Exploration de la traduction automatique neuronale espagnol-français : Pour une traductologie de corpus appliquée à l'analyse des outils de traduction

Cristian Valdez, Maria Lomeña Galiano

Revue de Traduction et Langues Volume 20Numéro 01/2021, pp. 85-111
Journal of Translation and Languages مجلة الترجمة واللغات
ISSN (Print): 1112-3974 EISSN (Online): 2600-6235

Exploration de la Traduction Automatique Neuronale Espagnol-Français : Pour une Traductologie de Corpus Appliquée à l'Analyse des Outils de Traduction

Exploring Spanish-French Neural Machine Translation: Towards a Corpus-Based Translation Studies Approach for the Analysis of Translation Tools

Dr. Cristian Valdez
valdezcristian1@gmail.com
LIDILE Université Rennes 2-France
0000-0003-2291-8085
Dr. María Lomeña Galiano
marialomena@gmail.com
LIDILE Université Rennes 2 –France
0000-0003-4187-7135

Pour citer cet article :
Valdez, C & Galiano, M-L. (2021). Exploration de la traduction automatique neuronale espagnol-français : Pour une traductologie de corpus appliquée à l'analyse des outils de traduction. *Revue Traduction et Langues 20 (1),* 85-111.

Reçu : 29/03/2021 ; **Accepté** : 19/06/ 2021**, Publié :** 31/ 08/ 2021

Abstract: *Neural machine translation (NMT) systems, based on large volumes of bilingual and monolingual data, represent a significant leap forward in the processing of linguistic data. They are capable of proposing a target text that is natural, fluid and idiomatic. The field of translation has gained much progress thanks to this technology wherein the working procedures are witnessing an important transformation. Automating the translation process requires not only to rethink the professional practices but also the objectives and the methods governing training in translation. In some professional settings, they are becoming a complementary tool for translation. The challenge is therefore to know the advantages and weaknesses of their use, which would make it possible to approach the post-editing phase more effectively. From a textometric and translation studies approach, this article proposes the exploration of a translation corpus composed of journalistic texts in Spanish and their translations into French by the generic NMT system DeepL. The exploration of the data comprises three stages: a textual approach, the analysis of the units selected by means of the cartographic representation of the correspondences and the return to the text, by examining the units in context. A sample of the most frequent lexemes in the corpus was qualitatively analysed using the MkAlign software in order to evaluate the output of the automatic*

L'auteur correspondant : Cristian Valdez

information transfer. This is meant to verify the accuracy of the automatic translation process between two languages and the extent to which, if any, it causes information loss. The analysis indicates that, in most cases, the information content of the lexemes studied was correctly transferred into the target language. Furthermore, appropriate translation choices were revealed when processing anaphoric expressions. However, the target text contains transfer errors concerning some neologisms, proper names and parasynonyms.
Keywords: *translation corpora - Spanish-French - neural machine translation - Corpus-Based Translation studies.*
Résumé : *Les systèmes de traduction automatique dite neuronale (TAN), basées sur de grandes masses de données bilingues et monolingues, représentent un saut qualitatif dans le traitement des données linguistiques. Dans certains milieux professionnels, ils se généralisent et deviennent ainsi un véritable outil de traduction. L'enjeu est donc de connaître les avantages et les faiblesses de leur utilisation, ce qui permettrait d'aborder la phase de post-édition de manière plus efficace. Cet article, fondé sur une approche textométrique et traductologique, propose l'exploration d'un corpus parallèle constitué de textes journalistiques en espagnol et de leur traduction en français réalisée par le système générique de TAN DeepL. Un échantillon formé des lexèmes les plus fréquents du corpus a été analysé qualitativement à l'aide du logiciel MkAlign dans le but d'évaluer le résultat du transfert automatique d'informations. L'analyse indique que, dans la majorité des cas, le contenu informationnel concernant les lexèmes étudiés a été correctement transféré en langue cible. De plus, ont été mis en évidence des choix de traduction pertinents au moment du traitement d'expressions anaphoriques. Cependant, le texte cible comporte des erreurs de transfert concernant certains néologismes, des noms propres et des parasynonymes.*
Mots clés : *corpus parallèle - espagnol-français - traduction automatique neuronale - traductologie de corpus.*

1. Introduction

En 2010, Matthieu Guidère faisait part du renouveau de la traduction automatique (TA) « après des hauts et des bas » (2010, p. 157). Une décennie plus tard, ils sont nombreux les traductologues et formateurs qui témoignent des avancées concernant la qualité des traductions automatiques, surtout lorsqu'il s'agit de la traduction dite neuronale, fondée sur l'apprentissage profond ou deep learning (Loock, 2018 ; Hernández-Morin, 2019 ; Yvon, 2019 ; Moorkens & Way, 2019 et Barbin, 2020). Ces systèmes de dernière génération sont actuellement capables de proposer un texte cible naturel, fluide et idiomatique. Certains travaux récents (cf. Esperança-Rodier et Becker, 2018, Poibeau, 2019) montrent que la qualité du résultat dépend de plusieurs variables : le volume d'information disponible pour chaque combinaison linguistique, la directionnalité de la traduction et le type de texte.

Le marché de la traduction profite de ces progrès technologiques. La traduction automatique neuronale (TAN), comme le suggère Loock (2019), est en train de transformer les méthodes de travail ainsi que le modèle économique du secteur. Désormais, une grande proportion d'entreprises de services linguistiques en Europe a recours à ce type d'outil de traduction (op. cit.).

L'automatisation concerne certains types de textes et des contextes d'application spécifiques. Elle intervient tantôt dans un but purement informatif (pour une traduction non professionnelle), tantôt comme la première étape d'une traduction professionnelle complétée ultérieurement par une révision humaine, car, malgré les améliorations récentes, la machine ne dépasse pas l'activité traduisante humaine en termes de qualité (cf. Hutchins, 2005 ; Hernández-Morin, 2019 ; Poibeau, 2019 ; Yvon et Sadaf, 2020). Par conséquent, la post-édition des sorties de TA, à savoir la correction des textes produits «

à l'état brut » par un système informatique (Robert, 2010, p. 139), reste une tâche incontournable pour garantir la qualité des services.

Les avancées dans l'automatisation du processus de traduction obligent à repenser non seulement les bonnes pratiques professionnelles, mais aussi les objectifs et les méthodes d'enseignement des formations en traduction. Tel qu'il a été préconisé par le dernier référentiel de compétences (Groupe d'experts EMT, 2017), TA devrait faire l'objet des enseignements spécifiques destinés aux apprentis traducteurs.

En ce sens, il s'avère nécessaire d'orienter la recherche vers une traductologie appliquée aux outils de traduction automatique. Ce travail vise donc à une meilleure compréhension de la TAN. Si les études ne manquent pas dans le domaine de la traduction anglais-français, la revue bibliométrique à partir de BRITA (Franco Aixelà, 2001-2020) montre qu'il existe peu de travaux associant les études en TAN au couple de langues espagnol-français. Par conséquent, cet article propose l'analyse outillée d'un corpus parallèle de textes économiques issus de la presse espagnole traduits en français de façon automatique avec le système DeepL. À partir d'une approche combinant exploration textométrique et analyse traductologique, la finalité est de révéler les points forts et les points faibles de l'outil de TAN pour ainsi préparer et étayer le travail de post-édition.

2. La traduction automatique neuronale

Depuis les années 1960, les outils de TA n'ont pas cessé d'évoluer. Les premières versions, basées sur des dictionnaires et sur l'application de règles grammaticales, ont ensuite été remplacées par des systèmes fondés sur le traitement statistique de corpus bilingues alignés. Ces derniers, dont le plus connu reste certainement Systran, s'attèlent à la recherche de la meilleure équivalence de traduction à partir de l'analyse des probabilités d'apparition des différentes séquences.

Parmi les aspects positifs des systèmes de TA basés sur des règles ou sur le traitement statistique, il a été question de l'uniformisation terminologique (Diéguez, 2001) et le gain en productivité, lorsque le résultat brut est exploitable pour la post-édition. Cependant, de nombreuses études ont tout de même relevé un certain nombre de faiblesses lors du transfert linguistique. Plus précisément, ces investigations ont mis en lumière des limitations concernant les omissions ou le traitement de phrases longues (Hutchins, 2005), la difficulté à lever des ambiguïtés lexicales liées à la polysémie ou des dysfonctionnements lors de l'identification du référent de certaines expressions anaphoriques (Ping, 2009, p. 166). Globalement, il semblerait que la TA produise davantage d'erreurs dans la traduction de textes spécialisés (Martínez, 2019, p. 322).

Certaines des imperfections identifiées pour les outils cités semblent être mieux gérées par le dernier avatar en date de la TA, à savoir la traduction automatique neuronale (TAN) ou par apprentissage profond. La TAN a signifié un saut qualitatif dans le traitement de données linguistiques, à tel point que des logiciels commercialisés ou gratuits fonctionnant auparavant uniquement sur la base d'un calcul statistique des équivalences font la transition vers l'approche neuronale. Cette dernière est adoptée, entre autres, par les services développés par Systran, Google Traduction, Prompt ou par le système eTranslation de la Commission européenne destiné aux administrations publiques.

Le processus développé par la TAN s'appuie sur l'analyse statistique et sémantique des mots lexicaux présents dans de larges corpus monolingues et parallèles multilingues, dans le but d'identifier les constructions les plus fréquentes ou, au contraire, les plus rares en fonction des contextes de production. Cette gestion des informations donne suite à une représentation lexico-sémantique de grande richesse qui rend possible la production d'une traduction en quelques secondes. Le fonctionnement de ce système, fondé sur des algorithmes d'Intelligence Artificielle, se caractérise en outre par la possibilité d'auto-apprentissage. Autrement dit, le système s'entraîne et progresse grâce aux données recueillies lors de la traduction de nouveaux segments ainsi qu'à partir des corrections apportées par les utilisateurs humains dans les sorties brutes.

Koehn et Knowles (2017) et Poibeau (2019) indiquent que si le système de TAN dispose des données importantes, la sortie brute sera de bonne qualité. En ce sens, pour certaines combinaisons de langues et certains types de textes, la traduction proposée par le système est exploitable. Poibeau (2019, p. 190) souligne, par exemple, que la traduction automatique des textes journalistiques, techniques ou juridiques est « très bonne » lorsque les langues impliquées sont l'anglais et le français, à condition qu'ils soient bien rédigés et édités en langue source. De même, la proximité entre les langues impliquées ainsi que la directionnalité de la traduction sont des éléments contribuant à un meilleur aboutissement du processus : la traduction de l'anglais vers l'allemand semble plus satisfaisante que celle réalisée dans le sens inverse dû à la complexité des mots composés en allemand (op. cit. , p. 137).

Pour ce qui est des limites concernant la TAN, elles portent entre autres sur la traduction des expressions figées (Poibeau, 2018, p. 148), ainsi que sur le traitement de néologismes et de la terminologie (Tinsley, 2017 ; Hernández-Morin, p. 2019), et des noms propres (Yvon & Sadaf, 2020). À l'instar des systèmes statistiques, la TAN semble aussi être peu performante pour la traduction de phrases longues (Toral & Sánchez-Cartagena, p. 2017). En outre, comme il en découle du type de traitement informatique développé par la TAN, le transfert linguistique est moins optimal lorsqu'il s'agit de traduire depuis ou vers des langues peu dotées, c'est-à-dire celles dont les données parallèles sont moins volumineuses. Enfin, certaines faiblesses concernent également les textes à caractère créatif et artistique (Guidère, 2010, p. 153) ou la localisation de jeux vidéo (Hernández-Morin, 2019, pp. 244-245).

En résumé, de par la nature même des systèmes de TAN, la qualité de la sortie brute est tributaire de nombreux paramètres (quantité de données disponibles, type de texte, directionnalité, entre autres). Si le but est d'étudier le processus de traduction de l'espagnol vers le français, il est risqué de transposer directement les conclusions obtenues dans une étude focalisée sur une combinaison de langues différentes. Cette étude se propose donc d'évaluer spécifiquement les résultats obtenus pour la paire espagnol-français. L'évaluation portera plus précisément sur le transfert du contenu informationnel en soi entre le texte source et le texte cible.

3. Méthodologie de la recherche

L'objectif de l'étude présentée dans cet article est de vérifier si le processus de traduction automatique entre deux langues typologiquement proches, à savoir l'espagnol et le français, entraîne une perte d'informations. Afin d'évaluer la sortie brute d'un outil de

TAN, un corpus parallèle, également appelé corpus de traduction, a été constitué. Cette recherche s'appuie donc sur l'étude d'un ensemble de données composé de textes originaux et de leur traduction. La méthodologie d'exploration, quant à elle, a été définie à partir des apports de la traductologie de corpus et, plus précisément, des études textométriques.

3.1 Constitution du corpus parallèle de traduction automatique espagnol-français

Les textes retenus pour cette étude appartiennent au genre journalistique, soit un type d'écrits pour lequel la TAN a fourni des résultats satisfaisants dans la combinaison linguistique français-anglais (Poibeau, 2018, p. 190). Plus exactement, le corpus de travail comprend vingt articles de presse rédigés en espagnol et leur traduction en français, réalisée avec la dernière version 2020, en ligne et gratuite, de DeepL. Les articles sélectionnés, ayant pour thématique commune la crise sanitaire provoquée par la COVID-19, ont été publiés entre mars et décembre 2020 dans la rubrique *Economía* du journal espagnol *El País*. Étant destinés au grand public, ces articles sont considérés comme des textes non techniques car le public cible n'est pas spécialisé dans le domaine abordé, ici l'économie (Hurtado Albir, 2001).

Pour ce qui est de la traduction automatique en elle-même, il convient de faire deux remarques. Premièrement, certains articles journalistiques ont été abrégés car la version de DeepL qui a été utilisée pour constituer le corpus n'accepte pas les textes au-delà de 5 000 caractères. Pour cette adaptation, le découpage initial en paragraphes et les enchaînements logiques ont été scrupuleusement respectés. Deuxièmement, partant du principe qu'il existe pour chaque texte « une pluralité de traductions parfaitement recevables » (Keromnes, 2016, p. 102), DeepL propose une multitude de choix de traduction. Pour cette étude, c'est la première version en langue cible proposée par le système de TAN qui a été choisie. Autrement dit, la sortie brute n'a subi aucun traitement.

Une fois les textes traduits, le logiciel MkAlign (Fleury & Zimina, 2007) a permis la création du bi-texte en alignant les segments de la version en espagnol (appelé volet source) avec ceux de la version en français (volet cible). En raison du type d'analyse envisagé, l'alignement a été fait en réalisant une segmentation du corpus par phrase. Une correction manuelle a été nécessaire afin d'éviter tout décalage dans l'alignement. DeepL est en effet en mesure de proposer une traduction ayant un découpage par phrase différent des textes originaux.

Le corpus parallèle ainsi constitué représente un total de 15 126 mots pour le volet source et de 15 230 mots pour le volet cible issu de la TAN.

3.2 Exploration outillée du corpus parallèle

S'inspirant des études précédentes réalisées dans le cadre de la traductologie de corpus, l'exploration des données comporte trois étapes : une approche textométrique, l'analyse des unités retenues par le biais de la représentation cartographique des correspondances et le retour au texte, soit l'examen des unités en contexte.

3.2.1 Première approche par le prisme textométrique

L'analyse textométrique prend appui sur les données textuelles d'un corpus numérique et vise à réaliser une description statistique basée sur le vocabulaire. Elle permet l'observation quantitative des cooccurrences, des segments répétés ou des concordances des unités choisies (*cf*. Fleury & Zimina, 2007).

L'analyse du vocabulaire du corpus parallèle faisant l'objet de cette étude a été réalisée avec le logiciel MkAlign. Ce système constitue des listes de mots (appelées « dictionnaires ») pour chacun des volets en ordonnant les lexèmes (« formes ») dans l'ordre croissant ou décroissant en fonction de la fréquence d'apparition. À partir de ces informations, un échantillon des termes les plus fréquents du volet source a été sélectionné. Les unités suivantes ont donc été retenues : *España, pandemia, economía, crisis, COVID-19, año, empresa*. Pour chacune d'elles, l'examen du dictionnaire du volet cible a permis ensuite d'identifier les correspondants directs, c'est-à-dire les mots qui ont une fréquence similaire et la même signification en langue. Il s'agit respectivement de *Espagne*, *pandémie*, *économie*, *crise*, *COVID-19, année* et *entreprise*.

Chaque paire de mots a été étudiée séparément afin de déterminer s'il existe des écarts en termes de fréquence entre les deux volets du bi-texte. Cette comparaison a permis de classifier les correspondances en trois catégories selon les critères définis par Zimina (2004, 2005) : univoques, quasi-univoques et multiples.

La première catégorie, soit les correspondances dites univoques, regroupe les cas où la fréquence d'une unité du volet source est identique à celle de son correspondant du volet cible. Dans ces conditions, il est possible de présumer que chaque apparition d'un lexème en espagnol trouvera systématiquement la même traduction en français (et vice-versa). Les unités à correspondance quasi-univoques, pour leur part, attestent d'une fréquence similaire, bien que différente, entre les deux volets du bi-texte. Dans ce cas, l'écart numérique peut indiquer des équivalences traductionnelles différentes pour un même mot ou, dans le corpus analysé, une erreur de transfert. Enfin, la catégorie des correspondances multiples concerne les écarts de fréquence importants. En règle générale, il s'agit de mots « dotés d'un large éventail de sens dans le corpus » (Zimina, 2004, p. 1196) qui forment de réseaux de correspondances complexes. L'écart peut non seulement s'expliquer par la polysémie propre aux langues naturelles, mais aussi par l'existence de parasynonymes de l'un des volets traduits par une seule et même unité dans l'autre partie du bi-texte.

Force est de souligner qu'à cette première étape de l'exploration, la catégorisation initiale prend uniquement en compte la fréquence d'apparition des unités. L'étiquette de correspondance univoque, quasi-univoque ou multiple repose sur une analyse quantitative et non nécessairement sémantique. Certes, l'identité ou les écarts de fréquences peuvent être une conséquence des valeurs sémantiques des mots sélectionnés. Cependant, les différences peuvent également avoir pour origine la création d'équivalences traductionnelles allant au-delà de la simple correspondance directe. Étant donné que le volet cible est issu de la traduction automatique, il faudra donc également intégrer l'existence de traductions inexactes ou incomplètes aux explications possibles d'un écart. Seules les deux autres étapes de l'exploration permettront d'affiner la description des correspondances.

3.2.2 Approche cartographique du bi-texte

La deuxième étape de l'exploration consiste à interroger le bi-texte à partir d'une représentation graphique des correspondances des unités sélectionnées dans la première étape. La démarche correspond à ce qui a été qualifié de « résonance textuelle » (Salem 2004 ; Zimina, 2004 ; Fleury et Zimina, 2014). Cette technique permet d'analyser la mise en relation entre les unités des textes du volet source et du volet cible « à travers une transition fondée sur une correspondance topographique entre ces textes » (Salem, 2004, p. 992).

Concrètement, MkAlign permet une représentation visuelle des segments du bi-texte. Chacune des phrases du volet source et du volet cible est représentée par un carré, construisant ainsi une carte du corpus parallèle, comme le montre la figure (1).

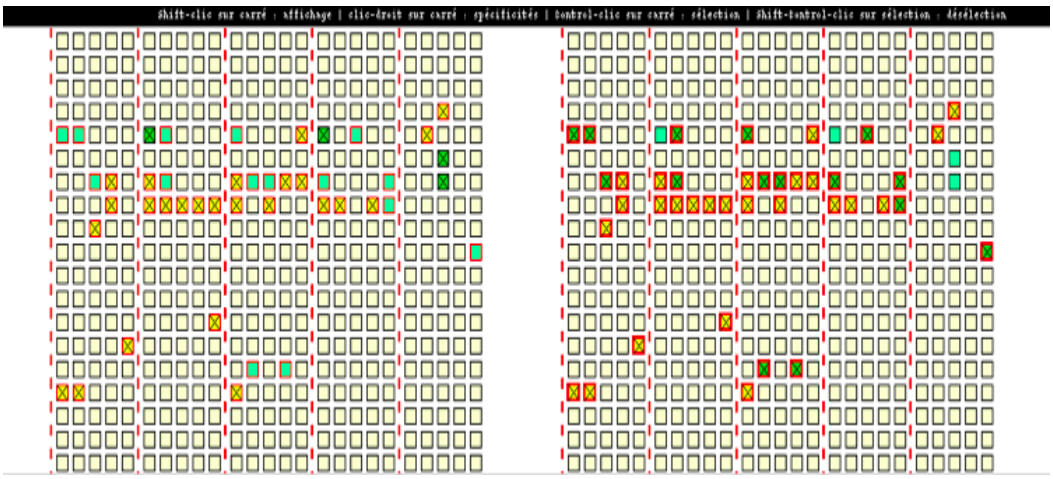

Figure 1. Carte représentant la mise en correspondance de empresa et entreprise

L'ensemble de carrés à gauche de l'image 1 correspond au volet source, tandis que celui de droite est associé au volet cible. De plus, comme il est question d'un corpus aligné, chaque carré d'un volet est associé à son homologue dans l'autre volet ; il s'agit d'une carte de sections parallèles (Zimina, 2005). C'est précisément là que réside l'intérêt de la représentation cartographique.

Une fois la carte créée, MkAlign permet de réaliser des requêtes par mots clés dans l'un des volets. Le système entoure alors les carrés où se trouvent les occurrences du lexème en question. Il est également possible de choisir un mot précis pour chacun des volets ; le terme ayant été saisi en premier dans le système est considéré comme le *terme d'induction* et constitue le repère qui amorce le processus de résonance. À l'aide d'un code couleur spécifique, le logiciel distingue alors l'absence ou la présence simultanée des unités recherchées dans chacun des volets. Par exemple, l'image 1 montre le résultat obtenu de la requête du couple de mots *empresa* (dans le volet source) et *entreprise* (dans le volet cible). Grâce à cette fonctionnalité, la carte devient un outil qui rend possible l'observation des correspondances de traduction.

Ce panorama peut être complété par l'observation des lexèmes dans leur contexte. Pour ce faire, en cliquant sur les différentes sections (les différents carrés), le système affiche la phrase du volet en question ainsi que l'énoncé de l'autre volet qui lui est

rattaché. Cet examen manuel se révèle utile dans deux cas importants. Premièrement, si le système indique, par le biais de la couleur correspondante, que le terme recherché se trouve dans une section, aucune information concernant la quantité d'occurrences n'est proposée. Afin de repérer les éventuelles répétitions à l'intérieur des phrases, une exploration manuelle de chaque section s'avère nécessaire. Cette vérification a été effectuée systématiquement dans l'analyse de tous les lexèmes retenus. Deuxièmement, quand la représentation topographique signale des cas d'asymétrie entre les unités étudiées, il est possible d'obtenir des informations supplémentaires sur la traduction. Plus concrètement, d'après l'image 1, on peut identifier une section où le mot *empresa* se trouve dans le volet source, sans que la forme *entreprise* ne soit présente dans la phrase homologue du volet cible. En sélectionnant l'un de ces carrés, on obtient les données de l'image 2.

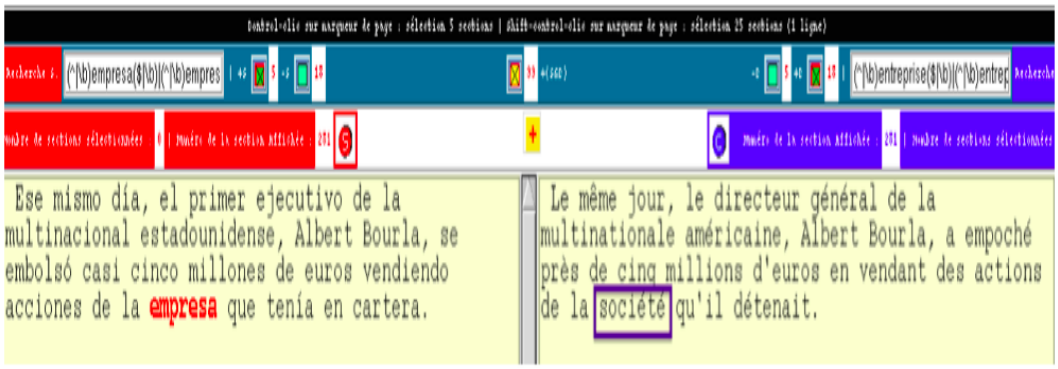

Figure 2. Inspection des segments en cas de non correspondance

Étant donné que le système affiche les mots recherchés en gras, il est facile d'examiner rapidement les phrases et d'analyser les correspondances établies. Dans l'exemple, le retour au texte dans une section qui n'intègre pas la correspondance visée (*empresa - entreprise)* fait ressortir un nouveau *terme induit* dans le volet cible, à savoir *société.* Le processus de résonance dévoile alors un réseau de correspondances plus large que prévu. Par conséquent, l'objectif est d'obtenir le panorama le plus complet possible.

Les résultats de l'exploration topographique peuvent alors conduire à un processus de résonance textuelle réitérée, surtout dans le cas des unités à correspondances multiples. Pour ce qui est de la correspondance *empresa-entreprise*, il sera dès lors possible de définir un nouveau terme d'induction pour amorcer un deuxième processus de résonance. Par exemple, si le mot *société* est défini comme terme d'induction, dans cette direction, la démarche d'exploration fait ressortir en espagnol *compañía* comme nouveau terme induit dans le volet source. Ce dernier devient ensuite terme d'induction pour recommencer les requêtes et ainsi de suite. Le logiciel permet de réaliser ce processus de manière successive.

En somme, le processus de résonance textuelle permet d'interroger le corpus à partir des correspondances lexicales retrouvées lors d'une première exploration quantitative. Il construit le réseau sémantique des unités analysées par un système de va-et-vient entre les volets. Les informations obtenues seront ensuite complétées par une analyse des unités en contexte. La représentation cartographique du corpus parallèle

constitue donc l'étape charnière entre l'analyse textométrique et la vérification qualitative des équivalences de traduction.

3.2.3 Vérification des équivalences de traduction par le retour au texte

Comme l'explique Pincemin (2020), pour analyser plus en détail les lexèmes ou les syntagmes sélectionnés dans un corpus, la perspective textométrique repose également sur le « retour au texte ». Cette démarche se traduit par la possibilité de faire une analyse qualitative et d'intégrer, par la même occasion, une approche traductologique contextualisée des unités étudiées.

Partant des résultats obtenus au cours des étapes précédentes, la troisième phase consiste à analyser les correspondances du point de vue de leur actualisation textuelle concrète. Autrement dit, l'objectif est de vérifier les équivalences de traduction entre les deux volets du bi-texte. En naviguant manuellement parmi les sections du corpus, il est possible d'analyser des phrases et des textes et donc d'observer les procédés de traduction (transposition, modulation, calque, emprunt, etc.) qui ont été opérés lors du processus de traduction automatique.

Le retour au texte permet de répondre à la question de départ de cette étude. En d'autres termes, il permet d'évaluer précisément si le système DeepL parvient à une équivalence qui respecte le contenu informationnel du texte source. Pour cette évaluation, le résultat de la traduction automatique sera examiné par rapport à la *fidélité* au texte original. Cette notion, largement débattue dans le champ de la traductologie, est à comprendre ici comme l'adéquation entre les énoncés du texte source et du texte cible quant à leur contenu informatif et leur message (Dancette, 1989, p. 99). Il s'agit de la fidélité au contenu et à la transmission des informations référentielles en langue cible. La lecture des contextes pour l'analyse des équivalences amènera donc à parler de fidélité au contenu informationnel ou, au contraire, de perte et/ou d'omission.

Outre l'évaluation de la fidélité, la lecture des contextes permet d'identifier en parallèle d'autres limites, ou d'autres aspects positifs, dans le produit de la TAN, lesquels ne seront que très brièvement évoqués dans cet article. À cet égard, il sera possible de parler de la *qualité* de la traduction. Il est important de rappeler que le jugement de qualité ne saura être circonscrit qu'à cette étude ; il dépendra en effet, dans d'autres contextes, de la fonction de la traduction et d'autres critères fixés à la demande du client ou d'autres parties prenantes de la prestation de service.

Par ailleurs, il convient de noter que cette troisième étape permet également d'affiner les résultats des deux autres phases. Par exemple, le retour au texte met en évidence les homographes de certaines unités retenues. Les occurrences qui sortaient du champ sémantique visé ont été écartées de manière à améliorer l'analyse ainsi que la représentation des résultats sur les tableaux présentés ci-après.

En résumé, les informations obtenues à chacune des trois étapes interagissent. Suivant cette méthodologie en trois phases à l'aide du logiciel de textométrie MkAlign, l'exploration a consisté à observer d'un point de vue quantitatif la mise en correspondance des unités entre les deux volets pour ensuite réaliser une évaluation qualitative du bi-texte. Autrement dit, l'analyse part des correspondances pour arriver aux équivalences traductionnelles proposées par DeepL.

4. Analyse exploratoire des unités. Discussion des résultats

Les lexèmes les plus fréquents du volet source ont été analysés en suivant la méthodologie tripartite décrite plus haut. Comme le rappelle Zimina (2005), l'étude des textes depuis une approche textométrique se fonde sur une analyse du vocabulaire sans *a priori.* Ainsi, la description qui suit regroupe les unités dans des catégories selon les résultats spécifiques obtenus dans cette recherche.

4.1 *Le l'univocité présumée à des équivalences instables*

Parmi les mots les plus récurrents du volet source figurent *España* et *pandemia*. Étant donné les critères de constitution du corpus, cette fréquence s'explique aisément : les articles sélectionnés ont été publiés dans ce pays ibérique et abordent la question de la crise liée au coronavirus. Les correspondants directs dans le volet cible (*Espagne* et *pandémie* respectivement) sont également fréquents.

Tableau 1.

Fréquences des unités à correspondance univoque

Fréquence	Volet cible (TAN)	Fréquence	Volet source
54	Espagne	54	España
40	Pandémie	40	Pandemia

Les données quantitatives du tableau 1 révèlent qu'il s'agit d'un cas où la fréquence des unités du volet source est équivalente à celle des correspondants du volet cible. Du point de vue du nombre d'occurrences, il serait possible d'affirmer qu'il s'agit de mots à correspondance univoque.[1] La traduction d'un toponyme connu tel que *España* n'est généralement pas contrainte à la variation. Ceci permettrait de justifier l'équivalence univoque décelée entre *España* et *Espagne* d'après les données chiffrées du corpus. Le caractère hautement spécialisé du terme *pandemia* a également une incidence dans la création de ce type de correspondance univoque.

Cette première approximation strictement quantitative semble indiquer que, pour les deux unités étudiées, il n'existe pas de perte d'information entre le volet source et le volet cible. Pour ce qui est du nom *España,* l'examen de la représentation visuelle construite avec MkAlign et le retour au texte permettent de compléter ce panorama. Selon les données du corpus, la traduction de ce toponyme par son correspondant connu avec une forme francisée (adaptation grapho-phonique) *Espagne* est pertinente et ce à deux niveaux. Premièrement, du point de vue du contenu informationnel, la version cible ne comporte pas de perte d'information essentielle, même dans le cas de remaniement de phrases complexes. Deuxièmement, du point de vue de la syntaxe, le système de TAN respecte les contraintes propres à chaque langue. En d'autres termes, il distingue les contextes où, pour la traduction vers le français, le toponyme doit être précédé d'un

[1] Le toponyme *España* pourrait être restitué dans une construction syntaxique faisant intervenir non pas le nom mais l'adjectif gentilé (*la crisis en España* par *la crise espagnole*, par exemple). Dans le corpus d'étude, ce type de transposition n'a pas été retrouvé : la traduction s'apparente plus à une traduction formelle. L'analyse d'un corpus de données plus large permettrait de confirmer s'il s'agit d'une stratégie de traduction programmée dans le traducteur automatique.

article (lorsqu'il constitue le noyau d'un syntagme nominal, par exemple *L'Espagne s'est montrée reconnaissante*) de ceux où l'omission de l'article est la norme (dans un syntagme prépositionnel, par exemple *En Espagne, ils prévoient une augmentation des crédits douteux*).

La représentation cartographique a permis néanmoins de détecter une différence entre les deux parties du bi-texte. Dans le volet source, il existe une occurrence de *España* n'étant pas traduite par *Espagne* dans le volet cible (exemple 1) ; et, inversement, dans le volet cible il y a une occurrence de *Espagne* sans que dans le volet source le mot *España* ne soit présent (exemple 2). Il s'agit de deux cas de figure où le nom *Espagne* prend un autre sens : il n'est pas un toponyme mais il fait partie du nom de l'association *España de noche,* organisation du secteur de la vie nocturne (exemple 1) ou du nom d'une entité réunissant des entrepreneurs et des investisseurs (exemple 2).

Así se manifiesta Ramón Mas, presidente de **España** de Noche, la asociación que agrupa a las empresas de ocio nocturno	C'est l'avis de Ramón Mas, président de *Spain* de Noche, l'association qui regroupe les entreprises de la vie nocturne

según datos del Mapa del Emprendimiento 2020, elaborado por el *Spain* Startup-South Summit	selon les données de la carte de l'entrepreneuriat 2020, produite par le Sommet de **l'Espagne** sur la création d'entreprises du Sud

Dans ces deux contextes, le nom *Espagne* n'apparaît pas comme désignateur spatial mais il intègre un nom propre. Dans l'exemple 1, le système de TAN arrive à déterminer que *España de Noche* ne correspond pas à un nom commun, mais la base de données de DeepL ne comporte vraisemblablement pas assez d'informations concernant ce nom. Le processus de traduction débouche sur une proposition hybride, mêlant l'espagnol et l'anglais : le toponyme est exprimé en anglais et le reste du nom propre est emprunté à l'espagnol.[2] Outre le mélange de codes linguistiques, la qualité de la traduction peut être jugée moindre car les marques typographiques indiquant un emprunt sont absentes. L'exemple 2 illustre un cas particulier car le nom propre est en anglais et non pas en espagnol dans le volet source (*Spain Startup-South Summit*). Le système opère une traduction littérale vers le français de chacun des mots.[3] D'un point de vue pragmatique, le résultat n'est pas pertinent car un faux-sens est créé : le *South Summit* correspond au nom d'une plate-forme d'innovation et non pas à un sommet proprement dit.

Le nombre réduit d'occurrences ne permet pas de déterminer si, en termes de stratégie traductionnelle, le système de TAN privilégie les emprunts des noms propres (comme dans une partie de l'exemple 1) ou la traduction littérale par calque sémantique

[2] Ce résultat de traduction est à certains égards similaire à celui d'une deuxième occurrence de *España de Noche* présente dans le corpus. Cette dernière est traduite par *l'Espagne de la Noche*. Même si dans ce dernier cas le toponyme retrouve sa forme francisée, le système emprunte à nouveau une partie à la version en espagnol, quelque peu modifiée.

[3] Dans ce même exemple, il existe un autre cas de traduction littérale d'un nom propre (qui ne concerne pas le toponyme analysé). Le titre du rapport *Mapa del emprendimiento 2020* est exprimé en français dans le volet cible. Il est possible de se demander si le système à réellement catégorisé ce syntagme en tant que nom propre étant donné que les majuscules ne sont pas respectées dans la version traduite.

(comme dans l'exemple 2). Cependant, les données du corpus montrent que les dysfonctionnements concernent le transfert de *España* dans des noms propres peu fréquents. Autrement dit, lorsque *España* apparaît dans le nom d'institutions reconnues, dont la traduction est largement généralisée, le système propose une équivalence usitée, reconnue par le public cible. C'est le cas, par exemple, de la traduction de *Banco de España* par *La Banque d'Espagne*.[4]

D'une manière générale, la traduction des noms propres fait partie des dimensions problématiques pour la TAN (Yvon & Sadaf, 2020). Selon les données correspondantes au nom propre le plus fréquent du corpus analysé, DeepL ne semblerait pas échapper aux dysfonctionnements pour la paire de langues espagnol-français — même s'il serait nécessaire d'examiner un corpus plus large de données. Le résultat obtenu ici signale qu'il s'agit d'un aspect à réviser attentivement lors d'une phase de post-édition.

Les imprécisions et les faux-sens n'ont pas été détectés à l'aide des informations quantitatives recueillies : le nombre d'occurrences de *España/Espagne* étant équilibré entre le volet source et le volet cible, seule la représentation cartographique et le retour au texte, soit la troisième étape de la méthodologie, ont fait émerger les problèmes. Cette troisième phase a également dévoilé des particularités de la traduction du lexème *pandemia* qui n'étaient pas prévisibles à partir des données chiffrées. En effet, si le dictionnaire révèle que la forme *pandemia* présente la même fréquence que celle de *pandémie* (F=40), la représentation visuelle des fragments alignés montre qu'il n'y a pas de correspondance dans deux contextes. La présence de la notion d'antériorité (une période *avant la pandémie*) crée des différences entre le volet source et le volet cible. Dans un cas, le syntagme *anterior a la pandemia* (volet source) est transposé par l'adjectif *prépandémique* (exemple 3). Dans un autre cas, l'expression *avant la pandémie* (volet cible) constitue la traduction de *prepandemia* en espagnol (exemple 4). L'écart entre *pandemia* et *pandémie* en termes de nombre d'occurrences est neutralisé car la transposition morphosyntaxique arrive une fois dans chaque volet.[5]

España tendrá que esperar como pronto a 2023 para recuperar el nivel de PIB anterior a la **pandemia**	L'Espagne devra attendre 2023 au plus tôt pour retrouver son niveau de PIB *pré-pandémique*

A España no le bastará con la vacuna y con dos ejercicios de crecimiento ininterrumpido […] para regresar al nivel de PIB *prepandemia*	Le vaccin et deux années de croissance ininterrompue […] ne suffiront pas à l'Espagne pour retrouver le niveau de PIB d'avant la **pandémie**

Le retour au texte montre qu'il n'existe pas de perte d'information. Le système de TAN arrive à traiter de manière pertinente l'information sur la préfixation et sa transposition possible. En ce sens, la requête spécifique concernant les termes

[4] D'autres exemples du corpus d'étude, qui ne concernent pas exclusivement la toponymie mais les noms propres en général, vont également dans ce sens. Par exemple, le nom de l'organisation internationale *Fondo Monetario Internacional* est traduit dans tout le corpus par son équivalent officiel.

[5] Dans le système, le mot *prepandemia* est catégorisé comme un lexème différent de *pandemia.* Autrement dit, il n'est pas pris en compte dans le calcul du nombre d'occurrences de ce dernier mot.

prepandemia/pré-pandémique (F=3) montre que le contenu informationnel est maintenu entre les deux volets du corpus. Le résultat produit par les transpositions est naturel et adapté au cotexte. En définitive, en ce qui concerne le terme spécialisé *pandemia*, la traduction de DeepL est fidèle à la version originale.

L'analyse menée pour les termes *España* et *pandemia* montre les limites de l'approche purement quantitative. Les fréquences attestées dans le corpus n'encouragent pas à présumer la perte d'information entre le volet source et le volet cible. Seules la représentation cartographique de MkAlign et l'exploration qualitative des contextes autorisent à nuancer la supposition d'une traduction stable. Cette méthodologie a permis d'entrevoir non seulement les dysfonctionnements de DeepL pour la traduction de noms propres composés peu (re)connus mais aussi la pertinence de certaines transpositions réalisées par le système.

4.2 Des unités quasi-univoques à des traductions souvent fidèles

Les critères de constitution du corpus expliquent que les lexèmes *economía*, *crisis* et *COVID-19* fassent également partie des unités les plus fréquentes. Selon les données quantitatives disponibles, les correspondants directs (*économie, crise* et *COVID-19* respectivement) possèdent une fréquence similaire à celle observée dans le volet source. Au vu de ces données, il s'agirait donc d'unités à correspondance quasi-univoque (Zimina 2004, 2005).

Tableau 2.

Fréquences des unités à correspondance quasi-univoque

Volet source	Fréquence	Volet cible (TAN)	Fréquence
Economía	41	économie	**44**
Crise	45	Crisis	**48**
COVID-19	18	COVID-19	**16**

Contrairement aux expectatives créées à partir des fréquences des unités univoques (cf. section précédente), l'écart entre les fréquences des unités quasi-univoques pourrait faire présumer une différence entre le volet source et le volet cible en termes de transfert des informations. L'analyse qualitative a permis de faire la distinction entre les unités où la traduction tend de manière générale vers la fidélité entre les textes source et cible et les cas où le système de TAN rencontre des difficultés menant à des propositions de traduction erronées.

4.2.1 Des traductions fidèles au contenu

En ce qui concerne le lexème *économie*, la différence de trois occurrences par rapport à *economía* s'explique aisément : il s'agit de transpositions allant de l'adjectif (dans le volet source) vers le nom (dans le volet cible). C'est le cas par exemple de la traduction de *futuro laboral y económico* par *l'avenir du travail et de l'économie*. Aussi, le volet cible comporte plus d'occurrences du mot *economía*. L'analyse qualitative de ces

contextes montre que les transpositions sont acceptables et que la transmission du message initial est assurée.

La représentation visuelle construite à l'aide de MkAlign permet d'approfondir dans l'analyse et d'identifier un contexte de non correspondance directe : la traduction de *endeudamiento de las economías domésticas* par *endettement des ménages*. DeepL reconnaît la locution *economía doméstica* et en propose une équivalence pertinente en termes de contenu informationnel. Cette transformation aurait pu créer un écart numérique, dans le sens où le mot *economía* apparaît dans le volet source mais son correspondant est absent du volet cible. Néanmoins, en termes de fréquence, cet écart est neutralisé par l'apparition d'un ajout du mot *économie* dans un autre contexte du volet cible (exemple 5).

Entre sus compañeros en la parte baja del listado hay **economías** tan destruidas como la de Libia o diminutas como Fiji, Maldivas o Aruba.	Parmi ses pairs en bas de la liste figurent des **économies** aussi détruites que celle de la Libye ou des **économies** minuscules comme celles des Fidji, des Maldives ou d'Aruba.

Le système de DeepL opte pour une répétition du mot *économie* afin d'expliciter le référent qui, dans la version originale, restait implicite. Si la transmission des informations est assurée, lors d'une post-édition approfondie, il serait légitime de questionner la qualité de la proposition de traduction par rapport au style et à la fluidité. Dans la prochaine section, la question des répétitions créées par le système de TAN est également abordée.

Seule l'exploration manuelle de tous les contextes a permis de dévoiler l'ajout dont il a été question car la représentation cartographique de MkAlign ne signale pas les segments avec des occurrences multiples du mot recherché. Afin de confirmer l'existence d'autres particularités non visibles à partir des tableaux de fréquence, une analyse étendue par champ lexical a été menée. Le corpus a donc été exploré en sélectionnant à la fois tous les lexèmes reliés à *economía/économie* (*económico/a*, *economista* ; *économique*, *économiste*).[6] Toutes les occurrences obtenues (F=102 pour le volet source, F=105 pour le volet cible) ont été passées en revue manuellement. Les résultats de cette analyse confirment que dans l'ensemble les informations sont correctement transférées ; le cas échéant, les transpositions s'adaptent au cotexte.[7] L'écart en termes de fréquence concernent les occurrences déjà citées, en plus de la traduction correcte de *ahorro* par *économies.*[8]

[6] Il est possible de créer des *Generalized types* ou *Tgen* en sélectionnant toute une série d'occurrences à partir d'une recherche d'expression régulière, telle que « adminst+ » (qui regroupe donc administration, administrateur, administrer…) (cf. Zimina et Fleury, 2007).

[7] Une seule occurrence de « *économie »* dans le volet cible correspond à une modulation d'un segment du volet source : l'expression *tejido empresarial* est traduite par *tissu économique*. Cette modulation n'implique pas de perte d'information grâce à d'autres éléments présents dans le cotexte. Cependant, le système crée une répétition innécessaire avec une occurrence déjà existante dans la version originale du mot *económico.*

[8] Il s'agit de deux occurrences dans tout le corpus. Elles n'intègrent pas le tableau 2, mais il s'agit bien d'une illustration de la polysémie du mot *économies*, ce qui justifie également que ce lexème puisse créer par ailleurs d'autres réseaux de correspondances. Force est de signaler que le retour au texte a permis

Quant au mot *economía,* il n'existe donc pas d'interprétation erronée de la part du système, ni de reformulation non adaptée. Pour ce lexème, l'écart entre la fréquence du volet source et celle du volet cible n'est pas un indicateur de traduction erronée ou non fidèle à l'original. Au contraire, l'analyse de tous les mots du même champ lexical indique que, malgré certaines maladresses au niveau du style, la traduction du contenu informationnel est réalisée correctement.

L'analyse du mot *crisis*, quant à elle, débouche sur des conclusions quelque peu similaires. Si la fréquence de *crisis* n'est pas équivalente à celle de *crise*, il n'existe pas de perte d'informations au cours du processus de traduction réalisé par le système de TAN. Le retour au texte a cependant fait ressortir trois cas de non correspondance directe.

Premièrement, le système DeepL traite correctement le préfixe dans *precrisis* pour en proposer la paraphrase *avant la crise*. Cette transposition, analogue à celles présentées plus haut pour *prepandemia*, s'insère de manière satisfaisante dans le nouveau cotexte, aussi bien du point de vue strictement grammatical que stylistique.

Deuxièmement, un cas d'ajout a été détecté à partir de l'exploration manuelle des contextes (exemple 6). Le système reformule une phrase et explicite le référent d'un pronom du volet source renvoyant au mot *crisis*. Le résultat est grammaticalement recevable, mais comporte une répétition qui, absente dans la version originale, a une incidence sur la qualité de la traduction, à l'image de l'exemple 5 présenté plus haut.

Una **crisis** tras otra sin apenas poder respirar hace que la que se está sufriendo actualmente adquiera una gravedad sin precedentes desde la Segunda Guerra Mundial.	Une **crise** après l'autre, sans grand répit, rend la **crise** actuelle plus grave que toutes celles survenues depuis la Seconde Guerre mondiale.

Enfin, l'exemple 7 illustre un cas de non correspondance directe. A partir du syntagme *en tiempos de covid*, le système de TAN propose la traduction *en période de crise* en s'appuyant sur la relation métonymique liant *crise* à *covid.* Cette dernière relation assure la transmission du message original. En lisant tout l'article traduit, on arrive à comprendre que la crise en question est celle qui se rattache à la COVID-19. Néanmoins, il convient de souligner que la proposition de traduction de DeepL passe complètement sous le silence la référence culturelle à l'œuvre de Gabriel García Márquez (*El amor en los tiempos del cólera*) évoquée par la version originale. Certaines informations extratextuelles du message sont donc perdues au cours du processus de traduction automatique.

Crear una empresa en tiempos de covid	Créer une entreprise en période de **crise**

Afin d'évaluer la modulation proposée par DeepL, il faudrait également prendre en compte que la phrase de l'exemple 7 constitue le titre de l'un des articles journalistiques

d'écarter les occurrences s'éloignant du champ sémantique visé pour les différentes unités analysées. A titre d'exemple, le lexème *firma* dans le sens de *signature* (F=1) n'a pas été retenu lorsque le but était d'analyser le mot *empresa*.

du corpus. Cela pourrait rendre acceptable une traduction libre ou peu littérale comme celle dont il est question. Reste néanmoins à déterminer si le système de TAN est en mesure de réaliser une distinction entre le texte et ses paratextes et en proposer des procédés de traduction différenciés. Une telle investigation nécessite un corpus de données spécifiques et excède le cadre de cet article.[9]

Les trois occurrences de *crise* décrites dans cette section justifient que le volet cible comporte une fréquence plus élevée de ce lexème par rapport à son correspondant direct dans le volet source. L'analyse montre que, dans l'ensemble, le volet cible est fidèle aux informations du volet source, malgré certaines inélégances stylistiques et la perte de références intertextuelles (dans un contexte particulier tel que le titre d'un article de presse). A l'instar de ce qui arrive avec le terme *economía*, dans le cas d'envisager un processus de post-édition, ce dernier devra donner priorité aux aspects discursifs par-dessus le plan strictement grammatical et du contenu informationnel.

4.2.2 Des traductions problématiques

L'acronyme *COVID-19* peut également être considéré comme une unité à correspondance quasi-univoque selon les données du tableau de fréquences présenté plus haut (cf. Tableau 2). En outre, étant donné que la fréquence est moindre dans le volet cible, l'hypothèse d'une perte d'informations durant le processus de traduction semblerait justifiée.

La représentation cartographique du corpus permet de localiser les deux contextes de non correspondance du bi-texte. Le premier a déjà été présenté dans la sous-section précédente. La traduction de *covid* par *crisis* se rapporte à un procédé qui, dans l'ensemble, permet de transférer de manière fidèle le message original. Le deuxième cas de non correspondance concerne une erreur grave de traduction (exemple 8).

¿la vacuna evita la **covid** grave?	le vaccin prévient-il la **covariectomie** sévère ?

La base de données de DeepL ne semblerait pas encore suffisamment volumineuse pour que le système puisse reconnaître les occurrences de *COVID-19* dans tous les contextes. L'exemple 8 illustre un problème de détection : le système de TAN ne parvient pas à repérer la forme courte de *COVID-19* et en propose une traduction erronée à partir d'un mot inexistant. Le non-sens de cette phrase empêche d'affirmer qu'il n'existe pas de perte d'information ; au contraire, le résultat ne fait pas sens.

Dans ce dernier cas, la post-édition du texte cible, qu'elle soit approfondie ou légère, est nécessaire. Il en va de même pour d'autres erreurs repérées à partir de l'examen manuel des traductions. Ces fautes ne concernent pas spécifiquement le lexème *COVID-19* mais la construction syntaxique choisie par DeepL. Il s'agit d'omissions de mots grammaticaux, tels que des articles et des prépositions. Le système de TAN traduit, par exemple, *La COVID-19 ha irrumpido con su trágica incidencia* par **COVID-19 a fait une percée avec son impact tragique* ; ou encore *La respuesta a la pandemia de la*

[9] Il existe dans le cotexte immédiat de la phrase de l'exemple 7 (premier paragraphe de l'article) une occurrence de *en tiempos de pandemia* traduite de manière littérale par *en période de pandémie*, même si le noyau du syntagme prépositionnel est le même que dans le titre illustré dans l'exemple 7. Ces données semblent appuyer l'hypothèse d'une traduction plus libre des paratextes.

COVID-19 par **La réponse à la pandémie COVID-19*. Les imprécisions syntaxiques des traductions précédentes n'empêchent pas complètement la transmission du message original, contrairement à ce qui arrive dans l'exemple 9 ci-dessous où l'omission de la préposition et de l'article rend la compréhension plus difficile.

El ejemplo más reciente es Pfizer y su anuncio de que su vacuna contra la **covid** tiene una eficacia "superior al 90%".	L'exemple le plus récent est celui de Pfizer et son annonce que son vaccin **covid** est "efficace à plus de 90%".

Les problèmes syntaxiques ont sans doute trait au fait que *COVID-19* correspond à un acronyme de création récente. Le traitement de néologismes est en effet l'une des limites déjà identifiées des systèmes de traduction neuronale - *cf.* Tinsley (2017) et Hernández-Morin (2019). En termes de propriétés grammaticales, cette unité est encore mal catégorisée par le système de TAN, suscitant les dysfonctionnements signalés.

La nouveauté de l'acronyme suscite d'autres problèmes mineurs issus des propriétés d'un terme d'apparition récente pas encore complètement stabilisé. Plus précisément, le lexème *COVID-19* présente des variantes orthographiques dans les textes du volet source : *covid-19*, *Covid-19* et la forme courte *covid*. Cette même fluctuation se retrouve dans le volet cible, sans qu'il y ait pour autant une relation directe entre les variantes. Par ailleurs, la variation concerne également le genre de l'acronyme. Malgré les discours institutionnels préconisant l'utilisation du féminin, le flottement entre le masculin et le féminin reste une caractéristique du terme *COVID-19*, aussi bien en espagnol qu'en français. Dans le corpus analysé, la variation est réduite dans le volet source : seule une occurrence de cet acronyme a été employée au masculin en espagnol. En revanche, dans le volet cible, les occurrences se répartissent de manière équivalente entre les deux genres. Une variation plus marquée a donc été introduite par le système de TAN. À cet égard, une procédure de post-édition visant l'application d'une norme d'usage déterminée, aussi bien pour l'orthographe que pour le genre de l'acronyme, serait souhaitable afin d'homogénéiser le résultat en fonction de critères de traduction délibérés.

Certes, il existe des occurrences de *COVID-19* correctement traitées par DeepL ; il est indéniable que ce système s'adapte avec une grande vitesse à l'évolution des langues. Cependant, contrairement aux autres unités analysées, l'acronyme en question pose de véritables problèmes requérant nécessairement l'intervention humaine afin d'aboutir à une traduction convenable. Au-delà de l'homogénéisation conseillée quant à la typographie et le genre grammatical, les erreurs d'ordre strictement grammatical et les traductions erronées portent préjudice à la qualité de la traduction.

Les unités analysées dans cette section (*crisis*, *economía* et *COVID-19*) manifestent un écart en termes de fréquence entre le volet source et le volet cible. Le retour aux textes montre que cette différence n'est pas due à l'existence de plusieurs lexèmes de l'un des volets permettant de traduire une seule unité de l'autre partie du bi-texte. Autrement dit, il ne s'agit pas d'un phénomène de quasi-univocité à proprement parler. Il existe des raisons distinctes à l'origine de la différence des fréquences. Pour ce qui touche aux lexèmes *economía* et *crisis,* l'écart observé surgit de transpositions morphosyntaxiques et d'explicitations des référents qui font accroître le nombre d'occurrences dans le volet cible. Dans ces cas, la perte d'information entre les parties du bi-texte est presque

inexistante. En revanche, la traduction de l'acronyme *COVID-19* est parfois problématique, en ce sens qu'elle témoigne d'erreurs morphosyntaxiques ou d'une perte d'information fondée sur des non-sens. Pour cette dernière unité le nombre d'occurrences est moins important dans le volet cible. Au vu des résultats obtenus dans cette section, il s'avère dès lors que l'identification des lexèmes avec une fréquence moindre dans le volet cible peut constituer une véritable stratégie de détection d'erreurs lors de la préparation d'un processus de post-édition.

4.3 De la correspondance multiple à l'uniformisation

Parmi les unités les plus fréquentes du corpus, se trouvent également les lexèmes *empresa* et *año*, ce dernier ayant la fréquence la plus élevée de tout le volet source. Le mot *empresa* se rattache directement à l'un des thèmes généraux du corpus, à savoir l'économie. En ce qui concerne *año*, à première vue, ce mot ne semblerait pas se relier nécessairement à la même thématique. Cependant, cette fréquence d'apparition s'explique par le fait que la question de la temporalité est centrale dans les analyses économiques : comparaisons entre les périodes, prévisions pour un laps déterminé, cadrage temporel des affirmations. Le tableau 3 ci-dessous présente les fréquences attestées pour chacun des lexèmes en question, ainsi que pour leurs correspondants directs dans le volet cible.

Tableau 3.

Fréquences des unités à correspondances multiples

Volet source	Fréquence	Volet cible (TA)	Fréquence
Año	69	année / an	**80**
Empresa	39	entreprise	**55**

La différence quantitative entre les deux parties du bi-texte n'est pas négligeable : il y a lieu de considérer qu'il s'agit d'un réseau de correspondances traductionnelles multiples (Zimina 2004, 2005). En d'autres termes, chaque unité de l'un des volets possède plusieurs correspondants dans le volet opposé. Des processus de résonance textuelle successifs deviennent donc nécessaires pour mettre en lumière le réseau de correspondances.

L'analyse du lexème *año* permet de révéler les choix de traduction réalisés par le système de TAN. Le tableau 4 résume les correspondances trouvées dans le corpus à partir du terme d'induction *año* et, ensuite, à partir des mots *année* et *an*.

Tableau 4.

Fréquences des correspondances multiples à partir de año – année

Volet source	Fréquence	Volet cible (TAN)	Fréquence
año	69	année	**60**
década	3	an	**20**
ejercicio	2		
interanual	1		
TOTAL	75		**80**

Le mot *año* a été traduit en français par DeepL soit par *année* soit par *an.* Le choix entre ces deux dernières variantes répond à des contraintes lexico-sémantiques et d'usage que le système de TAN a correctement identifiées. Aucune erreur de sélection n'a été identifiée à partir de l'examen manuel de tous les contextes d'apparition des mots en question. De plus, la perte d'informations essentielle est inexistante. A ce sujet, la différence de fréquences totales entre les deux volets est due à des ajouts réalisés dans les textes traduits afin d'expliciter des référents qui, dans la version originale, apparaissent pronominalisés ou restent implicites, comme dans l'exemple 10.

Las anteriores, publicadas en septiembre, anticipaban una caída del PIB de la eurozona este **año** del 8% y un repunte el siguiente del 5%.	Les précédents, publiés en septembre, prévoyaient une baisse de 8 % du PIB de la zone euro cette **année** et une hausse de 5 % **l'année** suivante.

Outre le mot *año*, les lexèmes *année* et *an* ont permis de traduire une série d'unités sémantiquement reliées : *década*, *ejercicio* e *interanual.* Ces dernières ont été intégrées dans le tableau 4 afin de présenter un aperçu complet du réseau de correspondances décelé à l'aide de MkAlign. Cependant, il convient de souligner que, pour *década* e *interanual*, la traduction correspond à un syntagme faisant intervenir *an* ou *année*. On trouve ainsi *dix ans* pour *década*, et *baisse de 2% par rapport à l'année précédente* à la place de *retroceso del 2% interanual.* Le retour au texte confirme que la traduction des segments où interviennent ces mots est fidèle à l'original en termes de dénotation.

Pour ce qui est de *década,* il ne semblerait pas exister de raison cotextuelle qui motiverait l'emploi privilégié de la paire *année/an* au détriment du mot *décennie* disponible dans le système de la langue d'arrivée. Si le contenu informationnel est maintenu, le choix réalisé par DeepL fait que le niveau de langue du texte traduit n'est pas équivalent à celui du texte original.

La situation est différente pour le lexème *ejercicio*, lequel, employé avec le sens de 'période d'exécution du budget', correspond à *exercice* en français. Certes, le choix réalisé par DeepL au profit du lexème *année* de la langue générale ferait que l'effet de spécialisation suscité par le texte d'origine ne soit pas recréée dans le texte cible. Néanmoins, l'examen des contextes révèle que la sélection n'est pas injustifiée (exemple 11).

A España no le bastará con la vacuna y con dos **ejercicios** de crecimiento ininterrumpido (5% el **año** que viene, muy por debajo de lo que proyecta el Gobierno con y sin fondos europeos; 4% el siguiente) para regresar al nivel de PIB prepandemia.	Le vaccin et deux **années** de croissance ininterrompue (5% **l'année** prochaine, bien en dessous de ce que le gouvernement prévoit avec et sans fonds européens ; 4% **l'année** suivante) ne suffiront pas à l'Espagne pour retrouver le niveau de PIB d'avant la pandémie.

En espagnol, le mot *ejercicio* peut faire référence à une période généralement d'un an lorsqu'il s'agit de décrire l'activité économique d'une entreprise ou d'une institution, d'après le *Diccionario de la lengua española* de la *Real Academia Española*. Selon le *Trésor de la langue française*, *exercice* peut être employé avec le même sens lorsqu'il concerne la gestion budgétaire d'une entreprise. Or, dans l'exemple 11, il n'est pas question d'une entreprise ou d'une institution, mais d'un pays. Au sens strict, le mot ne s'adapte donc pas au cotexte. Ce manque d'adéquation est corrigé par DeepL en ayant recours, comme il a été dit, au mot *année*. Les données disponibles dans le corpus ne sont pas suffisantes pour déterminer s'il s'agit d'une stratégie programmée dans le système de TAN.[10] Cette traduction mérite néanmoins d'être exposée en ce sens qu'elle illustre les capacités du système non seulement pour transférer le message dans son entièreté mais aussi pour le façonner le plus possible à l'image de la langue d'arrivée.

L'exemple 11 rend compte par ailleurs d'une faiblesse de DeepL. Le changement de *ejercicio* par *année* coïncide avec deux autres occurrences de ce dernier mot, l'une d'elles issue d'un ajout réalisé par le système. La répétition, inexistante dans la version originale, constitue donc une conséquence du processus de traduction. Le résultat témoigne d'une *pauvreté d'expression* (Martínez de Sousa, 2015, p. 134) qu'il serait souhaitable de corriger lors d'un processus de post-édition. Ce type de manque de variété dans l'expression se manifeste également pour la traduction du lexème *entreprise*, bien que les raisons soient quelque peu différentes.

Afin de construire le réseau de correspondances autour de la paire *empresa/ entreprise*, divers processus de résonance ont été réalisés successivement. A chaque fois, les termes induits devenaient à leur tour termes d'induction pour entamer un nouveau processus. Le tableau 5 présente les résultats obtenus.

Tableau 5.

Fréquences des correspondances multiples à partir de empresa – entreprise

Volet source	Fréquence	Volet cible (TAN)	Fréquence
empresa	39	entreprise	**56**
compañía	8	société	**11**
negocio	8	startup	**1**

[10] Si la deuxième occurrence de *ejercicio* traduite par *année* du corpus correspond à une situation semblable à celle de l'exemple 11, le nombre de cas analysés est très réduit.

startup	2	
corporación	1	
comercio	1	
emprendimiento	1	
firma	1	
sociedad	1	
TOTAL	64	**68**

Le tableau 5 suscite deux remarques d'ordre général. La première concerne l'écart entre les fréquences totales ; la deuxième, la moindre variété de lexèmes dans le volet cible.

À l'instar de ce qui a été constaté pour d'autres unités analysées dans cet article, la différence entre les fréquences des volets source et cible provient des ajouts réalisés par le système de TAN. En ce sens, en espagnol, il n'est pas peu fréquent d'employer *las farmacéuticas* à la place de *las empresas farmacéuticas* ; le nom reste implicite et l'adjectif est nominalisé. La traduction française, en revanche, aura tendance à expliciter le nom. Ce type d'explicitation a fait accroître le nombre d'occurrences du mot *entreprise* dans le volet cible. Dans le corpus, outre le mot *farmacéutica*, cette situation concerne les lexèmes *gestora de aeropuertos (entreprise de gestion aéroportuaire*) et *hotelera* (*entreprise hôtelière*).[11]

Il existe également un cas différent d'ajout dans le volet cible. L'exemple 12 illustre les capacités accrues de l'outil de TAN. En effet, élucider le référent du déterminant possessif de *su presidente* dans la deuxième phrase requiert une analyse fine excédant le cadre phrastique strict car le référent *entreprise* se trouve au début de la phrase précédente.

Dos meses después, cuando la **empresa** anunció que la vacuna pasaba a la última fase de ensayo, la subida semanal fue del 50%, batiendo un nuevo máximo histórico. Una circunstancia que su presidente, su consejero delegado y sus directores técnico y médico han aprovechado para deshacerse de buena parte de sus paquetes accionariales en ventas millonarias.	Deux mois plus tard, lorsque la **société** a annoncé que le vaccin entrait dans la phase finale des tests, l'augmentation hebdomadaire était de 50 %, battant un nouveau record. Le président, le PDG et les directeurs techniques et médicaux de la **société** en ont profité pour se défaire d'une grande partie de leur chiffre d'affaires de plusieurs millions de dollars.

L'analyse qualitative de tous les ajouts révèle que le transfert d'informations se déroule correctement.[12] Cependant, il en va sans dire que les explicitations réalisées dans

[11] La situation inverse existe également. Dans le corpus, on trouve un cas d'omission du nom dans la version française : *las grandes empresas fabricantes y distribuidoras Aecoc* est traduit par *les grands fabricants et distributeurs Aecoc*.

[12] Même si l'erreur ne concerne pas le mot *empresa*, notez cependant que la traduction automatique de l'exemple 12 n'est pas exempte de fautes. Le système de TAN réalise un contre-sens de par le mauvais traitement du terme spécialisé *paquete accionarial*. Ce dernier est traduit de manière erronée par *chiffre d'affaires* et non par *paquet d'actions* ou *bloc d'actions.* En outre, selon le cadre général du processus de

le volet cible peuvent donner suite à des répétitions absentes dans la version originale. L'exemple 12 illustre l'un des cas où la présence rapprochée de deux occurrences de *sociedad* nuit à la fluidité du texte.

La deuxième remarque d'ordre général concerne le manque de variété dans le choix des lexèmes. Selon le réseau de correspondances décelé, les neuf unités du volet source sont traduites par trois lexèmes dans la proposition de DeepL. L'analyse numérique stricte montre donc qu'une uniformisation s'opère au cours du processus de traduction. Certes il serait possible d'imaginer une situation où une série de parasynonymes en espagnol ne retrouverait pas en français un ensemble aussi fourni. L'équivalence devrait donc être assurée par d'autres moyens que la simple correspondance directe entre les unités. Cependant, pour la traduction de *empresa* et ses parasynonymes, le système du français dispose également de termes véhiculant des nuances similaires. Certaines occurrences de *entreprise* auraient pu être remplacées par des mots courants tels que *compagnie* ou *firme*. L'absence de ces derniers lexèmes dans le volet cible autorise à remettre en question le réseau de correspondances créé par le système de TAN. Par ailleurs, dans le but de rendre le résultat moins uniforme, le transfert des informations pourrait ne pas se restreindre au champ strict des parasynonymes, mais avoir recours à d'autres procédés. Par exemple, l'unité *hotelera* présentée plus haut aurait pu être traduite par *chaîne hôtelière* — les occurrences du corpus acceptent ce choix. Cependant, DeepL opte pour *entreprise hôtelière*, ce qui participe d'autant plus au manque de diversité dans le vocabulaire employé.

L'uniformisation terminologique réalisée par les outils de TA peut constituer un aspect positif car elle facilite « le travail des traducteurs professionnels spécialisés dans les domaines technico-scientifiques » (Diéguez, 2001, p. 206). Dans ces cas, conserver sans variation le même terme technique dans un texte serait un gage de qualité. Cependant, la pertinence de l'uniformité doit être évaluée en fonction du type de terme uniformisé, du type de texte et de son intention communicative, c'est-à-dire en tenant compte des facteurs intra- et extratextuels qui interviennent dans le processus de traduction. Étant donné que les textes analysés dans cette étude sont considérés comme non techniques, l'uniformisation a été interprétée comme un dysfonctionnement de l'outil de TAN.

Le retour au texte permet de constater que, au-delà de l'uniformisation, certains choix de traduction suscitent un décalage entre le texte source et le texte cible. Il existe des cas où la proposition de DeepL n'adopte pas le même point de vue choisi dans le texte original pour décrire une situation (exemple 13).

El Gobierno aprobó en el último Consejo de Ministros un paquete de medidas [...] para tratar de ayudar a los bares y restaurantes, que son los **negocios** que más han sufrido con las restricciones derivadas de la pandemia de coronavirus.	Lors du dernier Conseil des ministres, le gouvernement a approuvé un ensemble de mesures [...] pour tenter d'aider les bars et les restaurants, qui sont les **entreprises** ayant le plus souffert des restrictions résultant de la pandémie de coronavirus.

traduction (visé pragmatique et contexte socioculturel), il serait également légitime de se poser la question quant au degré de pertinence de l'équivalence entre *consejero delegado* et *PDG* proposée par le système.

Dans le texte source, le lexème *negocio* entre en relation d'hyperonymie avec *bares* et *restaurantes* et fait référence aussi bien au caractère concret des entités (l'établissement) qu'à leur fonctionnement (l'activité commerciale). En revanche, le correspondant choisi par DeepL dirige l'interprétation uniquement vers le domaine général de l'activité économique. Si ce changement n'a pas de conséquence en termes de contenu informationnel, l'effet produit est différent. Ce type de modulation n'est pas exclusif au lexème *negocio*. Il est également à l'œuvre dans la traduction d'autres unités du tableau 5, telles que *emprendimiento* (exemple 14).

María Benjumea, sin embargo, prefiere subrayar algunos aspectos más emocionales del **emprendimiento** en su fase inicial.	María Benjumea, préfère cependant mettre en avant certains aspects plus émotionnels de l'**entreprise** dans sa phase initiale.

Dans l'exemple 14, le mot *emprendimiento* fait allusion au processus de mise en place d'une activité économique. Cette idée ne se reflète pas dans le choix de traduction, lequel dirige l'attention sur le résultat du projet (l'existence d'une entreprise). Afin de trouver une alternative plus fidèle à l'original, il serait possible d'insister sur la *création de l'entreprise* dans la version cible. La focalisation d'une étape particulière de la mise en place d'une activité commerciale semble en effet occasionner des dysfonctionnements au cours du processus de traduction. A cet égard, il convient de souligner une autre correspondance peu appropriée réalisée par le système, à savoir la traduction d'une occurrence de *startup* par *création d'entreprise.* Le calque approximatif réalisé par DeepL ne permet pas de construire le même sens du texte source. Cette correspondance est en plus instable car dans un autre contexte, face à la même unité, DeepL conserve l'anglicisme.

En définitive, les modulations réalisées par DeepL peuvent être source d'une divergence entre la version originale et la version cible. Outre les changements du point de vue, certaines informations du texte source ne sont pas correctement transférées. Dans les cas présentés, la traduction de *negocio* et *emprendimiento* par *entreprise* ne permet pas de construire le même sens dans le texte cible. Par ailleurs, l'uniformisation opéré par le système empêche également de conserver les nuances qui distinguent la série de parasynonymes du tableau 5. Par conséquent, dans le cas de viser une qualité optimale, certaines modifications seraient donc à apporter à la sortie brute de DeepL lors d'une étape de post-édition.

5. Conclusion

De nombreuses études témoignent des progrès de la traduction automatique neuronale (TAN) ou par apprentissage profond. Cependant, les sorties brutes de cette génération de systèmes de traduction automatique ne sont pas infaillibles, notamment lorsqu'un niveau de qualité optimal est visé. Les services commercialisés alertent à ce sujet. DeepL, l'outil de TAN générique (Loock, 2018) employé dans cette étude en est un exemple. Si la qualité de la traduction effectuée par ce système est satisfaisante, elle est jugée imparfaite par les développeurs du système eux-mêmes. Dans sa version commerciale de 2020, il est rappelé, au moment de la signature des conditions générales de souscription, que « doe to its nature, machine translation may be imprecise ». D'une manière générale, la qualité des traductions dépendra des informations dont disposent les systèmes de TAN (Poibeau, 2019). Les données disponibles sont soumises à évolution et ne sont pas équivalentes

pour toutes les combinaisons de langues. Les systèmes peuvent en effet « apprendre » ou être nourris par les nouvelles informations reçues. Par conséquent, il convient de souligner que les conclusions de cette recherche ne sont pas transposables telles quelles à d'autres combinaisons linguistiques.

Cet article s'est proposé d'évaluer un corpus parallèle espagnol-français issu de la traduction automatique. L'attention s'est portée plus précisément sur le transfert d'informations au cours du processus de traduction. Il a été question d'interroger la fidélité de la traduction en termes de conservation des informations dénotatives. Pour ce faire, certaines fonctionnalités du logiciel MkAlign ont permis de définir une méthodologie tripartite. D'abord, sept unités du volet source et leurs unités correspondantes en volet cible ont été sélectionnées à partir d'une approche quantitative. Ensuite, une étape intermédiaire, fondée sur la représentation visuelle du corpus, a permis d'examiner de quelle manière les correspondances s'établissent entre les deux volets du bi-texte. La troisième étape, quant à elle, a eu pour vocation d'analyser les équivalences en contexte afin de déterminer si le contenu informationnel était correctement transféré.
Cette exploration du corpus a mis en lumière des aspects négatifs et positifs de la TAN. Le système a montré des déficiences à différents niveaux de gravité. Pour ce qui est des erreurs mineures, il convient de signaler les effets négatifs de certaines modulations réalisées par DeepL. Certains choix lexicaux ne focalisent pas exactement la même étape dans un processus de création d'activité économique (p. ex. la traduction de *emprendimiento* par *entreprise*). Il existe également une perte des nuances de certains parasynonymes, traduits par un seul et même lexème, entraînant l'uniformisation du texte cible (p. ex avec l'unité *entreprise*). Si l'uniformisation terminologique est souhaitée dans la traduction des textes spécialisés, elle est moins pertinente dans les textes journalistiques destinés au grand public à cause de l'effet produit, à savoir un écrit stylistiquement répétitif. D'autres répétitions dans des contextes rapprochés sont issues des ajouts réalisés par le système de TAN pour expliciter certains référents implicites dans le texte original. Selon les données disponibles, cette stratégie constitue un défaut de DeepL. Enfin, certaines unités pas encore complètement stabilisées du point de vue de l'usage, tel que le néologisme *COVID-19*, ont fait l'objet de différents choix de traduction au cours du processus. Par conséquent, le texte cible manque d'homogénéisation et devrait être soumis à post-édition.

L'analyse a également mis en relief des erreurs graves de non-sens. Elles concernent l'acronyme d'apparition récente *COVID-19* et certains noms propres peu connus (tel que *España de Noche*). Ne disposant vraisemblablement pas de données suffisantes, le système propose des unités inexistantes en langue cible. Il s'agit, de ce fait, des seules occasions illustrant la perte grave d'informations entre les deux volets du bi-texte. Au vu de ces données, dans le cas d'un niveau d'exigence irréprochable en termes de qualité, la TAN n'a pas réussi à égaler la production humaine.

Pour ce qui est des bons résultats, un point positif à mentionner est d'ordre lexical et syntaxique. Il porte sur la résolution de certaines expressions anaphoriques. Le système arrive à localiser des référents qui, dans la version source, se trouvent au-delà des limites phrastiques — ces explicitations créent cependant des répétitions dans certains cas. Un autre point à souligner concerne l'objectif de cette recherche. Selon les critères définis, il s'est avéré que la traduction automatique réalisée par le biais de DeepL

ne manifeste pas de perte systématique d'informations dénotatives concernant les lexèmes étudiés. Autrement dit, le texte cible est dans l'ensemble fidèle à l'original.

La démarche d'analyse outillée adoptée a permis d'organiser l'exploration du corpus parallèle de façon méthodique afin d'objectiver l'évaluation du résultat de traduction. Elle comporte cependant des limites qui devraient être prises en compte dans de futurs travaux. Une première remarque porte sur la représentation cartographique du bi-texte. Seule la vérification manuelle permet de rendre visibles les cas où il existe plusieurs occurrences des lexèmes étudiés au sein d'une même phrase. Dans le cadre d'une étude avec un corpus de travail plus large, cette vérification serait coûteuse en temps et le risque d'une analyse peu rigoureuse serait accru. Il serait en conséquence convenable de traiter le corpus en amont ou bien d'exploiter d'autres fonctionnalités disponibles dans MkAlign, telles que le calcul des spécificités ou la recherche de co-occurrents.

Par ailleurs, l'échantillon analysé est constitué à partir des unités les plus fréquentes du dictionnaire du volet source. Étant donné que le dictionnaire se fonde sur un découpage strict par mot (défini comme un ensemble de caractères entre deux blancs), la méthodologie adoptée laisse dans l'ombre certaines unités dont une analyse contextualisée pourrait s'avérer intéressante. Par exemple, le retour au texte a mis en lumière des formes composés (*estado de alarma*), des expressions idiomatiques (*balde de agua fría*) ainsi que des métaphores (*cóctel de tipos de interés*). Dans le but d'étudier la traduction de ce genre d'unités phraséologiques dont la restitution en langue cible par les systèmes de TAN présente des limites (Poibeau, 2019 : 148), il serait également nécessaire d'avoir recours à une annotation lexico-syntaxique et sémantique du corpus.

Les résultats de l'exploration du corpus ainsi que la méthodologie adoptée permettent d'envisager également des perspectives didactiques. Compte tenu de la demande croissante de travaux de post-édition ainsi que des recommandations en matière de formation aux outils de la traduction, de quelle manière la traductologie de corpus peut-elle inspirer de nouvelles approches didactiques de la post-édition ? À la lumière des limites et des avantages du système de TAN révélés ici, la construction d'une méthode pédagogique alimentée par l'exploration de corpus parallèles serait envisageable. Dans ce cadre général, certains outils d'exploration de corpus pourraient donc être intégrés dans la formation des apprentis traducteurs. Par exemple, la méthodologie de saisie par nombre d'occurrences pourrait être utile dans un processus de post-édition. Les cas où la fréquence est plus faible dans le volet cible feraient l'objet d'une révision prioritaire. La personne chargée de la post-édition pourrait alors consacrer moins de ressources attentionnelles au reste des unités, lesquelles selon l'analyse n'ont pas suscité d'erreurs graves.

De même, la méthodologie mise en œuvre dans cette étude pourrait trouver sa place dans un cours de traduction à visée professionnelle dans le but de travailler les compétences linguistiques et de traduction. D'une part, la fonctionnalité du dictionnaire et de la classification des unités par fréquences invite les étudiants à se questionner sur la notion de mot clés (sont-ils les plus fréquents du corpus ?) et *in fine* à travailler sur les domaines de spécialisation (les mots clés définissent-ils le domaine ? Les mots les plus fréquents sont-ils rattachés au domaine de spécialisation ?). D'autre part, l'exploration en trois étapes pourrait être réinvestie lors d'un travail sur la notion d'équivalence et sur les

procédés de traduction. En d'autres termes, il serait question de faire prendre conscience de la différence entre la signification en langue et le sens textuel, c'est-à-dire entre la correspondance et l'équivalence.

Références

[1] Barbin, F. (2020). La traduction automatique neuronale, un nouveau tournant ? *Palimpseste* (*4)*, 51-53.

[2] Dancette, J. (1989), La faute de sens en traduction, *TTR : Traduction, Terminologie, Rédaction 2(2),* 83-102. https://doi.org/10.7202/037048ar

[3] Franco Aixelá, J. (2001-2020). *BITRA (Bibliografía de Interpretación y Traducción)*. Base de datos en acceso abierto. https://dti.ua.es/es/bitra/ introduccion .html

[4] Groupe EMT (2017). *Référentiel de compétences de l'EMT*. Bruxelles.

[5] Guidère, M. (2010). *Introduction à la traductologie* (2e éd.). De Boeck.

[6] Hernández-Morin, K. (2019). Évolutions des technologies et des usages en traduction : Pratique et enseignement de la post-édition. In É. Lavault-Olléon & M. Zimina (Éds.), *Des mots aux actes. Traduction et technologie, regards croisés sur de nouvelles pratiques* (pp.239-255). Classiques Garnier.

[7] Hurtado Albir, A. (2001). *Traducción y traductología. Introducción a la traductología* (3e éd.). Cátedra.

[8] Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation 17(1-2),* 5-38.

[9] Keromnes, Y. (2016). La comparaison de traductions et de « textes parallèles » comme méthode heuristique en traductologie. In J. Albrecht & R. Métrich (Éd.), *Manuel de traductologie* (pp. 99-117). De Gruyter.

[10] Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28-39.

[11] Loock, R. (2016). *Traductologie de corpus*. Presses universitaires du Septentrion.

[12] Loock, R. (2018). Traduction automatique et usage linguistique : Une analyse de traductions anglais-français réunies en corpus. *Meta (Canada) 63(3),* 786-806.

[13] Loock, R. (2019). La plus-value de la biotraduction face à la machine. *Traduire (241),* 54-65.

[14] Mahadi, T. S. T., Vaezian, H., & Akbari, M. (2010). *Corpora in Translation. A practical guide*. Peter Lang.

[15] Martínez, L. (2019). La technologie et la traduction spécialisée. In *Des mots aux actes. Traduction et technologie, regards croisés sur de nouvelles pratiques* (pp. 309-326). Classiques Garnier.

[16] Martínez de Sousa, J. (2015). *Manual de estilo de la lengua Espanola* (5e éd.). Ediciones Trea.

[17] Mialet, E. B. (2010). The sociology of translation: Outline of an emerging field. *MonTi: Monografías de Traducción e Interpretación (2).*

[18] Moorkens, J., & Way, A. (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation*. https://doi.org/10.1007/ s10590019-09232-x

[19] Pincemin, B. (2020). Sémantique interprétative et textométrie – Version abrégée. *Corpus* (10), 259-269.

[20] Ping, K. (2009). Machine translation. In M. Baker & G. Saldanha (Éd.), *Routledge Encyclopedia of Translations Studies* (pp. 162-169). Routledge.

[21] Robert, A.-M. (2010). La post-édition : L'avenir incontournable du traducteur ? *Traduire* (222), 137-144.

[22] Salem, A. (2004). Introduction à la résonance textuelle. *Journées internationales d'Analyse statistique des Données Textuelles* (JADT 2014). Laboratoires ERTIM & CLESTHIA-SYLED, Jun 2014, Paris, France, 986-992.

[23] Tinsley, J. (2017). Neural MT and the legal field. *Multilingual*, 28-34.

[24] Toral, A., & Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 1*(1), 1063-1073.

[25] Yvon, F. (2019). Les deux voies de la traduction automatique. *Hermes, La Revue*, *3*(85), 62-68.

[26] Yvon, F., & Sadaf, A. R. (2020). Utilisation de ressources lexicales et terminologiques en traduction neuronale. *[Rapport de recherche] 2020-001, LIMSI-CNRS*.

[27] Zimina, M. (2004). Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve (Belgique), 10-12 mars 2004, 1195-1202.

[28] ------ (2005). Exploration textométrique de corpus de traduction. *Meta*, *50*(4), 1-11.

[29] Zimina, M., & Fleury, S. (2007). Exploring translation corpora with MkAlign. *Translation Journal 11(1)*. https://translationjournal.net/journal/39mk.html

[30] ------ (2014). Approche systémique de la résonance textuelle multilingue. *Journées internationales d'Analyse statistique des Données Textuelles* (JADT 2014). Laboratoires ERTIM & CLESTHIA-SYLED, Jun 2014, Paris : France., 717-728.