La traduction spécialisée basée sur les corpus : une expérience dans le domaine informatique

Jean-Pierre Colson
Institut libre Marie Haps (Bruxelles)
Université catholique de Louvain (Louvain-la-Neuve)

Synergies *Tunisie* n° 2 - 2010 pp. 115-123

Résumé: *Les études traductologiques basées sur les corpus ont aujourd'hui l'image d'un domaine de recherche théorique, sans incidence pratique réelle sur le travail quotidien des traducteurs. Cette contribution présente les résultats d'une expérience qui illustre toute l'utilité d'une telle approche dans le cas de la traduction spécialisée, en l'occurrence dans le domaine informatique. Lorsque les ouvrages de référence traditionnels tels que les dictionnaires et les banques de données terminologiques font défaut, comme c'est souvent le cas en matière de technologie, l'extraction linguistique de données à partir de la Toile ou de vastes corpus linguistiques permet d'offrir au traducteur des solutions rapides et efficaces. Ceci vaut particulièrement lorsque l'on a recours à des corpus indexés et à de grandes collections de données extraites de la Toile et filtrées avec précision.*

Mots-clés: *collocations, extraction automatique, traduction spécialisée, corpus.*

Abstract : *After a promising start, corpus-based translation is now often seen as a theoretical research line, with few practical implications for the everyday work of translators. This paper reports the results of an experiment illustrating the usefulness of this approach in the case of specialized translation within the field of computer science. When traditional works such as dictionaries or terminology databases are incomplete, as is often the case for technology, extracting linguistic results from the Web or from huge linguistic corpora may present the translator with quick and efficient solutions. This is especially the case if indexed corpora and large web-based collections with a high degree of filtering are used.*

Keywords: *collocations, automatic extraction, specialized translation, corpora.*

1. Introduction

La traduction spécialisée, lieu de rencontre entre professionnels de la langue, terminologues, spécialistes d'un domaine, monde des entreprises, se retrouve tout naturellement au cœur de la révolution technologique actuelle. Linguistique de corpus, linguistique computationnelle, théorie de l'information, lexicographie automatisée : autant de disciplines dont le traducteur technique espère tirer profit pour l'assister dans sa mission souvent périlleuse de traduire des textes de plus en plus pointus, qui recourent

à un jargon spécialisé en évolution constante. Néologismes, nouvelles technologies, nouvelle version d'un produit, découverte scientifique : autant d'écueils pour la traduction technique, d'autant que les ouvrages lexicographiques ont du mal à suivre le tempo.

Depuis quelques années, on admet généralement que les traducteurs généraux ou techniques bénéficient largement de l'apport des corpus linguistiques. Dans la pratique, par contre, il est rare de voir les bureaux de traduction utiliser quotidiennement des corpus linguistiques, hormis les traductions déjà effectuées. Et pourtant, le fondement même de la linguistique de corpus (Sinclair 2001) est une remise en question des dictionnaires traditionnels, presque toujours en-deçà de la véritable richesse des corpus authentiques.

En langue de spécialité, le problème est complexe pour le traducteur, car il doit dans ce cas recourir à un corpus spécialisé et récent émanant du domaine précis qu'il est occupé à traduire. A défaut, il peut tenter de le constituer lui-même, ce qui risque de lui prendre pas mal de temps, surtout s'il n'est ni informaticien ni linguiste… Ceci explique sans doute cela : les souhaits de la *traduction basée sur les corpus* (Laviosa 2002) restent souvent des vœux pieux. Dans la présente contribution, nous tenterons d'illustrer par une expérience dans le domaine informatique, que la traduction technique basée sur les corpus est bel et bien une alternative possible.

Le choix de l'informatique se justifie par deux éléments. Tout d'abord, la terminologie et la lexicographie spécialisée peuvent déjà s'appuyer dans ce domaine sur les brillants travaux de M.-C. L'Homme (2004, 2008, 2010) ; en outre, l'informatique constitue bien évidemment l'un des terrains de prédilection de la Toile, à partir de laquelle des corpus spécialisés peuvent être assemblés.

Afin de démontrer l'utilité des corpus pour la traduction spécialisée, nous partirons donc du domaine déjà bien balisé de l'informatique. Le *Dictionnaire fondamental de l'informatique et de l'internet* ou DiCoInfo, réalisé par l'université de Montréal sous la direction de M.-Cl. L'Homme (2008), ainsi que la banque de données terminologiques multilingue de l'Union Européenne, l'IATE ou *Interactive Terminology for Europe*[1] nous serviront de point de départ.

Le DiCoInfo représente un remarquable lexique spécialisé, qui fournira au traducteur anglais-français ou français-anglais une multitude de renseignements précieux en matière d'informatique et d'internet. L'IATE, quant à elle, constitue l'une des bases les plus complètes au monde dans tous les domaines de spécialité actuels. Quel pourrait être, dès lors, l'apport des corpus en la matière ?

Nous partirons du principe qu'un corpus pertinent pour notre expérience sera forcément extrait de la Toile. Il devra en outre être disponible en anglais et en français (à l'instar du DiCoInfo) et, idéalement, représenter une taille suffisante. Nous avons choisi de nous intéresser également aux collocations spécialisées, et notre corpus devra dès lors dépasser les quelques millions de mots. En matière de phraséologie en effet, Moon (1998) et Colson (2007, 2008) ont démontré la nécessité de corpus gigantesques.

Notons au passage que le statut précis des collocations spécialisées (par exemple, *créer un fichier, lancer un programme, une variable typée*) fait toujours débat : s'agit-il de réelles collocations telles que le langage général les utilise, ou plutôt de combinaisons

propres au domaine en question et liées à des restrictions sémantiques ? Selon L'Homme & Bertrand (2000), la plupart de ces structures présentent des co-occurrents liés à une classe sémantique (par exemple, *créer un fichier* mais aussi *créer une application, créer un répertoire*, etc.) et méritent dès lors plutôt l'étiquette de « combinaisons lexicales spécialisées ». Même si elles partagent avec les collocations un figement relatif, elles ne constituent pas, selon L'Homme & Bertrand, de véritables collocations, car elles n'associent pas des éléments pour former une combinaison unique, dotée d'une signification spécifique. Quoi qu'il en soit, ces combinaisons lexicales spécialisées se comportent largement comme les collocations du langage général : leur fréquence est généralement beaucoup plus faible que celle des termes simples, d'où la nécessité de corpus de grande taille.

Pour notre expérience, le meilleur candidat disponible parmi les corpus accessibles gratuitement est le projet WaCky (Baroni et al. 2009). En effet, les auteurs ont constitué à partir de la Toile et selon des critères linguistiques bien précis, des corpus gigantesques (environ 2 milliards de mots par langue) pour l'anglais, l'allemand, le français et l'italien[2].
Outre ce corpus linguistique extrait de la Toile, nous tenterons par ailleurs d'utiliser aussi directement la Toile comme corpus au sens large, afin d'en extraire le plus rapidement possible des informations pertinentes.
Contrairement à une recherche directe sur la Toile, les corpus du projet WaCky offrent le gros avantage d'un filtrage linguistique, qui élimine l'essentiel des données non-textuelles présentes sur les pages Web (images, tableaux, etc.).
Revers de la médaille : les corpus de 2 milliards de mots peuvent difficilement être manipulés par les outils classiques tels que les traitements de texte. Dans cette expérience, des programmes informatiques en langage Perl et Java ont été spécialement conçus pour extraire rapidement des informations ponctuelles dans l'ensemble du corpus, qui se présente sous la forme d'un fichier brut ou d'un ensemble de sous-corpus étiquetés, ce qui laisse toute latitude au chercheur.

2. Exemples et discussion

En matière de langue de spécialité, il faut bien comprendre que le traducteur est confronté presqu'à chaque page à des termes ou collocations spécialisées intraduisibles. Entendons par là qu'ils ne figurent dans aucun dictionnaire, ni même dans les banques de données terminologiques. Pour illustrer notre propos, nous nous intéresserons à quelques passages du *Java Cookbook* (Darwin 2004), un ouvrage de référence sur la programmation dans le langage informatique le plus populaire au monde, Java.

Extrait 1

> *Use an interconnected pair of piped streams and a Thread to read from the input half, and write it to the text area. You may also want to redirect System.out and System.err to the stream.*
> [Darwin 2004 : 391]

Qu'est-ce qu'un *piped stream*? Ce terme ne figure en effet pas dans la banque de données européenne IATE, et pas davantage dans le DiCoInfo. Quel est par ailleurs le caractère de cette cooccurrence : s'agit-il d'un terme ou d'une collocation spécialisée ? Quels sont les autres éléments figés ou techniques de cette phrase ?

Face à autant d'inconnues, il est difficile pour le traducteur technique de travailler de manière précise et efficace. Or, tant les corpus que la Toile peuvent lui fournir très rapidement des informations utiles, à condition qu'il maîtrise les bases de l'informatique, ou qu'il dispose des programmes adéquats.

Tout d'abord, il est utile de déterminer la fréquence relative de cette combinaison ainsi que son degré de fixité. Nous avons proposé (Colson 2010b) un algorithme permettant d'évaluer à partir de la Toile le degré de proximité moyenne, et donc la fixité relative, entre les divers éléments des n-grammes de niveau 2 à 10 (combinaisons de deux à dix mots). Couplé à un moteur de recherche, cet algorithme examine la proportion de proximité complète à partir des résultats renvoyés par le moteur de recherche[3]. Le score obtenu, le *WPR (Web Proximity Measure)*, varie de 0 à 1 selon le degré de fixité de la combinaison (le seuil de signifiance est de 0.15).

Dans le passage qui nous occupe, le traducteur se demandera par exemple si *interconnected pair, piped stream* et *you may also want to* constituent des structures figées et fréquentes. Le tableau 1 ci-dessous permet d'apporter une réponse à ces questions.

Combinaisons	WPR-score	Fréquence (Yahoo !)
interconnected pair	0,02	696
piped stream	0,41	1 860
you may also want to	0,32	11 100 000

Tableau 1 : Degré de fixité (score WPR) et fréquence de quelques combinaisons de l'extrait 1

En quelques secondes, le traducteur peut donc aboutir aux hypothèses de travail suivantes : *interconnected pair* est une combinaison peu courante (le chiffre de 696 occurrences sur l'ensemble de la Toile est très bas) et son degré de fixité est très faible également ; *piped stream* apparaît comme une structure très figée (et donc un terme composé, plutôt qu'une collocation spécialisée), de fréquence assez faible ; *you may also want to*, enfin, est une tournure nettement figée et très fréquente, ce qui confirme l'hypothèse d'une routine rédactionnelle à sens partiellement figuré.

Dans les trois cas, la méthode fournit une indication utile à la traduction. Pour *interconnected pair*, il ne s'agit pas d'un terme composé ni d'une collocation spécialisée, et *pair* a donc le sens littéral du langage courant ; le participe *interconnected* est, lui, technique, et une rapide vérification sur la banque de données de l'IATE nous livre son équivalent français (*interconnecté*). Etant donné que les scores nous ont révélé une combinaison non figée, la traduction pourra donc être : « une paire de… interconnectés », mais aussi : « deux… interconnectés ».

Pour *piped stream*, les scores ont indiqué un terme composé, et il s'agit donc de trouver un terme composé équivalent en français (non fourni par les l'IATE ou DiCoInfo). Ici encore, le traducteur peut recourir aux corpus ou à la Toile pour affiner sa traduction. Il s'agit tout d'abord, comme en extraction d'information, de sélectionner un « seed », un élément significatif qui permet de délimiter la recherche. En l'occurrence, l'ouvrage traitant du langage Java, il paraît évident de taper « java » suivi de « piped stream » dans le corpus français ou directement sur la Toile, en demandant des résultats en langue française. Les résultats nous montrent, à défaut d'une traduction, que le terme est parfois (rarement) utilisé tel quel en français ; en outre, en tapant séparément

pipe et *stream*, toujours précédés de *java*, le traducteur découvre dans les corpus que *stream* et *pipe* ont reçu également des traductions françaises, respectivement « flux » et « tube ». Selon le pays, le client et le souhait plus ou moins prononcé de purisme, le traducteur pourra donc laisser le terme anglais « piped stream », ou proposer par exemple « un flux-tube », traduction attestée également sur la Toile.

Toujours sur la piste des corpus, le traducteur découvre enfin que cette notion de *piped stream* intervient dans presque tous les contextes pour se référer à deux classes précises (au sens informatique) du langage Java : la *PipedInputStream* et la *PipedOutputStream*. Ces deux termes pourront donc figurer entre parenthèses après la traduction « flux-tube», ou en note.

L'une des grandes découvertes des recherches en phraséologie (Burger 1998, Mejri 2003, 2008) concerne par ailleurs l'omniprésence du figement au sens le plus large dans tous les registres et domaines de la langue, y compris en langue de spécialité. Dans l'extrait 1, *you may also want to* a pu être facilement identifié comme figement rédactionnel (par sa fréquence et son score statistique élevés). Il s'agira donc de le traduire par une routine courante, par exemple « vous pourriez également » ou « une autre solution consiste à ». Nombre de traducteurs techniques tombent dans le piège de traductions trop littérales pour les routines rédactionnelles.

Extrait 2

Implementing an Internet protocol from the ground up is not for the faint of heart. To get it right, you need to read and study the requisite Internet RFC pseudo-standards.
[Darwin 2004 : 547]

L'extrait 2 illustre également la présence importante de phraséologie en langue de spécialité. Le traducteur peut à nouveau hésiter quant au partage éventuel entre langue technique et tournures générales, et les scores statistiques lui sont fort utiles, comme indiqué au tableau 2, pour les combinaisons *from the ground up, for the faint of heart, get it right*, manifestement figées et très fréquentes.

Combinaisons	WPR-score	Fréquence (Yahoo !)
from the ground up	0,87	26 200 000
for the faint of heart	0,43	3 760 000
get it right	0,82	31 100 000

Tableau 2 : Degré de fixité (score WPR) et fréquence de quelques combinaisons de l'extrait 2

Une autre stratégie qui s'offre au traducteur technique consiste en la création d'un corpus de taille moyenne (quelques millions de mots) à partir de la Toile afin d'en extraire des informations pertinentes pour la traduction.

Parmi les moyens les plus récents et disponibles gratuitement, la suite Xaira Tools, mise au point par l'Université d'Oxford (http:www.oucs.ox.ac.uk/rts/xaira/) offre des possibilités étonnantes. Elle peut en effet être combinée à n'importe quel corpus pour créer dans un premier temps une version entièrement indexée du corpus. Ceci permet ensuite de réaliser une concordance quasi instantanée à partir de n'importe quel terme de recherche, contrairement à la plupart des programmes de concordance. Autrement dit, les résultats pertinents apparaîtront pratiquement aussi vite que sur la Toile.

Même s'il est toujours possible de constituer soi-même un corpus Toile en utilisant des programmes informatiques, l'opération prend du temps et se solde presque toujours par des résultats truffés de « bruit », c'est-à-dire des images, des données chiffrées inutilisables, du code html, etc. C'est à nouveau ici qu'intervient toute l'originalité du projet WaCky précité. La sélection et le filtrage des sources y sont particulièrement stricts, et la taille de chaque corpus linguistique est considérable (près de 2 milliards de mots dans chaque langue). Ceci permet dès lors de sélectionner des parties précises du corpus selon le domaine, ou par mot-clé. En combinaison avec la suite Xaira Tools, il est possible d'obtenir des données linguistiques très fines, ainsi que le montre le tableau 3 pour le terme anglais *pointer*.

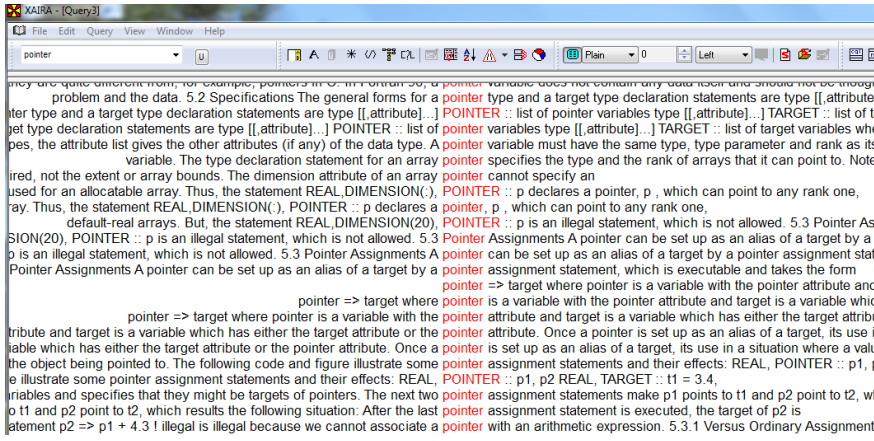

Tableau 3 : Association du WaCky corpus et de Xaira Tools pour créer une concordance Toile à partir du terme anglais « pointer »

Ce type d'information contextuelle dépasse largement ce que peuvent offrir les moteurs de recherche tels que Google ou Yahoo sur leur page web, et pourtant le laps de temps écoulé est tout aussi court. En outre, et contrairement aux résultats bruts de la Toile, le programme Xaira livre de nombreuses possibilités de tri des données (contexte de gauche et de droite par ordre alphabétique, etc.), ainsi que de précieux scores statistiques pour étudier les collocations liées au terme de recherche.

Ainsi, le z-score nous livre, parmi les collocations situées un mot à gauche, des combinaisons telles que : *array pointer, null pointer, dummy pointer, undefined pointer, buffer pointer, binary pointer, private pointer, file pointer*, et le même travail peut être effectué dans une autre langue pour rechercher des traductions possibles. Cette tâche était relativement fastidieuse à partir de corpus et de concordanciers traditionnels, mais l'indexation du corpus par Xaira Tools la rend particulièrement aisée.

Une autre méthode basée sur les corpus peut s'avérer très utile en matière de traduction informatique à partir de l'anglais (ce qui est évidemment presque toujours le cas). La position particulière de l'anglais, dans les sciences en général et au sein des nouvelles technologies en particulier, a mené à une situation où les termes anglais vivent en concurrence avec une traduction possible, même dans le jargon des spécialistes. A partir de cette constatation, il est loisible au traducteur d'imaginer des contextes pertinents

en langue cible, dans lesquels le terme figurera dans les deux langues. Très simplement, il suffira souvent, par exemple, de rechercher les contextes où le terme anglais sera suivi d'un autre entre parenthèses, ou inversement.

Prenons à nouveau l'exemple de *stream*. Si nous tapons ce terme sur Google en spécifiant les résultats en français, la lecture prendra un certain temps, car il n'est pas possible de demander que *stream* soit suivi d'une expression entre parenthèses (les moteurs de recherche ignorent en effet toute ponctuation). Par contre, à partir du gigantesque corpus français du projet WaCky, il est possible d'affiner la recherche par des expressions régulières. Cette technique informatique très simple est souvent accessible à partir d'un simple concordancier, ou elle peut être appliquée depuis un script ou programme, par exemple en Perl ou en Java. Ainsi, l'expression régulière suivante (en syntaxe Perl), fera apparaître en quelques secondes toutes les occurrences de *stream* suivi d'une formule entre parenthèses :

```
/stream \([^)]+\)/i
```

Parmi les résultats affichés à partir du corpus, nous trouverons bel et bien la traduction française :

stream (flux)

Dans d'autres cas, le traducteur pourra s'intéresser à la concurrence entre le terme anglais et son équivalent français, toujours sur la base de la même expression régulière. Ainsi, le corpus WaCky nous révèle que l'*operating system* de l'anglais est supplanté par sa traduction *système d'exploitation* ; par contre, le terme français est souvent concurrencé à son tour par l'abréviation anglaise OS. En effet, les exemples de type :

operating system (système d'exploitation)

sont beaucoup moins fréquents que l'inverse :

système d'exploitation (operating system) ;

par contre, nous trouvons de nombreux exemples du type :

système d'exploitation (OS : operating system)
système d'exploitation (ou OS pour " operating system ")
système d'exploitation (ou « OS » pour operating system).

3. En guise de conclusion

Longtemps rangée parmi les illusions, la révolution du TAL (Traitement Automatique des Langues) est bel et bien lancée. Linguistique de corpus et linguistique computationnelle s'associent de plus en plus pour offrir aux métiers de la langue des outils préfabriqués ou des techniques de plus en plus efficaces.
En langue de spécialité, en particulier dans le vaste domaine technologique, le traducteur mène une course incessante contre les avancées de la science et donc les néologismes. Dans cette contribution, nous avons tenté de montrer que la traduction basée sur les corpus et les techniques d'extraction d'information peuvent apporter une aide précieuse en cas de lacune des ouvrages lexicographiques et des banques de données.

Bibliographie

Burger H., 1998, *Phraseologie. Eine Einführung am Beispiel des Deutschen*, Schmidt Verlag, Berlin.

Baroni M., Bernardini S., Ferraresi A. & Zanchetta E., 2009, «The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora», *Journal of Language Resources and Evaluation*, 43, p. 209-226.

Colson J.-P., 2007, «The World Wide Web as a corpus for set phrases», *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*, H. Burger, D. Dobrovol'skij, P. Kühn et N. R. Norrick (eds.), p. 1071-1077, Walter de Gruyter, Berlin / New York.

Colson J.-P., 2008, «Cross-linguistic phraseological studies: An overview», *Phraseology. An interdisciplinary perspective*, S. Granger et F. Meunier (eds.), p. 191-206, John Benjamins, Amsterdam / Philadelphia.

Colson J.-P., 2010[a], «The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology», *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*, S. Ptashnyk, E. Hallsteindóttir et N. Bubenhofer (eds.), p. 23-35, Schneider Verlag, Hohengehren.

Colson J.-P., 2010[b], «Automatic extraction of collocations: a new Web-based method», *Proceedings of JADT 2010, Statistical Analysis of Textual Data*, Sapienza University , Rome, 9-11 June 2010, p. 397-408, LED Edizioni, Milan.

Darwin I.F., 2004, *Java Cookbook*, O'Reilly, Sebastopol (Californie).

Laviosa S., 2002, *Corpus-based Translation Studies: Theory, Findings, Applications*, Rodopi, Amsterdam / New York.

L'Homme M.-C., 2004, « Sélection des termes dans un dictionnaire d'informatique : comparaison de corpus et critères lexico-sémantiques », Actes d'*Euralex* 2004, 6-10 juillet 2004, p. 583-593, Lorient.

L'Homme M.-C., 2008, « Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés », *Traduire*, 217, p. 78-103.

L'Homme M.-C., 2010, «Designing Specialized Dictionaries with Natural Language. Examples of applications in the fields of computing and climate change», *Lexicography in the 21st Century: New challenges, new applications. Les Cahiers du Cental*, S. Granger et M. Paquot (dirs.), p. 203-216, Louvain-la-Neuve.

L'Homme M.-C. & Bertrand C., 2000, «Specialized Lexical Combinations: Should they be Described as Collocations or in Terms of Selectional Restrictions?», *Proceedings of the Ninth Euralex International Congress*, p. 497-506, Stuttgart University, Stuttgart.

Mejri S., 2003, «L'idiomaticité, problématique théorique», *L'espace euro-méditérranéen : une idiomaticité partagée*, S. Mejri (dir.), p. 231-243, CERES, Tunis.

Mejri S., 2008, «La traduction des jeux de mots», Jeux de mots et traduction, S. Mejri (dir.), *Equivalences*, 35, p. 85-101.

Moon R., 1998, *Fixed Expressions and Idioms in English*, Clarendon Press, Oxford.

Sinclair J., 1991, *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

Notes

[1] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load.
[2] voir : http://wacky.sslmit.unibo.it/doku.php?do=show&id=start.
[3] Pour plus de détails, voir Colson 2010[b].