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
[**]Université de Nantes - LINA UMR 6241
2, rue de la Houssinière BP 92208
44322 NANTES CEDEX 3
estelle.delpech(a)univ-nantes.fr

Résumé. Cet article argumente en faveur d'une évaluation applicative des terminologies bilingues et propose un protocole d'évaluation pour ce type de terminologies . Le protocole est appliqué aux terminologies issues de corpus comparables et l'application envisagée est la traduction humaine spécialisée. Le protocole consiste à faire traduire des textes spécialisés dans différentes situations de traduction : sans ressource spécialisée, avec une terminologie issue d'un corpus comparable, 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 spécialisée.

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 recherche.

En ce qui concerne l'évaluation des terminologies, Nazarenko et al. (2009) montrent que les terminologies sont des objets complexes et que leur évaluation peut être assez laborieuse. Ces auteurs distinguent trois modes d'évaluation :

Comparaison à une référence : les sorties du système sont comparées à une terminologie de référence. On calcule une mesure indiquant l'adéquation entre la référence et les sorties du système.

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 .

Évaluation applicative : on compare les résultats d'une application avec et sans la ressource terminologique. Les critères et le protocole d'évaluation dépendent de l'application considérée.

Nazarenko et al. (2009) indiquent des protocoles et des métriques pour les deux premiers modes d'évaluation et se concentrent uniquement sur l'évaluation des terminologies monolingues. Dans cet article, nous nous penchons sur l'évaluation applicative des terminologies bilingues et spécialement celles issues de corpus comparables.

Les terminologies issues de corpus comparables sont habituellement évaluées par comparaison à une référence (la mesure utilisée est une précision sur le $TopN$ - cette notion est expliquée en détails dans la section 2). Ce type d'évaluation est relativement peu coûteux à mettre en place (pour peu qu'une référence soit disponible) et est couramment utilisée pour évaluer, au jour au jour, l'effet de modifications apportées à l'algorithme d'alignement ou lorsque l'on souhaite comparer plusieurs systèmes d'alignement entre eux. Toutefois, nous pensons qu'il est important de rendre compte de l'impact et de l'utilité des outils d'extraction de terminologies bilingues *en contexte d'utilisation*, c'est-à-dire, lorsqu'ils sont intégrés à des applications ou lorsque leur produit est employé tel quel par les utilisateurs finaux.

Renders et al. (2003), par exemple, ont montré l'influence de lexiques bilingues issus de corpus comparables dans une tâche de recherche d'information cross-lingue. Nous souhaitons, dans ce travail, nous pencher sur le cas de la traduction spécialisée et mettre au point un protocole d'évaluation applicative qui permette de rendre compte de la valeur ajoutée de nos terminologies dans un contexte de traduction spécialisée humaine. Ce protocole se base sur des "situations de traductions" que l'on peut considérer comme des reconsitutions *in vitro* des situations dans laquelle un traducteur peut-être amener à effectuer son travail. L'approche est contrastive : le protocole consiste à faire traduire des textes avec différentes ressources spécialisées ; la qualité des éléments traduits via ces ressources est ensuite comparée, ce qui permet de déterminer l'apport effectif des terminologies.

L'article est organisé comme suit. La section 2 revient sur l'acquisition de terminologies et de lexiques bilingues en corpus comparable ainsi que sur les méthodes d'évaluation utilisées dans le domaine. Dans la section 3, nous nous penchons sur la question de l'évaluation de la qualité des traductions. La section 4 présente et argumente les choix méthodologiques ayant présidés à l'élaboration du protocole d'évaluation. La section 5 décrit les données et le cadre expérimental et la section 6 présentent les résultats obtenus.

2 Acquisition terminologique en corpus comparables

Si les techniques d'extraction terminologique et d'alignement en corpus parallèles sont aujourd'hui rôdées et bien connues[1], il n'en est pas de même pour l'alignement en corpus comparables. Nous décrivons dans la section 2 la technique la plus couramment employée pour acquérir des paires de traductions dans de tels corpus. La section 2.2 discute des méthodes d'évaluation.

2.1 Méthode d'alignement

L'intérêt pour les corpus comparables en extraction terminologique bilingue démarre avec les travaux de Rapp (1995) et Fung (1997). Un corpus comparable est un ensemble de textes

[1] Pour une revue des différents extracteurs terminologiques voir Castellví et al. (2001), pour un exemple d'aligneur de termes en corpus parallèle voir Gaussier et al. ou Lefever et al. (2009)

en deux langues, qui ne sont pas en relation de traduction mais qui traitent d'une même thématique. (Déjean et Gaussier, 2002, p. 2) les définissent comme des corpus de deux langues l_1 et l_2 pour lesquels « *il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1* ». A l'opposé, un corpus parallèle est un ensemble de textes écrits dans une langue source accompagnés de leur traduction dans une langue cible. L'intérêt pour les corpus comparables se justifie par leur plus grande disponibilité et leur facilité de constitution, ce qui permet de traiter de nouvelles paires de langues ainsi que des langues ou des domaines peu fournis en corpus parallèles. Un second avantage est le caractère spontané et naturel des termes et expressions rencontrés dans ces corpus : en effet, les textes en langue cible ne sont pas des traductions, leur production n'a pas été influencée par une langue source.

Les techniques d'extraction s'appuient sur la sémantique distributionnelle, initiée par Z. Harris. On considère que des termes de sens proches ont une distribution similaire et que ceci est valable de façon cross-lingue. Techniquement, la distribution des termes à traduire est représentée par un *vecteur de contexte* qui contient le nombre de fois où le terme tête du vecteur co-occurre avec chacun des mots du texte au sein d'une fenêtre de taille donnée. Le nombre de cooccurrences est normalisé à l'aide, par exemple, du taux de vraisemblance de Dunning (1993). Ensuite, 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 via une mesure de similarité comme le Cosinus ou le coefficient Jaccard - voir par exemple les travaux de Morin et al. (2004) pour une description de ces mesures. Plus les vecteurs de deux termes sont proches, meilleures sont les chances que ces termes soient des traductions l'un de l'autre.

2.2 Évaluation des alignements

Les algorithmes qui extraient des terminologies bilingues à partir de corpus comparables génèrent une liste d'alignements $1-n$: chaque terme source est aligné avec les n meilleures traductions candidates ; on parle alors des $TopN$ traductions candidates.

Actuellement, l'unique méthode d'évaluation utilisée dans le domaine est une évaluation par comparaison à une référence. La sortie de l'algorithme est comparée à un lexique ou une terminologie de référence et la métrique d'évaluation correspond à un taux de précision calculé sur les $TopN$ traductions candidates. Par exemple, une précision de 50% sur le $TopN$ signifie qu'une traduction correcte est trouvée parmi les N meilleurs candidats dans 50% des cas.

L'exactitude des alignements est nettement en dessous de ce que l'on obtient avec des corpus parallèles, qui serait plutôt de l'ordre de 80% à 90% sur le $Top1$. Morin et Daille (2009) donnent un panorama des performances des algorithmes d'alignement en corpus comparables : 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 extraites du corpus et qui l'aideront à comprendre et utiliser le terme. C'est le cas pour les terminologies évaluées que nous considérons comme des ressources bilingues "riches" dans le sens où elles ne se contentent pas de donner de simples liens de traductions mais aussi des variantes de termes, des informations de fréquence, des concordances, etc.

La méthode d'évaluation par comparaison à une référence a beaucoup d'avantages pratiques (facilité de mise en œuvre, reproductibilité) mais est critiquable sur plusieurs points :

- La critique de Blanchon et Boitet (2007) sur l'utilisation de référence pour l'évaluation en traduction automatique vaut aussi pour l'évaluation des alignements. La faible disponibilité des références oblige à réutiliser quasi-systématiquement les mêmes données, les algorithmes et les progrès réalisés sont toujours évalués sur la même référence, avec le risque d'une baisse des performances dès que l'on passe à un domaine ou un corpus différent, voire même de produire un système complètement ajusté à la référence.
- La notion de référence est discutable, particulièrement pour juger de traductions. D'une part, on sait bien qu'il existe rarement *une et un seule* traduction correcte. D'autre part, l'appariemment entre référence et sortie du système est binaire : la traduction correspond ou ne correspond pas ; or, il est évidemment que la traduction d'un terme peut-être plus ou moins bonne, plus ou moins adaptée au contexte. La notion de contexte est d'ailleurs totalement absente des évaluations par comparaison à une référence : l'exercice consiste à juger des correspondances *terme source* $\leftrightarrow$ *terme cible* en dehors de toute réalité de traduction, ce qui est totalement artificiel. La terminologie classique nous indique que le sens d'un terme est stable, que le terme entretient une relation bijective avec la notion ou l'objet du monde qu'il désigne. On pourrait alors en déduire qu'il n'existe qu'une seule traduction possible, indépendante du contexte. Mais cette vision de la terminologie est largement contestée depuis les années 90. Les nouvelles approches comme celle de Bourigault et Slodzian (2000) rappellent que le terme est aussi une unité linguistique et qu'en tant que tel, il est utilisé en discours où il est soumis à des variations aussi bien morpho-syntaxiques et que sémantiques [2].
- L'évaluation par comparaison à une référence est une évaluation incomplète. Elle laisse croire que la qualité d'une terminologie bilingue se réduit à l'exactitude des alignements de termes (pour peu que cela ait un sens). Or, comme indiqué plus haut, les terminologies s'enrichissent de plus en plus et se résument de moins en moins à des listes de paires de traduction. Et surtout, l'évaluation par comparaison à une référence ne rend pas compte de l'*utilité* de la ressource crée.

Tous ces points argumentent en faveur d'une évaluation en contexte, dans laquelle on juge de la capacité d'une ressource terminologie bilingue à remplir son rôle, c'est-à-dire permettre à un traducteur spécialisé de traduire mieux, plus vite et plus aisément.

Evidemment, on pourra toujours faire valoir le coût que représente une évaluation applicative, surtout lorsqu'elle implique une tâche de traduction humaine comme c'est le cas pour le protocole présenté ici. Il est clair que l'évaluation par comparaison a un avantage pratique indéniable et est certainement la méthode la plus adaptée pour évaluer les progrès faits par un système d'alignement dans le cadre d'un développement quotidien. Néanmoins, nous considérons que l'on ne peut pas faire l'économie d'une évaluation applicative à étapes régulières dans le développement d'un système d'acquisition terminologique (par exemple, à chaque nouvelle version majeure) ou dans le cadre de campagnes d'évaluation.

L'évaluation applicative est dépendante des protocoles propres à l'application finale. Dans notre cas, l'application finale est la traduction spécialisée. Il nous faudra donc pourra juger de la qualité de traductions produites par des humains, avec ou sans une terminologie bilingue.

[2] Sur la question de la polysémie des termes et des exemples de polyacception, voir les travaux sur le point de vue en langue spécialisée de Condamines et Rebeyrolle (1996)

Dans la section suivante, nous nous penchons sur les méthodes d'évaluation de qualité des traductions.

3 La qualité en traduction

La problématique de l'évaluation des traductions se retrouve dans deux domaines : la traduction automatique (TA) et la traductologie. Bien que nous souhaitions, dans ce travail, évaluer des traductions humaines, il a semblé intéressant de s'enquérir des techniques employées en TA. L'objet évalué par ces deux domaines est sensiblement différent. Alors que la traductologie cherche à évaluer les productions de traducteurs professionnels, de façon absolue et non relativement à une traduction "de référence", la TA évalue plutôt la capacité d'un système à produire une traduction la plus proche possible de ce qui aurait été produit par un traducteur. L'évaluation de la qualité des traductions dans ces deux domaines est étudiée respectivement dans les section 3.1 et 3.2.

3.1 Qualité des traductions et Traduction Automatique (TA)

L'évaluation en TA remplit deux objectifs. D'une part, il s'agit d'analyser, au jour au jour, les impacts d'une modification d'un système de TA sur la qualité des traductions. D'autre part, l'évaluation permet de comparer les systèmes entre eux, généralement lors de campagnes d'évaluation de grande envergure. À ces deux objectifs correspondent *grosso modo* deux techniques d'évaluation. Pour une évaluation au jour le jour, les développeurs de systèmes de TA utilisent des mesures calculables automatiquement à partir de traductions de référence, on parle alors d'*évaluation automatique* ou *évaluation objective*. Ces mesures, simples et peu coûteuses à mettre en œuvre, restent néanmoins perçues comme les substituts pratiques d'un autre type d'évaluation bien plus coûteux mais jugé meilleur. Ce second type d'évaluation, appelé *évaluation humaine* ou *évaluation subjective* est celui qui est utilisé dans les campagnes d'évaluation comme celle du *Statistical Workshop on Machine Translation* de l'ACL dont les résultats des dernières éditions sont donnés par Koehn et Monz (2006), Callison-Burch et al. (2007), Callison-Burch et al. (2008), Callison-Burch et al. (2009) et Callison-Burch et al. (2010). Il consiste à demander à des juges de noter la qualité des traductions. On imagine facilement le coût en termes de temps, d'organisation, de formation des juges, sans compter que les résultats ne sont pas reproductibles. Toutefois, le consensus actuel est en faveur de l'évaluation humaine, jugée comme plus à même de rendre compte de la qualité d'une traduction.

Dans les parties suivantes, nous rendons compte des techniques d'évaluation automatique (section 3.1.1) et des techniques d'évaluation par des humains (section 3.1.2).

3.1.1 Mesures pour l'évaluation automatique

L'évaluation automatique mesure la qualité d'une traduction de façon indirecte : on n'évalue pas la qualité de la traduction elle-même mais sa ressemblance avec une traduction de référence, produite par un traducteur professionnel. A défaut de pouvoir manipuler et comparer des paramètres linguistiques tels que la conservation du sens ou la fluidité du texte, les mesures d'évaluation emploient des indices de surface comme les mots ou suites de mots communs entre traduction évaluée et traduction de référence.

La mesure la plus connue et certainement la plus utilisée est BLEU de Papineni et al. (2002). Elle s'appuie sur les critères suivants :

- le nombre de n-grammes de mots communs à la traduction à évaluer et à la traduction de référence, pour n allant de 1 à 4
- les différences de taille (en nombre de mots)
- les possibilités de variation dans la traduction : un même texte pouvant être traduit de plusieurs façons différentes, le score BLEU peut être calculé avec plusieurs traductions de référence, de façon à autoriser plus de variation dans les formulations.

À la suite de BLEU, d'autres métriques ont été proposées dans le but d'améliorer la justesse de l'évaluation des systèmes de TA. Parmi les mesures concurrentes à BLEU, on trouve :

NSIT de Doddington (2002) : équivalente à BLEU, si ce n'est que les n-grammes sont pondérés en fonctions de leur fréquence (les n-grammes les plus fréquents étant jugés moins informatifs) et que la précision globale est calculée en utilisant la moyenne arithmétique au lieu de géométrique.

Une adaptation de la F-mesure de Turian et al. (2003) : Cette mesure a été conçue dans le but d'être facilement "interprétable" : elle se base sur les mesures standards de recherche d'information que sont le rappel et la précision[3]. Rappel et précision sont dans ce cas calculés sur le nombre de n-grammes communs à la traduction à évaluer et la traduction de référence.

Meteor de Banerjee et Lavie (2005) : associe précision et rappel calculés sur des unigrammes de mots à une mesure prenant en compte l'ordre des mots. En plus des mots identiques, Meteor considère également les mots semblables tels que les variantes morphologiques ou les synonymes. Un des buts de cette mesure est de permettre une évaluation au niveau de la phrase, alors que les autres mesures ne fonctionnent bien que lorsqu'on évalue tout un corpus de traductions.

TER de Snover et al. (2006) : calcule le nombre d'opérations d'édition nécessaires pour parvenir de la traduction évaluée à la traduction de référence.

Ces mesures d'évaluation peuvent elles-mêmes être méta-évaluées en calculant leur corrélation avec des jugements humains. Les métriques sont évaluées sur un corpus de traductions - elles sont dans ce cas plutôt fiables - ou des phrases. D'après Callison-Burch et al. (2009), l'évaluation automatique de traductions de phrases reste un problème ouvert : les meilleures métriques sont cohérentes avec les jugements humains dans 54% des cas, alors ligne basse est estimée à 50%.

Il semble aussi difficile d'identifier une technique d'évaluation automatique qui donnerait des résultats plus fiable qu'une autre. Par exemple, dans l'édition 2009 du *Workshop on Statistical Machine Translation* de Callison-Burch et al. (2009), les mesures les mieux corrélées sont plutôt des mesures combinant plusieurs mesures ou des mesures basées sur des correspondances entre structures sémantiques et syntaxiques. Dans l'édition 2010 du même workshop, décrite par Callison-Burch et al. (2010), les meilleures mesures sont celles qui emploient des informations de surface telles que des n-grammes de lettres, sachant qu'en plus, les évaluations sont faites sur des jeux de données quasi-similaires.

La stabilité du comportement de ces mesures "objectives" face aux données est aussi questionnable : les résultats de Callison-Burch et al. (2009, 2010) affichent d'importantes variations

[3]La F-mesure est la moyenne harmonique du rappel et de la précision

	Adéquation	**Fluidité**
5	tout le sens	anglais sans fautes
4	majeure partie du sens	bon anglais
3	une partie du sens	anglais non-natif
2	peu de sens	mauvais anglais
1	aucun sens	incompréhensible

TAB. 1 – *Échelles d'évaluation de l'adéquation et de la fluidité utilisées par Koehn et Monz (2006)*

dans les performances d'une même mesure selon le couple de langue, le sens de traduction ou le niveau de granularité de l'évaluation en jeu.

Les mesures d'évaluation objectives ont par ailleurs été critiquées par Blanchon et Boitet (2007) qui expliquent que ces dernières sont d'autant moins corrélées aux jugements humains que la qualité de la traduction augmente. Ils décrivent également une expérience consistant à faire évaluer des traductions automatiques post-éditées par des humains. Ces traductions sont jugées de qualité moindre que des traductions produites par des systèmes automatiques, et ce, sur la base de mesures telles que BLEU, NIST, etc. Les auteurs s'appuient sur cette expérience pour rappeler que ces mesures ne sont pas directement liées à la qualité des traductions mais qu'elles évaluent seulement la ressemblance avec une traduction de référence, sans compter que la référence est toujours discutable, tout particulièrement en TA.

3.1.2 Évaluation humaine de la TA

L'évaluation humaine consiste à présenter des traductions de phrases à des humains qui doivent alors juger de leur qualité. Cette méthodologie a évolué au cours des années. En 2006, Koehn et Monz (2006) demandent à des juges de donner deux notes aux traductions, sur une échelle de 1 à 5 donnée dans le tableau 1 [4] : l'une concerne l'adéquation entre traduction et texte d'origine (conservation du sens) et l'autre concerne la fluidité (bonne formation grammaticale). L'annotation des traductions se fait via une interface. Chaque juge peut voir le texte d'origine et annote cinq traductions à la fois, de façon à lui permettre de contraster les phrases et obtenir un meilleur jugement.

En 2007, Callison-Burch et al. (2007) testent deux autres méthodes :

classement des phrases : les juges doivent ordonner les phrases de la moins bien à la mieux traduite (avec la possibilité d'égalités)

classement de constituants syntaxiques : même principe que le classement des phrases, sauf qu'il s'applique à des traductions de syntagmes

Ce système de classement a été rajouté car il s'est avéré que les échelles d'adéquation laissent beaucoup de place à l'interprétation. Par exemple, il est difficile de cerner la valeur de *majeure partie du sens* (« *much meaning* ») dans l'échelle d'adéquation. De plus, les juges ont

[4] Les échelles ont été traduites. Originellement, on a : *all meaning, most meaning, much meaning, little meaning, none* pour l'adéquation et *flawless English, good English, non-native English, disfluent English, incomprehensible* pour la fluidité

	accord inter-annotateur	accord intra-annotateur	temps moyen par élément (secs)
fluidité	0,25	0,54	26
adéquation	0,23	0,47	26
classement des phrases	0,37	0,62	20
classement des constituants	0,54	0,74	11

TAB. 2 – *Accord intra- et inter- annotateur, temps d'annotation lors du 2007 Workshop on Statistical Machine Translation - Callison-Burch et al. (2007)*

du mal à noter séparément l'adéquation de la fluidité. A l'inverse, le classement, qui ramène l'évaluation à une simple comparaison, est plus simple à appréhender.

Les deux méthodes ont été comparées en mesurant le degré d'accord inter- et intra- annotateurs. La mesure utilisée est le $Kappa$ de Carletta (1996). Comme indiqué dans le tableau 2, la méthode de classement obtient un accord intra- et inter- annotateur plus élevé. De plus, elle permet une annotation plus rapide. Le classement des constituants syntaxiques est lui même plus fiable et plus rapide que le classement des phrases.

Dans l'édition 2008 du workshop, Callison-Burch et al. (2008) abandonnent la méthode d'évaluation basée sur l'adéquation et la fluidité. A la place, ils proposent une méthode plus simple, dans laquelle on présente aux juges des traductions de constituants syntaxiques et on leur demande d'indiquer si la traduction est acceptable ou pas. Les juges ont aussi la possibilité d'indiquer qu'ils ne sont "pas sûrs". Cette méthode a obtenu le plus haut taux d'accord : 0,64 et 0,86 - respectivement inter et intra annotateur. Finalement, dans les éditions 2009 et 2010, seule la méthode consistant à classer à été gardée. Une ultime méthode à été ajoutée, mais cette dernière sert uniquement à évaluer l'intelligibilité des traductions.

On voit que toute la difficulté de l'évaluation humaine touche à sa subjectivité et à son manque de reproductibilité, puisque, comme le montre les Kappa, une même traduction n'est pas toujours jugée de la même façon par les juges, ce qui peut faire douter de la fiabilité de ces jugements. La solution consiste alors à juger la traduction sur la base d'un grand nombre de jugements, ce qui permet de neutraliser les différences individuelles. Blanchon et Boitet (2007) remarquent que les juges ont tendance à devenir plus sévères sur la durée, ils indiquent aussi que le fait de former les juges augmente le taux d'accord. La préparation en question consiste à fournir aux juges une fiche d'instruction et à effectuer une première évaluation à blanc. Les divergences sont ensuite discutées afin de normaliser la notation.

3.2 L'évaluation en traductologie

En traductologie, la question de l'évaluation est en elle-même un sous-domaine. Williams (2004) y réfère par le vocable *Appréciation de la Qualité des Traductions* (AQT) - *Translation Quality Assessment* (TQA) en anglais. L'AQT trouve ses origines dans la critique de la traduction, activité qui consiste à commenter la qualité littéraire du texte traduit, avec ou sans référence au texte original. La discipline se développe dans les années 70 où la traductologie souhaite se doter de modèles avec un double objectif : donner à l'industrie de la traduction

		Nombre maximal de défauts dans une tranche de 4000 mots	
Côte	**Qualité**	**Défauts graves**	**Défauts mineurs**
A	supérieure	0	0 à 6
B	acceptable	0	7 à 12
C	à revoir	1	13 à 18
D	innaceptable	1 et +	18 et +

TAB. 3 – *Grille d'évaluation du modèle Sical - Larose (1998); Williams (2004)*

des moyens de contrôler la qualité de ses produits et permettre aux écoles de traduction d'évaluer leurs élèves. L'évaluation en traductologie est différente de l'évaluation en TA a plusieurs niveaux :

- le niveau d'exigence est supérieur : on évalue des traductions faites par des professionnels, et non pas la ressemblance avec une traduction humaine.
- la TA évalue les traductions en relation à d'autres, le but étant de classer des traductions de façon à classer les systèmes qui les ont produites ; la traductologie évalue les traductions en elles-mêmes, il ne s'agit pas de comparer les traducteurs professionnels entre eux
- la TA utilise une traduction professionnelle comme référence, la traductologie n'a pas de référence de qualité, le juge lui-même est la référence.

On trouve un panorama de l'AQT dans les articles de Williams (2004) et de Secară (2005). Larose (1998) propose une réflexion théorique sur la méthodologie de l'évaluation des traductions. Williams (2004) distingue deux types de modèles : les modèles quantitatifs et les modèles non-quantitatifs. Les modèles quantitatifs (section 3.2.1) sont plutôt pragmatiques, ils doivent permettre de donner un score de qualité à toute traduction. Ces modèles produisent des grilles d'évaluation utilisées dans l'industrie de la traduction ou dans l'enseignement. Les modèles non quantitatifs (section 3.2.2) constituent plutôt des approches théoriques du problème de l'évaluation et se concentrent surtout sur la définition de ce qu'est une "bonne" traduction.

3.2.1 Modèles quantitatifs et grilles d'évaluation

La plupart des modèles quantitatifs ont été conçus par et pour des organisations - gouvernementales ou commerciales - qui cherchaient un moyen de maîtriser la qualité de leurs traductions. Le premier modèle d'AQT a été crée par le Bureau de la traduction du Canada en 1976. Ce modèle, appelé Sical (Système canadien d'appréciation de la qualité linguistique) est décrit par Williams (2001) et Secară (2005). Il sépare erreurs de langue (intelligibilité, grammaticalité, idiomaticité) et erreurs de transfert (conservation du sens). Chaque erreur est jugée comme grave ou mineure, la gravité étant déterminée sur la base des conséquences supposées de l'erreur. La qualité globale de la traduction est estimée sur le nombre et le type d'erreurs rencontrées dans un passage de 4000 mots sélectionné aléatoirement (voir tableau 3).

La grille Sical a donné lieu à d'autres variantes par la suite. De même, diverses grilles d'évaluation ont été proposées par des organismes comme l'ATA (American Translators Association), la SAE (Society of Automotive Engineers) et le LISA (Localization Industry Standards Association) ou l'agence de traduction ITR.

Toutes ces grilles d'évaluation suivent le même schéma : elles consistent en une typologie d'erreurs de traduction, chaque type d'erreur étant associée à un poids représentant sa gravité. Certaines, comme le SEPT - décrit dans Larose (1998) - vont jusqu'à dénombrer 675 types d'erreurs. Larose (1998) remarque aussi à juste titre que tous les modèles séparent erreurs de transfert (sens ou contenu) et erreurs de langue (forme ou expression), avec une prédominance du sens sur la forme. On retrouve le même principe dans les premières versions de l'évaluation humaine de la TA, où l'on demande aux juges de noter séparément adéquation (erreurs de transfert) et fluidité (erreurs de langue).

En comparaison au domaine de la TA, on peut être surprise par l'absence de processus de validation ou de comparaison des différents modèles proposés. Bien que généralement conscients de la subjectivité des jugements humains, rien n'est fait pour tenter de la quantifier. On pourrait tout à fait envisager de comparer ces modèles sur la base d'un accord inter-annotateur. Il en est de même pour le coût en temps. Mise à part pour le Sical, qui est supposé prendre 1h pour évaluer un passage de 4000 mots, aucun auteur n'indique le temps que prend une évaluation en suivant telle ou telle grille.

Ces modèles quantitatifs, à visée opérationnelle, sont assez critiqués par les tenants des modèles théoriques, comme nous le verrons dans la partie suivante.

3.2.2 Approches théoriques

Une des principales critiques des approches théoriques envers les grilles d'évaluation utilisées dans l'industrie est le niveau d'analyse de ces grilles. En effet, la plupart les modèles quantitatifs restent au niveau des mots et de la phrase et se préoccupent rarement du niveau discursif. Les grilles d'évaluation sont monolithiques, supposées valables pour toutes les traductions, sans prendre en compte la fonction du texte, la situation de communication dans laquelle il a été produite ou les attentes du commanditaire de la traduction.

Williams (2004) par exemple, suggère de passer d'une approche micro-textuelle (celles des modèles quantitatifs, basée sur la phrase) à une approche macro-textuelle qui repose sur l'analyse et la comparaison de la structure argumentale des textes source et cible. S'appuyant sur la théorie de l'argumentation de S. Toulmin, il découpe chaque texte en six modules d'argumentation, qui sont indépendants du genre, type, fonction ou domaine du texte traduit. Dans ce cadre théorique, une bonne traduction est une traduction qui reprend chacun des modules présent dans le texte source et reproduit fidèlement leur contenu et relations. L'auteur considère l'absence d'un des modules comme une erreur majeure, mais ne donne pas plus de détails.

Reiss (1971) propose une approche fonctionnelle de la traduction. Elle affirme que les critères d'évaluation doivent dépendre de la fonction du texte. Pour cela, elle spécifie quatre types de textes :

textes centrés sur le contenu - « content-focused » Ce sont des textes dénotatifs, référentiels, qui privilégient la description de faits : articles de presse, travaux scientifiques, notices. Le traducteur adapte totalement la forme du texte à la langue cible, il amène le texte au lecteur, en respectant en priorité le sens du texte source.

textes centrés sur la forme - « form-focused » Ce sont les textes ayant une fonction poétique, par exemple les textes littéraires, artistiques. Le traducteur amène le lecteur au texte, en respectant en priorité la forme du texte source, le traducteur jouit d'une plus grande liberté au niveau du transfert du sens.

textes incitatifs - « appeal-focused » Ce sont les textes conatifs, destinés à provoquer une réaction chez leur lecteur : publicité, propagande. Dans ce cas, la traduction devient une adaptation libre : son but premier est de conserver l'effet du texte sur le lecteur, il n' y a pas d'obligation de respect de la forme ou du sens.

textes audio-médiaux « audio-medial » Ce sont les textes qui ne sont pas transmis par le support écrit : 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. Cette dernière catégorie semble assez bancale car elle se situe à un niveau de classification supérieur aux trois autres (oral vs. écrit) : un texte peut être incitatif et audio-médial (publicité radio vs. publicité sur affiche), centrés sur la forme et audio-médial (pièce de théâtre vs. œuvre littéraire), etc.

Comme tous les travaux, Reiss distingue sens et forme et donne des critères sur lesquels évaluer les traductions. Par contre, elle ne donne pas de grille d'évaluation à proprement parler. Elle différencie éléments linguistiques (sémantiques, lexicaux, grammaticaux, stylistiques) et extra-linguistiques (situation de communication, sujet, époque, lieu, audience, locuteur, enjeux affectifs). Chaque critère a une influence plus ou moins grande sur la qualité globale de la traduction, en fonction du type de texte traduit. Par exemple, dans le cas des textes orientés vers le contenu, l'équivalence totale entre éléments sémantiques du texte source et cible est obligatoire, alors que le non respect de l'équivalence stylistique est tolérable, voire recommandé si cela permet, en adaptant le texte source à la langue cible, un meilleur transfert du sens.

L'analyse de la question de la traduction, vu à travers deux domaines relativement éloignés que sont la TA et la traductologie, offre des pistes intéressantes pour la mise au point d'un protocole d'évaluation applicative des terminologies bilingues. Ces pistes sont exploitées dans la section suivante.

4 Protocole d'évaluation : considérations méthodologiques

La mise au au point d'un protocole d'évaluation de terminologies bilingues au travers d'une tâche de traduction spécialisée humaine soulève plusieurs questions :

- la fiabilité : comment obtenir à des résultats significatifs et une évaluation homogène et constante, voire reproductible ?
- la définition de la qualité : sur quels critères définit-on une bonne traduction ?
- d'une façon plus pragmatique, comment la traduction d'un texte spécialisé lorsque l'on n'a pas, à sa disposition, un expert du domaine (comme c'est bien souvent le cas) ?

Fiabilité des résultats Concernant la fiabilité des résultats, l'idéal serait de recourir à une méthode totalement reproductible, comme pour les évaluations objectives de la TA. Cette option, qui consisterait à utiliser une évaluation automatique, n'est pas intéressante. Certes l'évaluation à l'aide de mesures permet d'obtenir des scores reproductibles et invariants mais ceci ne concernerait que la seconde étape de l'évaluation. Or, l'évaluation globale des terminologies comprend aussi une tâche de traduction, qui elle, ne peut être faite que par des humains et n'est en aucun cas reproductible. L'apport de ce mode d'évaluation est donc très minime. De plus, comme vu plus haut, il s'agit d'une évaluation très indirecte de la qualité des traductions, considéré comme un substitut pratique mais imparfait de l'évaluation humaine. On

aura donc recours à une évaluation humaine en essayant de gérer au mieux possible la subjectivité inhérente à cette méthode. Une des solutions semble résider dans l'entraînement et la formation des évaluateurs. De même, le recours à plusieurs juges est un moyen de lisser les préférences individuelles. L'utilisation de mesures d'accord inter-juge de style Kappa, comme en TA, permettra de quantifier la fiabilité des jugements et de s'assurer qu'il existe un accord inter-annotateur suffisant.

Qualité des traductions Le sujet de la qualité de la traduction est plus ardu. Si on retrouve universellement les deux critères du sens et de la forme, il est difficile d'affiner plus avant la question. Dans le monde de la traduction, aucun barème ou un mode d'évaluation ne fait consensus. Et pour cause : en compilant les diverses grilles d'évaluation et travaux théoriques, on se rend compte que la qualité globale d'une traduction dépend de l'interaction complexe de nombreux paramètres linguistiques (orthographe, lexique, sémantique, style, structure argumentale) comme extra-linguistiques (lieu, époque, audience...). De plus, leur interaction et le poids de chaque paramètre seraient réglés par la fonction du texte et les attentes du commanditaire de la traduction. Pour reprendre l'expression de (Larose, 1998, p. 2), on se trouve face à un « *fol magma de variables variables* ».

Comment définir, dans ce cadre, la valeur ajoutée des terminologies bilingues ? S'attend-on à ce que ce qu'elles influent directement sur la qualité globale des traductions ou à ce qu'elles agissent uniquement sur quelques paramètres, qui à leur tour, influencent la qualité de la traduction ? Quels sont les paramètres les plus importants dans le cas d'une traduction spécialisée ?

Pour répondre à ces questions, nous poserons qu'une terminologie bilingue, lorsqu'elle est utilisée pour produire une traduction spécialisée, a pour but d'aider le traducteur lorsqu'il/elle bute sur un terme ou une expression propre au domaine de spécialité du texte. Deux cas deux figures sont possibles :

aide au décodage Il se peut que le sens du terme ou de l'expression soit opaque : la terminologie, étant enrichie d'informations extraites du corpus, donne accès aux contextes du terme, aux termes semblables, éventuellement à une définition. Toutes ses informations se conjuguent pour permettre au traducteur de cerner le sens du terme.

aide à l'encodage Il se peut que le traducteur comprenne le terme mais ne sache pas comment le traduire, i.e il ne connaît pas le terme consacré en langue cible : la terminologie propose alors des traductions candidates et chaque traduction candidate est assortie d'informations contextuelles permettant de faire le bon choix de traduction. Dans le cas où le traducteur a une intuition de traduction qui n'apparaît pas parmi les traductions candidates, le logiciel de gestion terminologique lui permet permet de chercher cette traduction potentielle dans le corpus duquel a été extraite la terminologie.

Les terminologies bilingues sont donc supposées agir sur les deux (méta-) critères de qualité que sont le transfert sens (décodage) et la production d'une forme adéquate (encodage). Nombre de paramètres de qualité sont cités par les travaux de traductologie, pourtant, les terminologies ne sont censées agir que sur quelques uns, par exemple l'orthographe, le respect des normes de vocabulaire, l'idiomaticité, l'interprétation correcte du terme source. On ne peut donc pas juger la valeur ajoutée des terminologies sur la base de la qualité globale de la traduction, puisque cette qualité dépend d'autres paramètres sur lesquels les terminologies ont

peu ou pas d'influence : grammaticalité, omissions / insertions, cohérence, respect de la structure argumentale, localisation, choix du registre... Nous définirons donc la valeur ajoutée des terminologies bilingues comme leur capacité à aider le traducteur à traduire des termes ou expressions spécialisées, dans le contexte de la traduction d'un texte. Par conséquent, la mesure de la valeur ajoutée sera calculée sur la base de la qualité de la traduction des termes ou expressions qui auront posé problème aux traducteurs et non sur la traduction d'un texte dans son ensemble. Nous aurons donc très probablement à évaluer la traduction de syntagmes, de composés syntagmatiques ou d'unités lexicales simples[5], comme l'on fait (Callison-Burch, Camerob, Koehn, Monz, and Schroeder 2008). En plus de permettre de mieux cibler l'évaluation, le recours à des segments inférieurs à la phrase aura également pour effet de réduire le temps d'annotation et de faciliter la tâche aux juges humains, comme l'ont montré (Callison-Burch, Camerob, Koehn, Monz, and Schroeder 2008).

Comme la traduction est faite en dehors de toute finalité professionnelle, si ce n'est celle d'évaluer les lexiques, on ne cherche pas à mettre au point une grille d'évaluation qui, dans le style de l'AQT, associerait des points différents aux fautes d'orthographe, au manque d'idiomaticité, etc. On se contente de rester les critères généraux du sens et de la forme. Pour cela, nous nous conformons aux recommandations de K. Reiss, qui recommande de donner la priorité au sens plutôt qu'à la forme lorsqu'on évalue des traductions de textes centrés sur le contenu (dont les textes spécialisés). Nous utiliserons trois catégories pour juger de la qualité des traductions (voir aussi tableau 4), ce qui permet également de sortir d'une évaluation binaire correct/incorrect comme c'est le cas dans les évaluations en comparaison à une référence :

exact : le terme choisi est le terme de référence ou l'expression consacrée en usage dans le domaine

acceptable : il ne s'agit pas du terme ou de l'expression de référence mais le traducteur est quand même parvenu à donner une équivalence sémantique : le sens est conservé

faux : la traduction est incorrecte : le traducteur n'a pas compris le terme et/ou il n'est pas parvenu à donner un équivalence sémantique

	transfert du sens	respect de la forme
exact	+	+
acceptable	+	-
faux	-	-

TAB. 4 – *Critères pour juger la qualité des traductions*

Expertise sur le domaine de spécialité Le fait de travailler avec des textes spécialisés rajoute un obstacle supplémentaire à l'évaluation. En plus de maîtriser les langues de départ et d'arrivée, le juge doit aussi être expert dans les domaines de spécialité des textes à traduire. En l'absence d'expert disponible, la solution retenue consiste à employer des textes spécialisés qui existent en langue source et en langue cible et qui ont été produits par un expert du domaine. Les résumés d'articles scientifiques constituent une ressource parfaite pour pallier l'absence

[5]Composées d'un seul mot.

de juge-expert. Dans le cas de résumés d'articles, la référence n'est pas une traduction, avec les risques d'infidélité au texte source que cela implique, mais bien une deuxième version du texte, produite par une même personne dans une autre langue. Le fait que l'auteur soit un expert du domaine assure sa légitimité en termes de choix terminologiques. Les articles étant nécessairement révisés avant publication, cela garantit que des faute de langue éventuelles ont été corrigées. En plus d'une traduction de référence, les juges peuvent aussi s'aider d'une base terminologique qui peut leur permettre de valider les cas où le traducteur n'a pas employé le terme attendu mais une variante ou une expression de sens équivalent. Les traductions à juger sont toujours montrées en contexte : le juge a accès aux phrases source et cible qui contiennent le terme ainsi qu'aux documents d'origine.

Situations de traductions En dernier lieu, nous rappellerons que la mise au jour de la valeur ajoutée des terminologies se fait par contraste : on compare le résultat de la traduction d'un même texte source traduit à l'aide de ressources linguistiques différentes. On appelle, dans la suite de l'article, ces situations où un texte est traduit à l'aide d'une ressource donnée, une *situation de traduction*. Idéalement, il convient d'employer au moins trois situations de traduction. D'abord, une situation servant de *ligne basse*, c'est-à-dire que les traductions sont faites à l'aide de ressources minimales, une sorte de "kit de survie" du traducteur. Dans ce cas, on considère que les traductions seront également de qualité minimale : il ne faut pas descendre sous ce seuil de qualité. Ensuite, une situation servant de *ligne haute*, c'est-à-dire que les traductions sont faites grâce à un maximum de ressources, on considère alors qu'il est impossible d'obtenir de meilleures traductions. Enfin, la dernière situation utilise la ressource évaluée : la qualité des traductions produites dans cette situation est comparée aux traductions produites dans la situation ligne basse et la situation ligne haute. On estime alors que les différences de qualité de traduction sont dues aux ressources qui aident plus ou moins le traducteur en lui fournissant, ou non, les traductions correctes des termes sur lesquels il/elle bute.

Toutefois, pour que toutes choses soient égales par ailleurs, il conviendrait que les textes soient traduits par une même personne. Or, ceci pourrait, paradoxalement, biaiser les résultats. En effet, lorsqu'un traductueur traduit un texte à l'aide d'une ressource donnée, il/elle garde forcément en mémoire une partie des traductions des termes sur lesquelles il/elle a buté. Si ce même traducteur doit ensuite retraduire le texte en question (ou un texte du même domaine de spécialité) dans une autre situation, il/elle réutilisera forcément les traductions apprises lorsqu'il/elle a traduit le texte pour la première fois. La deuxième situation de traduction est alors favorablement avantagée. Il faut donc faire en sorte qu'un traducteur ne traduise jamais des textes issus d'un même domaine dans des situations de traduction différentes. En retour, si les différentes situations de traductions sont jugées sur la base de traductions faites par différents traducteurs, comment s'assurer que les différences dans la qualité des traductions résultent bien dans de différences dans la qualité des ressources linguistiques et non pas de l'expertise des traducteurs ? Pour résoudre ce dilemne, on met en place une sorte de roulement qui permet que chaque situation de traduction soit jugée sur la base de textes traduits par différents traducteurs et qu'un traducteur ne traduise jamais deux fois le même texte ou deux fois des textes d'un même domaine. Ceci est schématisé dans le tableau 5. Idéalement, il faudra multiplier le nombre de traducteurs et de domaines de façon à collecter plus de données et augmenter la représentativité de l'évaluation.

	textes du domaine 0	textes du domaine 1	textes du domaine 2
traducteur 0	situation 0	situation 1	situation 2
traducteur 1	situation 1	situation 2	situation 0
traducteur 2	situation 2	situation 0	situation 1

TAB. 5 – *Roulement entre situations de traductions, traducteurs et domaines de spécialité des textes*

Les choix méthodologiques étant argumentés, nous décrivons dans la section suivante une première expérience d'évaluation applicative des terminologies bilingues.

5 Expérimentation

Cette partie décrit une première expérimentation de la méthode d'évaluation. Cette expérimentation vise uniquement à éprouver le protocole et si possible, mettre au jour d'éventuelles incohérences et difficultés de mise en œuvre. Les données utilisées sont décrites en section 5.1, le déroulement de l'évaluation est exposé en section 5.2

5.1 Données

Textes à traduire et corpus d'acquisition des terminologies 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 6). La thématique SCIENCES DE L'EAU est nettement moins précise que la thématique CANCER DU SEIN mais le corpus d'acquisition est plus volumineux (400k mots vs. 2M mots par langue). Les langues sont le français et l'anglais. Les corpus d'acquisition sont constitués uniquement de publications scientifiques pour le corpus SCIENCES DE L'EAU et d'un mélange de publications scientiques et de textes de vulgarisation pour le corpus CANCER DU SEIN. Les textes donnés à traduire sont équitablement répartis entre discours scientifique (résumés d'articles scientifiques) et discours vulgarisé (pages de sites Web bilingues).

Terminologies évaluées Deux terminologies ont été évaluées, l'une est extraite du corpus comparable CANCER DU SEIN, l'autre est extraite du corpus comparable SCIENCES DE L'EAU. La procédure d'acquisition est la suivante :

- Les termes sont extraits à l'aide du logiciel *Similis* de Planas (2005).
- Les termes, ainsi que chacune des lexies de catégorie nom, adjectif, adverbe, verbe 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, collocations, concordancier et si possible une définition.
- La terminologie est importée dans l'interface de consultation conçue par Delpech et Daille (2010), cette interface offre des fonctionnalités de recherche plein-texte dans le corpus d'acquisition.

	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. 6 – *Données : corpus d'acquisition des terminologies et textes à traduire*

Situations de traductions Nous avons comparé trois situations de traductions.

Situation 0 ou *ligne basse* Les textes sont traduits sans aucune ressource spécialisée. Le traducteur a uniquement accès à trois ressources génériques :
- Le Larousse bilingue français/anglais[6] et anglais/français[7]
- Le Larousse monolingue français[8]
- Le Cambridge monolingue anglais[9]

Situation 1 : terminologie issue de corpus comparables En plus des ressources génériques de la situation 0, le traducteur a accès à deux terminologies. L'une est extraite du corpus comparable SCIENCES DE L'EAU et l'autre du corpus comparable CANCER DU SEIN. Ces terminologies ont plusieurs caractéristiques :

Ambiguïté : chaque terme source est associé à 20 traductions candidates, ce sont des terminologies brutes, n'ayant subi aucune post-édition

Richesse : chaque terme est associé une entrée terminologique constituée automatiquement. Cette entrée stipule des informations comme la catégorie grammaticale, la fréquence, les collocations, les termes proches, parfois une définition

Outillage : les terminologies sont consultables dans une interface Web, cette interface permet également d'interroger le corpus duquel ont été extraits les termes : le traducteur a donc accès aux contextes du terme, il peut éventuellement tester une intuition de traduction et vérifier si elle est attestée dans le corpus

[6] http ://www.larousse.com/en/dictionaries/french-english
[7] http ://www.larousse.com/en/dictionaries/english-french
[8] http ://www.larousse.com/en/dictionaries/french/
[9] http ://dictionary.cambridge.org/

Situation 2 ou *ligne haute* En plus de ressources génériques de la situation 0, le traducteur a un accès total à Internet où il peut consulter les différentes ressources spécialisées, concordanciers, forums de traductions, etc. disponibles en ligne. Il peut aussi utiliser les moteurs de recherche pour contextualiser le terme à traduire ou vérifier une intuition. Cependant, on lui interdit les sites dont sont extraits les textes à traduire et les corpus d'acquisition ainsi que le site de la base de données terminologique TERMIUM qui est utilisée plus tard lors de l'évaluation des traductions.

Traducteurs et juges Disposant de peu de moyens humains (3 personnes : deux traducteurs professionnels et une personne non formée à la traduction) pour expérimenter le protocole, nous avons dû faire quelques entorses méthodologiques : il y a eu des collisions entre les rôles d'organisateur/traducteur et traducteur/juge. Le traducteur non professionnel est l'auteure de l'article et a aussi organisé l'évaluation. Les traducteurs professionnels étaient des étudiants de dernière année d'école de traduction, effectuant leur stage de fin d'année. 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. Aucun des traducteurs n'est familier avec les thématiques des sciences du l'eau ou du cancer du sein.

Répartition des textes et situations de traduction entre traducteurs Nous avons utilisé la personne non formée à traduction pour traduire uniquement dans la situation 0, qui est censée produire les moins bonnes traductions (la *ligne basse*). Les deux autres traducteurs ont traduit alternativement dans les situations 1 et 2.

	textes CANCER DU SEIN	textes SCIENCES DE L'EAU
situation 0 - ligne basse	non professionnel	non professionnel
situation 1 - terminologie corp. comp.	professionnel 1	professionnel 2
situation 2 - ligne haute	professionnel 2	professionnel 1

TAB. 7 – *Répartition des textes et situations de traduction entre traducteurs*

5.2 Déroulement de l'évaluation

L'évaluation s'est déroulée en deux phases : phase de traduction et phase d'évaluation de la qualité des traductions.

Phase de traduction Chaque traducteur reçoit les six textes à traduire accompagnés de l'instruction suivante :

> *Traduisez chaque texte selon la situation de traduction spécifiée. Indiquez le temps que vous avez mis pour traduire chaque texte. Une fois la traduction finie, listez les termes ou expressions qui vous ont posé problème. Indiquez quelles ressources vous avez utilisées pour trouver la traduction et notez la traduction finalement retenue.*

6	***mammogram***
VG-3	*Research has shown that women who have regular mammograms are more likely to survive breast cancer.*
7	***mammographie***
VG-3	*La recherche indique que les femmes qui passent régulièrement des mammographies sont plus susceptibles de survivre au cancer.*

ID	traduction	rang	exact	acceptable
8	mammogram	0	0	0
9	mammograph	0	0	0
10	mammograph	0	0	0

FIG. 1 – *Exemple de traductions de termes à annoter*

Chaque situation est décrite précisément au traducteur comme en section 5.1. La traduction se fait de langue seconde vers la langue maternelle du traducteur, dans notre cas, de l'anglais vers le français. Une fois les textes traduits, on collecte tous les termes relevés comme problématiques et la traduction retenue par le traducteur. Pour l'évaluation, on ne garde que les termes problématiques communs à au moins deux situations de traductions. Dans notre cas, nous avons collecté 148 groupes de termes problématiques (61 pour la thématique SCIENCES DE L'EAU et 87 pour la thématique CANCER DU SEIN).

Phase d'évaluation de la qualité des traductions Les deux juges notent la qualité des traductions des termes. Ils sont aidés par une traduction de référence, qui correspond au terme trouvé dans la version cible du texte. Terme source et traduction de référence sont contextualisés, c'est-à-dire présentés dans leur phrase d'origine. Les traducteurs ont aussi accès aux documents d'origine source et cible. Les juges peuvent recourir, en plus de la traduction de référence, à la base de données terminologique TERMIUM[10]. Les traductions sont anonymisées et mélangées aléatoirement, de façon à ce que le juge ne puisse pas savoir dans quelle situation ont été traduits les termes. Le tout est fourni dans un fichier tableur, où chaque groupe de traductions se présente comme dans la figure 1.

Les juges effectuent deux tâches d'évaluation :

- une tâche de classement : ils ordonnent les traductions de la meilleure à la moins bonne (les égalités sont autorisées)
- une tâche de jugement : ils notent séparément la qualité de chaque traduction selon les critères définis plus haut (exact, acceptable, faux).

Afin d'homonégéiser un maximum l'évaluation, des instructions d'annotation détaillées et quelques exemples d'annotations sur des cas difficiles ont été fournis aux juges. Compte-tenu du petit nombre de données (seulement 148 groupes de termes problématiques), nous n'avons pas procédé à une première évaluation "à blanc" qui aurait permis d'améliorer encore plus l'homogénéité de l'évaluation. Cette technique est courante dans le cadre de campagnes d'évaluation comme l'on fait, par exemple, Blanchon et Boitet (2007) dans leurs expériences.

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

6 Résultats

Nous détaillons ci-après les résultats obtenus suite à l'expérimentation de notre protocole d'évaluation.

Impressions de traducteurs Tout d'abord, il faut noter que la phase de traduction n'a pas été aisée. Les textes spécialisés se sont révélés très durs à traduire et, bien que les retours soient bons sur l'interface en elle-même, la qualité de la terminologie a été fortement critiquée par les traducteurs. 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".*

Il s'avère que les traducteurs s'attendaient à trouver directement la traduction d'un terme en le tapant dans le champ de recherche. Ils n'étaient pas suffisamment préparés à l'utilisation d'une terminologie ambiguë. Il a donc fallu leur suggérer d'autres façons de trouver des traductions, notamment d'exploiter les informations secondaires présentes dans les fiches terminologiques et la recherche plein-texte.

Ensuite, on note un problème lié à la couverture des terminologies. Celles-ci étaient loin d'être suffisantes pour permettre de traduire aisément les textes donnés. Comme l'indique les traducteurs, beaucoup de termes ou d'expressions nécessaires à la traduction des textes étaient absents des terminologies. On peut avoir une idée générale de la couverture des terminologies en calculant le rapport entre le nombre de mots communs aux textes à traduire et à la terminologie et le nombre de mots des textes à traduire :

$$couverture = \frac{|mots\ terminologie| \cap |mots\ textes|}{|mots\ textes|}$$

Le tableau 8 donne la couverture des terminologies dans chacune des thématiques. On y 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)	0,94	0,14
traductions de référence (FR)	0,67	0,78

TAB. 8 – *Couverture des terminologies par rapport aux textes à traduire et leurs traductions*

Utilisation des ressources Les traducteurs ont noté, pour chaque terme problématique, les ressources auxquelles ils avaient eu recours pour traduire le terme. Trois sources d'information ont été distinguées :

ressources génériques : les dictionnaires de langue générales bilingues et monolingues

ressources spécialisées : les terminologies issues de corpus comparables pour la situation 1, Internet pour la situation 2

intuition : recours à des heuristiques intuitives comme l'adaptation du terme source à la graphie de la langue cible, par exemple, la traduction de *sensitivity* par *sensitivité*.

Une traduction a pu être produite à l'aide de plusieurs sources, par exemple, en traduisant un terme mot à mot à l'aide de la ressource générale, puis en cherchant une attestation de la traduction candidate dans la ressource spécialisée.

Le tableau 9 montre des modes d'utilisation des ressources nettement différents d'une situation à l'autre. Dans les situations 1 et 2, les traducteurs ont eu très peu recours aux ressources générales : plus la ressource spécifique est large (terminologie issue de corpus comparable < Web) , plus elle a joué un rôle dans la production de la traduction, et moins l'intuition ou les ressources générales ont été utilisées.

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

TAB. 9 – *Utilisation des ressources en fonction des situations de traduction*

Temps de traduction Les six 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.

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).

Tâche de jugement Nous avons évalué séparément les traductions de la thématique CANCER DU SEIN et les traductions de la thématique SCIENCES DE L'EAU.

Thématique CANCER DU SEIN Le tableau 10 donne les jugements pour les traductions de la thématique CANCER DU SEIN. On y voit que la proportion de traductions jugées fausses est quasi-équivalente dans les trois situations. On voit que la terminologie issue du corpus comparable a permis d'augmenter nettement le nombre de traductions jugées exactes par rapport à la situation 0 (ligne basse) : 43% contre 38%. La situation 2 (ligne haute) est celle qui a permis de produire les meilleures traductions (47% de traductions exactes).

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

TAB. 10 – *Jugement de la qualité des traductions - thématique* CANCER DU SEIN

Thématique SCIENCES DE L'EAU Le tableau 11 donne les jugements effectués sur les traductions de la thématique SCIENCES DE L'EAU. On y voit que les traductions produites avec la terminologie 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 la terminologie auraient dû être au moins aussi bonnes que celles produites sans ressource spécialisée.

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

TAB. 11 – *Jugement de la qualité des traductions - thématique* SCIENCES DE L'EAU

Ces résultats s'expliquent par les raisons suivantes :

– comme le montre le tableau 9 sur l'utilisation des ressources, les traducteurs qui ont traduit dans la situation 1 se sont surtout servi de la terminologie et de leur intuition et ont peu utilisé les ressources générales pour chercher des traductions. Une utilisation plus systématique des ressources générales aurait sûrement mené à des résultats au moins aussi bons.
– pour la thématique SCIENCES DE L'EAU, la terminologie a une très faible couverture (seulement 14% des mots des textes cible se retrouvent dans la partie cible de la terminologie, contre 94% pour le corpus Cancer du sein), la terminologie n'a donc pas été d'une grande aide dans la traduction.

Tâche de 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 la thématique. Les traductions faites avec les terminologies sont meilleures que celles produites sans ressource spécialisée uniquement

dans la thématique CANCER DU SEIN et pas pour la thématique SCIENCES DE L'EAU, très probablement pour les raisons expliquées plus haut.

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

TAB. 12 – *Classement des traductions - thématique* CANCER DU SEIN

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

TAB. 13 – *Classement des traductions - thématique* SCIENCES DE L'EAU

7 Conclusion et perspectives

Après avoir discuté des limites des évaluations en comparaison à une référence, nous avons argumenté en faveur d'une évaluation applicative des terminologies bilingues. Un protocole d'évaluation applicative à été décrit puis expérimenté. 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 d'une terminologie dépend fortement de son 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. Nous avons mesuré cette adéquation entre textes à traduire et terminologie par un simple ratio. Une piste de recherche est d'améliorer cette mesure et éventuellement, de définir un seuil de couverture minimale.
- 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 : aucune ressource ou uniquement les ressources génériques,
  - situation 1 : avec les terminologies bilingues 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 et en incluant les remarques précédentes.

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 terminologies bilingues issues de corpus comparables. 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, corpus de documents confidentiels), contiendra toujours plus de solutions de traductions. Il faut donc inclure dans l'interface de consultation des terminologies bilingues 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.

Bourigault, D. et M. Slodzian (2000). Pour une terminologie textuelle. *Terminologies Nouvelles* (19), 29–32.

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.

Castellví, M. T. C., R. E. Bagot, et J. V. Palatresi (2001). Automatic term detection : A review of current systems. In *Recent Advances in Computational Terminology* (John Benjamins ed.)., pp. 53–88. Bourigault, Didier, Christian Jacquemin and Marie-Claude L'Homme.

Condamines, A. et J. Rebeyrolle (1996). Point de vue en langue spécialisée. *META 1*(42), 174–184.

Déjean, E. et E. Gaussier (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, 1–22.

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.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram Co-Occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California, pp. 128–145.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1), 61–74.

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.

Gaussier, E., D. A. Hull, et S. AÃŕt-Mokhtar. Term alignment in use : MachineAided human translation. In *Parallel Text Processing* (Kluwer Academic Publisher ed.)., pp. 253–274. London : VÃl'ronis, J.

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.

Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Méta : journal des traducteurs / Meta : Translators' Journal 43*(2), 163–186.

Lefever, E., L. Macken, et V. Hoste (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. Athens, Greece.

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.

Morin, E., S. Dufour-Kowalski, et B. Daille (2004). Extraction de terminologies bilingues à partir de corpus comparables. In *Actes, 11ème Conférence annuelle sur le Traitement*

Automatique des Langues Naturelles (TALN), Fès, Maroc, pp. 309–318.

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.

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.

Renders, M., H. Déjean, et E. Gaussier (2003). Assessing automatically extracted bilingual lexicons for CLIR in vertical domains : XRCE participation in the GIRT track of CLEF 2002. *Lecture Notes in Computer Science 2785/2003*, 363–371.

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

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, et J. Makhoul (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pp. 224–231.

Turian, J., L. Shen, et I. D. Melamed (2003). Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, New Orleans, USA, pp. 386—393.

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

Williams, M. (2004). *Translation quality assessment : an argumentation-centred approach*. University of Ottawa Press.

Summary

This paper argues in favor of an applicative evaluation of bilingual terminologies. It describes a protocol for the evaluation of such terminologies, which is experimented on terminlogies acquired from comparable corpora. The considered application is human specialized translation. The protocol consits in having specialized texts translated in various situations : without any specialized resource, with a terminology acquired from comparable corpora 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.