Exploitation de corpus pour la recherche terminologique ponctuelle

Cet article décrit la façon dont un traducteur peut utiliser un corpus électronique comme ressource documentaire, surtout pour faire des recherches terminologiques ponctuelles dans les domaines qui évoluent rapidement (par exemple, l'informatique). Nous présentons les résultats d'une étude comparative dans le cadre de laquelle un groupe d'étudiants a fait des recherches ponctuelles en utilisant deux types différents de ressources: d'une part les ressources documentaires classiques et d'autre part un corpus électronique.

Termes-clés:
recherche terminologique ponctuelle; corpus; outils d'exploitation de corpus; stratégies d'exploitation; traduction.

1 Introduction

Les sous-domaines de la linguistique travaillent de plus en plus à partir de corpus. En terminologie, les chercheurs comme Bourigault (1994), Daille (1994), Jacquemin (1994) et Lauriston (1993), entre autres, ont examiné la possibilité de l'extraction automatique de termes. Meyer et McHaffie (1994) et Condamines (1995) ont utilisé des corpus à l'aide de la construction de bases de connaissances terminologiques (BCT). Un grand nombre d'applications automatisées sont décrits dans les actes de *Computerm '98* (Bourigault *et alii* 1998). En ce qui concerne les applications moins automatisées, Pearson (1996) a démontré l'utilité des corpus pour mener à bien la recherche terminologique thématique et pour développer les glossaires terminologiques. Cet article vise à dévoiler l'utilité des corpus pour mener à bien la recherche terminologique ponctuelle.

Selon Rondeau (1984: 64), «On distingue la recherche terminologique *thématique* et la recherche terminologique *ponctuelle*, la seconde étant soumise aux contraintes de besoins immédiats à satisfaire». D'après Dubuc (1985: 45), «La recherche terminologique qui porte sur des problèmes isolés porte le nom de recherche ponctuelle». Rondeau (1984: 65) précise que «La terminologie ponctuelle a pour but de fournir des réponses de qualité, dans les délais les plus brefs, à des questions spécifiques localisées dans le temps et dans l'espace».

Malheureusement, certains se plaisent à dénigrer la recherche ponctuelle, considérant que ce type de travail n'apporte pas vraiment de solutions d'ensemble aux besoins terminologiques des divers milieux de travail. Il est vrai que la recherche ponctuelle à elle seule ne peut apporter de solutions globales et à long terme; néanmoins, cette activité compte pour une part importante du travail terminologique, et pour les traducteurs, elle est une réalité quotidienne.

On sait bien que les traducteurs doivent souvent travailler dans des délais très serrés et que donc ils ne peuvent pas souvent consacrer beaucoup de temps à la recherche ponctuelle. Malheureusement, les dictionnaires et les glossaires, même les plus récents, ne fournissent pas toujours les renseignements nécessaires, surtout dans les domaines qui ont l'air d'être dans un état de perpétuel changement (par exemple, l'informatique). Dans ce type de cas, le traducteur doit chercher d'autres sources, telles que des documents parallèles (Williams 1996). Un document parallèle est un document écrit dans la langue d'arrivée qui est comparable au texte de départ du point de vue du fond et de la fonction (type de texte). Cependant, faire une recherche dans des documents parallèles présente certains inconvénients: la consultation des documents prend beaucoup de temps et est difficile à faire de façon systématique. Souvent, le traducteur découvre des données utiles au hasard plutôt qu'à dessein (Miller 1993: 8).

À notre avis, les corpus électroniques constituent des ressources excellentes pour les traducteurs. Il est clair que les corpus bilingues fournissent une mine d'informations, mais ce type de corpus est encore assez difficile à trouver ou à construire pour le traducteur surchargé de travail. Néanmoins, un corpus monolingue peut fournir des informations très utiles. Dans cet article, nous décrivons une expérience que nous avons effectuée à la Dublin City University. Le but est de comparer les résultats des recherches ponctuelles fondées sur des ressources conventionnelles avec celles qui étaient fondées sur un corpus.

2 Les participants

Cinq étudiants de quatrième année ont participé à l'expérience. Ils suivaient un cours de traduction spécialisée (français-anglais), mais aucun d'entre eux n'était spécialiste en informatique. Tous les étudiants connaissaient l'outil d'exploitation de corpus *WordSmith Tools*(1), mais aucun d'entre eux ne l'avait encore utilisé pour mener une traduction ou une recherche ponctuelle.

3 Le texte et les termes

Le texte de départ se compose d'un extrait de 100 mots, tiré d'un article qui s'intitule «Microprocesseur et carte mère» et paru dans le journal *Science et Vie micro* (décembre 1997). Les cinq termes suivants ont été choisis pour être l'objet de recherches ponctuelles: *carte mère*, *carte d'extension*, *bus de données*, *fréquence d'horloge* et *connecteurs pour barrettes de mémoire vive*.

(1) http: //www.liv.ac.uk/~ms2928/wordsmith/index.htm.

4 Les ressources

L'expérience a été réalisée en deux parties, chacune d'entre elles exigeant des ressources différentes.

4.1 Les ressources conventionnelles

Lorsqu'ils ont fait la traduction et la recherche ponctuelle à l'aide des ressources conventionnelles, les étudiants pouvaient accéder à toutes les ressources de la bibliothèque, y compris aux dictionnaires (généraux et spécialisés, bilingues et monolingues), aux encyclopédies, aux livres et aux revues spécialisées ainsi qu'à l'Internet.

4.2 Le corpus

Lorsqu'ils ont fait la traduction et la recherche ponctuelle à l'aide d'un corpus, les étudiants avaient le droit de consulter un corpus monolingue spécialisé qui était construit spécialement pour l'expérience. Le corpus comporte un million de mots de textes authentiques écrits en anglais par des experts du domaine. Le corpus était tiré d'un cédérom intitulé *Computer Select*. Ce disque contient des articles tirés de centaines de revues qui traitent du domaine de l'informatique. Les textes choisis étaient similaires au texte de départ du point de vue du fond, de la fonction (type de texte) et de la date de parution. Il n'a fallu que trente minutes pour construire ce corpus. Pour une description plus détaillée de la conception et de la construction d'un corpus terminologique, voir Bowker (1996) ou Meyer et Mackintosh (1996).

Pour exploiter le corpus, les étudiants ont utilisé *WordSmith Tools*, qui est un type de logiciel qui permet à l'utilisateur de manipuler les mots dans le corpus de plusieurs façons. Par exemple, on peut accéder à toutes les occurrences des termes et à leur contexte d'apparition (les concordances); on peut regarder les collocations courantes d'un terme; on peut déterminer la fréquence d'un terme (Rézeau 1997: 167).

5 L'expérience

L'expérience s'est effectuée en deux stades. Au premier stade, les étudiants ont traduit le texte à l'aide des ressources conventionnelles dans un délai d'une heure et quart. Le deuxième stade s'est déroulé deux semaines plus tard. Les étudiants ont alors traduit à nouveau le même texte, mais cette fois-ci, à l'aide du corpus et de *WordSmith Tools*. Un délai de 45 minutes était imposé lors du deuxième stade (une heure et quart, moins la demi-heure nécessaire pour construire le corpus). À la fin de chaque stade, nous avons interrogé les étudiants pour découvrir leurs avis et leurs attitudes à l'égard des différentes ressources. De plus, en ce qui concerne le corpus, nous nous sommes attachés à découvrir leurs stratégies d'exploitation.

6 L'analyse des données

Les deux séries de traductions sont analysées à divers égards, mais dans cet article nous examinons les données uniquement dans l'optique de la recherche terminologique ponctuelle. Les tableaux 1 à 5 illustrent les résultats des recherches ponctuelles effectuées par les étudiants pour chacun des cinq termes en question. Si un terme est marqué d'un astérisque, il s'agit d'un terme incorrect.

Tableau 1

carte mère	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	*motherboard*	*motherboard*
étudiant 2	**mother card*	*motherboard*
étudiant 3	*(*)mother board*	*motherboard*
étudiant 4	*(*)mother board*	*motherboard*
étudiant 5	*motherboard*	*motherboard*

Tableau 2

carte d'extension	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	**extension card*	*expansion card*
étudiant 2	**extension card*	*expansion card*
étudiant 3	**extension card*	*expansion card*
étudiant 4	**extension card*	*expansion card*
étudiant 5	*expansion card*	*expansion card*

Tableau 3

fréquence d'horloge	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	**clock frequency*	*clock speed*
étudiant 2	**clock frequency*	*clock speed*
étudiant 3	**clock frequency*	*clock speed*
étudiant 4	**clock frequency*	*clock speed*
étudiant 5	**clock frequency*	*clock speed*

Tableau 4

bus de données	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	*data bus*	*data bus*
étudiant 2	*data bus*	*data bus*
étudiant 3	**information bus*	*data bus*
étudiant 4	*data bus*	*data bus*
étudiant 5	*data bus*	*data bus*

Tableau 5

connecteurs pour barrettes de mémoire vive	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	**connectors for *strips of RAM*	**connectors for RAM *strips*
étudiant 2	**connectors for * strips of RAM*	**connectors for RAM*
étudiant 3	**connectors for RAM *strips*	*RAM *connectors*
étudiant 4	**connectors for RAM *strips*	**connectors for *strips of *live memory*
étudiant 5	**connectors for *strips of RAM*	*SIMM slots for RAM*

Tableau 6

Nombre d'erreurs	Traduction 1 (ressources conventionnelles)	Traduction 2 (corpus)
étudiant 1	4	2
étudiant 2	5	1
étudiant 3	5	1
étudiant 4	5	3
étudiant 5	3	0
Total	22	7

Premièrement, nous avons constaté une diminution des erreurs, aussi bien chez tous les étudiants que chez chacun d'entre eux, quand le corpus est employé comme ressource terminologique. Les statistiques figurent dans le tableau 6.

Une analyse plus profonde indique que les erreurs particulières que les étudiants ont commises quand ils travaillaient à l'aide des ressources conventionnelles ont été corrigées pour la plupart quand les mêmes étudiants travaillaient à l'aide du corpus. Afin de déterminer pourquoi le nombre d'erreurs a diminué, nous avons questionné les étudiants sur leurs attitudes, leurs avis, et leurs stratégies d'exploitation.

6.1 Les inconvénients des dictionnaires

Nous avons déjà établi que les dictionnaires sont souvent vieillis, que les termes recherchés ne s'y trouvent pas toujours. Dans ce cas, les étudiants avaient parfois recours à la recherche d'éléments individuels d'une unité complexe et les combinaient selon les règles grammaticales. Résultat : un terme qui suit la syntaxe de la langue, mais qui n'a aucune valeur sémantique. C'est le cas, par exemple, des traductions suivantes. Le terme 'carte mère' ne se trouvait pas dans les dictionnaires disponibles, alors l'étudiant a opté pour les traductions des éléments individuels 'carte' = *card*, 'mère' = *mother*, et les a combiné de façon grammaticale pour produire le non-sens **mother card*. D'autres exemples similaires : 'carte d'extension' *(*extension card)*; 'bus de données'; *(*information bus)*; 'fréquence d'horloge'; *(*clock frequency)*.

6.2 Les inconvénients des documents parallèles

La plupart des étudiants ont reconnu qu'ils ne consultaient pas souvent les documents parallèles, bien qu'ils connaissent la valeur des informations qu'ils renferment. Ils ont tous avancé la raison du manque de temps. L'un des étudiants parlant de l'Internet, a rapporté que tant les périodes d'attente que le nombre de documents récupérés étaient excessifs. Pour un autre étudiant, la difficulté de repérer le terme ou le passage désiré représente le problème principal lié aux documents parallèles. Il a déclaré que les termes en question n'étaient pas toujours mis en évidence dans le texte, et qu'ils se trouvaient rarement dans les index ou les tables des matières. Il a ajouté que la lecture de longs passages hors du sujet l'avait quelque peu lassé. Un troisième étudiant a noté que, si on travaille dans des délais très courts, l'utilisation des documents parallèles n'est pas toujours réalisable. Ce sont des plaintes valables car, d'après Rondeau (1984 : 65), la contrainte de temps est toujours présente en terminologie ponctuelle et la durée temporelle entre la formulation d'une question et l'obtention de la réponse doit toujours être réduite au minimum.

Les documents parallèles présentent un autre inconvénient. Même si on peut trouver le terme désiré dans le document, il est probable qu'on n'en trouve qu'une ou deux occurrences. Dans ce cas, il peut être difficile d'établir un « crochet terminologique ». Selon Dubuc (1985 : 72), en terminologie comparée, « on entend par crochet terminologique les descripteurs communs aux contextes accompagnant les vedettes ». De plus, si on n'a que peu d'occurrences, il est difficile de savoir si le candidat équivalent est d'usage commun ou d'usage particulier.

6.3 Les avantages du corpus

Un corpus offre de grands avantages par rapport aux ressources classiques : il est plus d'actualité et plus complet que les dictionnaires, plus facile à consulter que les documents parallèles, et, d'après les étudiants, plus intéressant à exploiter. Par exemple, l'étudiant qui négligeait la recherche de termes dans les documents parallèles, s'intéressait beaucoup à la recherche dans le corpus. En outre, le corpus présente les termes, y compris les unités complexes, dans leurs contextes d'apparition. La possibilité de voir plusieurs exemples d'usage en même temps facilite l'établissement d'un crochet terminologique. En d'autres termes, le traducteur peut mieux vérifier l'exactitude et l'adéquation du terme comme équivalent en regardant un grand nombre de contextes. De plus, le traducteur peut déterminer si le candidat équivalent est d'usage commun ou s'il constitue une préférence particulière à peu d'auteurs.

6.4 Les stratégies d'exploitation du corpus

Pour bien exploiter un corpus, il faut avoir des stratégies. Dans le cadre de l'expérience susmentionnée, les étudiants ont employé les outils suivants, soit de façon autonome, soit en association avec d'autres : l'outil de collocation, le concordancier, et l'outil de fréquence.

Du fait que le corpus est monolingue (en langue d'arrivée), la plupart des étudiants ont commencé la recherche en cherchant un équivalent (ou un équivalent partiel) dans un dictionnaire bilingue général. Comme point de départ, il s'agit d'une stratégie bien fondée. Selon Dubuc (1985 : 46), si le traducteur possède le terme en langue de départ, il peut consulter les dictionnaires bilingues car ces ouvrages fournissent souvent des éléments de solution. Dubuc continue en précisant : « En [terminologie] ponctuelle, chaque fois qu'il y a consultation de documentation, il faut toujours procéder du général au spécialisé. Cette démarche se justifie par les contraintes de temps qui régissent cette activité. Les dictionnaires généraux, faits pour le grand public, sont habituellement plus faciles à consulter, l'information s'y retrouve plus rapidement et souvent ils présentent une vue d'ensemble du terme en rubrique où l'on peut trouver, à défaut de solutions, des jalons de recherches supplémentaires » (Dubuc 1985 : 47).

Après avoir trouvé un terme candidat (ou un terme candidat partiel), il suffit de saisir ce ou ces mot(s) dans la fenêtre « mot(s) à rechercher » dans l'outil de collocation. Le logiciel fournit une liste de tous le

mots qui figurent dans le corpus à proximité du mot à rechercher. Par exemple, prenons le terme 'carte mère'. On peut commencer en recherchant les collocations du terme *card*, que le dictionnaire bilingue général présente comme équivalent commun de 'carte'. Parmi les collocations proposées on trouve : *expansion card, fax/modem card, graphics card, network card, sound card, tuner card, video card, WinTV card*, etc. Si on considère que l'un de ces termes est un candidat, on peut chercher des concordances ; si non, on peut recommencer en recherchant les collocations du terme *mother*. Cette fois, le logiciel propose un candidat prometteur : *motherboard* (2).

À l'étape suivante, le traducteur regarde les concordances pour le terme candidat. Les contextes lui permettent d'établir le crochet terminologique. De plus, il peut découvrir l'usage correct (à l'égard de la syntaxe, de la phraséologie) du terme.

S'il existe plusieurs candidats, le traducteur peut comparer les concordances de chacune des possibilités. Si les contextes indiquent qu'il y a des synonymes, le traducteur peut utiliser l'outil de fréquence pour l'aider à choisir le terme correct. Par exemple, les concordances pour les termes *clock rate* et *clock speed* indiquent que les deux sont synonymes. L'outil de fréquence dévoile que *clock speed* paraît 157 fois dans le corpus, lorsque *clock rate* ne paraît que 49 fois.

La possibilité d'exclure les hypothèses incorrectes constitue un avantage supplémentaire. Par exemple, quand ils employaient les ressources conventionnelles, tous les étudiants ont proposé *strips* comme traduction de 'barrettes' dans le syntagme 'barrettes de mémoire vive'. Cependant, quand ils ont consulté le corpus, le mot *strip(s)* n'y figurait pas – pas une seule fois dans un corpus d'un million de mots ! Face à cette « preuve négative », les étudiants ont cherché d'autres façons d'exprimer la notion. Bien qu'ils n'aient pas tous trouvé la meilleure solution, ils ont tous amélioré leurs traductions en éliminant la mention de *strips*.

(2) Notons que le traducteur peut, avec l'expérience, court-circuiter les cheminements classiques pour arriver presque d'instinct à la solution cherchée sans compromettre la validité de sa recherche, mais le débutant doit être prudent.

6.5 Les inconvénients du corpus

Le corpus monolingue n'est manifestement pas une ressource terminologique parfaite. L'un des inconvénients principaux c'est qu'il faut avoir un point de départ adéquat. Sinon, autant chercher une aiguille dans une botte de foin ! Par exemple, l'un des étudiants ne pouvait pas trouver la traduction de 'mémoire vive' dans le corpus. Il a examiné les collocations et les concordances pour le terme *memory* ainsi que pour le terme *live*, mais le corpus n'a rien donné. C'est parce que la bonne traduction prend généralement le forme du sigle *RAM*.

7 Conclusion

L'exploitation d'un corpus n'est pas une panacée. Elle ne remplace jamais le travail consciencieux du traducteur ou du terminologue. Néanmoins, à notre avis, un corpus, même monolingue, peut être une ressource utile pouvant compléter les ressources terminologiques conventionnelles. Les possibilités des corpus comme aide à la recherche terminologique (soit ponctuelle, soit thématique, soit théorique), ne commencent qu'à être réalisées. C'est une époque bien excitante pour la terminologie et nous espérons que le travail décrit ici pourra servir de point de départ pour des recherches plus approfondies dans ce domaine.

Lynne Bowker,
School of Applied Language and Intercultural Studies,
Dublin City University,
Irlande.

Remerciements

Nous remercions les étudiants qui ont participé à l'expérience susmentionnée. Nous remercions également Danièle Tort, DCU.LS, pour la correction des épreuves.

Bibliographie

Blanc (C.), 1997 : « Microprocesseur et carte mère », dans Science et Vie micro, n° 155, décembre 1997, p. 260-265.

Bourigault (D.), 1994 : *LEXTER, Un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*, Thèse de doctorat, EHESS, Paris, France.

Bourigault (D.), Jacquemin (C.), L'Homme (M.-C.), dir., 1998 : *Computerm '98, Proceedings of the First Workshop on Computational Terminology, COLING-ACL '98*, 15 août 1998, Université de Montréal, Canada.

Bowker (L.), 1996 : « Towards a corpus-based approach to terminography », dans *Terminology*, n° 3(1), p. 27-52.

Condamines (A.), 1995 : « Analyse de textes spécialisés pour le recueil de données terminologiques », dans Terminologies nouvelles, n° 14, p. 35-42.

Daille (B.), 1994 : *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistique*, Thèse de doctorat, Université de Paris VII, France.

Dubuc (R.), 1985 : *Manuel pratique de terminologie*, 2e édition, Québec, Linguatech.

Jacquemin (C.), 1994 : « Quelques mécanismes spécifiques d'une grammaire d'unification adaptée à l'extraction terminologiques », dans *Actes du 9e congrès « Reconnaissance des formes et intelligence*

artificielle» (RFIA'94), Paris, AFCET, p. 385-396.

Lauriston (A.), 1993 : *Le repérage automatique des syntagmes terminologiques*, Thèse de maîtrise, Université du Québec à Montréal, Canada.

Meyer (I.) et Macintosh (K.), 1996 : « The Corpus from a Terminographer's Viewpoint », dans *International Journal of Corpus Linguistics*, n° 1(2), p. 257-285.

Meyer (I.) et McHaffie (B.), 1994 : « De la focalisation à l'amplification : nouvelles perspectives de représentation des données terminologiques », dans Clas (A.) et Bouillon (P.), dir., TA-TAO : *Recherches de pointe et applications immédiates, Actes des Troisièmes Journées scientifiques du réseau thématique «Lexicologie, Terminologie, Traduction»*, Montréal, 30 septembre-2 octobre 1993, Beyrouth, FMA, p. 425-440.

Miller (D.R.), 1993 : *Towards Knowledge-Base Systems for Translators*, Thèse de maîtrise, Université d'Ottawa, Canada.

Pearson (J.), 1996 : « Teaching terminology using electronic resources », dans Botley (S.), Glass (J.), McEnery (T.) et Wilson (A.), dir., *Proceedings of Teaching and Language Corpora 1996*, University of Lancaster, UCREL, p. 203-216.

Rézeau (J.), 1997 : « Concordances, cédérom et internet au service de l'enseignement du français aux adultes », dans *The Dong-eui International Journal*, n° 3, p. 166-192.

Rondeau (G.), 1984 : *Introduction à la terminologie* (2e édition), Québec, Gaëtan Morin.

Williams (I.A.), 1996 : « A Translator's Reference Needs : Dictionaries or Parallel Texts? », dans *Target*, n° 8(2), p. 275-299.