Analyse de la qualité des traductions automatiques du français vers l'anglais, d'Expressions Poly-Lexicales (EPL) à partir d'un corpus parallèle

Quelles sont les erreurs les plus fréquentes par type d'EPL ?

Emmanuelle Esperança-Rodier

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France ;
emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr

Résumé : Cette étude s'attache à évaluer la qualité d'un système de traduction probabiliste état de l'art, LIG-Moses (Potet, 2010), du français vers l'anglais, de la traduction d'expressions poly-lexicales, décrites selon Tutin *et al.* (2015), dans un document technique, à l'aide de la typologie d'erreurs de Vilar *et al.* (2006). Nous nous attachons à montrer si l'outil est utilisable par des apprenants pour la traduction d'EPL. De plus, les travaux de Hamilton (2015) ont montré que sur un corpus d'apprenants de L2, pour les erreurs d'acceptabilité textuelle, les erreurs de mise en phrase (i.e locutions figées, expressions idiomatiques, locutions semi-figée, collocations et unités multi-mots, ce qui correspond à la notion d'EPL que nous adoptons dans cet article) ont la plus haute fréquence soit 25,58%. Il s'agit donc d'étudier les types d'erreurs commis par l'outil de Traduction Automatique lors de la traduction d'EPL afin de guider les apprenants sur le choix de l'outil étudié pour la traduction d'EPL en fonction de leur niveau de connaissance de la langue. Après avoir décrit les différents éléments de notre étude, nous montrons que bien que traduites correctement à hauteur de 80%, les EPL qui n'ont pas été traduites correctement ne peuvent être corrigées que par des apprenants ayant un bon niveau de compétence dans la langue cible et une bonne connaissance du domaine du document.

Mots-clés : Évaluation de la qualité de traduction, Analyse d'erreurs de traduction, Expressions Poly-Lexicales, Corpus Parallèle.

1 Introduction

Depuis plusieurs années, l'évaluation des systèmes de traduction automatique (noté TA ci-dessous) est au cœur de nombreux travaux. Les plus connus proposent des outils permettant des évaluations automatiques telles que la métrique BLEU (Papineni *et al.*, 2002), qui vise à aider les développeurs afin d'évaluer leurs outils de manière rapide, peu coûteuse, et surtout avec une forte corrélation avec les jugements humains. Récemment, la dernière version de la métrique METEOR (Denkowski *et al.*, 2014) ou encore son amélioration par Servan (2016) proposent des caractéristiques similaires. D'autres études portent sur des critères d'évaluation tels que proposé par NIST (Doddington *et al.*, 2002), le FEMTI (King *et al.*, 2003), ou l'analyse d'erreurs de traduction suivant des critères linguistiques à partir de la typologie de Vilar *et al.* (2006). Plus récemment, le projet QT21 (2016) propose des métriques multidimensionnelles de qualité (MQM) alors que QuEst++ propose la prédiction de qualité de systèmes (Specia *et al.*, 2015). Dans le domaine de l'enseignement en traduction, l'analyse d'erreurs de traduction est notamment utilisée afin d'améliorer les méthodologies d'enseignement (Kübler et al, 2016) en utilisant la typologie d'erreurs MeLLANGE (Castagnoli *et al.*, 2011). Cette typologie se distingue des précédentes car elle a été "créée pour décrire les difficultés de traduction" et non pas pour évaluer la qualité d'un système de traduction ou de la traduction elle-même.

La majorité de ces travaux porte sur l'évaluation des systèmes de TA dans le but de les améliorer. Ils sont tous grandement tournés vers les développeurs afin qu'ils modifient les outils pour améliorer la qualité de leurs traductions. Lorsqu'il s'agit d'évaluer les outils selon des types d'erreurs à corriger, en termes de mesure de correction à mener pour fournir une meilleure traduction, les critères sont le plus souvent fortement liés au couple de langues utilisés ainsi qu'à la méthode de traduction. Le travail présenté ici, bien que basé sur l'analyse d'erreurs des sorties de traduction, ne se destine pas à évaluer des systèmes de TA pour viser à leur amélioration. Il s'agit plutôt d'évaluer la capacité de l'outil à fournir une traduction correcte de certains phénomènes linguistiques, ceci dans l'optique d'orienter les utilisateurs de ces traductions vers l'outil le plus adéquat à leur tâche de traduction et à leur compétence. Ainsi, cette étude s'intéresse à l'évaluation de la qualité de traduction des expressions poly-lexicales (notées ci-après EPL) dans le cadre de traductions automatiques destinées à être utilisées par des apprenants de langues. Pour un apprenant, l'utilisation d'outils de TA afin de produire des écrits de bonne qualité dans leur seconde langue vivante va être fortement liée aux compétences linguistiques qu'ils possèdent dans cette langue. Pour ce faire, le Cadre Européen Commun de Référence en Langues (CECRL, 2001) indique pour les compétences langagières ce qui est attendu par niveau de langue. Les niveaux de langue vont des niveaux débutants (notés A) en passant par les niveaux intermédiaires (notés B) pour terminer aux niveaux bilingues ou natifs notés C. La notion d'« expressions idiomatiques et de tournures courantes » n'est indiquée que dans la description du niveau C2 (niveau équivalent à bilingue). Pour un apprenant débutant (A2 du CECRL), savoir reconnaître une EPL n'est pas évident et repérer qu'une EPL est bien ou mal traduite l'est encore moins. Par conséquent, lors du choix d'outils d'aide à la rédaction ou à la traduction, l'apprenant débutant devra être guidé vers un outil sachant traduire correctement ces EPL. De plus, Hamilton

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

(2015) montre que sur un corpus d'apprenants de L2 de niveau de langue homogène, les erreurs de mise en phrase qui portent sur les locutions figées, expressions idiomatiques, locutions semi-figée, collocations et unités multi-mots, ce qui correspond à la notion d'EPL que nous employons, ont la plus haute fréquence pour les erreurs d'acceptabilité textuelle, soit 25,58%. Un quart des erreurs commises par les apprenants de L2 portent sur la traduction d'EPL. Ceci porte à croire qu'il sera difficile pour ces apprenants de corriger les erreurs de traduction des EPL puisqu'ils commettent ses erreurs eux-mêmes. L'étude décrite ici analyse la qualité des traductions des EPL par un système de Traduction Automatique Probabiliste afin de vérifier si ce type de système peut-être utilisé par des apprenants. Pour ce faire, dans la section suivante, après une explicitation de la méthodologie pour laquelle les éléments considérés comme des Expressions Poly-Lexicales, les outils de traduction et d'évaluation seront décrits, les résultats de l'évaluation seront présentés en fonction de chaque type d'EPL. Nous conclurons sur la possibilité d'utilisation de ce type de système par des apprenants sur ce point particulier qu'est la traduction des EPL.

2 Méthode et Outils

2.1 Expressions Poly-Lexicales

Les EPL étudiées dans cet article, sont décrites ci-après en reprenant les spécifications de Tutin *et al.* (2015). Tutin définit les EPL comme étant des « éléments multi-mots qui incluent plusieurs éléments graphiques, séparés par des espaces ou bien des tirets, ou bien séparés par d'autres mots ne faisant pas partie intégrante de l'EPL, par exemple : *pomme de terre, rendez-vous...* ». Tutin classe les EPL en neuf types. Parmi ces neufs types, nous nous concentrerons sur cinq d'entre eux, définis comme suit, à savoir, tout d'abord les mots fonctionnels, tels que *non seulement/even if* notés F dans le reste de l'article. Ces mots fonctionnels incluent les mots grammaticaux comme les conjonctions, les prépositions, les pronoms, les déterminants ainsi que les adverbes discursifs, modaux, intensifieurs ou de négation. Leur signifié n'est pas forcément clair et est principalement fonctionnel. Ensuite viennent les phrasèmes complets, par exemple *Peine de mort/Death penalty*, notés PH dans le texte. Le type PH comporte les EPL qui ne sont pas compositionnelles mais dont la signification ne peut être déduite de la signification des mots les composant. Ce type comporte également des mots, principalement des noms, qui font référence à un référent spécifique. L'exemple *Peine de mort/Death penalty* est facilement compréhensible et partiellement compositionnel mais comme il réfère à un fait spécifique il est considéré comme un phrasème complet. Le troisième type étudié est celui des collocations, comme *Gros fumeur/Heavy smoker*, notées C dans le corps de l'article. Le type Collocation comprend des expressions compositionnelles fréquentes, majoritairement des expressions binaires, dans lesquelles un mot, la base, garde son sens usuel alors que l'autre mot est plus imprévisible. En traduction, il s'agit souvent de traduction non littérale. L'avant-dernier type considéré ici est celui des entités nommées, comme *Université Grenoble Alpes*, notées EN, incluant les noms propres, les dates, les évènements, lieux, organisation. Et en dernier, le type termes complexes (notés T), qui sont souvent liés à un domaine précis, et qui peuvent être considérés comme un sous-type des phrasèmes complets, par exemple *Procédé de travail/Work process*. Ils sont principalement des

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

phrasèmes complets nominaux typiques des corpus spécialisés scientifiques ou professionnels. Les EPL contenues dans le corpus, que nous décrivons ci-dessous, ont été annotées semi automatiquement en type d'EPL, selon la description de Tutin que nous venons de reprendre.

2.2 Corpus

Le couple de langues étudié dans cet article est le couple français-anglais. Le corpus parallèle utilisé est composé d'un document rédigé en français d'environ 12 566 mots, issu du domaine technique (document BAF Citi 1 issu du corpus Baf) et de sa traduction en anglais, effectuée par un système de traduction probabiliste maison dont la description sera faite dans la section ci-dessous. Comme précisé précédemment, les EPL ont été annotées dans le corpus en fonction de leur type.

2.3 Outils de Traduction : LIG-Moses

Le système de traduction utilisé pour effectuer la traduction de notre document en français vers l'anglais est un système standard probabiliste basé sur des phrases où les unités de traduction sont des segments (séquences de n mots consécutifs). Tout d'abord, les corpus d'entrainement ont été alignés à l'aide de la boîte à outils GIZA++ (Och et Ney, 2003). Ensuite, les paires de phrases sources et les phrases cibles correspondantes ont été extraites des corpus d'entrainement bilingues alignés sur les mots en utilisant les scripts fournis avec le décodeur Moses (Koehn *et al.*, 2007). Le résultat est un tableau de phrases contenant tous les segments alignés. Cette table de phrases, produite par la modélisation de traduction, est utilisée pour extraire les quatorze fonctions par défaut, combinées dans un modèle log-linéaire pondéré. Ces pondérations peuvent être ajustées à l'aide du *Minimum Error Rate Training* – MERT (Och, 2003) qui minimise un critère d'erreur sur un corpus de développement. Néanmoins, cela n'a pas été fait sur le système de référence parce que cet ajustement détériore les résultats de performance de notre système sur les données de développement du WMT 2010.

Ce système de référence est conforme à l'état de l'art puisqu'il a été validé lors de la participation à la campagne d'évaluation internationale WMT 2010 en obtenant un score BLEU de 25,27. Il est décrit dans (Potet *et al.*, 2010) comme système (3). Le modèle de traduction a été entraîné sur un corpus composé du corpus Europarl (≈46 MWords) et du corpus News Commentary (≈2 MWords) contenant 1 640 463 énoncés dont deux préalablement normalisés. La phase de normalisation comprend la suppression de la casse, la tokenisation et la transformation du t euphonique. Le modèle de langue cible est un langage standard de 4-grammes.

2.4 Évaluation

2.4.1 BLAST

La traduction anglaise obtenue à l'aide de LIG-Moses décrit ci-dessus, a ensuite été évaluée à l'aide de BLAST (Stymne, 2011), un outil libre pour l'annotation des erreurs de traduction. Nous avons choisi cet outil parce qu'il a déjà été utilisé lors de précédentes évaluations de traductions d'EPL et qu'il utilise comme typologie d'erreurs, la

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

classification des erreurs de traduction de Vilar (2006), décrite dans la sous-section suivante, même si BLAST peut être utilisé avec toute autre classification hiérarchique des erreurs. Cet outil est également très bien adapté à toute fin d'évaluation car il n'est pas lié à l'information fournie par un outil de traduction automatique spécifique. De plus, BLAST est facile à utiliser grâce à son interface graphique. Nous avons vu dans cet outil d'évaluation la façon d'ajouter de nouvelles annotations, d'éditer des annotations existantes et aussi d'effectuer des recherches parmi les annotations.

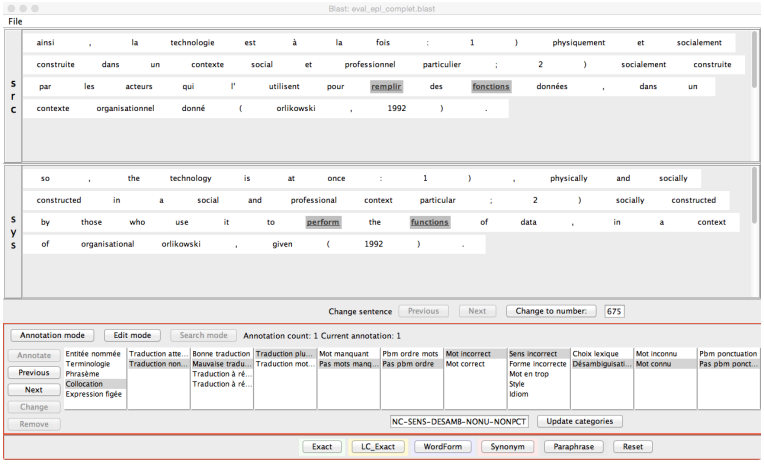


FIGURE 1: Capture d'écran de l'outil Blast

2.4.2 Typologie d'erreurs

La typologie d'erreurs de traduction de Vilar que nous avons utilisée pour l'évaluation, décrite dans la figure 2, identifie cinq catégories principales d'erreurs de traduction à savoir :

— Mots manquants, pour les mots qui n'ont pas été traduits,
— Ordre des mots, pour un mauvais ordre des mots dans la séquence traduite,
— Mots incorrects, pour une erreur de traduction,
— Mots inconnus, pour les mots restés dans la langue source,
— Ponctuation, lorsque les règles de ponctuation de la langue cible n'étaient pas respectées.

En ce qui concerne les deux premières catégories d'erreurs de traduction, les mots manquants et l'ordre des mots, des sous-catégories ont été créées pour affiner la classe d'erreur.

Pour la catégorie d'erreur Mots manquants, la distinction entre les mots signifiants et les mots outils permet de voir si le mot manquant était significatif ou non. Cette sous-catégorie illustre le fait que le sens complet de la phrase a été conservé ou non, ce qui est évidemment l'un des buts d'une évaluation de la qualité de traduction. En ce qui concerne la catégorie d'erreur Ordre des Mots, la sous-catégorie Mot ou Syntagme indique si l'erreur de traduction entraîne une réorganisation des mots eux-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

mêmes ou un ré-ordonnancement des segments dans la phrase. Il permet de localiser à quel niveau, lexical ou sémantique, le système a échoué.

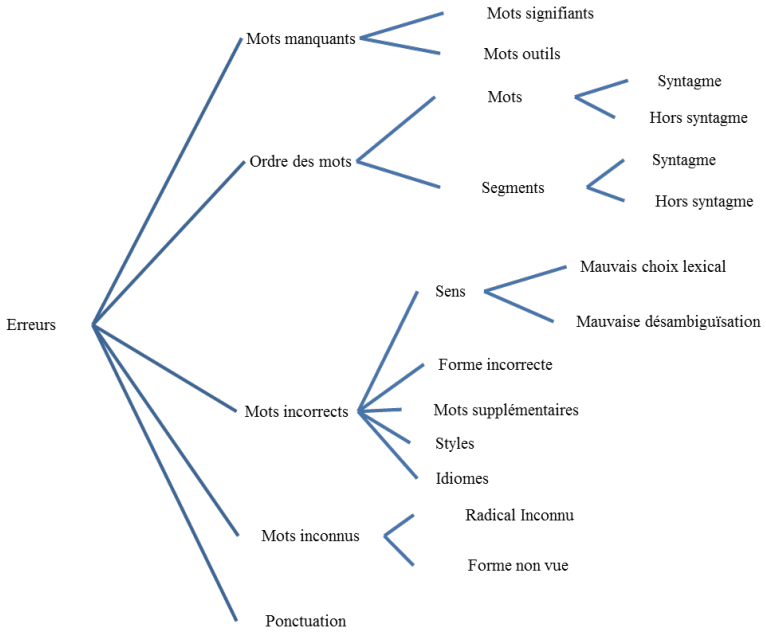


FIGURE 2: Typologie d'erreurs de Vilar *et al.* (2006)

En regardant la troisième catégorie d'erreur de traduction, Mots incorrects, nous pouvons voir qu'il existe plusieurs sous-catégories visant à distinguer la raison de l'erreur de traduction, ce qui peut être dû au fait que le système n'a pas été capable de désambiguïser correctement le sens d'un mot source ou de produire la bonne forme du mot, bien que le lemme du mot ait été bien traduit.

Pour la quatrième catégorie d'erreur de traduction, Mots inconnus, on peut distinguer si le lemme des mots était connu par le système ou non.

Et enfin, la cinquième catégorie d'erreur de traduction, qui est Ponctuation, n'a pas été utilisée dans cette étude.

Une fois la traduction évaluée en fonction de la typologie de (Vilar et al, 2006) que nous venons de décrire, nous avons pu observer les résultats décrits dans la section 3.

3 Résultats de l'évaluation

De manière globale, le système de traduction LIG-Moses traduit bien les EPL avec 80% d'entre elles traduites correctement.

Comme cet article s'attache à l'analyse d'erreurs, afin d'orienter les apprenants vers l'utilisation d'outils de traduction pour lesquels leurs compétences en langue cible leur permettra de corriger les erreurs de traduction, nous allons nous attacher dans

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

les sections suivantes à détailler les erreurs de traductions commises au niveau des EPL.

Dans cette section sur l'évaluation, nous parlons d'adéquation et de fluidité qui sont des mesures subjectives d'évaluation de la traduction automatique. La mesure de fluidité d'une traduction, notée entre 1 peu fluide et 5 très fluide, correspond à la correction grammaticale de la phrase. Il s'agit de déterminer si la phrase est bien formée et correspond au schéma grammatical de la langue. En ce qui concerne la mesure d'adéquation d'une traduction, notée sur une échelle identique à celle de la fluidité, allant de 1, peu adéquate à 5 tout à fait adéquate, la cohérence de sens entre la phrase traduite et la phrase à traduire est alors mesurée. Dans ce cas, c'est la préservation du sens initial qui est alors évaluée.

3.1 Erreurs rencontrées

La figure 3 ci-dessous montre par catégorie d'erreurs de traduction, représentée en abscisse, le pourcentage d'EPL annotées comme mal traduites en fonction de ces catégories d'erreurs.

En étudiant la figure 3, une première analyse quantitative est effectuée, suivie par une analyse de la répartition des types d'EPL par catégorie d'erreurs pour lesquelles des exemples sont donnés dans la table 1.

Nous remarquons tout d'abord, que les catégories d'erreurs annotées le plus fréquemment pour la mauvaise traduction des EPL sont les catégories MOTS INCORRECTS>SENS>MAUVAIS CHOIX LEXICAL et MAUVAISE DESAMBIGUÏSATION, soit respectivement 16% et un peu moins de 18% des annotations, ainsi que les MOTS MANQUANTS>SIGNIFIANTS à environ 11,5%. Puis, il est à noter qu'il n'y a pas d'erreurs de catégorie ORDRE DES MOTS>SYNTAGMES, ni de la catégorie MOT INCONNUS>FORMES NON VUES pour aucun type d'EPL. Ceci pourrait s'expliquer par le genre technique de notre texte source.

Nous pouvons déjà entrevoir qu'il sera difficile pour les apprenants des niveaux A et B du CECRL de corriger les EPL mal traduites au niveau du sens (mauvaise désambiguïsation et mots manquants signifiants), bien qu'ils puissent éventuellement corriger les mauvais choix lexicaux puisqu'il s'agit d'un souci de fluidité plus que d'adéquation, ainsi que les mots manquants. Cependant, en ce qui concerne les mots manquants, l'adéquation entrant en jeu, les apprenants auront certainement des difficultés de compréhension pour modifier de manière correcte les traductions. On peut donc prédire que l'outil n'aidera pas les étudiants des niveaux débutants à traduire correctement les EPL pour environ 20% des cas mal traduits.

Ensuite, nous nous intéressons à la répartition des EPL par catégorie d'erreurs. Nous constatons, dans un premier temps, que les erreurs de la catégorie ORDRE DES MOTS>MOTS ainsi que MOTS INCORRECTS>FORME ne concernent qu'un seul type d'EPL, les collocations (notées C). Dans la table 1, contenant des exemples d'erreurs par catégorie d'erreurs, l'exemple 1 montre que pour la catégorie ORDRE DES MOTS>MOTS, la traduction donnée par le système de TA et la traduction attendue sont très proches et que pour un apprenant de niveau A2 du CECRL, la

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

traduction du segment source travail de qualité par *work of quality* sera jugé comme acceptable alors que la traduction *quality work* est celle attendue. Pour le type MOTS INCORRECTS>FORME, l'exemple 2 de la table 1 indique une mauvaise construction de la forme verbale. Outre le fait que le temps employé n'est pas le bon, utilisation du *Present Perfect have [...] increased* alors qu'un présent simple *increase* était attendu, la forme de ce temps est mal construite. La traduction de la base verbale est effectuée à la troisième personne du singulier du présent *increases* alors qu'un participe passé était attendu, i.e. *increased*. L'apprenant doit connaître l'utilisation ainsi que la construction des temps en anglais afin de pouvoir modifier cette traduction d'EPL. De telles erreurs pourront être corrigées uniquement si la compétence en langue de l'apprenant lui permet de maîtriser l'utilisation des temps en anglais.

Dans un second temps, nous notons que les erreurs de la catégorie MOTS INCORRECTS>SUPPLEMENTAIRES n'apparaît que pour deux types d'EPL, les collocations (C $<$ 1%) et les mots fonctionnels (F$\approx$4%). L'exemple 3 de la table 1 montre une EPL fonctionnelle : au niveau de, traduite par l'amalgame de son expression correspondante en français, *in terms of*, et sa traduction mot à mot *the level of* pour donner la traduction erronée *in terms of the level of*. Les ajouts peuvent être repérés par tout apprenant, mais le choix entre la bonne traduction et la traduction mot à mot dépendra de son niveau de langue.

Dans un troisième temps, les erreurs des catégories MOTS MANQUANTS> SIGNIFIANTS, MOTS INCORRECTS>STYLE et MOTS INCONNUS>LEMMES ont été utilisées pour annoter les erreurs produites sur les classes d'EPL phrasèmes complets (PH), collocations (C) et mots fonctionnels (F). Dans l'exemple 4 de la table 1, nous pouvons remarquer que l'EPL *en cause*, n'a pas été traduite dans la phrase cible. De même que pour l'exemple 3 explicité ci-dessus, l'apprenant peut remarquer qu'il manque une partie de la phrase, mais corriger par la bonne traduction s'avère également relever de son niveau de langue. Pour l'exemple 5 concernant la catégorie d'erreurs MOTS INCONNUS>LEMMES, le mot *rancard*, dans l'EPL *mise au rancard*, n'a pas été traduit et a été laissé tel quel dans la traduction. Ceci est un cas typique des systèmes de traduction probabilistes qui, lorsqu'ils ne trouvent pas la traduction insèrent le mot de la phrase source. Il est intéressant de constater que l'élément non traduit porte lui-même à confusion dans la langue source car il peut s'écrire de manière différente pour ses deux sens mettre au rancart avec un T et non un D à la fin du mot, pour signifier mettre au rebus, et rancard avec un D à la fin du mot, signifiant rendez-vous ou renseignement. Les deux sens peuvent également s'orthographier rencard/rencart avec un E et non un A. Autant de confusion pour les systèmes de traduction probabilistes qui vont traduire en fonction de la forme de ce mot qu'ils ont rencontrée dans les corpus d'apprentissage.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

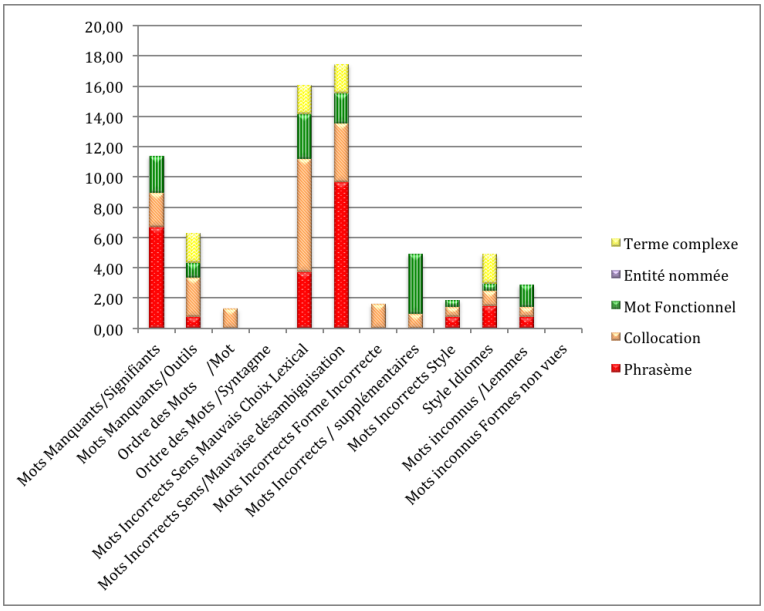


FIGURE 3: Pourcentage d'EPL étiquetées (axe des ordonnées) par catégorie d'erreurs (axe des abscisses)

Dans un quatrième et dernier temps, les erreurs des catégories MOTS MANQUANT>OUTILS et MOTS INCORRECTS>SENS>MAUVAIS CHOIX LEX, MOTS INCORRECTS>SENS>MAUVAISE DESAMBIGUISATION et STYLE>IDIOMES apparaissent lors de l'annotation des types d'EPL phrasèmes complets, collocations, mots fonctionnels ainsi que termes complexes. L'exemple 5 de la table 1, pour la catégorie d'erreurs MOTS INCORRECTS>MAUVAIS CHOIX LEXICAL, montre que le terme complexe *outils informatisés*, a été traduit par *computer tools*, alors que la traduction adéquate est *computerized tools*. Le dernier exemple de la table 1, l'exemple 6, indique que, pour le type STYLE>IDIOME, la traduction produite par LIG-Moses du terme *recherches appliquées* est *research carried out*, alors que *applied research* est le terme consacré en anglais.

Ces deux derniers exemples laissent supposer que pour les apprenants en langue de spécialité, le vocabulaire technique n'est pas forcément bien traduit et qu'il est donc à apprendre si l'on veut que l'apprenant puisse corriger ces traductions. De plus, la non correspondance terme à terme en langue de spécialité peut s'avérer une difficulté pour les apprenants. Ceci montre également l'importance de la connaissance du domaine en plus de la compétence en langue, afin de pouvoir repérer et corriger les erreurs sur les traductions des corpus techniques.

Maintenant que nous avons vu les catégories d'erreurs rencontrées en fonction du type d'EPL, et que nous avons pu dégager quelques observations sur l'utilisation de ce genre d'outils par les apprenants, nous allons, dans la section 3.2, observer les erreurs les plus fréquentes.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Tableau 1 : Exemple de traductions par catégorie d'erreurs

#	Type d'erreurs	Exemples
1	Ordre des mots>Mots	FR: […] une grande fierté d'un travail de qualité […] GB: *[…] *great pride of a work of quality* […] Attendu : […] *great pride of a quality work* […]
2	Mots incorrects>Forme	FR: […] on augmente considérablement les chances […] GB: *[…] *we have significantly increases the chance* […] Attendu : […] *we significantly increase the chance* […]
3	Mots incorrects>Supplémentaires	FR: […] au niveau de la communication interne […] GB: *[…] *in terms of the level of internal communication* […] Attendu : […] *in terms of internal communication* […]
4	Mots Manquants>Signifiants	FR: […] du domaine d'application en cause. GB: *[…] *the area of application.* Attendu : […] *the area of application involved.*
5	Mots Inconnus>Lemmes	FR: […] sa mise au rancard […] GB: *[…] *its implementation in rancard* […] Attendu : […] *being chucked out* […]
6	Mots Incorrects/ Mauvais Choix Lexical	FR: […] des outils informatisés leur permettant […] GB: *[…] *computer tools to enable* […] Attendu : […] *computerized tools to enable* […]
7	Style>Idiome	FR: […] résultats de recherches appliquées […] GB: *[…] *results of research carried out* […] Attendu : […] *results of applied research* […]

3.2 Erreurs les plus fréquentes

La Figure 4 ci-dessous montre par type d'EPL en abscisse, le pourcentage, en ordonnée, des catégories d'erreurs étiquetées pour ce type d'EPL.

Nous constatons que sur cinq types d'EPL, une seule n'a jamais été annotée comme comportant des erreurs : les entités nommées. Ceci n'est pas étonnant car typiquement un système de TA probabiliste retranscrit littéralement les mots qu'il ne peut traduire. Par conséquent, les entités nommées sont toujours bien traduites par ce genre de système.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

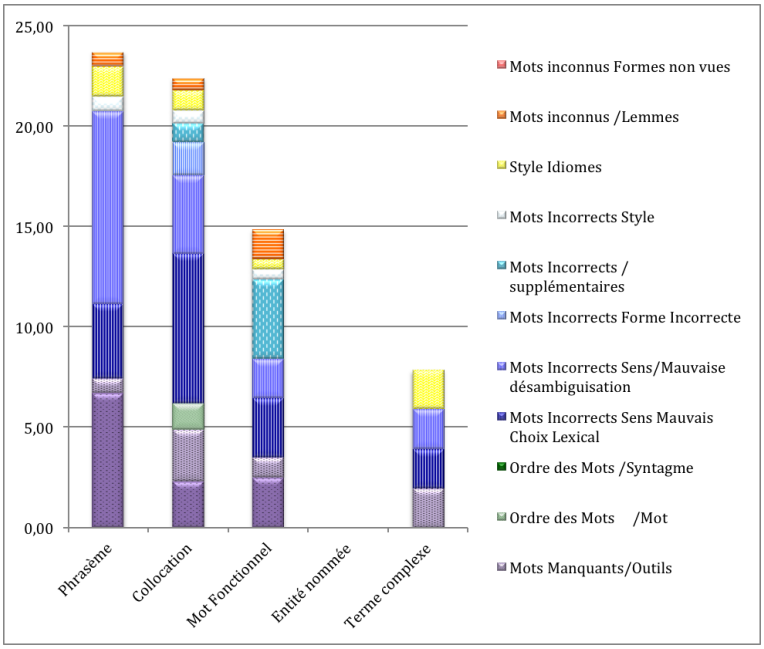


FIGURE 4: Pourcentage de catégorie d'erreurs (axe des ordonnées) par EPL (axe des abscisses)

Les quatre autres types d'EPL sont annotés pour plusieurs catégories d'erreurs chacune. Le second type d'EPL comportant le moins d'erreurs de traduction (un peu moins de 8% d'erreurs) est le type Termes Complexes. Les erreurs annotées pour ce type sont au nombre de quatre, présentes à égale mesure et sont MOTS MANQUANTS>OUTILS, MOTS INCORRECTS>SENS>MAUVAIS CHOIX LEXICAL, MOTS INCORRECTS>SENS>MAUVAISE DESAMBIGUISATION et STYLE> IDIOME. Ces résultats rejoignent ceux de la section précédente, et corroborent le fait que l'apprentissage du lexique lié au domaine du document à traduire est important afin qu'un apprenant puisse savoir si le Terme Complexe (T) traduit l'a été comme il se doit pour le domaine.

Lorsque l'on regarde la Figure 4, on s'aperçoit que l'on passe d'un peu moins de 8% pour le type T à presque 15% pour le type F, Mots Fonctionnels, soit le double d'erreurs. Huit catégories d'erreurs ont été associées à ce type d'EPL, principalement la catégorie MOTS INCORRECTS. On retrouve les quatre catégories d'erreurs précédentes, en moindre proportion cependant, pour les catégorie MOTS MANQUANTS>OUTILS et STYLE>IDIOME, dans approximativement les mêmes proportions pour MOTS INCORRECTS>SENS>MAUVAISE DESAMBIGUISATION, et pour un point de plus concernant la catégorie MOTS INCORRECTS>SENS> MAUVAIS CHOIX LEXICAL. Enfin, quatre catégories d'erreurs supplémentaires ont été utilisées pour annoter ce type, à savoir MOTS INCORRECTS>STYLE en très petite proportion, MOTS INCONNUS>LEMMES et MOTS MANQUANTS>

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

SIGNIFIANTS en petite proportion, et enfin MOTS INCORRECTS> SUPPLEMENTAIRES en proportion plus conséquentes.

Le type d'EPL collocations (C) a été annoté en fonction de dix catégories d'erreurs sur les quatorze catégories de Vilar. C'est le type d'EPL pour lequel le plus de catégories d'erreurs ont été utilisées. On retrouve les mêmes catégories d'erreurs pour les deux types d'EPL précédents, T et F. La catégorie d'erreurs principale est la catégorie MOT INCORRECTS>MAUVAIS CHOIX LEX et MOT INCORRECTS>MAUVAISE DESAMBIGUISATION. La catégorie MOTS MANQUANTS constitue également une partie conséquente des erreurs annotées pour le type d'EPL C. Cependant, c'est le seul type pour lequel les catégories d'erreurs ORDRE DES MOTS>MOTS et MOTS INCORRECTS>FORME INCORRECTE ont été utilisées pour l'annotation.

Finalement, le type phrasèmes complets (PH) contient le plus grand pourcentage d'erreurs environ 24%, pour seulement sept catégories d'erreurs. Il s'agit principalement des catégories d'erreurs MOTS INCORRECTS>SENS>MAUVAISE DESAMBIGUISATION comme dans l'exemple 8 ci-dessous :

FR : [...]sont encouragés à faire part[...]

GB : *[...]*are encouraged to take part*[...]

Attendu : [...]*are encouraged to share*[...]

où faire part n'a pas été traduit correctement par *to share* mais par *to take part* ce qui induit un contre-sens.

Ces résultats montrent que les collocations et les phrasèmes complets, qui relèvent plus de la langue générale, sont les plus difficiles à traduire pour un système de TA probabiliste. Ces résultats restent relatifs puisqu'ils sont mal traduits pour respectivement 23% et 24%, soit à peu prêt un quart des collocations et des phrasèmes complets présents dans le corpus à traduire. Cela signifie que LIG-Moses traduit correctement un peu plus de 75% des collocations et des phrasèmes. C'est un bon résultat, mais cependant de par la nature non compositionnelle du type PH, et du fait que souvent les collocations font l'objet de traduction non littérale, il est très difficile pour les apprenants de corriger ce genre d'erreurs. Lors d'un jugement subjectif de la qualité d'une traduction, le fait qu'un phrasème complet ou qu'une collocation est mal traduite impacte de manière négative le jugement de qualité de la traduction.

Par ailleurs, il ressort de cette étude que pour toutes les EPL, les catégories MOTS MANQUANTS et MOTS INCONNUS sont les plus fréquentes. Il existe des spécificités en fonction de la catégorie d'EPL, ainsi les catégories ORDRE DES MOTS>MOTS et MOTS SUPPLEMENTAIRES ont été utilisées uniquement afin d'annoter les collocations. La catégorie MOTS INCONNUS>LEMMES n'a été utilisée que pour les EPL phrasème complet, collocation et mots fonctionnels.

4 Conclusion et perspectives

Cette étude porte sur l'évaluation de la qualité pour un système standard de traduction automatique probabiliste, LIG-Moses, des traductions des EPL dans un texte

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

technique en fonction de la typologie de Vilar. Ces travaux montrent que les catégories d'erreurs les plus fréquentes pour l'annotation des erreurs sont les mots signifiants manquants dans la traduction (Mots manquants>Signifiants), la mauvaise désambiguïsation de lemme engendrant une traduction incorrecte (Mots incorrects>Mauvaise désambiguïsation) et enfin le mauvais choix lexical pour l'occurrence d'un lemme (Mots incorrects>Mauvais choix lexical).

Cette étude a mis en avant que les EPL les plus sujettes à erreur, par ordre décroissant, sont les suivantes :

— PhrasÈmes complets (PH)
— Collocations (C)
— Mots Fonctionnels (F)
— Termes COMPLEXES (T)

Il est à noter que du fait du comportement du système probabiliste, aucun problème n'a été remarqué pour la traduction des Entités Nommées (EN).

Le système LIG-Moses traduit correctement dans 80% des cas les EPL, ce qui laisse à penser que les systèmes probabilistes standards peuvent être proposés à des apprenants pour la traduction de documents contenant des EPL. Cependant dans les 20% d'erreurs de traductions sur les EPL, il s'avère que les erreurs commises sont difficiles à corriger par des apprenants, comme le suggèrent les travaux d'Hamilton (2015) puisque les apprenants de L2 commettent ses erreurs eux-mêmes. On peut donc en conclure qu'afin de corriger ces catégories d'erreurs, les apprenants doivent avoir une bonne connaissance du domaine ainsi qu'avoir un niveau de compétence élevé dans la langue cible. De plus, les types des EPL qui sont le plus mal traduits, induisent une bonne connaissance de la langue, niveau C2, comme l'indique le CECRL. Bien que ce système probabiliste standard traduise de manière satisfaisante les EPL, il reste encore des recherches à effectuer afin d'améliorer la traduction des types d'EPL les plus opaques à la compréhension que sont les phrasèmes complets et les collocations. Une étude sur les relations entre les erreurs commises par les apprenants en fonction de leur niveau du CECRL et les erreurs effectuées par les systèmes de TA reste également à mener.

Bibliographie

Castagnoli, Sara, Ciobanu Dragos, Kübler Natalie, Kunz, Kerstin, Volanschi, Alexandra (2011) Designing a Learner Translator Corpus for Training Purposes. Dans *Corpora, Language, Teaching and Resources : From Theory to Practice*. Édité par Kübler N. Bern : Peter Lang.

CECRL (2018) *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Édité par les Éditions Didier. 2001. ISBN 9782278058136.

Denkowski, Michael, and Lavie, Alon (2014) Meteor Universal : Language Specific Translation Evaluation for Any Target Language, Actes EACL 2014 Workshop on Statistical Machine Translation, 2014

Doddington, George (2002) Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. Dans *2nd Human Language Technologies Conference* (HLT-02). San Diego, CA. p. 128-132.

Hamilton, Clive, E. (2015) Cartographie des erreurs en anglais L2 : vers une typologie intégrant système et texte. Linguistique. Université Sorbonne Nouvelle – Paris 3, 2015. Français.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

King, Margaret, Popescu-Belis, Andrei and Hovy. Eduard (2003) FEMTI : Creating and Using a Framework for MT Evaluation. Dans *Machine Translation Summit IX*, New Orleans, LA. Sept. 2003. p. 224-231.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra and Herbst, Evan (2007) Moses : Open source toolkit for statistical machine translation. Dans *45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kübler, Natalie, Zimina, Maria, Fleury, Serge (2016) Origines des erreurs en Traduction Spécialisée : différentiation textométrique grâce aux corpus de textes cibles annotés. Dans *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 09 : ELTAL*, Paris.

Och, Franz Josef (2003) Minimum Error Rate Training in Statistical Machine Translation. Dans *41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan. 160–167.

Och, Franz Josef, and Ney, Hermann (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29 :19–51.

Papineni, Kishore Roukos, Salim, Ward, Todd and Zhu, Wei-Jing (2002) BLEU : a Method for Automatic Evaluation of Machine Translation. Dans *40th Annual Meeting of the Association for Computational Linguistics* (ACL-02). Philadelphia, PA. July 2002. p. 311-318.

Potet, Marion, Besacier, Laurent, Blanchon, Hervé (2010) The LIG machine translation system for WMT 2010. Dans *ACL Workshop on Machine Translation (WMT)*, 2010, Upsala, Sweden

QT21 : Quality Translation 21 : http ://www.qt21.eu/. (Consulté le 25 septembre 2016).

Servan, Christophe, Bérard, Alexandre, Elloumi, Zied, Blanchon, Hervé, and Besacier, Laurent (2016) Word2Vec vs DBnary : Augmenting METEOR using Vector Representations or Lexical Resources ?. Dans *International Conference on Computational Linguistics* (COLING 2016), Dec 2016, Osaka, Japan. 26th, 2016.

Specia, Lucia, Paetzold, Gustavo Henrique, et Scarton, Carolina (2015), Multi-level Translation Quality Prediction with QuEst++. Dans *ACL-IJCNLP 2015 System Demonstrations*, Beijing, China, 115-120.

Stymne, Sara (2011) Blast : A tool for error analysis of ma- chine translation output. Dans *Annual Meeting of the Association for Computational Linguistics* ACL 2011 System Demonstrations, pages 56–61, Portland, OR, USA, Jun. ACL.

Tutin, Agnès, Esperança-Rodier, Emmanuelle, Iborra, Manuel, Reverdy, Justine (2015) Annotation of multiword expressions in French. Malaga, Espagne, Dans *Europhras2015*, Juin 2015.

Vilar, David, Xu, Jia, D'Haro Luis Fernando (2006) Error analysis of statistical machine translation output. Dans *Language Resources and Evaluation Conference* LREC. 2006. p. 697-702.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454