DES OUTILS

À BASE DE CORPUS

À L'INTENTION DES TRADUCTEURS

Elliott Macklovitch

Centre d'innovation en technologies d'information

octobre 1993

Résumé

Cet article décrit un nouvel ensemble d'outils destinés à être intégrés au poste de travail du traducteur (PTT) conçu au Centre d'innovation en technologies de l'information (anciennement, le Centre canadien de recherche sur l'informatisation du travail). Jusqu'à présent, toutes les composantes du PTT ont été des produits commerciaux, dont bon nombre ne s'adressent pas spécifiquement aux traducteurs. Les nouveaux outils dont il est question dans cet article offrent par contre des fonctions conçues expressément pour les traducteurs et qu'on ne trouve pas, à notre connaissance, dans les logiciels commerciaux. Ces outils exploitent un type particulier de corpus, appelé bi-texte, qui est composé de textes sources et de leurs traductions, et où les unités de traduction de chaque langue sont explicitement appariées, ou reliées. Une fois de tels corpus alignés, il a été possible d'élaborer un nouvel outil d'aide à la traduction, le concordancier bilingue interactif. Nous montrerons comment cet outil permet aux traducteurs d'exploiter leur production antérieure et d'y trouver des solutions à une gamme de problèmes qui ne sont généralement pas traités dans les dictionnaires ou les banques de terminologie; nous en présenterons des exemples concrets, tirés d'un corpus de 26 millions de mots constitué des transcriptions et traductions des débats parlementaires. Nous décrirons enfin un nouveau type d'outil, basé lui aussi sur l'existence d'un bi-texte, qui permet de vérifier un texte source et sa traduction pour s'assurer que cette dernière possède certaines caractéristiques essentielles.

Cat. no. : Co 28-1/109 - 1993F

ISBN : 0-662-98725-X

Cet article a été présenté sous forme de communication à la 33ième conférence annuelle de la American Translators' Association, tenue à San Diego du 4 au 8 novembre 1992.

Les opinions exprimées dans cet article n'engagent que l'auteur.

This report is also available in English.

1. QUAND ON NE SAIT PLUS À QUEL SAINT SE VOUER

Vous êtes un traducteur qui traduisez de l'anglais au français et vous avez le privilège d'exercer votre profession à l'aide d'un poste de travail perfectionné comme le PTT[1], conçu au Centre d'innovation en technologies de l'information. Vous venez de recevoir un texte à traduire. Vous mettez votre ordinateur sous tension et commencez à travailler. Une première lecture du texte vous a permis de constater qu'il n'était pas très technique. Vous prévoyez donc pouvoir le traduire très rapidement.

Vous ouvrez une fenêtre de traitement de texte et commencez à traduire. Après quelques phrases, vous tombez sur le dialogue suivant : "I've already spoken to Lucien. The guy is completely out to lunch!" Qu'est-ce que "completely out to lunch" peut bien vouloir dire? D'après le contexte et la présence de l'adverbe de degré "completely", vous vous doutez bien qu'il s'agit d'une expression figurée et que la phrase ne signifie pas littéralement que Lucien est sorti déjeuner. Mais qu'est-ce que cela signifie au juste?

Votre premier réflexe est de consulter les ressources lexicales et terminologiques qui vous sont accessibles sur le PTT. Vous appuyez donc sur la touche qui active le gestionnaire de glossaires et tapez "lunch" dans le champ de recherche. Le système ne trouve rien dans le glossaire de travail, mais affiche un enregistrement du dictionnaire bilingue de référence qui vous apprend que "lunch" se traduit en français par le nom masculin "déjeuner". Vous zoomez sur l'index du dictionnaire et passez à l'enregistrement suivant, qui donne des équivalents des expressions "to have lunch" et "lunch hour", mais rien du tout pour "to be out to lunch". Vous voilà bien avancé. Vous passez ensuite à la fenêtre d'exploitation de la banque de données terminologiques *Termium*, où vous lancez quelques interrogations, qui ne donnent, hélas, aucun résultat pertinent. Bon, s'il n'y a rien dans les ressources électroniques du PTT, vous pouvez toujours vous rabattre sur la bibliothèque. Vous prenez le *Robert-Collins*, qui liste généralement de nombreuses expressions figurées d'usage courant, mais vous n'y trouvez rien qui réponde à votre question. En dernier recours, vous décidez de consulter un dictionnaire unilingue anglais[2], le *Webster's Ninth Collegiate*, et eurêka! Le *Webster's* vous apprend que "to be out to lunch" est une expression argotique qui signifie "to be out of touch with reality".

Parfait. Maintenant que vous savez enfin ce que signifie l'expression, vous avez quelques idées sur la façon de la traduire en français, entre autres : "être décroché de la réalité" et "être perdu dans

1. "PTT" est l'acronyme de "poste de travail du traducteur". On trouvera, à la référence bibliographique 10, une description détaillée du PTT et de ses composantes, ainsi que des principes qui sous-tendent ce projet du CITI.
2. Un dictionnaire unilingue anglais a été ajouté depuis au PTT-3.

la brume”. Mais ce sont là des traductions de “to be out of touch with reality”, et vous n’êtes pas tout à fait certain si cette expression peut s’utiliser dans les mêmes contextes que “to be out to lunch”. Quand vous êtes aux prises avec des problèmes de ce genre, vous aimez normalement consulter vos collègues, surtout ceux qui sont les plus versés en anglais familier. Malheureusement, c’est samedi et vous êtes seul au bureau.

Vous ne savez vraiment plus à quel saint vous vouer. Saint Antoine, pour qu’il guide vos recherches? Ou plutôt saint Jérôme, patron des traducteurs? En fait, l’idée de demander l’intercession de votre illustre prédécesseur n’est pas si bizarre, car elle fait allusion à un important élément de la solution que nous allons proposer ci-après à votre dilemme.

Supposons que le PTT contienne un système d’interrogation de base de données qui vous permettrait d’obtenir rapidement tous les équivalents traductionnels d’un mot ou d’un syntagme figurant dans tous les textes déjà traduits par tous les membres, anciens et actuels, de votre service de traduction, même si cela représentait des dizaines de millions de mots de production antérieure. Alors, au lieu de quitter votre bureau et d’aller déranger vos collègues (dans l’hypothèse où ils seraient présents), vous n’auriez qu’à ouvrir la fenêtre du PTT qui donne accès à cette base de données et, par le truchement d’une interface simple, vous pourriez demander au système de récupérer toutes les traductions de l’expression “to be out to lunch”. Pour illustrer comment cela fonctionnerait, nous avons fait la demande à votre place, en utilisant notre base de données de 26 millions de mots du Journal des débats de la Chambre des communes, couvrant les années 1986 et 1987. Les résultats de cette interrogation sont présentés à l’Annexe A.

Si vous examinez le listage de l’Annexe A, vous pourrez constater qu’il revêt la forme d’une série de concordances, c’est-à-dire que chaque entrée présentée du côté gauche de la page est une occurrence distincte de l’expression “out to lunch” trouvée dans un texte donné, accompagnée de son contexte immédiat[3], et que chaque entrée présentée en regard, du côté droit, est la traduction de ce segment de texte source. Nous appelons l’outil qui produit ce genre de sorties traductionnellement reliées un *concordancier bilingue*.

Prenez quelques minutes pour lire les diverses traductions présentées à l’Annexe A. Il est difficile de ne pas être impressionné par la variété des solutions que les traducteurs parlementaires ont trouvées pour l’expression “out to lunch”. Certaines sont d’un niveau de langue passablement plus relevé que l’expression anglaise, par exemple : “être tout à fait dans l’erreur”, tandis que

3. Chaque citation de la concordance est précédée de l’emplacement du répertoire, de la date de création et de la langue du fichier texte; les nombres entre crochets renvoient à la ligne du fichier où l’expression a été repérée. Depuis la rédaction de cet article, nous avons programmé une nouvelle interface sous *X-Windows*, ce qui a considérablement amélioré la présentation des sorties.

d'autres, comme "se fourvoyer", trahissent un léger glissement de sens et correspondraient davantage à "to be on the wrong track"[4]. On peut aussi observer certaines différences dialectales: il est en effet peu probable qu'un Canadien francophone emploie l'expression "être dans la choucroute", à laquelle il préférera plutôt "être dans les patates". Ce qui compte, cependant, c'est que notre traducteur dispose maintenant d'une gamme de solutions intéressantes, voire dans certains cas plutôt inspirées, à partir de laquelle il peut faire un choix. Voilà qui est plutôt agréable, n'est-ce pas?, quand on est en panne d'inspiration.

2. LE RECYCLAGE DES TRADUCTIONS

Comment fonctionne un concordancier bilingue? Si, comme nous l'espérons, les traducteurs conviennent de l'utilité de cet outil, ils voudront sans doute savoir ce qui est nécessaire pour pouvoir formuler des interrogations et obtenir des résultats comme ceux qui sont présentés à l'Annexe A.

Le premier point à souligner est qu'il n'existe actuellement, du moins à notre connaissance, aucun progiciel commercial capable de produire des concordances bilingues comme celles qui sont présentées en annexe. Il existe certes des concordanciers unilingues[5], mais nous ne connaissons aucun logiciel commercial capable de rechercher, en mode interactif, une expression donnée et d'en récupérer toutes les occurrences dans leur contexte, avec la traduction de chaque unité telle qu'elle apparaît dans un autre fichier texte. Les seuls concordanciers bilingues fonctionnels que nous connaissions se trouvent encore dans des laboratoires de recherche comme celui du CITI.

Pourquoi? Cela peut s'expliquer en partie par le fait qu'un logiciel de ce genre n'intéresserait que les traducteurs, qui constituent un marché relativement restreint comparativement à celui des rédacteurs unilingues. Mais il existe aussi une raison d'ordre technique, qui est encore plus importante : un concordancier bilingue présuppose l'existence de liens explicites entre les segments d'un texte source et ceux d'un texte cible qui en sont la traduction. Ces unités reliées sont généralement des phrases, mais il est également possible, en principe, d'établir des liens entre des unités de texte plus petites ou plus grandes[6]; et, bien sûr, une phrase du texte source peut être reliée de cette façon à deux phrases ou plus du texte cible, et vice-versa. On dit de deux textes ainsi reliés qu'ils sont *alignés* ou *appariés*, et l'on appelle parfois *bi-texte*[7] un ensemble de textes alignés. Pour

4. Il se pourrait aussi que l'expression anglaise "to be out to lunch" soit assez vague pour englober les deux sens.
5. Les distributeurs de *Mercury/Termex*, par exemple, ont déjà incorporé à leur logiciel de gestion de glossaire un concordancier unilingue appelé *TextSearch*.
6. Isabelle [6] appelle ce paramètre la *résolution* d'un bi-texte.

un petit corpus, comprenant disons quelques dizaines de pages, il peut être concevable (quoique cela serait certainement onéreux) d'établir manuellement de tels liens, peut-être à l'aide d'une procédure macro semi-automatique. Il devrait toutefois être évident que l'intérêt présenté par un concordancier bilingue augmente de façon directement proportionnelle à la taille du corpus qu'il permettra aux traducteurs d'interroger. Pour qu'un outil de ce genre soit vraiment utile, il faudrait qu'il puisse aligner de très grands corpus de textes traduits, en fait des millions de mots, et pour des volumes de cette ampleur, l'alignement manuel est évidemment hors de question. Il nous faut donc des procédures d'alignement entièrement automatiques, des programmes qui pourront produire, à partir de textes sources et de leurs traductions, des bi-textes correctement alignés. Ce n'est que tout récemment que des programmes efficaces capables d'exécuter cette tâche ont été mis au point par des linguistes-informaticiens. Les chercheurs du CITI s'appliquent, pour leur part, à perfectionner une approche algorithmique qui a d'abord été élaborée par Ken Church et William Gale aux laboratoires AT&T Bell[8]. Ces algorithmes sont basés sur des calculs statistiques de la longueur des phrases et sont donc indépendants d'une langue en particulier. Les premiers essais que nous avons faits indiquent qu'ils fonctionnement non seulement avec l'anglais et le français, mais également avec d'autres langues européennes.

On pourrait dire, en un sens, que tout texte à traduire est constitué de problèmes de traduction plus ou moins complexes, qui doivent tous être résolus. La très grande partie de ces problèmes sera résolue de façon presque inconsciente par le traducteur compétent, qui pourra également trouver la solution de certains problèmes plus ardus dans des dictionnaires, des banques de terminologie ou d'autres ouvrages de référence standard. Mais il arrive fréquemment que le traducteur ne puisse trouver de solution toute faite et qu'il doive faire appel à sa propre ingéniosité. Or, peu de traducteurs, même parmi les plus consciencieux, prennent la peine de rédiger des fiches terminologiques. Imaginer le nombre de fiches qu'il faudrait faire pour consigner ou documenter toutes les solutions créatives d'une manière qui permettrait à d'autres traducteurs de les consulter facilement. En fait, le seul enregistrement *intégral* des choix linguistiques effectués par le traducteur réside dans le texte cible lui-même, lequel est souvent mis de côté et oublié après que la traduction a été livrée au client. Après tout, à quoi bon encombrer ses classeurs ou son disque dur, s'il n'existe aucune façon de localiser rapidement l'information dont on a besoin? Même dans les services de traduction où l'on archive systématiquement la version électronique de toutes les traductions réalisées, le mieux qu'on puisse faire actuellement, c'est d'aider le traducteur à localiser les textes qui pourraient être utiles pour une nouvelle traduction, peut-être en recourant à

---

7. Le terme *bi-texte* a été créé par Brian Harris (voir les références bibliographique 4 et 5).
8. Voir la référence bibliographique [3]. Pour plus de détails sur les efforts que nous avons faits pour raffiner leur algorithme basé sur la longueur des phrases, voir la référence [12].

un logiciel de recherche plein texte comme celui que nous avons incorporé à notre PTT-2. Mais la mesure de pertinence sur laquelle sont basés ces outils n'est souvent pas plus raffinée qu'un décompte des mots ou des expressions répétés dans le texte source. Le logiciel de recherche indique au traducteur qu'il a repéré un document source dans lequel sont répétés certains des termes de son interrogation, mais il revient au traducteur de déterminer si ce texte est pertinent par rapport au texte à traduire, puis d'extraire ce qu'il peut de ce document et de sa traduction. En supposant, ne l'oublions pas, qu'il sache où se trouve la version électronique de la traduction, car le logiciel d'extraction n'établit pas lui-même de lien explicite avec le texte cible correspondant. Ce lien, c'est *le traducteur* qui doit le faire quand il enregistre les deux versions du document, peut-être en donnant aux fichiers des noms identiques se terminant par une extension différente pour chaque langue.

Le concordancier bilingue que nous proposons fonctionne de façon tout à fait différente. En alignant automatiquement les textes sources et leur traduction, nous pouvons transformer les archives d'un service de traduction en une **mémoire d'entreprise interactive** qui, à la différence de nos collègues humains, n'oublie jamais. Équipée d'une interface appropriée, cette base de données bi-textuelle permettrait aux utilisateurs de trouver des solutions à un vaste éventail de problèmes de traduction. Et de fait, nos premiers travaux avec le Journal des débats nous ont convaincus que ces traductions recèlent une mine de trouvailles jusqu'ici inexploitée. Dans la section 4, ci-après, nous décrirons brièvement d'autres outils de traduction qui sont basés sur la notion de bi-textes et qui vont au-delà des concordances bilingues.

Au cours de la dernière décennie, les pays occidentaux industrialisés ont enfin commencé à réaliser qu'ils ne pouvaient plus se permettre de jeter les bouteilles vides, les vieilles voitures et les vieux journaux qu'ils produisent en grande quantité, et qu'ils doivent plutôt recycler tous ces produits. Peut-être le temps est-il venu de comprendre que la même chose vaut pour une partie de notre production intellectuelle. Avec l'avènement des bi-textes, les traducteurs n'auront peut-être plus à réinventer constamment des solutions qui ont déjà été trouvées à grand peine par leurs prédécesseurs. Au lieu de jeter nos traductions, il se pourrait que nous soyons enfin en mesure de les recycler de façon rentable.

3. LES PRÉCURSEURS

Dans la section précédente, nous avons tenté de mettre en évidence certaines des différences qui existent entre les concordanciers bilingues et les logiciels de recherche plein texte. Alors que ces derniers créent généralement un seul index inverse de tous les fichiers textes d'une base de données, notre concordancier bilingue dépend d'un algorithme d'alignement automatique pour relier les unités de traduction entre des *paires* de fichiers textes. Le résultat est, à notre avis, un outil d'aide à la traduction radicalement nouveau. Évidemment, les innovations révolutionnaires surgissent rarement du néant et nous examinerons donc, dans cette section, certains des précurseurs, commerciaux et scientifiques, du concordancier bilingue.

Les concordanciers bilingues ne sont pas encore offerts dans le commerce, mais il existe d'autres types de logiciels d'aide à la traduction qui génèrent automatiquement des liens explicites entre les unités d'un texte source et celles d'un texte cible. Le mieux connu est peut-être le produit qu'ALPS commercialisait sous le nom de TSS[9]. En mode “AutoTerm”, ce logiciel traitait un texte source phrase par phrase; il cherchait la traduction de chacun des mots de la phrase source dans un glossaire spécifique au document, puis affichait les résultats dans une petite fenêtre située au bas de l'écran. Le traducteur pouvait ensuite insérer autant de ces termes qu'il le désirait dans sa traduction, puis n'avait qu'à appuyer sur la touche Enter pour afficher la prochaine phrase à traduire. À l'activation de cette commande, cependant, le système reliait automatiquement les unités source et cible précédentes. En fait, avant d'afficher chaque nouvelle phrase à traduire, le TSS parcourait la portion déjà traduite du document source (ainsi qu'un fichier de référence spécifié) à la recherche d'une phrase identique en langue source : s'il en trouvait une, il affichait automatiquement la traduction qui y était reliée. Grâce à ce fichier de répétitions, clamait la publicité d'ALPS, le traducteur n'avait jamais à traduire deux fois la même phrase[10].

Il s'agissait là d'une excellente idée, qui a d'ailleurs été reprise par la suite par au moins un autre concepteur de logiciels de TA. Mais comment se compare-t-elle à notre système de concordancier bilingue? La première différence évidente est que le TSS ne relie que des segments de textes qui ont été traduits à l'aide de ce logiciel, ce qui signifie que les archives électroniques d'un service de traduction qui auraient été stockées avant l'introduction du TSS ne seraient pas accessibles à la fonction de repérage des répétitions. Pourquoi? Parce que le TSS n'incorpore pas d'algorithme d'alignement automatique du genre de celui que nous avons décrit précédemment. Le TSS dépend plutôt du traducteur humain, qui doit lui indiquer l'unité ou les unités du texte cible qui constituent la traduction de chaque phrase du texte source; ce n'est qu'après cela que le système relie explicitement les deux segments. Par contre, les algorithmes d'alignement que nous utilisons dans notre programme de concordancier bilingue sont entièrement automatiques, en ce sens qu'ils

9. Je ne suis pas certain si ALPS commercialise toujours le TSS, ni même si l'entreprise existe encore. La société UNITRAN d'Ottawa offre un produit qui est très semblable à TSS.

10. ALPS a amélioré par la suite le fonctionnement de ce fichier de répétitions, en assouplissant la nécessité d'une identité stricte entre les deux phrases sources et en autorisant des correspondances floues dans le cas des nombres et des dates.

peuvent, à partir de n'importe quelle paire de fichiers source et cible, **calculer** les unités qui doivent être alignées. Aucune intervention humaine n'est nécessaire[11].

Deuxième différence importante entre les deux approches : avec le TSS, c'est le système qui prend l'initiative de consulter la base de données bi-textuelle. Le TSS recherche systématiquement une copie de chaque nouvelle phrase source dans le fichier de référence spécifié, et l'utilisateur n'a que très peu de choix en la matière. Par contre, dans notre façon d'envisager l'utilisation d'un concordancier bilingue, c'est l'utilisateur qui décide quand il veut consulter la base de données bi-textuelle et sur quelles unités portera la recherche – phrases complètes, syntagmes, expressions ou mots. Comme nous l'avons suggéré dans le petit scénario présenté à la première section, nous envisageons le concordancier bilingue comme un outil souple qui viendra compléter les ressources lexicologiques et terminologiques existantes, une source d'inspiration qui s'enrichira constamment par l'addition des nouvelles traductions réalisées.

En ce qui concerne les antécédents scientifiques du concordancier bilingue, nous avons déjà mentionné les articles de Brian Harris sur les utilisations d'un tel outil dans le domaine de la traduction, ainsi que les travaux de développement d'algorithmes d'alignement automatique réalisés par Church et Gale chez AT&T. Ces algorithmes sont basés sur l'observation qu'il existe une corrélation entre la longueur d'un texte et la longueur de sa traduction; pour aligner deux textes, Church et Gale utilisent une mesure de distance locale qui compare le nombre de caractères dans des régions concurrentes. Cette méthode ne constitue toutefois pas la seule façon d'aligner automatiquement des textes. D'autres algorithmes utilisent des mots correspondants dans les textes source et cible pour établir des alignements de phrases; c'est l'approche adoptée par Catizone et al [2], ainsi que par Kay et Röscheisen [8]. Ces derniers, soit dit en passant, semblent avoir été les premiers à proposer un concordancier bilingue. Fait intéressant à noter, c'est aussi Martin Kay qui fut le premier, il y a plus de dix ans, à suggérer l'idée d'un poste de travail du traducteur comme solution de rechange à la traduction entièrement automatique (voir la référence bibliographique 7). Le projet du CITI, qui vise à intégrer un concordancier bilingue au poste de travail de traducteur et qui utilise un algorithme d'alignement basé sur la longueur des phrases pour actualiser automatiquement la base de données bi-textuelle, emprunte librement à tous ces précurseurs.

11. Cela ne veut pas dire que les algorithmes statistiques utilisés produisent toujours des alignements parfaits; mais dans nos expériences sur le corpus du Journal des débats, nous obtenons des alignements corrects dans plus de 98 p.cent des cas.

4. AUTRES UTILISATIONS, AUTRES UTILISATEURS

La plupart des recherches scientifiques récentes dans le domaine des corpus bilingues sont destinées à d'autres utilisateurs que les traducteurs. Par exemple, les articles de Catizone et al. [2], Church et Gale [3] et Klavens et Tzoukermann [9] s'adressent tous aux lexicographes-informaticiens ou aux linguistes qui s'intéressent au problème des équivalences traductionnelles. Les traducteurs ne sont pas vraiment non plus la clientèle visée par le nombre croissant de nouveaux projets dans le domaine de la traduction automatique à base de corpus, que ces projets appliquent les approches purement statistiques pilotées par Brown et al.[1] ou les approches à base d'exemples introduites par Sato et Nagao [11]. Les systèmes mis au point suivant ces deux approches dépendent de la production des traducteurs – c'est-à-dire de vastes corpus de textes alignés – comme source première de leurs connaissances linguistiques et interlinguistiques; mais au bout du compte, ces chercheurs n'ont pas pour but de concevoir des outils de travail pour les traducteurs humains, mais plutôt d'élaborer des systèmes entièrement automatisés qui fonctionneront aussi bien que les traducteurs humains et finiront par les remplacer. Par contraste, les travaux du CITI sur les bi-textes sont carrément orientés vers la traduction humaine assistée par ordinateur et répondent, de ce fait, aux préoccupations immédiates des traducteurs. Nous décrirons brièvement, dans cette section, d'autres outils à base de corpus qui sont destinés aux traducteurs et que nous espérons pouvoir intégrer aux futures versions du PTT.

La conversion d'un ensemble de documents en une base de données bi-textuelle, où les segments d'un texte source et les segments du texte cible qui en constituent la traduction sont explicitement reliés, rend possible la vérification automatique de certaines caractéristiques propres aux textes qui présentent une relation traductionnelle réciproque. Considérons, par exemple, la caractéristique de l'intégralité. On exige normalement de la traduction d'un texte qu'elle n'omette aucun des éléments d'information contenus dans le texte source. Il arrive néanmoins, et peut-être même plus souvent qu'on ne voudrait l'admettre, que le traducteur omette sans s'en tenir compte une phrase ou un paragraphe ou même, dans un texte volumineux, une page entière. À l'heure actuelle, ces oublis ne peuvent être détectés que par une révision manuelle soigneuse des deux textes. Dans un bi-texte correctement aligné, cependant, ces omissions seraient immédiatement apparentes : la représentation bi-textuelle attirerait l'attention du traducteur ou du réviseur sur le fait qu'un segment particulier du texte source n'a **aucun segment** correspondant dans le texte cible. En fait, on peut aisément imaginer un outil de vérification qui signalerait ou imprimerait automatiquement ces alignements douteux.

Cela n'est qu'une des façons dont une représentation bi-textuelle permettrait d'éviter les erreurs de traduction les plus grossières. Considérons maintenant le problème de traduction classique des faux- amis. Les faux-amis sont des paires de mots dans des langues différentes qui ont une étymologie commune mais qui ont adopté, avec le temps, des sens différents; chaque mot est parfaitement licite dans sa langue respective, mais ils ne peuvent pas être utilisé comme équivalents traductionnels. Il existe de nombreuses études qui répertorient des centaines, voire des milliers de faux amis anglais-français; pour n'en citer qu'un exemple évident, prenons la paire de mots “library” et “librairie”, où le dernier terme désigne un magasin où l'on vend des livres, alors que “bibliothèque” est l'équivalent français correct de “library”. Il ne faut pas confondre les faux-amis avec les anglicismes ou les gallicismes, qui sont des emprunts illicites faits à une autre langue et qui ne font pas partie du vocabulaire reconnu; “scholarity”, par exemple, ne fait pas partie du lexique anglais, pas plus que “rewinder” n'est un verbe licite en français. En principe, un vérificateur orthographique ou stylistique unilingue sera en mesure de détecter de telles interférences, comme il peut détecter d'autres mots mal orthographiés ou inconnus. Les faux-amis, par contre, sont des mots corrects dans les deux langues; ce qui est illicite, c'est leur utilisation comme équivalents traductionnels réciproques. Les vérificateurs orthographiques ou stylistiques unilingues sont impuissants face à ce problème, car par définition ils ne vérifient qu'un fichier linguistique à la fois. Ce qu'il nous faut, très évidemment, c'est une représentation bi-textuelle, qui relie les segments correspondants dans les deux fichiers, et un anti-dictionnaire, ou liste de paires illicites qui ne peuvent être utilisées comme des équivalents traductionnels dans ces segments. Un programme comportant un détecteur de faux-amis vérifierait systématiquement tous les mots qui figurent dans des segments correspondants d'un bi-texte par rapport à une liste de faux-amis; si, par exemple, il repérait le mot “librairie” dans le segment français correspondant à un segment anglais renfermant le mot “library”, il signalerait une erreur possible.

Si tous les cas de faux-amis étaient aussi évidents que “library/librairie”, le problème de leur détection automatique serait alors relativement simplifié. Il s'avère toutefois que la plupart des faux-amis sont partiels et non absolus, c'est-à-dire que leur traduction réciproque est interdite, mais pas dans tous les contextes; et il n'est pas du tout simple de caractériser les contextes dans lesquels ces paires peuvent ou non constituer des traductions réciproques acceptables. Une deuxième difficulté concerne la résolution de l'alignement qui serait nécessaire pour permettre à un programme de détecter des faux-amis de manière précise et efficace (voir la note n^{o} 6, ci-dessus). Ce qui rend souvent les vérificateurs grammaticaux et stylistiques commerciaux si frustrants à utiliser, c'est qu'ils signalent trop de fausses erreurs. Si notre détecteur de faux-amis n'avait accès qu'à des alignements de phrases et signalait des erreurs potentielles chaque fois qu'il rencontrait des faux-amis n'importe où à l'intérieur de deux phrases complètes, on courrait le risque d'irriter les utilisateurs de la même façon. La solution à notre problème, semble-t-il, est de produire des alignements plus fins, ce que nous croyons pouvoir réaliser.

Vérifier l'intégralité d'une traduction ou détecter la présence de faux-amis fait appel à un outil d'aide à la traduction bien différent de ceux que nous avons réussi à intégrer jusqu'à présent à notre PTT. D'abord, les caractéristiques à vérifier sont inhérentes à l'activité traduisante elle-même. À la différence d'un vérificateur orthographique ou d'un programme de conjugaison, ces outils ne sont pas conçus pour l'ensemble des rédacteurs unilingues, mais bien pour le traducteur et son réviseur. Que le PTT réponde aux besoins spécifiques du réviseur constitue aussi un progrès important, car nous n'avons eu jusqu'à présent que très peu à offrir à cette catégorie de professionnels hautement rémunérés, dont le temps est trop précieux pour être consacré à des tâches mécaniques qui peuvent être exécutées plus efficacement par une machine.

En somme, nous pensons que les perspectives en ce qui concerne ce genre d'outils sont extrêmement intéressantes, et nous sommes convaincus que nous n'avons que commencé à effleurer la surface de la bi-textualité.

REMERCIEMENTS

Cet article résume les travaux réalisés au CITI par plusieurs de mes collègues, dont les noms n'apparaissent pas à côté du mien, même s'ils ont parfois contribué plus que moi à l'avancement du projet de bi-textualité. Il s'agit, d'abord et avant tout, de Pierre Isabelle, directeur du groupe de traduction assistée par ordinateur du CITI, ainsi que de George Foster, Marie-Louise Hannan, Jean-Marc Jutras, François Perrault, XiaoBo Ren, Benoît Robichaud et Michel Simard. J'aimerais leur exprimer ma reconnaissance et les prier d'excuser les erreurs que j'aurais pu commettre en résumant leurs travaux.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L Mercer, et Paul S. Roossin : "A statistical approach to machine translation", *Computational Linguistics*, vol. 16, no 2, juin 1990, p. 79-85.

[2] Catizone, Roberta, G. Russell, et S. Warwick : "Deriving Translation Data from Bilingual Texts", U. Zernick (éd.), *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum, 1992.

[3] Church, K. et W. Gale : "Concordances for Parallel Texts", *Proceedings of the 7th Annual Conference of the UW Centre for the NOED and Text Research*, Oxford, 1991.

[4] Harris, B. : "Bi-text: A New Concept in Translation Theory", *Language Monthly*, no 54, p. 8-10, 1988.

[5] Harris, B. : "Are you Bi-Textual?", *Language Technology*, no 7, p. 41, 1988.

[6] Isabelle, Pierre : "La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie", *META*, vol. 37, no 4, p. 721-737, décembre 1992.

[7] Kay, Martin : "The Proper Place of Men and Machines in Translation", CLS-80-11, Xerox Palo Alto Research Center, Palo Alto, Californie, 1980.

[8] Kay, Martin. et M. Röscheisen : "Text-Translation Alignment", manuscrit inédit, Xerox PARC, 1988.

[9] Klavens, J. et E. Tzoukermann : "The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries", *Proceedings of COLING-90*, vol. 3, p. 174-179, 1990.

[10] Macklovitch, E. : "PTT-2: A Second Version of the CWARC's Workstation for Translators", rapport technique, CCRIT, 1991.

[11] Sato, S. et M. Nagao : "Toward Memory-Based Translation", *Proceedings of COLING-90*, p. 247-252, 1990.

[12] Simard, M., G. Foster, et P. Isabelle : "Using Cognates to Align Sentences in Parallel Corpora", *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, p. 67-81, 1992.