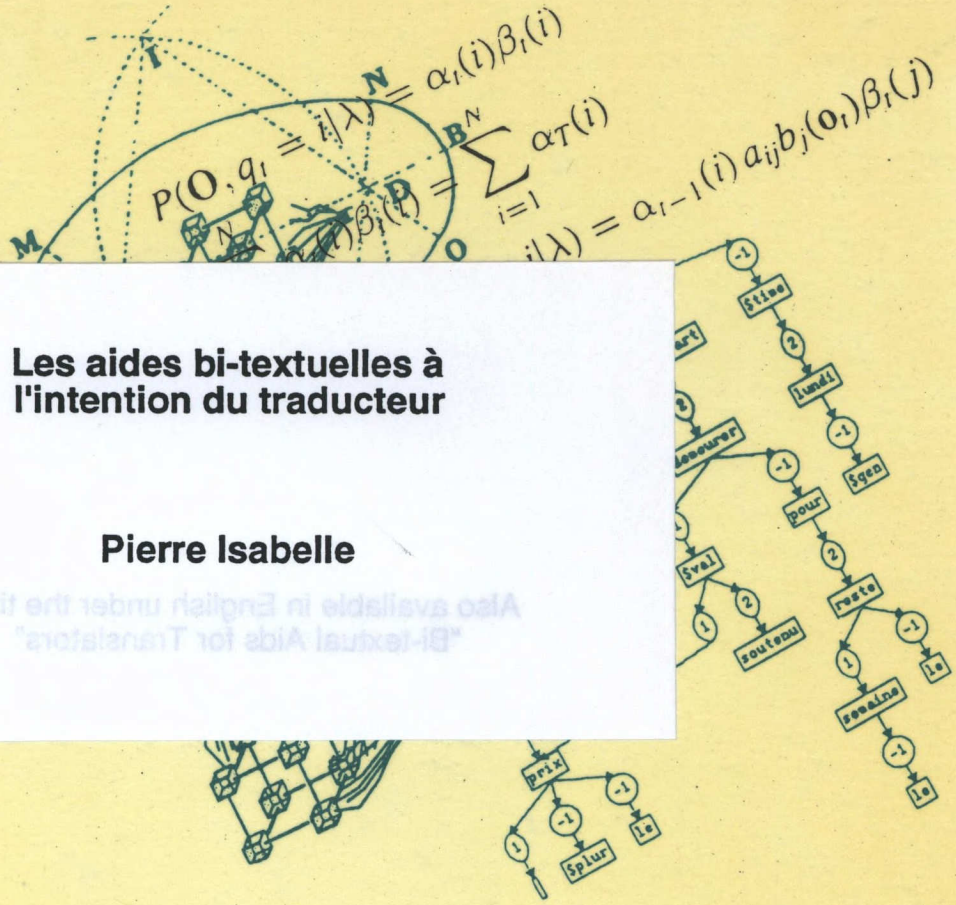
Industrie et Sciences Canada | Industry and Science Canada



Les aides bi-textuelles à l'intention du traducteur

Pierre Isabelle

CITI

Centre d'innovation en technologies de l'information

Centre for Information Technologies Innovation

QUEEN
P
308
.I3714
1993
c.2

Canada

Queen
P308
.I37
1993
c.2

Les aides bi-textuelles à l'intention du traducteur

Pierre Isabelle

Publié en version anglaise sous le titre de «Bi-Textual Aids for Translators», dans Proceedings of the 8th Annual Conference of the UW Centre for Text Research, University of Waterloo, 1992, p. 76-89.

Industry Canada
Library Queen
JUIL JUL 24 1998
Industrie Canada
Bibliothèque Queen

Laval
1993

Also available in English under the title:
"Bi-textual Aids for Translators"

Ce document fait état de travaux de recherche réalisés dans le cadre des activités du Centre d'innovation en technologies de l'information (CITI), Industrie et Sciences Canada. Les opinions exprimées dans ce document n'engagent que l'auteur.

© Copyright Industrie et Sciences Canada 1993
No catalogue Co28-1/106-1993F
No ISBN 0-662-98646-6

Les aides bi-textuelles à l'intention du traducteur

Pierre Isabelle

Centre d'innovation en technologies de l'information
1575, boul. Chomedey, Laval (Québec) H7V 2X2
Courrier électronique : isabelle@condor.citi.doc.ca

RÉSUMÉ

Les systèmes de traduction automatique fonctionnent bien avec certains sous-langages extrêmement restreints, mais il est plus productif, dans la plupart des cas, de recourir aux outils de soutien destinés au traducteur humain. Les fonctions réalisées par les postes de travail de traducteur sont encore plutôt accessoires par rapport aux aspects fondamentaux de la traduction. Cependant, des recherches récentes montrent qu'il est possible de produire automatiquement des représentations explicites (bien que partielles) des correspondances traductionnelles qui relient un texte source et sa traduction. Ces représentations, appelées bi-textes, peuvent servir de point de départ à la conception d'outils de travail qui relèvent davantage de la traduction proprement dite, par exemple : a) une mémoire traductionnelle exploitable par divers moyens, y compris un concordancier bilingue; b) des outils de critique de traductions capables de détecter les erreurs de correspondance, comme les omissions ou les faux-amis et c) des systèmes de reconnaissance de la parole orientés vers la traduction, qui sont capables d'exploiter les contraintes de correspondance par rapport aux textes de départ. Les perspectives qui s'ouvrent aux outils de soutien à la traduction semblent donc très prometteuses.

1. Introduction

Malgré les efforts massifs déployés depuis plusieurs décennies, la traduction automatique (TA) ne donne encore des résultats vraiment satisfaisants que lorsqu'elle est appliquée à certains sous-langages très restreints, comme celui qui est visé par le système TAUM-MÉTÉO (Isabelle [17]). De plus, le fait qu'on n'ait pas réussi à reproduire systématiquement l'éclatant succès de ce système

semble indiquer qu'il existe actuellement très peu de sous-langages simples qui génèrent un important volume de traduction.

Ainsi, à l'exception d'un très petit nombre de cas, la traduction automatique en est au même point qu'en 1951, quand Bar-Hillel [1] écrivit :

> Dans le cas des domaines cibles où la précision absolue est une condition *sine qua non*, il faut renoncer à la TA pure au profit d'une TA mixte, c'est-à-dire un processus traductionnel faisant intervenir l'intelligence humaine. Ce qui soulève la question suivante : Quelles étapes du processus devrait-on confier à un partenaire humain? (p. 230) [traduction]

Bar-Hillel lui-même favorisait les méthodes dans lesquelles l'intervention humaine précédait (pré-édition) ou suivait (post-édition) le processus mécanique, «mais de préférence pas quelque part au milieu». Il envisageait une démarche où la machine prenait en charge l'essentiel du processus traductionnel.

Après quatre décennies d'efforts acharnés, les chefs d'accusation qui s'accumulent contre la pré-édition et la post-édition en TA sont devenus très compromettants. D'une part, personne n'a encore réussi à mettre au point une méthode pratique de pré-édition des textes généraux qui puisse garantir la qualité constante des résultats produits par le processus ultérieur de TA. D'autre part, il a été démontré à de nombreuses reprises qu'il n'est pas rentable de recourir aux post-éditeurs humains pour récupérer les sorties médiocres qui sont produites, dans la plupart des cas, par les systèmes de TA actuels (voir, par exemple, Macklovitch [23]). Il n'est donc pas étonnant que la plupart des services de traduction considèrent que, dans l'état actuel de la technologie, ils ne peuvent tirer aucun profit de la TA, qui demeure un phénomène marginal sur le marché actuel de la traduction. Rien n'indique que cette situation soit sur le point de changer.

Il y a plus de dix ans, Martin Kay [18] proposait son *secrétaire ou alter ego du traducteur*, qui constituait une réponse très différente à la question de Bar-Hillel au sujet de la division du travail entre l'homme et la machine :

> Je veux préconiser une approche au problème selon laquelle on permettrait aux machines de prendre en charge graduellement, presque imperceptiblement, certaines fonctions du processus général de traduction. Les machines assumeraient d'abord des tâches non essentiellement reliées à la traduction. Puis, peu à peu, elles

> s'attaqueraient à la traduction elle-même. Tout reposerait sur la modestie. À chaque étape, la machine ne ferait que ce qu'elle sait bien faire. Et ainsi, petit à petit, l'oiseau ferait son nid! (p. 11) [traduction]

Au lieu de commencer par des systèmes moins qu'optimaux et de demander au traducteur de compenser leurs lacunes, il s'agirait de partir de la traduction humaine et de chercher des façons, même modestes, de tirer profit de la machine. C'est pour cette approche réaliste que le Centre d'innovation en technologies de l'information (CITI) a opté lorsqu'il a lancé son projet de poste de travail de traducteur en 1987 (Macklovitch [21], [22]). Dans sa plus récente incarnation, le poste de travail du CITI met à la disposition du traducteur un environnement multi-fenêtres qui lui donne simultanément accès à plusieurs outils, comme le traitement de texte en écran partagé, la vérification orthographique, la consultation terminologique et lexicographique, la comparaison de fichiers, le compte de mots, etc.

Ce poste de travail, comme la plupart des autres postes existants, en est encore au stade initial de développement, en ce sens que la plupart des fonctions réalisées par la machine relèvent plus de la bureautique que des aspects fondamentaux de l'activité traduisante. Pourtant, les résultats obtenus au Bureau de la traduction du Secrétariat d'État du Canada, où ce poste de travail est à l'essai depuis environ trois ans, ont été bien supérieurs à ceux qu'ont obtenus les systèmes de TA. À preuve : les utilisateurs sont si satisfaits de cet outil qu'ils veulent le garder!

Selon le scénario proposé par Kay, nous pourrions maintenant tirer profit de cet environnement bureautique pour mettre à la disposition du traducteur de nouveaux outils qui s'avanceront plus profondément dans le domaine de la traduction proprement dite. Nous avons besoin, pour ce faire, d'un cadre conceptuel qui nous procure des points d'accès appropriés pour la technologie. Nous croyons que le concept de *bi-texte* constitue justement un tel cadre, qui nous ouvre une multitude de nouvelles possibilités pour le soutien de l'activité traduisante. Nous décrivons ce concept dans la section suivante. Dans la section 3, nous décrivons comment il est possible de produire automatiquement des représentations bi-textuelles. Puis, dans les sections 4, 5 et 6, nous examinons trois types d'outils qui peuvent prendre appui sur ces représentations.

2. Le concept de bi-texte

Quelle est la caractéristique essentielle qui distingue le traducteur des autres langagiers? La réponse est évidente. Le traducteur agit non pas sur un mais bien sur deux textes : le texte source (TS) préexistant et le texte cible (TC) à produire dans une autre langue, en respectant la contrainte que le TS et le TC présentent une relation d'équivalence traductionnelle. La satisfaction de cette contrainte constitue le pivot de l'activité traduisante. On pourrait, par conséquent, s'attendre à ce que les outils destinés expressément au traducteur incorporent certaines connaissances relatives à l'équivalence traductionnelle.

Pour que la traduction soit possible, l'équivalence traductionnelle doit obéir au principe de *compositionnalité*, c'est-à-dire que la traduction d'un texte doit être une fonction de la traduction de ses parties, et ce, jusqu'au niveau d'un ensemble fini d'équivalences élémentaires (entre des mots et des syntagmes, par exemple). Les dictionnaires et banques de termes multilingues ont pour objet de répertorier certaines de ces équivalences élémentaires entre langues différentes. Ils constituent actuellement les meilleurs exemples d'outils spécifiques à la traduction qui sont incorporés aux postes de travail de traducteur existants.

Cependant, quiconque a déjà essayé de traduire des textes en langage naturel admettra que la traduction demeure une tâche difficile, même lorsqu'on dispose des meilleurs dictionnaires et banques de données terminologiques. Ces ressources lexicographiques ne décrivent que des équivalences virtuelles. En général, elles répertorient plusieurs équivalents possibles en langue cible (LC) pour chaque élément en langue source (LS), et il incombe au traducteur de choisir l'équivalent approprié à son texte, en tenant compte de divers facteurs contextuels. De plus, les ressources lexicographiques ne sont jamais exhaustives et ne réussissent jamais à épuiser la gamme complète des équivalences possibles.

Les traductions existantes constituent la seule source d'équivalences (c'est-à-dire de correspondances) réalisées. En vertu du principe de la compositionnalité, il est normalement possible de décomposer, par l'analyse, la

correspondance globale entre un texte (TS) et sa traduction (TC) en groupes de correspondances plus fines entre segments particuliers du TS et du TC. Comme le souligne Harris [15], les mises en page «côte-à-côte» ou «interlinéaires» communément utilisées pour les traductions présupposent que les correspondances traductionnelles (entre paragraphes, entre phrases, etc.) sont directement analysables. D'ailleurs, n'importe quel locuteur bilingue serait capable d'identifier la plupart, sinon toutes, des correspondances entre les éléments d'un texte source et sa traduction.

Harris [15], [16] propose le terme de *bi-texte* pour désigner tout couplage explicite des segments d'un texte source avec ceux de sa traduction. Nous adoptons ce terme, en lui donnant la définition technique suivante : un bi-texte est un quadruplet $<T_1, T_2, Fs, C>$, dans lequel T_1 et T_2 sont deux textes, Fs est une fonction qui décompose T_1 en un ensemble quelconque d'éléments $Fs(T_1)$ et T_2 en un ensemble quelconque d'éléments $Fs(T_2)$, et C est un sous-ensemble du produit cartésien $Fs(T_1) \times Fs(T_2)$.

Cette définition soulève plusieurs questions importantes. Premièrement, quelle est la nature des éléments produits par la fonction d'analyse *Fs*? Une possibilité relativement évidente serait d'assimiler *Fs* à une fonction d'analyse syntaxique. Dans ce cas, on peut présumer que *Fs* structurera chacun des textes en un système hiérarchique quelconque : les textes sont composés de sections, les sections de paragraphes, les paragraphes de phrases, les phrases de syntagmes, les syntagmes de mots et les mots, de morphèmes.

Une deuxième question, connexe à la première, concerne la nature de la fonction de mise en correspondance *C*. Nous avons déjà indiqué que cette fonction doit faire intervenir un certain degré de compositionnalité. Une fonction d'analyse hiérarchique satisfait naturellement à cette exigence, car son résultat constitue un domaine naturel pour les correspondances traductionnelles hiérarchisées : la traduction d'une section est composée des traductions de ses paragraphes, la traduction d'un paragraphe est composée des traductions de ses phrases, etc., comme l'illustre la Figure 1.

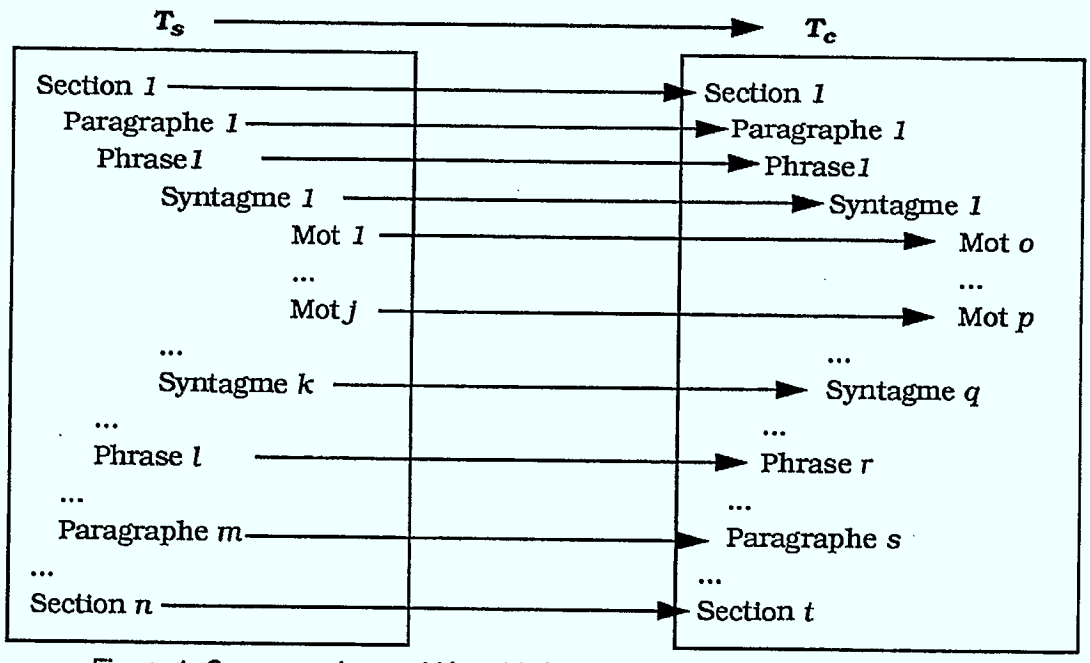


Figure 1: Correspondances hiérarchisées entre un texte source et sa traduction

Il suffit toutefois d'examiner des traductions pour constater que les correspondances traductionnelles ne se réduisent pas systématiquement à de simples correspondances biunivoques (un-pour-un). Cela est particulièrement vrai dans le cas des unités plus élémentaires. Par exemple, le mot anglais *potato* se traduit habituellement en français par *pomme de terre*, soit une suite de trois mots. Le modèle hiérarchique facilite la formulation des correspondances de ce genre, car il permet d'établir une correspondance entre unités syntagmatiques abstraites : le syntagme *pomme de terre* correspond au nom *potato*.

Mais il est facile de répertorier des cas où la structure syntaxique superficielle ne permet pas de caractériser toutes les unités qui entrent en jeu dans les correspondances traductionnelles. Par exemple, le constituant discontinu *ne... pas* n'est pas représenté comme une unité à ce niveau, mais il participe effectivement à une correspondance traductionnelle avec *not*, dans le couple de phrases suivant :

1) a) Max n'a pas vu Conrad.
   b) Max has not seen Conrad.

De même, le constituant discontinu *turned... on* est une unité qui se traduit par *alluma* dans le couple de phrases suivant :

2) a) Max turned the radio on.
   b) Max alluma la radio.

En général, il semble donc que certaines correspondances traductionnelles ne puissent être exprimées que par le type de représentations abstraites (représentations syntaxiques profondes, représentations sémantiques, représentations conceptuelles, etc.) auxquelles les systèmes de TA doivent faire appel pour produire des traductions. Notre connaissance de ces modèles de représentation est loin d'être complète et il est remarquablement difficile d'élaborer des algorithmes capables de mettre en correspondance des textes non restreints en langage naturel. Par conséquent, la construction d'un dispositif capable de produire automatiquement des bi-textes complets (c'est-à-dire qui rendent explicites toutes les correspondances traductionnelles) pour des couples de textes arbitraires pourrait se révéler très ardue.

Pourtant, les possibilités de production de bi-textes sont beaucoup plus prometteuses que ne le sont celles des systèmes de TA, et ce, pour au moins deux raisons. Premièrement, à la différence de la traduction, qui exige une compétence linguistique active, la reconstruction des correspondances entre un texte et sa traduction ne nécessite qu'une compétence linguistique passive, qui devrait en principe être plus facile à caractériser. Deuxièmement, contrairement aux traductions, qui doivent généralement être complètes pour être utiles à leurs destinataires, les représentations bi-textuelles même partielles (c'est-à-dire qui n'expriment qu'un sous-ensemble des correspondances existant entre un texte et sa traduction) demeurent très utiles, comme nous le verrons plus loin. En ce qui concerne le modèle hiérarchique dont nous avons parlé précédemment, il est possible de catégoriser les représentations bi-textuelles en fonction de leur niveau de *résolution*. Un bi-texte à très basse résolution n'exprimera que les correspondances entre les unités de rang supérieur, comme les sections ou les paragraphes.[1] La résolution augmente à mesure que l'on spécifie les

1 Les mises en page côte-à-côte n'expriment généralement que les correspondences au niveau des paragraphes.

correspondances entre unités de rang inférieur : phrases, syntagmes, mots et morphèmes. Cette notion de résolution n'a pas d'homologue en traduction : on ne saurait traduire une unité (disons une phrase) sans traduire en même temps ses éléments (syntagmes, mots).

Le niveau de résolution peut donc servir de critère d'évaluation de bi-textes particuliers. Un autre critère évident est le niveau de *précision*, c'est-à-dire la proportion de correspondances représentées qui sont, en réalité, correctes.

3. La production de représentations bi-textuelles

3.1 L'alignement de phrases

Martin Kay a été, à ma connaissance, le premier chercheur à proposer des méthodes de reconstruction des correspondances traductionnelles entre textes préexistants, et qui ont permis la production automatique de ce que nous avons appelé des représentations bi-textuelles. C'est en 1984 que je l'ai d'abord entendu décrire, de façon informelle, un algorithme dont il fit par la suite, avec Röscheisen [19], une description systématique. Cet algorithme n'a pas pour objet d'identifier toutes les correspondances, mais seulement de produire un «alignement» correct au niveau de la phrase. Ce qui vient évidemment compliquer le problème, c'est que les correspondances ne sont pas nécessairement biunivoques (un-pour-un) : une phrase du texte source peut être rendue par deux ou plusieurs phrases dans le texte cible; inversement, plusieurs phrases du texte source peuvent être «contractées» en une seule phrase dans le texte cible.

L'algorithme de Kay et Röscheisen examine d'abord tous les alignements de phrases possibles[2], puis il sélectionne celui qui maximise le nombre de correspondances systématiques supposées au niveau du mot. Supposons, par exemple, que le texte source contienne exactement dix occurrences du mot *dog* et que le texte cible contienne précisément dix occurrences du mot *chien*. Toutes

2 La gamme des alignements possibles est restreinte au départ. Les alignements croisés sont interdits et les alignements multivoques (*many-to-many*) ne sont autorisés que si leur nombre est limité, ne dépassant pas un petit nombre *n*.

autres choses étant égales par ailleurs, le système favorisera les alignements de phrases faisant correspondre les occurrences de *dog* et de *chien*. Étant donné que les alignements de phrases et les alignements de mots sont interdépendants, l'algorithme de Kay et Röscheisen est basé sur un processus de raffinement itératif.

Les auteurs affirment avoir obtenu des résultats presque parfaits avec leurs corpus d'essai (deux articles de *Scientific American* et leurs traductions allemandes, et 1000 phrases tirées des données bilingues du *Journal des débats de la Chambre des communes*). L'une des caractéristiques intéressantes de cette approche est qu'elle ne fait appel à aucune référence extérieure aux textes eux-mêmes, par exemple un dictionnaire bilingue.

Catizone *et al.* [5] soutiennent toutefois que l'espace de recherche est radicalement réduit lorsqu'on élargit la méthode de Kay et Röscheisen pour y intégrer l'utilisation d'un dictionnaire bilingue permettant de guider les hypothèses initiales sur les correspondances entre les mots.

Debili et Sammouda proposent, pour leur part, un algorithme d'alignement différent, mais également basé sur les correspondances lexicales établies à l'aide d'un dictionnaire bilingue.

Deux groupes de chercheurs, soit Brown, Lai et Mercer [3], et Gale et Church [12], abordent différemment le problème de l'alignement des phrases de textes parallèles. Ils proposent tous deux des méthodes qui sont basées sur l'observation très simple qu'il existe une forte corrélation entre la longueur d'un texte et la longueur de sa traduction. Ces méthodes présentent elles aussi l'avantage de ne recourir à aucune donnée extérieure. Elles exigent beaucoup moins de temps de calcul que la méthode de Kay et Röscheisen, et produisent d'étonnamment bons résultats avec les données du *Journal des débats*. Cependant, du fait qu'elle ne tiennent aucun compte du contenu des phrases qu'elles apparient, ces méthodes apparaissent moins fiables et moins robustes, car dès qu'un algorithme basé sur la longueur des phrases aligne incorrectement deux phrases, il tend à désaligner le reste du paragraphe.

Enfin, Simard, Foster et Isabelle [56] examinent un autre critère pouvant servir de base à l'alignement de phrases, soit le «degré de parenté» entre un texte source et sa traduction. Ils signalent qu'il y a généralement un fort «degré de parenté», c'est-à-dire une forte proportion de mots apparentés (*cognates*), entre un texte source et sa traduction. Il est facile de définir des approximations opérationnelles relativement fiables de la notion de «mots apparentés» (en vertu d'un préfixe commun, par exemple), ce qui permet de tester le «degré de parenté» sans recourir à des données extérieures. Les auteurs signalent que lorsqu'il est utilisé comme unique critère d'alignement, le degré de parenté ne produit pas de très bons résultats, mais que lorsqu'il est utilisé de concert avec le critère de la longueur, il améliore la précision et la robustesse sans pour autant augmenter radicalement le temps de calcul.

3.2 Les correspondances lexicales

Comme nous l'avons vu, certains algorithmes d'alignement de phrases (ceux de Kay et Röscheisen et de Debili et Sammouda) procèdent en posant des hypothèses relatives à certaines des correspondances lexicales reliant les textes source et cible. Il serait peut-être possible d'utiliser ces correspondances pour produire des bi-textes à plus haute résolution, mais les auteurs en disent très peu sur leur couverture et leur précision.

Le mécanisme d'alignement de phrases de Brown, Lai et Mercer n'est que la première étape d'une procédure visant à estimer les paramètres du système de TA stochastique décrit par Brown *et al.* [2]. Afin d'estimer les paramètres d'un dictionnaire de transferts probabiliste pour ce système, ils doivent ensuite rendre explicites les correspondances lexicales identifiées dans leur corpus de paires de phrases (une portion des données du *Journal des débats*). Pour ce faire, ils utilisent une version particulière de l'algorithme EM (Dempster *et al.* [10]), qui devrait leur permettre d'obtenir une couverture complète. Les auteurs ne fournissent toutefois aucun détail sur le niveau de précision de leurs résultats.

Gale et Church [13] présentent une méthode qui permet d'identifier certaines des correspondances lexicales existant entre des textes déjà alignés au niveau de la phrase. Ils commencent par définir un ensemble de paires de mots qui sont

fortement associées dans des paires de phrases, en appliquant une statistique de type χ^2 à des tableaux de contingence et en sélectionnant des paires de mots dont le degré d'association dépasse un certain seuil. Ils utilisent alors ces paires de mots pour marquer les correspondances lexicales probables dans les paires de phrases. Selon les auteurs, quand les seuils pertinents sont été fixés de façon à obtenir un taux de couverture de 60 p. cent, les correspondances sont correctes dans 95 p. cent des cas.

3.3 Conclusions

Bien que la bi-textualité soit un domaine de recherche très récent, il est déjà possible de produire automatiquement des bi-textes d'un niveau de résolution limité, mais d'un niveau de précision très élevé, à partir de textes source et cible préexistants. Même s'il reste encore beaucoup à faire pour améliorer la rapidité et la précision, il est déjà possible de réaliser des alignements de textes relativement satisfaisants au niveau de la phrase. Il est également possible de calculer les correspondances lexicales, même s'il faut, pour l'instant, accepter de faire des compromis au niveau de la couverture ou de la précision. À notre connaissance, personne ne s'est encore penché sur le problème des correspondances entre syntagmes ou entre mot et syntagme, pour ne rien dire des phénomènes tels que les constituants discontinus. Mais comme la recherche sur les textes parallèles fait l'objet d'une attention sans cesse croissante, nous pouvons nous attendre à des progrès rapides dans ces domaines.

Selon nous, ces progrès revêtent une importance considérable pour le développement futur des aides à la traduction. Plus précisément, nous croyons que ces progrès ouvrent la voie à au moins trois types d'outils entièrement nouveaux, orientés vers 1) la constitution d'une mémoire d'entreprise pour les traducteurs; 2) la critique de traductions et 3) la dictée en traduction.

4. Une mémoire d'entreprise pour les services de traduction

La plupart des traducteurs sont quotidiennement confrontés à des problèmes de traduction dont la résolution peut se révéler laborieuse. Les ressources dont ils disposent, comme les dictionnaires et les banques de terminologie, ne leur fournissent pas de réponses immédiates. Il serait souhaitable que les solutions qui ont déjà été trouvées pour des problèmes donnés soient conservées et mises à la disposition ultérieure des mêmes personnes ou de leurs collègues. Malheureusement, ce n'est pas du tout le cas actuellement. En général, les grands services de traduction ne sont même pas en mesure de garantir qu'un de leurs traducteurs n'est pas en train de refaire depuis le début une traduction qui a déjà été faite par un autre.

Si l'on considère l'énorme volume de traductions qui sont produites bon an mal an, il devient tout à fait évident que **la masse des traductions existantes contient infiniment plus de solutions à plus de problèmes** que tous les outils de référence existants. Malheureusement, les traducteurs profitent très peu de cette richesse virtuelle. Dans la plupart des cas, les traductions ne sont archivées que sous forme de copie papier. Les archives ordinolingues, dans les rares cas où elles existent, ne sont d'à peu près aucune utilité, faute d'outils capables d'aider le traducteur à en exploiter efficacement le potentiel.

Supposons par contre qu'un service de traduction verse systématiquement sa production dans une base de données bi-textuelles, dans laquelle, par définition, chaque segment d'un texte source serait relié à sa traduction dans l'autre langue. Ainsi, chaque segment qui constitue un problème de traduction serait relié à un segment solution. Le service de traduction disposerait par conséquent d'une mémoire centrale structurée et exploitable.

Il y aurait plusieurs façons d'exploiter cette mémoire centrale. Visant une perspective à long terme, certains chercheurs s'intéressent à la possibilité d'utiliser les bases de données bi-textuelles comme fondements de systèmes de traduction automatique à base mémorielle, à base d'analogie ou à base d'exemples (voir notamment Sato et Nagao [25]).

Une approche moins ambitieuse vise le développement de systèmes capables d'extraire, durant le processus de traduction manuelle, des exemples pertinents d'une base de données, laissant au traducteur le choix de les utiliser ou non. Certains systèmes commerciaux existants, comme ALPS TSS et UNITRAN, incorporent déjà des éléments de cette approche.[3]

Enfin, une solution encore moins ambitieuse mais probablement plus universellement utile consiste à fournir au traducteur le moyen d'interroger sélectivement et sur demande une base de données bi-textuelles. Plusieurs auteurs ont déjà suggéré qu'un outil capable de produire des concordances bilingues serait utile aux lexicographes bilingues (voir Kavans et Tzoukermann [20]; Catizone, Russell et Warwick [5]; Church [7]). Il est très évident qu'un concondancier bilingue serait également utile au traducteur. Il se peut, par exemple, qu'en rencontrant, dans un texte de départ anglais, une expression idiomatique comme *to be out to lunch* ou *to add insult to injury*, le traducteur ait quelque hésitation à choisir l'équivalent français approprié. Il se peut aussi qu'il ne trouve pas de réponse satisfaisante dans les dictionnaires bilingues traditionnels. S'il dispose d'un concordancier bilingue, il pourra alors interroger une base de données bi-textuelles et en extraire des exemples de ces expressions, accompagnées de leurs traductions. Voir Macklovitch [24] pour un exposé plus détaillé de cette question.

On trouvera, à l'Annexe A, un exemple des résultats affichés à l'écran par OCTA, le prototype de concordancier bilingue mis au point par le CITI, lors d'une recherche de segments anglais contenant la séquence discontinue *insult... injury*, dans une base de données renfermant des alignements de phrases tirés des données du *Journal des débats* de 1986. Après avoir examiné quelques exemples comme celui-là, les traducteurs en concluent généralement que les concordanciers bilingues leur seraient très utiles.

---

[3] Ces systèmes ne sont pas basés sur la production automatique de représentations bi-textuelles. Ils utilisent plutôt des logiciels de traitement de texte spécialisés qui permettent au traducteur d'entrer, à volonté, le segment qu'il ou elle est en train de traduire. Ces renseignements sont ensuite versés dans une base de données de paires de segments qu'on peut ensuite interroger pour repérer des segments source (quasi) identiques.

5. Les outils de critique de traductions

Bien entendu, les traductions ne sont pas toutes parfaites et la mémoire d'entreprise doit être utilisée avec discernement. Il est cependant intéressant de noter que l'approche bi-textuelle pourrait très bien mener au développement d'outils capables d'aider les traducteurs à améliorer la qualité de leur travail.

Depuis quelques années, on voit apparaître sur le marché du logiciel des outils critiques qui aident les rédacteurs à améliorer leurs textes en détectant les fautes d'orthographe, de grammaire et même de style. Certains traducteurs trouvent ces outils utiles, même s'ils ne sont pas expressément conçus pour la révision des traductions comme telles. Ces outils ne sachant examiner qu'un texte à la fois, ils n'ont aucun moyen de détecter les erreurs de correspondance entre deux textes. La détection des erreurs de correspondance ne peut s'effectuer que dans la mesure où ces correspondances ont été rendues explicites, c'est-à-dire dans la mesure où l'on dispose de représentations bi-textuelles.

Considérée dans son ensemble, l'évaluation de la qualité des traductions constitue un problème extrêmement complexe et frustrant. Il est malgré tout possible de progresser dans la bonne direction en isolant certaines des propriétés spécifiques auxquelles les correspondances traductionnelles sont censées satisfaire, et en tentant de caractériser précisément (quoique peut-être partiellement) ces propriétés.

L'une de ces propriétés, élémentaire, est que la correspondance entre un texte source et sa traduction doit être exhaustive : aucun des éléments du texte source ne devrait être omis dans la traduction. Et pourtant, il n'est pas rare que les traductions humaines pèchent par omission, car le traducteur aura oublié de traduire des phrases, des paragraphes, voire des pages entières. Il est donc à espérer que les algorithmes d'alignement auront bientôt la robustesse nécessaire pour repérer les omissions.

Les traductions sont également censées être exemptes de toute «interférence» de la langue de départ (LD). Lorsqu'un traducteur travaille sur une paire de langues étroitement apparentées (comme c'est le cas de l'anglais et du français),

les problèmes d'interférence peuvent devenir très prononcés. Prenons, par exemple, le problème des *faux-amis*. Il s'agit de mots qui, en dépit de liens étymologiques évidents, ont cessé d'être des équivalents sémantiques. Dans le cas des faux-amis «absolus», les significations sont totalement disjointes et la correspondance directe est toujours illicite (ex. : *definitely/définitivement, actual/actuel, ignore/ignorer*). Dans le cas des faux-amis «partiels», il existe un certain chevauchement sémantique, de sorte que la correspondance demeure possible dans certains contextes seulement (ex. : camera/caméra).

Une étude approfondie de ce problème, réalisée au CITI, a révélé que les faux-amis sont la source d'un grand nombre d'erreurs dans les traductions du *Journal des débats*. Le concordancier bilingue que nous avons mentionné précédemment nous a aidés à en documenter des centaines d'exemples, y compris certaines erreurs plutôt grossières (ex. : *library/librairie, physician/physicien*). Même si ces traductions sont réalisées par certains des meilleurs traducteurs au Canada, il semblerait que les délais de production très restreints qui leur sont imposés rendent les «interférences» linguistiques très difficiles à maîtriser. Des outils capables de détecter ce genre d'erreurs pourraient, par conséquent, se révéler extrêmement utiles.

Par ailleurs, la notion de faux-amis n'est pas parfaitement définie. Il existe certains ouvrages de référence qui sont utiles (voir, par exemple, Van Roey, Granger et Swallow [27]), mais l'on peut douter de leur exhaustivité. De plus, dans le cas des faux-amis «partiels», la gamme des correspondances illicites est souvent floue et soumise aux variantes dialectales.

Il serait donc logique de commencer par les cas les plus évidents, et les représentations bi-textuelles devraient faciliter le repérage des correspondances mettant en jeu une paire de faux-amis absolus. Le niveau de «bruit» affectant la détection des correspondances erronées sera fonction du niveau de résolution du bi-texte. Si les correspondances sont établies au niveau du mot, le niveau de bruit sera très faible. Nos analyses préliminaires tendent à indiquer que, même avec une résolution moindre, c'est-à-dire au niveau de la phrase, le niveau de bruit pourrait demeurer dans des limites acceptables.

Le problème des faux-amis partiels est bien sûr plus difficile. Il semble cependant raisonnable de croire que les méthodes probabilistes de désambiguïsation sémantique, comme celles qui sont proposées par Gale, Church et Yarowski [14], puissent constituer un mécanisme de détection approprié, du moins pour un sous-ensemble de ces faux-amis.

Dans tous les cas, l'utilisation d'un programme de catégorisation grammaticale, comme celui décrit par Church [6], améliorerait probablement la précision, du fait que la catégorisation correcte de certains faux-amis (absolus ou partiels) doit reposer sur des informations concernant la catégorie grammaticale.

Nous croyons en outre que les représentations bi-textuelles permettraient de vérifier de nombreuses autres propriétés des correspondances traductionnelles.

6. Un système de dictée pour traducteur

Une des lacunes des postes de travail de traducteur existants réside dans le fait que de nombreux traducteurs préfèrent dicter leurs traductions, plutôt que de les entrer au moyen d'un clavier. Pour ces traducteurs, un poste de travail complet et pleinement intégré devrait donc comprendre un dispositif de reconnaissance de la parole.

Malheureusement, les systèmes de reconnaissance de la parole n'ont pas encore atteint le degré de perfectionnement qui en ferait des outils de travail pratiques aux yeux de nombreux traducteurs. La parole ne peut être décodée de façon plus ou moins fiable que si nous appliquons certaines contraintes relativement strictes au contenu du signal sonore. Les systèmes existants utilisent généralement une ou plusieurs des contraintes suivantes : vocabulaire et syntaxe restreints, parole discontinue (par opposition à parole enchaînée), entraînement du système en fonction d'un utilisateur particulier, faible niveau de bruits de fond, etc.

En 1989, nous avons réalisé une expérience de traduction de parole-à-parole (Cochard, Isabelle et Simard [8]), qui nous avait convaincus que la technologie

actuelle ne se prêtait pas encore aux applications pratiques, car il fallait contraindre l'entrée vocale de façon beaucoup trop artificielle. Mon collègue Marc Dymetman a fait, par la suite, une observation capitale, en posant que la situation pourrait fort bien être différente si la parole était entrée dans la langue cible par le traducteur. Il existerait, dans ce cas, une source naturelle de contraintes imposées au signal acoustique : **on sait, *a priori*, que le signal encode un texte qui est la traduction d'un texte source donné.**

Il serait donc possible, en principe, de concevoir un système de reconnaissance de la parole orienté spécifiquement vers la traduction. Un tel système utiliserait un modèle de traduction (complet ou partiel) qui permettrait de partir du texte source pour prédire certaines des caractéristiques de la traduction parlée. Par exemple, à partir de la présence du mot *government* dans une phrase de départ anglaise, le modèle de traduction pourrait prédire que la phrase correspondante en français renfermera probablement une réalisation parlée du mot *gouvernement*. Il est évident qu'un tel mécanisme faciliterait considérablement la tâche de reconnaissance de la parole.

Brown et al. [4] sont arrivés, indépendamment, à la même conclusion. Ils décrivent une expérience dans laquelle ils ont comparé la perplexité par mot d'un modèle langue-cible non assisté avec la perplexité par mot du même modèle combiné avec un modèle de traduction. Selon eux, la perplexité est de 63,61 dans le premier cas, comparativement à 17,2 dans le second. Leur conclusion :

> Il est raisonnable d'espérer, à la lumière de ces résultats, que la technologie actuelle permette la reconnaissance précise de la parole continue, lorsqu'on applique la contrainte suivante : le texte parlé est la traduction d'une séquence connue en langue source. (p. 10) [traduction]

Le CITI réalise actuellement un projet qui vise à explorer certains aspects de cette très intéressante possibilité (Dymetman et al. [11]), projet qui est étroitement relié à nos recherches sur la bi-textualité, en ce qu'il englobe l'élaboration d'un modèle de traduction probabiliste dont les paramètres sont extraits d'une vaste base de données bi-textuelles.

7. Conclusions

Dans l'état actuel des connaissances et de la technologie, les systèmes de TA réellement efficaces sont encore l'exception, plutôt que la règle, et les outils visant à soutenir le travail des traducteurs humains constituent généralement une utilisation plus judicieuse de la technologie. Les postes de travail de traducteur existants offrent principalement des fonctions de bureautique, et l'on n'a pas encore abordé directement les aspects essentiels de l'activité traduisante. Nous soutenons que la notion de bi-texte, c'est-à-dire la représentation explicite des correspondances traductionnelles entre un texte source et sa traduction, est extrêmement pertinente à la réalisation de cet objectif.

L'élaboration d'algorithmes capables de structurer automatiquement un texte source et sa traduction en une représentation bi-textuelle progresse à grand pas. Les résultats déjà obtenus ouvrent la voie à la mise au point d'une panoplie d'outils entièrement nouveaux à l'intention du traducteur. Parmi ceux-ci, mentionnons les concordanciers bilingues, qui permettront au traducteur d'exploiter les ressources d'une mémoire centrale constituée de représentations bi-textuelles de traductions existantes. Un peu plus loin à l'horizon, nous pouvons envisager l'apparition d'outils de critique de traductions qui aideront les traducteurs à détecter, dans leurs traductions, les erreurs de correspondance telles que les omissions et les faux-amis. Enfin, nous avons des raisons de croire que, dans un avenir relativement rapproché, les postes de travail de traducteur en viendront àincorporer des systèmes de reconnaissance de la parole spécialement conçus, qui incorporeront certaines connaissances traductionnelles tirées de bases de données bi-textuelles.

Nous en concluons qu'à court terme les outils de soutien destinés au traducteur sont susceptibles de progresser beaucoup plus rapidement et d'avoir une incidence beaucoup plus importante sur les langagiers que les systèmes de traduction automatique traditionnels.

Références bibliographiques

[1] Bar-Hillel, Y. : *The State of Machine Translation in 1951*, **American Documentation**, vol. 2, 1951, p. 229-237.

[2] Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L. et Paul S. Roossin : «A statistical approach to machine translation», *Computational Linguistics*, vol. 16, n° 2, juin 1990, p. 79-85.

[3] Brown, Peter F., Lai, Jennifer C. et Robert L. Mercer : «Aligning sentences in parallel corpora», *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley (Californie), juin 1991.

[4] Brown, Peter F., Chen, Stanley F., Della Pietra, Stephen A., Della Pietra, Vincent J., Kehler, Andrew S. et Robert L. Mercer : «Automatic speech recognition in machine aided translation», manuscrit inédit, IBM T.J. Watson Research Center, 1992.

[5] Catizone, R., Russell, G. et S. Warkwick : «Deriving Translation Data from Bilingual Texts», Zernick (éd.), *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum, 1992.

[6] Church, K. : «A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text», *Proceedings of the 2nd ACL Conference on Applied Natural Processing*, Austin, 1988.

[7] 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.

[8] Cochard, J.L., Isabelle, P. et M. Simard : «IRMA: an Agricultural Market Report Interpreter», rapport technique n° Co28-1/49-1990E, CCRIT, Laval, 1990.

[9] Debili, F. et E. Sammouda : «Appariement des phrases de textes bilingues français-anglais et anglais-arabes», *Proceedings of COLING-92*, Nantes, 1992.

[10] Dempster, A., Laird, N. et D. Rubin : «Maximum Likelihood from Incomplete Data via the EM Algorithm», *Journal of the Royal Statistical Society*, 39(B), 1977.

[11] Dymetman, M., Foster, G. et P. Isabelle : «Towards Automatic Dictation Systems for the Professional Translator», mémoire technique, CCRIT, Laval, 1992.

[12] Gale, W. et K. Church : «A program for aligning sentences in bilingual corpora», *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley (Californie), 1991.

[13] Gale, W. et K. Church : «Identifying Word Correspondences in Parallel Texts, *Proceedings of the DARPA SLS Workshop*, 1991.

[14] Gale, W., Church, K. et D. Yarowski : «Using Bilingual Materials to Develop Word Sense Disambiguation Methods», *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, 1992.

[15] Harris, B. : «Are you Bi-Textual?», *Language Technology*, n° 7, p. 41, 1988.

[16] Harris, B. : «Bi-text, a New Concept in Translation Theory», *Language Monthly*, n° 54, p. 8-10, 1988.

[17] Isabelle, P. : «Machine Translation at the TAUM Group», Margaret King (éd.), *Machine Translation Today: The State of the Art*, Edinburgh University Press, 1987.

[18] Kay, M. : «The Proper Place of Men and Machines in Translation», CLS-80-11, Xerox PARC, 1980.

[19] Kay, M. et M. Röscheisen : «Text-Translation Alignment», manuscrit inédit, Xerox PARC, 1988.

[20] Klavans, J. et E. Tzoukermann : «Linking Bilingual Corpora and Machine Readable Dictionaries with the BICORD System», *Proceedings of the 6th Annual Conference of the UW Centre for the NOED and Text Research*, University of Waterloo, 1990.

[21] Macklovitch, E. : «An Off-the-shelf Workstation for Translators», D. Hammond (éd.), *Proceedings of the 30th Annual Conference of the ATA*, Washington DC, 11-14 octobre 1989, p. 4891-4898.

[22] Macklovitch, E. : «A Second Version of the CWARC's Workstation for Translators», rapport technique, CCRIT, 1991.

[23] Macklovitch, E. : «Evaluating Commercial MT Systems», communication présentée au Evaluator's Forum on MT Systems, Ste-Croix, Suisse, 1991.

[24] Macklovitch, E. : «Corpus-Based Tools for Translators», à paraître dans *Proceedings of the ? Conference of the ATA*, San Diego, 1992.

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

[26] Simard, M., Foster, G. 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, 1992.

[27] Van Roey, J., Granger, S. et H. Swallow : *Dictionnaire des faux-amis français-anglais*, Paris, Duculot, 1988.

Annexe A

Recopie d'écran du système OCTA, le concordancier bilingue mis au point par le CITI

OCTA: Query

>> insult...injury:

OCTA: Properties

Default database:
path: /TAO/tao_stuff/octa
file: odb-hans-86

Interface language: English | French

Query:
Search mode: ☑ accent sensitive ☐ case sensitive

OCTA alpha: odb-hans-86

File ▽ View ▽ Find ▽ Props... 1 1 8

To add **insult to injury** we in Nova Scotia -- a have-not province -- are being discriminated against by the Government .	Pour ajouter l'insulte à l'injure , nous , en Nouvelle-Écosse -- une province défavorisée -- sommes objet de discrimination de la part du gouvernement .
Why have the Minister and the Prime Minister added **insult to injury** by breaking the promise of a 3 per cent real increase in Canada's defence budget to which they committed themselves ?	Comble de l'injure , pourquoi le ministre et le premier ministre trahissent-ils l'engagement qu'ils avaient pris de décréter une hausse réelle de 3 p. 100 du budget de la défense ?
It has added **insult to injury** as far as the offshore is concerned .	En ce qui concerne la prospection au large il a été doublement injuste .
It is shameful to hear this kind of nonsense coming from government Members who have described Katimavik in those terms and have added **insult to injury** by describing it as a pork-barrelling employment project .	Il est honteux d'entendre ces absurdités de la part des députés ministériels qui ont décrit Katimavik en ces termes et non contents de cela , l'ont qualifié d'assiette au beurre .
The Government has now added **insult to injury** .	Le gouvernement vient de doubler ses torts d'un affront .
That is how the Government has added **insult to injury** .	Voilà comment le gouvernement tourne le fer dans la plaie .
To add **insult to injury** the Government is allowing British Telecommunications to lift the research and development of the Mitel empire out of Canada in spite of the fact that Canadian taxpayers have a major investment in that research and development .	Par dessus le marché , le gouvernement permet à British Telecommunications de transférer la recherche et le développement de l'empire Mitel hors du Canada en dépit du fait que les contribuables canadiens ont beaucoup investi dans cette activité .
Mr. Bill Blaikie (Winnipeg -- Birds Hill) : Mr. Speaker , **insult has been added to injury** in Manitoba .	M. Bill Blaikie (Winnipeg -- Birds Hill) : Monsieur le Président , on a porté l'insulte à son comble au Manitoba .

Ready

CACC / CCAC
35699

P308 .I37f 1993 c.2 JOUR

QUEEN P 308 .I3714 1993 c.2
Isabelle, Pierre
Les aides bi-textuelles à l

DATE DUE

PRINTED IN U.S.A.

1575, boulevard Chomedey, Laval (Québec), H7V 2X2 Téléphone: (514) 973-5700 Télécopieur: (514) 973-5757

1575, Chomedey Boulevard, Laval, Quebec, H7V 2X2 Telephone: (514) 973-5700 Facsimile: (514) 973-5757