Un Protocole d'Évaluation Applicative des Terminologies Bilingues Destinées à la Traduction Spécialisée

Estelle Delpech

Un Protocole d'Évaluation Applicative des Terminologies Bilingues Destinées à la Traduction Spécialisée

Estelle Delpech[*,**]

[*]Lingua et Machina
c/o Inria Rocquencourt BP 105 Le Chesnay Cedex 78153
ed(a)lingua-et-machina.com - www.lingua-et-machina.com
[**]Université de Nantes - LINA UMR 6241
2, rue de la Houssinière BP 92208 44322 NANTES CEDEX 3
estelle.delpech(a)univ-nantes.fr - www.lina.univ-nantes.fr

Résumé. Cet article propose un protocole pour l'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée. Le protocole consiste à faire traduire des textes spécialisés dans différentes situations de traduction : sans ressource spécialisée, avec une terminologie adéquate, à l'aide d'Internet. La qualité des éléments traduits via ces ressources est ensuite comparée, ce qui permet de déterminer la valeur ajoutée des terminologies bilingues dans le cadre d'une tâche de traduction.

1 Introduction

L'évaluation est une étape cruciale dans le développement des outils de traitement automatique des langues : elle permet de rendre compte de la qualité des outils, en signale les limites, met en lumière les progrès accomplis et dégage de futures pistes de recherches. Nous nous penchons dans cet article sur l'évaluation des terminologies bilingues destinées à la traduction spécialisée. Nous avons mis en place un protocole d'évaluation applicative qui nous permet de déterminer la valeur ajoutée de ces terminologies lorsqu'elles sont utilisées pour traduire des textes spécialisés. Le protocole consiste à faire traduire des textes dans différentes situations de traduction : sans ressource spécialisée, avec une terminologie adéquate, à l'aide d'Internet. La qualité des éléments traduits via ces ressources est ensuite comparée, ce qui permet de déterminer l'apport effectif des terminologies.

Ce travail s'appuie sur les travaux en alignement de termes en corpus comparables, évaluation des outils terminologiques et évaluation de la qualité des traductions. Ces trois domaines sont présentés dans la section *2. Contexte*. La section *3. Protocole* décrit le protocole d'évaluation et argumente les choix méthodologiques. La section *4. Expérimentation* rend compte des résultats obtenus lors de la première mise en œuvre du protocole. Enfin, *Conclusions et perspectives* sont données dans la section *5*.

2 Contexte

2.1 Alignement de termes en corpus comparables

Les terminologies bilingues évaluées sont extraites de corpus comparables. Nous y référons par la suite en utilisant le sigle TBCC (Terminologies Bilingues issues de Corpus Comparables). Un corpus comparable est un ensemble de textes en deux langues, qui ne sont pas en relation de traduction mais qui traitent d'une même thématique. Un corpus parallèle est un ensemble de textes écrits dans une langue source accompagnés de leur traduction dans une langue cible. Les corpus comparables ont l'avantage d'offrir des termes et expressions produits spontanément et non influencés par une langue source. Ils sont aussi plus faciles à obtenir et permettent de travailler sur des couples de langues plus variés.

L'alignement d'éléments lexicaux en corpus comparables a débuté avec les travaux de Rapp (1995) et Fung (1997). La méthode consiste à aligner les termes qui apparaissent dans des contextes similaires. Le contexte d'un terme est représenté par un vecteur qui contient le nombre de fois où ce terme co-occurre avec chacun des mots du texte au sein d'une fenêtre de taille donnée. Le nombre de cooccurrences est normalisé, les vecteurs de contexte des termes sources sont traduits en langue cible à l'aide d'un dictionnaire-amorce, puis on calcule la distance entre vecteurs sources et vecteurs cibles. Plus les vecteurs de deux termes sont proches, meilleures sont les chances que ces termes soient des traductions l'un de l'autre.

L'exactitude des alignements est nettement en dessous de ce que l'on obtient avec des corpus parallèles. D'ailleurs, la précision des alignements se mesure sur un *TopN*, c'est-à-dire sur les *N* premières traductions proposées par l'algorithme d'alignement : une précision de 0,5 sur le Top20 signifie que la traduction exacte du terme source est présente parmi les 20 premiers candidats dans 50% des cas. Morin et Daille (2009) donnent un panorama des performances des algorithmes d'alignement : on peut attendre entre 80% et 42% de précision sur le Top20 en fonction des langues en jeu, du type et de la taille des corpus, des unités traduites (mots simples, termes simples, termes complexes). On doit donc considérer ces alignements comme des alignements ambigus, dans lesquels un terme source est généralement associé à une vingtaine de traductions candidates. Il est primordial de ne pas les livrer tels quels au traducteur, mais de les accompagner de connaissances linguistiques qui l'aideront à comprendre et utiliser le terme comme c'est le cas pour la terminologie évaluée (voir section 4.1.

2.2 Evaluation applicative

Le terme d'*évaluation applicative* est repris de Nazarenko et al. (2009) qui distinguent trois scénarios d'évaluation :

- évaluation via une référence : les sorties du système sont comparées à une référence. On calcule une mesure indiquant l'adéquation entre la référence et les sorties du système (pour les TBCC : la précision sur le TopN).
- évaluation de l'interaction : on compare les sorties du systèmes avant et après validation par un utilisateur, ce qui permet de déterminer un coût de post-édition (pour les TBCC : coût de la désambiguisation des alignements).
- évaluation applicative : on compare les résultats de l'application avec et sans la ressource terminologique. Les critères et le protocole d'évaluation dépendent de l'application considérée.

Les TBCC étant destinées à de la traduction spécialisée, on comparera la qualité de traductions produites avec et sans cette ressource.

2.3 Évaluation de la qualité des traductions

L'évaluation de la qualité des traductions est une problématique récurrente en traduction automatique (section 2.3.1) et en traductologie (section 2.3.2).

2.3.1 Qualité des traductions et Traduction Automatique (TA)

La TA a recours à deux techniques : évaluation automatique (ou objective) et évaluation humaine (ou subjective).

L'évaluation automatique est surtout utilisée pour une évaluation au jour le jour, afin de mesurer rapidement l'impact de modifications faites au système de TA. Ce mode d'évaluation utilise des mesures calculables automatiquement, simples et peu coûteuses à mettre en oeuvre. La plus connue est BLEU de Papineni et al. (2002) qui repose principalement sur le nombre de n-grammes de mots communs entre traduction à évaluer et traduction(s) de référence. D'autres mesures prennent en compte les variantes morphologiques ou les synonymes comme Meteor de Banerjee et Lavie (2005), d'autres comparent les structures syntaxiques et/ou sémantiques comme RTE de Pado et al. (2005) ou combinent plusieurs mesures comme UCL de Giménez et Márquez (2008).

Ces mesures sont méta-évaluées en calculant leur corrélation avec des jugements humains. Les résultats donnés lors des campagnes d'évaluation de Callison-Burch et al. (2009) et Callison-Burch et al. (2010) indiquent que ces mesures sont fiables lorsqu'on évalue tout un corpus de traductions. Leur utilisation sur des segments plus courts comme des phrases reste un problème ouvert. Il est aussi difficile d'identifier une mesure qui serait, de façon générale, plus fiable que les autres car les performances d'une même mesure varient selon le couple de langue, le sens de traduction et la granularité de l'évaluation.

L'évaluation humaine est la méthode utilisée dans les campagnes d'évaluation de TA. Elle consiste à faire évaluer la qualité des traductions par des juges. Cette méthode est nettement plus coûteuse et contraignante mais considérée comme plus fiable que l'évaluation automatique. Les éditions 2006 à 2010 du *Workshop on Statistical Machine Translation* font état de plusieurs protocoles et critères d'évaluation : Koehn et Monz (2006) demandent aux juges de noter l'adéquation et la fluidité des traductions sur une échelle allant de 1 à 5. Callison-Burch et al. (2007) proposent aux juges d'ordonner les phrases de la moins bien à la mieux traduite. Ils appliquent aussi ce protocole à des traductions de syntagmes. Callison-Burch et al. (2008) demandent aux juges de noter comme "acceptable" ou "pas acceptable" des traductions de syntagmes. Callison-Burch et al. (2009).

L'évaluation humaine étant subjective, on s'assure de la fiabilité des jugements en mesurant le degré d'accord inter- et intra- annotateurs. La mesure utilisée est le Kappa de Carletta (1996). Callison-Burch et al. (2007) démontrent que plus la tâche d'évaluation est complexe (beaucoup de catégories, segments évalués longs), plus les juges passent du temps à annoter et plus l'accord baisse.

2.3.2 Qualité des traductions et traductologie (AQT)

Dans sa version pragmatique, l'AQT produit des grilles d'évaluation destinées à l'industrie de la traduction qui y voit un moyen de contrôler la qualité de ses produits. Secară (2005) décrit plusieurs de ces grilles. Bien qu'aucune ne fasse consensus, toutes se basent sur le même principe : les erreurs sont organisées en une typologie, chaque type d'erreur est associé à un coût en points et une traduction de qualité ne doit pas dépasser un certain coût global. Les théoriciens de l'AQT reprochent à ces grilles de rester au niveau lexical et syntaxique et de ne pas prendre en compte les niveaux d'analyse supérieurs. Par exemple, Williams (2001) propose un mode d'évaluation basé sur la comparaison des structures argumentales. Ces grilles ont aussi l'inconvénient d'être monolithiques, supposées valables pour toutes les traductions, sans prendre en compte la fonction du texte, la situation de communication ou les attentes du commanditaire alors que des travaux comme ceux de Reiss (1971) préconisent au contraire d'adapter les critères d'évaluation à la fonction du texte à traduire.

Reiss (1971) établit une typologie fonctionnelle des textes à traduire. Les critères d'évaluation acquièrent plus ou moins de poids en fonction du type de texte traduit :

- textes centrés sur le contenu : articles de presse, travaux scientifiques, notices. Le traducteur adapte totalement la forme du texte à la langue cible en respectant en priorité le sens du texte source.
- textes centrés sur la forme : les textes littéraires, artistiques. Le traducteur respecte en priorité la forme du texte source, il jouit d'une plus grande liberté au niveau du transfert du sens.
- textes incitatifs : publicité, propagande. La traduction est une adaptation libre : son but premier est de conserver l'effet du texte sur le lecteur, sans obligation de respect de la forme ou du sens.
- textes audio-mediaux : pièces de théâtres, discours. Le traducteur doit adapter le texte à son environnement et à la manière dont il sera prononcé : mouvement des lèvres dans le sous-titrage, rythme dans les chansons.

3 Protocole

L'avantage des mesures d'évaluation automatique, qui est la reproductibilité, est faible : l'évaluation applicative des TBCC comportera une partie non reproductible (la tâche de traduction) ce qui rend, de toute façon, le processus d'évaluation non reproductible. On aura donc recours à une évaluation humaine. Les grilles d'évaluation de l'AQT ont le défaut d'être complexes à appliquer et peu documentées. On leur préférera le protocole d'évaluation de la TA, qui a l'avantage de simplifier la tâche d'évaluation, tant du côté des organisateurs que des juges. On retient deux tâches : la tâche de classement et la tâche de jugement décrites dans Callison-Burch et al. (2007).

Les travaux de traductologie démontrent que la qualité d'une traduction dépend de l'interaction de nombreux paramètres linguistiques et extra-linguistiques. Or nous savons bien que les TBCC n'agissent que sur certains d'entre eux (par ex. : orthographe, respect de la terminologie, interprétation du terme source). On ne peut donc pas juger la valeur ajoutée des TBCC sur la base de la qualité globale de la traduction, puisque cette qualité dépend d'autres paramètres sur lesquels les TBCC ont peu ou pas d'influence. C'est pourquoi on a choisi de

mesurer l'apport des TBCC uniquement sur la qualité de la traduction des éléments lexicaux qui ont posé problème aux traducteurs, soit là où précisément on attend un apport des TBCC.

Cette valeur ajoutée est mesurée par contraste avec deux autres situations :

Situation 0 Il s'agit de la ligne basse : les traductions sont produites par une personne qui n'est pas un-e professionnel-le de la traduction et uniquement avec des ressources génériques (dictionnaires bilingues et monolingues de langue générale).

Situation 1 les traductions sont produites par un traducteur professionnel avec les ressources génériques et la TBCC.

Situation 2 les traductions sont produites par un traducteur professionnel avec les ressources génériques et la possibilité d'utiliser Internet pour trouver les traductions, à l'exception des sites dont sont issus les textes à traduire et la base de donnés terminologiques en ligne *Termium* qui est utilisée par la suite pour aider à juger les traductions.

Dans les trois situations, les ressources génériques sont les mêmes. La traduction se fait de langue seconde vers la langue maternelle du traducteur. De façon à lisser les différences de qualité de traduction qui pourraient apparaître du fait de l'expérience du traducteur plutôt que grâce à la qualité de la ressource spécialisée, les situations 1 et 2 seront jugées sur la base de traductions faites par les deux traducteurs professionnels. En retour, il faut éviter qu'un traducteur traduise un texte issu d'un domaine donné dans le scénario 1 puis s'attèle à un autre texte issu du même domaine dans le scénario 2, car il y a un risque qu'il réutilise les connaissances acquises dans la première situation. Afin de contrer ce risque, chacun des traducteurs professionnels aura des textes de domaines différents pour chaque situation de traduction (voir tableau 1). L'idéal, bien sûr, est de multiplier le nombre de traducteurs de façon à juger les différentes situations sur la base de nombreuses traductions, ce qui donne plus de représentativité à l'évaluation.

	textes du domaine 1	textes du domaine 2
Situation 0	traducteur non professionnel	traducteur non professionnel
Situation 1	traducteur professionnel 1	traducteur professionnel 2
Situation 2	traducteur professionnel 2	traducteur professionnel 1

TAB. 1 – *Répartition des domaines et situations de traductions entre traducteurs*

Une fois le texte traduit, le traducteur note le temps passé à traduire, les termes qui ont posé problème, la traduction finalement retenue et le type de ressources utilisées pour produire la traduction. Les termes problématiques sont regroupés, anonymisés et présentés aux juges avec les traductions retenues dans les différentes situations. Les juges classent les traductions de la meilleure à la moins bonne (les égalités sont autorisées) et notent séparément la qualité de chaque traduction selon des critères inspirés des travaux de Reiss (1971) qui recommande de donner la priorité au sens dans le cas des textes centrés sur le contenu (voir tableau 2).

Sans expert du domaine, il est difficile de juger du transfert du sens et de l'idiomaticité des traductions. Les textes utilisés sont donc des textes pour lesquels il existe déjà une traduction. Les juges, qui sont des étudiants de dernière année d'école de traduction, ont accès aux docu-

	transfert du sens	respect de la forme
exact	✔	✔
acceptable	✔	
faux		

TAB. 2 – *Critères pour juger la qualité des traductions*

ments d'origine et aux phrases dans lesquelles apparaîssent le terme source et sa traduction ; ils peuvent aussi s'aider de la base de données terminologiques *Termium* [1].

Mis à part les instructions d'annotation et quelques exemples d'annotations sur des cas difficiles, les juges n'ont reçu aucune formation. Il est courant, dans le cadre de campagnes d'évaluation, d'avoir recours à une première évaluation à blanc. Par exemple, Blanchon et Boitet (2007) préparent leurs juges en leur fournissant une fiche d'instructions et en effectuant une première évaluation à blanc. Les divergences sont ensuite discutées afin de normaliser la notation.

4 Expérimentation

4.1 Terminologie évaluée

Les TBCC sont construites de la façon suivante :
- Les termes sont extraits à l'aide du logiciel *Similis* de Planas (2005).
- Les termes, ainsi que chacune des lexies du corpus d'acquisition ayant un nombre d'occurrences supérieur à 5 sont alignés en utilisant l'algorithme de Fung (1997).
- Pour chaque terme et lexie, on génère automatiquement une fiche terminologique indiquant sa partie du discours, fréquence, termes proches, variantes, définition, collocations et un concordancier.
- La terminologie est consultable via l'interface décrite par Delpech et Daille (2010) qui offre aussi une fonctionalité de recherche des termes et lexies dans le corpus d'acquisition.

4.2 Données

Le protocole est testé avec des textes issus du domaine médical, thématique CANCER DU SEIN et des textes issus du domaine des sciences de l'environnement, thématique SCIENCES DE L'EAU (voir tableau 3). La thématique SCIENCES DE L'EAU est nettement moins précise que la thématique CANCER DU SEIN mais le corpus d'acquisition est plus volumineux (0,4M mots vs 2M mots). Les langues sont le français et l'anglais.

4.3 Traducteurs et juges

Disposant de peu de moyens humains pour expérimenter le protocole, nous avons dû faire quelques entorses méthodologiques : il y a eu des collisions entre les rôles d'organi-

[1] http ://www.termiumplus.gc.ca/

	CANCER DU SEIN	SCIENCES DE L'EAU
corpus d'acquisition	$\simeq$ 400k mots par langue portail *Elsevier*[I]	$\simeq$ 2M mots par langue revues *Siences de l'eau*[II] et *Water Science and Technology*[III]
textes scientifiques	3 résumés d'articles 508 mots portail *Elsevier*	3 résumés d'articles 499 mots revue *Sciences de l'eau*
textes de vulgarisation	1 page web 613 mots site *Société canadienne du cancer du sein*[IV]	1 page web 425 mots site *Lenntech* sur le traitement des eaux[V]

[I] http ://www.elsevier.com/
[II] http ://www.rse.inrs.ca/
[III] http ://www.iwaponline.com/wst/
[IV] http ://www.cbcf.org/
[V] http ://www.lenntech.com/

TAB. 3 – *Données : corpus d'acquisition des TBCC et textes à traduire*

sateur/traducteur et traducteur/juge. Le rôle du traducteur non-professionel a été tenu par l'auteure de l'article qui a aussi organisé l'évaluation. Les deux autres traducteurs étaient des étudiants en dernière année d'école de traduction, on peut les considérer comme semi-professionnels. Ils ont aussi jugé et classé les traductions (l'anonymisation empêchant les juges de savoir qui ou dans quelle situation avait été produites les traductions). La langue maternelle des trois personnes est le français.

4.4 Résultats

4.4.1 Impressions de traducteurs

Les traducteurs ont eu du mal à accepter le fait qu'une TBCC soit ambiguë. Bien que le but et le contexte de l'évaluation ait été expliqués, les traducteurs s'attendaient plutôt à obtenir la traduction des termes sur simple clic. Citons une des réactions :

> *En gros, 75% des mots techniques ne figurent pas dans le glossaire, et sur les 25% restants, 99% ont entre 10 et 20 traductions candidates, mais aucune de validée. Du coup, dans le meilleur des cas on est "à peu près sûr", mais jamais totalement. Et dans le pire des cas (très fréquemment, malheureusement) on y va "à l'instinct".*

Un second problème est que les TBCC, notamment celle des SCIENCES DE L'EAU, couvraient peu le vocabulaire des textes à traduire. Le tableau 4 indique le pourcentage de mots des textes à traduire qui se trouvent effectivement dans les terminologies. On voit que la terminologie CANCER DU SEIN, bien qu'acquise sur un corpus plus petit, couvre plus de vocabulaire que la terminologie SCIENCES DE L'EAU. La thématique des sciences de l'eau est beaucoup trop large et ne permet pas d'extraire un vocabulaire ciblé. Il faut donc favoriser des thématiques fines plutôt que des corpus volumineux.

	Cancer du sein	Sciences de l'eau
textes à traduire (EN)	94%	14%
traductions de référence (FR)	67%	78%

TAB. 4 – *Couverture des TBCC par rapport aux textes à traduire et leurs traductions*

4.4.2 Temps de traduction

Les 8 textes à traduire totalisent 2147 mots. La situtation impliquant uniquement les ressources génériques est celle qui demande le moins de temps de traduction (7,15 mots/sec.), ce qui est normal car le traducteur a moins de ressources à parcourir. Il n'y a pas de différence notable entre les deux autres situations : 11,18 mots/sec. et 11,6 mots /sec. pour les situations 1 et 2 respectivement.

4.4.3 Accord inter-annotateur

L'accord inter-annotateur a été calculé avec la mesure Kappa de Carletta (1996). Cette mesure prend en compte l'accord observé P(A) et la probabilité d'un accord aléatoire P(E).

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

L'accord a été meilleur pour la tâche de classement : 0, 65 (accort fort) que pour la tâche de jugement : 0,36 (accord faible), ce qui confirme les résultats de Callison-Burch et al. (2007). Il a été meilleur pour les textes de vulgarisation : 0,57 (accord modéré) que pour les textes scientifiques : 0,48 (accord modéré).

4.4.4 Jugement

Le tableau 5 donne les jugements pour les traductions du domaine CANCER DU SEIN. On y voit que la proportion de traductions jugées fausses est quasi-équivalente dans les trois situations. Les traductions produites avec la TBCC sont plus exactes que celles produites sans aucune ressource spécialisée et moins exactes que celles produites avec l'aide d'Internet. Le tableau 6 donne les jugements effectués sur les traductions du domaine SCIENCES DE L'EAU. On y voit que les traductions produites avec la TBCC sont plus souvent fausses que celles traduites sans aucune ressource spécialisée. Ceci n'est pas normal car les deux situations partagent un socle commun de ressources génériques. Les traductions produites avec les TBCC auraient dû

	ress. gén.	ress. gén. + TBCC	ress. gén. + Web
exact	38%	43%	47%
acceptable	42%	38%	35 %
faux	20%	19%	18%

TAB. 5 – *Jugement de la qualité des traductions - domaine* CANCER DU SEIN

	ress. gén.	ress. gén. + TBCC	ress. gén. + Web
exact	59%	56%	77%
acceptable	23%	23%	16 %
faux	18%	21%	7%

TAB. 6 – *Jugement de la qualité des traductions - domaine* SCIENCES DE L'EAU

être au moins aussi bonnes que celles produites sans ressource spécialisée. Il se trouve que, selon la situation dans laquelle ils étaient, les traducteurs n'ont pas utilisé les ressources de la même façon. Grâce aux données collectées durant la phase de traduction, on peut savoir, pour chaque traduction, si celle-ci a été obtenue à l'aide de la ressource générique ou de la ressource spécialisée ou en utilisant l'intuition (non exclusif). Le tableau 7 montre que les traducteurs ayant à leur disposition des ressources spécalisées ont eu très peu recours à la ressource générique. Il se peut qu'il ne leur ait pas paru utile d'employer cette ressource, sachant qu'elle est susceptible de ne pas contenir de termes techniques et qu'ils avaient, à leur disposition, une ressource spécialisée. Or, la TBCC SCIENCES DE L'EAU ayant une très faible couverture, cette dernière n'a pas été d'une grande aide. Une utilisation systématique de la ressource générique dans la situation 1 aurait donné des résultats "au moins aussi bons" que la situation 0.

	Situation 0	**Situation 1**	**Situation 2**
ress. gén.	43%	14%	3%
ress. spéc. (TBCC ou Web)	-	25%	56 %
intutition	79%	77%	44%

TAB. 7 – *Utilisation des ressources en fonction des situations de traduction*

4.4.5 Classement

On retrouve des résultats similaires pour la tâche de classement. Lorsque les traductions d'un même terme sont comparées entre elles, celles produites avec l'aide d'Internet sont toujours les meilleures, quel que soit le domaine. Les traductions faites avec les TBCC sont meilleures que celles produites sans ressource spécialisée uniquement dans le domaine CANCER DU SEIN et pas pour le domaine SCIENCES DE L'EAU, très probablement pour les raisons expliquées plus haut.

	ress. gén. vs TBCC	ress. gén. vs Web
meilleur	28%	26%
idem	47%	42%
moins bon	26%	32%

TAB. 8 – *Classement des traductions - domaine* CANCER DU SEIN

	ress. gén. vs TBCC	ress. gén. vs Web
meilleur	18%	16%
idem	49%	41%
moins bon	33%	43%

TAB. 9 – *Classement des traductions - domaine* SCIENCES DE L'EAU

5 Conclusion et perspectives

Nous avons décrit un protocole d'évaluation applicative pour les terminologies bilingues destinées à la traduction spécialisée. Ce protocole propose de comparer diverses situations de traductions, dans lesquelles les traducteurs ont à leur disposition des ressources différentes : soit uniquement des ressources génériques, soit des ressources génériques et la terminologie bilingue, soit des ressources génériques et un accès à Internet. Les différences de qualité entre les traductions produites avec ou sans la terminologie bilingue permettent de mesurer la valeur ajoutée de cette dernière dans le cadre d'une tâche de traduction spécialisée.

Une première expérimentation du protocole, bien qu'effectuée avec un jeu restreint de données et de participants, a permis de tester sa faisabilité et d'identifier les points problématiques :

- La valeur ajoutée des TBCC dépend fortement de leur degré de couverture des textes avec lesquels elle est évaluée. Toute mesure de valeur ajoutée doit aussi indiquer cette couverture ainsi que le degré de comparabilité des corpus source et cible, sinon elle n'est pas interprétable. Une piste de recherche est de mettre au point une mesure d'adéquation entre la terminologie et les textes à traduire.
- L'utilisation conjointe de plusieurs ressources dans une situation de traduction vient parasiter les résultats. Il est préférable de n'avoir qu'une seule ressource par situation de traduction, par exemple : situation 0 sans aucune ressourc, situation 1 avec les TBCC uniquement, situation 2 avec Internet uniquement.
- Les traducteurs doivent être mieux préparés à utiliser des terminologies ambiguës. L'idéal serait de recourir à une première traduction à blanc pour recueillir leurs impressions et les aider à appréhender ce type de ressource.

La prochaine étape dans la mise au point de ce protocole sera de le tester à plus grande échelle. Nous songeons notamment à l'expérimenter sur une classe entière d'étudiants traducteurs, sans collision entre les rôles d'organisateur, de traducteur et de juge, avec plus de variété dans les textes traduits et des thématiques mieux définies.

Enfin, même si ce n'est pas le but de ce travail, cette première évaluation donne des pistes de recherche pour améliorer l'apport des TBCC. D'une part, le corpus d'acquisition doit être constitué en fonction des textes à traduire et en suivant une thématique très fine. D'autre part, on se rend bien compte que le Web, sauf dans le cas de domaines très restreints (ex. : terminologie propre à une entreprise), contiendra toujours plus de solutions de traductions. Il faut donc inclure dans l'interface de consultation des TBCC des appels à des programmes de traduction à la volée qui puissent, lorsqu'un terme n'est pas présent dans la base, soit générer une traduction candidate et la filtrer sur Internet, soit aller chercher la traduction dans des ressources en ligne définies par le traducteur.

Remerciements

Ce travail a été financé par la société Lingua et Machina et l'ANR (subvention n° ANR-08-CORD-009). Je tiens également à remercier Clémence De Baudus et Mathieu Delage de l'Institut Supérieur d'Interprétariat et de Traduction (ISIT) pour leur participation aux tâches de traduction et d'évaluation.

Références

Banerjee, S. et A. Lavie (2005). METEOR : an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics*, Ann Arbor, Michigan, pp. 65–72.

Blanchon, H. et C. Boitet (2007). Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *Traitement Automatique des Langues 48*(1), 33–65.

Callison-Burch, C., F. Camerob, P. Koehn, C. Monz, et J. Schroeder (2008). Further Meta-Evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, pp. 70–106.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, et J. Schroeder (2007). (Meta-) evaluation of machine translation. In *Proceedings of the 2nd workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 136–158.

Callison-Burch, C., P. Koehn, C. Monz, K. Peterson, M. Przybocki, et O. Zaidan (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. Uppsala, Sweden.

Callison-Burch, C., P. Koehn, C. Monz, et J. Schroeder (2009). Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, pp. 1–28. Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics 22*(2), 249–254.

Delpech, E. et B. Daille (2010). Dealing with lexicon acquired from comparable corpora : validation and exchange. In *Proceedings of the 2010 Terminology and Knowledge Engineering Conference (TKE 2010)*, Dublin, Ireland, pp. 211–223.

Fung, P. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong, pp. 192–202.

Giménez, J. et L. Márquez (2008). A smorgasbord of features of automatic MT evaluation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, 195-198.

Koehn, P. et C. Monz (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pp. 102–121.

Morin, E. et B. Daille (2009). Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation (LRE)* (Springer Netherlands ed.), Volume 44 of

Multiword expression : hard going or plain sailing, pp. 79–95. P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón.

Nazarenko, A., H. Zargayouna, O. Hamon, et J. V. Puymbrouk (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique des Langues 50*(1), 257–281.

Pado, S., M. Galley, D. Jurafsky, et C. D. Manning (2005). Machine translation evaluation with textual entailment features. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, Greece. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, et W. Zhu (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp. 311–318.

Planas, E. (2005). Similis : un logiciel d'aide à la traduction au service des professionnels. *Traduire* (206), 41–48.

Rapp, R. (1995). Identifying word translations in Non-Parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Boston, Massachussets, USA, pp. 320–322.

Reiss, K. (1971). *Translation criticism, the potentials and limitations : categories and criteria for translation quality assessment*. Manchester, GB : St. Jerome Pub.

Secară, A. (2005). Translation evaluation - a state of the art survey. In *eCoLoRe / MeLLANGE Workshop*, Leeds, UK, pp. 39–44.

Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta : journal des traducteurs / Meta : Translator's Journal 46*(2), 326–344.

Summary

This paper describes a protocol for the evaluation of bilingual specialized terminologies through an application to specialized translation. The protocol consits in having specialized texts translated in various situations : without any specialized resource, with an adequate terminology or using Internet. By comparing the quality of the segments translated using the various resources, we are able to determine the added-value of bilingual terminologies in specialized translation.