

Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus

Rudy Loock

Volume 63, numéro 3, décembre 2018

Traductologie de corpus : 20 ans après

URI : <https://id.erudit.org/iderudit/1060173ar>

DOI : <https://doi.org/10.7202/1060173ar>

[Aller au sommaire du numéro](#)

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

1492-1421 (numérique)

[Découvrir la revue](#)

Citer cet article

Loock, R. (2018). Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta*, 63(3), 786–806.
<https://doi.org/10.7202/1060173ar>

Résumé de l'article

Cet article propose une analyse linguistique d'un corpus de français traduit de façon automatique depuis l'anglais, en comparaison d'un corpus de français original. Deux outils de traduction automatique ont été retenus pour cette étude, l'un générique, grand public et neuronal tandis que l'autre est un outil spécifique, utilisé par une grande organisation internationale et à base de statistiques. Selon la méthodologie de la traductologie de corpus, à travers une analyse quantitative de phénomènes linguistiques (lexicaux et grammaticaux) connus pour poser problème aux traducteurs anglais-français, nous montrons que l'usage linguistique, au-delà des règles et dont le respect permet d'atteindre la fluidité et l'idiomaticité de la langue cible attendues sur le marché, n'est pas pris en compte par les outils de traduction automatique actuels. L'objectif est de mettre au jour la valeur ajoutée de la traduction humaine, tout particulièrement auprès des traducteurs en formation.

Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus

RUDY LOOCK

Université de Lille, Lille, France

Laboratoire « Savoirs, Textes, Langage » du CNRS, Lille, France

rudy.loock@univ-lille.fr

RÉSUMÉ

Cet article propose une analyse linguistique d'un corpus de français traduit de façon automatique depuis l'anglais, en comparaison d'un corpus de français original. Deux outils de traduction automatique ont été retenus pour cette étude, l'un générique, grand public et neuronal tandis que l'autre est un outil spécifique, utilisé par une grande organisation internationale et à base de statistiques. Selon la méthodologie de la traductologie de corpus, à travers une analyse quantitative de phénomènes linguistiques (lexicaux et grammaticaux) connus pour poser problème aux traducteurs anglais-français, nous montrons que l'usage linguistique, au-delà des règles et dont le respect permet d'atteindre la fluidité et l'idiomaticité de la langue cible attendues sur le marché, n'est pas pris en compte par les outils de traduction automatique actuels. L'objectif est de mettre au jour la valeur ajoutée de la traduction humaine, tout particulièrement auprès des traducteurs en formation.

ABSTRACT

In this article we provide a linguistic analysis of a corpus of machine-translated texts from English into French, in comparison with a corpus of original French. Two machine translation tools have been selected: one of them is a generic, general public, neural system while the other is a specific, statistical-based tool developed by a major international organization. Following the corpus-based translation studies approach, we provide a quantitative analysis of a series of linguistic features (lexical and grammatical) which are known to be problematic for English-French translators. We aim to show that linguistic usage, which goes beyond grammatical correctness and needs to be taken into account to provide natural, idiomatic translations in order to meet the demands of today's translation market, is not taken into account by current machine translation systems. Our aim is to show human translators' added value over such systems, in particular for translation trainees.

RESUMEN

Este artículo propone un análisis lingüístico de un corpus en francés traducido de manera automática del inglés, comparado con un corpus en francés (lengua de origen). Para este estudio se seleccionaron dos herramientas de traducción automática: una genérica, disponible a todo tipo de público, y neuronal; mientras que la otra es una herramienta específica utilizada por una gran organización internacional y basada en estadísticas. Según la metodología de la traductología de corpus, por medio de un análisis cuantitativo de los fenómenos lingüísticos (léxicos y gramaticales) que se sabe de antemano que plantean problemas a los traductores inglés-francés, demostramos que el uso lingüístico, más allá de las reglas y cuya consideración permite alcanzar la fluidez e idiomaticidad de la lengua meta que se esperan en el mercado, no es tenido en cuenta por las actuales herramientas de traducción automática. El objetivo es destacar el valor añadido de la traducción humana, especialmente para los traductores en formación.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVES

traduction automatique, corpus, évaluation, usage linguistique, didactique de la traduction
 machine translation, corpus, evaluation, linguistic usage, translation didactics
 traducción automática, corpus, evaluación, uso lingüístico, didáctica de la traducción

1. Introduction

Les récents progrès de la traduction automatique (TA), ou traduction machine¹, sont réels et ne peuvent plus être ignorés et considérés comme sans conséquence sur l'activité des traducteurs professionnels. Il semble loin le temps où les traductions proposées par les différents systèmes, à base de règles ou statistiques, déclenchaient l'incrédulité, voire l'hilarité. L'arrivée de la traduction automatique neuronale en 2015 a permis un bond qualitatif indiscutable, au point que la presse généraliste évoque régulièrement ces progrès, qui entraîneraient selon elle la disparition du métier de traducteur. Les spécialistes savent pourtant qu'il n'en est rien : le marché de la traduction se porte bien, comme le montrent différents rapports sur le sujet, par exemple les rapports annuels du *Common Sense Advisory*² ou encore les rapports du groupe de réflexion TAUS³. Il est en revanche certain que l'arrivée de la traduction automatique neuronale est déjà en train de bouleverser le secteur, tout comme l'intelligence artificielle bouleverse moult secteurs d'activité : pour la première fois en 2018, plus de la moitié des entreprises de services linguistiques en Europe déclarent utiliser la traduction automatique⁴. La façon de travailler et le modèle économique vont devoir faire l'objet d'adaptations. Il semble donc important à ce stade de réfléchir à la façon dont (bio)traducteur⁵ et machine peuvent coexister, la seconde n'ayant pas vocation à remplacer le premier, qui en revanche se doit de définir sa plus-value, sa valeur ajoutée par rapport à une traduction obtenue de façon automatique.

Cette question est également très importante pour les formateurs en traduction, qui se doivent de former les traducteurs et les traductrices de demain. En dehors de cas spécifiques comme certains cas de traduction littéraire, on attend en effet du traducteur aujourd'hui qu'il soit invisible et que ses traductions soient rédigées dans une langue qui soit la plus fluide, la plus idiomatique possible, au point où celles-ci sont censées témoigner d'une homogénéisation linguistique avec des textes rédigés directement en langue cible originale par des natifs de la langue. C'est ainsi que sont formés la plupart des traducteurs en ce début de XXI^e siècle, et cela nécessite une qualité de langue cible qui aille au-delà du grammaticalement correct, en respectant l'usage de la langue en plus de ses règles, pour une invisibilité maximale. Devant les progrès de la TA neuronale, qui met l'accent sur la fluidité de la langue cible, cet enjeu devient crucial : la traduction humaine doit se distinguer de la traduction automatique, et nous considérerons ici que cela passe par le respect de l'usage de la langue cible. Notre contribution propose donc également une réflexion didactique.

Dans cet article nous analysons un corpus de traductions dites « automatiques », à savoir effectuées par une machine, de l'anglais vers le français, et comparons ce corpus à des textes rédigés en français original, afin de mettre au jour des différences relatives à divers phénomènes linguistiques touchant au lexique et à la morphosyntaxe. Notre analyse se situe dans le cadre de la traductologie de corpus (Loock 2016), traduction en français de *corpus-based translation studies* (Laviosa 2002), dont l'objectif est d'analyser quantitativement et qualitativement des traductions réunies

en corpus de travail, afin notamment d'effectuer des comparaisons avec la langue originale et de mettre au jour des différences de fonctionnement. Nous exploiterons donc deux corpus, un de français original et un de français traduit automatiquement depuis l'anglais. Pour ce second corpus, nous avons sélectionné deux systèmes de TA, l'un neuronal et grand public, l'autre à base de statistiques et réservé à une organisation internationale. Il s'agira donc, en fonction des phénomènes linguistiques retenus pour l'analyse, de comparer les performances des deux outils de TA, en comparaison de la langue originale. Cette évaluation sera quantitative à partir de résultats chiffrés obtenus lors de l'analyse des différents sous-corpus de notre corpus de travail : elle cherchera à mettre au jour des différences de fréquence pour toute une série de phénomènes linguistiques ayant la réputation de poser problème aux traducteurs anglais-français ; la question du lien avec l'influence de la langue source (anglais) sera également abordée.

L'article est organisé de la façon suivante : la deuxième partie dresse un état des lieux des différents types d'évaluation des systèmes de traduction automatique et fournit les résultats d'études récentes sur la TA neuronale ; la troisième partie explique notre méthodologie et détaille notre corpus de travail ; la quatrième partie fournit les résultats détaillés des analyses de corpus ; la cinquième partie, enfin, propose une discussion de ces résultats et étudie le lien avec l'interférence de la langue source mais aussi ses conséquences en termes d'enseignement de la traduction.

2. L'évaluation de la traduction automatique : des métriques à l'évaluation linguistique

Tout au long du développement depuis le milieu du siècle dernier des différents systèmes de traduction automatique, d'abord à base de règles (rule-based machine translation ou RBMT) puis à base de statistiques (statistical-based machine translation ou SBMT) ou hybrides, avant l'arrivée en 2015 de la traduction automatique neuronale (neural machine translation ou NMT), des méthodes d'évaluation des traductions obtenues ont été mises au point. L'objectif est alors de comparer les performances des différents systèmes entre eux mais aussi de mesurer les progrès d'un même système en fonction des modifications apportées aux règles, aux algorithmes, ou encore aux corpus sous-jacents. Grâce à cette évaluation constante, il a été possible de mettre au jour les faiblesses et les erreurs des différents systèmes afin de les corriger et de permettre une amélioration continue des outils de traduction automatique.

2.1. Les métriques

Parce que l'évaluation humaine est très coûteuse, en temps et en argent, la recherche appliquée s'est orientée principalement vers des méthodes d'évaluation non humaines, automatisées, appelées « métriques ». La plus célèbre d'entre elles est certainement BLEU (*BiLingual Evaluation Understudy*), proposée par Paineni, Roukos, *et al.* (2002). Il s'agit alors de mesurer la proximité entre le résultat fourni par un système de traduction automatique et une ou plusieurs traductions humaines dites « de référence » : plus la traduction fournie par la machine se rapproche d'une traduction humaine, plus elle est de bonne qualité. D'autres métriques ont été développées, comme ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), NIST (*National*

Institute of Standards and Technology), ou encore METEOR (*Metric for Evaluation of Translation with Explicit ORdering*).

2.2. L'évaluation humaine

Les métriques ont fait et font encore l'objet de critiques puisque l'évaluation, qui se concrétise sous forme d'un score, est indépendante des langues source et cible, et n'évalue que la forme et non le contenu (Hartley et Popescu-Belis 2004; Koehn 2010) par le biais de comparaisons entre n-grammes (séries de n mots). Se sont donc développées en parallèle et en complément des évaluations humaines, comme (i) le classement des traductions par des professionnels du secteur ou non permettant de désigner les meilleures et les moins bonnes selon la fidélité au texte source et/ou la fluidité de la langue cible (Bojar, Chatterjee, *et al.* 2015); (ii) la quantité de postédition nécessaire pour rendre la traduction acceptable (Koehn et Germann 2014; Bentivogli, Bisazza, *et al.* 2016 constatent que les besoins de post-édition diminuent de plus de 25 % avec la TA neuronale pour la paire de langues anglais-allemand par rapport à un système de TA statistique); ou encore (iii) la classification des types d'erreurs identifiées (erreurs lexicales, syntaxiques; ajouts, omissions; ordre des mots...) comme dans Federico, Negri, *et al.* (2014), certaines études ayant recours conjointement aux différentes méthodes comme dans Popović, Avramidis, *et al.* (2013).

Cette évaluation humaine peut alors venir compléter une évaluation automatique obtenue grâce aux métriques. Ainsi, Castilho, Moorkens, *et al.* (2017) propose une comparaison entre un système de TA statistique et un système de TA neuronale dans le cadre de la traduction de documents éducatifs, en l'occurrence des MOOC (*Massive Open Online Courses*) de l'anglais vers l'allemand, le portugais, le russe et le grec. Les auteurs ont eu pour cela recours à des métriques (BLEU, chrF3, METEOR), à la quantification de post-édition nécessaire, mais aussi à des évaluations humaines par des traducteurs professionnels (expression de préférences, évaluation de type Likert pour la fidélité et la fluidité, repérage d'erreurs). Leurs résultats montrent que le système de TA neuronale testé dans le cadre de leur étude permet une augmentation générale des métriques, une amélioration de la fluidité et une diminution du nombre de segments devant être post-édités, la préférence des traducteurs s'orientant vers les traductions obtenues par la TA neuronale. Néanmoins, les auteurs observent que l'effort de post-édition reste globalement le même, et que le nombre d'omissions et d'erreurs de traduction reste également sensiblement le même. On notera que l'évaluation par les métriques et l'évaluation humaine semblent converger. C'est également cette approche hybride (évaluation métrique et humaine) qui a été utilisée pour promouvoir les performances de l'outil de TA neuronale d'une grande multinationale américaine⁶: les scores BLEU, mais aussi une évaluation de type humain par classement de traductions de pages Wikipédia et de textes de presse pour différentes paires de langues (anglais-français, anglais-espagnol et anglais-chinois), confirment le gain qualitatif avec la TA neuronale pour cet outil, permettant une diminution du nombre d'erreurs de traduction pouvant aller jusqu'à 60 %.

2.3. L'analyse linguistique

Plus spécifiquement, certains chercheurs ont souhaité se concentrer sur une évaluation *linguistique* du produit des différents systèmes de traduction automatique. Il s'agit alors de se concentrer sur des phénomènes linguistiques spécifiques, lexicaux ou grammaticaux, afin de constater la façon dont ils sont gérés par les différents systèmes. Ceci peut alors se faire en comparaison des traductions humaines ou encore avec des textes rédigés directement en langue cible. Pour cela, les traductions sont réunies en corpus de travail et analysées selon les méthodes développées dans le cadre de la traductologie de corpus, inspirées de la linguistique de corpus et appliquées aux textes traduits (Laviosa 2002), afin de comparer des échantillons de textes en y quantifiant les phénomènes linguistiques retenus ; l'objectif est alors de mettre au jour des différences ou similitudes de fréquence pour les différents sous-corpus considérés. Il est toutefois important de noter qu'il est également possible de mener des évaluations de type linguistique sans recourir à l'analyse de corpus, comme dans Isabelle, Cherry, *et al.* (2017), qui soumettent différents systèmes de traduction automatique à un *challenge set*, à savoir un ensemble de phrases isolées mettant en jeu une série de phénomènes linguistiques précis (p. ex. : position des pronoms, expression du mouvement, présence de prépositions dites orphelines) et connus comme posant problème pour la traduction du fait des différences entre les deux systèmes linguistiques considérés (ici, anglais et français).

L'analyse linguistique de traductions réunies en corpus, approche qui reste minoritaire, se retrouve par exemple chez Macketanz, Avramidis, *et al.* (2017), qui proposent une évaluation des trois systèmes de TA (RBMT, SMT et NMT) qui est nettement linguistique. L'analyse porte sur un échantillon de 100 segments extraits de traductions de documents techniques de l'anglais vers l'allemand et a été menée manuellement, par un linguiste. L'objectif est d'observer la façon dont les différents systèmes gèrent certains phénomènes linguistiques, qui dans le cas présent relèvent de la morphosyntaxe, de la sémantique, mais aussi du formatage et du style : impératifs, mots composés, points d'interrogation, verbes à particule, choix terminologiques, séparateurs « > » (propres au type de texte analysé), omissions de verbes, soit toute une série de phénomènes connus pour poser problème (Macketanz, Avramidis, *et al.* 2017 : 32). L'analyse du corpus de travail montre que les résultats obtenus pour les trois systèmes sont en moyenne équivalents, ce qui peut à première vue paraître surprenant, mais chaque système semble se distinguer en fonction du phénomène linguistique considéré : ainsi le système de TA neuronale testé fournit de meilleurs résultats s'agissant de la traduction des verbes, tandis que le système à base de règles est celui qui obtient les meilleurs résultats pour la traduction des mots composés.

On retrouve cette approche chez Lapshinova-Koltunski (2015), qui propose une méthode d'analyse de la variation au sein des textes traduits de l'anglais vers l'allemand en comparaison avec des textes en langue originale (textes sources en anglais ou textes comparables en allemand) en fonction des outils utilisés : (i) traductions humaines effectuées sans aucun outil ; (ii) traductions humaines effectuées à l'aide d'un logiciel de traduction assistée par ordinateur (TAO) ; (iii) traductions automatiques (un système à base de règles et deux systèmes statistiques, pas de système neuronal). Les auteurs, qui font nettement référence aux méthodes développées dans le cadre de la traductologie de corpus ainsi qu'aux concepts développés par les chercheurs ayant travaillé dans ce cadre, notamment ce que l'on a appelé les « universaux

de traduction», tels que conceptualisés dans l'article fondateur de Baker (1993), proposent d'analyser leurs différents sous-corpus en quantifiant certains phénomènes linguistiques afin de mettre au jour la présence ou non de certains de ces universaux de traduction : la simplification (via la variété et de la densité lexicales), l'explicitation (via la présence des marqueurs explicites de cohésion), la normalisation vs l'interférence de la langue source (grâce à l'analyse d'un phénomène discriminant entre les deux langues, à savoir la présence accrue de verbes en allemand). À partir des différents sous-corpus (textes originaux; textes traduits sans outils, avec un logiciel de TAO ou par différents systèmes de TA) couvrant 7 registres différents, les auteurs fournissent des résultats détaillés permettant de mettre au jour certains phénomènes en lien avec les universaux de traduction. Par exemple, si la densité lexicale (ratio entre mots lexicaux et grammaticaux) est assez homogène entre les différents corpus, la variété lexicale (ratio type/token) est inférieure dans les textes traduits à l'aide d'un outil de TAO ou de TA par rapport aux textes originaux et aux traductions humaines, ce qui pourrait selon les auteurs être le signe d'une simplification des textes traduits lorsque la traduction est outillée, bien que les résultats pour la densité lexicale viennent contredire l'hypothèse.

2.4. Enjeux pour l'évaluation de la TA

Les études publiées sur l'évaluation de la TA neuronale ces dernières années montrent globalement une progression de la qualité des textes traduits par rapport à la TA statistique, même si les résultats sont parfois largement exagérés, notamment dans la presse généraliste et dans les discours marketing. Les progrès sont réels, et l'analyse des erreurs commises par la traduction machine s'avère cruciale si l'on souhaite savoir ce que la machine sait faire et ne sait pas faire, et par conséquent si l'on souhaite dégager la plus-value de la traduction humaine. En particulier, l'analyse linguistique de traductions automatiques réunies en corpus de travail permet de mettre au jour les «manquements» de la machine. Elle doit également permettre une meilleure formation des traducteurs de demain. C'est cette approche que nous souhaitons utiliser ici pour des textes traduits automatiquement de l'anglais vers le français, et nous proposons spécifiquement une analyse linguistique qui se concentre sur l'usage de la langue, au travers de toute une série de phénomènes linguistiques ayant la réputation de poser problème au traducteur anglais-français du fait d'une différence de fréquence entre la langue anglaise originale et la langue française originale comme l'ont montré des travaux en grammaire comparée et en traductologie (voir ci-dessous). Précisément, notre analyse portera sur des phénomènes lexicaux (fréquence des lemmes *chose* et *dire*) et grammaticaux (adverbes dérivés en *-ment*, préposition *avec*, coordination par *et*, structures existentielles en *il y a*).

3. Corpus de travail et méthodologie

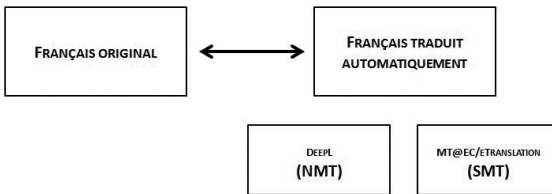
3.1. Le corpus de travail

L'objectif étant de comparer langue originale et langue traduite (automatiquement), nous avons compilé deux corpus distincts, le second étant lui-même divisé en deux sous-corpus. Nous avons en effet souhaité comparer français original et français traduit automatiquement depuis l'anglais par deux outils de TA différents : un outil

générique neuronal grand public et un outil à base de statistiques conçu pour des besoins spécifiques et non accessible au grand public. Pour le premier, nous avons sélectionné l'outil *DeepL*⁷, disponible gratuitement en ligne et connu pour ses résultats parfois impressionnants en ce qui concerne la qualité de la langue cible, même si cela se fait parfois aux dépens de la fidélité au texte source. Cet outil exploite les corpus parallèles bilingues de l'outil *Linguee*⁸ compilés à partir de traductions existantes et disponibles sur l'internet (sites multilingues, romans libres de droits, textes législatifs internationaux, etc.). Lancé en 2017, *DeepL* affirme, scores BLEU à l'appui, être le meilleur outil de TA au monde⁹ et obtenir des résultats trois fois supérieurs à celui de son concurrent principal. Les textes à traduire peuvent directement y être copiés/collés dans la limite de 5 000 mots; la traduction en langue cible s'affiche en quelques secondes. Pour le second sous-corpus de français traduit depuis l'anglais, nous avons eu recours à l'outil *MT@EC/eTranslation* de la Direction générale de la traduction (DGT) de la Commission européenne. Actuellement en transition avec le déploiement de la TA neuronale, l'outil interne de la DGT qui s'appelle désormais *eTranslation* depuis l'été 2018 mais qui s'appelait *MT@EC* jusque là, s'appuie sur les traductions déjà effectuées au sein de la Commission et est donc particulièrement conçu pour la traduction de textes institutionnels. À l'époque où nous avons mené notre étude (printemps 2018), l'outil recourait à la traduction statistique pour la paire de langues anglais-français. Nos deux sous-corpus permettent donc la comparaison entre deux systèmes de TA : neuronal pour *DeepL*, statistique pour *MT@EC/eTranslation*. La figure 1 synthétise la composition du corpus de travail.

FIGURE 1

Composition générale du corpus



Les textes utilisés afin de compiler le corpus ont été extraits du TSM Press Corpus (Loock, à paraître), un corpus de textes journalistiques originaux en anglais et en français compilé à l'Université de Lille dans le cadre de la formation de master Traduction Spécialisée Multilingue. Ce corpus contient des textes de presse rédigés dans les deux langues originales (anglais britannique et américain, français de France) et répartis en différentes thématiques (économie, environnement, sports, voyages, crime, culture, etc.). Les textes sont issus de la presse généraliste britannique, américaine, et française (par exemple *The Guardian*, *The Independent*, *The New York Times*, *USA Today*, *Le Monde*, *Libération*). Le corpus, dont la compilation a débuté en 2014, est un corpus « ouvert », de nouveaux textes étant ajoutés chaque année par les étudiants de la formation dans le cadre d'un cours de grammaire comparée anglais-français. Au moment où nous avons mené notre étude (printemps 2018), le corpus contenait environ 1,6 million de mots. Le tableau 1 fournit la composition détaillée du TSM Press Corpus.

TABLEAU 1

Composition du TSM press corpus

	Anglais américain original	Anglais britannique original	Français original
Économie & Finance	27 487	6 136	36 964
Crime	44 315	43 710	93 347
Culture	30 570	46 839	78 897
Environnement	41 500	32 367	88 574
Santé	34 790	28 170	65 024
Nouvelles internationales	33 767	29 168	65 354
Politique	45 840	46 901	98 540
Science & Technologies	45 269	47 213	97 252
Sports	45 156	43 766	97 367
Voyages	40 748	50 056	97 351
TOTAL	389 442	374 326	818 670
Nombre de fichiers	437	490	1 094

Pour notre étude, nous avons sélectionné l'ensemble des 1 094 textes en français, ainsi que les 490 textes en anglais britannique que nous avons soumis aux deux outils de traduction automatique sélectionnés. Notre corpus de travail correspond donc à 1 094 articles de presse rédigés en français original et à 980 articles traduits en français de façon automatique, pour un total d'un peu plus de 1,7 million de mots. La composition détaillée du corpus de travail est résumée dans le Tableau 2.

TABLEAU 2

Composition détaillée du corpus de travail

	Français original	Français traduit depuis l'anglais avec <i>DeepL</i> (NMT)	Français traduit depuis l'anglais avec <i>MT@EC/eTranslation</i> (SMT)
Nombre de textes	1 094	490	490
Nombre de mots	816 338	442 439	445 914

Il importe à ce stade d'effectuer une remarque importante sur la question du genre des textes. Comme pour les mémoires de traduction, le type de texte utilisé a nécessairement une influence sur les résultats fournis par les systèmes de TA, qui associent algorithme et corpus de données bilingues et parfois monolingues de façon complémentaire. Nous avons retenu ici les textes journalistiques car ni informels ni trop formels ou spécialisés. Or, aucun des deux outils de TA retenus n'est spécialisé dans la traduction automatique de textes de presse: *DeepL* est un outil générique; l'outil *MT@EC/eTranslation* est entraîné sur des textes institutionnels. Ne disposant pas de nos propres outils de traduction automatique, il nous a fallu faire des choix, en optant pour une position équilibrée qui ne privilégie ni l'un ni l'autre des outils de TA retenus, en leur soumettant des textes proches de la langue dite «générale» même si en tant que telle celle-ci n'existe pas (il ne s'agit en effet que d'une langue abstraite qui correspond en fait à l'addition des différents registres existants puisque tout acte de communication appartient à un registre spécifique). Il n'empêche qu'il s'agit là d'une des limites de notre étude, qu'il convient de ne pas passer sous silence.

3.2. *Compilation du corpus de travail*

L'ensemble des fichiers du corpus a été enregistré au format .txt avec encodage UTF-8 (*Universal Character Set Transformation Format - 8 bits*) permettant la prise en charge de l'ensemble des caractères utilisés pour les différents systèmes d'écriture (norme ISO/CEI 10646). Les fichiers en français original ont été directement extraits du TSM Press Corpus et n'ont subi aucune modification. Les fichiers en français traduit ont été obtenus de deux façons différentes. Pour l'outil *DeepL*, les textes ont été traduits en ligne via le site internet dédié en indiquant l'anglais comme langue source et le français comme langue cible. Les textes ont été copiés/collés depuis les fichiers en anglais britannique original du TSM Press Corpus; il a été nécessaire de les segmenter lorsque le nombre de mots des textes sources était supérieur à 5 000. S'agissant de l'outil de la DGT *MT@EC/eTranslation*, nous avons pu avoir accès à l'outil interne. Les 490 fichiers ont été déposés de façon individuelle (dans une limite de 50 fichiers par jour) via la fonctionnalité « Traduire des documents », et après avoir indiqué l'anglais comme langue source et le français comme langue cible, les textes traduits ont pu être téléchargés et convertis en fichiers .txt (encodage UTF-8).

Le corpus ainsi compilé a ensuite été exploité à l'aide du logiciel *AntConc*¹⁰ version 3.5.7, concordancier permettant l'exploitation de corpus hors ligne. Pour les besoins de notre étude (voir ci-dessous), un étiquetage grammatical a été nécessaire (*pos-tagging*); celui-ci a été effectué grâce à l'outil *TreeTagger* (Schmid 1997/2003) pour le français via le logiciel *TagAnt*¹¹ version 1.2.0. Le corpus a donc pu être exploité de façon automatique, mais un contrôle systématique des données a eu lieu et, lorsque cela fut nécessaire, les exemples dits « bruyants », à savoir les faux positifs, ont été retirés manuellement, comme les exemples du type *il y a passé beaucoup de temps* ou encore *il y a 3 ans* lors de la recherche des structures existentielles en *il y + AVOIR* puisque ces exemples ne sont pas des structures existentielles⁷. Les trois sous-corpus (français original, français traduit *DeepL*, français traduit *MT@EC/eTranslation*) étant de tailles différentes, les fréquences brutes ont été normalisées en fréquences par million de mots (pmm) afin que les comparaisons soient pertinentes. Enfin, afin de déterminer si les différences observées étaient le fruit du hasard ou significatives, un test statistique a été utilisé, en l'occurrence le test de significativité entre proportions indépendantes, avec calcul d'un z-ratio, la valeur p retenue pour le seuil de significativité étant $p < 0,01$ pour rejeter l'hypothèse nulle (Cappelle et Loock 2013).

3.3. *Méthodologie: phénomènes linguistiques retenus pour l'analyse*

Notre objectif étant de vérifier dans quelle mesure les deux outils de TA retenus sont capables de prendre en compte l'usage linguistique, nous avons concentré notre étude de corpus sur une série de phénomènes connus pour poser problème aux traducteurs de l'anglais vers le français, à savoir toute une série de phénomènes pour lesquels il existe une différence interlangagière significative entre les deux langues originales (Cappelle et Loock 2013; Loock 2016, à paraître) et souvent mentionnés dans les manuels de traduction ou de grammaire comparée anglais-français (par exemple *Stylistique comparée du français et de l'anglais* de Jean-Paul Vinay et Jean Darbelnet, *Approche linguistique des problèmes de traduction* d'Hélène Chuquet et Michel Paillard, ou encore *Syntaxe comparée du français et de l'anglais* de Jacqueline Guillemin-Flescher). En effet, une différence importante de fréquence entre deux

langues originales pour un même phénomène linguistique peut donner lieu à une sur-/sous-représentation de ces phénomènes en langue cible lors de la traduction (interférence de la langue source). Nous avons ici retenu différents phénomènes linguistiques, lexicaux et grammaticaux, pour lesquels nous avons déjà mis au jour ailleurs ces différences interlangagières significatives, en l’occurrence une fréquence bien plus élevée en anglais original de leur équivalent direct en français :

- La fréquence du lemme *chose* (vs son équivalent direct thing) ;
- La fréquence du lemme *dire* (vs son équivalent direct say) ;
- La fréquence du coordonnant *et* (vs son équivalent direct and) ;
- La préposition *avec* (vs son équivalent direct with) ;
- Les adverbes dérivés en *-ment* (vs leurs équivalents directs en -ly) ;
- Les structures existentielles (structures en *il y a*) (vs leur équivalent direct en there+BE)¹².

4. Résultats

De façon systématique, les résultats témoignent d’une surreprésentation en français traduit automatiquement depuis l’anglais par rapport au français original, pour l’ensemble des phénomènes linguistiques étudiés, mais dans des ordres de grandeur qui ne sont pas toujours similaires. Nous détaillons ces résultats ci-dessous.

S’agissant du lemme *chose*, on constate une surreprésentation très significative en français traduit automatiquement, avec des fréquences normalisées de 556,01 et 410,39 occurrences par million de mots en français traduits au moyen de *DeepL* et de *MT@EC/eTranslation* respectivement, contre seulement 248,67 pour le français original (z -ratio = -8,718 ; $p < 0,0001$; z -ratio = -4,967 ; $p < 0,0001$). On constate le même résultat pour le lemme *dire* avec des fréquences de 3 315,71 et 1 170,63 occurrences par million de mots en français traduit automatiquement contre 946,91 en français original (z -ratio = -30,106 ; $p < 0,0001$; z -ratio = -3,753 ; $p < 0,0001$). À noter toutefois que la surreprésentation est dans les deux cas bien plus importante pour les traductions effectuées à l’aide de *DeepL*, ce qui peut paraître étonnant pour un outil qui met l’accent sur la fluidité de la langue cible. Les figures 2 et 3 illustrent ces résultats pour deux lemmes que l’on peut qualifier d’hyperonymes et dont la fréquence en anglais original est bien plus élevée qu’en français original (Loock 2016 : 145-146, 153), la langue française utilisant davantage de termes précis (hyponymes) tout particulièrement pour les verbes de dire permettant d’introduire une citation. Il sera donc intéressant de comparer la fréquence des équivalents de *chose* et de *dire*, à savoir thing et say, dans les textes sources, ce que nous ferons dans la Section 5.

FIGURE 2

Fréquence du lemme *chose* en français original et traduit automatiquement (pmm)

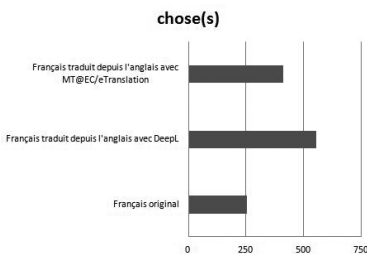
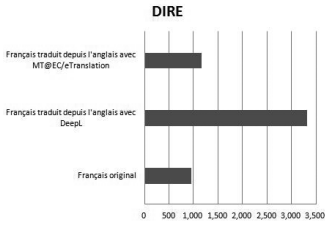


FIGURE 3

Fréquence du lemme *dire* en français original et traduit automatiquement (pmm)



S'agissant du coordonnant *et*, là encore, on constate une surreprésentation significative en français traduit automatiquement depuis l'anglais : là où sa fréquence est de 18 079,52 occurrences par million de mots en français original, elle est de 21 435,72 et de 21 129,63 en français traduit automatiquement au moyen des outils *DeepL* et *MT@EC/eTranslation* respectivement (z -ratio = -13,081 ; $p < 0,0001$; z -ratio = -11,349 ; $p < 0,0001$). Cette fois, en revanche, les fréquences en français traduit sont très proches et les deux outils semblent donc se comporter de façon similaire.

S'agissant de la préposition *avec*, les résultats vont dans le même sens, à savoir une surreprésentation en français traduit avec 4 289,86 et 4 184,66 occurrences par million de mots en français traduit automatiquement au moyen de *DeepL* et de *MT@EC/eTranslation* respectivement, contre 3 689,65 pour le français original (z -ratio = - 5,158 ; $p < 0,0001$; z -ratio = -4,284 ; $p < 0,0001$). Cette fois encore, les fréquences en français traduit restent proches.

Les Figures 4 et 5 illustrent ces résultats.

FIGURE 4

Fréquence du coordonnant *et* en français original et traduit automatiquement (pmm)

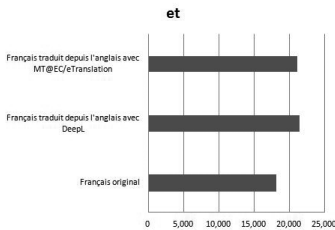
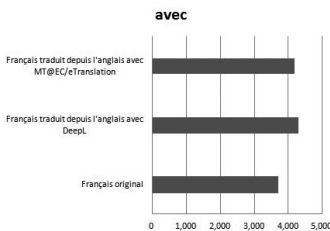


FIGURE 5

Fréquence de la préposition *avec* en français original et traduit automatiquement (pmm)



En ce qui concerne les phénomènes morphosyntaxiques à présent, on constate une surreprésentation en français traduit des adverbess dérivés en *-ment* avec des fréquences de 8 148,02 et 9 071,26 occurrences par million de mots en français traduit automatiquement au moyen de *DeepL* et de *MT@EC/eTranslation* respectivement, contre seulement 6 978,73 en français original ($z\text{-ratio} = -7,313$; $p < 0,0001$; $z\text{-ratio} = -12,481$; $p < 0,0001$). On note une surreprésentation légèrement plus marquée avec *MT@EC/eTranslation* cette fois (TA statistique).

Enfin, on note également une surreprésentation des structures existentielles en *il y a* en français traduit automatiquement grâce à l'outil *DeepL* avec une fréquence de 1 573,10 occurrences par million de mots contre 850,14 en français original, soit près de deux fois moins ($z\text{-ratio} = -11,66$; $p < 0,0001$), tout comme pour l'outil *MT@EC/eTranslation* mais dans des proportions bien différentes : avec une fréquence de 1000,19 occurrences par million de mots, la différence entre les deux sous-corpus reste significative ($z\text{-ratio} = -2,683$; $p = 0,0036$) mais se rapproche de celle en français original. Les Figures 6 et 7 illustrent ces résultats.

FIGURE 6
Fréquence des adverbess dérivés en *-ment* en français original et traduit automatiquement (pmm)

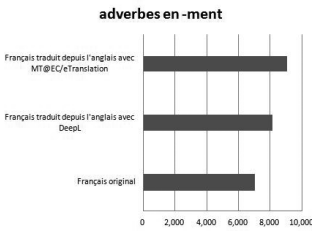
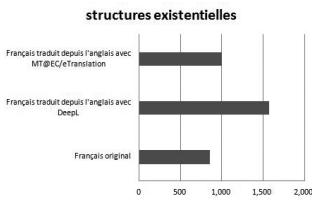


FIGURE 7
Fréquence des structures existentielles en français original et traduit automatiquement (pmm)



5. Discussion

5.1. L'interférence de la langue source ?

L'analyse de corpus dont les résultats détaillés viennent d'être fournis montre que les phénomènes linguistiques retenus sont en surreprésentation en français traduit automatiquement, et ce, de façon systématique. On constate donc des différences intralangagières quel que soit l'outil de TA utilisé, neuronal ou statistique, même si l'on note une différence de résultats en fonction de l'outil utilisé. Il semblerait – le nombre de phénomènes étudiés étant restreint, il convient d'être prudent – que

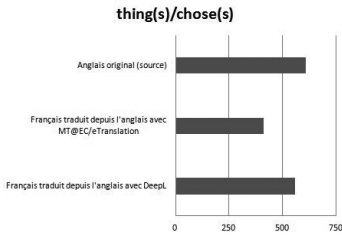
l'outil de TA neuronale *DeepL* engendre une surreprésentation plus importante des phénomènes plus fréquents en anglais original qu'en français original. Quel que soit l'outil, l'usage de la langue, illustré ici par la fréquence de ces différents phénomènes, ne semble donc pas respecté, dans des proportions plus importantes par l'outil de TA neuronale. Cette surreprésentation en langue traduite par rapport à la langue originale de phénomènes linguistiques, lexicaux ou grammaticaux, peut entraîner un problème de qualité, révélant au lecteur le statut traduit du texte, *contra* les exigences actuelles du marché de la traduction qui prône l'invisibilité du traducteur, comme nous l'évoquions en introduction. Pour une invisibilité maximale, il conviendra de tendre vers une forme d'homogénéisation linguistique entre langue originale et langue traduite, même si celle-ci ne peut être parfaite, au-delà du grammaticalement correct et qui prenne en compte l'usage de la langue.

Dans la mesure où tous ces phénomènes sont caractérisés par une différence interlangagière entre anglais original et français original (voir ci-dessus), il serait assez logique d'y voir une influence de la langue source, et donc d'une interférence de celle-ci lors du passage à la langue cible, y compris dans le cas de la TA neuronale dont la réputation est pourtant de mettre l'accent sur la fluidité de la langue cible (dans le cadre de cette hypothèse, il resterait néanmoins à expliquer les différences d'ampleur de la surreprésentation constatées entre les deux sous-corpus de français traduit automatiquement). Toutefois, si l'on souhaite explorer l'hypothèse de l'influence de la langue source sur la langue cible (la surreprésentation en français traduit a pour origine une fréquence élevée dans les textes sources), il convient de comparer textes originaux et textes traduits. C'est ce que nous avons fait, en comparant la fréquence des équivalents en anglais (textes sources) des différents phénomènes linguistiques à l'étude avec les fréquences dans les deux sous-corpus de français traduit automatiquement.

Ce que nous avons constaté, c'est que l'explication de l'influence de langue source seule ne suffit pas : on note en effet dans la majorité des cas des différences de fréquence significatives entre l'anglais original et le français traduit, alors que l'hypothèse de l'interférence de la langue source pourrait laisser penser qu'anglais original et français traduit automatiquement seraient plus proches linguistiquement. Ainsi pour ce qui est du lemme *chose* et de son équivalent en anglais *thing*, on note certes une absence de différence significative entre les textes sources en anglais original et leur traduction en français au moyen de *DeepL* (609,09 contre 556,01 occurrences par million de mots; $z\text{-ratio}=0,993$; $p\text{-value}=0,164$), soit en termes de fréquences brutes puisque les échantillons sont ici directement comparables, avec 228 et 246 occurrences. En revanche, on note une différence significative avec les traductions effectuées au moyen de *MT@EC/eTranslation* qui témoignent d'une fréquence de *chose* de 410,39 occurrences par million de mots, soit une fréquence brute de 183 seulement ($z\text{-ratio} = 4,005$; $p < 0,0001$). Dans le cadre de notre hypothèse, l'ampleur des non-traductions de *thing* par *chose* serait donc différente selon l'outil utilisé, et l'influence de langue source ne peut donc pas expliquer les résultats qui sont illustrés par la figure 8 (ou alors elle s'exercerait de façon non systématique).

FIGURE 8

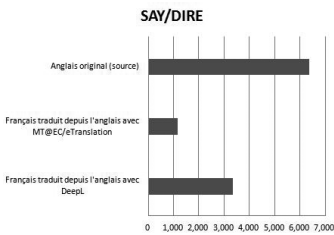
Fréquence des lemmes thing et chese en anglais original et en français traduit automatiquement (pmm)



L'influence de la langue source ne permet par ailleurs pas d'expliquer des résultats comme ceux obtenus pour le lemme *dire*. La fréquence en français traduit est très inférieure à la fréquence du verbe *say* dans les textes originaux : 3 315,71 et 1 170,63 occurrences par million de mots pour *dire* contre 6 355,42 pour *say* en anglais original. Les deux outils de TA semblent toutefois se comporter très différemment : dans les textes traduits avec *MT@EC/eTranslation*, la fréquence de *dire* est cinq fois inférieure à celle de *say* dans les textes originaux ; pour *DeepL*, elle n'y est que deux fois inférieure (Figure 9). Si influence de la langue source il y a, celle-ci n'est pas systématique et il conviendrait d'analyser les résultats de plus près (Section 5.2).

FIGURE 9

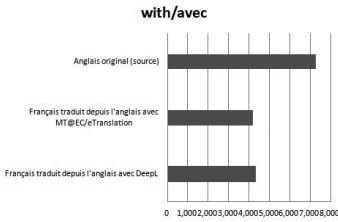
Fréquence des lemmes say et dire en anglais original et en français traduit automatiquement (pmm)



Même dans les cas où les deux outils de TA semblent se comporter de la même façon, on constate une différence significative avec les textes originaux. Ainsi, pour ce qui est de la fréquence des prépositions with et avec, on constate que les textes traduits, quel que soit l'outil, se situent entre les textes originaux en anglais (7 279,75 occurrences par million de mots) et les textes en français original (3 689,65 occurrences par million de mots), comme l'illustre la Figure 10. L'hypothèse de l'interférence de la langue source peut certes venir expliquer cet état de fait, mais il serait faux de penser que les outils de TA restituent systématiquement with par avec : en termes de fréquences brutes, la différence entre les textes originaux et les textes traduits est de 827 (*DeepL*) et 859 (*MT@EC/eTranslation*) occurrences.

FIGURE 10

Fréquence des prépositions with et avec en anglais original et en français traduit automatiquement (pmm)



Un cas a particulièrement retenu notre attention : celui des coordonnants and et et. En effet, l'analyse des fréquences normalisées laisse entendre qu'il existe une différence significative à la baisse entre les textes originaux anglais et leurs deux traductions en français : 25 127,83 occurrences par million de mots pour and, contre 21 435,72 (*DeepL*) et 21 129,63 (*MT@EC/eTranslation*). Or, si l'on observe les fréquences brutes, on constate que ce qui rend la différence significative est le nombre total de mots du corpus, qui est en augmentation lors du passage en français (+18,2 % pour *DeepL* et +19,1 % pour *MT@EC/eTranslation*), phénomène que l'on constate également pour la traduction humaine, sans que ceci ne puisse être attribué à la traduction de and : les fréquences brutes sont en effet de 9 406 pour and dans les textes originaux et de 9 484 et 9 422 pour et dans les textes traduits en français au moyen de *DeepL* et de *MT@EC/eTranslation* respectivement. On constate donc que la fréquence brute de et dans les textes cibles est en fait légèrement plus importante que celle de and dans les textes sources. Si l'on imagine que le coordonnant and a systématiquement été traduit par et (interférence de la langue source), ce qui resterait à démontrer, il reste à expliquer l'« ajout » de et lors de la traduction (78 et 16 occurrences respectivement). Ceci montre qu'une analyse quantitative seule ne suffit pas ; nous y revenons ci-dessous.

5.2. Le besoin d'une analyse qualitative

L'analyse chiffrée de notre corpus de travail fournit des résultats intéressants, mais n'apporte certainement pas toutes les réponses si l'on souhaite expliquer le comportement des deux outils de traduction automatique. Comme pour les traductions humaines, si l'on souhaite mettre au jour la systématisme d'un phénomène ou encore identifier les cas d'interférence en fonction du contexte, cette analyse doit être couplée à une analyse qualitative. Au-delà de l'analyse de notre corpus comparable, il conviendrait donc d'exploiter un corpus dit « parallèle »¹³ aligné au niveau de la phrase en comparant textes originaux et leurs traductions. Ceci permettrait de repérer et d'analyser les cas de traductions dites « littérales » (thing traduit par *chose* ; structure existentielle conservée en langue cible) comme en (1)-(2) et les traductions « non littérales » comme en (3-5), dont certaines peuvent être fautives (5) :

- 1) The fruit and flowers were a colourful joy, and in the relentless heat and humidity **things** grew while we watched, which had its good and bad sides.

(O’Cuneen 2015)¹⁴

Les fruits et les fleurs étaient une joie colorée, et dans la chaleur et l’humidité incessantes, les **choses** grandissaient pendant que nous observions, ce qui avait ses bons et ses mauvais côtés. (DeepL)

- 2) Referring to a letter sent by George Osborne that urges “rapid progress” on fracking, he said: “**There are** too many people digging their fingers into affairs that are not theirs.”

(Carrington 2015)¹⁵

Se référant à une lettre envoyée par George Osborne qui demande “des progrès rapides” sur le fracking, il a dit: “**Il y a** trop de gens qui se creusent les doigts dans des affaires qui ne sont pas les leurs”. (DeepL)

- 3) In reality, any lynx can be either or none of these **things**.

(Mathiesen 2015)¹⁶

En réalité, n’importe quel lynx peut être l’un ou l’autre de ces **éléments**. (DeepL)

- 4) Among other **things**, Greece would be able to contract a special credit line from the European Stability Mechanism.

(Hope 2015)¹⁷

La Grèce pourrait **notamment** contracter une ligne de crédit spéciale au titre du mécanisme européen de stabilité. (DeepL)

- 5) Now it seems **there will be** almost twice as many extreme La Niñas as well.

(Slezak 2015)¹⁸

Il semble qu’aujourd’hui, **ce seront** presque deux fois plus extrêmes La Niñas également. (MT@EC/eTranslation)

Une telle analyse permettrait de déterminer la proportion de traductions littérales et de traductions non littérales, d’identifier les différents types de traductions non littérales, tout en distinguant traductions acceptables et non acceptables, ce qui permettrait au final une approche plus fine du comportement des deux outils de TA. Une analyse qualitative sur un corpus parallèle aligné permettrait également de quantifier les « ajouts » au-delà des traductions littérales et des reformulations, ce que les résultats issus du corpus comparable ne permettent pas: dans quels types de contexte ces ajouts se produisent-ils et sont-ils acceptables ?

5.3. Conséquences pour la formation des futurs traducteurs ?

L’arrivée de la traduction automatique neuronale représente un (nouveau) défi pour les formateurs en traduction, après l’informatisation importante qu’a déjà connue le secteur dans les années 1990 et 2000 avec l’arrivée des outils de traduction assistée par ordinateur (TAO). En particulier, la traduction automatique a mauvaise presse auprès des futurs traducteurs, comme auprès des traducteurs en exercice d’ailleurs. Avec l’arrivée de la TA neuronale, ce sentiment de crainte en regard d’une technologie dont on lit parfois qu’elle les remplacerait s’est accentué, du fait des gains de fluidité en langue cible mais aussi des applications et accessoires développés pour le

grand public et aux performances parfois étonnantes. La spécificité de cette nouvelle technologie qui relève de l'intelligence artificielle (réseaux neuronaux) est qu'elle est opaque pour la plupart des acteurs du secteur : autant il était possible d'expliquer le fonctionnement de la TA à base de règles et de la TA statistique (Rossi 2017), autant expliquer le fonctionnement des réseaux neuronaux et la façon dont la machine propose une traduction est aujourd'hui difficile, en particulier auprès d'un public en formation. Les traductions proposées par les outils de TA sont aujourd'hui difficiles à prévoir ou à expliquer. La présence accrue de ce type d'outils sur le marché oblige néanmoins les formations en traduction à aborder le sujet et il importe de former les étudiants à ce nouvel outil. Une réflexion doit donc avoir lieu sur la façon d'intégrer cette nouvelle technologie au sein des formations (Moorkens 2018 ; Massey et Ehrensberger-Dow 2017).

Une approche pertinente consiste justement à tester la machine et à montrer ce qu'elle est capable de faire, mais aussi ce qu'elle n'est pas capable de faire. Cela permet non seulement de démystifier la machine (Moorkens 2018 : 2), mais aussi de préparer le terrain pour la post-édition qui doit suivre : pour quels phénomènes puis-je faire confiance à la machine ? À l'inverse, que dois-je absolument vérifier ? Une expérience telle que nous l'avons menée ici pour la paire de langues anglais-français, même si elle ne fournit pas toutes les réponses, permet (i) de montrer aux étudiants que la machine ne peut pas tout et est loin de respecter l'usage linguistique de la langue cible, pourtant attendu pour des traductions de haute qualité ; (ii) d'identifier les phénomènes linguistiques pouvant conduire à un manque d'idiomaticité ; et donc (iii) de leur faire prendre conscience des éléments sur lesquels porter leur attention dans le cadre de la post-édition. L'objectif final est de mettre au jour la plus-value du traducteur humain par rapport à la machine en ayant conscience de ses forces et de ses faiblesses. Nous avons pu nous-mêmes expérimenter cette approche et avons constaté qu'elle rassurait les étudiants. Massey et Ehrensberger-Dow (2017 : 307-308) parlent alors de « métacognition » : il s'agit d'amener les étudiants à prendre de la distance par rapport aux outils utilisés afin de maîtriser la technologie et non de la subir.

6. Conclusions

Dans cet article, nous avons proposé une analyse linguistique d'un double corpus de français traduit de façon automatique au moyen de deux outils différents. L'analyse a concerné toute une série de phénomènes linguistiques ayant la réputation de poser problème aux traducteurs anglais-français du fait d'une différence de fréquence entre les deux langues originales, en l'occurrence une fréquence plus élevée en anglais (notre langue source) qu'en français (notre langue cible). Les résultats montrent que de façon systématique, on constate une surreprésentation de ces phénomènes lexicaux et grammaticaux en français traduit automatiquement par rapport au français original, mais dans des proportions qui ne sont pas toujours similaires. L'hypothèse explicative la plus évidente semble alors être l'interférence de la langue source, mais nous avons montré qu'elle ne pouvait pas à elle seule expliquer les résultats obtenus. Une analyse qualitative complémentaire permettrait de mesurer l'ampleur de cette influence de la langue source sur les résultats obtenus. De la même manière, une comparaison avec du français traduit non pas par des outils de TA mais par des (bio)

traducteurs pourrait apporter un éclairage intéressant. En effet, l'existence de différences entre langue originale et langue traduite est un phénomène que l'on retrouve aussi du côté de la traduction humaine : de très nombreux travaux en traductologie de corpus ont mis au jour l'existence de ces différences intralangagières pour de nombreux phénomènes linguistiques dans de nombreuses langues du monde (Laviosa 2002 ; Olohan 2004), en d'autres termes l'existence du « troisième code » tel qu'il est défini par Frawley (1984). Ceci montre que même la traduction humaine professionnelle n'atteint pas cet idéal d'homogénéisation linguistique entre langue originale et langue traduite, sans que cela ne représente nécessairement un problème de qualité. Il s'agit là d'une étude à venir pour nos travaux de recherche.

REMERCIEMENTS

Je souhaite remercier la Direction générale de la traduction de la Commission européenne de m'avoir donné accès à l'outil *MT@EC/eTranslation*. Mes remerciements iront également aux collègues de l'UNINT de Rome, où j'ai pu présenter ce travail dans le cadre d'un *Translating Europe Workshop* (mai 2018). Je remercie également les étudiants du parcours de master Traduction Spécialisée Multilingue de l'Université de Lille pour leur contribution à la compilation du TSM Press Corpus.

NOTES

1. Le choix du terme *traduction automatique* peut prêter à débat, certains traducteurs réfutant la légitimité du mot *traduction* pour désigner le résultat d'un algorithme et lui préférant des termes comme *transposition* ou *équivalent*. Nous n'entrons pas dans le débat ici et utiliserons le terme *traduction automatique*.
2. *CSA Research* (anciennement *Common Sense Advisory*). Consulté le 17 décembre 2018, <<https://csa-research.com/>>.
3. *TAUS, the language data network*. Consulté le 17 décembre 2018, <<https://www.taus.net/think-tank/reports>>.
4. European Union of Associations of Translation Companies (2018) : *2018 Language Industry Survey – Expectations and Concerns of the European Language Industry*. Consulté le 30 juillet 2018, <https://www.euatc.org/images/2018_Language_Industry_Survey_Report.pdf>.
5. Le néologisme *biotraduction* et son dérivé *biotraducteur* que nous utilisons ici pour distinguer l'homme de la machine trouvent leur origine dans le roman de science-fiction *Le revenant de Fomalhaut* publié par Jean-Louis Trudel en 2002 (Froeliger 2013 : 20). Là encore, le choix de ce terme peut prêter à débat. Voir TRUDEL, Jean-Louis (2002) : *Le revenant de Fomalhaut*. Montréal : Éditions Mediaspaul.
6. WU, Yonghui, SCHUSTER, Mike, CHEN, Zhifeng, *et al.* (2016) : Google's Neural Machine Translation System: Bridging the gap between human and machine translation. *arXiv*. Consulté le 17 décembre 2018, <<https://arxiv.org/abs/1609.08144>>.
7. DEEPL GMBH (Dernière mise à jour : 5 décembre 2018) : *DeepL Translator*. Cologne : DeepL GmbH. Consulté le 17 décembre 2018, <<https://www.deepl.com/>>.
8. DEEPL GMBH (2008-) : *Linguee*. Cologne : DeepL GmbH. Consulté le 17 décembre 2018, <<https://www.linguee.com/>>.
9. DEEPL GMBH (2017-) : *How do we compare to the competition?* Cologne : DeepL GmbH. Consulté le 17 décembre 2018, <<https://www.deepl.com/press.html>>.
10. ANTHONY, Laurence (2018) : *AntConc*. Version 3.5.7. Tokyo : Waseda University. <<http://www.laurenceanthony.net/software>>.
11. ANTHONY, Laurence (2015) : *TagAnt*. Version 1.2.0. Tokyo : Waseda University. <<http://www.laurenceanthony.net/software>>.
12. Par structure existentielle, il convient de comprendre les structures en *il y a* + AVOIR permettant l'introduction en discours d'un référent nouveau. Sont donc exclues les occurrences du *il y a* dit prépositionnel (Grevisse 1936/1986) du type *il y a 6 mois*. Voir GREVISSE, Maurice et GOOSSE, André (1936/1986) : *Le Bon Usage*. 12^e éd. Paris/Gembloux : Ducolot.

13. Nous opérons ici une distinction désormais classique entre corpus comparables et parallèles: les premiers sont des corpus de données qui appartiennent à deux langues originales ou à deux variétés de la même langue (langue originale vs langue traduite, comme nous l'avons fait ci-dessus), tandis que les seconds, parfois appelés corpus de traduction, sont des corpus bilingues/multilingues qui contiennent textes originaux et leur(s) traduction(s) dans une ou plusieurs langue(s).
14. O'CUNNEEN, Pamela (23 janvier 2015): Suriname: the mysterious land of seven cultures. *The Telegraph*. Consulté le 5 décembre 2018, <<https://www.telegraph.co.uk/expat/expatlife/11357602/Suriname-the-mysterious-land-of-seven-cultures.html>>.
15. CARRINGTON, Damian (28 janvier 2015): Lancashire council defers Cuadrilla fracking decision. *The Guardian*. Consulté le 1^{er} décembre 2018, <<https://www.theguardian.com/environment/2015/jan/28/lancashire-council-defers-cuadrilla-fracking-decision>>.
16. MATHIESEN, Karl (29 octobre 2015): Will reintroduced lynx hunt Britain's sheep? *The Guardian*. Consulté le 2 décembre 2018, <<https://www.theguardian.com/environment/2015/oct/29/will-reintroduced-lynx-hunt-britains-sheep>>.
17. HOPE, Kerin (15 janvier 2015): Greece risks cash crunch if Syriza wins, finance minister warns. *The Financial Times*. Consulté le 2 décembre 2018, <<https://www.ft.com/content/67baa242-9ca8-11e4-a730-00144feabdc0>>.
18. SLEZAK, Michael (26 janvier 2015): La Niñas on the rise in climate change double whammy. *New Scientist*. Consulté le 2 décembre 2018, <<https://www.newscientist.com/article/dn26858-la-ninas-on-the-rise-in-climate-change-double-whammy/>>.

RÉFÉRENCES

- BAKER, Mona (1993): Corpus linguistics and translation studies: Implications and applications. In: Mona BAKER, Gill FRANCIS et Elena TOGNINI-BONELLI, dir. *Text and Technology*. Amsterdam/Philadelphie: John Benjamins, 223-250.
- BENTIVOGLI, Luisa, BISAZZA, Arianna, CETTOLO, Mauro, *et al.* (2016): Neural versus phrase-based machine translation quality: a case study. In: Jian SU, Kevin DUH et Xavier CARRERAS, dir. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. (EMNLP 2016: Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 1-5 novembre 2016). Stroudsburg: Association for Computational Linguistics, 257-267. Consulté le 17 décembre 2018, <<http://www.aclweb.org/anthology/D16-1000>>.
- BOJAR, Ondřej, CHATTERJEE, Rajen, FEDERMANN, Christian, *et al.* (2015): Findings of the 2015 workshop on statistical machine translation. In: Ondřej BOJAR, Rajan CHATTERJEE, Christian FEDERMANN, *et al.*, dir. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. (EMNLP 2015: Tenth Workshop on Statistical Machine Translation, Lisbonne, 17-18 septembre 2015). Stroudsburg: Association for Computational Linguistics, 1-46. Consulté le 17 décembre 2018, <<http://www.statmt.org/wmt15/pdf/WMT01.pdf>>.
- CAPPELLE, Bert et LOOCK, Rudy (2013): Is there interference of usage constraints? A frequency study of existential *there* is and its French equivalent *il y a* in translated vs. non-translated texts. *Target*. 25(2):252-275.
- CASTILHO, Sheila, MOORKENS, Joss, GASPARI, Federico, *et al.* (2017): A comparative quality evaluation of PBSMT and NMT using professional translators. In: Sadao KUROHASHI et Pascale FUNG, dir. *Proceedings of MT Summit XVI*. (Machine Translation Summit XVI, Nagoya, 18-22 septembre 2017). Vol. 1. Tokyo: Asia-Pacific Association for Machine Translation, 116-131.
- CHUQUET, Hélène et PAILLARD, Michel (1987): *Approche linguistique des problèmes de traduction anglais-français*. Paris: Ophrys.
- FEDERICO, Marcello, NEGRI, Matteo, BENTIVOGLI, Luisa, *et al.* (2014): Assessing the impact of translation errors on machine translation quality with mixed-effects models. In: Alessandro MOSCHITTI, Bo PANG et Walter DAELEMANS, dir. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (EMNLP 2014: Conference on Empirical Methods in Natural Language Processing, Doha, 25-29 octobre 2014). Stroudsburg: Association for Computational Linguistics, 1643-1653. Consulté le 17 décembre 2018, <<http://www.aclweb.org/anthology/D14-1172>>.

- FRAWLEY, William (1984): Prolegomenon to a theory of translation. In: William FRAWLEY, dir. *Translation: Literary, Linguistic and Philosophical Perspectives*. Newark: University of Delaware Press, 250-263.
- FROELIGER, Nicolas (2013): *Les Noces de l'analogique et du numérique – De la traduction pragmatique*. Traductologiques. Paris: Les Belles Lettres.
- GUILLEMIN-FLESCHER, Jacqueline (1986): *Syntaxe comparée du français et de l'anglais, Problèmes de traduction*. Paris: Ophrys.
- HARTLEY, Anthony et POPESCU-BELIS, Andrei (2004): Évaluation des systèmes de traduction automatique. In: Stéphane CHAUDIRON, dir. *Évaluation des systèmes de traitement de l'information*. Sciences et technologies de l'information. Paris: Hermès, 311-335.
- ISABELLE, Pierre, CHERRY, Colin et FOSTER, George (2017): A challenge set approach to evaluating machine translation. In: Martha PALMER, Rebecca HWA et Sebastian RIEDEL, dir. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. (EMNLP 2017: Conference on Empirical Methods in Natural Language Processing, Copenhagen, 7-11 septembre 2017). Stroudsburg: Association for Computational Linguistics, 2486-2496.
- KOEHN, Philipp (2010): *Statistical Machine Translation*. Cambridge: Cambridge University Press.
- KOEHN, Philipp et GERMANN, Ulrich (2014): The impact of machine translation quality on human post-editing. In: Ulrich GERMANN, Michael CARL, Philipp KOEHN, et al., dir. *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*. (EACL 2014: 14th Conference of the European Chapter of the Association for Computational Linguistics, Göteborg, 26-30 avril 2014). Stroudsburg: Association for Computational Linguistics, 38-46. Consulté le 17 décembre 2018, <<http://www.aclweb.org/anthology/W14-0300>>.
- LAPSHINOVA-KOLTUNSKI, Ekaterina (2015): Variation in translation: evidence from corpora. In: Claudio FANTINUOLI et Federico ZANETTIN, dir. *New Directions in Corpus-based Translation Studies*. Berlin: Language Science Press, 93-114.
- LAVIOSA, Sara (2002): *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam/New York: Rodopi.
- LOOCK, Rudy (2016): *La traductologie de corpus*. Villeneuve d'Ascq: Presses universitaires du Septentrion.
- LOOCK, Rudy (à paraître): Parce que «grammaticalement correct» ne suffit pas: le respect de l'usage grammatical en langue cible. In: Michel BERRÉ, Céline LETAWE, Hedwig REUTER, et al., dir. *La formation grammaticale du traducteur: enjeux didactiques et traductologiques* (Quelle formation grammaticale pour de futurs traducteurs? Mons, 9-10 mars 2017). Villeneuve d'Ascq: Presses universitaires du Septentrion.
- MACKETANZ Vivien, AVRAMIDIS, Eleftherios, BURCHARDT, Aljoscha, et al. (2017): Machine translation: Phrase-Based, Rule-Based and Neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*. 17(2):28-43.
- MASSEY, Gary et EHRENSBERGER-DOW, Maureen (2017): Machine learning: Implications for translator education. *Lebende Sprachen*. 62(2):300-312.
- MOORKENS, Joss (2018): What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*. 12(4):375-387.
- OLOHAN, Maeve (2004): *Introducing Corpora in Translation Studies*. Londres/New York: Routledge.
- PAPINENI, Kishore, ROUKOS, Salim, WARD, Todd, et al. (2002): Bleu: a method for automatic evaluation of machine translation. In: Pierre ISABELLE, Eugene CHARNAK et Dekang LIN, dir. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. (ACL-02: 40th Annual Meeting of the ACL, Philadelphie, 7-12 juillet 2002). Stroudsburg: Association for Computational Linguistics, 311-318. Consulté le 17 décembre 2018, <<https://dl.acm.org/citation.cfm?doid=1073083.1073135>>.
- POPOVIC, Maja, AVRAMIDIS, Eleftherios, BURCHARDT, Aljoscha, et al. (2013): Learning from human judgments of machine translation output. In: Khalil SIMA'AN, Mike L. FORCADA,

- Daniel GRASMICK, *et al.*, dir. *Machine Translation Summit XIV (Nice, France, 2-6 September 2013): Proceedings*. (Machine Translation Summit XIV, Nice, 2-6 septembre 2013). Allschwil: European Association for Machine Translation, 231-238.
- ROSSI, Caroline (2017): Introducing statistical machine translation in translator training: From uses and perceptions to course design and back again. *Revista Tradumàtica. Tecnologies de la Traducció*. 15:48-62.
- SCHMID, Helmut (1997/2003): Probabilistic part-of-speech tagging using decision trees. In: Danny B. JONES et Harold SOMERS, dir. *New Methods in Language Processing*. (NeMLaP: International Conference on New Methods in Language Processing, Manchester, 14-16 septembre 1994). Londres/New York: Routledge, 154-164. Consulté le 17 décembre 2018, <<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf>>.
- VINAY, Jean-Paul et DARBELNET, Jean (1958): *Stylistique comparée du français et de l'anglais*. Paris: Didier.