ÉCOLE DOCTORALE INTERDISCIPLINAIRE
Faculté de Lettres

Andreea GHIȚĂ

La qualité du produit sémantique de l'intelligence artificielle dans la traduction des textes de vulgarisation scientifique

Résumé

Directeurs de Doctorat

Prof. univ. HDR Alexandru MATEI - Université Transilvania de Brașov
Prof. univ. HDR en cotutelle Sonia BERBINSKI – Université de Bucarest

BRAȘOV, 2025

Table des matières

1. Introduction

La traduction automatique, mécanisée ou mécanique qui existe depuis 1950 a toujours essayé de « comprendre » la langue écrite sous tous ses aspects au moyen des outils informatiques. Aujourd'hui, nous assistons à la naissance des logiciels de traitement de langue extrêmement puissants et fiables en matière de qualité. Ceux-ci vont au-delà de la simple incorporation des règles de la langue ou d'autres formalismes avec lesquels on peut opérer des calculs sur les unités lexicales, les syntagmes ou encore, les phrases.

La nouveauté dans ce domaine-ci est représentée par la traduction automatique neuronale faite par une MTN qui, juste comme le nom l'indique, repose sur le principe des réseaux neuronaux du cerveau. Elle utilise d'innombrables processeurs qui créent des liens sur le modèle des neurones humains. En mettant en œuvre un processus d'apprentissage qu'elle peut réaliser toute seule à l'aide de l'intelligence artificielle, la machine génère la traduction.

Cette technologie ne se contente plus de découper les phrases pour aboutir au décryptage de la langue source, d'après le modèle de ses prédécesseurs. Elle prend en considération la phrase dans son intégralité, pouvant ainsi apporter une meilleure traduction de la signification. Ce qui est plus important encore est que ce logiciel de traduction se renforce tout seul, bâtissant des voies neuronales qui apprennent comme les êtres humains. De cette façon, il devient de plus en plus puissant et performant au fur et à mesure que le temps passe et qu'on lui fournisse des données.

Il est donc impératif d'établir s'il existe des indices qui laissent présager d'ores et déjà un scénario où un jour, ce genre de logiciels n'auront plus besoin de la surveillance et de l'intervention humaine. La manière dans laquelle on peut réfuter ou attester la valeur de vérité de cette hypothèse est à travers l'analyse linguistique du degré de qualité sémantique du texte fourni en langue cible. C'est ce qui nous a amenée à mettre sous la loupe les fautes sémantiques commises par Google Translate et sa machine neuronale. On trouve qu'il est essentiel de mener ce travail si l'on prend en compte que cette machine de traduction neuronale a vu le jour fin 2016, ce qui signifie qu'elle se trouve à un stade précoce ; elle avait l'âge de 3 ans quand nous l'avons alimentée avec notre corpus.

Par la suite, notre travail s'insère parmi les premières recherches linguistiques sur le sujet de la qualité de la traduction automatique neuronale. En outre, l'approche que nous allons emprunter pour répondre à nos objectifs de recherche sera d'inspiration cognitive. De surcroît, il faut noter que la tendance dominante est que toute méthodologie cognitive, quel que soit le domaine, réponde aux besoins de l'IA.

Nous cherchions donc à exploiter ce type de traduction automatique munie d'intelligence artificielle à l'aide des instruments qui appliquent des principes du même ordre. Sous cet angle, notre thèse a le potentiel de devenir un travail de référence aussi bien pour le sujet abordé que pour la méthode d'analyse qui y est associée.

2. Contexte actuel de la recherche

Tout d'abord, nous tenons à préciser avec fierté que notre thèse s'inscrit dans un contexte actuel de recherche, et ce, grâce à deux axes directeurs. Le premier est relatif à la traduction obtenue par le biais de la GTNM et le deuxième a trait aux outils linguistiques que nous allons employer à nos fins de recherche.

Le premier argument quant à l'actualité de notre recherche consiste donc dans le fait que la technologie de traduction automatique de Google que nous voulons évaluer est pourvue d'intelligence artificielle. En fait, cette désignation d'« intelligence artificielle » semble avoir envahi notre quotidien. Toute nouvelle technologie cherche à opérer sur des mécanismes intelligents.

Le deuxième aspect qui situe notre recherche dans l'actualité est l'emploi des théories et principes d'analyse en provenance de la linguistique cognitive, particulièrement de la sémantique cognitive. Il est communément admis que la LCog. occupe depuis des années le devant de la scène en matière d'analyse du langage, au détriment de la linguistique structuraliste.

Ces deux domaines principaux qui se chevauchent dans notre recherche à savoir la traduction automatique neuronale et la SC partagent l'emphase mise sur a) le sens lexical et b) le contexte des lexies. Cela nous fait penser à l'approche contextualiste du domaine de la linguistique du corpus qui postule que :

> « Chaque occurrence d'une unité lexicale véhicule sa propre histoire textuelle, un environnement collocationnel particulier qui s'est construit lors de la création du texte et qui fournit le contexte dans lequel l'unité s'incarne pour cette occasion précise. Cet environnement détermine le sens instancié, ou le sens textuel de l'unité, sens qui est unique pour chaque utilisation spécifiée. » (Halliday & Hasan, 1976, p. 289 dans Williams, 2003, p. 35).

Dubreil (2008, p. 13) rajoute que « le mot dérive son sens du contexte et influe simultanément sur ce contexte pour créer l'environnement textuel ». C'est pourquoi le sens des unités lexicales simples et polylexicales que nous allons analyser sera examiné en relation étroite avec leur contexte.

Étant donné le fait que la machine neuronale de traduction se propose d'extraire le sens des mots et des phrases pour améliorer la compréhension globale du texte, nous sommes également dans la lignée de l'hypothèse de Jakobson (1959) qui affirmait que c'est la signification qu'on traduit. Cette théorie s'est vue renforcée par d'autres linguistes dès l'amorçage de la TA. Cela nous porte à croire qu'en définitive, le côté sémantique et conceptuel du texte devrait être considéré l'objectif – clé de la traduction automatique. Et comment mieux l'analyser qu'à travers des théories conceptuelles sémantiques issues de la linguistique cognitive ?

3. Hypothèse, questions de recherche et plan de la thèse

Toujours assiégés par des nouvelles couvrant les prouesses incontestables de l'intelligence artificielle dans de nombreux domaines d'activité, on ne peut pas s'empêcher de formuler une question assez angoissante : Serait-il possible de formaliser tout ce qui est l'empreinte humaine, en l'occurrence le comportement humain et langagier ?

En ce qui concerne le domaine de recherche auquel nous nous intéressons, notamment la TA neuronale, on craint que le traducteur ainsi que l'interprète soient remplacés définitivement par celle-ci. Les développeurs de la MTN et les linguistes-informaticiens prédisent que dans les années à venir on ne devrait plus employer des experts linguistiques pour livrer des traductions spécialisées de qualité. On estime davantage qu'un jour, il sera possible que les traducteurs professionnels n'interviennent plus dans les traductions écrites afin de les affiner, les retravailler. Bien qu'on puisse imaginer de tels scénarii, le niveau de complexité de la langue parlée, du langage littéraire et technique reste haut et c'est exactement ce défi que l'IA devra relever, à moins qu'elle ne l'ait déjà fait…

En conjuguant l'hypothèse ci-dessus avec le fait qu'on veut nous faire croire que la TAN est investie avec des pouvoirs sémantiques gigantesques, nous avons conçu une question de recherche pivot : Quel est le niveau de qualité du produit sémantique de l'intelligence artificielle dans la traduction des textes de vulgarisation scientifique ?

Selon cette question clé légitime, que le titre de notre thèse sous-tend, il est facile à comprendre que notre objectif principal est de découvrir à quel point la traduction neuronale est qualitative sur le plan sémantique dans le cas d'un corpus de spécialité. Afin d'atteindre notre objectif, nous allons examiner en profondeur le transfert sémantique du français en roumain des lexies simples et composées et des syntagmes phraséologiques.

Notre question de recherche principale cherche au premier abord un résultat d'ordre quantitatif, que nous allons certainement apporter. En fait, la première étape de notre recherche consistera dans l'inventaire et le groupement en classes de toutes les erreurs de traduction dans le fichier Excel « Pré-analyse Corpus aligné.xls ». Cela nous permettra d'en extraire des statistiques faisant montrer un pourcentage de fiabilité du Transformer de Google Translate.

Dans une deuxième étape, indéniablement plus importante encore, nous allons essayer de comprendre les aspects qui ont posé un challenge aux neurones artificiels au niveau de la traduction :

1. des termes et collocations de spécialité de notre corpus ;
2. des lexies simples et composées de notre corpus relevant du langage général et qui peuvent faire également partie des constructions élargies et plus ou moins fixes ;
3. des idiomes encodants et décodants de notre corpus dont la construction dépend la plupart du temps du phénomène de la métaphorisation et des tropes conceptuels.

Par ailleurs, toute la communauté linguistique reconnaît que la discrimination du sens ainsi que la compréhension/production d'une lexie nouvelle, des termes, des collocations et des idiomes font appel aux habiletés cognitives humaines qui sont chargées de la conceptualisation. Les linguistes cognitivistes vont plus loin que ce constat et postulent que cette dernière est influencée par des facteurs contextuels divers et la physiologie humaine.

Par voie de conséquence, afin de découvrir à quel point les « habiletés cognitives » de la machine sont proches de celles des êtres humains pendant l'affectation de la traduction des éléments linguistiques de 1 à 3 ci-dessus, nous avons conçu les questions ci-après :

1. De quelle manière et dans quelle proportion la polysémie influence-t-elle la production des erreurs aux niveaux LexS, LexC, Co et Loc ?
2. De quelle manière et dans quelle proportion le contexte influence-t-il la production des erreurs aux niveaux LexS, LexC, Co et Loc ?
3. De quelle manière et dans quelle proportion les tropes conceptuels influencent-ils la production des erreurs aux niveaux LexS, LexC, Co et Loc ?
4. De quelle manière et dans quelle proportion l'anglais, langue d'entraînement du réseau influence-t-elle la production des erreurs aux niveaux LexS, LexC, Co et Loc ?

Tout comme l'hypothèse initiale a servi de fil conducteur pour la formulation des questions de recherche, ces dernières ont inspiré le plan de la thèse qui fait suite à cette introduction :

- Chapitre 2 : décrit au départ les travaux antérieurs relatifs à la sémantique cognitive et ensuite, met en lumière les concepts et les théories issues de la sémantique cognitive qui sont employées dans l'exécution de l'analyse (i.e. la polysémie en contexte, la métaphore, la métonymie et l'intégration conceptuelle) ainsi que les perspectives nouvelles sur les constructions phraséologiques, inspirées du courant cognitif ;

- Chapitre 3 : détaille dans un premier temps le contexte historique de la traduction automatique et dans une deuxième étape, situe l'apparition de la machine neuronale de Google Translate sur la scène de l'intelligence artificielle ;

- Chapitre 4 : justifie les choix méthodologiques, présente à la fois le corpus et les méthodes d'analyse ;

- Chapitre 5 : interprète les origines des erreurs relevant du langage général et spécialisé au niveau des unités lexicales simples et composées et au niveau des collocations et locutions idiomatiques sous l'éclairage des théories passées en revue au Ch. 2 ;

- Chapitre 6 : fait connaître les contributions originales et les résultats obtenus à partir des tableaux, figures et explications du Ch. 5 en les reliant aux questions de recherche ;

- Chapitre 7 : explore la possibilité de mener de futures études dans la continuité de celle qui a été déjà réalisée.

4. Méthodologie de la recherche

4.1 Description et fiabilité du corpus

Notre corpus contient 100 articles portant sur les technologies émergentes sur le marché ou dont l'apparition est attendue dans un futur pas trop lointain dans le monde entier. Premièrement, il faut souligner que le corpus de notre thèse est un « corpus parallèle et multilingue » parce qu'il est composé de textes écrits en français et leurs équivalences supposées en roumain et en anglais. De plus, le corpus est aussi « homogène » grâce au fait que tous les textes appartiennent au même style d'écriture, à savoir celui de vulgarisation scientifique. Les articles comprennent par la suite des détails sur les parties constitutives de nouvelles technologies, leurs processus de fonctionnement ainsi que d'autres spécifications qui s'y rattachent. Par conséquent, selon quelques représentants de la linguistique du corpus, notamment Bowker et Pearson (2002, pp.11-13), nous pouvons affirmer que nos articles sont un « corpus de spécialité ». De plus, comme nous n'allons pas ajouter d'autres textes avant la fin de la recherche, il peut aussi être qualifié de « corpus clos ».

Étant donné que notre étude porte sur une « traduction automatique spécialisée », nous sommes censée inclure dans cette section des détails concernant : a) la TS et les défis posés par ce genre de traduction ; b) l'entrecroisement des concepts de TS et VS ; c) la définition, le rôle et les caractéristiques de la VS ; d) le rôle du vulgarisateur et du public ; e) les changements les plus récents dans la manière dans laquelle on fait de la VS.

Pour commencer, « la langue de spécialité » ou « le langage spécialisé » est soumis à un examen approfondi à partir des années 60 lors de l'apparition de la linguistique appliquée au sein de laquelle se sont développés les champs intitulés terminologie et linguistique du corpus. Il est important de noter que le langage de spécialité est vu comme un langage restreint, destiné exclusivement aux professionnels d'un certain domaine contrairement à la langue courante avec laquelle elle est contingente.

Même si au début, on croyait qu'à cause de son lexique spécialisé, seulement ces professionnels pouvaient le comprendre, Müller (1985, p. 187) a avancé la notion de « degrés de spécialisation ». Tant que le message, le destinataire et la culture de ceux qui sont les bénéficiaires de ce transfert de savoir sont différents, il va y avoir de différents degrés de spécialisation. À cause du progrès technologique et surtout du numérique qui a eu comme premier bénéfice l'accès au savoir, on constate que même le savoir limité jusqu'alors à une certaine communauté scientifique se voit partager.

En effet, le public général intéressé à n'importe quel sujet commence à accéder au langage spécialisé à l'aide de la vulgarisation, son langage de diffusion. La vulgarisation scientifique peut davantage être qualifiée de « traduction intralangue », par sa technique principale qui est la reformulation du contenu sémantique et de la terminologie parfois lourde et encombrante pour le lecteur qu'on trouve dans un texte fort spécialisé. Que ce soit la TS ou la VS, elles ont toutes les deux la même fonction, soit celle de sensibiliser le grand public aux informations émanant de l'univers scientifique. Quant aux textes de

vulgarisation scientifique, il est sûr qu'il va y avoir des pertes sémantiques, mais cela est vraiment nécessaire pour aboutir à leur composition.

Selon Bruno Dufay (2005, p. 33), on peut identifier quatre objectifs de la VS : informer, expliquer, éveiller le sens critique et attirer l'attention. La chose la plus difficile à faire par le journaliste scientifique est d'allécher le lecteur tout en restant objectif et en résistant à la tentation de créer du sensationnel. À présent, il y a une tendance de le faire afin de devenir commercial et vendable. Dufay (2005) formule les techniques à appliquer pour faire de la bonne vulgarisation : la comparaison et la métaphore, l'exemple, l'association des sujets, le condensé et la synthèse, l'histoire et l'anecdote, le raisonnement logique, le débat contradictoire et le questionnement, la méthode pointilliste, le jeu et l'astuce, un vocabulaire accessible, l'accroche et la surprise.

En ce qui concerne son rôle, la vulgarisation est essentielle dans les pays du tiers monde ou en voie de développement où l'information scientifique fait défaut. La science est la clé du développement et par conséquent, l'objectif central des élites qui y sont consacrées devrait être celui de vulgariser. Pour les scientifiques, cette activité devrait être une obligation morale, une dette qu'ils sont tenus de payer pour avoir été investis du savoir. Ils sont les seuls qui peuvent assurer la rectitude et la rigueur de l'information scientifique. Quand même, il est question de savoir si ceux qui en font le partage sont des savants, de vrais experts ou au moins, des personnes expérimentées dans le domaine qu'ils font découvrir aux autres. On peut rappeler à titre d'exemple William Herschel qui considérait qu'il y avait « de bonnes raisons de penser que le Soleil pouvait être habité » (Meadows, 1986, p. 397). Malgré son manque de base réelle, cela n'a pas empêché que cette idée loufoque a été relayée à son époque.

Cela montre que « le vulgarisateur » peut avoir un rôle décisif dans le monde. Il pourrait, tout aussi bien, changer le cours d'une crise humanitaire, la façon dont elle est perçue par les gens à cause de la manière de présenter les faits. Prenons des exemples concrets, les problèmes pressants auxquels l'humanité s'est vue confronter récemment : le Covid, la grippe aviaire, la grippe porcine, la vache folle, le réchauffement climatique, le clonage. Si la personne qui mène un débat sur un de ces sujets ou rédige un discours écrit pour le public général est un non avisé, cela peut facilement faire la différence entre la vie et la mort de toute une communauté ou un peuple. Imaginez-vous qu'un nouveau médicament pour traiter le Covid ou même éradiquer le cancer vient de sortir sur le marché. À part les institutions gouvernementales et internationales concernées, c'est aussi le rôle du vulgarisateur de présenter les avantages ainsi que les périls auxquels la population s'expose en le prenant et surtout, ses effets secondaires. Être en possession d'une telle expertise est quelque chose de vital compte tenu du fait qu'il y a de plus en plus de magazines, sites Internet, blogs scientifiques qui prétendent faire de la vulgarisation scientifique. Il s'avère, en conséquence, vraiment difficile à séparer le bon grain de l'ivraie.

Par ailleurs, on s'inquiète du fait qu'un vulgarisateur n'est plus à même de traiter un sujet de façon impartiale parce que maintenant, l'accent n'est plus mis sur le fait d'« expliquer » la science, mais sur la « médiatisation des faits scientifiques et techniques qui repose sur une mise en scène des

controverses entre des sphères d'activité qui n'ont pas les mêmes intérêts » (Moirand, Reboul-Touré, Pordeus Ribeiro, 2016, p. 144).

De plus, les auteurs ci-dessus nous informent que nous assistons à un basculement du rôle du public de la VS. Si au début de la VS, il détenait un rôle purement « interprétatif » et par excellence « passif », à présent, le public a une « instance productive ». Cela est dû aux avancées technologiques et à l'avènement des réseaux sociaux parce que maintenant le public peut répondre à des articles écrits sur des blogs scientifiques, détenus quelquefois par des amateurs, ou sites Internet de VS en envoyant des tweets, des messages sur Facebook ou des messages aux courriers des lecteurs. Donc, on est face à un public qui est devenu réactif face à l'information qu'on lui met sous les yeux, ce qui est une bonne chose parce qu'il peut dépister le vulgarisateur profane.

En ce qui concerne les articles qui constituent notre corpus, on se situe bien à l'écart des soupçons de ce genre. Cela découle du fait que les vulgarisateurs qui écrivent pour la revue Science et Vie sont des journalistes scientifiques à double spécialisation. En fait, la plupart ont suivi des études de licence et/ou de master dans le domaine auquel ils s'adonnent. Pour compléter ces compétences, ils ont fait également des études de journalisme scientifique, chose qui les qualifie sans équivoque de spécialistes de l'écriture de VS. Il en découle que nous sommes obligée de reconnaître le caractère véridique et qualitatif de leurs textes. Et manifestement, nous ne pouvons que les considérer comme de vrais promoteurs de la découverte scientifique. Finalement, nous estimons que leurs articles représentent de l'or pour notre recherche sur le plan de la qualité, inventivité, objectivité, fiabilité des sources et ainsi de suite.

4.2 Modalités de collecte, format et prétraitement du corpus

Les articles qui constituent notre corpus ont été extraits du site Internet de Science et Vie et de son magazine sur support numérique grâce à l'amabilité de la direction de la revue qui veut évidemment faire avancer les connaissances dans le domaine de la linguistique. Les 100 articles que nous avons recueillis ont été sauvegardés en annexe (Annexe A– Corpus français) en vue de leur utilisation au sein de notre recherche.

Les articles composant l'Annexe A sont munis premièrement d'un numéro de 1 à 100, ce qui signale l'ordre de collecte. Cette notation est suivie par les lettres « FR » qui veulent dire qu'il s'agit d'un texte en français. Ensuite viennent le nom de l'auteur, le mois ainsi que l'an de parution de l'article et finalement, le nombre de caractères avec espaces que le texte comporte. Toutes ces caractéristiques concernant l'article se trouvent dans la partie gauche qui précède le titre. Quant au nom du vulgarisateur, celui-ci peut figurer soit de façon intégrale, soit en initiales (i.e. par L.B.). Le choix appartient clairement à l'auteur et nous devons informer le lecteur de notre thèse qu'il y a des vulgarisateurs qui publient dans la revue et sur le site sous leur vrai nom, les initiales de leur vrai nom, un pseudonyme ou tout simplement, les initiales de leur pseudonyme. Qu'ils publient sous un

pseudonyme ou non, ils sont tous incontestablement, de vrais maîtres de ce qui est l'art de faire de la vulgarisation.

De l'autre côté, dans l'Annexe B – Corpus roumain, les textes qui reprennent la traduction automatique en roumain des textes source comportent le même numéro d'ordre accompagné, cette fois-ci, de l'abréviation « RO » qui est un raccourcissement du mot roumain. Le lecteur de notre thèse va pouvoir constater aussi que les trois détails, dont on a discuté dans le paragraphe antérieur, ne se retrouvent pas dans les deux autres annexes de corpus. En fait, nous avons jugé bon de ne pas les inclure pour les *tertia comparationis* de peur que l'information ne devienne redondante.

Sur les trois éléments indiqués ci-dessus, l'un d'entre eux est extrêmement important, à savoir, le nombre de caractères et espaces qui fonctionne également comme un indice de la longueur de l'article. Mentionner cela compte énormément dans le cadre de notre travail à cause du mécanisme de traduction de GTNM. L'interface de Google Translate permet la traduction d'un segment de texte ayant 5 000 caractères et espaces, la limite supérieure. Partant de cette règle qui nous est imposée et par souci de cohésion du texte et homogénéité du sens, nous avons décidé de ne découper aucun texte lors de sa traduction. Conséquemment, même si initialement nous avions collecté des textes dont l'étendue dépassait les 5 000 caractères et espaces, nous les avons finalement enlevés du corpus. La MTN a traité donc des textes intégraux de façon à ce que cet attribut d'entièreté soit également transféré à la signification globale du texte traduit. Il faut rappeler que GTNM crée et rend compte de la sémantique de l'ensemble des phrases qui constituent un article, ce qui justifie notre choix des textes ayant sous 5 000 caractères et espaces. En ce qui est de la limite inférieure, nous avons estimé qu'un texte ayant moins de 500 caractères et espaces ne pourrait pas être légitimement qualifié d'article. Cette option s'est traduite dans un éventail d'articles partant de 546 caractères et espaces (i.e., l'article 1FR) et allant jusqu'à 4 919 (i.e. article 78FR).

Dernier point, mais pas le moindre, comme notre but est celui d'examiner la qualité de la traduction automatique de ces articles et que GTNM ne peut pas traiter du paratexte (i.e., sigles, schémas, diagrammes, dessins, photos), tout élément s'y rapportant va être ignoré. C'est vrai pourtant que le paratexte assurerait la complétude du texte dans un article de vulgarisation scientifique et qu'il est un critère de bonne vulgarisation. Mais le fait que nous ne pouvons analyser aucun constituant du paratexte n'est pas de notre faute. Ce n'est pas une faiblesse qui peut nous être attribuée. Malheureusement, notre recherche est limitée par la machine neuronale même qui nous oblige, pour le moment, de nous contenter de faire seulement l'analyse du corps de l'article.

Outre ces deux annexes principales, notre corpus a été également traduit en anglais, ce qui correspond à l'Annexe C – Corpus anglais. Nous avons fait cela parce que l'une de nos questions de recherche vise à découvrir si l'anglais, qui reste la langue principale d'entraînement des réseaux de neurones artificiels, joue un rôle dans la production des erreurs du français en roumain. La raison principale pour laquelle l'anglais joue un rôle important pour la traduction automatique neuronale est l'existence d'un grand volume de données digitalisées et des corpus (traduits) parallèles de et vers l'anglais.

En effet, l'anglais se comporte comme une « langue pivot » de la TAN de Google, mais sans vraiment l'être ; au moins, au sens de ce qui implique le mécanisme *par interlingua*, c'est-à-dire, des paires de langues individuelles comme FR →/EN/→RO.

Cependant, dans le cadre de GTNM, un système qui traduit à partir de et vers plusieurs langues, on peut qualifier l'anglais de « langue intermédiaire nec plus ultra ». La machine utilise plusieurs encodeurs et décodeurs. Toute l'information linguistique appartenant à toutes les paires de langues et que le système trouve relevante est utilisée par la machine afin d'apprendre et arriver de la langue source à la langue cible. L'équipe Google Brain envisage une architecture neuronale *Zero-Shot* qui devrait être à même de traduire directement entre deux langues que le système n'a pas du tout vues.

Selon un article datant de septembre 2021 par Wang et al. (p. 4321), la réalisation d'une traduction de haute qualité en mode « zéro » est un défi de taille et ne représente, pour l'instant, qu'une fonction prometteuse de la machine de traduction neuronale multilingue. Même si l'architecture qui se base sur l'anglais comme lien implicite ou *implicit bridging* (i.e. le *Zero-Shot translation*), n'est pas encore performante, celle qui emploie l'anglais comme lien explicite ou *explicit bridging* est en place. Cela nous oblige à examiner son rapport avec les erreurs de sortie.

4.3 Traitement du corpus

Le mot « exploitation » du syntagme qui constitue le titre du présent sous-chapitre pourrait amener le lecteur à imaginer un chapitre incluant des explications portant sur l'analyse linguistique proprement dite. En réalité, il concerne les étapes de 1 à 3 ci-dessous, réalisées en amont de l'analyse et qui sont des éléments du fichier Excel « Pré-analyse_Corpus aligné » :

1. alignement du corpus multilingue

2. annotation du corpus

3. structuration des niveaux sémantiques

Ce travail préanalytique a été mis en forme à l'aide de quelques logiciels et coagulé dans un tableur Excel, une méthode obligatoire étant donnée la nature de nos objectifs. Ce n'est qu'en alignant, annotant et structurant les plans sémantiques dans un fichier Excel que nous avons pu ensuite mettre en œuvre des méthodes d'analyse quantitative.

Il est indispensable de produire des statistiques, car elles montrent que nos résultats ne sont pas de simples suppositions. En outre, elles nous aident à mieux structurer les données, mieux comprendre les résultats de la recherche, à découvrir des modèles et à faire des prévisions.

4.3.1 Alignement automatique – méthode et outils

Le premier stade de la préparation de notre corpus en vue de l'analyse a été l'utilisation de l'outil LF Aligner [1] (Farkas, 2015), une application capable d'aligner des textes dans plusieurs langues simultanément et avec précision sans faire recours à un outil TAO. En fait, elle nous a permis de générer un fichier XLS à partir de trois documents .docx (i.e. le document contenant le discours source et les deux autres documents contenant les traductions automatiques en anglais et français). Le logiciel en version 4.1 s'appuyant sur Hunalign[2] a réalisé automatiquement l'appariement des phrases dans les cellules du tableur.

Après avoir vérifié et corrigé les quelques inadvertances d'alignement, nous avons obtenu une base de données en Excel, chose qui nous a permis dans une étape ultérieure d'identifier et annoter chaque erreur manuellement. Cela a facilité l'utilisation du logiciel AntpConc[3], outil d'analyse linguistique pour le corpus parallèle. Nous avons réussi de cette façon à mieux traiter notre recueil de documents. Parmi les nombreuses façons dont ce programme nous a aidée, nous pouvons citer la récupération de toutes les occurrences d'un certain lexème dans le corpus français, roumain ou anglais et l'identification de son co-texte.

4.3.2 Annotation manuelle – procédure et conventions

Le format de l'erreur est de type « erreur_nombre1.nombre2 ». L'erreur se trouve surlignée en rouge et porte le même numéro de référence dans les trois langues qui entrent dans la composition de notre corpus. Par ailleurs, le format « ERREUR _nombre1.nombre2 » donne des indications additionnelles concernant l'emplacement de l'erreur. Il nous indique que l'erreur est extraite soit d'un titre ou d'un sous-article qui est écrit en majuscules. Conséquemment, tout item linguistique qui figure en majuscule dans les feuilles de notre fichier Excel « Pré-analyse_Corpus aligné » se situe au niveau du titre ou du sous-titre des textes.

Lorsque nous n'avons pas pu déceler le moindre détournement de sens que nous pouvons imputer au LI, nous n'avons ni numéroté ni surligné en rouge l'élément linguistique en question. En ce qui a trait à l'absence d'interférence du LI dans la production d'une erreur, celle-ci est signalée par une ligne oblique tandis que la non-traduction d'un élément linguistique est indiquée par le nombre « 0 ».

---

[1] LF Aligner est a priori une application qui permet aux traducteurs de générer des mémoires de traduction à partir de plusieurs types de documents et de leurs traductions. Elle parvient à générer des fichiers TXT, TMX ou XLS. L'application peut être téléchargée à l'adresse https://sourceforge.net/projects/aligner/.

[2] Varga D., Németh L., P. Halácsy P., A. Kornai A., V. Trón V., V. Nagy V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, pp. 590-596. https://kornai.com/Papers/ranlp05parallel.pdf ;

[3] Anthony, L. (2013). AntPConc (Version 1.0.2) [Computer Software]. Tokyo, Japan : Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Par souci de rapidité et de concision, tout au long de notre analyse, nous utiliserons souvent les abréviations en usage dans le fichier Excel « Pré-analyse_Corpus Aligné ». Le numéro de l'erreur peut être suivi parfois par FR, RO ou EN en fonction de la langue du corpus dont il était issu (ex. : 48.11_FR ; 48.11_RO ; 48.11_EN). De même, la lettre T ou G peut être attachée à la suite de l'erreur, indiquant l'appartenance au langage technique ou général (ex. : 48.11_G ; 48.11_EN_G). Ensuite, les notations LexS, LexC, Co et Loc interviennent en complément de celles mentionnées auparavant afin de préciser le niveau sémantique auquel l'erreur se situe (ex : 1.2_RO_G_LexS ; 12.8_EN_T_Co ; 5.7_Loc ;6.2_G_LexC). Il faut toutefois préciser que les trois types de notations offrant des informations complémentaires peuvent apparaître simultanément, individuellement ou sous différentes combinaisons entre elles.

4.3.3 Structuration - critères spécifiques d'analyse

Notre recherche s'articule autour de 4 niveaux sémantiques, chacun répertorié dans une feuille Excel distincte (i.e. LexS, Lex C, Co, Loc). Chaque feuille, à l'exception de la feuille Excel Loc, présente un ensemble uniforme de 10 colonnes, chacune mettant en évidence des caractéristiques spécifiques à son niveau respectif. Il convient de noter que la dernière colonne « Référent dans la réalité extra-linguistique » ne revêt pas de pertinence pour la catégorie Loc de sorte que les locutions ne sont pas associées à des référents dans le monde réel.

Ceci n'implique pas que chaque lexie des trois catégories restantes soit associée à un référent. Cette corrélation ne s'applique qu'à un nombre limité de lexies, susceptibles de poser problème lorsqu'on essaie de les relier à des objets physiques qu'on peut percevoir dans notre environnement quotidien. De plus, le référent est représenté par une photographie illustrant l'objet, tirée de l'article faisant partie de notre corpus et où l'on en parle.

Le nombre de lignes, quant à lui, varie en fonction du nombre d'erreurs identifiées pour chaque groupe. La première catégorie d'erreurs où nous avons listé des lexies individuelles figure dans la feuille Excel « LexS ». Pour chaque membre de cette catégorie, nous avons opté pour une description préanalytique comprenant des spécificités telles que : chiffres de référence, appartenance au langage général ou technique, typologie syntaxique, cause de l'erreur, brève pré-analyse, forme graphique en LS et LC, forme en LI si et seulement si elle trahit une participation à la mauvaise traduction et finalement, une réparation qui n'est conçue que pour le contexte particulier de l'erreur en question.

La réparation, que nous appelons « remédiation contextuelle », est le résultat de nombreuses consultations de sources roumaines fiables surtout pour les unités appartenant à la langue technique. Lorsque nous l'avons jugé utile, ces sources ont été incluses dans la « brève pré-analyse ». Dans les cas où nous avons découvert plusieurs variantes en usage, l'équivalent roumain sélectionné a été celui dont la fréquence d'utilisation est nettement supérieure. Lorsque des variantes présentent des taux d'utilisation presque égaux dans la LC, elles sont toutes incluses dans la fiche de l'unité en question. Ce dernier cas se rencontre plutôt chez les unités lexicales ou polylexicales qui constituent des éléments

de la langue technique. Comme vous pouvez le constater dans le cas de 41,9, il y a deux variantes d'équivalence principales en roumain, notamment *învățare prin consolidare* et *învățare prin întărire*. L'occurrence 41.9 reflète une notion néologique et c'est la raison pour laquelle il n'existe pas encore de terminologie précise en roumain pour la désigner. Il y a beaucoup de cas comme celui-ci et même des cas où les concepts sont si nouveaux que le roumain n'a pas encore eu le temps de les nommer. La plupart des fois, les dénominations sont empruntées à l'anglais dans leur entièreté ou du moins l'un de ses constituants l'est quand il s'agit des collocations du domaine spécialisé. On peut citer à titre d'exemple la collocation roumaine *eoliană offshore*, où le collocatif a été pris en tant que tel de l'anglais.

Mais, on insistera sur le sujet des remédiations et de la régulation des termes étrangers dans le roumain dans le chapitre dédié aux prolongations de la thèse. Nous allons y montrer en quoi notre travail est une source précieuse de données pour la constitution d'un glossaire terminologique dédié aux dénominations que portent les concepts novateurs selon l'usage courant. D'ailleurs, notre travail peut servir de base à la création d'un glossaire des propositions linguistiques de substitution qui puissent fournir des alternatives roumaines à la néo-terminologie anglaise.

Ensuite, les erreurs qui impliquent des unités linguistiques composées se regroupent dans la feuille Excel « LexC ». Leurs fiches descriptives renferment le même type d'information, à l'exception de l'encadrement syntaxique. Il est plus pertinent d'accompagner ces lexies par un critère de genre « typologique morphologique » parce que cela permet de considérer la décomposition en morphèmes. De cette façon, nous accédons à un schéma brut de construction de sens, d'entassement des couches de signification. La mise en exergue de cet entrelacement de sens permet de mieux appréhender leur non-compréhension par la MTN.

Les erreurs appartenant au domaine des combinaisons fréquentes des lexies qui s'associent naturellement dans le langage général et spécialisé à la fois figurent dans la feuille Excel « Co ». Les collocations sont pourvues des mêmes filtres de caractérisation que les lexies simples. Nous avons ajouté dans le cadre de la feuille, les constructions à verbe causatif et celles à verbe support (i.e. colligations). Celles-ci jouissent d'un sémantisme actualisé soit par un verbe à l'infinitif soit par un nom prédicatif. Et c'est en raison du fait que certains de leurs formants sont totalement ou partiellement vides de sens qu'elles font l'objet d'une sous-liste.

Deux autres feuilles Excel sont également incluses dans le fichier dédié à la pré-analyse des données. Elles sont spécifiquement conçues pour les légendes qui expliquent les raisons de production des erreurs de traduction. L'une est consacrée à la légende propre aux lexies simples et composées, tandis que l'autre concerne le niveau des syntagmes phraséologiques. Sans ces feuilles Excel supplémentaires, l'acquisition de quelques-unes de nos données statistiques serait impossible.

5. Conclusions générales

Cette thèse interdisciplinaire a été conçue de manière à ce qu'elle arrive à constituer avant tout, un modèle polyvalent d'étude sémantique d'inspiration cognitive appliquée à une technologie de TA avec IA sur un corpus de spécialité, multilingue. À cet effet, 100 articles de VS ont été alignés automatiquement à l'aide de LF Aligner alors que les résultats de traduction incorrecte fournis par la GTNM ont été annotés et inventoriés manuellement en fonction de plusieurs facteurs (i.e. dimension syntaxique, morphologique, langage technique ou général, cause de l'erreur, interférence du LI, remédiation contextuelle, etc.). Cela nous a permis davantage de les convertir en données statistiques afin d'établir le coefficient précis de qualité sémantique qui leur revient.

En fait, notre thèse a réussi à dépister toutes les fautes de traduction livrées en 2018/2019 par la technologie débutante renforcée de Google en matière de *deep learning* (i.e. le système Transformateur en libre accès) au niveau des constructions sémantiques réduites et élargies. Ces constructions ont été ensuite explorées en profondeur grâce à la mobilisation des outils d'analyse issus de la sémantique cognitive. Étant parfaitement adaptée aux particularités de la TAN, la sémantique cognitive s'est avérée le choix adéquat en ce qui concerne la stratégie d'analyse des données. Nous rappelons que le mécanisme d'attention de Badhanau incorporé en 2017 dans la machine de TAN de Google met l'accent sur le contexte et sur la compréhension du texte au niveau global. En outre, cette représentation dynamique du sens effectuée par les réseaux de neurones artificiels est un modèle comparable à l'approche dynamique proposée par la SC par rapport à la construction du sens.

Par la suite, pour pouvoir optimiser l'intégration des phénomènes sémantiques complexes dans les traducteurs automatiques neuronaux, les futures recherches en traitement du langage naturel et IA peuvent s'appuyer sur nos résultats de recherche. Pour aboutir à ces résultats, nous avons formulé des questions de recherche capables de nous montrer à quel point les choix de la TA neuronale se différencient des constructions mentales humaines et dérivent de la langue d'entraînement du réseau. Pour cette raison, la présentation des résultats génériques de notre recherche s'articule autour des questions de recherche de la thèse et sera composée des sections I à IV ci-après :

I. L'INFLUENCE DE LA POLYSÉMIE DANS LA PRODUCTION DES ERREURS

Un principe qui régit la SC établit que la représentation du sens est encyclopédique, ce qui équivaut à un rejet de l'approche définitionnelle. Compte tenu du fait que dans le cadre de la SC, la sémantique va de pair avec la pragmatique, la sélection d'un sens particulier ne peut pas se faire en dehors du contexte. Vu que le réseau neuronal artificiel de Google partage cette approche à l'égard de la désambiguïsation lexicale, l'enjeu principal qu'il est appelé à soulever est le choix d'une variante parmi les alternatives sens propre/figuré/commun/spécialisé/modulé, qui ne se révèlent qu'en contexte.

Au niveau LexS, où les lexies nominales prédominent avec 181 occurrences, la polysémie a été responsable d'un tiers des ruptures de sens, fait qui a conduit à 86 erreurs sur un total de 288. Ce groupe comporte 47 lexies issues du langage général contre 39 lexies spécialisées. La polysémie s'est

manifestée de plusieurs façons, mais plus de la moitié des cas sont liés au LI. De surcroît, 6 erreurs de ce type portent sur les lexies françaises qui sont des emprunts à l'anglais. En fait, la polysémie du LI est la première cause d'apparition des fautes au niveau des lexies françaises empruntées à l'anglais. Sur les 3 emprunts à l'anglais dont le transfert faussé est dû au LI, record_56.2/93.3 est un exemple de problème de TA liée à une polysémie conventionnelle alors que les emprunts intégraux flash_38.6/38.10 et brownie_64.6 acquièrent un microsens contextuel qui n'a pas été correctement décodé par la machine.

Au niveau LexC, où l'on ne trouve presque que des lexies nominales, nous avons repéré 5 situations de traduction déficiente sur 57 dont la source est la polysémie. Dans cette catégorie entrent deux termes, c'est-à-dire l'emprunt à l'anglais X-plane_18.4/18.6 et prise de vue_37.6/94.3. La dernière lexie composée fait l'objet d'une modulation au sein du contexte source.

Au niveau Co, où les erreurs classées sous les typologies syntaxiques N Adj et N prép N sont prévalentes, la polysémie n'intervient que rarement, à savoir 19 fois sur 152 :

- 11 cas de polysémie qui se manifeste soit au niveau de la base de la collocation, soit au niveau du collocatif ou au niveau de l'ensemble phraséologique livré en LI (voir 11A+11B+11C) ;
- 8 cas de polysémie soit du collocateur, soit du collocant de la construction collocative de la LS (voir 10A+10B).

En ce qui concerne la polysémie du LI, elle porte sur cinq G_Co_std, 4 T_Co_std, une T_Collig de type « faire-vb support » et une G_Collig de type « avoir-vb support ». Quant à la polysémie qui ne fait pas intervenir le LI, elle touche quatre collocations relevant du lexique général, trois collocations terminologiques et une T_Collig. Chaque typologique syntaxique au sein des Co_std faisant partie du podium relatif à la fréquence des erreurs y est représentée. Une construction de type « faire causatif » leur est également rattachée.

Au niveau Loc, où la plupart des erreurs constituent des locutions verbales, la polysémie n'a engendré qu'une erreur. Elle a trait à l'unité polylexicale figée « mettre en service_8.1 » dans le cas de laquelle la polysémie de la lexie obtenue en LI entraîne l'erreur de sortie.

II. L'INFLUENCE DU CONTEXTE DANS LA PRODUCTION DES ERREURS

La première section qui offre une vue d'ensemble sur l'influence de la polysémie dans l'émergence des sens compromis nous a fait comprendre qu'en définitive, dans le cas de toute erreur, nous avons affaire à un manque du traitement correct du contexte par l'architecture neuronale de Google. La signification des constructions faisant partie de notre corpus en LS est fidèlement reproduite par le cerveau artificiel en LC uniquement s'il maîtrise les aspects ci-dessous :

- les sens institutionnalisés d'une construction lexicale petite ou large de la LS (i.e. sens connotatif, dénotatif, spécialisé) ;
- les sens nouveaux attachés aux constructions préexistantes en LS ;

- les néotermes et les néophrasèmes de la langue source ;
- la signification des sigles et abréviations ;
- la forme et les particularités d'un titre/sous-titre ;
- les particularités des noms propres.

Or, le fait que la machine ne peut pas encore traiter adéquatement tous les aspects ci-dessus a généré :

- des erreurs de traduction au niveau des lexies et syntagmes phraséologiques émanant à la fois du LG et du LT, outre celles liées au LI ou à la polysémie ;
- une altération du sens des titres/sous-titres qui contiennent souvent des jeux de mots ou des sous-entendus fondés sur des tropes conceptuels ;
- des transferts incorrects au niveau des noms propres, majuscules, sigles et abréviations ;
- des segments vides de traduction ;
- des erreurs de traduction récurrentes ;
- une fabrication de référents fictifs ;
- fausse interprétation des constructions causatives et à verbe support.

Les traductions sans rapport au LI ou à la polysémie sont groupées sous 6A pour le niveau LexS et LexC et 14A/B/C pour le niveau des syntagmes phraséologiques. La cause 6A totalise 36 erreurs au niveau des LexS dont 25 sont des lexies générales. Ensuite, le niveau Co_std dénombre 22 erreurs de ce type qui portent majoritairement sur les collocations spécialisées tandis que le niveau Collig. en recense 5. Au niveau LexC, la cause 6A renferme 11 erreurs parmi lesquelles 9 sont des termes composés. En ce qui concerne les constructions idiomatiques, parmi elles ont été identifiés seulement trois cas de traduction dépourvus de toute liaison au LI ou à la polysémie, à savoir deux locutions nominales (i.e. rappel des faits_96.3 et coup d'éclat_49.5) et une locution verbale (i.e. monter le son_83.1). Leur transfert incorrect en LC a cependant une connexion aux tropes conceptuels qu'elles abritent.

La déformation du sens des titres et des sous-titres résulte principalement du contexte limité qui les caractérise. Les traducteurs neuronaux traitent parfois les titres et les sous-titres de manière isolée par rapport au contexte global, ce qui entraîne, bien entendu, des erreurs dans l'interprétation de l'implicite. Au niveau des titres et des sous-titres, le vulgarisateur insère parfois des constructions idiomatiques dont les nuances ne sont pas capturées par le mécanisme de TA neuronale (i.e. prendre le large_5.1, viser le large_7.1). En raison de la courte taille des titres et des sous-titres, des traductions déformées y surviennent. Mis à part les CVF précédentes, voici les autres constructions qui participent à la signification des titres et qui ont été mal comprises : LexS (i.e. chargé_10.2, aller_ 10.3, supersonique_18.1, arriver_24.1, duvet_33.1, bourdon_43.1, feu_44.1, embouteillage_54.1, drone_55.1, mollusque_60.1, durer_60.2, crécelle_63.1, fractale_75.1_végétal_79.1, tenir_82.1, skate_85.1, connecté_81.3, s'enrouler_86.1, se lacer_88.1, métro_92.1, colorant_99.1), LexC (valise accordéon_10.1, monoroue_72.2, éthylotest_74.1, autonettoyante_81.2), Co (i.e. lave-linge à pédale_11.1, partie mobile_36.1, essaim de minirobots_47.1, journal télévisé_68.1, véhicule tout-

terrain_70.1, trottinette électrique_72.1, litière pour chat_81.1, salto arrière_90.1) et Collig (faire bang_18.2, faire des gâteaux_64.1). Une observation qui s'impose ici également est le fait qu'il y a des cas où les items linguistiques sont traités correctement par la machine dans le corps de l'article, mais de façon fautive dans le titre ou sous-titre de l'article (ex. métro). Un autre aspect à relever est que quelques-unes de ces unités linguistiques génèrent plusieurs fois la même erreur ou des erreurs différentes tout au long de l'article (ex. bourdon, mollusque, duvet, crécelle, monoroue, salto arrière).

Les erreurs de traduction récurrentes du premier niveau d'analyse sémantique sont recouvertes par les sous-chapitres 5.1.2.3_LexS et 5.1.3.3_LexC. Parmi les 23 erreurs qui se répètent et qui ont généré chacune entre 2 et 10 erreurs au niveau LexS, 12 sont des lexies spécialisées et 9 des lexies courantes. Celles-ci ont eu de multiples causes de production, mais l'interférence du LI a mis son empreinte sur la majorité (i.e. 5C et 4A). En ce qui est du niveau LexC, il regroupe 13 lexies, à savoir 7 qui livrent le même résultat de sortie (X-plane, memristor, tout-en-un, prise de vue, feu rouge, mille-pattes, éthylotest) et 6 dont la traduction finale varie. Bien que les erreurs récurrentes qui sont attachées aux structures collocatives ne fassent pas l'objet d'un sous-chapitre distinct dans le cadre de la thèse, nous les discuterons ici. Elles se subdivisent en deux groupes, un qui recense les associations lexicales qui produisent le même résultat en LC et l'autre qui englobe celles qui donnent lieu à deux résultats différents. Le premier groupe comprend 6 collocations spécialisées, dont 3 comportent un sigle, ainsi qu'une collocation issue du vocabulaire général qui aboutit à trois traductions identiques (voir première mondiale_42.1/68.2/94.4). L'autre groupe affiche 5 collocations terminologiques et une collocation du LG qui sont majoritairement transférées en LC de manière viciée à cause de l'interférence du LI ; 4 cas de traduction littérale du LI (voir apprentissage par renforcement_41.8, station émettrice_43.13 & 43,11, trajet travail-domicile_72.6), 3 cas de reproduction du LI (voir affichage tête haute_80.2, salto arrière_90.1/90.3 & 90,6) et deux cas de polysémie du LI au niveau du collocatif (voir apprentissage par renforcement_41.9, essaim de minirobots_47.1). Une remarque finale qui me semble s'imposer, c'est le fait que, de toutes ces erreurs répétitives, certaines constituent des néologismes, emprunts à l'anglais, éléments de signification des titres et sous-titres et générateurs de lexies inexistantes.

Les segments vides de traduction sont répertoriés sous la cause 7 et 17 et ils sont présents uniquement au niveau LexS et Co. L'absence de la traduction au niveau des lexies simples est étudiée plus en profondeur dans le cadre du sous-chapitre 5.1.2.4. Comme nous l'avons démontré, elle peut avoir lieu seulement en LC ou en LI ainsi qu'en LC et elle peut découler du fait que :

- les lexies à traduire font partie du domaine technique (voir matériau _28.11, laser_38.8, houle_96.14) ;
- le sens des lexies est étranger à la machine (voir intémperie_26.3, adapté_48.12, lacé_88.3) ;
- les lexies recèlent un caractère métaphorique (voir ligne_86.7).

Il semblerait également que le mécanisme de limitation des pertes de traduction dicte parfois des reproductions des segments voisins au niveau LexS de façon à ce que la cohérence du texte de sortie ne soit pas ruinée (voir intémperie_26.3, lacé_88.3).

De l'autre côté, les colligations offrent uniquement 4 situations dans lesquelles la non-traduction peut être aperçue. Trois de ces quatre sont des collocations terminologiques dont le collocant n'est pas du tout transposé en LC (voir éolienne offshore_7.2, avion à réaction_36.13, courant continuu_93.2).

Les traductions par des lexies inventées se réunissent sous la cause 6B pour le niveau des LexS et LexC et sous la cause 18A/B pour le niveau Co & Loc. Les unités lexicales simples et polylexicales fabriquées par GTNM en LT et LG sont listées ci-après :

- G_LexS : OBSTACULURI_15.1, trântitor_18.5, ștafari_50.1, tobacconiști_50,5, tacamici_50.8, schiț_51.2, RETIREA_62.5, indisociat_72.19, Andele_76.4, Oița_91.1 ;
- T_LexS : etanșitate_5.5, PROZETA_16.3, moluscul_60.1/60.3/60.10, cubă_73.2, reedită_97.1 ;
- T_LexC : monoplaz_31.3, CUPURI DE BARBE_50.4, difenilamino-clorarsină_69.6 ;
- G_Co : haina *zarbată*_47.6
- T_Co : orezuri în terase_6.5, tentativă de intruzare_45.3, timp de decenare_49.10.

Lorsque les modèles neuronaux se servent d'un corpus d'entraînement limité/bruité et ne disposent pas d'un contexte suffisant, ils peuvent inventer des solutions plausibles qui respectent les règles morphologiques et phonétiques de la langue cible ou créer des approximations. D'autres fois, une segmentation incorrecte en *subwords* peut conduire à un réassemblage erroné qui produit à son tour, une construction inexistante (i.e. zarbată_47.6). En outre, un système de TAN multilingue peut mélanger des structures entre deux ou plusieurs langues, fait qui entraîne de mots hybrides (i.e tobacconiști_50.5). Parfois, dans le cas des traductions du français en roumain, l'absence de la diacritisation dans le fragment à traduire peut obliger la machine à deviner la lexie en question, générant ainsi une lexie arbitraire.

Les transferts incorrects au niveau des noms propres naissent en raison d'un algorithme de reconnaissance des noms propres qui n'est pas encore suffisamment performant. Dans le cas de deux noms propres au niveau LexS le modèle neuronal crée des approximations qui obéissent aux bases morphologiques et phonétiques du roumain (i.e. Andele_76.4, Oița_91.1). Ensuite, les cas où un nom propre est également un nom commun créent de l'ambiguïté, cela étant la raison pour laquelle le cerveau artificiel décide souvent de les traduire au lieu de les laisser inchangés (i.e LexS_Zefir_26.1). Nous pouvons déceler ce phénomène dans le cas des noms propres composés X-plane_18.4/18.6, Bulk Jupiter_96.6, Cake Factory_64.3, Emerald Star_96.5, Hunday Kite_12.3 qui ont reçu les suivantes traductions littérales : plan X, Jupiterul Bulk, fabricare a prăjiturilor, steaua de smarald, Hunday Zmeul.

La fausse interprétation qui vise les constructions en majuscules fait l'objet d'une étude en connexion avec les titres et les sous-titres étant donné qu'elles ont été détectées exclusivement dans le cadre de ces parties de texte. Par voie de conséquence, elles découlent du contexte condensé inhérent aux titres et sous-titres rédigés en capitales qui est parfois dénué d'accents diacritiques. La majorité des unités écrites en lettres majuscules sont des LexS, voire six G_LexS et cinq T_LexS. Quant aux déviations sémantiques qu'elles déclenchent, les suivantes peuvent être retenues :

- des lexies simples inventées : OBSTACUL, DRAFĂ, PROZETA, ALTE, RETIREA pour OBSTACLE_15.1, DRONE_15.2, PROTHÈSE_16.3, ETHER_50.2, RÉTICENCE_62.5 ;
- des lexies composées fictives : CUPURI DE BARBE pour CODE-BARRES_50.4 ;
- des reproductions du LI : BIRD pour OISEAU_15.3, ROUTE pour TRAJET_17.5 ;
- choix du sens spécialisé inadéquat du fait d'une polysémie du LI : nervură pour NERF_16.1 ;
- traduction littérale du LI : emisie zero pour ZÉRO ÉMISSION_17.3, CERERE pour JUSTIFICATIF_50.6.

Il convient de noter que la machine tente de convertir les lettres majuscules en minuscules dans le cas du terme 16.1 et du binôme spécialisé 17.3 dans le but d'améliorer la qualité de la traduction.

Les inexactitudes dans la traduction des sigles et des troncations apparaissent au niveau des T_Co. Le modèle neuronal est en butte à des difficultés dans le décodage des concepts que véhiculent les sigles et les formes tronquées. En fait, les néophrasèmes sigliques simples (i.e. EDP, MSRR, RNA, TER) ou hybrides (i.e. brins d'ARN, e-EDP, dalle OLED) de notre corpus sont des situations métonymiques de type FORM A – CONCEPT A POUR FORM B – CONCEPT A qui sont expliqués à fond dans le cadre de notre analyse.

Les constructions causatives et à verbe support sont fautivement interprétées parce qu'elles représentent des structures plus complexes qui présentent des difficultés et des particularités syntaxiques ainsi que sémantiques. D'ailleurs, les verbes « faire » et « avoir » qui entrent dans la composition de 15 colligations sur 20 posent des problèmes de contextualisation, car ils peuvent bénéficier de plusieurs traductions. C'est pourtant le « faire causatif » qui occasionne le plus d'erreurs, notamment 5 T_Collig et 3 G_Collig, car le roumain n'utilise pas un verbe spécifique tel que « faire » suivi de l'infinitif pour marquer la causation. En ce qui concerne les causes de la mauvaise traduction de cette construction causative, nous avons observé que la moitié des cas sont déterminés par une traduction littérale du LI tandis que l'autre moitié n'est pas du tout atteinte par le LI.

III. L'INFLUENCE DES TROPES CONCEPTUELS DANS LA PRODUCTION DES ERREURS

Grâce à notre analyse, nous avons pu remarquer que l'idiomaticité est présente de manière transversale à tout niveau linguistique. Elle repose sur des tropes conceptuels, soit métaphoriques, soit métonymiques, qui structurent la pensée et contribuent à la naturalité du langage. Les locuteurs natifs utilisent les tropes conceptuels intuitivement après les avoir assimilés à travers leur corporéité et leur expérience dans le monde réel alors que pour la MTN, les comprendre et les utiliser efficacement sous-tend un processus d'apprentissage dans un monde virtuel où il n'y a plus de corps connecté à une perception sensorielle. En outre, le sens métaphorique est enraciné dans un cadre culturel spécifique et c'est difficile pour le cerveau artificiel d'intégrer tous ces schémas cognitifs à cause de ses limites technologiques et d'une existence dans un monde fragmenté et médié.

Au niveau LexS et LexC, il faut premièrement vérifier si le contexte des lexies en question trahit une relation d'association entre un domaine concret et un domaine abstrait ou de contiguïté dans le même

domaine pour qu'on puisse détecter les sens figurés. Bien évidemment, notre corpus contient plusieurs lexies simples ayant un sens connotatif dans la langue de départ. Néanmoins, les erreurs en matière de figuralité prennent de multiples formes, parmi lesquelles les plus fréquentes sont :

- lexies simples employées métaphoriquement dans la LS, mais qui ont été traduites par un sens propre en LC sans interférence du LI, par exemple, engin_9.2 & 17,1 & 48,5 & 36,12/36,15 (*mașină* au lieu de *dispozitiv & vehicul & robot/aeronavă*), bridage_58.4 (*prindere* au lieu de *limită*), tomber_96.1 (*a cădea* au lieu de *a publica*) ;
- lexies simples rendues en LC par un sens figuré au lieu du sens propre sans interférence du LI, par exemple, carrefour_44.3/44.5/44.10 (*răscruce* au lieu de *intersecție*) ; embouteillage_54.1 (*blocaj* au lieu de *ambuteiaj*), acheminer_5.10 (*a transporta* au lieu de *a orienta*) ;
- lexies simples employées métaphoriquement, mais qui ont été reproduites en LC sans interférence du LI, par exemple, las_95.1 (las au lieu de *bătrân/învechit*) ;
- lexies simples employées métaphoriquement dans la LS, mais dont la traduction en LC a été altérée par le LI, par exemple, engin_36.6/36.17/36.18/72.26 (*ambarcațiune* au lieu de *aeronavă*), logé_9.4/24.3 (*adăpostit* au lieu de *amplasat*), surenchère_3.4 (*escaladare* au lieu de *concurs/competiție*), alléger_91.3 (*a lumina* au lieu de *a reduce sarcina/greutatea*), arriver_18.8/24.1 (*a ajunge* au lieu de *a apărea*) ;
- lexies simples employées métaphoriquement en LS qui ont été traduites de façon erronée par des constructions idiomatiques en LI, ce qui a influencé l'erreur finale en LC, par exemple rogner_78.6 (voir Le blend conceptuel_COUPER LES COINS EST OBTENIR UN RÉSULTAT DE MOINDRE QUALITÉ) ;
- lexies simples employées métaphoriquement en LS qui n'ont reçu de traduction ni en LI ni dans la LC, par exemple, ligne_86.7 (voir La métaphore conceptuelle_ LA LIGNE ÉCRITE EST UNE TRAJECTOIRE) ;

Au niveau Co, la métaphorisation est également présentée, bien que moins souvent que dans le cas des lexies simples ou des EIs. Dans le cas des collocations générales de notre corpus touchées par la métaphorisation, c'est d'habitude le collocatif qui permet l'établissement des correspondances entre des domaines différents de signification. Les fautes en lien avec la métaphorisation représentent environ 10 % du nombre total des collocations et elles touchent les collocations générales ainsi que spécialisées en parts presque égales. Nous pouvons citer en exemple les suivantes collocations générales :

- utilisation fluide _17.7 = La métaphore conceptuelle_LA SIMPLICITÉ EST LA FLUIDITIÉ ;
- calories dépensées_25.2 = La métaphore conceptuelle_ L'ARGENT EST DE L'ÉNERGIE CORPORELLE ;
- gros pic_45.2 = La métaphore conceptuelle_L'IMPORTANCE EST LA HAUTEUR ;
- gâteau au cœur coulant_64.7 = La métaphore conceptuelle_ LE CŒUR EST UN CENTRE ;

- franchir un obstacle_40.2 = La métaphore conceptuelle_ UN OBSTACLE/DÉFI EST UNE BARRIÈRE ;

Dans le cas des cooccurrences techniques, il n'est pas rare de constater que des termes- pivot qui abritent le concept central jouent un rôle clé dans la construction de la référence métaphorique (ex. dalle OLED, réservoir computing). De l'autre part, un effet de métaphorisation réalisé par le collocatif est perceptible dans le cas des collocations ci-après :

- flotte captive_ 1.6 ;
- essaim de minirobots_47.1 (voir L'amalgame L'ESSAIM DE MINIROBOTS EST UNE STRUCTURE QUI RÉPOND COLLECTIVEMENT).

Au niveau Loc, parmi les tropes conceptuels sur lesquels reposent les locutions dont la signification a été brisée, nous pouvons énumérer :

a) G_Loc verbales : les métaphores conceptuelles LE PROGRÈS EST UNE MARQUE PHYSIQUE LAISSÉE (creuser un sillon_5.11), LA CLARTÉ DE L'INFORMATION EST LA LUMIÈRE (laisser dans l'ombre_49.2), UN POINT CULMINANT EST UN COURONNEMENT (pour couronner le tout_95.13), UNE IDÉE EST UN OBJECT FIXÉ (enfoncer le clou_ 5.6), LE PROGRÈS EST LE DÉPASSEMENT D'UNE LIGNE DE DÉMARCATION PHYSIQUE (franchir un cap_23.1 & 76.1), LA VICTOIRE EST L'ARRIVÉE À UNE LIGNE DE DÉMARCATION PHYSIQUE ET LA TÊTE EST LA DOMINANCE (coiffer au poteau _49.12), LE DÉNONCIATEUR EST UN DOIGT (pointer du doigt_78.3) ;
b) G_Loc nominales : la métaphore conceptuelle UN ÉVÉNEMENT NOTABLE EST UN COUP D'ÉCLAT (coup d'éclat_49.5) et la métaphore conceptuelle FAIRE DU VÉLO EST UN SPORT NOBLE qui part de la base métonymique LE VÉLO EST LA REINE (la petite reine_49.5) ;
c) T_Loc verbales : L'OBJECTIF EST UN ENDROIT (viser le large_7.1), L'IMPORTANCE EST LA HAUTER (tenir le haut de l'affiche 28.2) ;
d) T_Loc nominales : UN ÉVÉNEMENT SOUDAIN ET DÉSÉQUILIBRANT EST UN COUP (coup de roulis_96.12) ;
e) T_Loc Adjectivales : LA CONSOMMATION RAPIDE DE RESSOURCES ÉNERGÉTIQUES EST UNE DÉVORATION DE NOURRITURE (être gourmand en energie_43.4 & 43,8/43,10).

De surcroît, en passant au crible les erreurs portant sur les constructions idiomatiques, nous avons observé comment le contexte leur ajoute parfois des microsens enracinés dans des chaînes métaphoriques ou *blends* conceptuels (découvrir le pot-aux-roses_54.5, faire figure d'Arlésienne_1.4, poids plume_26.2).

En somme, les tropes conceptuels s'insèrent à tous les niveaux sémantiques, quoique le discours spécifique à la diffusion scientifique soit essentiellement concerné par l'information technique exposée et moins par la manière littéraire dans laquelle elle est transmise. Le constat général qui s'impose est que lorsqu'il est confronté à des projections métaphoriques, le modèle de TA neuronale de Google livre

systématiquement des fautes de traduction au niveau LexS, LexC et Co et quasiment toujours des fautes au niveau des idiomes décodants, qui restent marginaux du point de vue de la fréquence.

IV. L'INFLUENCE DU LI DANS LA PRODUCTION DES ERREURS

D'après nos analyses, tout indique que les sorties impropres produites par la TAN sont accompagnées d'une prépondérance écrasante de la langue d'entraînement du réseau. En fait, le taux atteint par les anomalies de traduction qui sont provoquées, en totalité ou en partie, par le LI à chaque niveau sémantique est listé ci-après :

- ✓ LexS : ~75 % (i.e. 214 fois sur 288)
- ✓ LexC : ~ 64 % (i.e. 36 fois de 57)
- ✓ Co : ~ 60 % (i.e. 91 fois sur 152)
- ✓ Loc : ~ 68 % (i.e. 17 fois sur 25)

Au niveau LexS, les manières dont le LI est intervenu dans la déformation du sens de départ sont reprises dans la liste suivante qui va de la plus grande à la plus petite valeur :

1. Polysémie du LI : 103 err. (4A – 57 err., 2A - 28 err., 2B – 12 err., 1A – 6 err.) ;
2. Trad. littérale du LI : 71 err. (5C – 34 err., 2C – 18 err., 4C – 13 err., 1C – 4 err., 1E – 2err.) ;
3. Reproduction du LI : 19 err. (4B – 9 err., 5B – 6 err., 1B – 3 err., 1D – 1err.) ;
4. Interprétation du LI : 4 err. (5A) ;
5. Trad. erronée du LI : 2 err. (4D).

Au niveau LexC, les erreurs en connexion avec le LI sont attribuables aux causes suivantes, classées par ordre décroissant de fréquence :

1. Trad. littérale du LI : 25 err. (1C -10 err., 5C – 8 err., 4C – 6 err., 1E – 1err.) ;
2. Reproduction du LI : 5 err. (1B/4B – 2 err., 5B – 1 err.) ;
3. Polysémie du LI : 4 err. (1A/2A/2B/4A – 1 err.) ;
4. Trad. erronée du LI : 2 err. (4D).

Au niveau Co, la traduction littérale du LI est la cause principale des erreurs, suivie par la polysémie du LI et la reproduction du LI :

1. Trad. littérale du LI : 72 err. (16A/B/C) ;
2. Polysémie du LI : 11 err. (11A/B/C) ;
3. Reproduction du LI : 8 err. (13A/B/C).

Au niveau Loc, c'est toujours la traduction littérale qui est responsable de la majorité des erreurs :

1. Trad. littérale du LI : 16 err. (16A) ;
2. Polysémie du LI : 1 err. (11A).

En résumé, toutes les limitations de la TAN de Google que nous avons détaillées de I à IV proviennent d'une insuffisance de données, d'un contexte restreint ou isolé, des problèmes de découpage en *tokens* (des sous-unités), surinterprétation des relations entre les mots, ambiguïté sémantique, méconnaissance du domaine, valeur métaphorique des constructions, variations culturelles et idiomatiques des idiomes décodants.

À travers l'exploit des limites de la TAN de Google de 2019, nous espérons avoir pu proposer des pistes d'amélioration en ce qui concerne les mécanismes de gestion de la polysémie en contexte, la néonymie, les majuscules, les sigles et de la valeur métaphorique de toute construction. Compte tenu de nos résultats de recherche, il ne nous reste qu'à dire que l'hypothèse redoutée par les traducteurs humains autour de laquelle nous avons articulé notre thèse, à savoir que l'IA saura prendre le dessus sur eux, était une vision romanesque, digne de gros bonnets de la fiction en 2019. Notre analyse a pointé que la TA neuronale appliquée sur des textes semi-spécialisés n'est pas sans faille et qu'elle a encore un long chemin à parcourir pour qu'un jour, elle puisse devancer la cognition humaine sur ce terrain. Cette issue de notre étude exige cependant aux acteurs de la traduction de renforcer ou d'optimiser l'hybride courant traducteur – machine au lieu d'envisager les futurs changements technologiques avec appréhension. Mieux dit, toute TAO devrait se muer en TAIA (i.e., Traduction assistée par l'IA) ou au moins, faciliter la participation de la TAN dans le processus de traduction. En fait, nous avons déjà passé ce dernier cap parce que dans les 3-4 derniers ans, les logiciels de TAO, à l'instar de TRADOS, un des leaders du marché de logiciels de traduction essaye d'élever ce nouveau système au rang de norme courante. Dans le cadre de cette nouvelle formule, le traducteur humain remplit toujours le rôle de dépanneur des irrégularités linguistiques émanant cette fois de l'automatisation intelligente.

Le constat qui se profile pour le stade de la technologie de TAN de Google de 2019 est qu'on ne peut pas se passer du traducteur humain averti. Il est le seul qui peut retisser les liens de sens et style brisés par la TAN et nos textes de VS fourmillent d'exemples à cet égard. Quand même, la TAN peut donner de bons résultats lorsque le texte ne renferme pas de valeur hautement technique ou figurée. C'est un fait que nous avons pu également déceler dans le cas de quelques articles ou de paragraphes de notre corpus. Reste à suivre de près les avancées technologiques dans ce domaine de la TA neuronale afin de déterminer si ce genre de problèmes seront réglés pour de bon.

6. Prolongations possibles

Ce chapitre se penche sur les possibles prolongations de cette étude, en mettant en lumière les défis persistants dans les deux axes de recherche principaux qui sont au cœur de notre thèse : la TAN et la SC.

Comme nous avons vu, notre étude a exploité pleinement le résultat qualitatif de la traduction automatisée neuronale de Google sur le plan sémantique lorsqu'elle réjouissait du mécanisme introduit en 2017. Cependant, depuis ce moment-là, Google continue à faire des recherches, à améliorer et à enregistrer des avancées en ce qui concerne ses réseaux de neurones et ses modèles d'entraînement, la gestion des langues et ainsi de suite. Pour le constater, il suffit de consulter ses blogs et publications officielles, comme nous l'avons fait pour notre thèse ces dernières années.

Par la suite, nous pourrions faire une analyse comparative entre le modèle de TA neuronale de Google de 2018/2019 et son modèle le plus récent à partir de ce même corpus et en mettant à l'épreuve les mêmes instruments d'analyse sémantique. Bien évidemment, nous considérons que pour pouvoir attribuer un « pourcentage de conformité sémantique » à une traduction transculturelle, tous les plans que nous avons soumis à l'analyse devraient être réexaminés. Cependant, le traitement de la polysémie en contexte par la TAN ou seulement le traitement des tropes conceptuels peut être envisagé, car ces théories cognitives, elles aussi, sont assujetties constamment à des changements, à des modifications et à des progrès. Dans les deux domaines de recherche qui s'entremêlent dans notre thèse, même s'il y a des exploits réguliers, il reste encore beaucoup à perfectionner. Les chercheurs de la TAN s'efforcent à dépasser les limites de la traduction en ce qui concerne le sens grammatical, le vocabulaire terminologique existant ou émergent ou encore la dépendance à la langue d'entraînement tandis que les linguistes d'orientation cognitive essaient de rendre leurs théories aussi fiables que possible sur le plan scientifique et de se mettre d'accord sur les méthodes d'analyse les plus valides. Dans la TAN ainsi que dans le cadre de la SC, des métriques toujours plus complexes sont mises en place afin de forger une architecture/un cadre d'analyse optimal(e).

Ensuite, une analyse comparative entre GTNM et le modèle de traduction neuronale de DeepL sur le même corpus ou un autre corpus spécialisé pourrait être réalisée. Sur leur page en ligne[4], DeepL présente fièrement le témoignage de TechCrunch, USA qui prétend que l'entreprise quoique petite par rapport au géant Google a surpassé le dernier et placé la barre plus haut en matière de traduction. Cette comparaison est d'autant plus valide scientifiquement que les périodes de lancement de DeepL et de GTNM coïncident. La compagnie allemande vante les mérites de sa technologie en matière de réseaux neuronaux et garantit de continuer à innover et rendre la communication plus rapide et efficace. De plus, ce sera une étude qui se plie parfaitement sur un corpus à langage technique parce que DeepL

[4] https://www.deepl.com/de/whydeepl

Translate déclare avoir introduit en 2020-2021 de nouveaux modèles capables de restituer plus fidèlement le sens des phrases traduites et même de maîtriser avec succès le jargon propre à certains secteurs d'activité.

Une autre prolongation pourrait être l'analyse du niveau de fidélité sémantique assurée par la machine neuronale de Google seulement dans le cas des titres et sous-titres d'articles. Nous avons remarqué que dans certains cas, lorsque le titre ou le sous-titre est marqué en haut de casse, la traduction est plus erronée que dans le cas des titres ou sous-titres écrits en minuscules. Après ce constat, j'ai injecté dans le traducteur automatique les titres et les sous-titres comportant des lettres de grand format sous une forme en minuscules afin de voir s'il y a des changements au niveau de la qualité. Passer d'un texte de majuscule en minuscule a amélioré l'exactitude du transfert sémantique opéré par GTNM. Par déduction, un facteur qui a un énorme poids dans le contrôle de la conformité de la traduction sont les « accents français ». Ils servent à la réalisation d'une discrimination lexicale et à l'enlèvement de l'ambiguïté lexicale. Pour les signifiants « ou » vs » où », par exemple, nous pouvons avoir la certitude qu'ils seront dûment traités par une machine de TAN seulement s'ils se retrouvent dans le titre/sous-titre dans leur forme graphique correcte.

Quant à notre corpus, il y a plein d'exemples où des items lexicaux qui ont été édités en capitales dans un certain titre/sous-titre ne sont pas bien traduits. Nous avons procédé à l'extraction de toutes ces erreurs et nous avons remarqué que certaines d'entre elles ont été :

- traduites par la reprise de la traduction en LI (i.e. 15,3, 17,5) ;
-traduites par des lexies non existantes en LC (i.e. 15,1, 15,2, 16,3, 50,2, 62,5, 50,4).

De plus, sur un total de 13 erreurs de ce type, 11 se trouvent au niveau des lexies simples (6 x G_LexS et 5 x T_LexS). Les deux autres erreurs font partie du langage technique. Il en découle qu'il y a une quasi-égalité d'erreurs en ce qui concerne le rapport entre la langue générale et la langue spécialisée.

Conjointement, un autre constat a été celui que les titres et les sous-titres des articles démasquent une autre faiblesse de la TA neuronale, à savoir la traduction des mots ayant un contexte limité et une référence ambiguë. Les mots dans les titres introduisent des nuances parfois métaphoriques afin d'attirer l'attention du lecteur et ne fournissent pas une référence assez précise. C'est le corps de l'article qui clarifiera le sens de la référence initiale et aidera le lecteur à désambiguïser. C'est à cause de ces facteurs que la fréquence des erreurs qui figurent dans les titres et sous-titres est assez élevée. En fait, plus d'un tiers de la totalité des titres de notre corpus contient au moins une erreur.

Il s'ensuit qu'une autre analyse potentielle serait celle du niveau de fidélité sémantique assurée par la machine neuronale de Google dans le cas des articles intégraux comparés à leurs variantes fragmentées ou scindées en paragraphes. Cela est une autre possibilité de prolongation d'étude qui s'est dessinée à la suite du constat précédent. Ce serait intéressant de mettre en exergue les changements de sens issus de la traduction d'un texte dépecé par opposition à un texte intégral ou au

mois, découvrir à quel point la traduction d'un texte morcelé s'avère plus déficitaire. Cette idée peut soit servir de complément de recherche soit de thème pour une recherche autonome.

Une étude visant à tester la traduction des abréviations (i.e. sigles, acronymes, etc.) et des noms propres figurant dans des articles serait très intéressante en raison des défis qu'ils posent. Qu'il s'agisse de noms des lieux, des personnes, des dénominations de certains appareils, instruments, technologies ainsi que d'autres, les abréviations et les noms propres suscitent bien des problèmes pour un traducteur neuronal. Comme notre analyse l'a fait valoir, le sens de ces unités de langue doit être préservé parfois par une adaptation subtile ou le plus souvent, elles doivent être laissées inchangées au niveau de la forme pour ainsi réduire les malentendus culturels et favoriser leur reconnaissance correcte. Notre analyse a relevé que tous les noms propres du FR qui ont subi une traduction incorrecte ont la même forme en anglais. Une recherche focalisée sur la capacité actuelle de 2025 du traducteur neuronal de Google comparativement à sa capacité initiale de 2018/2019 indiquerait sans équivoque le niveau d'amélioration de la qualité de traduction de la machine. De plus, elle pourrait même contribuer à l'évolution des mécanismes neuronaux de détection des noms propres pour ainsi dépasser leur stade actuel décevant. Tel qu'il ressort de notre thèse, le traducteur neuronal de Google de 2018/2019 a tendance à faire la correspondance des noms propres avec des noms communs et livre par la suite, des traductions littérales ou des adaptations incorrectes.

En outre, on pourrait examiner le niveau de fidélité sémantique assurée par la machine neuronale de Google dans le cas des textes de VS qui visent à traiter la question de l'innovation dans un secteur précis, comme la médecine par exemple. Il est important de voir comment le langage spécialisé est abordé dans ce cas. Dans notre corpus, constitué de multiples domaines de spécialisation où l'innovation est présente, nous avons remarqué qu'un terme acquiert plusieurs traductions en fonction du domaine spécialisé dont il fait partie. Un exemple concret est le terme « enceinte » qui est présent dans l'article N°2 portant sur un assistant vocal, dans l'article N°5 qui traite de la centrale nucléaire, et dans l'article N° 45 qui parle d'un stéthoscope à Web. Dans l'article N°2, cette lexie doit avoir la traduction *boxă* (fr. : haut-parleur) parce qu'elle fait partie du domaine acoustique. Pourtant, GTNM opte dans le cas de 2,1 pour la traduction du terme par *carcasă* (fr. : boîtier). Par contre, dans le cas de 5,6, la lexie est traduite par *incintă* (fr. : espace clos, fermé à l'intérieur d'une construction) qui restitue bien le sens contextuel.

De plus, le fichier Excel dans lequel nous avons introduit les erreurs provenant de la TAN de Google peut servir à la construction d'un glossaire terminologique FR → RO dans le secteur de nouvelles technologies qui même après la fin de notre étude pourra être actualisé régulièrement. Nous sommes convaincue que notre mission en tant que chercheur ayant choisi ce genre de thèse est aussi celle de proposer la meilleure solution terminologique pour les concepts nouveaux ou pour ceux qui ont été déjà adoptés en roumain sous la forme d'un emprunt partiel ou total de l'anglais, une solution rapide à usage répandu. Cela n'est pas surprenant parce que la majorité des innovations ont été mises en œuvre dans des pays anglophones, ce qui fait que l'anglais domine le monde de la communication par rapport au

progrès scientifiques. Le fait que l'anglais est devenu la langue principale de dissémination scientifique de nouvelles technologies est un aspect indéniable. Pourtant, chaque nation devrait tenir à l'écart cette tendance de supprimer la signature nationale apposée sur un concept parce qu'elle mène à une perte d'identité culturelle et cela n'est qu'un autre effet de la globalisation. Malheureusement, l'influence de la culture anglophone ne se fait pas ressentir uniquement dans le cas des appellations des concepts récents.

L'Académie roumaine, un organisme qui peut influencer ce phénomène grâce à sa mission de protection de l'identité langagière roumaine n'a que peu de force à son encontre. La diffusion des emprunts dans la pratique des professionnels, qui préfèrent l'uniformité et utilisent les termes sous leur forme originale pour éviter toute confusion, est difficile à arrêter. C'est pourquoi nous pensons que l'Académie roumaine devrait se tenir au courant des dernières innovations dans chaque secteur spécialisé pour pouvoir publier en permanence des propositions linguistiques et les transmettre à la population, mais surtout, aux professionnels du secteur d'activité concerné. Il est indispensable de le faire, car le monde technologique progresse à une vitesse hallucinante chaque année. Parmi des stratégies qui vont dans ce sens-là, nous pourrions mentionner la mise en place des programmes de sensibilisation à l'importance de la préservation de la langue roumaine à travers l'emploi des termes roumains. Ensuite, la réalisation des partenariats avec les médias dans le but de promouvoir cet emploi pourrait apporter de bons résultats. Quant aux mesures un peu plus drastiques, on peut concevoir des lois qui réglementent la terminologie utilisée dans les documents officiels, la publicité et ainsi de suite. Des subventions pourraient inciter les entreprises du domaine scientifique à utiliser une terminologie roumaine.

En ce qui me concerne, comme je suis impliquée dans la recherche linguistique, je me sens tenue de m'engager dans le combat contre les emprunts anglais. Par la suite, j'ai créé un glossaire-échantillon, intitulé « Propositions linguistiques de substitution à la néo-terminologie technique anglaise », dans lequel je propose des équivalents 100 % roumains pour les termes et phraséotermes modernes de notre corpus qui ont été mal traduits et qui circulent en roumain sous leur forme anglaise originale, entière ou partielle (voir Annexe E).

Pour conclure, selon les traductions fournies en 2018/2019, il semblerait que l'intelligence artificielle ne fera pas disparaître le métier de traducteur, mais qu'elle le transformera à nouveau et équipera le traducteur de nouveaux acquis technologiques. Il deviendra de la sorte un ingénieur de la communication bilingue ou multilingue. En poursuivant cette voie de développement, la traduction automatique continuera à tirer parti, de plus en plus intelligemment, des glossaires terminologiques, des corpus de référence et surtout, de la contribution des traducteurs humains. *In fine*, nous restons des optimistes qui ne se laissent pas captiver par le côté sombre des scénarios et qui considèrent que la mise en scène de la TAN est un tribut bien mérité rendu au traducteur qui ne périra pas, mais qui changera effectivement de rôle.

Bibliographie

Dufay, B. (2005). *Apprendre à expliquer : l'art de vulgariser*. Eyrolles ;

Halliday, M. A. K., Hasan, R. (1976). Cohesion in English. Longmans ;

Jakobson, R. (1959). On linguistic aspects of translation. R. A. Brower (Éd.), *On translation.* Harvard University Press ;

Müller, B. (1985). *Le français d'aujourd'hui*. Éditions Klincksieck ;

Dubreil, E. (2008). Collocations : définitions et problématiques. *Texto !, Vol. XIII* (1). http://www.revue-texto.net/docannexe/file/126/dubreil_collocations.pdf ;

Meadows, J. (1986). Histoire succincte de la vulgarisation scientifique. *Impact : science et société, n° 144*, pp. 395-401. https://unesdoc.unesco.org/ark:/48223/pf0000071156_fre ;

Moirand, S., Reboul-Touré, S. Pordeus Ribeiro, M. (2016). La vulgarisation scientifique au croisement de nouvelles sphères d'activité langagière. *Bakhtiniana, Vol. 11* (2), pp.137-161. http://dx.doi.org/10.1590/2176-45732387 ;

Williams, G. (2003). Les collocations et l'école contextualiste britannique. *Travaux et recherches en linguistique appliquée, série E,* n°1, pp. 33-45 ;

Wang, W., Zhang, Z., Du, Y., Chen, B., Xie, J., Luo, W. (2021). Rethinking Zero-shot Neural Machine Translation:From a Perspective of Latent Variables. *Findings of the Association for Computational Linguistics : EMNLP 2021*, pp. 4321–4327. doi : 10.18653/v1/2021.findings-emnlp.366 ;