L'utilisation de la bitextualité pour la validation des traductions : le système *TransCheck* (1)

Cet article décrit la première version prototype de *TransCheck*, système qui permet de détecter automatiquement certains types d'erreurs de traduction et qui est basé sur la notion de bitexte, ou corpus composé de textes sources, alignés avec leur traduction.

Nous y présentons également l'analyse des résultats de traductions publiées, la description de certains des problèmes qui échappent actuellement au système, ainsi que les grandes lignes des projets de développement du premier prototype de *TransCheck*.

Termes-clés : traduction; détection d'erreur; alignement de corpus; bitexte; français; anglais.

(1) Article déjà paru en 1995 dans *L'Actualité terminologique*, vol. 28, n°3, publication du Bureau de la traduction de Travaux publics et services gouvernementaux Canada, et reproduit avec la permission des éditeurs.

(2) À leur décharge, il faut dire que ces traducteurs travaillent dans un climat de grande tension, car ils doivent traduire les débats parlementaires du jour même pour qu'ils puissent être publiés le lendemain matin. Dans ces conditions, et compte tenu de la quantité de textes que nous avons examinés, ce qui est réellement étonnant, c'est la rareté des erreurs relevées dans le *Journal des débats*.

(3) À part les exemples (xv) et (xvi). Voir les notes 10 et 14 plus loin.

(4) Pour plus de détails sur l'algorithme d'alignement utilisé actuellement au Citi, et qui présente un taux de succès de plus de 98%, voir Simard *et al.* (1992).

1 Détecteur d'erreurs

Comme tous les gens qui rédigent beaucoup, les traducteurs passent généralement leurs traductions au correcteur orthographique ou stylistique avant de les livrer, afin de déceler les erreurs simples qui auraient pu leur échapper. Étant unilingues, ces outils rédactionnels sont cependant tout à fait incapables de détecter les erreurs de traduction même les plus flagrantes, car ces erreurs sont bilingues de nature et dépendent des relations qui existent entre deux textes de langues différentes. Cet article décrit un nouveau genre d'outil rédactionnel expressément conçu pour détecter les erreurs comme celles qui sont présentées au tableau n°1, que l'on trouvera à la fin de l'article.

Tous les exemples énumérés au tableau n°1 sont authentiques : ils sont tirés du *Journal des débats* de la Chambre des communes du Canada, de sorte que les erreurs qu'ils contiennent ont en réalité été commises par des traducteurs qui sont parmi les plus compétents au Canada (2).

Les erreurs énumérées au tableau n°1 ont été détectées automatiquement par le système de vérification de traduction appelé *TransCheck* (3), qui fait partie d'une nouvelle famille d'aides à la traduction mises au point au Centre d'innovation en technologies de l'information (Citi) (Isabelle *et al.* 1993). Ces outils sont tous basés sur des algorithmes d'alignement bitextuel comme ceux qui ont été proposés initialement par Gale et Church (1991), et par Brown *et al.* (1991). Que font ces algorithmes ? En deux mots, ils calculent avec précision les liens entre les segments correspondants d'un texte source et de sa traduction, sans égard à la longueur des textes (4). Suivant la terminologie utilisée par Harris (1988), un corpus de textes sources/cibles, alignés de cette façon, est souvent appelé un bitexte. Dans les travaux que nous venons de citer, les segments alignés d'un bitexte sont généralement des phrases, mais il est également possible, en principe, d'établir des liens entre des unités plus longues ou plus courtes. Bien sûr, une phrase du texte source peut être explicitement reliée à deux ou plusieurs phrases du texte cible, et vice-versa.

Dans plusieurs publications récentes du Citi – notamment Isabelle *et al.* (1993), Isabelle (1992) et Macklovitch (1992) –, on a avancé que le concept de bitextualité permet le développement d'une toute nouvelle génération d'aides à la traduction, dont la plus évidente est peut-être le concordancier bilingue,

comme le système *TransSearch* mis au point au Citi. Un programme de concordance standard est une sorte de système d'interrogation de base de données qui recherche, extrait et affiche dans leur contexte toutes les occurrences d'un terme ou d'une expression contenues dans une base de données textuelles. Ce qu'un concordancier bilingue comme *TransSearch* fait, en outre, c'est afficher, en regard de chacune de ces occurrences, la ou les phrases qui en constituent la traduction, selon les calculs préalables des programmes d'alignement. Ce simple (du point de vue conceptuel) couplage de segments en langue de départ et en langue d'arrivée présente cependant d'énormes avantages : il permet aux traducteurs d'accéder facilement, au moyen de l'élégante interface graphique de *TransSearch*, à toutes les richesses enfouies dans leur production antérieure. Et, comme l'a justement souligné Pierre Isabelle (1992), «la masse des traductions produites chaque année contient infiniment plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables».

Les requêtes que l'utilisateur peut soumettre à *TransSearch* peuvent porter sur des mots simples ou des expressions complexes, et elles peuvent être unilingues ou bilingues. Par exemple, l'utilisateur peut demander au système d'afficher toutes les phrases de sa base de données qui renferment le terme *gridlock*, ou toutes les paires alignées dans lesquelles la phrase anglaise contient le terme *librairie*. Dans ce cas, les occurrences repérées par le système constitueraient des erreurs de traduction, car *library* et *librairie* sont de faux amis : bien qu'ils soient étymologiquement apparentés et encore semblables morphologiquement, ces deux termes ne peuvent jamais être utilisés comme des équivalents traductionnels, car ils ont maintenant des sens complètement différents (*librairie* correspond au terme anglais *bookstore*).

Supposons qu'on compile une liste exhaustive de ces faux amis. Supposons en outre qu'on puisse aligner automatiquement une ébauche de traduction avec son texte source et qu'on dispose aussi d'une procédure automatique permettant d'appliquer cette liste de faux amis au bitexte ainsi obtenu, en tant que requêtes en lot soumises à *TransSearch*. Supposons enfin qu'on dispose d'une interface nous permettant d'examiner et d'éditer les résultats de ces requêtes. Ce qu'on obtiendrait alors, en fait, c'est un simple vérificateur de traduction. Simple, en ce sens qu'il pourrait vérifier un premier jet de traduction pour y détecter un type évident d'erreurs de traduction, c'est-à-dire les faux amis. Et en fait, c'est exactement de cette façon que *TransCheck* a d'abord été conçu et développé au Citi – comme une extension de notre concordancier bilingue.

2 Erreurs de traduction

Les faux amis complets comme *library/librairie* constituent un bon point de départ pour le développement d'un vérificateur de traduction car, par définition, les mots qui constituent ces paires ne peuvent jamais être des traductions réciproques. Mais ce n'est pas là le seul critère d'inclusion dans un vérificateur de traduction; si ce l'était, le lexique bilingue du programme (ou anti-lexique, comme il serait plus juste d'appeler une base de données de couples illicites) contiendrait un nombre énorme d'entrées. Il est évident que le nombre de mots qui ne peuvent jamais être des équivalents traductionnels réciproques est infiniment plus grand que le nombre de paires de mots qui peuvent l'être. Pour n'en donner qu'un exemple farfelu, nous sommes tout à fait certains que le mot anglais *very* ne peut jamais être traduit en français par le mot *courgette* et, pourtant, nous ne voudrions pas inclure cette paire dans l'anti-lexique de *TransCheck*. Ce qui distingue la paire *library/librairie* de la paire *very/courgette*, c'est qu'il est beaucoup plus probable que la première apparaisse dans une traduction de l'anglais au français, précisément parce que la similitude morphologique des deux mots peut aisément amener le traducteur à oublier leur dissimilitude sémantique. D'une façon plus générale, les erreurs que nous voulons que notre programme détecte sont précisément les erreurs courantes les plus susceptibles de se glisser dans des traductions ou qui, selon les réviseurs, sont effectivement souvent commises par les traducteurs.

Pour nous aider à compiler un inventaire de ce genre d'erreurs pour le premier prototype de *TransCheck*, nous avons consulté plusieurs ouvrages de référence bien connus traitant des difficultés de traduction de l'anglais au français (Copron 1982, Dagenais 1984, De Villers 1988, Rey 1984, Van Roey *et al.* 1988). Nous avons extrait les descriptions de plus de 2 800 difficultés de traduction, que nous avons transcrites et stockées dans de simples enregistrements de base de données comprenant les champs suivants :

– Français incorrect;
– Anglais (correspondant);
– Français correct;
– Type d'erreur;
– Source;
– Commentaires.

Ces descriptions statiques ont ensuite été converties par un autre programme en requêtes *TransSearch*, sous la forme générale suivante :

(i) {e(mot+/cat}=/=f(mot+/cat)

où «e(mot)» désigne un ou plusieurs mots anglais, «f(mot)» désigne un ou

plusieurs mots français, «cat» désigne une catégorie morphosyntaxique et le symbole facultatif «+» correspond à la forme de base du mot, plus toutes ses variantes flexionnelles. Les accolades ({}) indiquent le caractère facultatif du segment en langue source [5].

Lorsque les e(mot) et f(mot) sont des éléments uniques et légitimes de leur vocabulaire respectif et qu'ils sont étymologiquement apparentés, la requête correspond à une recherche de faux amis. Mais, bien sûr, il n'est pas toujours nécessaire que ces mots soient étymologiquement apparentés; dans les ouvrages de référence que nous avons consultés, nous avons trouvé de nombreux exemples de traductions illicites entre des paires qui n'étaient pas morphologiquement similaires ou dérivées d'une racine commune. Pour reprendre la terminologie utilisée par une de ces autorités, ces paires sont appelées des *traductions impropres* et elles ont fait l'objet, dans notre base de données, de la même description formelle que les faux amis. Le mot *cabaret*, par exemple, est tout à fait correct en français, mais ce n'est pas une traduction acceptable du mot anglais *tray*; le mot *cabaret* désigne un type de bar et la traduction française correcte de *tray* est *plateau*. Ainsi, si notre vérificateur de traduction repérait deux segments alignés dont l'anglais contiendrait le mot *tray* ou *trays* et le français contiendrait le mot *cabaret*, il aurait selon toute vraisemblance détecté une erreur de traduction, en supposant, bien sûr, que le mot *cabaret* ne soit pas la traduction d'un autre mot de la phrase anglaise. (Nous reviendrons plus loin sur cette question.) On trouvera d'autres exemples de traductions impropres repérées dans le *Journal des débats* aux entrées (v-vii) du tableau n°1, à la fin de l'article.

En dépouillant les ouvrages de référence traitant des difficultés de traduction *(op.cit.)*, nous avons relevé d'autres types de difficultés de traduction que nous avons décidé d'inclure dans le prototype de *TransCheck*. Lorsque deux langues sont en contact étroit, comme c'est le cas de l'anglais et du français au Canada, les emprunts illicites constituent une source très fréquente d'erreurs de traduction. Pour les fins de nos travaux, un emprunt est simplement défini comme un mot anglais apparaissant dans un texte français, et vice-versa. En réalité, les choses ne sont pas si simples : toutes les langues font constamment des emprunts et, dans certains cas, les autorités en la matière ont des opinions divergentes quant au statut de naturalisation des mots étrangers. Dans la grande majorité des cas, cependant, il n'y a pas de controverse. Dans le cas de mots tels que *lunch*, *cool* ou le verbe *backer*, le vérificateur de traduction n'a qu'à détecter leur présence dans un texte français pour signaler un emprunt. Il n'est pas essentiel ici de déterminer leur catégorie lexicale, ni même de vérifier la présence de la forme correspondante dans le segment anglais aligné [6]. En fait, on peut même se demander s'il est nécessaire d'inclure, dans la base de données de *TransCheck*, des entrées décrivant de tels emprunts illicites; étant donné que ces emprunts ne font pas partie du lexique français, un correcteur orthographique unilingue devrait pouvoir les détecter en tant que mots inconnus. Cependant, ce que les entrées de la base de données de *TransCheck* peuvent faire de plus qu'un correcteur orthographique, c'est indiquer à l'utilisateur la forme correcte qu'il devrait utiliser en français au lieu de l'emprunt illicite. Découvrir, par exemple, que *briefing* n'est pas un mot correct en français est une chose, en connaître l'équivalent français correct en est une autre. (On trouvera, au tableau n°1 (xi-xiv), d'autres exemples d'emprunts illicites.)

Les calques constituent un autre type courant d'interférence linguistique. Un calque est la traduction littérale, c'est-à-dire mot à mot, d'une expression complexe en langue source, et dont le résultat est inacceptable en langue cible. Dans certains cas, l'expression calquée peut être parfaitement correcte sur le plan syntaxique, et transparente sur le plan sémantique, mais elle ne correspond tout simplement pas au terme accepté ou normalisé en langue cible. Ainsi, *certificat de naissance* est composé de mots français corrects et combinés conformément aux règles de formation des syntagmes nominaux français; le seul problème, c'est que le français utilise un autre terme pour désigner le *birth certificate* anglais, soit *extrait d'acte de naissance*. Dans d'autres cas, l'expression calquée contrevient aux règles de la grammaire française; c'est le cas de l'expression *à la journée longue*. En français, l'adjectif temporel *long* ne suit pas habituellement le nom qu'il modifie et la façon correcte de

---

(5) En réalité, l'opérateur «=/=» n'apparaît pas dans les requêtes soumises à *TransSearch*. Nous l'avons inclus ici pour souligner le fait que les couples anglais-français ne sont pas des traductions réciproques possibles. Plusieurs des champs de base de données mentionnés plus tôt n'apparaissent pas non plus dans les requêtes formelles, mais sont fournis à titre de renseignements supplémentaires durant la séance d'édition avec *TransCheck*.

(6) Cela ne veut pas dire que la détection des emprunts soit toujours aussi simple. Parmi les complications auxquelles nous nous sommes heurtés, signalons la difficulté de générer les variantes morphologiques des formes non françaises comme *backer*, ainsi que le problème posé par les formes empruntées qui se trouvent coïncider avec un mot parfaitement légitime en langue d'arrivée, comme le mot *pin* qui est utilisé incorrectement en français pour désigner une broche, par opposition au mot français *pin*, qui désigne un type de conifère.

traduire l'expression anglaise *all day long* est *à longueur de journée.* Pour ces deux types de calques, comme dans le cas des emprunts mentionnés ci-dessus, *TransCheck* n'a pas nécessairement besoin de vérifier la présence de l'expression anglaise correspondante dans le segment aligné pour pouvoir détecter une erreur. En fait, en règle générale, lorsqu'un mot ou un syntagme douteux peut être clairement identifié sans recourir au segment aligné de l'autre texte, les requêtes unilingues sont probablement préférables, car on court toujours le risque, dans une relation aussi libre que la traduction, que l'expression spécifiée ne se trouve pas dans le texte source et, par conséquent, que le système omette une erreur faute d'avoir pu établir la correspondance. (On trouvera, au tableau n° 1 (viii-x), d'autres exemples de calques.)

3 Résultats préliminaires

Afin de tester le premier prototype de *TransCheck*, nous l'avons appliqué à cinq échantillons de 100 000 mots de texte traduit : trois de ces échantillons étaient tirés du *Journal des débats* de la Chambre des communes et les deux autres, de manuels opérationnels utilisés par le ministère de la Défense nationale (MDN) [7]. Évidemment, *TransCheck* est conçu pour vérifier des ébauches de traduction et, dans tous ces cas, les échantillons soumis à la validation étaient des traductions peaufinées et publiées. C'est donc dire qu'on ne peut pas accorder beaucoup d'importance au nombre absolu d'erreurs détectées, pas plus qu'on envisagerait d'évaluer l'efficacité d'un correcteur orthographique en fonction des résultats obtenus en l'appliquant à des textes unilingues publiés. De plus, comme il s'agissait d'un premier prototype que nous étions en train de développer, nous désirions l'utiliser pour évaluer un certain nombre d'hypothèses, par exemple, la possibilité de relever toutes les expressions multilexicales (syntagmes) en n'utilisant que des requêtes unilingues. L'adoption de telles hypothèses introduit aussi un élément de distorsion dans les données relatives à la performance de *TransCheck*.

Malgré tout, il était important pour nous d'avoir un aperçu du genre d'erreurs que *TransCheck* était capable de détecter, ainsi que du niveau de bruit [8] qu'il produisait, ne serait-ce que pour être en mesure d'évaluer la viabilité générale d'un vérificateur de traduction qui utilise un programme d'alignement de phrases pour détecter des correspondances lexicales illicites. Le tableau n° 2, à la fin de l'article, résume les résultats obtenus par *TransCheck* sur les cinq corpus d'essai.

Ce qu'il faut d'abord remarquer à propos de ces résultats, ce sont les taux de précision impressionnants obtenus par *TransCheck* sur les corpus tirés du *Journal des débats* et, dans une moindre mesure, sur le premier corpus tiré des manuels du MDN. Ce que ces résultats nous apprennent, en d'autres termes, c'est que la grande majorité des erreurs potentielles repérées dans ces corpus s'avèrent des erreurs véritables, en ce sens qu'elles correspondent exactement aux difficultés linguistiques décrites par les autorités que nous avons consultées. Nous exposerons plus loin les raisons des résultats inférieurs obtenus par le système avec les manuels du MDN. Ce qu'il faut souligner ici, cependant, c'est que, parmi tous les cas de bruit relevés dans les cinq corpus, il n'y en a qu'un ou deux qui sont attribuables à la grossièreté des alignements réalisés par le système, et aucun qui n'est attribuable à un mauvais alignement des phrases source et cible. Cela confirme les premiers résultats décrits dans Isabelle *et al.* (1993) et semble justifier l'affirmation qui y est faite qu'un vérificateur de traduction basé sur un programme d'alignement de phrases et un programme de catégorisation grammaticale pourrait bien constituer une plate-forme suffisante pour le développement d'applications pratiques.

Plusieurs autres aspects des résultats présentés au tableau n°2 méritent qu'on s'y attarde, notamment le nombre considérablement moindre d'erreurs repérées dans le corpus *Journal des débats-2*, comparativement au corpus *Journal des débats-1*. Cette baisse importante découle du fait que nous avions omis de spécifier l'anglais comme langue source dans le premier échantillon du *Journal des débats*, de telle sorte qu'un grand nombre des erreurs détectées dans ce corpus apparaissaient en réalité dans les discours des parlementaires francophones. Comme nous nous intéressions davantage aux erreurs des traducteurs français qu'à celles des députés francophones, nous avons réglé ce paramètre pour que l'anglais soit exigé comme langue source dans les quatre autres échantillons d'essai [9]. Le nombre total d'erreurs

---

(7) L'un de ces manuels traitait de l'utilisation des bataillons d'infanterie, tandis que l'autre portait sur la formation des tireurs d'élite.

(8) Le mot *bruit* est employé ici dans un sens analogue à celui qui est le sien en recherche documentaire pour désigner des résultats qui ne sont pas pertinents, dans ce cas-ci, parce qu'ils ne correspondent pas à des erreurs de traduction véritables.

(9) Bien qu'un vérificateur de traduction puisse, en principe, être bidirectionnel, *TransCheck* présente une nette orientation anglais-français, principalement attribuable au fait que les ouvrages de référence que nous avons consultés se concentrent presque exclusivement sur les problèmes d'usage du français.

détectées dans ces échantillons serait par conséquent beaucoup plus élevé si nous n'avions pas modifié le paramètre langue source.

Une autre constante qui se dégage de l'analyse des échantillons d'essai est la suivante : la plupart des erreurs détectées dans un texte donné sont attribuables à un nombre relativement restreint de difficultés linguistiques spécifiées dans la base de données de *TransCheck*. En d'autres termes, la plupart des requêtes soumises par le système ne trouvent aucune correspondance dans le texte traduit. Nous avons mentionné plus tôt que la base de données actuelle de *TransCheck* contient environ 2 800 entrées; dans chacun des quatre derniers échantillons d'essai, à peine 25 de ces entrées ont permis de détecter des erreurs (bien que les requêtes n'aient pas été nécessairement les mêmes pour tous les échantillons et qu'une même requête ait pu évidemment repérer plusieurs erreurs dans le même texte). Nous ne nous attendons pas à ce que cela change beaucoup, même en appliquant *TransCheck* à des ébauches de traduction plus grossières. Ce phénomène est attribuable à la nature de la base de données de *TransCheck*, qui vise à décrire le plus grand nombre possible d'erreurs de traduction attestées et les plus probables, quel que soit le domaine ou le type de texte. Il reste qu'une traduction donnée ne contiendra normalement qu'une fraction de ces erreurs; donc, il est tout à fait naturel que la plupart des requêtes traitées par *TransCheck* ne donnent aucun résultat, car même les premiers jets les plus grossiers contiennent beaucoup plus de segments rendus correctement que de segments contenant des erreurs.

Du relativement petit nombre d'entrées qui produisent effectivement des correspondances, il y en a quelques-unes qui tendent à réapparaître, parfois très fréquemment, dans chacun des échantillons d'essai. Dans le *Journal des débats*, par exemple, le terme *caucus* est repéré à de nombreuses reprises, ce qui est également le cas du couple *deception/déception* dans les corpus tirés des manuels du MDN. Ce ne sont donc pas des fautes d'inattention; leur répétition indique plutôt une sérieuse divergence d'opinion entre au moins l'une des autorités que nous avons consultées et les traducteurs qui produisent ces textes. D'ailleurs, en consultant d'autres références (comme *Termium* ®), nous pouvons constater que ces deux termes sont acceptés et considérés comme corrects. Ce genre de divergence entre autorités suscite un problème épineux pour le développement d'un vérificateur automatique de traduction, et l'attitude que nous avons adoptée peut se résumer comme suit : ce n'est pas à nous de trancher. Nous n'avons pas la prétention d'être des experts, et encore moins des prescripteurs, en matière d'usage correct du français, et nous ne prenons pas position vis-à-vis des erreurs répertoriées par les autorités que nous avons consultées. Pour chaque erreur qu'il détecte, le système fournit un renvoi à l'ouvrage dans lequel cette difficulté est décrite. Si les utilisateurs éventuels ne sont pas d'accord avec l'opinion de cette autorité, il faudra qu'ils puissent neutraliser l'entrée correspondante dans *TransCheck*, tout comme les gens qui utilisent des correcteurs stylistiques peuvent désactiver certaines règles ou certains ensembles complets de règles. Plus généralement, *TransCheck* est un outil qui contribuera à l'application des normes de traduction, normes au pluriel, car il n'existe sûrement pas de norme de traduction unique qui s'applique à tous les domaines et à tous les types de textes. Voilà pourquoi la constitution d'une base de données élémentaire sera certainement plus difficile pour les concepteurs d'un vérificateur de traduction qu'elle ne le sera, disons, pour les concepteurs d'un correcteur orthographique. Mais quelle que soit la norme (ou les normes) que les utilisateurs décideront d'adopter pour leur usage personnel ou pour celui de leurs divers clients, nos résultats préliminaires portent à conclure qu'un outil comme *TransCheck* les aidera néanmoins à détecter un nombre appréciable d'erreurs dans leurs textes.

Les niveaux de bruit plus élevés qu'on peut constater, au tableau n°2, dans les résultats obtenus par *TransCheck* sur les deux échantillons tirés des manuels du MDN, sont principalement attribuables à un petit nombre d'erreurs de description dans la base de données du système, dont certaines renvoient à des termes qui reviennent parfois très fréquemment dans les manuels militaires. Notre analyse de ces résultats révèle qu'il s'agit, dans la plupart des cas, de règles de traduction dépendant du contexte, qui ont été incorrectement caractérisées, soit par nous, soit par une de nos autorités, comme des interdictions traductionnelles absolues [(10)]. Nous avons développé *TransCheck* en nous servant du *Journal des débats* comme corpus de référence et il se peut que nous ayons involontairement adapté, jusqu'à un certain point, les entrées lexicales du système au langage parlementaire. Il n'est donc pas tout à fait surprenant

(10) Selon une de nos autorités, par exemple, le nom français *raid* n'est pas une traduction correcte de l'anglais *raid*, quand il s'agit d'une intervention policière, bien qu'il le soit, en l'occurrence, quand il s'agit d'une intervention militaire. Quant à nous, nous avons entré, à tort, le verbe *graduer* comme faux ami complet, ignorant son emploi correct au sens de «calibrer», soit la façon dont ce verbe est employé dans le manuel de formation du MDN à l'intention des tireurs d'élite.

que, lorsque le système est appliqué à des textes relevant d'un domaine très différent, certains des ajustements qui entrent inévitablement dans l'élaboration d'un dictionnaire automatisé ne donnent pas les résultats attendus. L'autre principale source de bruit dans les résultats obtenus des échantillons du MDN est attribuable à certaines lacunes des modules de découpage et d'analyse lexicale de *TransCheck* : par exemple, l'abréviation *s/off* (sous-officier) est incorrectement découpée en deux unités et l'unité *off* est identifiée comme un emprunt. Nous devrions être en mesure de corriger ces problèmes sans trop de difficulté.

Voici maintenant quelques mots sur la forme actuelle des sorties de *TransCheck*. Pour ce premier prototype de démonstration, nous avons utilisé comme interface un éditeur *Emacs* standard (voir les deux recopies d'écran présentées à la fin du texte). Bien qu'elle ne soit pas aussi élégante que l'interface graphique utilisée pour *TransSearch*, l'interface *Emacs* offre beaucoup de souplesse aux programmeurs et aux linguistes qui travaillent au développement de *TransCheck*. La fenêtre supérieure de chaque écran affiche le bitexte aligné côte à côte, le segment visé par la requête étant affiché en gras et l'erreur potentielle détectée par le système étant soulignée. La fenêtre inférieure affiche l'entrée correspondante dans la base de données, la requête compilée et le résultat de l'analyse grammaticale des segments source et cible, qui se résume actuellement à une analyse de la catégorie grammaticale. Ainsi, dans la première recopie d'écran, tel type d'erreur détectée est un calque : selon le *Multidictionnaire des difficultés de la langue française*, l'expression *en devoir* n'est pas une traduction acceptable de l'anglais *on duty*; le syntagme prépositionnel qui devrait être utilisé est *de service* ou *en service*. Dans la deuxième recopie d'écran, l'entrée de la base de données nous apprend que le verbe *nominer* est un emprunt illicite en français; la forme correcte est *mis en nomination*.

4 Extension du prototype actuel

Les données présentées au tableau n°2 sont révélatrices de la qualité des résultats produits par *TransCheck*, pour ce qui est du rapport bruit/erreurs véritables détectées. Malheureusement, nous ne disposons pas actuellement de données sur le «silence» du système, c'est-à-dire sur la proportion d'erreurs véritables qui échappent à *TransCheck*. Pour obtenir de telles données, il nous faudrait appliquer le système à une traduction-étalon, dont toutes les erreurs auraient été préalablement décelées par un expert; cet exercice nous permettrait de calculer le nombre d'erreurs détectées et le nombre d'erreurs omises par le système. Néanmoins, d'après les mois de travaux préliminaires consacrés au développement du premier prototype et d'après l'analyse des premiers résultats décrits ci-dessus, nous savons que la version actuelle du système détecte à tort certains segments qui ne contiennent pas d'erreurs et, surtout, qu'il existe de nombreux types de difficultés de traduction qui ne peuvent être saisis par son modèle descriptif rudimentaire.

Comme nous l'avons déjà indiqué, nous savons que l'alignement de deux textes au niveau de la phrase n'est pas toujours adéquat; à l'occasion, les phrases alignées contiennent des paires de mots ou d'expressions interdites, mais qui ne sont pas des traductions réciproques [(11)]. La solution à ce problème est évidente : nous devons obtenir des alignements bitextuels plus subtils, qui s'établissent non plus au niveau de la phrase, mais au niveau du syntagme et du mot. Des progrès ont été réalisés dans ce sens, au Citi et ailleurs (voir, en particulier, Dagan *et al.* 1993).

Ensuite, il faut considérer les nombreuses difficultés de traduction dont la description exige plus d'information contextuelle que la catégorisation grammaticale élémentaire actuellement réalisée par *TransCheck*. Prenons, par exemple, le verbe français *débuter* : on peut l'utiliser pour traduire le verbe anglais *to start*, mais seulement lorsqu'il est intransitif. Pour que *TransCheck* arrive à ne détecter que les occurrences erronées de *débuter*, il faudrait enrichir le système d'un analyseur syntaxique capable de faire la distinction entre verbes transitifs et intransitifs. Et bien sûr, le même raisonnement s'applique aux contraintes sémantiques visant les types d'arguments. Par exemple, selon une des autorités que nous avons consultées, le verbe français *touer* ne peut accepter que des compléments directs appartenant à la classe des embarcations, contrairement à son sosie anglais *to tow*, qui peut avoir pour complément n'importe quel type de véhicule. Là encore, pour que *TransCheck* arrive à distinguer les occurrences correctes des occurrences incorrectes du verbe français, il lui faudra évidemment incorporer un analyseur sémantique.

En fait, si Martin Kay a raison d'affirmer qu'«il n'y a rien qu'une personne puisse savoir, ressentir ou rêver qui ne soit crucial à la bonne traduction d'un texte» [(12)], alors il n'y

---

(11) On trouvera un exemple de ce genre de bruit au tableau n°1 (xv). Dans ce cas, la paire illicite est constituée des noms *grant/octroi*, car le français *octroi* désigne uniquement l'action d'accorder une subvention, et non la subvention elle-même. Le problème est que la phrase française renferme aussi le nom *subvention*, qui est la traduction correcte de *grant*.

(12) Tiré (et traduit) de l'avant-propos de Hutchins et Somers (1992).

a, en principe, aucune limite aux connaissances linguistiques et extralinguistiques qu'un vérificateur de traduction devrait incorporer afin de pouvoir détecter toutes les erreurs de traduction possibles. Réviser une traduction n'exige pas moins de connaissances qu'il n'en faut pour produire un premier jet, et un vérificateur de traduction qui vise à émuler parfaitement un réviseur humain devra posséder un niveau de compréhension qui va bien au-delà du texte littéral pour englober toute l'intelligence qu'un traducteur compétent apporte à sa tâche. Il est bien évident qu'un tel système n'est pas près de faire irruption sur le marché, et il ne fait aucun doute que son apparition devra attendre l'avènement de la traduction entièrement automatique de grande qualité. Il y a, cependant, une différence majeure entre une approximation partielle du vérificateur de traduction idéal, comme le modeste prototype décrit dans cet article, et une approximation partielle d'un système de traduction automatique idéal. Un vérificateur de traduction moins qu'exhaustif, tout comme un correcteur orthographique incomplet, peut quand même être utile : dans un cas comme dans l'autre, les erreurs que le système arrive à détecter automatiquement permettent à l'utilisateur d'améliorer la qualité du texte final. Il n'est pas du tout évident que les systèmes de traduction automatique actuels, qui fonctionnent nécessairement à partir d'une compréhension incomplète du texte à traduire, puissent rendre un service comparable au traducteur humain. Il faut souvent plus de temps et d'effort pour corriger et remanier une approximation partielle de traduction produite par un système de traduction automatique qu'il n'en faut à un traducteur humain pour produire une version correcte à partir de zéro. Voir Pierre Isabelle (1993), où cet argument en faveur des outils de traduction, par opposition à la traduction automatique classique, est présenté de façon plus détaillée.

Par rapport aux exemples relativement simples de complémentation syntaxique et sémantique cités à la page précédente, on pourrait légitimement se demander si la détection de ces difficultés nécessite le recours à un vérificateur de traduction bilingue, ou si un vérificateur grammatical unilingue français ne ferait pas tout aussi bien l'affaire. Après tout, quand les experts, par exemple, qualifient d'«anglicisme» l'emploi transitif du verbe *contribuer*, en réalité, ils nous informent de la genèse du problème, c'est-à-dire du fait que les francophones ont emprunté la façon dont le verbe se construit en anglais. Cependant, quelle que soit son origine, cette difficulté demeure essentiellement un problème d'usage du français dont la détection automatique ne nécessite pas absolument le recours à la source d'une traduction. On pourrait facilement imaginer exactement le même problème survenant dans un document unilingue français n'ayant fait l'objet d'aucune traduction. En fait, on pourrait très bien considérer tous les problèmes traités par *TransCheck* dans l'optique d'un vérificateur de langue cible unilingue, même les faux amis comme *library/librairie*. De ce point de vue, le mot français *librairie* a deux sens, un sens correct et un incorrect. Quand ce mot apparaît dans une traduction, nous pouvons utiliser le texte source pour faire la distinction entre les deux sens : si *librairie* est utilisé pour traduire le mot anglais *library*, il est utilisé incorrectement. Mais, au fond, le recours au texte source n'est qu'une façon d'élucider le sens du mot, et on peut en imaginer d'autres.

En pratique, cependant, la question soulevée ci-dessus est quelque peu hypothétique (13). Si nous disposions d'une grammaire générative complète et calculable d'une langue, nous pourrions peut-être l'utiliser pour valider automatiquement la correction de tout texte rédigé dans cette langue, en fonction des formes correctes énumérées par cette grammaire et en fonction du complément de cet ensemble constitué par les formes incorrectes (14). Il va sans dire que cette grammaire n'existe pas et que tant qu'il en sera ainsi, la meilleure chose à faire consiste peut-être à décrire explicitement certaines des erreurs qui apparaissent fréquemment dans cette langue. C'est exactement ce que le système *TransCheck* tente de faire, particulièrement pour les erreurs qui sont susceptibles d'être commises par suite du contact entre deux langues; et pour chaque erreur de ce genre, il propose une correction possible.

Les extensions du modèle linguistique de *TransCheck*, auxquelles nous avons fait allusion plus haut, exigeront des efforts de recherche considérables, surtout si nous voulons que les résultats de l'analyse syntaxique et sémantique atteignent le niveau d'exactitude et de robustesse offert par le programme de catégorisation grammaticale que le système emploie actuellement. À court terme, nous prévoyons de procéder à des perfectionnements moins ambitieux, comprenant la vérification des expressions numériques entre

---

(13) En premier lieu, à notre connaissance, les correcteurs de grammaire française qu'on trouve actuellement sur le marché ne vérifient guère que l'accord grammatical, par exemple entre le verbe et le sujet, et même à ce niveau, leur performance est loin d'être brillante.

(14) Précisons que même cette grammaire ne pourrait pas nous aider à repérer les phrases de cette langue qui sont correctes, mais ne communiquent pas correctement le sens voulu par l'auteur.

textes source et cible [15], la détection des omissions d'unités importantes dans le texte traduit (c'est-à-dire des omissions de phrases et de paragraphes) et, peut-être, la vérification de la cohérence terminologique. Ces nouvelles fonctions seront décrites de façon plus détaillée dans les prochains articles faisant le point sur *TransCheck*.

5 Conclusion

Nous avons affirmé, au début de cet article, que l'avènement de la bitextualité permettait d'envisager un nouveau type d'outil rédactionnel pour traducteurs, un outil capable de détecter les erreurs de traduction qui échappent, de par leur nature, aux correcteurs orthographiques ou de grammaire unilingues. Au cours de l'exposé, nous avons quelque peu nuancé cette affirmation. Nous avons vu plusieurs types d'erreurs de traduction, notamment les emprunts illicites et les calques grammaticaux, qui pourraient, en principe, être détectées mais peut-être pas facilement corrigées par un outil rédactionnel unilingue. Pour d'autres types d'erreurs de traduction, en particulier celles qui sont caractérisées par le fait que le texte cible demeure grammaticalement correct mais ne véhicule pas le même sens que son pendant en langue source, un outil de validation bitextuelle comme *TransCheck* semble constituer la meilleure solution possible à l'heure actuelle. Nous avons avancé que même une approximation partielle d'un tel vérificateur de traduction pourrait être utile au traducteur ou au réviseur humain, et nous avons esquissé nos plans à court et à long terme pour augmenter les capacités du premier prototype de *TransCheck*.

Elliott Macklovitch,
Centre d'innovation en technologies de l'information,
Laval (Québec),
Canada.

Remerciements

Le système *TransCheck* décrit dans cet article est le fruit des efforts concertés de plusieurs chercheurs du Citi, dont j'aimerais souligner ici l'apport. Ce sont : Pierre Isabelle, directeur du groupe de traduction assistée par ordinateur du Citi, Monique Chevalier, George Foster, Marie-Louise Hanna, François Perrault, Xiaobo Ren, Michel Simard et Caroline Viel, que je tiens à remercier de l'excellent travail qu'ils ont accompli. J'assume évidemment l'entière responsabilité des erreurs qui auraient pu se glisser dans ce rapport.

Bibliographie

Brown (P.), Lai (J.) et Mercer (R.), 1991 : «Aligning sentences in parallel corpora», dans *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Californie, p. 169-176.

Colpron (Gilles), 1982 : *Dictionnaire des anglicismes*, Laval (Québec), Éditions Beauchemin.

Dagan (Ido) Church (Kenneth) et Gale (William), 1993 : «Robust Bilingual Word Alignment for Machine-Aided Translation», dans *Proceedings of the Workshop on Very Large Corpora*, Columbus, Ohio.

Dagenais (Gérard), 1984 : *Dictionnaire des difficultés de la langue française au Canada*, Boucherville (Québec), Les Éditions françaises.

De Villers (Marie-Éva), 1988 : *Multidictionnaire des difficultés de la langue française*, Montréal, Éditions Québec/Amérique.

Gale (William) et Church (Kenneth), 1991 : «A Program for Aligning Sentences in Bilingual Corpora» dans *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Californie, p. 177-184.

Harris (Brian), 1988 : «A New Concept in Translation Theory», dans *Language Monthly*, n°54.

Hutchins (John) et Somers (Harold), 1992 : *An Introduction to Machine Translation*, Academic Press.

Isabelle (Pierre), 1992 : «Bi-text: Toward a new Generation of Support Tools for Translation and Terminology», dans *Citi Technical Report*, 29. Publié en français sous le titre «La bitextualité : vers une nouvelle génération d'aides à la traduction et à la terminologie» dans *Meta*, vol. 37, n° 4, p. 721-737.

Isabelle (Pierre), 1993 : «Machine-Aided Human Translation and the Paradigm Shift», dans *Proceedings of the Fourth Machine Translation Summit*, Kobe.

Isabelle (Pierre), Dymetman (Marc), Foster (George), Jutras (Jean-Marc), Macklovitch (Elliott), Perrault (François), Ren (Xiaobo et Simard (Michel), 1993 : «Translation Analysis and Translation Automation», dans *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto.

Macklovitch (Elliott), 1992 : «Corpus-Based Tools for Translators», dans *Proceedings of the 33rd Annual Conference of the American Translators Association*, San Diego, Californie, p. 317-328.

Rey (Jean), 1984 : *Dictionnaire sélectif et commenté des difficultés de la version anglaise*, Paris, Éditions Ophrys.

Simard (Michel), Foster (George) et Isabelle (Pierre), 1992 : «Using Cognates to Align Sentences in Parallel Corpora», in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, p. 67-81.

Van Roey (Jacques) Granger (Sylviane) et Swallow (Helen), 1988 : *Dictionnaire des faux amis français-anglais*, Paris, Duculot.

(15) On trouvera, au tableau n°1 (xvi), un exemple de ce genre d'erreur, relevé dans le *Journal des débats*, que nous espérons pouvoir détecter ultérieurement avec *TransCheck*.

TABLEAU 1

	SOURCE	TRADUCTION
(i)	... the overall long term survival of any type and **stage** of childhood cancer ...	... le taux de survie à long terme, pour tous types et à tous les **stages** de cancer ... (19/05/92)
(ii)	... to ensure progress toward a world free of injustice, discrimination and **prejudice**.	... notre recherche d'un monde exempt d'injustice, de discrimination et de **préjudice**. (10/12/91)
(iii)	Since the passage of that **infamous** bill in 1987, Bill C-22 ...	Depuis l'adoption de **l'infâme** projet de loi C-22 en 1987 ... (17/11/92)
(iv)	In this country ... with our reputation for **affluence**, with our reputation for humanity ...	... dans notre pays reconnu ... pour son **affluence** et son empressement à aider ... (10/05/93)
(v)	Also, the polls say that regardless of the position of the Tories in England before and during the campaign – and it was up and down and **neck and neck** for a while ...	... quelle qu'ait été la position du Parti Conservateur en Angleterre avant et durant la campagne, et on peut dire que la position des partis a beaucoup fluctué, que ces derniers ont été **nez à nez** pendant un certain temps ... (10/04/92)
(vi)	They ... drive their children to practices and **games**.	Ils ... conduisent leurs enfants à des pratiques et à des **joutes**. (12/02/93)
(vii)	It is about $190 going over $200 **depending upon** how much money they earned ...	Il est d'environ 190 $ et peut dépasser les 200 $, **dépendant** du montant gagné ... (15/02/93)
(viii)	... the destruction of private property cannot be tolerated under **any circumstances**.	... la destruction de la propriété privée ne peut être tolérée **en aucun temps**. (5/05/92)
(ix)	... profits from a **used** equipment sale ...	... recettes d'une vente d'équipement de **seconde main** ... (4-06-90)
(x)	... his interest rate projections ... are on a **calendar year basis**.	... les taux d'intérêt sont basés sur l'année de **calendrier** (4/04/90)
(xi)	It is also an **essential condition** for the confidence ... that will ensure a strong recovery.	Il s'agit également d'un **prérequis** au retour de la confiance ... (25/02/92)
(xii)	You are resting on **technicalities.**	Vous vous arrêtez à des **technicalités**. (12/03/92)
(xiii)	... many of whom were historic **supporters** of the Conservative party ...	... dont beaucoup ont toujours été des **supporteurs** du Parti conservateur ... (21/03/91)
(xiv)	A real **rip-off** occurred during the long weekend.	Lorsqu'il y a des longues fins de semaine, on assiste à un véritable **racket** ... (25/05/93)
(xv)	Canada Employment has recently announced a section 25 **grant** to help it upgrade its facilities ...	Emploi Canada a récemment annoncé **l'octroi** à cette entreprise d'une subvention en conformité de l'article 25, afin de l'aider ... (27/04/93)
(xvi)	The European Economic Community's research and development project has a budget of **$17 million** spread over four years ...	Le projet de recherche et de développement de la Communauté économique européenne bénéficie d'un budget de **70 millions de dollars** réparti sur quatre ans ... (17/04/86)

TABLEAU 2

Corpus	Nombre total d'erreurs repérées	Erreurs véritables	Bruit	Précision
Débats 1	63	60	3	95 %
Débats 2	33	28	5	85 %
Débats 3	50	48	2	96 %
MDN 1	76	52	24	68 %
MDN 2	60	31	29	52 %

TransCheck : recopie d'écran I

TransCheck : recopie d'écran II