le signifié d'un terme, les cooccurrences ou des constructions particulières de phrases (Bowker 2001 : 349).

Il existe deux types d'affichage de concordance : le KWIC et l'affichage plein texte. L'affichage KWIC (*key-word in context)* extrait les contextes du motif recherché et les aligne de manière à mettre tous les motifs en évidence, au centre de la ligne. Comme le précise L'Homme (2004 : 144), le contexte à droite et à gauche est normalement paramétrable par l'utilisateur, c'est-à-dire que l'on peut choisir combien de mots afficher avant et après l'occurrence recherchée et trier les résultats par ordre alphabétique par rapport au mot qui suit ou qui précède le patron. La figure 2.3 montre le résultat dans l'ordre original de la recherche du patron « chocolat » dans le corpus CHOCOLAT. La lecture des occurrences est, dans ce cas, difficile car le contexte n'est pas trié.

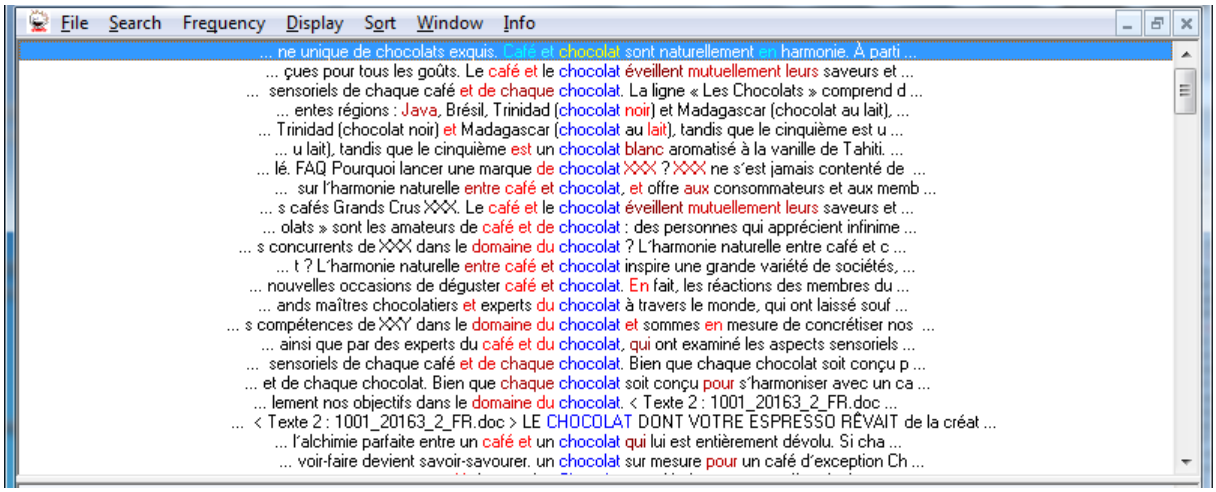


Figure 2.3 : *ParaConc* recherche de « chocolat » avec affichage KWIC et occurrences dans l'ordre original

En triant les résultats par le premier et le deuxième mot à droite, nous voyons clairement que les occurrences ayant le même contexte à droite sont réunies, ce qui permet de focaliser notre attention uniquement sur les occurrences pertinentes (figure 2.4).

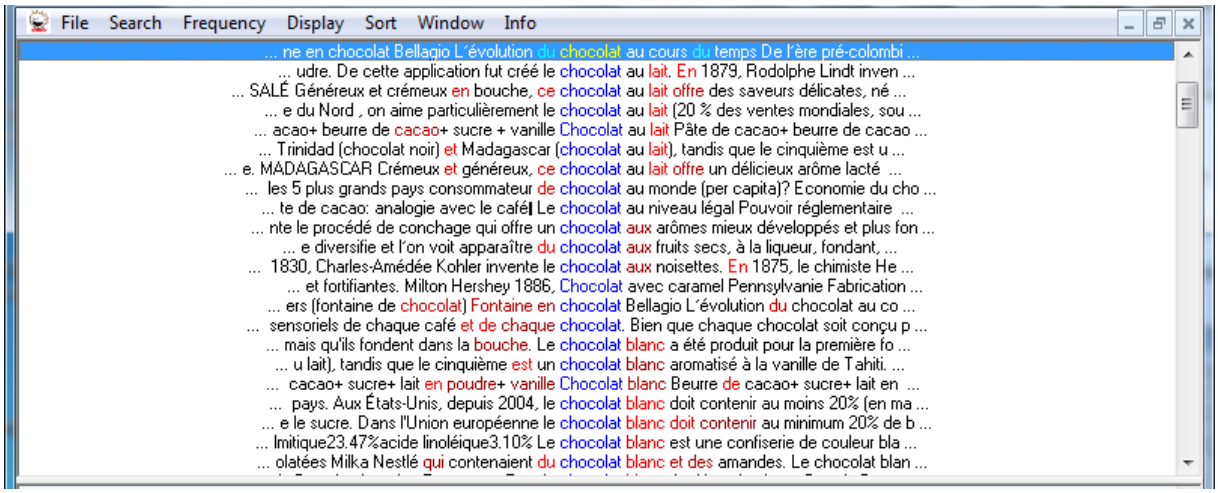


Figure 2.4 : *ParaConc*, recherche de « chocolat » avec affichage KWIC trié par le premier mot à droite et le deuxième mot à droite

L'affichage plein texte reprend tout le texte d'origine dans lequel se trouve le patron en mettant en évidence la séquence recherchée. Ceci permet d'avoir un contexte plus ample, mais rend parfois la recherche moins immédiate. Voilà pourquoi certains concordanciers combinent les deux techniques en offrant un affichage KWIC qui permet de voir l'occurrence en plein texte si nécessaire.

3.4. Affichage d'informations spécifiques

Nous présentons, dans cette partie, des fonctions particulières des concordanciers permettant d'obtenir des informations spécifiques sur l'ensemble du corpus, comme dans le cas de la liste de mots ou sur une séquence recherchée, comme dans le cas des clusters et des collocations.

3.4.1. Liste de mots (wordlist)

Une liste de mots permet de découvrir de quelles unités différentes est composé un corpus et combien de fois elles apparaissent (Bowker, 2002 : 47), dans le but de se familiariser avec le langage utilisé dans le corpus (Bowker et Pearson, 2002 : 119). Elle représente un point de départ pour des recherches plus approfondies sur les termes et leurs collocations (Adolphs, 2006 : 40).

Zanettin (2012 : 117) explique qu'il est possible d'afficher la wordlist selon deux modalités : par ordre alphabétique ou selon la fréquence.

L'affichage par ordre alphabétique permet de rassembler tous les mots appartenant à la même famille sémantique et les formes fléchies (à l'exception des formes verbales

irrégulières). Zanettin (ibidem) explique que parfois il est possible de classer les mots en partant de la dernière lettre, ce qui permet de regrouper tous les termes avec le même suffixe.
L'affichage par fréquence trie les mots selon le nombre de fois qu'ils apparaissent dans le corpus en ordre croissant ou décroissant.

Comme nous l'avons vu, la wordlist contient tous les mots du corpus, y compris les mots grammaticaux. Ces mots sont généralement les plus fréquents et, comme l'observe Zanettin (ibid. : 118) en se référant à la loi de Zipf sur la fréquence des mots dans un texte, ils occupent la première partie de la liste, alors qu'à la fin on retrouve les mots qui n'apparaissent qu'une fois (hapax).
Afin d'avoir un affichage plus ciblé, il est possible de filtrer les résultats à travers une stoplist (ibid. : 119) ou liste d'exclusion (L'Homme, 2004 : 150) qui permet de dresser la liste des mots à ignorer. Ceci fait « émerger plus clairement les thèmes centraux d'un corpus » (L'Homme, 2004 : 150) en permettant donc d'ignorer les mots grammaticaux et peu intéressants et en ne gardant que les termes qui caractérisent le corpus.
Comme l'explique Bowker (2001 : 349) les listes de mots peuvent aider le traducteur à choisir quel terme utiliser face à des potentiels synonymes ou à des équivalents de traduction. En particulier, elles aident à évaluer si un terme est utilisé couramment par les experts du domaine ou s'il s'agit simplement d'une préférence d'un auteur (ibidem).

3.4.2. Cluster

Le terme « cluster » indique une séquence d'unités qui apparaît avec une certaine fréquence dans le même contexte que l'unité ou l'expression recherchée (Zanettin 2012 : 132). Les résultats sont affichés sous forme de liste. Généralement, il est possible de paramétrer le nombre d'unités à repérer et de choisir de chercher à droite ou à gauche de la séquence.
Ce type de recherche permet d'analyser de quelle manière sont associées les unités. Par exemple, si nous prenons la partie en français de notre corpus FORMATION[13] et nous cherchons les clusters pour le mot « formation » dans un groupe de deux ou trois mots à droite, nous voyons que la chaîne de caractères « formation professionnelle » apparaît 460

[13] Le corpus FORMATION est le deuxième corpus que nous utilisons dans le chapitre 5 pour l'évaluation des outils. Il est composé de documents d'un institut de formation professionnelle suisse, parmi lesquels le règlement de l'institut et un rapport de gestion.

fois, suivie de « formation continue » (121 fois) et de « formation professionnelle initiale » (77 fois).

3.4.3. Collocations

Bowker (2002 : 64) définit ainsi les collocations :

> « collocations are words that appear together with a greater than random probability. »

En d'autres termes, il s'agit d'une liste dans laquelle figure un patron accompagné des unités qui lui sont associées le plus souvent, mais de manière statistiquement significative. Il est possible de voir aussi dans quelle position (à droite ou à gauche de l'unité observée) ils se trouvent le plus souvent.

Bowker (2002 : 64-65) explique que cette association est le résultat d'un calcul statistique. En bref, ce calcul détermine combien de fois les deux termes apparaissent dans le corpus et compte le nombre de fois qu'ils apparaissent ensemble ou avec d'autres termes. Statistiquement, si le nombre de fois où ils apparaissent ensemble est plus élevé par rapport au nombre de fois où ils apparaissent avec d'autres termes, ceci signifie que les deux unités sont liées entre elles. Ceci permet d'exclure les associations fortuites (ibidem).

Les résultats sont affichés sous forme de tableau par ordre alphabétique ou de fréquence.

Les collocations permettent de voir immédiatement les cooccurrents d'un terme et le nombre de fois qu'ils apparaissent ensemble. Cela peut être vu comme une première approche pour identifier les cooccurrents d'une séquence pour ensuite les analyser dans leurs contextes à travers des recherches plus spécifiques.

4. Conclusions

Les trois types d'outils classiques d'exploitation des corpus parallèles présentés dans cette section ont chacun une fonction différente.

Le principe de base d'une mémoire de traduction est celui de « recycler » les traductions précédentes afin de ne pas devoir traduire deux fois le même segment. Dans ce but, les anciennes traductions sont stockées dans une base de données et récupérées au moment de la traduction à l'aide d'un algorithme qui est en mesure d'établir des correspondances entre le segment à traduire et les segments contenus dans la mémoire de traduction. Ces correspondances peuvent être parfaites ou partielles.

Outre la récupération automatique des segments, les MT disposent d'une interface de recherche permettant d'effectuer des recherches simples (recherche par mot avec ou sans joker), des recherches avancées (recherche par troncature et recherche par filtres) ou des recherches par « *fuzzy match* ».

L'aligneur est généralement intégré dans les outils de TAO. Il permet de segmenter les textes et de mettre en relation les segments des textes sources avec leurs correspondants dans les textes cibles. Il existe principalement deux méthodes d'alignement automatique : l'alignement par phrase et l'alignement par mot. Nous constatons que l'alignement par phrase est actuellement plus fiable, car il permet d'atteindre des résultats satisfaisants (jusqu'à 99 % sur les corpus segmentés). Nous présentons trois méthodes d'alignement par phrase (l'ancrage lexical, l'alignement par longueur de phrase et les cognats) et deux méthodes d'alignement par mots (les cooccurrences parallèles et les catégories grammaticales).

Le concordancier parallèle est un outil d'analyse des corpus permettant de retrouver des occurrences dans des corpus parallèles. Il permet donc d'effectuer des recherches ponctuelles en mettant à disposition des modalités de recherche et d'affichages différentes. Nous pouvons subdiviser les modalités de recherche en recherche simple (recherche par mot et recherche par troncature), recherche avancée (recherche par opérateur OU, recherche par cooccurrents et recherche parallèle) et recherche par expressions régulières. Nous présentons les deux types d'affichage des concordances (affichage KWIC et affichage plein texte) et pour finir les fonctions permettant d'obtenir des informations spécifiques sur le corpus entier (listes de mots) et sur des unités spécifiques (clusters et collocations).

Partie II : Évaluer pour aller vers un outil idéal

Chapitre 3 : Choix des outils

Afin de poursuivre nos recherches vers le concordancier idéal, il est maintenant temps de tester quelques outils disponibles sur le marché. Nous présentons tout d'abord quelques outils afin de dresser un état des lieux des logiciels existants (section 1), sans pour autant les analyser de manière approfondie. Nous les présentons ici dans le seul but de montrer la diversité des instruments disponibles sur le marché. Notons qu'aucun d'entre eux n'est parmi les outils que nous testons.

Ensuite, nous décrivons dans les détails les trois outils que nous avons choisis pour notre évaluation : *MultiTrans Prism, ParaConc* et *myCAT* (section 2). Cette évaluation nous permet de comprendre si les outils répondent réellement aux besoins des utilisateurs et d'établir quelles sont les fonctionnalités indispensables pour un outil d'exploitation des corpus.

À la fin de ce chapitre, nous mettons en relation ces trois outils à travers une grille d'analyse qui en résume les caractéristiques (section3).

1. Outils disponibles

Comme nous l'avons vu dans le chapitre précédent, les outils d'exploitation de corpus peuvent être regroupés en trois grandes familles :

- **mémoires de traduction** : outils d'assistance à la traduction, qui permettent de récupérer automatiquement les segments des traductions précédentes pendant la phase de traduction (voir chapitre 2 section 1) ;
- **aligneurs** : outils qui permettent de segmenter les textes et de mettre en relation les segments des textes sources avec leurs correspondants dans les textes cibles, afin de préparer les corpus pour l'exploitation avec d'autres outils (voir chapitre 2 section 2) ;
- **concordanciers parallèles** : outils d'analyse des corpus parallèles qui permettent de retrouver des occurrences dans deux (ou plusieurs) langues simultanément et d'afficher les résultats en parallèle (voir chapitre 2 section 3).

Nous présentons, dans cette section, divers outils d'exploitation de corpus parallèles disponibles sur le marché, en nous limitant aux aligneurs et aux concordanciers parallèles. Nous ne présentons donc aucune mémoire de traduction, car nous retenons qu'il existe à ce sujet une vaste littérature. Nous citons, à titre d'exemple, les logiciels *SDL Trados Studio*, *Similis* et *Wordfast*. Pour de plus amples informations sur les mémoires de traduction, nous renvoyons le lecteur, par exemple, à Massion (2005).

Pour connaître les sites internet des logiciels, nous renvoyons le lecteur à la webographie qui se trouve à la fin de ce mémoire.

1.1. Aligneurs

Comme nous l'avons vu dans le chapitre précédent, les aligneurs sont intégrés dans la plupart des outils d'exploitation de corpus. Toutefois, il existe des outils entièrement dédiés à l'alignement. Nous en présentons ici deux : *Alinéa* et *AlignFactory*.

1.1.1. Alinéa

Olivier Kraif, directeur du Département d'informatique pédagogique de l'Université Stendhal de Grenoble, a développé un outil appelé *Alinéa*, qui est à la base un aligneur.

Kraif (2006) part de la théorie de Kay et Röscheisen (voir chapitre 2 section 2.2.1.2) et en développe une nouvelle basée donc sur l'ancrage lexical, mais en utilisant des points d'ancrage entre les deux langues.

Le logiciel cherche le « meilleur chemin d'alignement possible » dans un couple de textes. Pour faire cela, il passe par deux étapes : dans un premier temps, *Alinea* extrait des points d'ancrage « faibles » tels que les balises, les nombres, les noms propres et les emprunts; ensuite, le logiciel effectue l'alignement proprement dit en cherchant le chemin le plus probable. Il y arrive en évaluant tous les éléments disponibles : « *rapport des longueurs*, appariement de chaînes identiques (*transfuges*), appariement de mots ressemblants (*cognats*), de mots possédant des *distributions* similaires ou de lexèmes équivalents. » (Kraif, 2006 : 25). Le reste est fait par un algorithme de programmation dynamique qui finalise l'alignement en tenant compte des points d'ancrage et en calculant les alignements possibles entre les phrases contiguës. Les alignements admis sont donc ceux que nous avons décrits dans le chapitre précédent (section 2.2.1.1) auxquels s'ajoutent les rapports 3:1 et

1:3. Lorsque ces indices sont insuffisants, il est possible d'utiliser des dictionnaires bilingues afin de résoudre le problème (Kraif, 2006 : 25).

Bien que ce logiciel propose également des fonctions de recherche complexes et des concordances avec des critères bilingues, il suffit de faire quelques essais pour comprendre qu'uniquement les recherches monolingues sont disponibles. Il semble en effet impossible de faire des recherches sur les corpus parallèles.

En outre, *Alinéa* se bloque facilement et son utilisation en tant que concordancier devient vite pénible. Il reste quand même un excellent aligneur.

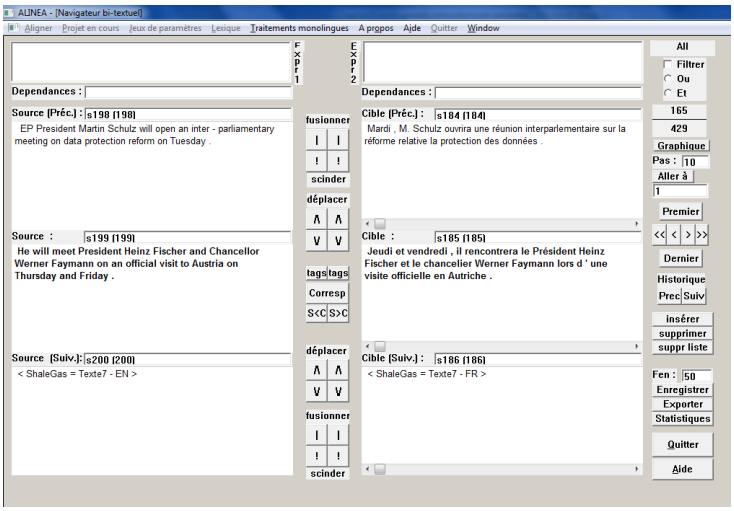


Figure 3.1 : Interface d'*Alinéa*

1.1.2. AlignFactory

AlignFactory est un aligneur automatique développé par la société canadienne Terminotix Inc. Il accepte les principaux formats de textes et traite environ 40 langues.

Étant donné qu'il permet de créer des bitextes au format BitextesLogiTerm, HTML et TMX, il est compatible avec les principaux outils de TAO.

Il est possible de créer un projet d'alignement, ce qui donne l'avantage d'apparier un grand nombre de textes automatiquement, même à partir de plusieurs répertoires ou sous-répertoires, et d'obtenir un seul fichier TMX par couple de langues. *AlignFactory* donne la possibilité de paramétrer les options de segmentation (par exemple : segmentation par paragraphe, par phrase, en prenant en compte les points-virgules et les deux points) et d'ajouter des filtres pour exclure, par exemple, les segments identiques.

L'interface est très simple, ce qui permet une utilisation intuitive du logiciel.

Suite à plusieurs tests, on s'aperçoit qu'*AlignFactory* propose effectivement des alignements corrects avec un taux d'erreur très bas pour les documents du même format (par exemple deux documents Word), mais qu'il rencontre d'énormes difficultés lorsqu'il faut aligner un PDF et un Word.

Il inclut également un éditeur TMX qui permet d'afficher les documents sous forme de tableau et de les modifier. Il est possible de déplacer, ajouter, supprimer, fusionner ou séparer des cellules dans le tableau afin de faciliter la modification des segments alignés ainsi que de modifier le texte à l'intérieur des segments.

Les avantages résultent, sans doute, de la rapidité du logiciel qui permet d'aligner automatiquement en quelques minutes un grand volume de documents et du fait qu'il accepte une grande variété de langues et de types de fichiers. De plus, si une phrase n'est pas alignée correctement, ceci ne compromet pas entièrement la tâche, car *AlignFactory* fait une sorte de compensation avec les phrases suivantes pour rattraper l'erreur et reprendre normalement quelques segments après.

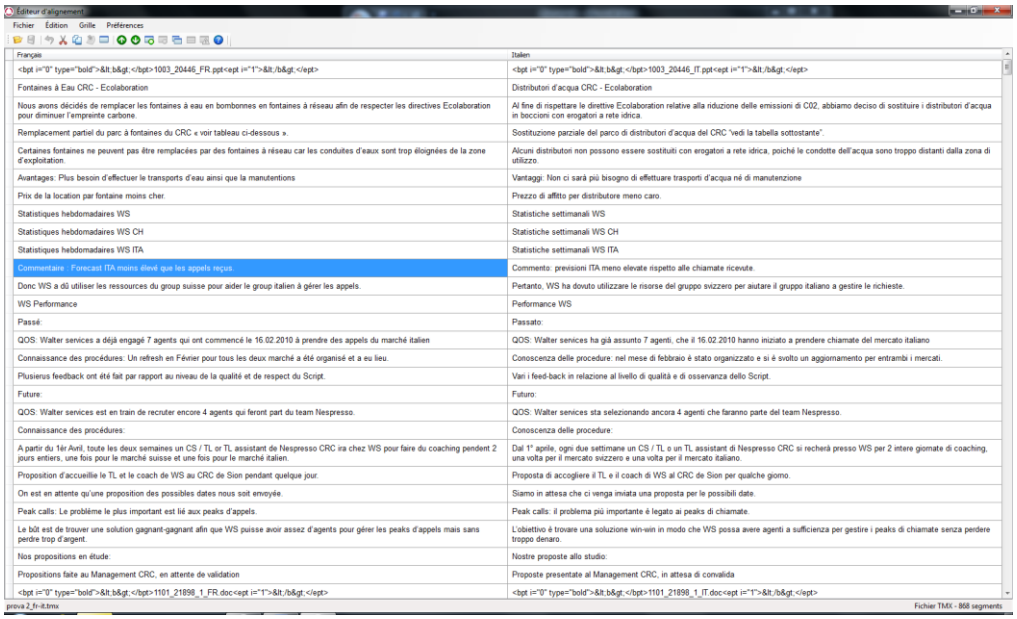


Figure 3.2 : *AlignFactory* : interface de l'éditeur d'alignement

1.2. Concordanciers

De nombreux corpus parallèles sont disponibles en ligne. Nous en citons deux à titre d'exemple. Le premier est le *Canadian Hansard corpus* développé par le Natural Language

Group à l'Université de Californie du Sud qui regroupe les transcriptions des débats du parlement Canadien en anglais et en français et contient 283 millions de mots. Le deuxième est l'*Europarl*, développé par l'Université d'Édimbourg qui rassemble les procès-verbaux du Parlement européen. La langue pivot est l'anglais est le corpus compte 21 langues.

Pourtant, face à cette richesse de ressources en ligne, il suffit d'approfondir les recherches pour s'apercevoir qu'il existe peu de concordanciers parallèles permettant d'analyser exclusivement des corpus parallèles privés. Nous en présentons dans cette section quelques-uns parmi les principaux, à savoir : *AntPConc, MultiConcod, Concordancier* et *TradooIT*.

1.2.1. AntPConc

Laurence Anthony, professeur à la « Faculty of Science and Engineering » à la Waseda University (Japon) et créateur du concordancier monolingue *AntConc*, propose un nouveau concordancier parallèle appelé *AntPConc* version 1.0.2 daté de 2013. L'interface rappelle celle du concordancier monolingue *AntConc*, mais on s'aperçoit immédiatement que les fonctions proposées sont nettement réduites : aucune possibilité de recherche avec des expressions régulières, pas de clusters, pas de collocations, ni de liste de mots. En bref, aucune fonction avancée. Les seules modalités de recherche proposées sont celles par mot ou par chaîne de caractères, l'affichage est en KWIC dans la langue source et par phrase dans la langue cible. Il est possible d'inverser les langues directement depuis l'interface de recherche, de trier les résultats par rapport aux contextes droits ou gauches et de modifier le nombre de mots à afficher dans le contexte.

Le logiciel accepte uniquement des fichiers TMX et ne permet pas de modifier l'alignement.

Pour l'instant, aucun manuel et aucune présentation d'*AntPConc* ne sont disponibles, ni à partir du logiciel, ni en ligne. Toutefois, s'agissant d'un outil récent, il est probable qu'une nouvelle version plus complète sera disponible dans les mois à suivre.

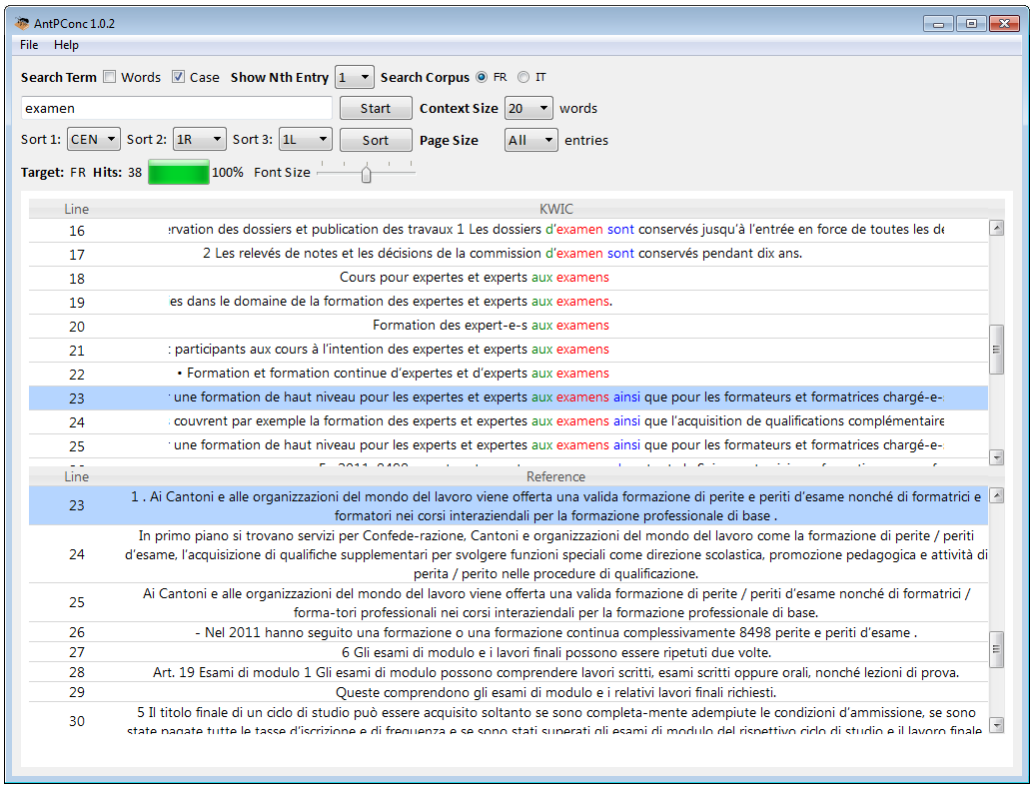


Figure 3.3 : Interface d'AntPConc

1.2.2. MultiConcord

Le portail de l'enseignement supérieur de la recherche et de l'innovation en Lorraine propose *MultiConcord,* un concordancier parallèle multilingue conçu par Francine Roussel, maître de conférences en anglais à l'Université Nancy 2. Il s'agit d'un outil payant vendu uniquement sur CD à 62,10 EUR TTC.

Il accepte les langues suivantes : français, anglais, allemand, italien, espagnol, portugais, grec moderne, danois, suédois, finnois, néerlandais et russe.

Il est fourni avec des corpus contenant des débats du Parlement européen, des extraits du Guide Michelin, des lettres et courriels et *Alice au pays des merveilles* de Lewis Carroll. L'utilisateur peut y ajouter ses propres textes via un logiciel annexe. L'analyse de corpus privé est donc possible, mais est présentée comme un atout et non pas comme un aspect essentiel, car *MultiConcord* est conçu principalement comme outil de consultation des corpus déjà fournis.

1.2.3. Concordancier

L'équipe Langue et Dialogue du Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) a développé un concordancier parallèle appelé tout simplement *Concordancier*, logiciel couplé à *XAlign* pour l'alignement multilingue.

Concordancier est un outil open source que l'on peut télécharger depuis le site du LORIA. L'installation se fait dans un environnement *Java*, mais il nous a été impossible de comprendre quelle est la démarche à suivre !

Nous avons tout de même accès au mode d'emploi, c'est pourquoi nous avons constaté que l'interface est basique, mais il est possible, a priori, d'utiliser des expressions régulières et des opérateurs booléens.

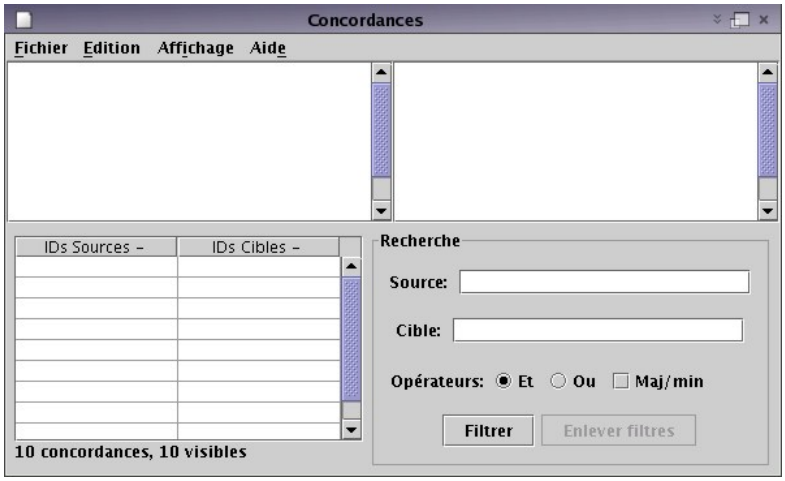


Figure 3.4 : Interface de *Concordancier*

1.2.4. TradooIT

La société Okidoo Inc. s'est associée à Logosoft pour commercialiser *TradooIT*[14], un ensemble d'outils d'aide à la traduction accessibles gratuitement en ligne comprenant, entre autres, un concordancier bilingue. Les langues disponibles sont l'anglais, le français et l'espagnol. Peu d'informations sont disponibles en ce qui concerne son architecture. Toutefois, on peut déduire que l'alignement est statistique et que les corpus ne sont pas annotés[15]. Après un enregistrement gratuit, *TradooIT* permet d'accéder au concordancier bilingue. Ce dernier met à disposition des utilisateurs des mémoires de traduction issues du web, en particulier des sites gouvernementaux et des sites d'organisations internationales, ainsi que des corpus contenant des sous-titres de films. La seule partie en anglais compte plus de 250 millions de mots[16]. Il est également possible de charger des MT privées et de décider de les partager ou non avec le grand public.

[14] http://www.tradooit.com/ - Consulté le 23/05/2014
[15] http://blog.tradooit.com/search?updated-max=2012-09-10T18:09:00-07:00&max-results=7&reverse-paginate=true - Consulté le 23/05/2014
[16] Ibidem

Lorsqu'on lance une recherche, le logiciel récupère les segments des MT publiques et privées ainsi que les termes dans les banques terminologiques publiques (TERMIUM Plus, ONTERM, Wikipédia et le Portail linguistique de Microsoft) et privées, et les présente sous forme de tableau. Les occurrences sont surlignées dans les deux langues et l'affichage s'effectue par phrases. C'est là la grande particularité de ce logiciel. Il s'agit d'un des seuls corpus non annoté qui est en mesure de repérer automatiquement les séquences ainsi que leur traduction. L'hypothèse est qu'il y parvient en passant par les bases terminologiques.

Le concordancier présente également une statistique des différentes formes et traductions trouvées et met à disposition des filtres en fonction de la forme et de la source. Les utilisateurs avancés ont la possibilité de rechercher les radicaux des mots, d'inclure des métacaractères et disposent de fonctions de suggestion de recherche.

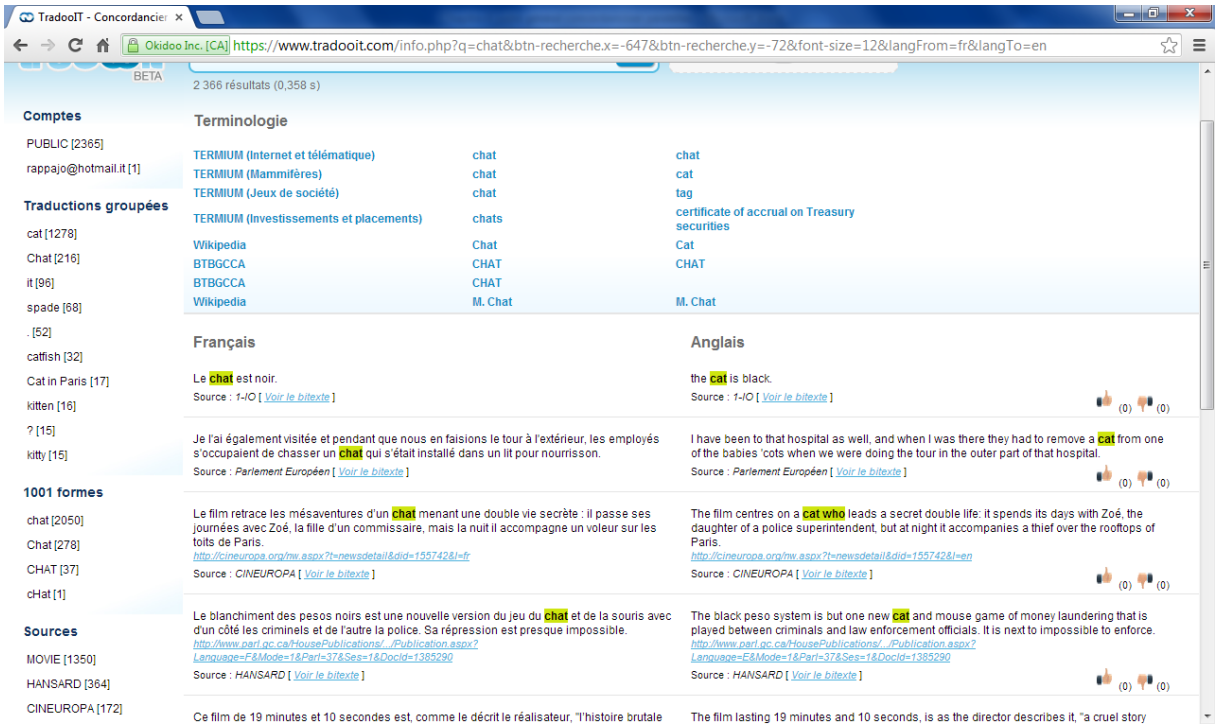


Figure 3.5 : Résultat d'une recherche dans *TradooIT*

2. Choix des outils

Après avoir vu les différents types d'outils disponibles, nous choisissons, dans cette section, les logiciels à utiliser pour notre tâche d'évaluation.

Notre évaluation se base sur les fonctionnalités de recherche, c'est pourquoi nous essayons de choisir des outils de différente nature, afin de comprendre quelles fonctionnalités sont effectivement utiles du point de vue d'un traducteur.

Nous décidons de comparer deux concordanciers parallèles avec une mémoire de traduction parce que ce sont les deux instruments qui permettent d'analyser les corpus parallèles, mais avec une grande différence dans l'usage. Les concordanciers peuvent être vus comme des outils de recherche terminologique ponctuelle. Ils servent donc principalement à la consultation, raison pour laquelle on s'attend à avoir des modalités de recherche avancées. Par contre, les mémoires de traduction sont des outils d'assistance à la traduction, qui permettent donc de récupérer automatiquement les segments des traductions précédentes pendant la phase de traduction. Étant donné qu'elles contiennent des bitextes et que parfois les correspondances sont trop basses pour être relevées automatiquement, les MT mettent à disposition une interface de recherche qui est souvent très basique. Le but de cette comparaison est donc de voir dans la pratique comment ces modalités de recherche différentes influencent la tâche de recherche et d'affichage des résultats.

Les trois outils que nous avons choisis sont *MultiTrans Prism, ParaConc* et *myCAT*.

MultiTrans Prism est un des logiciels de TAO les plus répandus. Il s'agit d'une mémoire de traduction de nouvelle génération qui garde en mémoire non pas des segments isolés mais des bitextes. Ceci le rapproche aux concordanciers parallèles. L'autre raison est que, contrairement à d'autres MT, *MultiTrans Prism* offre plusieurs options de recherche.

Parmi les concordanciers parallèles existants sur le marché, *ParaConc* est le plus connu et le plus complet, mais également l'un des plus anciens. Aucun des autres concordanciers parallèles que nous avons testés ne permet d'effectuer des recherches aussi complexes et d'afficher, par exemple, des listes de mots ou des clusters.

Enfin, le choix de *myCAT* est dû au fait qu'il s'agit d'un outil nouveau qui nous a été présenté directement par l'un de ses créateurs, M. Benzineb, que nous avons eu l'occasion de rencontrer à plusieurs reprises. Le logiciel en lui-même présente une conception un peu

différente par rapport aux autres concordanciers. Il cherche à optimiser les temps de recherche tout en réduisant au minimum les tâches d'entretien des corpus[17].

Nous présentons ici les trois logiciels. Pour chacun, nous décrivons tout d'abord les caractéristiques techniques, puis nous parlons du codage et de l'alignement pour ensuite passer aux fonctions de recherche. Pour terminer, nous montrons les autres fonctions disponibles.

2.1. MultiTrans Prism

MultiTrans Prism est un outil d'aide à la traduction et à la gestion langagière développé par la société canadienne MultiCorpora R&D Inc. Fondée en 1999, MultiCorpora[18] est une organisation internationale qui se consacre exclusivement au développement et au soutien de *MultiTrans Prism* et qui compte parmi ses clients plus de 60 organisations en Amérique du Nord et en Europe. Le 10 mars 2014, la société MultiCorpora a été acquise par la société R. R. Donnelley & Sons Company.[19] Cette dernière est une société américaine qui dispose de plusieurs bureaux dans le monde entier et qui fournit des services multilingues, des services d'assistance et des solutions de communication[20].

MultiTrans Prism est composé de 4 modules :

- la mémoire de traduction *TextBase* et son *Agent d'alignement* ;
- la *TermBase* pour le développement de bases terminologiques ;
- l'*Agent de traduction*;
- l'*Agent d'analyse* qui comporte des outils d'analyse de projet et de production de rapports.

Dans le cadre de ce mémoire, nous nous intéressons uniquement au module de la mémoire de traduction (*TextBase*), car c'est la partie qui permet d'exploiter les bitextes et donc de lancer des recherches terminologiques.

[17] Les tâches d'entretien des corpus sont représentées par le nettoyage des textes, l'alignement et la mise à jour des corpus.
[18] http://multicorpora.com/fr/societe/ - Consulté le 25/05/2014
[19] http://www.rrdonnelley.com/languagesolutions/fr/news/2014/03102014.aspx - Consulté le 23/05/2014
[20] http://www.rrdonnelley.com/languagesolutions/fr/locations/ - Consulté le 23/05/2014

2.1.1. Caractéristiques techniques

Les informations concernant les prix du logiciel ne sont pas disponibles sur le site officiel et nous n'avons pas réussi à les obtenir. À titre purement indicatif, nous pouvons signaler qu'il est possible d'acheter la version *MultiTrans Prism Freelance* au prix de 453.55 EUR[21], auquel il faut ajouter les Options de maintenance et de soutien pour 12 mois au prix de 113.58 EUR ou de 265.01 EUR.

En plus des licences clients, MultiCorpora propose aussi des versions serveur et web, dédiées aux entreprises, qui peuvent être adaptées aux besoins en variant le nombre d'accès, le volume de textes et les services complémentaires.

Nous testons ici la licence client *MultiTrans Prism Expert version 5.5.2388.0*.

Pour cette version, la configuration minimale requise est la suivante (MultiCorpora, 2013 : 18) :

- *Windows XP, Vista* ou *Windows 7* avec mises à jour;
- *Internet Explorer 7*;
- *Microsoft Word XP, 2003, 2007* ou *2010*, 32 bits, installé localement;
- Pentium IV 2 GHz (au minimum);
- 2 Go de RAM, dont au moins 1 Go de RAM disponible;
- 500 Mo d'espace disponible sur le disque dure pour l'installation et les données;
- *Microsoft .Net Framework 4.0*;
- *SQL Server 2008 Express*

MultiTrans Prism supporte une grande quantité de langues, des plus communes aux plus rares.

En ce qui concerne le module *TextBase,* il est possible d'ajouter des documents avec les formats suivants : doc, docx, txt, rtf, ppt, pptx, xlsx, xml, html, pdf, wpd ; tout autre format pouvant être ouvert et enregistré en format texte dans *Microsoft Word*, INX, DITA, MIF, RESX et autres formats balisés; formats de fichiers bilingues TMX, XLIFF et RTF.

2.1.2. Codage et alignement

La *TextBase* est une mémoire à reconnaissance avancée qui indexe tous les documents. Ces derniers ne sont donc pas divisés en segments indépendants, mais stockés en tant que

[21] Le prix indiqué se réfère à la version Freelance en choisissant la France comme pays. C'est la seule version que l'on peut acheter en ligne au lien https://www.multicorpora.ca/CARTStep2_f.php - Consulté le 15/12/2013

bitexte. En outre, les mémoires sont multidirectionnelles, c'est-à-dire qu'il est possible d'inverser la direction des langues.

L'alignement est statistique et se base sur des algorithmes complexes. Les erreurs d'alignement peuvent être corrigées manuellement et on peut établir une liste d'exceptions pour chaque langue. Il se fait à partir de l'*Assembleur de TextBase* qui permet, à travers une démarche guidée, de sélectionner, apparier et aligner les documents. S'il y a plusieurs documents à aligner, il est possible de les apparier automatiquement avec la fonction *ListBuilder.*

La figure 3.6 montre le module *TextBase* après avoir créé la mémoire de traduction. Comme on peut le voir, la fenêtre est divisée en deux grandes parties principales : un volet pour les recherches et, juste à côté, les textes source et cible alignés sur deux colonnes. Notons qu'il est possible d'inverser les langues.

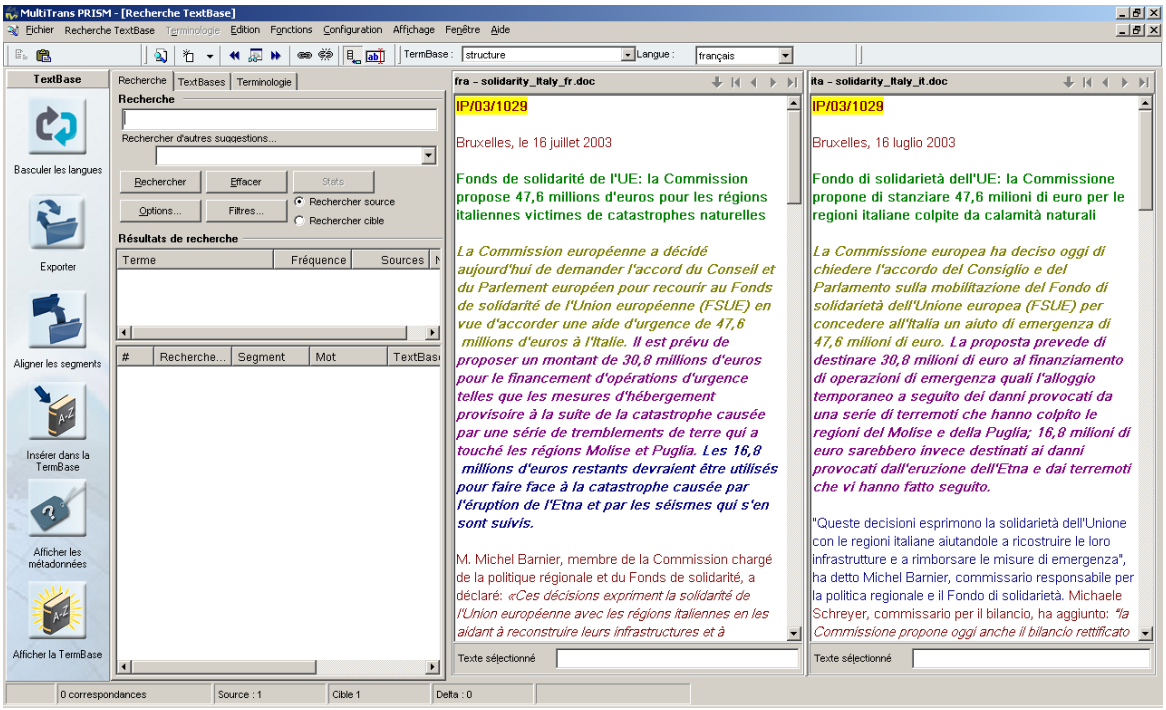


Figure 3.6 : Aperçu de la *TextBase* de *MultiTrans Prism* suite à l'alignement de deux textes

2.1.3. Recherche

MultiTans Prism permet d'effectuer principalement deux types de recherche : la recherche par mots qui permet de chercher des mots entiers, et celle par radicaux qui permet d'effectuer la recherche par troncature. Cependant, les radicaux ne sont pas linguistiques, ce qui veut dire qu'ils ne correspondent pas forcément à des radicaux morphologiques et les résultats peuvent donner des mots qui n'appartiennent pas à la famille d'un point de vue

sémantique. Cette modalité de recherche reste toutefois la seule manière de trouver la forme fléchie d'un mot. Dans les deux cas, les opérateurs booléens et les jokers ne sont pas admis.

MultiTrans Prism propose également la recherche par fuzzy, appelée ici « *Mots consécutifs* ». Si cette option est cochée, *TextBase* affichera tous les résultats contenant l'ordre exact des mots, dans le cas contraire, tous les segments contenant ces mots seront affichés, quel que soit leur ordre d'apparition. « *Tous les mots* » permet d'afficher uniquement les segments contenant tous les mots introduits dans la recherche. Si cette option n'est pas cochée, *MultiTrans Prism* affichera tous les segments contenant au moins deux des mots recherchés.

L'affichage des résultats est plein texte avec la possibilité de passer d'un segment à l'autre en cliquant sur la liste qui est affichée sous « Résultat des recherches » dans le deuxième volet. Dans le volet « Terme » apparaissent toutes les variantes de la séquence recherchée. Lorsqu'on lance une recherche, les segments correspondants sont surlignés en jaune et le terme trouvé est affiché en bleu et en gras. En bas, à gauche de la fenêtre principale, apparaît le nombre d'occurrences trouvées ainsi que la position des segments source et cible (figure 3.7 encadré en rouge).

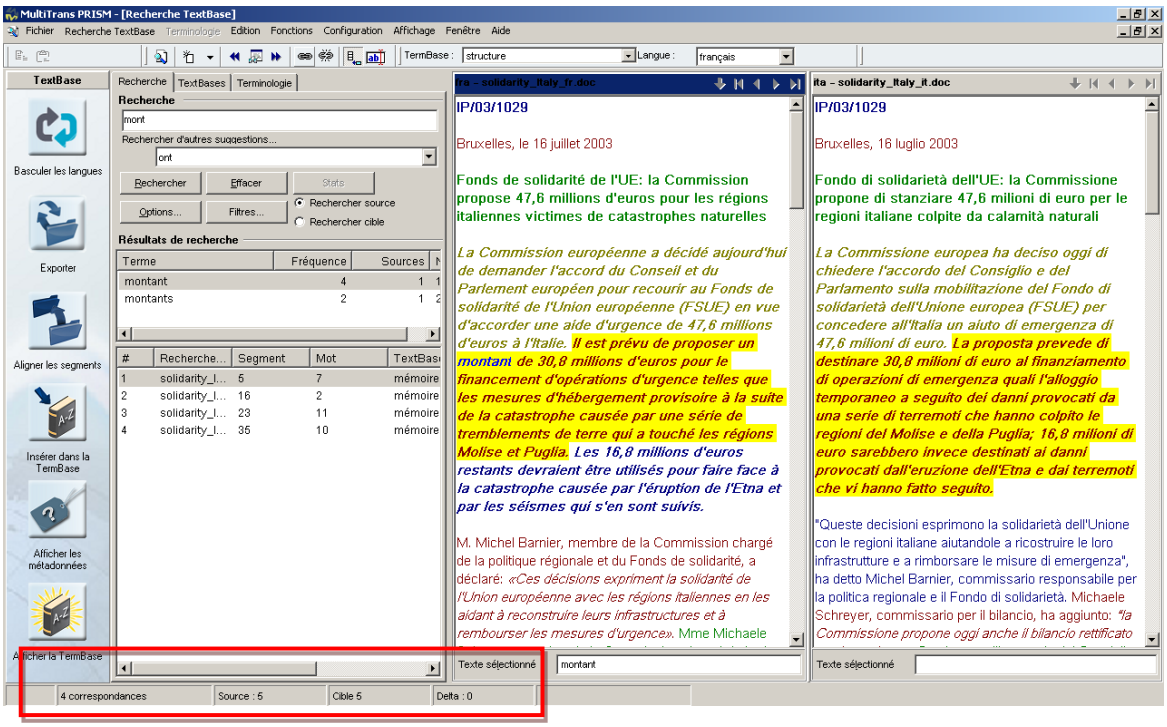


Figure 3.7 : *MultiTrans Prism, TextBase,* recherche par radicaux

2.1.4. Autres fonctions

Il est possible de sélectionner un terme et sa traduction et de les introduire dans la base terminologique (*TermBase*).

Sous l'onglet « *Terminologie* », la liste des mots est disponible par ordre de fréquence ou alphabétique.

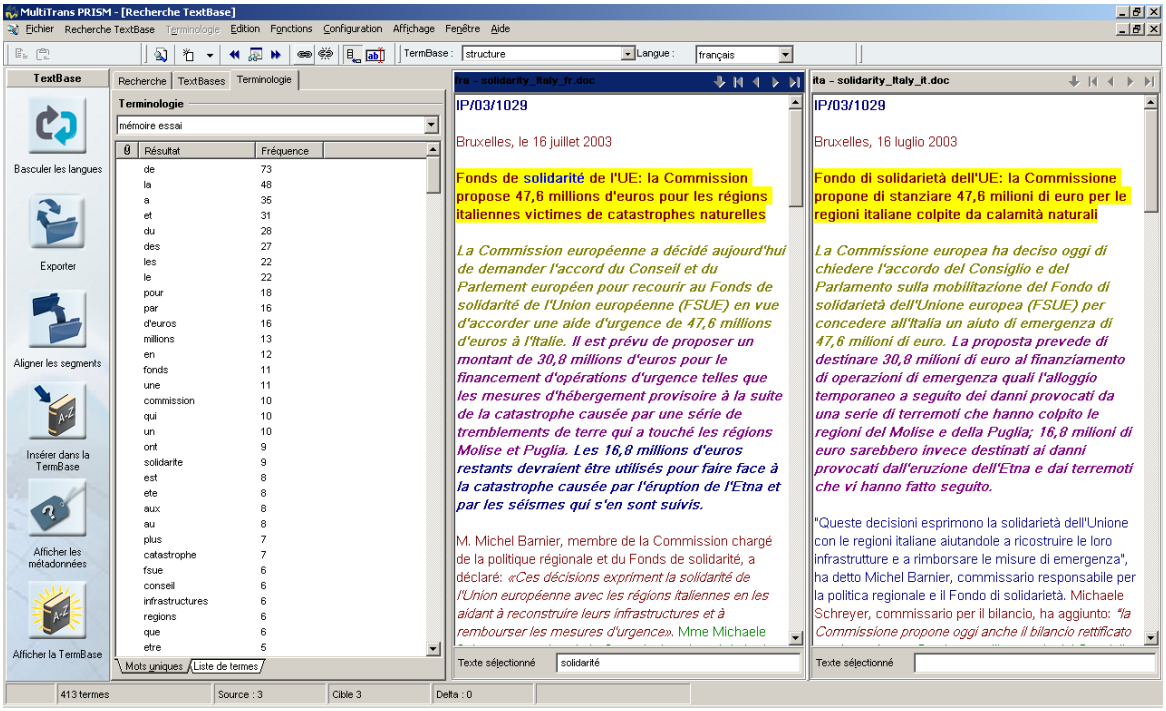


Figure 3.8 : *MultiTrans Prism, TextBase*, liste des mots par ordre de fréquence

2.2. ParaConc

ParaConc est un concordancier parallèle multilingue développé par la société américaine Athelstan[22]. Ce logiciel est conçu pour l'analyse contrastive, l'enseignement des langues et l'étude et enseignement de traduction. Les acheteurs principaux de ce logiciel sont des enseignants ou des chercheurs.

2.2.1. Caractéristiques techniques

ParaConc est payant et coûte 89 USD. Il est possible de le télécharger directement sur le site, ou de demander une copie sur CD avec un manuel imprimé. C'est un logiciel mono-utilisateur et monoposte, mais on peut acheter plusieurs licences à la fois.

Nous testons ici *ParaConc- Beta Version 1.0 (Build 269).*

[22] http://athel.com/product_info.php?products_id=81&osCsid=95cd6a3d58b331ddc04a162c383fa3d2 - Consulté le 23/05/2014

ParaConc est compatible avec les systèmes d'exploitation Windows 32 bit à partir de la version 95, nécessite de 2 à 20 Mo d'espace sur le disque dur pour gérer les corpus et de 16 à 32 Mo de RAM.

Il accepte uniquement les fichiers au format texte. Il supporte 35 langues qui sont principalement des langues européennes ainsi que l'afrikaans, le turc et le russe. Il ne peut pas être utilisé facilement avec le chinois, le japonais et le coréen. Il est capable de gérer jusqu'à quatre textes à la fois, ce qui veut dire qu'il est possible d'analyser quatre langues différentes ou le texte source et trois traductions dans la même langue (Barlow, 2004 : 1).

2.2.2. Codage et alignement

ParaConc gère aussi bien les corpus annotés que les corpus non annotés en donnant la possibilité d'ajouter des étiquettes pour la segmentation (Barlow, 2004 : 2).

L'alignement est semi-automatique et se fait au niveau des paragraphes, des segments ou des étiquettes. Pour les corpus non annotés, l'alignement par phrase se base sur l'algorithme de Gale et Church (Barlow, 2004 : 2), dont nous avons parlé dans le chapitre 2 à la section 2.2.1.3.

Les options d'alignement doivent être sélectionnées au moment du chargement des textes. Il est possible de choisir entre :

- *non aligned*, pour les textes qui ne sont pas alignés précédemment ;
- *new line delimiter,* qui segmente les textes par rapport aux retours à la ligne ;
- *delimiter* qui permet d'introduire dans les options des symboles au choix pour délimiter les segments ;
- *start/stop tags* qui permet de gérer l'alignement avec les tags.

La figure 3.9 montre le résultat de l'alignement. Comme on peut voir, si les textes donnés en entrée ne sont pas prétraités ou ne contiennent pas le même nombre de phrases, la qualité de l'alignement n'est pas satisfaisante : un grand nombre de segments n'a aucune correspondance et d'autres ne semblent pas correctement alignés. Il est donc nécessaire de modifier manuellement l'alignement à l'aide d'options très basiques[23], ou de prétraiter les textes avant de les charger dans *ParaConc*. Le taux d'erreur est influencé négativement par le fait qu'il n'y a aucun moyen d'introduire des listes d'exclusions pour la segmentation.

[23] Il est possible de diviser une phrase ou un segment, de fusionner une phrase ou un segment avec la phrase ou le segment suivant ou précédent, d'ajouter un segment vide ou d'effacer la dernière modification.

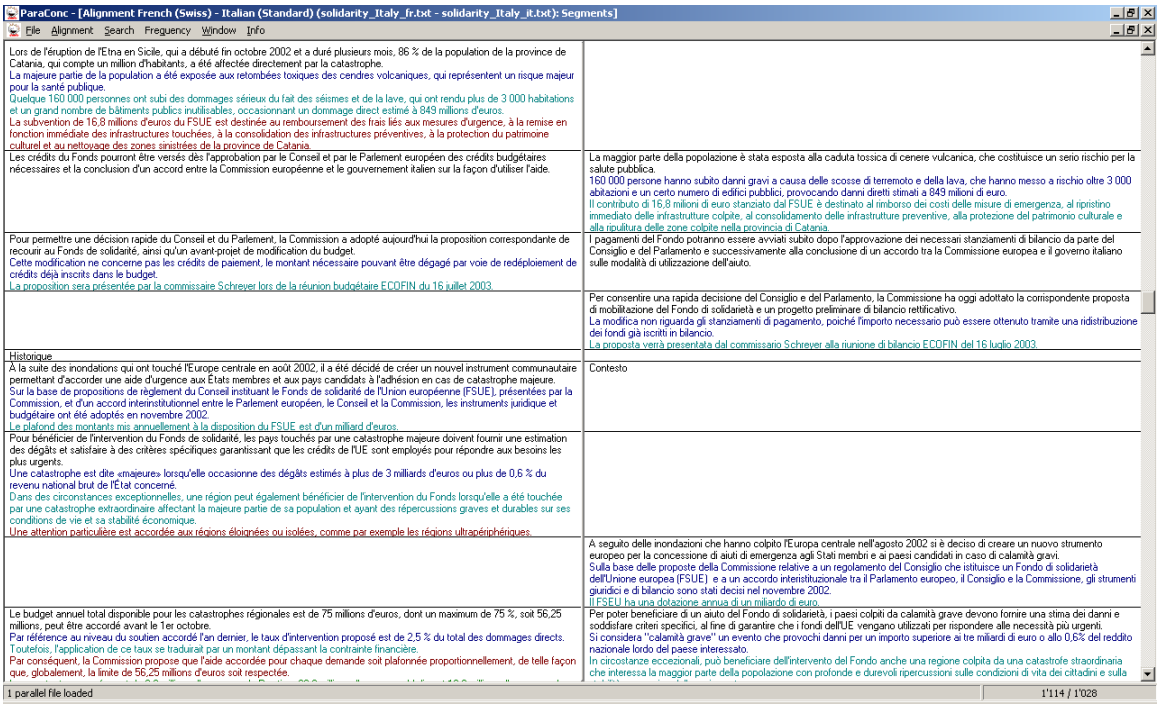


Figure 3.9 : L'alignement dans *ParaConc*.

En bas à gauche s'affiche le nombre de documents chargés et à droite le nombre de mots contenus dans les deux corpus (Barlow, 2004 : 3).

2.2.3. Recherche

Les recherches peuvent être effectuées sur les deux langues. Il est possible de choisir entre les options suivantes :

- ***Simple text search*** **:** permet d'effectuer des recherches simples de mots ou de groupes de mots en utilisant au besoin des métacaractères que l'utilisateur peut personnaliser (Barlow, 2004 : 3) ;
- ***Regular expression search*** : permet d'effectuer des recherches avec des expressions régulières ;
- ***Tag search*** : permet d'effectuer des recherches sur les corpus étiquetés et elle peut être combinée avec la « *Textsearch »* ;
- ***Parallel Search*** **:** permet de lancer une recherche sur les deux langues en même temps, c'est-à-dire de chercher une séquence dans la langue source et une autre séquence dans la langue cible et d'afficher seulement les occurrences dans lesquelles le mot A en langue source correspond au mot B dans la langue cible. Si la case « Not » est cochée, *ParaConc* affichera toutes les occurrences dans lesquelles le mot B n'apparaît pas.

Les options de recherche « *Simple text search* », « *Tag search* » et « *Parallel Search* » admettent les métacaractères suivants :

- * n'importe quel nombre de caractères alphanumériques ;
- % zéro ou un caractère (par exemple avec chat% on obtient chat ou chats) ;
- ? n'importe quel caractère.

Les résultats de la recherche sont affichés en KWIC pour la langue source et par segment dans la langue cible. La fonction « *Hot Words »* suggère des traductions potentielles en se basant sur la fréquence des mots (Barlow, 2004 : 4). On peut choisir un des mots suggérés ou en introduire un nouveau. Une fois que cela est fait, il est possible de sélectionner l'affichage KWIC pour les occurrences de la langue cible. La fonction *« Short »* permet de paramétrer les contextes à droite et à gauche du patron.

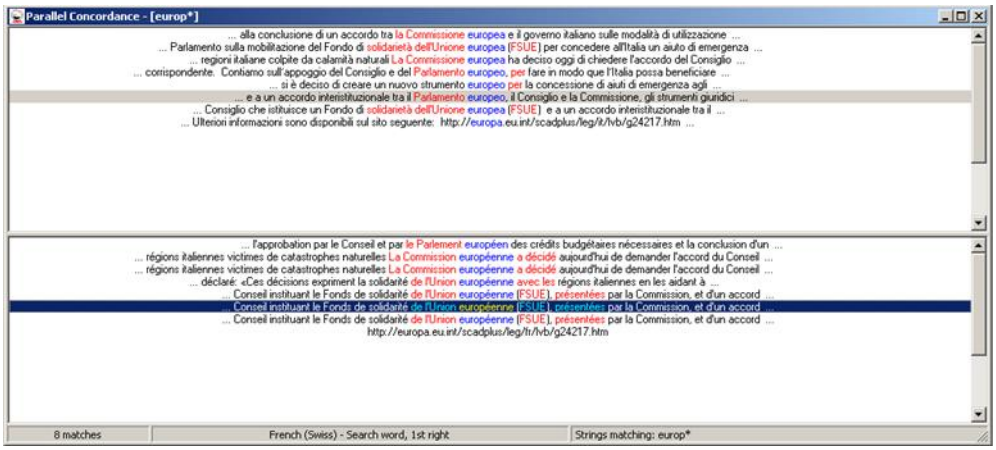


Figure 3.10 : Recherche simple avec suggestion de traduction et affichage KWIC pour les deux langues

2.3.4. Autres fonctions

Lorsqu'une fenêtre de recherche est active, on peut afficher une liste des collocations du patron ou les clusters. Comme on peut le voir à la figure 3.11, *ParaConc* montre les collocations par rapport à la position du terme (contexte à droite ou à gauche). Pour chaque langue, il est possible de modifier les paramètres du contexte et de la fréquence sous « *Frequency Options* ».

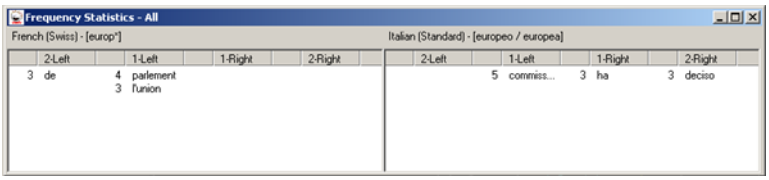


Figure 3.11 : Liste des collocations des termes « européens » et « européenne »

Les clusters sont accessibles à partir de l'onglet « *Frequency* » à travers l'option « *Advances collocations* ». Le nombre d'unités et le contexte sont paramétrables.

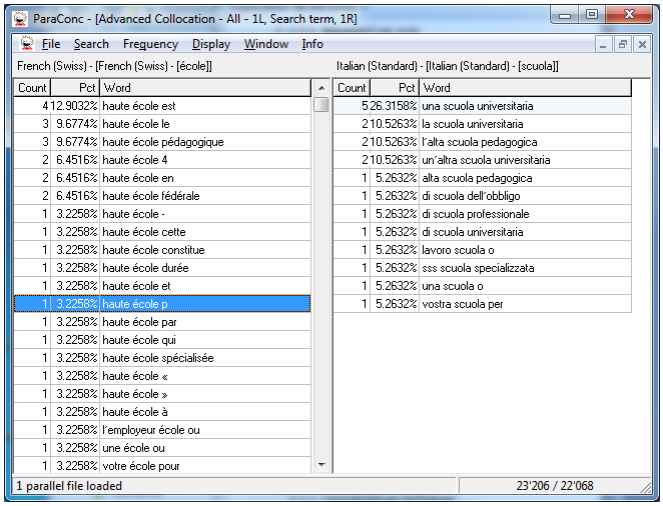


Figure 3.12 : *ParaConc,* cluster de l'unité « école »

Sous l'onglet « *Frequency* », il est possible d'afficher une liste de mots (« *Frequency List* ») par ordre alphabétique ou de fréquence. L'option « *Stop List* », permet d'indiquer les mots à exclure de cette liste.

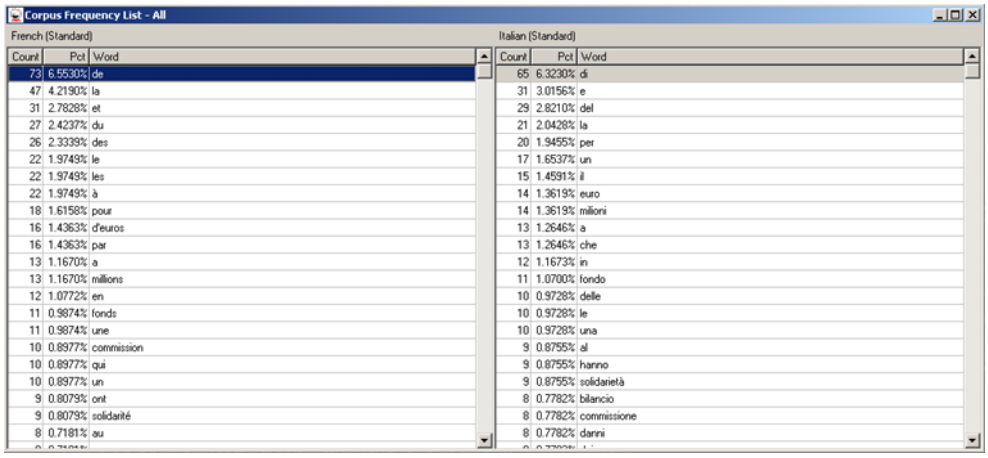


Figure 3.13 : *ParaConc*, Frequency List

2.3. myCAT

Le logiciel *myCAT* est un concordancier parallèle open source développé par la société Olanto[24], fondation suisse à but non lucratif dont la mission est de créer et distribuer des logiciels libres dans les domaines de la traduction assistée par ordinateur (TAO), de la traduction automatique (TA), de la recherche multilingue et d'autres domaines liés aux langues.

2.3.1. Caractéristiques techniques

L'outil *myCAT* présente une architecture client-serveur multi-utilisateurs. Il est installé sur une machine virtuelle et est accessible depuis n'importe quel ordinateur à travers une interface web. Ce logiciel n'est donc pas conçu pour une utilisation monoposte, mais pour les entreprises.

Il est compatible avec les navigateurs *Mozilla Firefox* ou *Internet Explorer*. Bien que *myCAT* soit développé avec la technologie GWT de *Google*, le logiciel rencontre quelques problèmes avec *Chrome* pour la partie appelée « *Quote Detector* » dont nous parlons à la section 2.3.4.

Il peut être installé sur un serveur avec les systèmes d'exploitation Windows 7, Windows 2008 Server, version 64 bits, sur un serveur avec *Ubuntu 12.04 Desktop*, version 64 bits. Il nécessite de 6 à 8 Go de RAM et tourne à l'aide des logiciels suivants *: Java JRE 6* ou *7, Apache Tomcat 6.0* et *OpenOffice 3*. Il supporte des fichiers PDF, *Microsoft Office* ou *OpenOffice*.

La version de base est fournie avec les cartes d'alignement suivantes : anglais-arabe, anglais-espagnol, anglais-français, anglais-portugais et anglais-russe. Il est possible de télécharger à partir du site les cartes d'alignement anglais-allemand et anglais-chinois. Théoriquement, *myCAT* est en mesure de gérer n'importe quelle combinaison linguistique, il suffit d'entraîner le logiciel et d'introduire de nouvelles cartes d'alignement.

En mars 2014, la société Olanto a publié une nouvelle version de *myCAT*, la version 3.2.2[25]. Nous testons ici la version précédente publiée en 2013. Pour la partie concordancier (*Text Aligner*), les différences entre ces deux versions concernent la possibilité de redimensionner l'interface afin de l'adapter à la fenêtre du navigateur de recherche, la possibilité de choisir la langue de l'interface et l'indication du total des occurrences contenues dans le document

[24] http://olanto.org/fr/fondation - Consulté le 23/05/2014
[25] http://olanto.org/fr/logiciels/mycat - Consulté le 23/05/2014

sous « *Hit List* ». Il s'agit donc de petites améliorations qui ne changent pas le fonctionnement de base du logiciel.

2.3.2. Codage et alignement

Tous les fichiers sont transformés automatiquement au format texte afin d'effacer la mise en page. À chaque fin de phrase est ajouté un retour à la ligne. Ces textes sont ensuite segmentés et alignés par phrases.

L'alignement est automatique et repose sur des cartes d'alignement qui sont construites avec *Moses*[26]. Le concordancier *myCAT*, en se fondant sur le modèle de traduction par phrase de *Moses*, fait une traduction automatique de la phrase qu'il utilise ensuite pour retrouver la phrase correspondante dans le texte cible. Le logiciel fait donc une comparaison entre la phrase traduite par *Moses* et toutes les phrases contenues dans le texte cible et établit un lien entre les deux phrases qui ont le plus de probabilités de correspondre par rapport au nombre de mots équivalents. Une fois cette correspondance établie, il garde le lien entre les deux phrases et efface la traduction automatique qui n'est pas nécessaire pour le reste du fonctionnement du logiciel.

Cette méthode d'alignement permet donc de surmonter les problèmes liés à l'ordre des phrases dans un texte ainsi qu'aux suppressions ou aux adjonctions.

Pour que ce système d'alignement fonctionne, il faut que le logiciel soit entraîné avec des corpus d'entraînement, afin qu'il soit capable de faire des statistiques correctes et de définir, par rapport au contexte du mot (qui se limite à la phrase), quel est le meilleur choix d'alignement dans le texte cible.

Si aucune carte d'alignement n'est disponible pour la paire de langues choisie, *myCAT* utilise un algorithme d'alignement géométrique simple.

Une fois que cet alignement est terminé, les textes sont indexés, à l'exception des articles, des prépositions et des mots très courts, comme, par exemple, certaines abréviations composées de deux ou trois lettres.

[26] *Moses* est un outil open source de traduction automatique statistique développé par Hoang et Koehn à l'Université d'Édimbourg. Il permet de gérer n'importe quelle paire de langue, à condition qu'il soit d'abord entraîné par un corpus parallèle. Une fois que le système contient un modèle d'entraînement, un algorithme se charge de trouver la traduction la plus probable parmi le nombre exponentiel de choix (Koehn, 2014 : 11).

2.3.3. Recherche

Text Aligner est l'interface de recherche. À côté de la case de recherche, il est possible de sélectionner un couple de langue indiqué selon les codes de la norme ISO-639-1. Sous « *Collections OFF »,* il est possible de sélectionner les sous-corpus que l'on veut utiliser. L'ordre de préférence des corpus est établi selon l'ordre de sélection. Si aucune collection n'est sélectionnée, *myCAT* travaillera sur l'ensemble des documents disponibles.
La recherche se fait par mots ou par groupe de mots. La séquence recherchée est en surbrillance dans le texte source et le segment est aligné en plein texte dans le texte cible.

Le nombre total d'occurrences trouvées est affiché sous « *Hit List »* et, en bas à droite, on peut voir à quelle occurrence nous sommes dans le texte. En d'autres termes, pour connaître le nombre exact d'occurrences dans un seul texte, il faut arriver à la dernière[27].

En cliquant sur le bouton « *Original »*, il est possible de voir le document dans son format d'origine.

Le concordancier *myCAT* ne fait aucune différence entre majuscules et minuscules, mais ne recherche que les mots exacts. Ce qui veut dire que pour obtenir un nom au singulier et au pluriel il faut utiliser un astérisque. L'astérisque peut remplacer un ou plusieurs caractères, mais ne peut être utilisé que pour la recherche d'un seul mot.
Il est possible de chercher un document en indiquant /*
Trois opérateurs booléens sont disponibles :

- AND : il permet de chercher deux unités et il est implicite lorsqu'aucun opérateur booléen n'est indiqué ;
- OR : il permet de chercher deux unités qui ne figurent pas ensemble dans un document, comme par exemple les formes fléchies d'un terme ou des synonymes ;
- NEAR : il effectue une recherche de cooccurrences. La deuxième unité peut se trouver au maximum 5 mots après. Les mots non indexés ne sont pas comptés.

[27] Dans la version 3.2.2 de myCAT, le nombre total d'occurrences dans un document est indiqué sous « Hit List ».

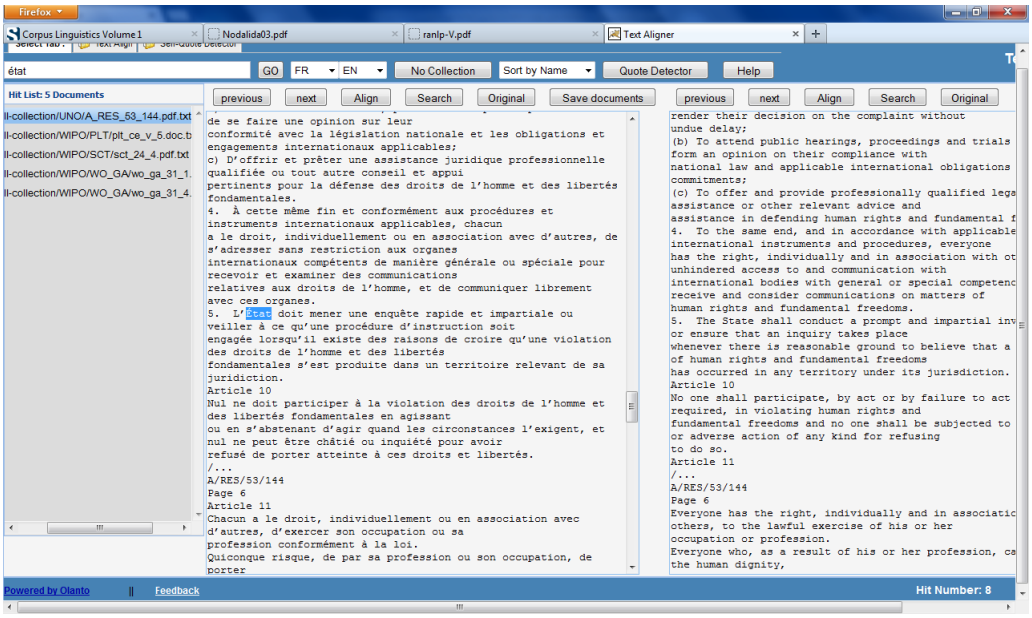


Figure 3.14 : Aperçu de *Text Aligner*

2.3.4. Autres fonctions

En plus du *Text Aligner, myCAT* dispose de deux autres parties.

La première est le *Quote Detector* qui compare le texte à traduire avec les documents contenus dans le corpus et surligne en jaune les parties déjà traduites. En cliquant dessus, on peut voir l'alignement des versions source et cible du document de référence, comme le montre la figure 3.15.

Quote Detector affiche également une statistique, notamment du total de mots, du nombre de mots déjà traduits et du nombre de textes de référence trouvés.

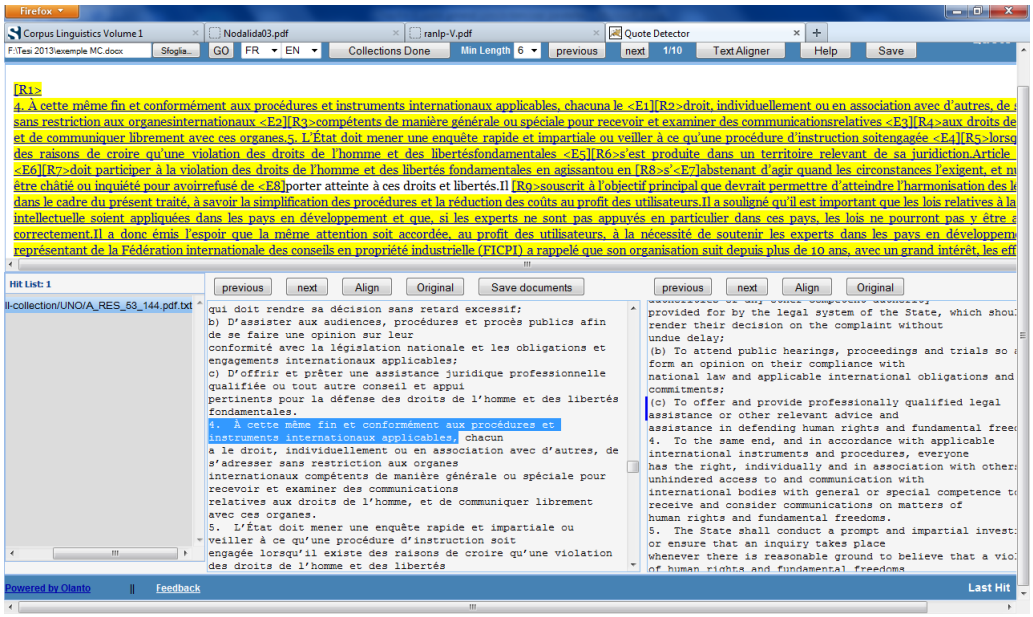


Figure 3.15 : Recherche dans *Quote Detector* de *myCAT* des parties de texte déjà traduites

La troisième partie est le *Self-Quote Detector* qui détecte les séquences d'unités qui se répètent le plus souvent dans le texte. Cette séquence est paramétrée par l'utilisateur et elle peut contenir entre 3 et 9 unités. Le logiciel *myCAT* affiche ensuite un tableau statistique indiquant le nombre d'occurrences de la chaîne de caractères et d'autres informations.

3. Grille d'analyse des trois outils

Nous venons de présenter les trois outils dont nous nous servons pour la tâche d'évaluation dans les prochains chapitres : la MT de *MultiTrans Prism*, ainsi que les concordanciers *myCAT* et *ParaConc*. Nous pouvons mettre en relation ces outils à travers six aspects : la famille, la conception technique, le codage du texte, l'alignement, la recherche et autres fonctions. Nous les présentons un à la fois en mettant en évidence les différences entre nos trois outils. À suivre, notre grille d'analyse qui réunit tous ces aspects.

3.1. Famille

Cet aspect indique si l'outil est un logiciel spécifique en commerce, un logiciel mandaté, donc développé pour un client spécifique, ou un logiciel open source, c'est-à-dire un logiciel libre.

Comme nous décrivons dans la section 2, *MultiTrans Prism* et *ParaConc* sont des outils spécifiques en commerce, tandis que *myCAT* est un outil open source.

3.2. Conception technique

La conception technique rassemble les aspects techniques des outils. Tout d'abord, nous distinguons le nombre d'utilisateurs, entre mono-utilisateur, si le logiciel peut être utilisé par une seule personne à la fois, et multi-utilisateurs, si plusieurs personnes peuvent l'utiliser en même temps. Étant un outil qui existe en plusieurs versions, *MultiTrans Prism* est à la fois mono-utilisateur et multi-utilisateurs, *ParaConc* est mono-utilisateur, alors que *myCAT* est multi-utilisateurs.

La deuxième distinction concerne l'architecture qui peut être monoposte (installé sur un ordinateur), client-serveur (installé sur un serveur interne ou externe à l'entreprise qui l'utilise) et interface web (accessible depuis un navigateur web). *MultiTrans Prism* est disponible dans les trois versions. *ParaConc* est exclusivement monoposte, alors que *myCAT* est disponible uniquement en version client-serveur.

La troisième distinction concerne les caractéristiques techniques. Nous avons vu que *MultiTrans Prism* et *ParaConc* sont compatibles avec les systèmes d'exploitation *Microsoft*

Windows 32 bits, alors que *myCAT* est compatible avec *Windows 7 64 bits, Windows 2008 Server 64 bits* et *Ubuntu 12.04 Desktop 64 bits*.

Concernant les formats de fichiers supportés, *ParaConc* n'accepte que les fichiers au format texte, *myCAT* accepte les PDF ainsi que les fichiers des paquets *Microsoft Office* et *OpenOffice*, alors que *MultiTrans Prism* accepte au total 17 formats de fichiers.

Au niveau des langues supportées, le plus faible semble être *ParaConc* avec 35 langues, alors que les deux autres logiciels sont en mesure de toutes les supporter.

Enfin, concernant le fonctionnement de base, *MultiTrans Prism* et *myCAT* sont des outils statistiques, alors que *ParaConc* est hybride.

3.3. Codage du texte

Le codage du texte indique la manière dont les corpus sont stockés dans le logiciel. *MultiTrans Prism* et *myCAT* indexent les unités, alors que *ParaConc* permet d'annoter ou de lemmatiser les corpus.

3.4. Alignement

L'aspect de l'alignement est repris tels que décrit dans le chapitre 2 (section 2). *MultiTrans Prism* et *myCAT* disposent d'un alignement automatique, tandis que *ParaConc* dispose d'un alignement semi-automatique.

Concernant le niveau d'alignement, nous avons vu que les trois alignent au niveau de la phrase, mais que *ParaConc* utilise l'alignement par longueur de phrases. Quant à *myCAT*, il utilise pour certaines langues des cartes d'alignement qui sont construites avec *Moses* et, pour les autres langues, pour lesquelles ces cartes d'alignement ne sont pas disponibles, il fait un alignement géométrique. *MultiTrans Prism* dispose d'un alignement statistique.

3.5. Recherche

MultiTrans Prism permet d'effectuer des recherches par mot, par troncature et par *fuzzy match*. Il affiche les concordances en plein texte et propose une liste de mots.

ParaConc propose des recherches par mot avec ou sans joker, par troncature, avec l'opérateur OU, par cooccurrent, par expressions régulières ainsi que des recherches parallèles. Il dispose d'un affichage KWIC avec possibilité de voir l'occurrence sélectionnée en plein texte dans le document d'origine. Il permet également d'obtenir des listes de mots et des collocations.

Le logiciel *myCAT* permet d'effectuer des recherches par mot, par mot avec joker (mais seulement avec une unité), par troncature, par opérateur OU et par cooccurrent. L'affichage est en plein texte avec la possibilité de voir le document dans son format d'origine.

Nous approfondissons l'analyse de l'aspect de la recherche dans les chapitres 4 et 5 car cet aspect est au centre de notre évaluation (voir section 1 du chapitre 4).

3.6. Autres fonctions

Parmi les autres fonctions, *MultiTrans Prism* dispose, dans l'interface de recherche, d'une fonction permettant de sélectionner et d'ajouter des termes dans la *TermBase*. Le concordancier *myCAT* est composé de deux autres parties permettant d'analyser le document à traduire : *Quote Detector* qui compare le texte à traduire avec les documents contenus dans le corpus et surligne en jaune les parties déjà traduites et *Self-Quote Detector* qui détecte les séquences de mots qui se répètent le plus souvent.

Nous présentons ci-dessous la grille qui résume les aspects tels que décrits dans cette section.

				MT Multi Trans Prism	ParaConc	myCAT
1. Famille	1. logiciel spécifique en commerce			X	X	
	2. logiciel spécifique mandaté					
	3. logiciel open source					X
2 .Conception techniques	2.1. nombre d'utilisateurs	mono-utilisateur		X	X	
		multi-utilisateurs		X		X
	2.2. architecture	monoposte		X	X	
		client-serveur		X		X
		interface Web		X		
	2.4. caractéristiques techniques	OS		Windows XP, Vista ou Windows 7	Windows 95 et versions suivantes	Windows 7 64 bits,Windows 2008 Server 64 bits, Ubuntu 12.04 Desktop 64 bits
		format de fichier supporté		doc, docx, txt, rtf, ppt, pptx, xlsx, xml, html, pdf, wpd,INX, DITA, MIF, TMX, XLIFF et RTF	txt	PDF, Microsoft Office, OpenOffice
		langues supportées		Toutes	35	Toutes
		fonctionnement de base	statistique	X		X
			linguistique			
			hybride		X	
3. Codage du texte	3.1 indexation			X		X
	3.2 annotation				X	
	3.4 lemmatisation				X	
4. Alignement	4.1. Type d'alignement		manuel			
			semi-automatique		X	
			automatique	X		X
	4.2. Niveau d'alignement	1.2.1. mot	technique avancée			
		1.2.2. phrase	ancrage lexical			
			longueur des phrases		X	
			cognats			
			alignement géométrique			X
			autres	X		X
5. Recherche	5.1. Modalités de recherche	5.1.1. Recherche simple	recherche par mot	X	X	X
			recherche par mot avec joker		X	oui, mais seulement avec un mot
			recherche par trocature	X	X	X
		5.1.2. Recherche avancée	opérateur OU		X	X
			cooccurrent		X	X
			recherche parallèle		X	
			recherche par filtre			
		5.1.3. Recherche spécifique	recherche par fuzzy	X		
			recherche par expressions régulières		X	
	5.2 Affichage des données	5.2.1. Concordances	affichage plein texte	X		X
			affichage KWIC		X	
			possibilité de voir le document d'origine		X	X
	5.3. Recherche d'informations spéc.	5.3.1. Liste de mots		X	X	
		5.3.2. Cluster			X	
		5.3.3. Collocations			X	
6. Autres fonctions				Ajouter les mots dans TermBase		Quote Detector
						Self-Quote Detector

Chapitre 4 : Présentation de la méthode et des critères d'évaluation

Dans ce chapitre, nous proposons une méthode d'évaluation des outils d'exploitation de corpus parallèles, notamment en ce qui concerne l'aspect de la recherche dans les textes, dans le but de comprendre si les caractéristiques de ces outils répondent réellement aux besoins du traducteur et dans la perspective de définir quelles sont les fonctions indispensables afin d'améliorer les outils existants.

Nous commençons donc par définir le but de l'évaluation (section 1), puis nous définissons ce qu'est une bonne évaluation (section 2). Dans la section 3, nous présentons notre méthode d'évaluation et nous passons ensuite à la définition des exigences de qualité (section 4) et à la préparation de l'évaluation (section 5).

1. But de l'évaluation

Le but de cette évaluation est celui de voir dans la pratique si, en l'état actuel, les outils d'exploitation de corpus sont en mesure d'apporter à la traduction les avantages que nous avons vus dans la théorie (voir section 3 du chapitre 1), c'est-à-dire s'ils sont réellement adaptés aux besoins des traducteurs.

Afin de limiter cette étude à un travail de Master, nous concentrons notre évaluation des outils uniquement sur la partie de la recherche des occurrences, tout en sachant que chacun de ces logiciels possède d'autres fonctionnalités, comme nous l'expliquons dans le chapitre précédent.

Nous voulons comparer les fonctions de recherche des concordanciers parallèles par rapport à celles des mémoires de traduction pour comprendre lequel de ces outils permet de mieux cibler la recherche et lequel dispose d'un meilleur affichage des résultats et des informations. Nous voulons également comprendre lesquelles de ces fonctions et de ces affichages sont effectivement utiles au traducteur et de quelle manière il serait possible de les améliorer.
Pour répondre à ces questions, nous comparons deux concordanciers parallèles (*ParaConc* et *myCAT*) à la mémoire de traduction de *MultiTrans Prism.*

2. Critères d'une bonne évaluation

Pour construire une bonne méthode d'évaluation, il faut tout d'abord comprendre ce qu'est une évaluation. Le rapport final du projet EAGLES (1996 : 15) la définit ainsi :

> « To *evaluate* is to determine what something is worth to somebody. »

Cette définition donne la notion d'utilité de quelque chose pour quelqu'un. Si l'on exprime le même concept dans des termes d'évaluation de logiciel, on pourrait dire qu'évaluer signifie définir dans quelle mesure un logiciel est utile à un utilisateur. C'est-à-dire dans quelle mesure un logiciel est capable de répondre aux besoins exprimés ou implicites de l'utilisateur (ISO/EIC 9126 : 2000, 2000 : 3).

En accord avec la norme ISO/IEC 9126 : 1991 (1991) et le rapport final du projet EAGLES (1996), nous déduisons que les besoins des utilisateurs dépendent tout d'abord de la catégorie d'utilisateurs et du contexte d'utilisation. Les outils d'exploitation de corpus peuvent être utilisés par différentes catégories d'utilisateurs, par exemple un développeur ou un traducteur (utilisateur final). Chacun l'utilise de manière différente et pour des raisons différentes. Avant de commencer une évaluation, il est donc essentiel de définir la catégorie d'utilisateurs, pour pouvoir ensuite identifier ses besoins par rapport au contexte d'utilisation.
Dans ce travail, nous décidons de tester nos trois outils en adoptant le point de vue du traducteur, donc de l'utilisateur final du logiciel. Notre évaluation est donc orientée vers l'utilisateur.

Un autre point fondamental est représenté par la définition des critères d'évaluation et par l'attribution des échelles d'évaluation. Pour qu'une évaluation soit valable, il faut éviter la subjectivité.
Höge (2002 : 70) et les auteurs du projet EAGLES (1996 : 14) concordent sur le principe que les critères d'évaluation ainsi que l'attribution des métriques doivent donner lieu à des résultats valables et répétables. D'après le projet EAGLES (ibidem), cette validité doit être interne (basés sur les contenus) et externe (basée sur les critères). La validité interne, soumise au jugement de l'expert, est obtenue en veillant à ce que toutes les métriques évaluent correctement les attributs de l'objet à évaluer (ibid.).

Une métrique est répétable si on obtient les mêmes résultats lorsqu'on l'applique aux mêmes phénomènes et elle peut être obtenue avec un coefficient de corrélation entre les résultats obtenus dans deux essais en appliquant la même métrique (ibid.).

Nous construisons donc notre méthode d'évaluation en respectant ces lignes directrices et en nous basant sur des travaux existants.

3. Méthode d'évaluation

Afin de construire une méthode d'évaluation objective, nous nous basons sur le rapport final du projet EAGLES (1996). Ce dernier décrit des méthodes permettant d'évaluer les systèmes de traitement automatique des langues, sans pour autant en définir les critères.
En partant de la norme ISO/IEC 9126 : 1991 (ISO, 1991), les auteurs (EAGLES, 1996 : 12-13) suggèrent les étapes principales à suivre lors de la mise en place d'une méthode d'évaluation. Ces étapes sont : la définition des exigences de qualité, la préparation de l'évaluation et l'exécution de l'évaluation.

La première étape consiste à analyser le contexte de l'évaluation et la relation entre l'utilisateur et le logiciel afin de définir un ensemble de caractéristique et de sous caractéristiques de qualité que doit respecter le logiciel afin de répondre aux besoins exprimés ou implicites de l'utilisateur (ISO/IEC 9126 : 1991, 1991). Nous en parlons dans les détails dans la section suivante.

La phase de préparation de l'évaluation consiste à définir les critères d'évaluation sur la base des caractéristiques et des sous-caractéristiques issues de l'étape précédente, auxquelles on assigne des métriques, ainsi qu'à préparer le matériel nécessaire à l'évaluation. Nous en parlons dans la section 5.

Enfin, l'exécution de l'évaluation rassemble l'exécution de la tâche ainsi que l'interprétation des résultats. À cette phase est dédié le chapitre 5.

4. Définition des exigences de qualité : adaptation de travaux existants

Dans cette partie, nous expliquons comment définir les exigences de qualité qui sont à la base des critères d'évaluation. Pour mieux comprendre comment est structurée notre approche, nous parlons tout d'abord de l'évaluation dans le domaine du génie du logiciel en expliquant brièvement la théorie du modèle de tâche et le concept d'exigences de qualité tel

que décrit dans la norme ISO/IEC 9126 : 1991 (1991). Ensuite, nous retournons au domaine de la traduction avec l'étude de Höge (2002) qui a appliqué la norme ISO/IEC 9126 : 1991 au modèle de tâche en définissant ainsi des caractéristiques de qualité qui sont à la base de nos critères d'évaluation.

Avant de continuer, il est nécessaire de faire une précision concernant la norme ISO/IEC 9126 : 1991. Dans le cadre de ce travail, nous décrivons la norme ISO/IEC 9126 : 1991 de 1991 qui est à la base du rapport final du projet EAGLES (1996) ainsi que de l'étude de Höge (2002). Or, cette norme a été d'abord mise à jour en 2000 et subdivisée en deux nouvelles normes : la norme ISO/IEC 9126 (Software product quality) et la norme ISO/IEC 14598 (Software producte valuation) (ISO/IEC 9126, 2000 : v), puis remplacée et enrichie par la norme ISO/IEC 25010 : 2011 (2011 : iv).
Bien que cela puisse être contestable, nous choisissons de fonder nos critères d'évaluation sur la norme ISO/IEC 9126 : 1991 de 1991, dans la mesure où cette dernière est à la base de l'étude de Höge de laquelle nous nous inspirons, car parfaitement adaptable à notre cas d'étude.

Les travaux cités soulignent qu'il est nécessaire de définir et structurer le contexte d'évaluation avant de passer à la définition des exigences de qualité. Ayant déjà fixé au préalable que notre évaluation est basée sur la tâche de recherche dans un outil d'exploitation des corpus, nous n'approfondissons pas le sujet et focalisons notre attention sur les méthodes de définition des critères d'évaluation. Nous dédions simplement quelques lignes au modèle de tâche (*generic task modelling*), théorie issue du domaine du génie du logiciel, pour mieux comprendre l'approche de Höge fondée sur cette théorie.

4.1. Évaluation dans le domaine du génie du logiciel (Software Engineering)

Le génie du logiciel rassemble toutes les activités qui concernent la conception, le développement et la production des logiciels ainsi que leur suivi, y compris l'évaluation.
Comme l'explique Höge (2002 : 60), dans le génie du logiciel, le but de l'évaluation est celui de détecter et de définir les propriétés dont doivent disposer les logiciels afin d'être acceptés par les clients. En faisant une comparaison avec notre étude, nous pouvons observer que le but de l'évaluation est différent, car nous souhaitons évaluer dans quelle mesure le logiciel est capable de satisfaire les exigences de l'utilisateur. Compte tenu de

cette observation, nous montrons au cours de ce chapitre que l'évaluation dans le domaine du génie du logiciel s'adapte à l'évaluation des outils d'exploitation de corpus.

Nous décrivons ici, dans les grandes lignes, les principes fondamentaux du modèle de tâche, qui est l'une des théories à l'origine de la méthode de définition des caractéristiques de qualité de Höge, puis nous parlons de la définition des exigences de qualité au sein de la norme ISO/IEC 9126 : 1991 (1991).

4.1.1. Modèle de tâche

Selon Caelen et Villaseñor (1997 : 89) :

> « l'objectif de base d'un modèle de tâche est de représenter et de permettre de contrôler la suite des actes d'un utilisateur devant réaliser une tâche. »

En d'autres termes, il s'agit d'identifier une série d'actions accomplies par un utilisateur lors du déroulement de la tâche.

Comme l'explique Höge (2002 : 83), dans le contexte du développement d'un logiciel, le but principal du modèle de tâche est celui de faciliter la transformation des exigences des utilisateurs en des primitives exécutables, à travers l'identification de fonctions.

Ce modèle permet donc de définir des tâches auxquelles sont associées des sous-tâches et des activités, c'est-à-dire une suite d'actions effectuées par l'utilisateur pour résoudre un problème (Caelen et Villaseñor, 1997 : 89).

Nous sélectionnons dans ce modèle deux concepts clés : *action* et *objet*. Selon Höge (2002 : 84), l'action est l'activité accomplie par une personne vers un objet spécifique qui peut être représenté par un objet physique (par exemple le clavier ou la souris) ou par un objet informatique (par exemple un logiciel ou un fichier).

En partant de cette théorie, nous pouvons affirmer que la recherche des occurrences représente l'action et que l'outil d'exploitation des corpus parallèles, dans ce cas l'interface de recherche, est l'objet sur lequel a lieu l'action.

Nous retrouvons ces deux concepts dans la méthode d'évaluation de Höge dont nous parlons à la section 4.2.

Comme nous venons de voir, le modèle de tâche permet d'identifier une série d'actions accomplies par l'utilisateur afin de résoudre un problème. La tâche relève donc de l'action. L'objet a un rôle passif dans l'action, mais il offre un ensemble d'attributs ou caractéristiques permettant à l'utilisateur d'accomplir la tâche. L'évaluation consiste donc à définir quelles sont ces caractéristiques dites de qualité qui rendent possible l'accomplissement de l'action.

4.1.2. Norme ISO/IEC 9126 : 1991

En 1991, dans le but d'uniformiser l'évaluation des logiciels, l'ISO (Organisation internationale de normalisation) a développé la norme ISO/IEC 9126 : 1991 qui définit les caractéristiques de qualité et donne les lignes directrices pour leur application.

Au sens de la norme ISO/IEC 9126 : 1991 (1991), le concept d'exigences de qualité indique un certain nombre de caractéristiques et de sous-caractéristiques qu'un logiciel doit respecter afin d'être considéré comme bon. Ces caractéristiques sont les suivantes :

- **Capacité fonctionnelle (*functionality*) :** attributs portant sur l'existence d'un ensemble de fonctions et leurs propriétés données. Ces fonctions sont celles qui permettent de satisfaire les besoins exprimés ou implicites. Les sous-caractéristiques sont : aptitude, exactitude, interopérabilité, conformité réglementaire et sécurité.
- **Fiabilité (*reliability*) :** attributs portants sur la capacité d'un logiciel à maintenir son niveau de performance à certaines conditions et pendant une période déterminée. Les sous-caractéristiques sont : maturité, tolérance aux fautes et possibilité de récupération.
- **Facilité d'utilisation (*usability*) :** attributs portant sur l'effort nécessaire pour l'utilisation et sur l'évaluation individuelle de cette utilisation, par un ensemble d'utilisateurs défini ou implicite. Les sous-caractéristiques sont : facilité de compréhension, facilité d'apprentissage et facilité d'exploration.
- **Rendement (*efficiency*) :** attributs portant sur le rapport entre le niveau de performance d'un logiciel et la quantité de ressources utilisées à des conditions déterminées. Les sous-caractéristiques sont : temps de réponse et débit, quantité de ressources nécessaires et durée d'utilisation.
- **Maintenabilité (*maintainability*) :** attributs portant sur la mesure de l'effort nécessaire pour faire des modifications données. Les sous-caractéristiques sont :

facilité d'analyse, facilité de modification, stabilité, robustesse, non régression et facilité de test.

- **Portabilité (*portability*) :** attributs portant sur l'aptitude d'un logiciel à être transféré d'un environnement à un autre. Les sous-caractéristiques sont : facilité d'adaptation, facilité d'installation, conformité aux règles de portabilité et d'interchangeabilité.

Ces caractéristiques représentent des lignes directrices à suivre lors de l'évaluation d'un logiciel. Tout d'abord, il faut identifier les caractéristiques et les sous-caractéristiques pertinentes avec l'outil et avec le but de l'évaluation. À partir de ces sous-caractéristiques, on définit ensuite les critères d'évaluation auxquels on attribue une échelle d'évaluation.

Bien que ces caractéristiques soient orientées vers le logiciel, nous montrons dans la section 5 qu'il est possible de les appliquer à nos tâches d'évaluation.

En accord avec Höge (2002 : 97), nous estimons que la maintenabilité ne joue pas un rôle central dans notre évaluation, car la modification d'un logiciel relève de la compétence du développeur et non pas de l'utilisateur du logiciel. Nous excluons également la portabilité car, bien que ce soit une caractéristique importante pour le choix d'un logiciel, elle ne rentre pas dans le cadre de notre évaluation, car centrée sur la recherche dans les textes.

4.2. Évaluation des outils d'aide à la traduction : l'étude de Höge

Dans son étude, Höge (2002 : 7-8) définit deux méthodes d'évaluation pour les outils d'aide à la traduction : la première est adressée à l'industrie de la traduction et a pour but de définir des critères pour l'évaluation d'un logiciel avant l'achat ; la deuxième a pour but de décrire des techniques et de donner des lignes directrices afin de fournir un support pour l'évaluation d'un logiciel dans la phase de développement.

Dans le cadre de ce mémoire, nous nous basons sur le premier type d'évaluation, c'est-à-dire l'évaluation pour l'achat d'un logiciel de TAO. Soulignons le fait que, dans notre étude, le but et le type de logiciel à évaluer ne sont pas identiques, car nous devons comparer des outils d'exploitation de corpus parallèles afin de comprendre s'ils répondent aux besoins des utilisateurs. Toutefois, les deux évaluations sont très proches pour deux raisons : premièrement, dans les deux cas, le domaine est bien la traduction et, parmi les outils que nous évaluons, *MultiTrans Prism* est un outil d'aide à la traduction ; deuxièmement, le fait

qu'un outil répond aux besoins d'un utilisateur est à la base du choix entre plusieurs logiciels au moment de l'achat.

Notre intérêt pour l'étude de Höge (2002) se focalise sur le procédé de définition des exigences de qualité. Le point de départ est l'évaluation dans le domaine du génie du logiciel. Höge reprend les caractéristiques de qualité telles que décrites dans la norme ISO/IEC 9126 : 1991 (1991) et les passe en revue une à la fois afin de définir lesquelles sont pertinentes à l'évaluation des outils de TAO. Elle considère que les caractéristiques de maintenabilité et de portabilité ne sont pas appropriées à une évaluation orientée vers l'utilisateur (ibidem : 100) et décide donc de prendre en compte quatre caractéristiques : capacité fonctionnelle, fiabilité, facilité d'utilisation et rendement.
Elle ajoute à la capacité fonctionnelle la sous-caractéristique de personnalisation qu'elle définit comme étant des attributs du logiciel qui permettent à l'utilisateur d'établir des contraintes spécifiques dans un système d'entrées (de données), traitement des entrées et ressource de données (ibid. : 90). Ensuite, elle applique ces caractéristiques de qualité à l'approche du modèle de la tâche, en particulier au concept d'action et d'objet (ibid. : 100). Elle obtient ainsi une liste d'aspects de qualité relatifs à l'action (figure 4.1) et une liste d'aspects de qualité relatifs à l'objet (figure 4.2). Ces listes étant très génériques, elles sont applicables à d'autres évaluations de logiciels orientées à l'utilisateur (ibid. : 105).

Dans l'étape de préparation de l'évaluation (section 5), nous partons de l'étude de Höge que nous venons de décrire, en particulier des listes d'aspects de qualité relatifs à l'action et à l'objet. Pour chacun de ces tableaux, nous sélectionnons les caractéristiques de qualité pertinentes à notre évaluation et nous les adaptons à notre cas d'étude afin de définir nos critères auxquels nous attribuons des échelles d'évaluation.

ASPECT RELATED TO **ACTIONS**	QUALITY CHARACTERISTIC
processes involved to perform action different options to perform action appropriateness of processes for actions suitability of outcome of action objects handled during action data accessed during action type of input necessary to perform action type of constraints on action interaction with other actions interaction with objects/functions similarity of actions security of actions type of result as output of action customisation of actions	functionality
effort to perform action different ways to perform action difficulties in performing actions difficulties in understanding actions typical sequence of actions help necessary during performance of action	usability
possibilities to undo actions failures during performance of actions types of failures possibility to stop actions	reliability
time needed for action resources needed for action correctness of action output	efficiency

Figure 4.1 : Aspects de qualité relatifs à l'action (« *Qualitative Aspects Related to Actions* » : Höge, 2002 : 100)

ASPECT RELATED TO **OBJECTS**	QUALITY CHARACTERISTIC
function in which it is involved characteristics of object type of object size of object operation modes appropriateness of object for purpose objects with which it interoperates similarity with other objects constraints on object importance of object within use case adaptation of object to specific needs	functionality
failures in objects types of failures action which leads to failure of object	reliability
naming of object mnemonic labels to objects understandability of object names understandability of object function frequency of usage of object layout/shape of objects handling of object presentation of object (interface)	usability
time needed to operate resources needed to operate amount of data processed	efficiency

Figure 4.2 : Aspects de qualité relatifs à l'objet (« *Qualitative Aspects Related to Objects* » : Höge, 2002 : 101)

4.3. Méthodes d'attribution des scores

Afin d'évaluer un logiciel, il est nécessaire de voir comment les critères d'évaluation se combinent entre eux. Pour le faire, il faut les quantifier, c'est-à dire attribuer des scores. L'attribution des scores sert donc à mesurer les critères afin de pouvoir les comparer.

Comme rappelé dans le projet EAGLES (1996 : 29-30), l'attribution des scores est une étape très importante, car si les scores ne sont pas établis correctement, l'évaluation entière perd toute sa valeur. Les scores doivent donc être fiables. Pour le faire, il faut suivre une méthode fiable d'attribution des scores (ibidem).

Les scores sont attribués par rapport à une échelle qu'il faut établir dans la phase de préparation de l'évaluation. Höge (2002 : 66-70) explique qu'il existe plusieurs types d'échelle d'évaluation :

- **échelle nominale** : elle établit la relation des valeurs nominales entre différents systèmes. Elle est composée d'une réponse libre ;
- **échelle binaire** : elle mesure la présence ou l'absence de quelque chose. Elle est caractérisée par deux valeurs (1/0, oui/non, vrai/faux) ;
- **échelle ordinale** : elle est composée de valeurs prédéfinies que l'évaluateur peut choisir. Ces valeurs peuvent être des chiffres (par exemple de 1 à 5) ou des valeurs nominales (insuffisant, suffisant, excellent) ;
- **échelle d'ordre de classement** : elle consiste à donner une liste d'éléments que l'évaluateur doit classer selon un certain critère. Par exemple, on peut donner une liste de villes que l'évaluateur doit classer selon sa préférence ;
- **échelle de proportion (ou de ratio) :** elle est caractérisée par des valeurs équidistantes. Elle fixe un point zéro entre les valeurs et elle est composée par des valeurs numériques issues d'une estimation, comme par exemple, le nombre de clics pour accomplir l'action.

5. Préparation de l'évaluation

Nous avons présenté dans la section 4 le modèle des tâches duquel nous retenons les concepts d'action et d'objet. Nous rappelons que l'action est l'activité accomplie par une personne vers un objet spécifique (objet physique ou informatique). Dans notre cas, l'action est la tâche de recherche et l'objet est représenté par l'outil que nous évaluons.

En suivant cette distinction entre action et objet, nous définissons nos critères d'évaluation pour la recherche (action) (section 5.1.1) et pour l'outil (objet) (section 5.1.2). Comme nous l'expliquons dans les sections respectives, ces critères sont basés sur les tableaux de Höge présentés à la section 4.2 contenant les aspects de qualité relatifs à l'action et à l'objet.

Après avoir défini nos critères d'évaluation, nous présentons nos deux grilles d'évaluation : une pour l'action (section 5.2.1) et une pour l'objet (section 5.2.2). Ensuite, nous passons à la préparation des corpus dont nous nous servons dans le prochain chapitre pour l'exécution de l'évaluation (section 5.3).

5.1. Critères d'évaluation

Comme présenté dans la section 4.2, nous définissons nos critères d'évaluation en partant des deux tableaux de Höge (2002) concernant les aspects de qualité relatifs à l'action (figure 4.1) et les aspects de qualité relatifs à l'objet (figure 4.2).
De cette manière, nous obtenons deux grilles d'évaluation : la première renvoie à la recherche (action) et la deuxième à l'outil (objet). Nous utilisons ces deux grilles de manière différente dans la phase d'exécution de l'évaluation. Comme nous le précisons dans la section 5.2, la grille d'évaluation des tâches sert à évaluer les trois tâches, une à la fois, alors que la grille d'évaluation des outils est utilisée une seule fois pour évaluer les caractéristiques des outils d'exploitation de corpus qui ne peuvent pas être relevées pendant l'exécution des tâches.

5.1.1. Action : définition des critères d'évaluation pour la recherche

En partant du tableau de Höge (2002) concernant les aspects de qualité relatifs à l'action (figure 4.1), nous définissons les critères d'évaluation pour l'action de recherche.

Nous excluons la caractéristique de fiabilité, car nous estimons qu'elle s'applique plutôt à l'évaluation de l'objet. Nous choisissons donc les caractéristiques de qualité suivantes :

- Capacité fonctionnelle
- Facilité d'utilisation
- Rendement.

Parmi les aspects qualitatifs, nous ne retenons que ceux qui sont pertinents pour la recherche et à partir desquels nous définissons nos critères d'évaluation.

5.1.1.1. Capacité fonctionnelle

La capacité fonctionnelle indique la présence d'un ensemble de fonctions que l'on peut utiliser pendant l'accomplissement de l'action. Nous sélectionnons quatre aspects que nous adaptons à notre étude.

Le premier est *different option to perform action*. Cet aspect qualitatif nous permet de dresser une liste des types de recherche (que nous avons décrits dans le chapitre 2) dont le traducteur peut se servir pour accomplir la tâche. Nous l'appelons donc « **différentes options de recherche** ». Les critères correspondants sont :

- recherche par mot
- recherche par mot avec joker
- recherche par troncature
- opérateur OU
- cooccurrents
- recherche parallèle
- recherche par filtre
- recherche par fuzzy
- recherche par expressions régulières.

Le deuxième aspect, *objects handled during action* permet d'identifier l'option d'affichage choisie par le traducteur parmi celles disponibles. Nous l'appelons « **choix de l'affichage** » et elle inclut :

- Liste de mots
- Cluster
- Collocations
- Concordances.

Le troisième aspect que nous sélectionnons à partir du tableau de Höge est*, interaction with object/functions* que nous appelons « **interaction avec d'autres objets/fonctions** ». Il permet de voir si le logiciel peut interagir avec un autre objet ou avec une autre fonction. Dans le cas de la recherche dans un outil d'exploitation des corpus, cette interaction pourrait être représentée par la nécessité de copier les résultats. Cependant, certains logiciels ne

permettent pas de copier automatiquement les informations, ce qui pourrait constituer une perte de temps pour l'utilisateur. Nous définissons donc le critère « Possibilité de copier/coller les informations ».

Enfin, l'aspect de la **personnalisation de l'action** (*customisation of actions*) joue un rôle important, car il permet à l'utilisateur d'adapter le logiciel à ses besoins. La personnalisation peut se faire au moment de la saisie des occurrences ou au moment de la consultation des résultats. Au moment de la saisie des occurrences, on peut avoir besoin de changer la direction des langues ou d'utiliser des filtres de recherche. Ces filtres permettent de limiter le champ de recherche, par exemple, à une partie du corpus ou, au contraire, de l'élargir en sélectionnant plusieurs corpus à la fois. Au moment de la consultation des résultats, l'utilisateur peut avoir l'exigence de trier les résultats selon certains critères afin de faciliter leur lisibilité, par exemple en les triant par ordre alphabétique ou selon une date ou un contexte.
En suivant cette logique, nous réunissons dans cette catégorie les critères concernant la personnalisation de la recherche :

- Changement de la direction des langues
- Possibilité de lancer les recherches simultanément dans plusieurs corpus à la fois
- Possibilité d'affiner la recherche par rapport à la date, au corpus, etc.
- Possibilité de passer d'un corpus à l'autre.

À ces derniers, nous ajoutons les critères concernant le tri des résultats qui permettent, comme nous venons d'expliquer, d'en améliorer la lisibilité :

- Possibilité de choisir l'ordre d'affichage des occurrences selon le document/corpus d'origine
- Possibilité de trier les résultats par rapport à la date, l'auteur, etc.
- Possibilité de trier les occurrences selon le contexte
- Possibilité de régler le nombre de caractères/mots affichés
- Concordance : possibilité de voir le document d'origine.

5.1.1.2. Facilité d'utilisation

La facilité d'utilisation permet de relever les difficultés liées à l'accomplissement de l'action. Parmi les aspects de qualité définis par Höge, nous choisissons l'« **aide nécessaire pendant l'accomplissement de l'action** » (*help necessary during performance of action*). Lorsqu'un logiciel n'est pas intuitif ou contient un grand nombre de fonctions, un utilisateur peu expérimenté pourrait devoir avoir recours au manuel ou à l'aide proposée par le logiciel. Cela joue un rôle important d'un point de vue du temps nécessaire à maîtriser le logiciel. En effet, un nombre excessif de recours au manuel ou à l'aide indique que le logiciel n'est pas intuitif et son utilisation risque d'être perçue par l'utilisateur comme une perte de temps, ce qui le conduira à ne plus s'en servir.

Nous définissons donc le critère « Nombre de recours au manuel ou à l'aide ».

5.1.1.3. Rendement

Le rendement réunit les aspects qui permettent de comprendre si le logiciel est rentable, c'est-à-dire s'il permet d'accomplir la tâche en peu de temps et en utilisant le minimum de ressources.

Parmi les aspects définis par Höge, nous sélectionnons le « **temps nécessaire à l'accomplissement de l'action** » (*time needed for action*). Dans ce cas, le concept de temps peut être mesuré en période (par exemple, minutes, secondes) ou en nombre d'actions (nombre de passages ou de clics). Nous fixons donc les critères suivants :

- Temps nécessaire à accomplir la tâche (en minutes[28])
- Nombre d'essais nécessaires pour accomplir la tâche (maximum 6[29])
- Nombres de passages nécessaires afin de voir toutes les occurrences différentes
- Nombre de clics nécessaires afin de voir toutes les occurrences et leur contexte.

Nous passons ensuite à l'aspect de l'**exactitude des résultats** (*correctness of action output*). Le critère principal concerne la réussite de la tâche qui a une grande importance au moment de l'interprétation des résultats. En effet, l'échec dans l'accomplissement de la tâche

[28] Le choix des minutes est arbitraire. Nous retenons que, par rapport aux secondes, les minutes ont un impact plus immédiat lors de l'interprétation des résultats.

[29] Ce choix est arbitraire. Il s'explique par le fait que, d'après notre expérience, si le traducteur n'obtient pas la réponse souhaitée au bout de 6 essais, il préfère consulter une autre ressource.

signifierait que le logiciel n'est pas en mesure de satisfaire les besoins de l'utilisateur. Le critère est donc « Réussite de la tâche ».

Nous définissons ensuite deux critères complémentaires :

- Nombre d'occurrences trouvées
- Nombre d'occurrences pertinentes.

La comparaison de ces deux critères permet de mesurer le bruit et le silence. En faisant la différence entre le nombre d'occurrences trouvées et le nombre d'occurrences pertinentes, on obtient le bruit, c'est-à-dire le nombre d'occurrences récupérées par le logiciel qui ne sont pas pertinentes. Le silence n'est pas directement mesurable, mais, en comparant pour chaque tâche le nombre d'occurrences pertinentes dans les trois logiciels, il est possible de voir approximativement si un des outils a trouvé moins d'occurrences et donc s'il y a du silence.

Le bruit et le silence peuvent être influencés négativement par l'inexactitude de la recherche (et dépendre donc de l'utilisateur), par l'impossibilité d'affiner la recherche ou par un problème de conception de l'outil (et dépendre dans ces deux derniers cas du logiciel).

Pour finir, nous décidons d'ajouter un nouvel aspect qui est celui de la **fiabilité des résultats**. Cet aspect vient du fait que certains logiciels, comme par exemple OPUS[30], ne donnent pas toujours les mêmes résultats ou ne les présentent pas dans le même ordre. Or, il est important d'avoir des outils fiables et d'être sûr qu'une seule recherche soit suffisante pour obtenir la réponse à notre problème. Le critère correspondant est **«** Stabilité des résultats à la répétition de la tâche ».

5.1.2. Objet : définition des critères d'évaluation pour l'outil

En partant du tableau de Höge (2002) concernant les aspects de qualité relatifs à l'objet (figure 4.2), nous définissons plusieurs critères pour l'évaluation de l'outil.

[30] OPUS est un recueil de corpus parallèles, constitués à partir de textes issus du web, développé par l'Université d'Uppsala. Il contient plusieurs corpus sur des sujets variés, parmi lesquels l'Europarl, l'OpenSubs2013 composés de sous-titres de films, ou le MEA - European Medicines Agency documents. Tous les corpus sont alignés et parfois annotés. OPUS met à disposition tous les corpus pour le téléchargement sous les formats TXT, Moses et XML, ainsi qu'une interface permettant de faire des requêtes en ligne. OPUS est disponible sur le lien suivant : http://opus.lingfil.uu.se/ - Consulté le 28/05/2014.

Par rapport aux propositions de Höge, nous excluons la caractéristique de capacité fonctionnelle qui est plus pertinente pour l'action (section 5.1.1.1). Nous menons donc notre étude sur les caractéristiques de qualité de fiabilité, de facilité d'utilisation et de rendement.

5.1.2.1. Fiabilité

La fiabilité rassemble les aspects permettant de comprendre si un outil présente des défauts de programmation et donc si des actions peuvent provoquer des erreurs. Lorsqu'un logiciel est souvent sujet à des erreurs ou à des dysfonctionnements, l'utilisateur perd la confiance en l'outil et arrive même à ne plus s'en servir. Bien entendu, les erreurs peuvent être dues à des actions externes, qui se produisent en dehors du logiciel même (par exemple dues à un blocage de l'antivirus). Nous ne tenons compte ici que des erreurs engendrées par les actions accomplies sur le logiciel même.

Par rapport à la grille de Höge, nous choisissons l'aspect « **type d'erreur** » (*type of failure*) auquel nous attribuons les critères :

- Erreur d'affichage
- Blocage du logiciel.

Nous sélectionnons également le critère « **action qui provoque l'erreur** » (*action which leads to failure of object*) qui nous permet, le cas échéant, d'en détecter l'origine. Nous identifions trois critères :

- Mauvaise manipulation de l'utilisateur
- Annulation d'une action
- Utilisation d'une option du logiciel.

5.1.2.2. Facilité d'utilisation

La facilité d'utilisation permet d'observer la manière dont est conçu le logiciel. Nous sélectionnons l'aspect « **présentation de l'objet (interface)** » (*presentation of object (interface*)) et nous définissons des critères concernant la présentation des concordances et des informations liées aux occurrences :

- Concordance : affichage plein texte
- Concordance : affichage KWIC

- Nombre maximum de caractères pour la recherche
- Mise en évidence de l'élément recherché
- Repérage du segment traduit
- Informations concernant la source de la séquence affichée
- Affichage du nombre total d'occurrences
- Affichage des flexions d'une occurrence.

5.1.2.3. Rendement

Le rendement réunit l'ensemble des aspects liés à l'efficacité du logiciel, vue comme le rapport entre les ressources utilisées, le temps nécessaire à la performance et le nombre de données traitées.

En ce qui concerne l'action de recherche, il est intéressant de voir l'aspect « **temps nécessaire à la performance** » (*time needed to operate*). Nous identifions deux critères :

- Nombre de passages nécessaires pour arriver à la fenêtre de recherche
- Temps (sec.) nécessaire à l'affichage des résultats.

5.2. Présentation des grilles et attribution des échelles d'évaluation

Sur la base de ces choix, nous synthétisons les critères et les échelles d'évaluation dans les deux grilles suivantes. Ces grilles sont utilisées dans le prochain chapitre pour la phase d'exécution de l'évaluation. Pour chaque grille, nous montrons les critères d'attribution des échelles d'évaluation.

5.2.1. Grille d'évaluation des tâches

Cette première grille rassemble tous les critères d'évaluation pour la recherche que nous définissons à la section 5.1.1. Nous l'utilisons pour l'évaluation de chaque tâche que nous décrivons dans les détails dans le chapitre 5.

Afin de rendre cette évaluation la plus objective possible, nous attribuons uniquement des échelles d'évaluation binaires (oui/non) et de ratio.

Dans les critères issus de la capacité fonctionnelle que nous évaluons avec une échelle binaire, les réponses affirmatives correspondent aux fonctions ou aux options utilisées pour accomplir la tâche. En considérant l'hétérogénéité des outils et pour des raisons pratiques,

nous ajoutons à l'échelle binaire une troisième réponse qui est ND (non disponible). Ceci nous permet de comprendre, lors de l'interprétation des résultats, si la fonction ou l'option n'a pas été choisie parce que l'utilisateur en a préféré une autre ou parce qu'elle n'est pas présente sur l'outil.

Caractéristiques de qualité	Aspects qualitatifs	Critères	Échelle d'évaluation
Capacité fonctionnelle	différentes options de recherche	recherche par mot	binaire
		recherche par mot avec joker	binaire
		recherche par troncature	binaire
		opérateur OU	binaire
		cooccurrents	binaire
		recherche parallèle	binaire
		recherche par filtre	binaire
		recherche par fuzzy	binaire
		recherche par expressions régulières	binaire
	choix de l'affichage	liste de mots	binaire
		cluster	binaire
		collocations	binaire
		concordances	binaire
	interaction avec d'autres objets/fonctions	possibilité de copier/coller les informations	binaire
	personnalisation de l'action	changement de la direction des langues	binaire
		possibilité de lancer les recherches simultanément dans plusieurs corpus à la fois	binaire
		possibilité d'affiner la recherche par rapport à la date, au corpus, etc.	ratio
		possibilité de passer d'un corpus à l'autre	binaire
		possibilité de choisir l'ordre d'affichage des occurrences selon le document/corpus d'origine	ratio
		possibilité de trier les résultats par rapport à la date, l'auteur, etc.	ratio
		possibilité de trier les occurrences selon le contexte	binaire
		possibilité de régler le nombre de caractères/mots affichés	binaire
		concordance: possibilité de voir le document d'origine	binaire
Facilité d'utilisation	aide nécessaire pendant l'accomplissement de l'action	nombre de recours au manuel ou à l'aide	ratio
Rendement	temps nécessaire à l'accomplissement de l'action	temps nécessaire à accomplir la tâche (en minutes)	ratio
		nombre d'essais nécessaires pour accomplir la tâche (maximum 6)	ratio
		nombres de passages nécessaires afin de voir toutes les occurrences différentes	ratio
		nombre de clics nécessaires afin de voir toutes les occurrences et leur contexte	ratio
	exactitude des résultats	réussite de la tâche	binaire
		nombre d'occurrences trouvées	ratio
		nombre d'occurrences pertinentes	ratio
	fiabilité des résultats	stabilité des résultats à la répétition de la tâche	binaire

Figure 4.3 : Grille d'évaluation des tâches

5.2.2. Grille d'évaluation des outils

Cette deuxième grille rassemble tous les critères d'évaluation pour l'outil que nous définissons à la section 5.1.2. Elle permet d'évaluer les caractéristiques des outils qui ne sont pas paramétrables au moment de l'accomplissement de la tâche, car elles sont orientées vers l'outil même. Dans la phase d'exécution de l'évaluation, nous utilisons cette grille une seule fois pour les trois logiciels, après avoir accompli les tâches.

Caractéristiques de qualité	Aspects qualitatifs	Critères	Échelle d'évaluation
Fiabilité	type d'erreur	erreur d'affichage	binaire
		blocage du logiciel	binaire
	action qui provoque l'erreur	mauvaise manipulation de l'utilisateur	binaire
		annulation d'une action	binaire
		utilisation d'une option du logiciel	binaire
Facilité d'utilisation	présentation de l'objet (interface)	concordance: affichage plein texte	binaire
		concordance: affichage KWIC	binaire
		nombre maximum de caractères pour la recherche	ratio
		mise en évidence de l'élément recherché	binaire
		repérage du segment traduit	binaire
		informations concernant la source de la séquence affichée	ratio
		affichage du nombre total d'occurrences	binaire
		affichage des flexion d'une occurrence	binaire
Rendement	temps nécessaire à la performance	nombre de passages nécessaires pour arriver à la fenêtre de recherche	ratio
		temps (sec.) nécessaire à l'affichage des résultats	ratio

Figure 4.4 : Grille d'évaluation des outils

5.3. Préparation des corpus

Enfin, après avoir défini les critères d'évaluation et avoir attribué des échelles d'évaluation, dans cette partie nous terminons par la présentation des deux corpus que nous utilisons pour l'évaluation.

5.3.1. Choix des corpus

Notre évaluation se base sur deux corpus parallèles italien-français. Ils résultent d'un stage que nous avons effectué auprès de l'agence de traduction Transpose SA de Genève dont l'objectif principal était de classer tous les travaux disponibles dans l'archive de 2004 à 2012. Nous avons extrait de l'archive, classé et renommé plus de 40 000 fichiers. Le résultat est un recueil subdivisé par type d'entreprises appartenant à des domaines variés (par exemple : *Food Industry, Education, Medical, Legal, Watch Industry, Security*, etc.), puis, à l'intérieur de ces mêmes catégories, par client. La plupart de ces collections de documents sont multilingues et peuvent contenir plusieurs centaines de documents sous les formats Excel, Word, PowerPoint ou PDF.

Après avoir obtenu les autorisations nécessaires, nous analysons le recueil afin de former les corpus à utiliser pour ce mémoire. Tout d'abord, nous sélectionnons toutes les collections contenant les couples de langues français-italien. Parmi ces dernières, nous excluons toutes celles contenant un grand nombre de données et d'informations sensibles afin de ne pas courir le risque de violer les normes sur la confidentialité. Nous retenons donc 5 entreprises. Parmi ces dernières, il y a trois sociétés horlogères que nous décidons d'exclure car les documents à notre disposition sont des catalogues et ne garantissent donc pas assez d'hétérogénéité. Nous retenons, au final, les collections de textes d'une grande entreprise qui produit du chocolat (corpus CHOCOLAT) et d'un institut de formation professionnelle (corpus FORMATION). À l'intérieur de ces collections, nous sélectionnons les textes à intégrer à nos corpus que nous présentons dans les sections suivantes.

5.3.2. Contenu des corpus

5.3.2.1. Corpus CHOCOLAT

Le corpus CHOCOLAT contient des documents que l'agence Transpose SA a traduits pour une grande société qui produit, entre autres, des tablettes de chocolat. Les textes sélectionnés

ont pour sujet le processus de production des tablettes de chocolat et la description des produits finis.

Le corpus est composé de 6 textes par langue (12 au total) dont la langue source est le français et la langue cible est l'italien. Le traducteur est le même pour tous les textes.

Le contenu des textes étant très varié, nous les présentons tous sous forme de tableau.

Texte	Description
Texte 1	Brochure de présentation des produits destinée aux consommateurs
Texte 2	Présentation en PowerPoint destinée à la formation du personnel composée de : - présentation de la gamme de produits - histoire du chocolat - pays producteurs de cacao - description des phases d'élaboration du chocolat (de la culture du cacaoyer jusqu'à la transformation en produit fini) - notions de dégustation du chocolat et association entre chocolat et café, thé et alcool
Texte 3	Petit lexique destiné au personnel, en particulier aux vendeurs, contenant les termes clés du domaine du chocolat
Texte 4	Interview de la personne de contact entre la société en question et les pays producteurs de cacao
Texte 5	foire aux questions adressées aux consommateurs pour expliquer l'origine des matières premières et la politique de la société
Texte 6	Présentation adressée aux consommateurs du coffret « Expert »

Tableau 4.1 : Descriptif du contenu du corpus CHOCOLAT

5.3.2.2. Corpus FORMATION

Le corpus FORMATION contient des documents d'un institut de formation professionnelle suisse. Il est composé de 13 textes par langues (26 au total) traduits de l'allemand vers le français et vers l'italien. Ce ne sont donc pas des traductions directes et il y a plusieurs traducteurs. Toutefois, les textes sont pour l'institut des versions « officielles ».

Le corpus contient :

- le règlement de l'institut ;
- deux présentations de projets ;
- quatre descriptifs de cours ;
- un descriptif des fonctions au sein de l'institut ;
- un descriptif des objectifs stratégiques de l'institut ;

- un mandat de prestation du Conseil fédéral ;
- un rapport de gestion ;
- un glossaire de termes d'évaluation.

Pour finir, voici les caractéristiques des deux corpus dans les détails :

	Corpus CHOCOLAT	Corpus FORMATION
Nombre de textes	12	26
Nombre de mots partie FR	13.751	23.206
Nombre de mots partie IT	13.115	22.068
Total	26.866	45.274

Tableau 4.2 : Contenu des corpus

5.3.3. Nettoyage des corpus

Comme l'expliquent Bowker et Pearson (2002 : 96), la phase de nettoyage des textes est indispensable pour réduire au minimum les erreurs dans la phase d'alignement, car les logiciels d'alignement présupposent que les textes et leurs traductions sont composés du même nombre de phrases. Pour cette raison, elle doit être faite avec attention.

Une fois les corpus construits, nous ouvrons en parallèle les textes originaux et leur traduction et vérifions si le document est traduit dans toutes ses parties. Nous effaçons les en-têtes, les pieds de page ainsi que les logos, les adresses et les parties des tableaux inutilisables, ainsi que toutes sortes d'éléments pouvant gêner l'alignement des textes.
Pour des raisons de confidentialité, nous rendons ensuite anonymes les documents en remplaçant tous les noms propres par des initiales et en enlevant le nom de la société dans le corpus CHOCOLAT.
Le texte 2 du corpus CHOCOLAT subit un traitement différent. Il s'agit à l'origine d'une présentation en PowerPoint. Certaines diapositives du fichier source contiennent des commentaires qui ne sont pas traduits dans le texte cible et les images du texte source sont traduites en tant que commentaires aux diapositives. Nous effaçons donc tous les commentaires et toutes les images non traduites. Ensuite, nous supprimons les doublons et convertissons le fichier au format TXT. Il s'agit donc du seul texte qui n'est pas intégral.

5.3.4. Alignement des corpus

La dernière phase de préparation des corpus est l'alignement. Ce critère ne faisant pas l'objet d'une évaluation dans le cadre de ce mémoire, nous décidons de pré-aligné tous les documents avant de les charger sur *MultiTrans Prism* et *ParaConc* (dans *myCAT* l'alignement est automatique). Trois raisons principales nous portent à ce choix : premièrement, nous voulons éviter de corriger deux fois l'alignement, vu la dimension de nos corpus ; deuxièmement, *ParaConc* ne permet pas de garder en mémoire une paire de documents alignés, ce qui implique une correction de l'alignement à chaque ouverture du logiciel ; troisièmement, nous devons garantir les mêmes résultats en cas de test de la part de tiers ou de plusieurs utilisateurs sur nos corpus qui ne peuvent êtes obtenus que si l'alignement est identique.

Nous décidons donc de réduire au minimum le nombre de fichiers à charger dans chaque outil. Or, nous considérons le fait que *MultiTrans Prism* accepte des fichiers TMX, alors que *ParaConc* accepte uniquement des fichiers de texte. Ceci veut dire que nous ne pouvons pas utiliser les mêmes formats de documents pour ces deux outils. Au final, nous devons obtenir un fichier TMX français-italien pour *MultiTrans Prism* et deux fichiers au format TXT (dont un pour le français et l'autre pour l'italien) pour *ParaConc*. Pour le faire, nous passons par deux étapes. Pour commencer, nous utilisons le logiciel *AlignFactory* (que nous présentons à la section 1.1.2 du chapitre 3) afin d'aligner les textes de chaque corpus et d'obtenir un seul document par corpus au format TMX que nous utilisons pour *MultiTrans Prism.*

Une fois le fichier TMX obtenu, nous le scindons en deux fichiers au format de texte, un pour le français et l'autre pour l'italien. Ceci nous permet de construire des documents pré-alignés à utiliser pour *ParaConc* sans avoir besoin d'aligner manuellement les documents d'origine. En suivant cette méthode, après le premier chargement des fichiers sur *ParaConc*, seules quelques petites corrections à nos fichiers TXT sont nécessaires.

Pour finir, nous précisons que nous choisissons l'encodage de caractères ANSI pour les TXT et ajoutons les balises sous le format <Texte 1 : *[nom du document]*>.

Chapitre 5 : Évaluation

Le présent chapitre est dédié à l'exécution de l'évaluation. Nous testons nos trois logiciels dans l'accomplissement de trois tâches de recherche sur nos deux corpus parallèles français-italien afin de vérifier si ces outils répondent aux besoins potentiels des traducteurs.
Nous commençons par décrire dans les détails nos trois tâches d'évaluation (section 1), puis nous passons au déroulement de ces tâches (section 2) et à l'évaluation avec notre grille d'évaluation des outils (section 3). Nous concluons avec la discussion des résultats (section 4).

1. Définition des tâches

Nous définissons, dans cette partie, trois tâches de recherche afin de pouvoir tester nos trois outils à l'aide de la grille d'évaluation des tâches que nous présentons dans le chapitre 4 (section 5.2.1). Ces tâches nous permettent de voir dans la pratique comment interviennent les corpus parallèles dans les différentes phases de traduction. Nous rappelons que ces phases sont la pré-traduction, la traduction et la révision (section 3 chapitre 1).

1.1. Présentation de la tâche 1 : documentation

La première tâche concerne la phase de pré-traduction. Dans cette phase, le traducteur peut se servir des corpus, entre autres, pour se documenter, c'est-à-dire pour se familiariser avec le sujet et trouver des informations concernant les concepts clés (chapitre 1 section 3.2.1.1).
La situation imaginée est que nous avons reçu un nouveau texte à traduire du français vers l'italien concernant le processus de transformation de la fève de cacao en chocolat. N'étant pas des experts en la matière, avant de commencer notre traduction, nous avons besoin de nous documenter. Nous savons que d'autres textes du même domaine ont déjà été traduits par notre agence pour un autre client. Nous décidons donc de nous servir de nos outils d'exploitation de corpus parallèles afin de trouver des informations sur le sujet dans notre corpus CHOCOLAT. En particulier, nous cherchons une réponse à la question suivante : « Qu'est-ce que le *tempérage* et à quel stade de l'élaboration du produit intervient-il ? Quel est son équivalent en italien ? ».

1.2. Présentation de la tâche 2 : traduction

La deuxième tâche est la recherche des collocations d'un terme. Ce type de recherche peut s'effectuer tout au long du processus de traduction, mais on y recourt surtout dans la phase de traduction.

La situation imaginée est la suivante : nous sommes en train de traduire le rapport annuel 2013 d'un institut de formation professionnelle. La notion « décerner un titre » revient plusieurs fois dans une même page. Nous souhaitons éviter des répétitions, c'est pourquoi nous voulons savoir avec quels verbes se construit « titre ». Cet institut de formation est notre client depuis des années, c'est pourquoi nous consultons notre corpus FORMATION afin de voir les collocations du terme.

La question à laquelle nous devons répondre est donc la suivante : « Quelles sont les collocations de *titre* en français et de son équivalent en italien ? ».

1.3. Présentation de la tâche 3 : vérification de la terminologie

La dernière tâche concerne la phase de révision. Comme nous l'avons vu dans le chapitre 1 (section 3.2.3), dans la phase de révision, les corpus parallèles peuvent être utilisés afin de fournir des réponses à des doutes et de vérifier l'harmonisation du texte par rapport aux traductions précédentes.

La situation imaginée est la suivante : nous avons traduit le rapport annuel 2013 de l'institut de formation professionnelle. En relisant le texte, nous nous apercevons que nous avons traduit « école supérieure » par « *scuola superiore* ». Nous doutons du fait que ce soit la bonne traduction, mais aucune alternative ne nous vient à l'esprit. Ainsi, nous décidons de consulter notre corpus FORMATION.

La question est donc la suivante : « Est-ce que *scuola superiore* est bien le meilleur équivalent d'*école supérieure* ? »

2. Déroulement des tâches

Une fois les trois tâches de recherche définies, nous pouvons passer à l'exécution de l'évaluation.

Nous présentons dans cette section le déroulement des trois tâches. Chaque tâche est évaluée à l'aide de la grille d'évaluation des tâches que nous présentons dans le chapitre 4

(section 5.2.1). Les versions intégrales des grilles pour les trois tâches peuvent être consultées à l'annexe A.

2.1. Tâche 1 : documentation

Comme nous le précisons dans la section 1.1, la question à laquelle nous devons répondre est la suivante : « Qu'est-ce que le *tempérage* et à quel stade de l'élaboration du produit intervient-il ? Quel est son équivalent en italien ? ».

2.1.1. MultiTrans Prism

2.1.1.1. Capacité fonctionnelle

Dans le choix des options de recherche, nous considérons le fait que, étant donné que nous devons chercher une phase de l'élaboration du chocolat, le mot sera sûrement au singulier. S'agissant d'un seul mot, il n'est pas nécessaire d'utiliser la recherche par fuzzy (option tous les mots). Nous cherchons « tempérage » dans *MultiTrans Prism* avec une recherche par mot exacte (qui, nous le rappelons, s'oppose à la recherche par radicaux). Nous choisissons l'affichage des concordances et nous obtenons 5 occurrences, dont la première est :

> « Tempérage : Faire passer le chocolat à 3 paliers de température différent pour le préparer au moulage ou à l'enrobage ».

Le segment correspondant en italien est bien surligné en jaune et nous montre que le terme équivalent est « *temperaggio* ». Nous obtenons donc la traduction du terme ainsi que l'information que le tempérage précède la phase de moulage ou d'enrobage.

La deuxième occurrence explique dans les détails ce qu'est le tempérage et précise qu'il suit la phase de concassage.

Aucune option de personnalisation de l'action n'a été utilisée.

2.1.1.2. Facilité d'utilisation et rendement

Aucun recours au manuel n'est nécessaire. Nous voyons donc qu'avec une simple recherche par mot, sans utiliser les fuzzy ni aucune option de personnalisation et en choisissant l'affichage des concordances, le temps nécessaire à l'accomplissement de cette recherche est de 1,15 minute. Pour voir toutes les occurrences dans leur contexte, nous effectuons 5 passages, un par occurrences. Nous remarquons également que l'on obtient les mêmes résultats en modifiant les options de recherche, c'est-à-dire en faisant la recherche par radicaux ou par fuzzy.

2.1.2. ParaConc

2.1.2.1. Capacité fonctionnelle

Nous cherchons « tempérage » en utilisant la recherche par mot avec joker et avec un affichage par concordances, tout en sachant que l'on devrait obtenir les mêmes résultats avec la recherche par mot sans joker. Au premier essai, le contexte par défaut (40 caractères) apparaît trop court pour lire le contexte des occurrences. Nous recourons alors à une première option de personnalisation de l'action pour augmenter le nombre de caractères affichés jusqu'à 100 caractères. Nous utilisons l'option *possibilité de voir le document d'origine* pour 3 occurrences sur 5 car l'affichage en KWIC ne nous permet pas d'avoir des phrases complètes.

Pour voir le terme équivalent en italien, il n'y a pas besoin de mettre l'affichage en KWIC car il apparaît clairement qu'il s'agit de « *temperaggio* ».

2.1.2.2. Facilité d'utilisation et rendement

Le temps nécessaire pour la recherche est de 3,30 minutes. Du point de vue du rendement, l'avantage de voir toutes les occurrences en une seule page disparaît, car nous devons utiliser deux options de personnalisation et consulter trois fois l'affichage plein texte, ce qui porte le nombre d'actions à 5.

2.1.3. myCAT

2.1.3.1. Capacité fonctionnelle

Nous sélectionnons le corpus CHOCOLAT et nous lançons la recherche par mot « tempérage » en choisissant le seul affichage possible, c'est-à-dire celui par concordances et sans utiliser les options de personnalisation. Nous voyons que le terme apparaît dans 2 documents. Nous cliquons sur le premier document pour voir la première occurrence et nous voyons tout de suite que le tempérage vient après le concassage et avant le moulage.

Nous remarquons que l'occurrence correspondante en italien n'est pas correctement signalée, mais cela ne nous empêche pas de retrouver facilement le bon équivalent qui est « *temperaggio* ».

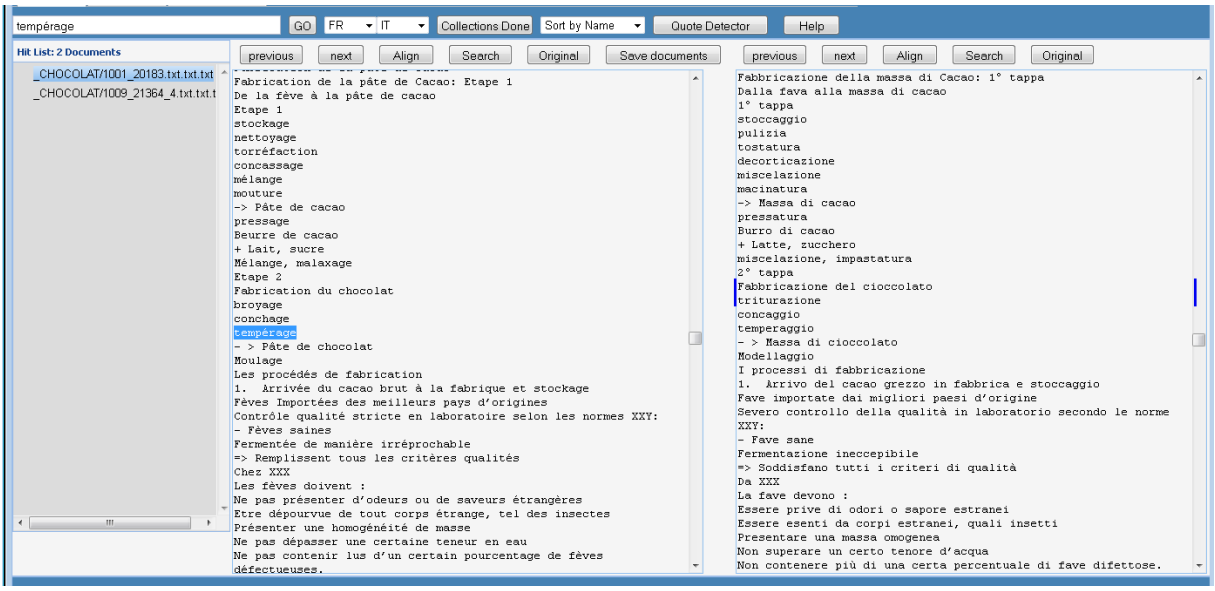


Figure 5.1 : *myCAT* recherche de « tempérage »

La quatrième et dernière occurrence du document nous donne les détails de la phase de tempérage. Le segment correspondant en italien est correctement repéré.

Dans le document 2, nous voyons qu'il n'y a qu'une seule occurrence qui est la suivante :

> « Tempérage : Faire passer le chocolat à 3 paliers de température différent pour le préparer au moulage ou à l'enrobage ».

L'indication du segment correspondant est encore une fois erronée.

2.1.3.2. Facilité d'utilisation et rendement

Nous n'avons aucun recours au manuel ou à l'aide. Le temps nécessaire pour accomplir la tâche est de 1,10 minute, par contre 6 passages sont nécessaires afin de voir toutes les occurrences différentes et de les voir dans leur contexte.

2.2. Tâche 2 : traduction

Comme nous l'avons précisé dans la section 1.2, la question à laquelle nous devons répondre est la suivante : « Quelles sont les collocations de *titre* en français et de son équivalent en italien ? ».

2.2.1. MultiTrans Prism

2.2.1.1. Capacité fonctionnelle

Nous lançons la recherche par troncature « titre » dans le but de retrouver les occurrences au singulier et au pluriel et nous trouvons 31 occurrences pour « titre » et 6 pour « titres ».

Nous choisissons l'affichage des concordances et nous n'utilisons aucune option de personnalisation.

2.2.1.2. Facilité d'utilisation et rendement

Le temps nécessaire pour accomplir la tâche est de 7,20 minutes, mais il faut 37 passages pour voir toutes les occurrences dans leur contexte.

Après avoir analysé les résultats, nous observons que, pour le français, « titre » s'associe avec « décerné » (6 occurrences), « décernés » (3 occurrences), « obtention » et « visé », (2 occurrences), suivi de « conféré », « délivré », « délivrés » et « porter » (une occurrence). Pour l'équivalent italien « *titolo* », nous observons qu'il est associé à « *rilasciato* » (3 occurrences), suivi de « *portare* », « *conseguire* », « *acquisito* », « *rilascio* », « *conferiti* » et « *assegnato* » (une occurrence).

2.2.2. ParaConc

2.2.2.1. Capacité fonctionnelle

Nous effectuons une recherche avec l'expression régulière « \btitres?\b » et nous choisissons l'affichage des concordances. Dans la partie en italien, nous affichons les « *Hot Words* » et sélectionnons « *titolo »* et « *titoli* », ensuite nous passons à l'affichage KWIC. Maintenant que la fenêtre de recherche est active et que nous avons sélectionné les « *Hot Words* » pour l'italien, nous pouvons avoir recours aux collocations. Dans « *Frequency Options* » nous baissons le nombre minimum d'occurrences à 1 et nous choisissons un intervalle de 3 unités à gauche et 3 à droite. Nous lançons la recherche mais nous obtenons des résultats trop hétérogènes. Cependant, si l'on analyse les résultats de la première et de la deuxième unité à droite, on obtient 3 occurrences pour « décernés », 2 occurrences pour « visé », une occurrence pour « conférer », « décerné », « délivrés », « délivré », mais nous trouvons aussi « d'acquérir » qui pourrait être pertinent selon le contexte. Les résultats pour l'italien sont difficiles à interpréter. Nous voyons qu'à la deuxième unité à gauche nous avons « *conseguire* » et à la première à droite « *rilascio* ».

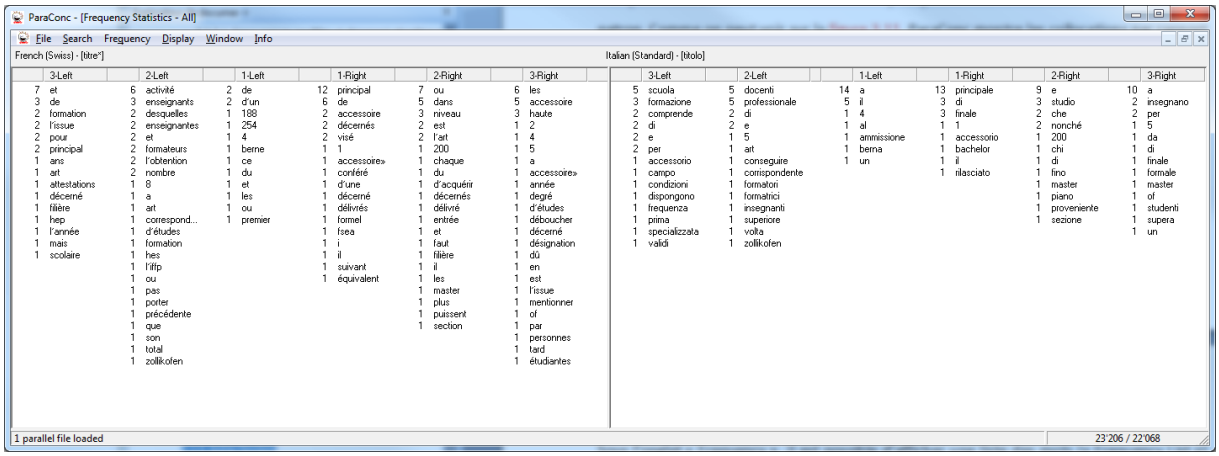


Figure 5.2 : *ParaConc*, collocations pour « \btitres?\b » et pour « titolo/titoli »

Cette hétérogénéité est signe que nous ne pouvons pas espérer de meilleurs résultats avec les clusters, c'est pourquoi nous décidons d'analyser les résultats obtenus avec les concordances. Afin d'améliorer la lisibilité des occurrences, nous avons recours à deux options de personnalisation de l'affichage : tout d'abord, pour la partie en français nous trions les résultats par rapport au contexte, plus précisément par rapport au premier mot à gauche (*Short/1st Left/Search Term*). Nous devons ensuite passer à l'affichage plein texte pour une occurrence.

Figure 5.3 : *ParaConc* : recherche de « \btitres*\b » avec tri du contexte gauche et affichage KWIC dans les deux langues

2.2.2.2. Facilité d'utilisation et rendement

Nous avons eu recours au manuel pour vérifier comment accéder aux options de tri du contexte.

Nous obtenons les mêmes résultats qu'avec *MultiTrans Prism*. Nous effectuons deux essais, mais le nombre de passages est nettement inférieur (4 passages). Par contre, il nous faut 9 minutes pour compléter la tâche, car la lisibilité des résultats en une seule colonne n'est pas optimale.

2.2.3. myCAT

2.2.3.1. Capacité fonctionnelle

Nous sélectionnons le corpus FORMATION et nous lançons la recherche par mot avec joker « titre* ». Nous n'utilisons aucune option de personnalisation. Nous obtenons ainsi les occurrences au singulier et au pluriel dans 6 documents.

2.2.3.2. Facilité d'utilisation et rendement

Afin d'afficher les occurrences, 45 passages sont nécessaires. Nous obtenons les mêmes résultats qu'avec *MultiTrans Prism*, mais le temps employé pour l'accomplissement de la tâche est de 13,08 min. Cette différence de temps s'explique par le fait que le repérage des collocations de l'équivalent italien « *titolo* » est difficile à cause des erreurs de mise en évidence des segments.

2.3. Tâche 3 : vérification de la terminologie

Dans cette tâche, la question à laquelle nous devons répondre est la suivante : «Est-ce que *scuola superiore* est bien le meilleur équivalent d'*école supérieure*? ».

2.3.1. MultiTrans Prism

2.3.1.1. Capacité fonctionnelle

Dans *MultiTrans Prism*, nous lançons la recherche par *fuzzy* (mots consécutifs) « école supérieure », mais nous n'obtenons aucun résultat. Nous lançons donc une deuxième recherche avec « écoles supérieures » et nous trouvons 14 occurrences. Nous n'utilisons aucune option de personnalisation.

2.3.1.2. Facilité d'utilisation et rendement

Aucun recours au manuel ou à l'aide n'est nécessaire. Le temps employé pour accomplir la tâche est de 4,10 minutes et 16 passages ont été effectués afin de voir toutes les occurrences différentes et de les voir dans leur contexte.

Nous trouvons 14 occurrences, dont 8 correspondances pour « *scuole specializzate superiori* », 4 pour « *scuola specializzata superiore* », 1 pour « *scuole professionali di base specializzate superiori* » et 1 pour « *scuole superiori* ».

2.3.2. ParaConc

2.3.2.1. Capacité fonctionnelle

Nous lançons une recherche parallèle pour voir combien de fois « école(s) supérieure(s) » est traduit par « *scuola/e superiore/i* » à l'aide des expressions régulières « \bécoles*\b \bsupérieures*\b » pour le français et « \bscuol[ae]\b \bsuperior[ei]\b » pour l'italien. Nous obtenons une seule occurrence.

Nous lançons donc une nouvelle recherche parallèle avec les mêmes expressions régulières, en utilisant cette fois l'opérateur booléen SAUF pour retrouver toutes les occurrences dans lesquelles « école(s) supérieure(s) » n'est pas traduit par « *scuola/e superiore/i* ». Nous obtenons ainsi 12 occurrences. Afin de rendre plus lisibles les occurrences en italien nous personnalisons le contexte d'affichage en sélectionnant dans « *Hot Words* » le mot « *superiori* » et nous mettons l'affichage en KWIC.

2.3.2.2. Facilité d'utilisation et rendement

Le temps nécessaire à l'accomplissement de l'action est de 3 minutes et nous avons effectué 4 passages au total.

Par rapport à *MultiTrans Prism, ParaConc* affiche 13 occurrences, car il ne retrouve pas « ecoles supérieures » à cause d'une faute d'orthographe dans le document d'origine. Nous avons donc 9 occurrences pour « *scuole specializzate superiori* », 2 occurrences pour « *scuola specializzata superiore* », 1 occurrence pour « *scuole professionali di base specializzate superiori* » et 1 pour « *scuola superiore* ».

2.3.3. myCAT

2.3.3.1. Capacité fonctionnelle

Nous lançons la recherche par cooccurrents « école supérieure », mais n'obtenons aucune occurrence. Nous cherchons donc « écoles supérieures » et nous obtenons 4 documents. Nous rappelons que la recherche par joker n'est possible qu'avec un seul mot, c'est pourquoi nous devons lancer deux recherches distinctes.

2.3.3.2. Facilité d'utilisation et rendement

Le temps nécessaire pour la recherche est de 6,30 minutes et nous effectuons 16 passages pour voir toutes les occurrences dans leur contexte.

Nous obtenons 10 occurrences, dont 7 pour « *scuole specilazzate superiori* », 2 pour « *scuola specializzata superiore* » et 1 pour « *scuola superiore* ». Par rapport à *MultiTrans Prism*, nous remarquons que le nombre d'occurrences trouvées est inférieur. Nous déduisons donc que dans les résultats de *myCAT* il y a du silence (10 occurrences contre 14 pour *MultiTrans Prism*).

3. Évaluation avec la grille des outils

Une fois les trois tâches accomplies, nous remplissons notre grille d'évaluation des outils (Annexe B) et nous en analysons les résultats.

3.1. Fiabilité

Suite au déroulement des tâches, nous remarquons qu'aucun des trois outils testés n'a été sujet à un blocage lors de ces trois tâches. Par contre, nous voyons que *myCAT* connaît des erreurs d'affichage dont nous n'avons pas détecté l'origine : parfois le segment de la langue cible n'est pas correctement repéré ou n'est pas du tout mis en évidence (surtout sur *Internet Explorer*).

Pour aller plus loin et voir ainsi les limites des outils, nous cherchons une séquence se terminant par un point pour voir si les trois logiciels la reconnaissent. Nous précisons que cette séquence est présente dans le corpus et se termine bien par un point. *ParaConc* affiche un message d'erreur nous prévenant que nous avons utilisé un caractère non admis. Le nombre et le type de caractères admis sont d'ailleurs paramétrables. Au contraire, *MultiTrans Prism* et *myCAT* ne donnent pas de limites de nombre de caractères, mais

n'admettent pas les points, pourtant, aucun message d'erreur ne s'affiche au moment de la recherche.

Nous concluons donc que *ParaConc* est plus fiable, car, contrairement à *MultiTrans Prism* et à *myCAT*, nous n'avons détecté aucune erreur d'affichage.

3.2. Facilité d'utilisation

3.2.1. Affichage des concordances

En ce qui concerne l'affichage des concordances, nous remarquons que *MultiTrans Prism* et *myCAT* disposent d'un affichage plein texte, alors que *ParaConc* dispose d'un affichage KWIC et donne la possibilité de voir l'occurrence en plein texte seulement en cliquant dessus. Toutefois, *ParaConc* ne permet pas de passer d'une occurrence à l'autre en plein texte.

Les trois outils mettent en évidence l'élément recherché et repèrent le segment traduit, mais, grâce à un jeu de couleurs, *MultiTrans Prism* en facilite la lisibilité.

3.2.2. Affichage des informations complémentaires

La distinction la plus importante concerne les informations complémentaires aux occurrences retrouvées. *MultiTrans Prism* propose un tableau contenant le nombre total d'occurrences ainsi que les flexions accompagnées du nombre de fois qu'elles apparaissent. *ParaConc* affiche uniquement le nombre total des occurrences, mais pour voir toutes les flexions, il faut comparer toutes les occurrences. L'interface de *myCAT* est moins complète, car elle indique uniquement le nombre de documents contenant l'occurrence, mais ne donne aucune indication concernant le nombre total des occurrences. Lorsqu'on sélectionne un document, le nombre total d'occurrences contenues dans le document même n'apparaît que lorsqu'on arrive à la dernière séquence. Pour trouver toutes les flexions, il faut consulter tous les documents et comparer toutes les occurrences.

En ce qui concerne les informations concernant la source de la séquence affichée, *ParaConc* indique le corpus et la ligne, alors que *myCAT* permet de voir le nom complet du document, y compris le corpus. *MultiTrans Prism* n'est pas évaluable, car nous avons construit la TermBase à partir d'un fichier TMX et non pas à partir des documents d'origine. De ce fait, les seules informations disponibles sont les coordonnées du mot, car le corpus est indexé, et la TermBase.

3.3. Rendement

Du point de vue du rendement, il n'y a pas de différences remarquables en ce qui concerne les temps d'affichage des résultats, cependant, alors que *MultiTrans Prism* et *myCAT* permettent d'accéder directement à la fenêtre de recherche, pour *ParaConc* au moins deux passages sont nécessaires. Bien sûr, la différence vient du fait que plusieurs modalités de recherche sont disponibles, mais cela montre également les limites de son interface qui influence négativement aussi bien les temps de recherche que la convivialité du logiciel.

4. Discussion

Dans cette partie, nous commentons les résultats des trois tâches de recherche et de l'évaluation avec la grille des outils. La première observation est que les trois logiciels sont en mesure d'accomplir les trois tâches de recherche, cependant, ces dernières nous permettent également de mettre en évidence les différences entre les trois outils.

4.1. Rapport entre affichage, facilité d'utilisation et rendement

Le rapport entre affichage, facilité d'utilisation et rendement est très étroit et il est observable dans les trois tâches. Dans la première, nous voyons qu'avec *MultiTrans Prism* et *myCAT* les temps pour accomplir la tâche sont presque identiques (respectivement 1,15 minute et 1,10 minute) alors qu'avec *ParaConc* ils doublent (3,30 minutes). Pour les trois outils, nous obtenons les mêmes résultats. La différence en termes de temps nécessaire à l'accomplissement de la tâche avec *ParaConc* est due à l'affichage, car, pour obtenir une réponse satisfaisante, dans trois cas sur cinq nous sommes passés par l'affichage plein texte. *ParaConc* ne permet pas de passer d'une occurrence à l'autre en plein texte, donc cela nous oblige à ouvrir le document d'origine à chaque fois. Il est donc clair que pour cette tâche, il est préférable d'avoir un affichage plein texte, ou au moins la possibilité de passer de l'affichage KWIC au plein texte sans devoir à chaque fois ouvrir le document d'origine.

Dans la tâche de recherche des collocations, si l'on considère uniquement les temps, l'outil le plus performant semble être *MultiTrans Prism* avec 7,20 minutes, suivi par *ParaConc* avec 9 minutes et par *myCAT* avec 13,08 minutes. Si l'on observe mieux les résultats, on voit que pour *ParaConc* nous avons effectué deux essais : dans un premier temps, nous avons lancé une recherche avec affichage des concordances, recherche obligatoire pour pouvoir ensuite passer aux collocations. Comme les collocations ne nous ont pas donné une réponse

satisfaisante, nous avons au final analysé les concordances en triant le contexte des occurrences. Les temps nécessaires sont donc plus importants par rapport à *MultiTrans Prism*, mais le nombre de passages au total n'est que de 4, et ce grâce à l'affichage KWIC. Mais, encore une fois, *ParaConc* paye le prix d'un affichage sur une seule colonne pour les deux langues qui ne permet pas d'identifier l'occurrence correspondante dans la langue cible de manière immédiate, comme c'est le cas avec *MultiTrans Prism*, qui, malgré les 37 passages pour consulter toutes les occurrences, reste le plus rapide grâce à une bonne mise en évidence des segments source et cible ainsi que de la séquence recherchée. La mauvaise performance de *myCAT* n'est pas due seulement au nombre de passages (45), mais surtout aux problèmes de repérage des segments cibles qui oblige à lire attentivement le document pour repérer la bonne portion de texte. Ceci pénalise l'outil dans deux tâches sur trois, d'autant plus que l'écart concernant le nombre de clics nécessaires à voir toutes les occurrences et leurs contextes entre *myCAT* et *MultiTrans Prism* est relativement bas : 1 clic d'écart dans la première tâche (5 pour *MultiTrans Prism* et 6 pour *myCAT*), 8 dans la deuxième (37 pour *MultiTrans Prism* et 45 pour *myCAT*) et zéro dans la troisième (16 clics). Pourtant, hormis pour la première tâche où *myCAT* est plus rapide de 5 secondes, dans les deux autres, il perd respectivement 5,48 minutes et 2,20 minutes.

Enfin, lors de la tâche de révision on voit que l'affichage KWIC de *ParaConc* permet d'analyser les résultats en 3 minutes et 4 passages, alors que pour *MultiTrans Prism* et *myCAT* 16 passages sont nécessaires et le temps double pour *myCAT* (6,30 minutes).

4.2. Rapport entre options de recherche et de personnalisation et rendement

L'importance de disposer de plusieurs options de recherche est bien visible dans les deux dernières tâches. En effet, dans la recherche des collocations de « titre » avec *ParaConc*, le tri des occurrences par contexte nous permet de prendre en compte uniquement les occurrences pertinentes et d'exclure toutes celles contenant, par exemple, « à titre », alors que *MultiTrans Prism* et *myCAT* nous obligent à consulter toutes les 37 occurrences.

Dans la troisième tâche, la recherche parallèle et par expressions régulières de *ParaConc* nous permet immédiatement de voir que « *scuola superiore* » n'est pas la bonne réponse car nous n'obtenons qu'une seule occurrence. À la deuxième recherche, le tri des occurrences et l'affichage KWIC dans les deux langues nous permettent d'identifier immédiatement la bonne séquence. Au contraire, avec les deux autres logiciels nous devons d'abord chercher

le singulier et ensuite le pluriel, car, comme il s'agit d'une séquence de deux termes, nous n'avons aucune option nous permettant d'effectuer les deux recherches simultanément (dans *myCAT*, la recherche « école supérieure OR écoles supérieures » ne fonctionne pas dans ce cas, car l'opérateur OU prend en compte seulement deux termes).

Nous n'avons à vrai dire utilisé aucune des options de personnalisation de *MultiTrans Prism* et de *myCAT*, tout simplement parce qu'aucune ne nous semblait utile pour améliorer les résultats dans les trois tâches que nous avons définies. Nous soulignons que, dans d'autres cas, ces options peuvent se révéler très utiles.

4.3. Fiabilité

Nous remarquons que *ParaConc* est plus fiable, car *MultiTrans Prism* et *myCAT* ne retrouvent aucun résultat lorsqu'on effectue une recherche avec des caractères interdits (par exemple les points) mais ne donne aucun message d'erreur. Le concordancier *myCAT* est fortement instable du point de vue de l'affichage, en particulier en ce qui concerne la mise en évidence des segments cibles. Cette instabilité peut s'expliquer par le fait que c'est un outil récent, qui a sûrement besoin d'être mis au point.

4.3. Bruit et silence

La tâche de recherche des collocations de « titre » (tâche 2) met en évidence une divergence entre le nombre d'occurrences trouvées (37 pour les trois outils) et le nombre d'occurrences pertinentes (16 pour les trois outils). Nous observons donc une grande présence de bruit. Toutefois, dans ce cas, le bruit n'est pas lié à un problème des outils, mais au fait que le terme « titre » est polysémique. Cela explique également pourquoi les collocations et les clusters ne nous aident pas à trouver la bonne réponse.

En revanche, dans la tâche 3 nous remarquons que *MultiTrans Prism* trouve 14 occurrences, contre 13 de *ParaConc* et 10 de *myCAT*. En effet, *MultiTrans Prism* est le seul outil qui retrouve « ecoles supérieures » qui contient une faute d'orthographe dans le document d'origine. Par contre, *myCAT* ne trouve que 10 occurrences, ce qui indique du silence pour au moins trois occurrences. Dans ce cas, il s'agit d'un problème dû au logiciel.

4.4. Synthèse des points forts et des points faibles des outils

Le point fort de *MultiTrans Prism* est sûrement l'interface, tout d'abord parce que la mise en évidence du segment source et de la séquence ainsi que le repérage du segment cible rend les résultats très lisibles en plein texte. De plus, c'est le seul des trois outils qui propose un aperçu du nombre total d'occurrences trouvées et de toutes les flexions. Le point faible est représenté par le nombre réduit d'options de recherche et de personnalisation de l'action.

ParaConc est un outil complet, mais il n'est pas intuitif, car l'interface de recherche est vide : il faut passer par la barre des menus pour avoir accès aux options de recherche et d'affichage ainsi qu'aux clusters et aux collocations. Il propose le nombre total d'occurrences trouvées, sans indiquer les flexions. En outre, le fait d'afficher les résultats sur une seule colonne pénalise leurs lisibilités. Il faut quand même souligner que la dernière version de cet outil remonte à 2001, ce qui explique pourquoi l'affichage n'est pas en ligne avec les outils plus récents.

Par contre, *myCAT* est un outil récent, ce qui n'explique pas les choix d'affichage des développeurs, car le repérage des segments cible n'est pas optimal. De plus, le logiciel propose le nombre de documents contenant l'occurrence recherchée, mais ne donne à aucun moment le nombre total d'occurrences. Pour voir le nombre d'occurrences par document, il faut arriver à la dernière occurrence, lorsque « Last Hit» apparaît. En ce qui concerne la partie du logiciel que nous avons testé, les points forts de *myCAT* sont le caractère intuitif et le fait de pouvoir effectuer la recherche sur plusieurs corpus simultanément ainsi que de passer d'un corpus à l'autre directement depuis l'interface de recherche, même si nous n'avons pas testé cette option dans cette étude.

4.5. Perspectives d'amélioration de l'évaluation

Notre évaluation pourrait sans doute être améliorée, en faisant accomplir les trois tâches à un groupe de traducteurs, ce qui permettrait d'avoir un cadre plus ample et donc plus objectif en ce qui concerne les temps nécessaires pour accomplir l'action, mais également en observant les options choisies par un groupe d'utilisateurs expérimentés afin de vérifier quelle est la méthode la plus utilisée et laquelle permet d'obtenir les meilleurs résultats en moins de temps. On pourrait également tester les outils en définissant d'autres tâches d'évaluation et diviser les groupes de test par rapport aux tâches à accomplir ou aux outils à

tester. Une option envisageable serait celle de tester toutes les fonctionnalités pour arriver ensuite à concevoir un outil idéal.

Néanmoins, notre méthode d'évaluation nous a permis de comparer nos trois outils, de mettre en évidence leurs caractéristiques principales ainsi que de tester leur efficacité.

Conclusion

Dans ce travail, nous avons cherché à répondre aux questions concernant l'apport concret des concordanciers parallèles au processus de traduction et leurs atouts en plus d'autres outils classiques auxquels le traducteur recourt normalement tels que les dictionnaires, les bases terminologiques, les documents écrits et les ressources en ligne. En particulier, nous avons vu dans la pratique comment les différentes modalités de recherche influencent la tâche de recherche et d'affichage des résultats.

Nous pouvons affirmer que les corpus parallèles sont sans doute utiles pour les traducteurs, car ils interviennent tout au long du processus de traduction. Nous avons présenté les trois types d'outils classiques d'exploitation des corpus parallèles (les mémoires de traduction, les aligneurs et les concordanciers parallèles) et nous avons choisis les trois outils à tester parmi ceux disponibles sur le marché : *MultiTrans Prism*, *ParaConc* et *myCAT*.
Nous avons ensuite proposé une méthode d'évaluation des outils d'exploitation de corpus parallèles, notamment en ce qui concerne l'aspect de la recherche dans les textes, dans le but de comprendre si les caractéristiques de ces outils répondent réellement aux besoins du traducteur et dans la perspective de définir quelles sont les fonctions indispensables.

Notre méthode se fonde sur le projet EAGLES (1996) et se divise en trois étapes : la définition des exigences de qualité, la préparation de l'évaluation et l'exécution de l'évaluation.
Lors de la première étape, nous avons défini les exigences de qualité en partant de la norme ISO/IEC 9126 : 1991 (1991), du modèle de tâche issu du génie du logiciel et de l'étude de Höge (2002).
Dans la préparation à l'évaluation, nous avons présenté nos deux grilles d'évaluation basées sur l'étude de Höge : une grille pour l'évaluation des tâches (action) ainsi qu'une grille d'évaluation des outils (objet). Ensuite, nous avons préparé les deux corpus dont nous nous sommes servis pour l'évaluation : le corpus CHOCOLAT et le corpus FORMATION.
Lors de l'exécution de l'évaluation, nous avons défini trois tâches de recherche, une pour la pré-traduction (documentation), une pour la traduction (recherche de collocations) et une

pour la révision (vérification de la terminologie) et nous sommes enfin passés au déroulement des tâches.

Après avoir testé la fonction de recherche dans nos trois outils (*MultiTrans Prism, ParaConc* et *myCAT*), nous concluons qu'aucun de ces trois n'est en mesure de répondre entièrement aux besoins des traducteurs. Il est vrai que l'outil idéal dépend de la tâche à accomplir, mais il est tout aussi vrai que ces trois outils sont peu flexibles et s'adaptent donc difficilement aux besoins des traducteurs. *MultiTrans Prism* dispose certes d'un bon affichage plein texte grâce au fait que les phrases sont distinguées par des couleurs différentes et que le segment recherché ainsi que le segment correspondant sont surlignés. Il offre plus d'informations concernant les occurrences trouvées, mais, en tant que mémoire de traduction, il permet d'effectuer uniquement des recherches de base et ne dispose pas d'un affichage KWIC. Or, comme nous l'avons vu, l'affichage plein texte semble plus efficace pour la phase de documentation, mais il pénalise l'utilisateur lors de la recherche de concordances. Dans ce cas, un affichage KWIC permet une meilleure lisibilité des résultats.

ParaConc est, parmi les concordanciers parallèles existants, le plus complet, mais il est pénalisé par un alignement insatisfaisant et par une interface trop « ancienne », sans aucune information complémentaire, donc peu intuitive. L'affichage KWIC des occurrences rassemble les deux langues sur une seule colonne, ce qui rend difficile le repérage des segments cibles. Nous déduisons donc que la solution idéale serait de disposer d'un affichage sur deux colonnes et d'avoir la possibilité de choisir à tout moment entre un affichage plein texte et KWIC.

Le concordancier *myCAT* est instable et présente une interface trop basique et des fonctions de recherche qui fonctionnent correctement avec un seul terme ou avec une séquence exacte, mais qui échouent lorsqu'on tente d'utiliser une séquence de mots avec un joker ou un opérateur booléen. D'ailleurs, le manuel indique bien que les jokers et les opérateurs booléens ne sont pas conçus pour gérer des séquences de mots. En revanche, l'aspect innovant de *myCAT* est représenté par un alignement automatique basé sur le traducteur automatique statistique *Mose* et par le fait qu'il est capable d'accepter plusieurs formats de textes (PDF, Microsoft Office ou OpenOffice).

En ce qui concerne les fonctions de recherche, sur la base de notre évaluation, nous pouvons affirmer qu'un outil complet devrait offrir un maximum de flexibilité et permettre d'effectuer au moins les recherches par mot avec ou sans joker, par troncature, avec l'opérateur OU, par cooccurrents ainsi que par expressions régulières et les recherches parallèles. Les recherches par expressions régulières apparaissent à première vue compliquées, mais elles permettent de mieux cibler les recherches. Quant aux recherches parallèles, elles sont fondamentales dans un outil d'exploitation de corpus parallèle, car c'est la seule fonction qui le distingue des concordanciers monolingues. Comme nous l'avons vu, c'est d'ailleurs la meilleure solution pour vérifier immédiatement si l'équivalent d'un terme que nous avons en tête est correct.

L'idée d'avoir un outil complet nous suggère également qu'il serait utile d'avoir la possibilité d'afficher des informations complémentaires et donc de disposer des listes de mots, des collocations ainsi que des clusters. Comme le précise Bowker (2001 : 349), les listes de mots peuvent aider le traducteur à choisir quel terme utiliser face à de potentiels synonymes ou à des équivalents de traduction. Les clusters et les collocations permettent d'analyser de quelle manière sont associées les unités et donc de voir immédiatement les cooccurrents d'un terme et le nombre de fois qu'ils apparaissent ensemble. Cela peut être vu comme une première approche pour identifier les cooccurrents d'une séquence pour ensuite les analyser dans leurs contextes à travers des recherches plus spécifiques.

Dans une perspective d'évolution des concordanciers parallèles, nous nous sommes interrogés sur le fait que, contrairement aux autres outils de TAO, à l'état actuel, il n'existe que très peu que de concordanciers parallèles et qu'ils sont peu diffusés parmi les traducteurs. La raison pourrait être justement que, comme nous l'avons vu, actuellement aucun des outils existants n'est réellement capable de répondre aux besoins des utilisateurs et que la phase de préparation et de nettoyage des textes est très longue.

Une autre raison non négligeable est sans doute à rechercher dans le fait que la plupart des traducteurs et des agences de traduction qui utilisent des logiciels autres que des éditeurs de textes, se penchent sur les outils d'aide à la traduction qui incluent au moins une mémoire de traduction avec une interface de recherche, une base terminologique et un éditeur de texte. Or, comme nous l'avons vu, les mémoires de traduction sont considérées

elles-mêmes des bitextes qu'il est possible d'exporter au format TMX. Ce point nous permet d'envisager deux évolutions différentes.

La première consisterait à concevoir un concordancier parallèle capable d'exploiter le format TMX afin de permettre l'importation et l'exportation des bitextes, ce qui le rendrait à même d'interagir avec les autres outils de TAO. Ceci permettrait donc d'exploiter les corpus parallèles déjà stockés dans les mémoires de traduction en évitant de devoir répéter les phases de sélection, nettoyage et d'alignement des textes. C'est d'ailleurs la solution proposée par le concordancier *AntPConc* (voir la section 1.2.1 du chapitre 3), qui n'accepte cependant aucun autre format et qui ne permet pas de corriger l'alignement.

La deuxième hypothèse serait d'intégrer le concordancier parallèle à l'interface de recherche dans les outils de TAO, de manière à les enrichir en exploitant les fonctions de recherche des concordanciers tout en utilisant des mémoires de traduction constamment mises à jour.

Une intégration du concordancier parallèle dans les outils d'aide à la traduction permettrait également d'améliorer nettement le repérage des segments cibles lors de la recherche en développant un algorithme capable d'exploiter les bases terminologiques pour la mise en évidence du terme correspondant dans la langue cible, à l'instar de *TradooIT* (voir la section 1.2.4 du chapitre 3).

Les possibilités d'évolution sont donc multiples et sans doute nécessaires afin de permettre aux concordanciers parallèles et aux mémoires de traduction de devenir des outils indispensables au traducteur.

Bibliographie

ADOLPHS, Svenja, 2006 : *Introducing – Electronic Text Analysis – A practical guide for language and literary studies.* London and New York : Routledge.

AIJMER, Karin, 2008 : “Parallel and comparable corpora” dans *Corpus Linguistics An International Handbook – Volume 1* (eds.) LÜDELING Anke & KYTÖ, Merja. Berlin : Walter de Gruyter.

ASTON, Guy, 2000 : "I corpora come risorse per la traduzione e per l'apprendimento" dans *I corpora nella didattica della traduzione. Corpus Use Learning to Translate. Atti del Seminario di Studi Internazionali* (eds.) BERNARDINI, Silvia & ZANETTIN, Federico. Bologna : Cooperativa Libraria Universitaria Editrice.

AUSTERMÜHL, Frank, 2001 : *Electronic tools for translator*. Manchester : St. Jerome Publishing.

BAKER, Mona, 2000 : “Linguistica dei corpora e traduzione. Per un'analisi del comportamento linguistico dei traduttori professionisti” dans *I corpora nella didattica della traduzione. Corpus Use Learning to Translate. Atti del Seminario di Studi Internazionali* (eds.) BERNARDINI, Silvia & ZANETTIN, Federico. Bologna : Cooperativa Libraria Università Editrice Bologna.

BAKER, Paul – HARDIE, Andrew – MCENERY, Tony, 2006 : *A Glossary of Corpus Linguistics*. Edinburgh : Edinburgh University Press Ltd.

BARLOW, Michael, 2004 : “Parallel Concordancing and Translation” dans *Translating and the computer 26 – Proceeding of the Twenty-sixth International Conference on Translating and the Computer, 18-19 November* 2004, London. London : Aslib/IMI.

BOWKER, Lynne, 2001 *: “*Towards a Methodology for a Corpus-Based Approach to Translation Evaluation” dans *Meta : journal des traducteurs / Meta: Translators' Journal, Vol. 46 num. 2.* Montréal : Les Presses de l'Université de Montréal*.*

BOWKER, Lynne, 2002 : *Computer-aided translation technology : a practical introduction*. Ottawa : University of Ottawa Press, Didactic of Translation Series.

BOWKER, Lynne & PEARSON, Jennifer, 2002 : *Working with Specialized Language: A practical guide to using corpora*. London & New York : Routledge.

BROWN, Peter F., LAI, Jennifer C. & MERCER Robert L., 1991 : “Aligning Sentences in Parallel Corpora” dans *Proceedings ALC ‘91 of the 29th Annual Meeting of the Association for Computational Linguistics.* Stroudsburg : Association for Computational Linguistics.

CAELEN, Jean & VILLASEÑOR-PINEDA, Luis, 1997 : "Dialogue homme-machine et apprentissage" dans *Apprentissage par l'interaction* (eds.) ZREIK, Khaldoun. Paris : Europia Productions .

CHEN, Yu & EISELE, Andreas, 2012 : “MultiUN v2: UN Documents with Multilingual Alignments” dans *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (eds.) CALZOLARI, Nicoletta, CHOUKRI, Khalid, DECLERCK, Thierry, DOĞAN, Mehmet Uğur, MAEGAARD, Bente, MARIANI, Joseph, ODIJK, Jan & PIPERIDIS, Stelios. Istanbul : European Language Resources Association (ELRA).

CHIAO, Yun-Chuang, KRAIF, Olivier, LAURENT, Dominique, HUYEN NGUYEN, Thi Minh, SEMMAR, Nasredine, STUCK, François, VERONIS, Jean & ZAGHOUANI, Wajdi Z, 2006 : "Evaluation of multilingual text alignment systems : the ARCADE II” dans *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).* Genova : European Language Resources Association (ELRA).

CHURCH, Kenneth Ward, 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*. Stroudsburg : Association for Computational Linguistics.

DÉJEAN, Hervé & GAUSSIER, Éric, 2002 : “Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables” dans *Lexicometrica No spécial 2002*.

DIPPER, Stefanie, SEISS, Melanie, ZINSMEISTER, Heike, 2012 : “The Use of Parallel and Comparable Data for Analysis of Abstract Anaphora in German and English” dans *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (eds.)* CALZOLARI, Nicoletta, CHOUKRI, Khalid, DECLERCK, Thierry, DOĞAN, Mehmet Uğur, MAEGAARD, Bente, MARIANI, Joseph, ODIJK, Jan & PIPERIDIS, Stelios. Istanbul : European Language Resources Association (ELRA).

ESSELINK, Bert, 1998 : *A Practical Guide to Software Localisation*. Amsterdam : John Benjamins Publishing Company.

GALE, William A. & CHURCH, Kenneth, Ward, 1993 : “A program for aligning sentences in bilingual corpora” dans *Computational Linguistics Vol. 19, Num. 1*. Trier : Univerersität Trier.

HABERT, Benoît, 2001 : "Des corpus représentatifs : de quoi, pour quoi, comment ?" dans *Actes de la journée Linguistique et corpus* (eds.) BILGER, Mireille. Perpignan : Presses de l’Université de Perpignan.

HABERT, Benoît – FABRE, Cécile – ISSAC, Fabrice, 1998 : *De l’écrit au numérique – Constituer, normaliser et exploiter les corpus électroniques*. Paris : InterÉditions/Masson.

HABERT, Benoît, NAZARENKO, Adeline, SALEM, André, 1997 : *Les linguistiques de corpus*. Paris : A. Colin.

HANSEN-SCHIRRA, Silvia & TEICH, Elke, 2009 : “Corpora in human translation” dans *Corpus Linguistics – An International Handbook* (eds.) LÜDELING, Anke & KYTÖ, Merja. Berlin : Walter de Gruyter GmbH & Co.

HÖGE, Monika, 2002 : *Towards a framework for the evaluation of translators' aids systems*. Helsinki : Helsinki University Press.

ISABELLE, Pierre, 1992 : “La bi-textualité : vers une nouvelle génération d’aides à la traduction et la terminologie” dans *Meta : journal des traducteurs / Meta: Translators' Journal, Vol. 37 num. 4*. Montréal : Les Presses de l'Université de Montréal.

ISABELLE, Pierre & WARWICK-ARMSTRONG, Susan, 1993 : "Les corpus bilingues : une nouvelle ressource pour le traducteur" dans *La traductique - Etudes et recherches de traduction par ordinateur* (eds.) BOUILLON Pierette & CLAS André. Montréal : Les Presses de l’Université de Montréal.

ISO/IEC 25010 : 2011 (Standard), 2011: *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models.* Genève: ISO copyright office.

ISO/IEC 9126 : 1991 (Standard), 1991 : *Information Technology - Software Product Evaluation, Quality Characteristics and Guidelines for their Use.* International Organisation for Standardization. Genève: ISO copyright office.

ISO/IEC 9126 : 2000 (Standard), 2000 : *Software Engineering-Product Quality : External Metrics International Standard ISO/IEC 9126*. Genève: ISO copyright office.

JOHANSSON, Stig, EBELING, Jarle & HOFLAND, Knut, 1996 : "Coding and aligning the English-Norwegian parallel corpus" *dans Languages in Contrast : Papers from a Symposium on Text-based Cross-linguistic Studies (Lund, 4-5 March 1994)* (eds.) AIJMER, Karin, ALTENBERG, Bengt & JOHANSSON, Mats. Lund : Lund University Press.

KAY, Martin & RÖSCHEISEN, Martin, 1988: *Text-translation alignment. Technical Report.* Palo Alto : Xerox Palo Alto Research Center.

KAY, Martin & RÖSCHEISEN, Martin, 1993 : "Text-translation alignment" dans *Journal Computational Linguistics - Special issue on using large corpora: I archive Volume 19 Issue 1.* Cambridge : MIT Press Cambridge.

KILGARRIFF, Adam, 2010 : "Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project" dans *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010.* Malta.

KOEHN, Philipp, 2014 : *Moses – Statistical Machine Translation System - User Manual and Code Guide*.

KRAIF, Olivier, 2002 : "Translation alignment and lexical correspondences : a methodological reflection" dans *Lexis in Contrast*, (eds.) ALTENBERG, Bengt & GRANGER, Sylviane. Amsterdam : John Benjamins Publishing Company.

KRAIF, Olivier, 2006 : "Qu'attendre de l'alignement de corpus multilingues" dans *Revue Traduire, 4e Journée de la traduction professionnelle*, Société Française des Traducteur, N° 210. Version en ligne. Grenoble : Université Stendhal Grenoble 3.

KÜNZLI, Alexander, 2001 : "Experts versus novices : l'utilisation de sources d'information pendant le processus de traduction" dans *Meta : journal des traducteurs / Meta: Translators' Journal, Vol. 46, num. 3*. Montréal : Les Presses de l'Université de Montréal.

L'HOMME, Marie-Claude, 2004 : *La terminologie : principes et techniques.* Montréal : Les Presses de l''Université de Montréal, Coll. « Paramètres ».

L'HOMME, Marie-Claude, 2008 : *Initiation à la traductique (2e éd. Rev. et augm.)*. Montréal : Les Presses de l'Université de Montréal, Coll. «Paramètres».

LAPORTE, Eric, 2000 : "Mots et niveau lexical" dans *Ingénierie des langues* (eds.) PIERREL, Jean-Marie. Paris : Editions Hermès.

LAVIOSA, Sara 2003 : "Corpora and the translator" dans *Computers and Translation : A translator's guide* (eds.) SOMERS, Harold. Amsterdam, Philadelphia : John Benjamins Publishing Company.

LE SERREC, Annaïch, 2012 : *Analyse comparative de l'équivalence terminologique en corpus parallèle et en corpus comparable : application au domaine du changement climatique*. Montréal : Université de Montréal – Papyrus : Dépôt institutionnel numérique.

MASSION, François, 2005 : *Translation Memory Systeme im Vergleich*. Reutlingen : Doculine Vertrags GmbH.

MCENERY, Anthony M. & OAKES, Michael P., 1995: "Sentence and word alignment in the CRATER project : methods and assessment" dans *Proceedings of the EACL-SIGDAT Workshop* (eds.) WARWICK-ARMSTRONG, Susan. Dublin : ACL.

MCENERY, Tony & WILSON, Andrew, 1996 (second edition 2001) : *Corpus linguistics.* Edinburgh : Edinburg University Press.

MCENERY, Tony & XIAO, Richard, 2008 : "Parallel and Comparable Corpora : What is Happening ?" dans *Incorporating Corpora : The Linguist and the Translator* (eds.) ANDERMAN, Gunilla & ROGERS, Margaret. Clevedon : Multilingual Matters.

MELAMED, I. Dan, 1996 : *A Geometric Approach to Mapping Bitext Correspondence*. Philadelphia : University of Pennsylvania.

MELAMED, I. Dan, 2001 : *Empirical methods for exploting parallel text*. Massachusetts : The TIM Press.

MEYER, Ingrid, 2001 : "Extracting Knowledge-rich Context for Terminography" dans *Recent Advances in Computational Terminology* (eds.) BOURIGAULT, Didier, JACQUEMIN, Christian & L'HOMME, Marie-Claude. Amsterdam/Philadelphia : John Benjamins Publishing Company.

MULTICORPORA, 2012 : *MultiTrans Prism Didacticiel*. Montréal : MultiCorpora R&D INC.

MULTICORPORA, 2013 : *Configuration minimale pour MultiTrans Prism*. Montréal : MultiCorpora R&D INC.

MUNDAY, Jeremy, 1998 : "A Computer-assisted Approach to the Analysis of Translation" dans *Meta : journal des traducteurs / Meta: Translators' Journal, Vol. 43, num. 4*. Montréal : Les Presses de l'Université de Montréal.

PIPERIDIS, Stelios, BOUTSIS, Sotiris & PAPAGEORGIOU, Harris , 2000 : "From sentences to words and clauses" dans *Parallel Text Processing : Alignment and use of translation corpora* (eds.) VÉRONIS, Jean. Dordrecht : Kluwer Academic Publishers.

QUAH, Chiew Kin, 2006 : *Translation and Technology.* Basingstoke : Palgrave Textbooks in translation and interpreting.

SELJAN, Sanja, TADIĆ, Marko, AGIĆ, Željko, ŠNAJDER, Jan, DALBELO BAŠIĆ & Bojana, OSMANN, Vjekoslav, 2010 : "Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora" dans *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (eds.) CALZOLARI, Nicoletta, CHOUKRI, Khalid, MAEGAARD, Bente, MARIANI, Joseph, ODIJK, Jan, PIPERIDIS, Stelios, ROSNER, Mike & TAPIAS, Daniel. Malta : European Language Resources Association (ELRA).

SIMARD, Michel, FOSTER, George F., ISABELLE Pierre, 1992 : "Using cognates to align sentences in bilingual corpora" dans *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*. Montréal.

SINCLAIR, John, 1996 : *Preliminary recommendations on Corpus Typology. EAG-TCWG-CTYP/P, Version May 1996*, Pisa : Eagles.

SOMERS, Harold, 2003 : "Translation memory systems" dans *Computer and translation : a translator's guide* (eds.) SOMERS, Harold. Amsterdam/Philadelphia : John Benjamins Publishing Company.

VÉRONIS, Jean, 2000 : “Alignement de corpus multilingues” dans *Ingénierie des langues* (eds.) PIERREL, Jean-Marie. Paris : Editions Hermès.

VÉRONIS, Jean & LANGLAIS, Philippe, 2000 : “Evaluation of parallel text alignment systems : ARCADE” dans *Parallel Text Processing: Alignment and use of translation corpora* (eds.) VÉRONIS, Jean. Dordrecht : Kluwer Academic Publishers.

WEBB, Lynn E., 1992 : *Advantages And Disadvantages Of Translation Memory : A Cost/Benefit Analysis*. Monterey : Monterey Institute of International Studies.

ZANETTIN, Federico, 2012 : *Translation-Driven Corpora - Corpus Resources for Descriptive and Applied Translation Studies*. Manchester : St. Jerome Publishing.

ZWEIGENBAUM, Pierre & HABERT, Benoît, 2006 : "*Faire se rencontrer les parallèles : regards croisés sur l’acquisition lexicale monolingue et multilingue*" dans *Glottopol – revue de sociolinguistique en ligne n°* 8 – juillet 2006. Rouen : Université de Rouen.

Webographie

Aligned Hansards of the 36th Parliament of Canada: http://www.isi.edu/natural-language/download/hansard/ - Consulté le 23/05/2014

AlignFactory: http://www.terminotix.com/index.asp?name=AlignFactory&content=item&brand=1&item=4&lang=fr - Consulté le 23/05/2014

Alinéa : Site personnel de Olivier Kraif : http://olivier.kraif.u-grenoble3.fr/ - Consulté le 23/05/2014

British National Corpus : http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro - Consulté le 15/01/2014

Concordancier : Loria / Logiciels développés et ressources: http://led.loria.fr/outils.php.html#5- Consulté le 22/05/2013

EAGLES WORK GROUP, 1996 : *EAGLES Evaluation Group. Final Report*. Rapport n°EAG-EWG-PR.2. Copenhague : **Center for Sprogteknologi.** http://www.issco.unige.ch/en/research/projects/ewg96/index.html - Consulté le 25/05/2014

Europarl : http://www.statmt.org/europarl/ - Consulté le 23/05/2014.

European Telecommunications Standard Institute (ETSI) : News & Events > Latest News > New Industry Specification Group on Localisation Industry Standards : http://www.etsi.org/news-events/news/364-news-release-13-july-2011?highlight=YToxOntpOjA7czo0OiJsaXNhIjt9 - Consulté le 04/12/2013

Laurence Anthony's Website : http://www.antlab.sci.waseda.ac.jp - Consulté le 23/05/2014 AntPConc

Multiconcord : un logiciel lorrain pour trouver le mot adapté au contexte, dans 12 langues : http://ceres.univ-lorraine.fr/contentId%3D9272 - Consulté le 23/05/2014

MultiConcord : un logiciel lorrain pour trouver le mot adapté au contexte… dans 12 langues : http://eureka.lorraine.eu/jahia/Jahia/fr/pid/1968?actu=19835 - Consulté le 23/05/2014

MultiCorpora : http://multicorpora.com/fr/societe/ - Consulté le 25/05/2014

MultiCorpora : https://www.multicorpora.ca/CARTStep1_f.php - Consulté le 15/12/2013

MultiCorpora / Mémoire de traduction : http://www.multicorpora.com/fr/multitrans-prism-fr/memoire-de-traduction/- Consulté le 15/06/2013

myCAT : Fondation Olanto : http://olanto.org/fr/fondation - Consulté le 23/05/2014

myCAT - Quote Detector User Manual : http://srv1.olanto.org/TranslationText/# - Consulté le 25/05/2014

myCAT - Self-Quote Detector User Manual : http://srv1.olanto.org/mySelfQD/# - Consulté le 25/05/2014

myCAT - Text Aligner User Manual : http://srv1.olanto.org/TranslationText/# - Consulté le 25/05/2014

OPUS : http://opus.lingfil.uu.se/ - Consulté le 28/05/2014

OSCAR Recommendation, 2005 : *OscarTMX 1.4b Specification* : http://www.gala-global.org/oscarStandards/tmx/tmx14b.html#refLISA - Consulté le 23/05/2014

ParaConc : http://athel.com/product_info.php?products_id=81&osCsid=95cd6a3d58b331ddc04a162c383fa3d2 - Consulté le 23/05/2014

RR Donnelley : Acquisition du fournisseur de solutions technologiques de traduction MultiCorpora : http://www.rrdonnelley.com/languagesolutions/fr/news/2014/03102014.aspx - Consulté le 23/05/2014

RR Donnelley : http://www.rrdonnelley.com/languagesolutions/fr/locations/ - Consulté le 23/05/2014

SDL Trados Studio : http://www.sdl.com/it/products/sdl-trados-studio/ - Consulté le 27/05/2014

Similis : http://similis.fr/ - Consulté le 27/04/2014

TradooIT : http://blog.tradooit.com/search?updated-max=2012-09-10T18:09:00-07:00&max-results=7&reverse-paginate=true - Consulté le 23/05/2014

TradooIT : https://www.tradooit.com/index.php - Consulté le 23/05/2014

Wordfast : http://www.wordfast.net/ - Consulté le 27/05/2014

Annexes

Annexe A : Grilles d'évaluation des tâches

Grille d'évaluation de la tâche 1 : documentation

Caractéristiques de qualité	Aspects qualitatifs	Critères	MultiTrans Prism	ParaConc	myCAT
Capacité fonctionnelle	différentes options de recherche	recherche par mot	oui	non	oui
		recherche par mot avec joker	ND	oui	non
		recherche par troncature	non	non	non
		opérateur OU	ND	non	non
		cooccurrents	ND	non	non
		recherche parallèle	ND	non	ND
		recherche par filtre	ND	ND	ND
		recherche par fuzzy	non	ND	ND
		recherche par expressions régulières	ND	non	ND
	choix de l'affichage	liste de mots	non	non	ND
		cluster	ND	non	ND
		collocations	ND	non	ND
		concordances	oui	oui	oui
	interaction avec d'autres objets/fonctions	possibilité de copier/coller les informations	oui	non	non
	personnalisation de l'action	changement de la direction des langues	non	non	non
		possibilité de lancer les recherches simultanément dans plusieurs corpus à la fois	non	ND	non
		possibilité d'affiner la recherche par rapport à la date, au corpus, etc.	ND	ND	non
		possibilité de passer d'un corpus à l'autre	ND	ND	non
		possibilité de choisir l'ordre d'affichage des occurrences selon le document/corpus d'origine	ND	ND	non
		possibilité de trier les résultats par rapport à la date, l'auteur, etc.	non utilisé	ND	non utilisé
		possibilité de trier les occurrences selon le contexte	ND	non	ND
		possibilité de régler le nombre de caractères/mots affichés	ND	oui	ND
		concordance: possibilité de voir le document d'origine	ND	oui	non
Facilité d'utilisation	aide nécessaire pendant l'accomplissement de l'action	nombre de recours au manuel ou à l'aide	0	0	0
Rendement	temps nécessaire à l'accomplissement de l'action	temps nécessaire à accomplir la tâche (en minutes)	1.15 min.	3.30 min	1.10 min
		nombre d'essais nécessaires pour accomplir la tâche (maximum 6)	1	1	1
		nombres de passages nécessaires afin de voir toutes les occurrences différentes	0	0	6
		nombre de clics nécessaires afin de voir toutes les occurrences et leur contexte	5	5	6
	exactitude des résultats	réussite de la tâche	oui	oui	oui
		nombre d'occurrences trouvées	5	5	5
		nombre d'occurrences pertinentes	5	5	5
	fiabilité des résultats	stabilité des résultats à la répétition de la tâche	oui	oui	oui

Grille d’évaluation de la tâche 2 : traduction

Caractéristiques de qualité	Aspects qualitatifs	Critères	MultiTrans Prism	ParaConc	myCAT
Capacité fonctionnelle	différentes options de recherche	recherche par mot	non	non	non
		recherche par mot avec joker	ND	non	oui
		recherche par troncature	oui	non	non
		opérateur OU	ND	non	non
		cooccurrents	ND	non	non
		recherche parallèle	ND	non	ND
		recherche par filtre	ND	ND	ND
		recherche par fuzzy	non	ND	ND
		recherche par expressions régulières	ND	oui	ND
	choix de l’affichage	liste de mots	non	non	ND
		cluster	ND	non	ND
		collocations	ND	oui	ND
		concordances	oui	oui	oui
	interaction avec d'autres objets/fonctions	possibilité de copier/coller les informations	non	non	non
	personnalisation de l'action	changement de la direction des langues	non	non	non
		possibilité de lancer les recherches simultanément dans plusieurs corpus à la fois	non	ND	non
		possibilité d'affiner la recherche par rapport à la date, au corpus, etc.	ND	ND	non
		possibilité de passer d'un corpus à l'autre	ND	ND	non
		possibilité de choisir l'ordre d'affichage des occurrences selon le document/corpus d'origine	ND	ND	non
		possibilité de trier les résultats par rapport à la date, l'auteur, etc.	non utilisé	ND	non utilisé
		possibilité de trier les occurrences selon le contexte	ND	oui	ND
		possibilité de régler le nombre de caractères/mots affichés	ND	non	ND
		concordance: possibilité de voir le document d'origine	ND	oui	non
Facilité d'utilisation	aide nécessaire pendant l'accomplissement de l'action	nombre de recours au manuel ou à l'aide	0	1	0
Rendement	temps nécessaire à l'accomplissement de l’action	temps nécessaire à accomplir la tâche (en minutes)	7.20 min	9.00 min	13.08 min
		nombre d'essais nécessaires pour accomplir la tâche (maximum 6)	1	2	1
		nombres de passages nécessaires afin de voir toutes les occurrences différentes	37	4	45
		nombre de clics nécessaires afin de voir toutes les occurrences et leur contexte	37	6	45
	exactitude des résultats	réussite de la tâche	oui	oui	oui
		nombre d'occurrences trouvées	37	37	37
		nombre d'occurrences pertinentes	16	16	16
	fiabilité des résultats	stabilité des résultats à la répétition de la tâche	oui	oui	oui

Grille d'évaluation de la tâche 3 : vérification de la terminologie

Caractéristiques de qualité	Aspects qualitatifs	Critères	MultiTrans Prism	ParaConc	myCAT
Capacité fonctionnelle	différentes options de recherche	recherche par mot	non	non	non
		recherche par mot avec joker	ND	non	non
		recherche par troncature	non	non	non
		opérateur OU	ND	non	non
		cooccurrents	ND	non	oui
		recherche parallèle	ND	oui	ND
		recherche par filtre	ND	ND	ND
		recherche par fuzzy	oui	ND	ND
		recherche par expressions régulières	ND	oui	ND
	choix de l'affichage	liste de mots	non	non	ND
		cluster	ND	non	ND
		collocations	ND	non	ND
		concordances	oui	oui	oui
	interaction avec d'autres objets/fonctions	possibilité de copier/coller les informations	non	non	non
	personnalisation de l'action	changement de la direction des langues	non	non	non
		possibilité de lancer les recherches simultanément dans plusieurs corpus à la fois	non	ND	non
		possibilité d'affiner la recherche par rapport à la date, au corpus, etc.	ND	ND	non
		possibilité de passer d'un corpus à l'autre	ND	ND	non
		possibilité de choisir l'ordre d'affichage des occurrences selon le document/corpus d'origine	ND	ND	non
		possibilité de trier les résultats par rapport à la date, l'auteur, etc.	non utilisé	ND	non utilisé
		possibilité de trier les occurrences selon le contexte	ND	oui	ND
		possibilité de régler le nombre de caractères/mots affichés	ND	non	ND
		concordance: possibilité de voir le document d'origine	ND	non	non
Facilité d'utilisation	aide nécessaire pendant l'accomplissement de l'action	nombre de recours au manuel ou à l'aide	0	0	0
Rendement	temps nécessaire à l'accomplissement de l'action	temps nécessaire à accomplir la tâche (en minutes)	4.10 min.	3 min.	6.30 min.
		nombre d'essais nécessaires pour accomplir la tâche (maximum 6)	2	2	2
		nombres de passages nécessaires afin de voir toutes les occurrences différentes	16	4	16
		nombre de clics nécessaires afin de voir toutes les occurrences et leur contexte	16	4	16
	exactitude des résultats	réussite de la tâche	oui	oui	oui
		nombre d'occurrences trouvées	14	13	10
		nombre d'occurrences pertinentes	14	13	10
	fiabilité des résultats	stabilité des résultats à la répétition de la tâche	oui	oui	oui

Annexe B : Grille d'évaluation des outils

Caractéristiques de qualité	Aspects qualitatifs	Critères	MultiTrans Prism	ParaConc	myCAT
Fiabilité	type d'erreur	erreur d'affichage	oui	non	oui
		blocage du logiciel	non	non	non
	action qui provoque l'erreur	mauvaise manipulation de l'utilisateur	non	non	non
		annulation d'une action	non	non	non
		utilisation d'une option du logiciel	oui	non	oui
Facilité d'utilisation	présentation de l'objet (interface)	concordance: affichage plein texte	oui	non	oui
		concordance: affichage KWIC	non	oui	non
		nombre maximum de caractères pour la recherche	indéfini mais sans point	parametrable	indéfini mais sans point
		mise en évidence de l'élément recherché	oui	oui	oui
		repérage du segment traduit	oui	oui	oui
		informations concernant la source de la séquence affichée	non évaluable	corpus et ligne	document
		affichage du nombre total d'occurrences	oui	oui	non
		affichage des flexions d'une occurrence	oui	non	non
Rendement	temps nécessaire à la performance	nombre de passages nécessaires pour arriver à la fenêtre de recherche	0	2	0
		temps (sec.) nécessaire à l'affichage des résultats	< 2 sec.	< 1 sec.	< 2 sec.