Traduction automatique statistique vs. neuronale : Comparaison de MTH et DeepL à La Poste Suisse

Volkart, Lise

LISE VOLKART

TRADUCTION AUTOMATIQUE STATISTIQUE VS. NEURONALE

COMPARAISON DE MTH ET DEEPL A LA POSTE SUISSE

Directrice : Pierrette Bouillon

Jurée : Lucía Morado Vásquez

Mémoire présenté à la **Faculté de traduction et d'interprétation** (Département de traitement informatique multilingue) pour l'obtention de la **Maîtrise universitaire en traduction, mention Technologie de la traduction**

Université de Genève

Juin 2018

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la *Directive en matière de plagiat des étudiant-e-s, le Règlement d'études des Maîtrises universitaires en traduction et du Certificat complémentaire en traduction de la Faculté de traduction et d'interprétation ainsi que l'Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction*).

J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome.

Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis consciente que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Lise Volkart

Genève, le 23 mai 2018

REMERCIEMENTS

« L'amour, la bienveillance et l'encouragement sont les leviers de l'âme humaine. »
Maria Montessori

Je tiens à remercier les personnes qui m'ont accompagnée tout au long de la rédaction de ce mémoire.

Tout d'abord, merci à ma directrice de mémoire, Pierrette Bouillon, de m'avoir donné l'opportunité de participer à ce projet du Département TIM et de La Poste Suisse et de m'avoir guidée dans la rédaction de mon mémoire. Merci à elle pour sa patience, sa disponibilité et ses précieux conseils.

Merci à Sabrina Girletti qui a partagé avec moi les hauts et les bas de ce long projet et m'a apporté une aide inestimable. Je te souhaite le meilleur pour la suite de ce projet et pour la rédaction de ta thèse.

Merci à ma jurée, Lucía Morado Vásquez pour sa disponibilité et à Paula Estrella et Jonathan Mutal pour leur aide.

Merci à La Poste Suisse et tout particulièrement à Beatrice Bircher et Martina Bellodi pour leur belle initiative et leur disponibilité.

Merci à Valentine, Evelyne, Lise, Cécile et Marie qui ont consacré un temps précieux aux évaluations humaines.

Enfin, un grand merci à Yannick pour son immense soutien tout au long de mes études et de mon mémoire. Tes encouragements et ta bienveillance m'insufflent au quotidien la confiance dont je manque parfois.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

TA : traduction automatique

PE : post-édition

MTH : Microsoft Translator Hub

TAS : traduction automatique statistique

TAN : traduction automatique neuronale

TEAHQ : traduction entièrement automatique de haute qualité

TAAH : traduction automatique assistée par un humain

THAO : traduction humaine assistée par ordinateur

TH : traduction humaine

TAO : traduction assistée par ordinateur

LS : langue source

LC : langue cible

MT : mémoire de traduction

1. INTRODUCTION

Dans un monde de plus en plus globalisé et de plus en plus connecté, chacun de nous est inévitablement confronté aux barrières linguistiques. Dans ce contexte, la traduction automatique apparait comme un moyen de faire tomber ces barrières et s'immisce de plus en plus dans notre quotidien. Mais la traduction automatique n'est pas un domaine nouveau et les technologies dont nous disposons aujourd'hui sont le fruit d'un travail de plusieurs décennies.

Bien qu'ayant fait des progrès considérables, la traduction automatique est aujourd'hui encore mal perçue du grand public, qui critique allègrement sa qualité. Les traducteurs professionnels en ont aussi généralement une perception négative, qui est alimentée à la fois par la conviction que ses performances sont médiocres, mais aussi par la peur que cette technologie parvienne un jour à les remplacer. Cette méfiance des professionnels envers la traduction automatique semble aujourd'hui s'accentuer encore avec l'intégration progressive de l'intelligence artificielle dans les systèmes de traduction automatique. Mais lorsqu'elle s'invite dans le monde de la traduction professionnelle, la traduction automatique est le plus souvent un outil au service du traducteur plutôt qu'un outil destiné à le remplacer. Il est fort probable que la traduction automatique bouleverse profondément le métier du traducteur dans un futur très proche, mais pour le moment, rien n'indique qu'elle entrainera sa disparition.

1.1. Contexte

Ce mémoire a été réalisé dans le cadre d'une collaboration entre La Poste Suisse et le Département de Traitement informatique multilingue (TIM) de la Faculté de Traduction et d'Interprétation de l'Université de Genève. Désireux d'intégrer la traduction automatique (TA) dans son environnement de travail, le Service Linguistique de La Poste a fait appel au Département TIM pour réaliser différents tests en vue de l'intégration de la traduction automatique. Différentes personnes au sein du Département ont été impliquées dans ce projet.

Ce projet visait à effectuer des tests pour les scénarios suivants :

1. ***Gisting*** : intégration de la TA dans l'Intranet pour permettre aux collaborateurs de traduire automatiquement des contenus dans le but de les comprendre. Pour ce scénario, la traduction doit être d'une qualité suffisante pour permettre la compréhension.
2. ***Pre-translation in the order screen*** : intégration de la TA comme outil à disposition du personnel qui utilise les services de traduction. Avant d'envoyer un mandat au Service Linguistique, le donneur d'ordre a la possibilité de faire une pré-traduction automatique, s'il juge que la qualité de cette traduction est suffisante, il peut l'utiliser telle qu'elle, sinon, il peut demander une post-édition au Service Linguistique. Pour ce scénario, la traduction doit être de bonne qualité.
3. ***MT suggestions in editing tool*** : les suggestions de TA apparaissent dans l'outil de TAO des traducteurs lorsqu'il n'y a pas de correspondances avec la mémoire de traduction et le traducteur peut choisir d'utiliser ou non les suggestions de la TA pour les post-éditer. Pour ce scénario aussi, la traduction doit être de bonne qualité.
4. ***Full MT with PE as default*** : les textes urgents sont traduits de manière automatique et post-édités par les traducteurs. Là encore, la TA doit être de bonne qualité.

Dans le cadre de ce mémoire, nous nous sommes concentré sur les scénarios dans lesquels la TA s'adresse aux traducteurs (3 et 4).

Le projet prévoyait initialement l'entrainement de deux systèmes de TA statistique (*Microsoft Translator Hub* et *Moses*) pour différentes paires de langues (DE vers FR et IT) et leur comparaison afin de sélectionner le meilleur des deux pour poursuivre les tests avant de faire des propositions pour les différents scénarios. Une éventuelle comparaison des systèmes de TA statistique avec un système neuronal avait aussi déjà été envisagée. Le projet prévoyait ensuite un test de productivité avec le meilleur des systèmes. Notre rôle dans le cadre de ce projet consistait à entrainer et à évaluer *Microsoft Translator Hub* (MTH) pour la paire de langues DE>FR.

1.2. Objectif de ce mémoire

Le projet dans son ensemble a pour objectif d'évaluer la faisabilité de l'intégration de la TA à La Poste Suisse pour différents scénarios ainsi que d'identifier les avantages que cette intégration pourrait avoir pour La Poste et pour le Service Linguistique. En ce qui concerne ce

mémoire, l'objectif initial était d'évaluer les performances d'un système Microsoft Translator Hub dans un contexte d'entreprise, à savoir celui de La Poste Suisse et de voir dans quelle mesure l'intégration d'un tel système pourrait être bénéfique pour les traducteurs. Cet objectif a légèrement évolué au fur et à mesure de l'avancement du projet. Au cours de ces deux dernières années, le domaine de la TA a été marqué par l'émergence d'une nouvelle technologie, à savoir, la traduction automatique neuronale (TAN) ou « neural machine translation » (NMT). Certains fournisseurs de service de traduction automatique neuronale promettent une qualité de traduction surpassant celle de la plupart des autres systèmes[1]. La TAN éveille donc l'enthousiasme des chercheurs et du grand public, mais certaines études (Castilho et al., 2017) ont montré que pour des paires de langues et des domaines donnés, la TAS peut obtenir de meilleurs résultats que la TAN. Face à l'engouement que suscite la TAN, il nous a paru essentiel de l'intégrer dans ce travail de recherche. Nous avons donc décidé de comparer les performances du système MTH que nous avions entrainé pour La Poste avec celles de *DeepL Pro*, un système de traduction neuronale non spécialisé destiné aux traducteurs professionnels. L'idéal aurait certainement été de comparer notre système MTH avec un système neuronal entrainé avec les mêmes données, mais nous ne disposons pas des ressources informatiques suffisantes pour entrainer un tel système. En effet, la mise en place et l'entrainement d'un système neuronal requièrent d'importantes ressources informatiques et les utilisateurs potentiels de TA, tout comme nous, ne possèdent pas forcément ces ressources. Nous avons donc décidé de comparer la TA statistique standard avec la TAN telle qu'elle est accessible maintenant si l'on ne possède pas de grandes ressources informatiques. *Nous allons donc chercher à identifier si les systèmes statistiques spécialisés sont encore en mesure de rivaliser avec les systèmes neuronaux généralistes dans le contexte qui est le nôtre.*

1.3. Démarche

Cette section présente la démarche que nous avons suivie pour la réalisation de ce travail.

Nous avons tout d'abord entrainé des systèmes de TA statistique sur *Microsoft Translator Hub* avec des données de La Poste. Avant de réaliser ces entrainements, nous avons dû anonymiser les mémoires de traduction de La Poste, car certaines d'entre elles contenaient des données

[1] https://www.deepl.com/press.html (2018b)

confidentielles que nous ne pouvions pas mettre en ligne sur MTH. Nous avons effectué des entrainements par domaine, puis des entrainements avec les données de tous les domaines ainsi que des entrainements avec et sans terminologie. Nous avons créé des corpus de test à partir de textes de La Poste que nous avons traduits à l'aide des systèmes que nous avions entrainés. Nous avons utilisé une métrique automatique, le score BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), pour évaluer nos différents systèmes et sélectionner le meilleur d'entre eux. Nous avons ensuite créé un nouveau corpus de test contenant 250 phrases issues, elles aussi, de textes de La Poste. Nous avons traduit ce corpus avec notre meilleur système MTH ainsi qu'avec *DeepL Pro* dans le but d'effectuer une comparaison entre ces deux systèmes. Pour comparer ces deux systèmes, nous avons tout d'abord procédé à une évaluation automatique en utilisant la métrique BLEU, avant d'effectuer une évaluation humaine portant sur l'effort de post-édition. En adoptant cette démarche, nous avons tenté de répondre à la question de recherche suivante :

Les systèmes statistiques spécialisés sont-ils en mesure de rivaliser avec les systèmes neuronaux généralistes lorsque la traduction automatique est utilisée par le traducteur professionnel comme outil d'aide à la traduction ?

Au cours de notre travail, nous avons constaté qu'il existait peu d'études portant sur la corrélation entre les évaluations humaines et les évaluations automatiques (notamment avec la métrique BLEU) pour les systèmes de traduction neuronale. Néanmoins, deux études successives conduites par Shterionov et al. (2017; 2018) ont montré que le score BLEU ne reflétait pas forcément la qualité de la traduction neuronale et qu'il avait tendance à la sous-estimer. Nous avons donc décidé de nous intéresser à la question suivante en marge de notre question de recherche principale :

Le score BLEU est-il une métrique fiable pour l'évaluation des systèmes de traduction neuronale ?

1.4. Plan

Après avoir présenté le contexte dans lequel s'inscrit ce mémoire et détaillé nos objectifs et notre démarche, nous allons, dans le chapitre 2, présenter la TA et donner quelques définitions importantes. Nous évoquerons aussi brièvement l'histoire de la traduction automatique et

donnerons un aperçu de l'état actuel de la TA. Nous détaillerons les objectifs de la TA et présenterons la post-édition. Enfin, nous évoquerons les différentes approches en matière de traduction automatique et les différents types de systèmes.

Dans le chapitre 3, nous nous pencherons sur les modes d'évaluation de la TA et nous donnerons un aperçu des différentes méthodes d'évaluation humaine et automatique.

Dans le chapitre 4, nous présenterons plus en détail les systèmes de traduction automatique utilisés pour ce travail. Nous détaillerons le fonctionnement des systèmes statistiques et présenterons l'outil *Microsoft Translator Hub*. Nous décrirons aussi le fonctionnement des systèmes de traduction automatique neuronale et nous présenterons *DeepL*.

Dans le chapitre 5, nous détaillerons les méthodes mises en place pour l'entrainement des systèmes MTH et pour la sélection du meilleur système et du domaine sur lequel portera notre évaluation comparative.

Le chapitre 6 portera sur l'évaluation comparative de *DeepL* et MTH. Nous présenterons notre évaluation automatique ainsi que nos évaluations humaines.

Dans le chapitre 7, nous nous intéresserons à la corrélation entre nos évaluations humaines et notre évaluation automatique.

Le chapitre 8 nous permettra de présenter les conclusions de notre étude. Nous évoquerons aussi les limites de cette dernière et proposerons des pistes pour de futures recherches. Pour terminer, nous formulerons les recommandations qui s'adressent à l'entreprise avec laquelle nous avons collaboré dans le cadre de ce projet.

2. QU'EST-CE QUE LA TA ?

Ce chapitre vise à présenter le domaine de la traduction automatique dans ses grandes lignes et à donner au lecteur quelques concepts essentiels à la compréhension de ce travail. Nous définirons tout d'abord la notion de traduction automatique (section 2.1) et donnerons un aperçu de l'évolution de ce domaine à travers un bref historique et une présentation de l'état actuel (section 2.2 et 2.3). Nous parlerons ensuite des objectifs de la TA (Section 2.4) et définirons la post-édition (section 2.5). Nous terminerons ce chapitre en mentionnant les différentes approches et les différents systèmes de TA (section 2.6).

2.1. Définition

En 1993, Bouillon donne la définition suivante de la traduction automatique :

> *La traduction automatique (TA) se définit comme l'application de l'informatique à la traduction des textes d'une langue naturelle de départ (ou langue source LS) dans une langue d'arrivée (ou langue cible LC).* (Bouillon, 1993 : 15)

Cette définition suggère que la traduction automatique est un domaine assez large qui regroupe différents intermédiaires. La classification des types de TA donnée par Hutchins et Somers en 1992 nous aide à mieux cerner le domaine de la traduction automatique et à comprendre où il se situe parmi les différents types de traduction. Cette classification se présente sous la forme d'un continuum présenté ci-dessous (Figure 1) :

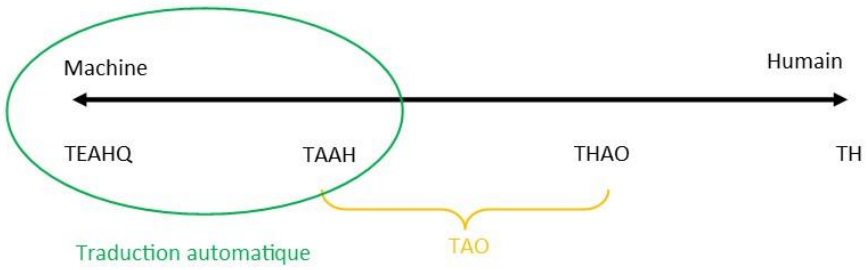


Figure 1 - *Adaptation de la classification des types de traduction de Hutchins et Somers (1992 : 148) « Human and machine translation »*

Ce continuum classe les différents types de traduction en fonction de l'importance plus ou moins grande que prennent la machine et l'humain dans le processus.

La **TEAHQ** (Traduction entièrement automatique de haute qualité) est une traduction entièrement réalisée par une machine et dont le résultat est de très bonne qualité. L'humain n'intervient pas du tout dans ce processus. La **TAAH** (traduction automatique assistée par un humain) désigne une traduction automatique dans laquelle intervient un humain, mais cette personne ne réalise pas la traduction à proprement parler (il peut intervenir avant, pendant ou après le processus de traduction). La **THAO** (traduction humaine assistée par ordinateur) désigne une traduction réalisée par un humain avec l'aide d'outils informatiques (par exemple une mémoire de traduction). Enfin, la **TH** (traduction humaine) désigne la traduction entièrement réalisée par un humain, sans aucune aide de la machine (Hutchins et Somers, 1992). Sur ce continuum, la traduction automatique inclut tout ce qui se trouve entre la TEAHQ et la TAAH. La TAAH et la THAO sont souvent regroupées sous l'appellation **TAO** (traduction assistée par ordinateur) (Hutchins et Somers, 1992). Dans le cas de la THAO, c'est le traducteur humain, qui réalise la traduction à proprement parler, c'est-à-dire le transfert entre les deux langues, en utilisant des outils qui lui apportent une aide. Dans le cas de la TAAH, la traduction est effectuée par la machine et l'humain intervient pour aider la machine avant, après ou pendant la tâche de traduction (Hutchins et Somers, 1992).

En résumé, nous pouvons dire que la traduction automatique désigne une traduction dans laquelle la machine réalise la transposition entre les deux langues. Cette machine peut-être ou non assistée par un humain. Les outils d'aide à la traduction tels que les mémoires de traduction ou les bases terminologiques n'appartiennent donc pas au domaine de la TA, mais à celui de la

THAO (Bouillon et Clas, 1993). Jusqu'à il y a peu, la TEAHQ semblait être un objectif ambitieux que l'on ne parvenait à atteindre qu'en se limitant à des domaines très restreints dans lesquels on trouve des **sous-langages** bien définis (les bulletins météorologiques par exemple) (Bouillon et Clas, 1993). Le traducteur humain a donc bien souvent un rôle à jouer dans la TA. On parle de **pré-édition** lorsque le traducteur intervient avant la traduction et de **post-édition** (PE) lorsqu'il intervient après. L'intervention du traducteur pendant le processus de traduction est dite **interactive**. Les systèmes qui permettent une intervention interactive du traducteur humain sont peu courants aujourd'hui (Hutchins et Somers, 1992) et le plus souvent l'intervention humaine en la TA se fait via la pré-édition et la post-édition.

L'objectif de la pré-édition, ou normalisation, est d'agir sur le texte source avant de le soumettre à la TA afin de tenter d'écarter ce qui pourrait poser problème à la machine (Hutchins et Somers, 1992). La pré-édition peut consister à ajouter des marqueurs dans le texte afin d'aider la machine à résoudre les ambigüités (Hutchins, 2005). Elle peut aussi inclure la reformulation (ou l'écriture) du texte dans un **langage contrôlé**, c'est-à-dire un langage spécialement conçu pour être adapté aux capacités du système. Le langage contrôlé peut, par exemple, se limiter à certaines structures syntaxiques ou à un certain vocabulaire. Le post-éditeur, quant à lui, est chargé de réviser et de corriger la traduction produite par la machine afin d'en améliorer la qualité (Hutchins et Somers, 1992). Nous présenterons la post-édition plus en détail dans la section 2.5.

2.2. Bref historique

L'histoire de la TA a été marquée par des périodes de grand enthousiasme et de déception (Koehn, 2010). Les premières tentatives en la matière ont été réalisées peu après l'apparition des ordinateurs numériques (Koehn, 2010). Ces derniers ont vu le jour aux Etats-Unis et au Royaume-Uni au cours de la Première Guerre mondiale. Ces machines ont été développées notamment pour calculer des tables de tir et décrypter les communications ennemies. Mais les scientifiques ont rapidement compris que ces machines avaient un grand potentiel et qu'elles pourraient servir à bien d'autres applications. Il semblerait que l'idée d'utiliser la puissance de ces ordinateurs pour traduire ait été formulée pour la première fois par le mathématicien Warren Weaver en 1947 (Hutchins, 1986).

En 1949, Weaver publie un mémorandum qui est aujourd'hui considéré comme le point de départ de la traduction automatique (Quah, 2006). Dans son mémorandum, il émet l'idée d'utiliser les ordinateurs modernes pour traduire. Il propose d'utiliser les méthodes de cryptographie en s'appuyant sur l'idée qu'un texte écrit dans une langue étrangère peut être vu comme un texte qui a été codé dans une autre langue (Weaver, 1949). Ainsi, de la même manière que les ordinateurs servaient à déchiffrer le langage codé des ennemis durant la guerre, ils pourraient servir à « décoder » une langue étrangère. A l'époque, ce mémorandum est considéré par beaucoup comme simpliste et naïf, mais il suscite suffisamment d'intérêt au sein de la communauté scientifique pour que des projets de recherche voient le jour (Quah, 2006). Suite à ce mémorandum, des fonds importants sont consacrés à la recherche dans le domaine de la traduction automatique et de nombreux groupes de recherche sont créés, notamment aux Etats-Unis, en Europe et en Union soviétique (Quah, 2006). Cet intérêt pour la TA peut s'expliquer en partie par le besoin d'avoir des traductions fiables des publications scientifiques dans des délais assez courts et par le manque d'experts bilingues pour les réaliser, de même que par la nécessité d'obtenir des traductions rapides des publications (journaux, outils de propagande, etc) des pays ennemis dans le contexte de la guerre froide (Bar-Hillel, 1951).

En 1951, le MIT confie à Yeoshua Bar-Hillel la mission d'étudier les possibilités en matière de traduction automatique afin d'orienter les recherches à venir (Hutchins, 1986). A la fin de cette même année, Bar-Hillel publie ses premières conclusions. Selon lui, la traduction entièrement automatique ne peut, pour le moment, qu'être réalisée qu'aux dépens de la qualité. Il estime que la machine n'étant pas capable d'éliminer les ambigüités sémantiques, elle ne peut pas produire une traduction entièrement automatique de haute qualité. Il explique cependant que cette conclusion ne devrait pas décourager les chercheurs du domaine, mais les inviter à poursuivre des objectifs plus modestes. Pour lui, il faut abandonner la TA pure pour se tourner vers une TA mixte dans laquelle l'humain intervient dans le processus (Bar-Hillel, 1951).

En juin 1952, la première conférence sur la traduction automatique se tient au MIT. Cet évènement rassemble 18 participants issus de différents domaines de recherches (Hutchins, 1986). A la suite de la conférence de 1952, un groupe de recherche en traduction automatique est formé par Leon Dostert à l'Université de Georgetown. Cette équipe travaille en partenariat avec IBM et cette collaboration aboutit à la première démonstration publique de traduction automatique en janvier 1954. Il s'agit d'un système de TA du russe vers l'anglais contenant 250 mots de vocabulaire en russe et six règles de grammaire. Ce système, bien que limité, montre que la traduction automatique est possible, encourageant ainsi la recherche dans ce domaine.

L'engouement suscité par cette première démonstration entraine la création de différents groupes de recherche aux Etats-Unis et en URSS et les publications dans ce domaine se multiplient (Hutchins, 1986).

A la fin des années 50, quelques systèmes opérationnels voient le jour, mais la qualité des traductions obtenue n'est pas à la hauteur des attentes (Hutchins, 2014). En 1964, face au manque de résultats concrets en matière de TA, les agences qui subventionnaient massivement ces recherches créent un comité baptisé ALPAC (Automatic Language Processing Advisory Committee) dont la mission est d'étudier la faisabilité de la TA (Koehn, 2010; Hutchins, 1986). Dans son rapport publié en 1966, l'ALPAC préconise de réduire les investissements dans le domaine de la traduction automatique et de consacrer les efforts sur la recherche en linguistique et sur le développement de méthodes pour améliorer la traduction humaine (Koehn, 2010; ALPAC, 1966). Ce rapport aura pour conséquence une diminution drastique des investissements dans le domaine de la TA et un ralentissement important de la recherche (Koehn, 2010). Quelques groupes de recherche subsistent cependant et poursuivent leurs travaux notamment aux Etats-Unis, au Canada et en Europe (Quah, 2006). Le groupe TAUM (Traduction Automatique à l'Université de Montréal) est l'un d'entre eux et il met au point, en 1976, un système de traduction automatique baptisé TAUM-METEO, destiné à traduire les bulletins météorologiques de l'anglais vers le français (Quah, 2006).

A partir de la fin des années 70, on assiste à une revitalisation de la recherche dans le domaine de la traduction automatique (Quah, 2006). Aux Etats-Unis, la Pan American Health Organization (PAHO) et la United States Air Force se lancent chacune dans le développement de systèmes de traduction automatique (Arnold et al., 1992). En Europe, la Commission de la Communauté acquiert le système de TA SYSTRAN en 1976 (Bouillon et Clas, 1993). De nombreux systèmes de traduction automatique font leur apparition dans les années 80 et la recherche se poursuit dans le domaine (Hutchins, 2014). A cette époque, les chercheurs se concentrent sur les méthodes de traduction « indirectes » (Hutchins, 2014) (voir section 2.6.1.). Au début des années 90, deux nouvelles approches basées sur les corpus font leur apparition (l'approche basée sur des méthodes statistiques et l'approche basée sur l'exemple) (Hutchins, 2014). Cette période a été marquée par l'arrivée d'internet et des nouveaux modes de communication qui l'accompagnent (Quah, 2006). La naissance et la croissance rapide de ce réseau mondial vont à la fois accroitre et modifier les besoins en traduction. Durant cette décennie, les chercheurs s'intéressent aux applications pratiques de la TA, notamment pour les

traducteurs professionnels. A cette époque, les ordinateurs personnels deviennent accessibles au grand public et le recours aux outils de TA et de TAO (Hutchins, 2014).

Au début des années 90, la recherche en TA se penche aussi sur la traduction de la parole, ce qui implique le développement de la reconnaissance et de la synthèse vocale (Hutchins, 2014). C'est aussi à cette époque que les outils de TAO connaissent un véritable essor (Rochard, 2017). A la fin des années 90, les ventes de logiciels de TA connaissent une croissance importante et l'on voit apparaitre les premiers services de TA gratuits en ligne qui permettent aux internautes d'obtenir une traduction rapide des contenus web (Hutchins, 2014).

Dans les années 2000, l'utilisation de la TA continue d'augmenter et les outils de traduction automatique s'intègrent désormais dans l'environnement de travail des traducteurs professionnels au même titre que les outils de TAO (Hutchins, 2014). Pour illustrer cette apparition de la TA dans l'environnement de travail du traducteur professionnel, nous pouvons par exemple citer SDL qui, à partir de 2008, inclus petit à petit des options de TA dans ses logiciels de TAO et Google qui lance en en 2009 le Google Translator Toolkit, un outil destiné au traducteur qui mêle TA et TAO (Robert, 2010).

En ce qui concerne le grand public, les outils de TA accessibles gratuitement se multiplient depuis le début des années 2000 (on peut notamment citer *Google Translate*, *Bing* et plus récemment *DeepL*). Les applications mobiles proposant des services de TA se multiplient elles aussi et sont toujours plus sophistiquées (traduction de la parole, reconnaissance optique des caractères avec traduction instantanée…).

2.3. Etat actuel et récents progrès

La TA est aujourd'hui omniprésente dans notre quotidien, notamment sur internet. Il nous est désormais possible de traduire automatiquement les contenus publiés sur les réseaux sociaux ou les messages que l'on reçoit sur certaines applications mobiles. A lui seul, Google Translate traduirait plus de 100 milliards de mots par jour. [2]

Malgré cet essor important, la TA est encore souvent mal perçue (à la fois par les traducteurs et le grand public) car nombreux sont ceux qui estiment que la qualité des traductions reste

[2] https://www.blog.google/products/translate/ten-years-of-google-translate/ (Turovsky 2016)

médiocre. Mais depuis peu, la TA semble attirer l'attention du plus grand nombre. Ce coup de projecteur soudain est dû à l'apparition des systèmes de traduction neuronaux (ou Neural Machine Translation) dont les performances surpasseraient grandement celles des autres systèmes[3]. L'utilisation des réseaux neuronaux dans le domaine de la TA est très récente, mais elle semble en mesure de bouleverser le monde de la TA. D'après Koehn (2017), qui n'hésite pas à parler de la TAN comme du « nouvel état de l'art » en TA, la recherche s'est vue complètement métamorphosée en à peine deux ans.

2.4. Objectifs de la TA

Comme nous l'avons vu, les systèmes de TA ne sont pour le moment pas capables de produire seuls des traductions d'une qualité équivalente à celle de la traduction humaine (sauf pour des domaines très restreints). Cependant, comme Koehn le souligne, *la traduction automatique n'a pas besoin d'être parfaite pour être utile* (2010 : 20) et une traduction de qualité moyenne peut tout à fait être suffisante pour certaines utilisations. Selon Koehn (2010), on peut diviser les utilisations de la TA en trois grandes catégories : **l'assimilation**, la **dissémination** et la **communication**. La traduction doit répondre à des critères différents en fonction de l'utilisation à laquelle il est destiné.

La TA a une fonction d'assimilation lorsque l'objectif de l'utilisateur est de comprendre un texte et d'accéder aux informations qu'il contient (on parle aussi de *gisting*). Dans ce cas, la traduction n'a pas besoin d'être parfaite, elle doit simplement être suffisamment bonne pour permettre à l'utilisateur de comprendre de quoi parle le texte. On parle de fonction de dissémination lorsque la traduction doit être de grande qualité, comme dans le cas où le texte est destiné à être publié. Dans le cas de la dissémination, il est souvent nécessaire de recourir à une intervention humaine pour obtenir le niveau de qualité souhaité (Hutchins, 2005). On aura donc souvent recours à un pré-éditeur ou à un post-éditeur. Enfin, on parle de fonction de communication lorsque la traduction est destinée à permettre la communication entre plusieurs personnes. Il peut s'agir par exemple d'un échange par e-mail ou par téléphone. Dans ce cas, la

[3] https://www.latribune.fr/technos-medias/internet/quand-l-intelligence-artificielle-revolutionne-la-traduction-automatique-626326.html

traduction n'a pas besoin d'être parfaite, mais elle doit permettre aux utilisateurs de comprendre les informations qui leur sont transmises (Hutchins, 2005).

2.5. La post-édition (PE)

Comme nous l'avons expliqué dans la section 2.1, la post-édition désigne la révision et la correction de l'output de la TA en vue d'améliorer sa qualité. Le *Center for Next Generation for Localisation* (CNGL) et la *Translation Automatic User Society* (TAUS) ont donné en 2011 une définition plus précise de la PE :

> *Post-editing is the correction of machine generated translation output to ensure it meets a level of quality negotiated in advance between client and post-editor.* (cité dans Melby et al., 2014 : 278).

Cette définition souligne que l'objectif de la PE est d'atteindre un certain niveau de qualité. Comme nous l'avons déjà expliqué, le niveau de qualité attendu d'une traduction dépend de la fonction à laquelle elle est destinée. Le niveau de qualité à atteindre par la PE dépend donc de la fonction de la traduction. Si le texte à post-éditer remplit une fonction d'assimilation, la degré qualité attendu sera probablement moins élevé que pour un texte qui a une fonction de dissémination. En fonction de la qualité souhaitée, on rencontre parfois la distinction entre une PE rapide (ou légère), dans laquelle le post-éditeur effectue le minimum de changements afin de produire une traduction d'une qualité suffisante pour l'assimilation, et une PE maximale (ou complète), dans laquelle le post-éditeur effectue tous les changements requis pour obtenir une traduction de haute qualité (en vue de la publication notamment) (Allen, 2003; TAUS et CNGL, 2010). Or, cette distinction peut être trompeuse, car l'effort produit par le post-éditeur dépend certes de la qualité attendue, mais aussi de la qualité de l'output de la TA. Ainsi, si la TA est de très bonne qualité, une post-édition légère peut suffire à atteindre un niveau de qualité élevée. Il semble donc plus logique d'effectuer une distinction en fonction de la qualité attendue plutôt que de l'effort demandé au post-éditeur (TAUS et CNGL, 2010).

Le travail du post-éditeur sera donc déterminé par l'objectif de qualité fixé. Taus (2010) propose un certain nombre d'instructions afin de guider le post-éditeur dans son travail en fonction de cet objectif. Peu importe la qualité attendue au final, le post-éditeur doit toujours veiller à ce que la traduction soit correcte d'un point de vue sémantique et doit s'assurer qu'aucune

information n'a été omise ou ajoutée. Il doit aussi corriger les éventuels contenus offensants ou inappropriés et corriger l'orthographe. Enfin, il doit faire en sorte d'utiliser le plus possible l'output de la TA. Si le texte a une fonction d'assimilation et que l'objectif est d'atteindre une qualité « suffisante », il est demandé au post-éditeur de ne pas tenir compte du style et de la fluidité du texte. En revanche, si l'on attend un texte de haute qualité, on demandera au post-éditeur de corriger la grammaire et la syntaxe, de tenir compte de la terminologie, de rectifier la ponctuation et de veiller à la bonne mise en page du texte cible (TAUS et CNGL, 2010).

Le travail de PE est en général réalisé par des traducteurs (Hutchins, 2005). Cela s'explique notamment par le fait que l'activité de PE requiert un certain nombre de compétences que l'on retrouve aussi chez les traducteurs professionnels, comme la connaissance de la langue source (LS) et de la langue cible (LC), ainsi qu'une bonne connaissance du domaine dont relève le texte qui est traduit (Seewald-Heeg, 2017). Cependant, le travail de PE nécessite certaines compétences qui lui sont spécifiques, d'où le besoin de former et d'entrainer les traducteurs à la PE (Seewald-Heeg, 2017). Seewald-Heeg (2017) explique notamment que la PE implique un changement dans le mode de travail du traducteur, car ce dernier doit orienter son travail en fonction de l'objectif final de qualité. La traduction post-éditée ne sera pas nécessairement parfaite, mais doit simplement correspondre à ce qui est attendu. Le post-éditeur doit aussi veiller à réduire au maximum l'effort de post-édition afin de maximiser le bénéfice de la TA. En outre, il doit avoir une bonne compréhension de la TA et de ses capacités et être conscient du fait que les erreurs commises par la machine sont différentes de celles d'un traducteur humain. Il doit aussi être capable de décider rapidement si une phrase peut-être post-éditée ou si une retraduction est plus appropriée. Enfin, le post-éditeur doit pouvoir fournir un retour sur la qualité de la TA et sur le type d'erreurs commises afin de permettre aux développeurs d'améliorer le système (Seewald-Heeg, 2017).

En résumé, nous pouvons dire que le travail de post-édition est principalement guidé par l'objectif à atteindre du point de vue de la qualité. Nous avons aussi vu qu'un post-éditeur doit avoir à la fois des compétences dans le domaine de la traduction, mais aussi certaines compétences spécifiques à la PE.

2.6. Les différentes approches et différents systèmes

Les systèmes de TA peuvent être classés selon différents critères, nous pouvons, par exemple, distinguer les systèmes bilingues (qui ne traitent qu'une seule paire de langues) des systèmes multilingues (qui traitent plusieurs paires de langues) ou encore les systèmes unidirectionnels (qui traduisent dans une seule direction) des systèmes bidirectionnels (qui peuvent traduire dans plusieurs directions). On peut aussi classer les systèmes en fonction du type de données qu'ils utilisent pour traduire. Il existe des systèmes qui reposent sur des données linguistiques comme des règles syntaxiques ou morphologiques, et des systèmes qui sont basés sur des informations non linguistiques, comme des corpus (Bouillon et Clas, 1993). Cette section visant à donner un aperçu du mode de fonctionnement des systèmes de TA, nous avons décidé de classer les systèmes en fonction du type de données qu'ils utilisent. Nous présenterons les différents systèmes qui s'inscrivent dans chaque catégorie et donnerons un aperçu de leur fonctionnement.

2.6.1. Systèmes basés sur les règles

Parmi les systèmes qui reposent sur des données linguistiques, on retrouve les systèmes *directs* et *indirects*. Les systèmes statistiques, les systèmes basés sur l'exemple, ainsi que les systèmes neuronaux reposent quant à eux sur des informations non linguistiques (Bouillon et Clas, 1993).

Systèmes directs

Les premiers systèmes de TA mis au point sont des systèmes dits *directs*, ces systèmes sont conçus pour une seule paire de langues et pour une seule direction de traduction. Ils ont été développés à la fin des années 50 et au début des années 60, à une époque où les ordinateurs disposaient de capacités encore assez limitées (Hutchins et Somers, 1992). Les systèmes directs sont donc des systèmes assez simples. Leur fonctionnement repose principalement sur un dictionnaire bilingue qui permet de trouver l'équivalent dans la langue cible des mots de la langue source (Quah, 2006). Ces systèmes n'analysent pas la structure syntaxique, ni les

relations sémantiques dans le texte source (Hutchins et Somers, 1992). Ils ne possèdent pas non plus de connaissances approfondies de la grammaire de la langue cible (Arnold et al., 1992).

Ces systèmes sont qualifiés de directs, car ils réalisent la traduction en passant directement de la langue source à la langue cible, sans passer par une représentation intermédiaire (Hutchins et Somers, 1992). Le système réalise tout d'abord une analyse morphologique de la phrase source afin d'attribuer à chaque mot sa catégorie grammaticale (Hutchins et Somers, 1992; Arnold et al., 1992). Pour ce faire, le système consulte un dictionnaire unilingue en langue source. Le système se base sur des règles morphologiques afin d'identifier les formes fléchies (Arnold et al., 1992). Il consulte ensuite un dictionnaire bilingue pour trouver les équivalences en langue cible des termes de la phrase source. Enfin, le système procède enfin à un réarrangement des mots dans la langue cible (Hutchins et Somers, 1992). Les systèmes directs produisent donc une traduction que l'on peut qualifier de « mot-à-mot », avec simplement une réorganisation de l'ordre des mots (Hutchins et Somers, 1992). Ces systèmes ne procédant pas à une analyse syntaxique du texte source, ils sont incapables de résoudre les ambigüités et ne peuvent être utilisés que pour traduire entre des langues qui sont assez proches (Quah, 2006). Parmi les systèmes directs, on peut citer les systèmes METEO, CULT, Weidner, ainsi que l'ancienne version de Systran (Quah, 2006).

Systèmes indirects

Contrairement aux systèmes directs, les systèmes indirects ne passent pas directement de la langue source à la langue cible, mais ils créent une représentation intermédiaire (Quah, 2006). Ils disposent aussi d'une connaissance approfondie de la langue source et de la langue cible, ainsi que des relations entre elles. Ils sont dotés d'une grammaire détaillée pour chaque langue et d'une grammaire comparative (Arnold et al., 1992). On distingue deux types de systèmes indirects : les systèmes par **interlangue** et par **transfert**. Ces systèmes se différencient par le type de représentation intermédiaire qu'ils utilisent.

Dans les systèmes par interlangue, la traduction est effectuée en deux étapes : la phrase source est analysée et transformée en une représentation intermédiaire (l'interlangue) qui contient les principales informations syntaxiques et sémantiques, la phrase cible est ensuite générée à partir de cette représentation. La représentation intermédiaire est indépendante des langues de départ et d'arrivée, et donc commune à toutes les langues, elle permet ainsi de générer des phrases

sources dans plusieurs langues. Cette approche facilite donc la création de systèmes multilingues et l'ajout de nouvelles langues (Trujillo, 1999). Pour chaque langue, le système possède un module capable d'analyser et de transformer la phrase en une représentation abstraite et un module capable de générer une phrase à partir de cette représentation (Trujillo, 1999). Ainsi, si l'on a un système qui traduit du français vers l'anglais, on peut ajouter un module de génération de l'allemand et ainsi ajouter une paire de langues sans avoir à toucher aux autres éléments du système (Hutchins et Somers, 1992).

La principale difficulté que posent les systèmes par interlangue réside dans la création de l'interlangue (Quah, 2006). Elle doit être totalement indépendante de la langue et permettre de représenter toutes les informations pour la génération d'une phrase cible dans toutes les langues (ou au moins dans toutes les langues du système). Elle doit donc représenter toutes les relations, ainsi que les concepts qui sont exprimés par le lexique (Hutchins et Somers, 1992). En ce qui concerne le lexique, il faut prendre en compte le fait qu'une langue peut avoir un seul mot pour désigner un certain concept alors qu'une autre langue peut en avoir plusieurs (Hutchins et Somers, 1992). Par exemple, en allemand il existe deux équivalents pour le verbe *manger*. Lorsque le sujet est un humain, on utilise le verbe *essen* et lorsque le sujet est un animal, on utilise le verbe *fressen.* L'interlangue devra donc pouvoir distinguer ces deux concepts, même si la distinction n'existe pas en français.

Le fonctionnement des systèmes par transfert repose aussi sur l'analyse et sur la transformation du texte en une représentation intermédiaire. Mais contrairement aux systèmes par interlangue, les systèmes par transfert utilisent des représentations qui dépendent de la langue (Hutchins et Somers, 1992). Le système analyse la phrase source et crée une représentation spécifique à la langue source, cette représentation est ensuite transformée en une représentation spécifique à la langue cible qui permet de générer la phrase cible (Hutchins et Somers, 1992). La traduction se fait donc en trois étapes : une première étape d'analyse de la LS, puis une étape de transfert (de la représentation de la LS à la représentation de la LC) et enfin une étape de génération de la phrase cible. Un système de transfert est donc composé au minimum de trois modules, un module d'analyse de la LS, un module de transfert et un module de génération de la langue cible (Trujillo, 1999).

En utilisant des représentations intermédiaires dépendantes de la langue, les systèmes par transfert contournent les difficultés mentionnées dans la section précédente concernant la création d'une interlangue (Hutchins et Somers, 1992). Cette représentation dépendante de la

langue source est moins abstraite, ce qui rend les étapes d'analyse et de génération bien moins complexes (Hutchins et Somers, 1992). En revanche, l'ajout d'une nouvelle paire de langues est plus compliqué qu'avec un système par interlangue, car il nécessitera un nouveau module d'analyse, un nouveau module de génération ainsi qu'un nouveau module de transfert (Hutchins et Somers, 1992).

Le recours à des connaissances purement linguistiques pour la TA pose plusieurs problèmes, tout d'abord il nécessite de créer un grand nombre de règles syntaxiques et lexicales pour chaque langue et ce travail est long et fastidieux. Ensuite, du fait de l'abondance et de la complexité de ces règles, ces systèmes sont difficiles à manipuler et enfin, il est difficile de définir le niveau d'analyse nécessaire pour produire des traductions de bonne qualité (Arnold et al., 1992). Face à ces problèmes, et avec l'accessibilité croissante des corpus de texte électroniques à la fin des années 90, des approches « empiriques », c'est-à-dire basées sur des corpus et non plus sur des connaissances linguistiques, sont apparues (Arnold et al., 1992). Dans la section suivante, nous allons donner un aperçu de ces approches et nous présenterons deux types de systèmes qui reposent sur des données empiriques.

2.6.2. Systèmes basés sur les corpus

Les systèmes basés sur les corpus n'utilisent pas de données linguistiques pour produire une traduction, mais ils se basent sur des corpus bilingues en utilisant une *approche statistique* ou une *approche basée sur l'exemple* (Bouillon et Clas, 1993). Parmi les systèmes basés sur des corpus, il existe aussi depuis peu des systèmes qui utilisent des réseaux de neurones. Ces systèmes peuvent être fondés sur l'approche statistique classique ou sur une approche purement neuronale (Koehn, 2017). Nous verrons qu'il existe aussi des systèmes hybrides qui combinent plusieurs approches.

Systèmes basés sur l'exemple

Dans l'approche basée sur l'exemple, le système parcourt le corpus à la recherche de phrases ou de fragments de phrases très proches de la phrase source à traduire, autrement dit, des exemples de traduction. Il utilise ensuite ces segments et leur traduction comme modèles et les combine entre eux pour produire la phrase cible (Quah, 2006). Etant donné qu'il y a peu de

chance de trouver dans le corpus une phrase correspondant exactement à celle qu'il faut traduire, le système va devoir chercher des correspondances portant sur des petites parties de phrases et les exemples trouvés sont souvent très courts, ce qui complique l'étape de génération de la phrase cible. Le système va devoir combiner plusieurs segments qui ne sont pas forcément compatibles. De plus, il se peut qu'il y ait plusieurs exemples pour un même fragment de phrase et se pose alors la question de comment en choisir un (Poibeau, 2017).

Lorsqu'ils ont été développés, les systèmes basés sur l'exemple étaient majoritairement utilisés pour traduire entre des langues très éloignées (le japonais et l'anglais par exemple), car il permettait d'éviter d'avoir à formaliser des règles de transfert très compliquées pour développer des systèmes basés sur les règles. Ils donnaient aussi de meilleurs résultats pour la traduction de textes traitant de domaines très spécifiques (comme l'informatique) dans lesquels la terminologie est bien définie et les répétitions sont fréquentes (Poibeau, 2017).

Systèmes statistiques

Dans le cas de l'approche statistique, le système utilise des corpus pour *trouver la phrase cible qui a le plus de probabilités d'être la traduction de la phrase source* (Bouillon, 1993 : 13). Le système utilise un corpus bilingue pour créer un modèle statistique de traduction, ce modèle peut être vu comme une sorte de grand dictionnaire bilingue qui contient toutes les traductions possibles (contrairement à un dictionnaire bilingue classique, qui lui ne contient que les traductions très probables), chacune de ces traductions se voit assigner une probabilité en fonction de sa fréquence d'apparition dans le corpus (Hearne et Way, 2011). A l'aide d'un corpus monolingue (de la langue cible), le système crée aussi un modèle statistique de langue qui assigne des probabilités aux séquences de mots de la langue cible (appelées **n-grammes**) (Hearne et Way, 2011). Le système cherche ensuite parmi toutes les possibilités, la traduction qui dont la probabilité est la plus élevée selon le modèle de traduction et le modèle de langue (Hearne et Way, 2011). Nous présenterons les systèmes de TAS de manière plus approfondie dans la section 4.1.1.

Systèmes neuronaux

Les systèmes basés sur les réseaux de neurones peuvent avoir différents modes de fonctionnement. Les premiers systèmes faisant appel à des réseaux de neurones visaient à

intégrer des modèles neuronaux à des systèmes de traduction statistique classiques afin d'en améliorer les performances (Koehn, 2017).

D'autres approches visent à créer des systèmes de TA purement neuronaux (en anglais : *pure neural machine translation*) dont le fonctionnement ne repose plus sur les méthodes statistiques (Koehn, 2017 : 6). Différents modèles de traduction neuronale pure ont été proposés et la recherche dans ce domaine est très dynamique (Koehn, 2017). En théorie, les systèmes entièrement neuronaux sont capables d'apprendre à l'aide de corpus d'entrainement de manière entièrement autonome. Ils sont composés d'un encodeur qui analyse le corpus d'entrainement et d'un décodeur qui génère une traduction à partir des données analysées. L'encodeur et le décodeur sont entièrement composés de réseaux de neurones (Poibeau, 2017). Contrairement aux systèmes statistiques classiques, les systèmes neuronaux prennent en compte la phrase dans son ensemble sans la décomposer en plus petites unités (Poibeau, 2017). Ces systèmes de traduction neuronale sont très complexes et nous tenterons de les présenter de manière plus détaillée dans la section 4.2.1.

Systèmes hybrides

Parmi les systèmes basés sur des corpus, on trouve aussi des systèmes dits *hybrides*, qui combinent des éléments issus de systèmes basés sur des informations linguistiques et des systèmes basés sur les corpus. L'objectif est de concevoir des systèmes qui puissent tirer le meilleur des deux approches. Les systèmes hybrides peuvent par exemple combiner l'approche linguistique et l'approche basée sur la connaissance en utilisant une méthode linguistique pour traduire et une méthode statistique pour post-éditer automatiquement le résultat brut de la traduction. Il existe aussi des systèmes hybrides qui utilisent deux systèmes de TA en parallèle et qui combinent ensuite les résultats. D'autres sont conçus en ajoutant des composantes d'un système à un autre ou encore en combinant toutes les composantes de deux systèmes (Thurmair, 2009).

Conclusion

Dans cette partie, nous avons défini le domaine de la TA et nous avons retracé son évolution de ses débuts jusqu'à aujourd'hui. Nous avons vu que l'on peut classer les différents types de TA en fonction de la place du traducteur humain dans le processus et nous avons présenté

l’activité de post-édition. Cette partie nous a aussi permis de donner un aperçu des différents objectifs de la TA en fonction des utilisations à laquelle elle est destinée. Enfin, nous avons mentionné les différentes approches et les principaux types de systèmes de TA.

3. EVALUER LA TA

Dans ce chapitre, nous allons donner un aperçu de quelques méthodes d'évaluation de la TA. Cette présentation nous permettra de comprendre les différentes manières d'évaluer la TA afin de sélectionner le ou les types d'évaluation les plus adaptés dans le cadre de ce travail. Nous commencerons par présenter brièvement les techniques d'évaluation humaine (section 3.1) et nous présenterons ensuite différentes métriques automatiques (section 3.2).

A l'heure actuelle, il n'existe pas de méthode d'évaluation standard et universelle dans le domaine de la TA, mais on trouve de nombreuses méthodes, parfois très différentes les unes des autres (Quah, 2006). Cette situation peut s'expliquer par le fait que l'évaluation de la TA est un enjeu pour différents acteurs et que chacun de ces acteurs à des besoins différents (Quah, 2006). Un traducteur, par exemple, ne sera pas forcément intéressé par les mêmes informations qu'un chercheur ou qu'un développeur, il est donc essentiel d'adapter les méthodes d'évaluation aux besoins des destinataires de l'évaluation (White, 2003). On peut notamment distinguer les évaluations réalisées en « boite de verre » (*glass box* en anglais) des évaluations réalisées en « boite noire » (*black-box* en anglais). Les premières visent à évaluer le système en fonction de ses caractéristiques internes, elles sont particulièrement intéressantes pour les chercheurs et les développeurs, tandis que les secondes portent sur l'évaluation de l'output du système sans tenir compte de ses éléments constitutifs. Les évaluations en boites noires répondent en général aux attentes des utilisateurs finaux des systèmes de TA (Dorr et al., 2011; Trujillo, 1999).

La multiplication des méthodes d'évaluation peut aussi s'expliquer par le fait que l'évaluation de la TA est une tâche compliquée, car en traduction, il n'y a, en général, pas *une seule et unique* bonne réponse. On peut faire traduire une phrase par différents traducteurs, aucun de produira la même traduction, mais il est possible que toutes les traductions soient correctes (White, 2003). Face à la complexité du travail d'évaluation et au grand nombre de méthodes existantes, il semble tout d'abord essentiel de connaitre les critères qui définissent une bonne métrique d'évaluation. Koehn (2010) donne quatre critères pour les méthodes d'évaluation de la qualité de l'output : tout d'abord, une métrique doit avoir un **faible coût**, c'est-à-dire qu'elle doit nécessiter un investissement en temps et en argent le plus faible possible. Ensuite, elle doit être **significative**, en d'autres termes, elle doit donner un résultat que l'on peut interpréter en termes de qualité de l'output. Une métrique doit aussi être **consistante**, ce qui signifie qu'un même

juge doit toujours évaluer l'output avec la même sévérité (si c'est le cas, on dit que la métrique est **stable**), mais aussi que plusieurs juges qui effectuent la même évaluation doivent obtenir les mêmes résultats (il doit y avoir un **accord entre les juges**). Et enfin, une métrique doit donner un **jugement correct**, c'est-à-dire que le résultat doit refléter correctement la qualité réelle de l'output.

On peut classer les méthodes d'évaluation de l'output en deux grandes catégories : **l'évaluation humaine** et **l'évaluation automatique** (Koehn, 2010). Ces deux catégories présentent chacune des avantages et des inconvénients et peuvent répondre à des objectifs différents. Nous allons présenter ces deux grandes catégories et donner des exemples de méthodes d'évaluations pour chacune d'elles, nous mentionnerons aussi les caractéristiques de ces méthodes.

3.1. Evaluation humaine

Parmi les méthodes d'évaluation humaine, Kit et Wong font une distinction entre les méthodes d'évaluation **intrinsèque** et les méthodes d'évaluation **extrinsèque**. L'évaluation intrinsèque vise à déterminer la qualité de l'output tandis que l'évaluation extrinsèque vise à déterminer l'utilité de l'output dans le contexte dans lequel il est utilisé (Kit et Wong, 2015). Pour réaliser une évaluation intrinsèque, on demande donc à des juges de déterminer le niveau de qualité du produit de la TA. Plusieurs approches basées sur différents critères ont été développées pour effectuer ce genre d'évaluations.

Jugement intuitif (fluidité & fidélité)

L'une des approches communément utilisées consiste à juger la **fluidité** (ou intelligibilité) et **l'adéquation** (ou fidélité). La fluidité concerne la qualité de la langue cible, indépendamment du contenu sémantique de la phrase, il s'agit d'évaluer si la phrase est lisible et si elle respecte les règles de construction de la LC. L'adéquation, quant à elle, se rapporte au contenu sémantique de la traduction, la tâche du juge est ici de déterminer si la phrase cible contient les mêmes informations que la phrase source (Kit et Wong, 2015; Koehn, 2010; Blanchon et Boitet, 2008). Une traduction fluide, ou intelligible, est donc une traduction qui se lit bien et qui respecte les règles de la langue cible et une traduction adéquate, ou fidèle, est une traduction qui restitue parfaitement le sens de la phrase source. La fluidité ne concerne donc

que la langue cible et est indépendante de la phrase source, elle peut donc être évaluée par des juges monolingues. Pour juger de l'adéquation, on peut faire appel à des juges bilingues qui se baseront sur la phrase source ou à des juges monolingues auxquels on fournira une traduction humaine de référence (Kit et Wong, 2015). La fluidité et l'adéquation sont en général mesurées à l'aide d'une échelle de 3, 5 ou 7 valeurs (Gerlach, 2015). Koehn (2010 : 219) propose, par exemple, une échelle de 5 valeurs allant de « all meaning » à « none » pour la fidélité et de « flawless English » à « uncomprehensible » pour la fluidité. Les évaluateurs ne jugent pas forcément la fluidité et la fidélité avec la même sévérité, ce qui peut donner des résultats avec un faible accord entre les juges, de plus les juges peuvent avoir du mal à être constant dans leur évaluation, car ces notions sont assez vagues. En outre, la fidélité et la fluidité, bien qu'indépendantes l'une de l'autre, sont souvent liées (une phrase inintelligible à peu de chance de transmettre beaucoup d'information) ce qui peut entrainer de la confusion chez les juges (Koehn, 2010).

Evaluation par comparaison

Face aux difficultés que pose l'évaluation de la fidélité et de la fluidité, il semble plus aisé d'évaluer le résultat de la TA par comparaison (Koehn, 2010). Dans l'évaluation par comparaison, on demande aux juges de classer plusieurs traductions (obtenues avec différents systèmes) de la meilleure à la moins bonne (Kit et Wong, 2015). Une étude conduite par Callison-Burch et al. a montré que cette méthode d'évaluation donne lieu à un meilleur accord entre les juges ainsi qu'à un jugement plus constant (Callison-Burch et al., 2007). Mais l'évaluation par comparaison ne fournit pas d'indication sur le type d'erreur, ni sur les faiblesses du système (Gerlach, 2015). De plus, l'évaluation par comparaison ne reflète pas forcément correctement la qualité d'un système, car un système peut être meilleur qu'un autre sans pour autant être bon.

Analyse des erreurs

La méthode d'analyse des erreurs peut être vue comme un moyen de réaliser une évaluation assez objective. Cette méthode consiste à demander aux évaluateurs de compter le nombre de modifications nécessaires pour arriver à une traduction correcte (Hutchins et Somers, 1992). Le nombre de modifications reflète l'effort de post-édition que requiert la TA

brute. Cependant, pour bien refléter cet effort, il est nécessaire d'attribuer un coefficient à chaque erreur en fonction des conséquences qu'elle a sur la fidélité et la fluidité, de la difficulté de sa correction et du temps nécessaire à la correction (Trujillo, 1999). L'évaluation par l'analyse des erreurs est considérée comme plus fiable que le jugement intuitif, car elle est plus objective et donne un meilleur accord entre les juges. De plus, elle donne de bonnes indications sur le type d'erreur et sur le niveau de qualité de l'output (Kit et Wong, 2015 : 223). Cependant, l'identification et la classification des erreurs peut poser problème pour les juges. Tout d'abord, les évaluateurs peuvent avoir une tolérance variable face aux erreurs et, ensuite, les catégories d'erreurs ne sont pas toujours claires et certaines erreurs peuvent appartenir à plusieurs catégories différentes (Kit et Wong, 2015 : 223).

Evaluations fondées sur la tâche

Les méthodes d'évaluation intrinsèque que nous venons de présenter visent à déterminer la qualité de l'output de la TA. Les méthodes fondées sur la tâche sont des méthodes d'évaluation extrinsèques qui cherchent à estimer dans quelles mesures l'output de la TA est utile dans le contexte dans lequel il est destiné à être utilisé. Ces méthodes d'évaluation dépendent donc fortement de la fonction de la TA.

Si la TA à une fonction d'assimilation, on peut par exemple réaliser une évaluation basée sur un test de compréhension du contenu en demandant à des évaluateurs de lire un texte traduit automatiquement et de répondre à des questions portant sur le contenu du texte. Le nombre de réponses correctes permet alors d'avoir une idée de la quantité d'information transmise par l'output (Koehn, 2010). Dans le cas où l'output de la TA est destiné à être post-édité par un traducteur pour atteindre un niveau de qualité donné, on peut mesurer l'effort fourni par le post-éditeur pour corriger l'output de la TA. Cet effort de PE peut être déterminé en mesurant le temps nécessaire à la PE et/ou le nombre de modifications effectuées par le post-éditeur pour arriver au niveau de qualité requis (Kit et Wong, 2015). Ce type d'évaluation basée sur la mesure de l'effort de PE vise à évaluer la qualité de la TA du point de vue de son utilité pour le traducteur, on cherche alors à déterminer si l'output est suffisamment bon pour pouvoir servir de base au traducteur (Kit et Wong, 2015). Mesurer l'effort de PE est aussi un moyen d'évaluer la qualité intrinsèque de la TA puisque, comme le soulignent Kit et Wong (2015), l'effort de PE est en général moins important lorsque la TA est de bonne qualité.

Il semble assez logique d'avoir recours à des évaluateurs humains pour juger les performances d'une machine, d'une part, car les traductions sont destinées à des humains, ce sont donc les humains eux-mêmes qui sont les mieux placer pour juger de leur qualité et, d'autre part, car les humains sont capables d'identifier le degré de gravité des erreurs (Dorr et al., 2011 : 751). Mais l'évaluation humaine a aussi ses inconvénients. Tout d'abord, comme nous l'avons vu, elle est inévitablement empreinte de subjectivité. Ensuite, elle implique d'avoir à disposition des évaluateurs et nécessite souvent beaucoup de temps (et donc potentiellement d'argent) (Koehn, 2010). Ces inconvénients peuvent être un vrai problème si l'on souhaite évaluer rapidement plusieurs systèmes pour se faire une idée de leurs performances respectives. Par exemple, lorsque des chercheurs tentent de développer un système et qu'ils veulent pouvoir connaitre les effets de tel ou tel paramètre sur les performances de leur système durant la phase de développement. Ils seront alors surement amenés à faire un grand nombre d'évaluations et n'auront donc pas forcément le temps et les ressources nécessaires à une évaluation humaine. Pour pallier ce type de problème, il existe des méthodes d'évaluation automatique, nous allons en présenter quelques-unes dans la section suivante.

3.2. Evaluation automatique

Toutes les métriques d'évaluation automatique reposent sur une base commune : la comparaison du résultat de la TA est avec une ou plusieurs traductions de référence (traduites par un humain) (Koehn, 2010). Pour cela, on part du principe qu'une traduction qui ressemble à la référence à plus de chance d'être correcte qu'une traduction qui ne lui ressemble pas (Koehn, 2010). Bien que, comme expliqué plus haut, il soit possible d'avoir des dizaines de traductions d'une même phrase source qui soient toutes correctes, mais toutes différentes.

Précision et rappel

On peut évaluer automatiquement l'output de la TA en mesurant la précision et le rappel. La précision est obtenue en divisant le nombre de mots corrects dans la traduction candidate (c'est-à-dire l'output de la TA) par le nombre total de mots qu'elle contient et le rappel correspond au nombre de mots corrects divisé par le nombre de mots de la traduction de référence (Koehn, 2010). Ces métriques peuvent être trompeuses, car on peut avoir un output

qui ne contient que des mots corrects (et qui a donc une précision élevée), mais dans lequel il manque de nombreux mots de la référence. Et, à l'inverse, on peut avoir un output qui contient un grand nombre de mots corrects (et qui a donc un rappel élevé), mais qui contient aussi plusieurs mots qui ne figurent pas dans la référence (Koehn, 2010). Dans l'idéal, une traduction doit donc avoir à la fois une précision et un rappel élevés. La **f-measure** est une métrique communément utilisée pour combiner précision et rappel, elle est obtenue en divisant le produit de la précision et du rappel par la moitié de la somme de la précision et du rappel (Koehn, 2010). La précision, le rappel, ainsi que la f-measure sont des métriques qui ne prennent pas en compte l'ordre des mots dans la phrase.

Word Error Rate

Le **Word Error Rate** (WER) est une autre métrique automatique, elle repose sur ce que l'on appelle la **distance de Levenshtein**. Cette distance correspond au nombre minimum de modifications (insertion, suppression et remplacement) qu'il faut apporter à une phrase pour la rendre identique à la référence (Koehn, 2010). Le WER est calculé en divisant le nombre de modifications par le nombre de mots de la traduction de référence (Koehn, 2010). Le principal problème du WER, c'est qu'il tend à attribuer de mauvais scores à des traductions qui sont correctes, mais dont la formulation diffère sensiblement de la référence (Gerlach, 2015 : 103).

Translation Edit Rate et Human-targeted Translation Edit Rate

Tout comme le WER, le **TER** (***Translation Edit Rate***) (Snover et al., 2006) mesure le nombre minimum de modifications à apporter à la TA pour qu'elle corresponde exactement à la traduction de référence. Mais en plus des insertions, des suppressions et des remplacements, le TER prend aussi en compte les déplacements (***shifts***) de séquences de mots. Peu importe le nombre de mots de la séquence, le déplacement est comptabilisé comme une modification au même titre qu'une insertion, qu'une suppression ou qu'un remplacement (Snover et al., 2006; Snover et al., 2009). Ainsi, dans certain cas, le TER comptabilisera une seule modification (un déplacement) là où le WER en compterait deux (une suppression et une insertion). Cette manière de comptabiliser revient à considérer que le post-éditeur humain déplace la séquence de mots qui est au mauvais endroit en faisant un couper/coller, plutôt qu'en la supprimant et en la réécrivant à la bonne place (Koehn, 2010). De la même manière que le WER, le TER est

calculé en divisant le nombre de modifications par le nombre de mots dans la référence (Snover et al., 2006).

Le **HTER** (***Human-targeted Translation Edit Rate*** ou ***Human-mediated Translation Edit Rate***) (Snover et al., 2006; Snover et al., 2009) est une métrique semi-automatique qui repose sur le même calcul que le TER. Mais pour calculer le HTER, on demande à un post-éditeur de générer une nouvelle traduction de référence en éditant l'output de la TA. Le HTER cherche ainsi à calculer le TER en se basant sur une référence la plus proche possible de la TA. Le HTER prendra ainsi mieux en compte les équivalences sémantiques ainsi que les variations dans la formulation que le TER. En effet, si l'output de la TA est correct, mais très éloigné de la TA, le TER sera très mauvais, alors que si l'on crée une référence correcte en se basant sur la TA, le TER sera bien meilleur. Pour calculer le HTER, on demande donc à un traducteur humain de post-éditer l'output de la TA pour créer une nouvelle référence qui est ensuite utilisée pour calculer le score TER (Snover et al., 2006; Snover et al., 2009). Le HTER est qualifié de métrique semi-automatique, car il requiert l'intervention d'un humain pour créer les nouvelles traductions de référence (Snover et al., 2009).

BLEU

Le **BLEU** (Bilingual Evaluation Understudy) est une métrique automatique basée sur la précision (Papineni et al., 2002). Dans le cas de BLEU, la précision est calculée pour des séquences de *n* mots (des **n-grammes**) et non pour des mots uniquement (**unigramme**) (Papineni et al., 2002). La longueur maximum des n-grammes prise en compte peut varier, mais elle est généralement de 4 (Gerlach, 2015). Tandis que la précision des 1-grammes indique le bon choix des mots et reflète donc la fidélité, la précision des séquences de mots plus longues donne des indications sur l'ordre des mots, et donc sur la fluidité (Papineni et al., 2002). Dans le calcul de BLEU, la précision est dite **modifiée** (Papineni et al., 2002), c'est-à-dire qu'elle exclut les n-grammes corrects qui apparaissent plus de fois dans la phrase candidate que dans la phrase de référence. L'exemple donné par Papineni et al. (2002 : 312) permet de mieux comprendre en quoi consiste la précision modifiée :

Phrase candidate : the the the the the the the

Référence : The cat is on the mat

Pour simplifier cet exemple, nous considérons ici la précision sur la base des mots. Comme nous l'avons expliqué précédemment, la précision correspond au nombre de mots corrects divisé par le nombre de mots de la phrase candidate. La phrase candidate obtient ici une précision de 1 (7/7). Pour obtenir la précision modifiée, on part du principe que puisqu'il n'y a que deux « the » dans la référence, on ne considère comme corrects que deux « the » dans la phrase candidate. Dans cet exemple, la précision modifiée de la phrase candidate est donc de 0.29 (2/7). Cette précision modifiée pénalise les phrases trop longues, mais pas les phrases trop courtes, une « pénalité de brièveté » a donc été intégrée au BLEU afin de diminuer le score des phrases trop courtes (Papineni et al., 2002). Un autre aspect important du BLEU est qu'il peut être calculé en utilisant plusieurs traductions de référence, ce qui augmente la tolérance de la métrique aux variations dans le choix et l'ordre des mots, car la phrase candidate n'est plus évaluée par rapport à une solution unique, mais par rapport à plusieurs (Papineni et al., 2002; Koehn, 2010). Le score BLEU est compris entre 0 et 1, 0 signifiant que la phrase candidate n'a aucun n-gramme en commun avec aucune des références et 1 signifiant qu'elle est parfaitement identique à l'une des références (Papineni et al., 2002).

BLEU est une métrique largement utilisée dans le domaine de l'évaluation de la TA et elle est généralement considérée comme une métrique dont les résultats corrèlent avec ceux des évaluations humaines (Doddington, 2002; Coughlin, 2003). Mais elle n'en est pas moins critiquée pour autant. On lui reproche notamment de traiter tous les mots de la même manière sans tenir compte du fait que certains mots peuvent avoir un impact plus important que d'autres sur la qualité de la traduction (l'oubli ou l'ajout d'une négation, par exemple, peut modifier profondément le sens de la phrase). On reproche aussi au BLEU de ne pas considérer la cohérence grammaticale de la phrase dans son ensemble. Une autre critique porte sur le fait que les scores BLEU obtenus ne sont pas forcément significatifs, car il est difficile d'interpréter un score BLEU en termes de qualité de l'output. Enfin, des expériences ont montré que des traducteurs humains peuvent obtenir des scores BLEU très proches de ceux de la TA alors que leurs traductions sont en réalité de bien meilleure qualité, en d'autres termes, cela signifie que le score BLEU n'est pas nécessairement correct (Koehn, 2010 : 229).

Il existe différentes métriques dérivées de BLEU, on peut notamment citer les métriques **NIST** (Doddington, 2002) ou **METEOR** (Banerjee et Lavie, 2005). Alors que le BLEU donne la même importance à tous les n-grammes, le NIST donne plus de poids aux n-grammes dont la valeur informative est supérieure, c'est-à-dire les n-grammes dont la fréquence d'apparition est plus faible. La pénalité de brièveté a aussi été modifiée dans le NIST afin de pénaliser moins

sévèrement les faibles variations dans la longueur de la phrase (Doddington, 2002). Le METEOR quant à lui n'identifie pas seulement les n-grammes identiques (comme le fait le BLEU), mais prend aussi en compte les variantes morphologiques ainsi que les synonymes (Banerjee et Lavie, 2005).

Les métriques automatiques que nous avons mentionnées ici ont toutes l'avantage d'être beaucoup moins couteuses que les évaluations humaines, mais leur fiabilité est cependant sans cesse remise en question. Pour évaluer une métrique automatique, on part du principe qu'une bonne métrique est une métrique dont les résultats corrèlent avec l'évaluation humaine, cette dernière étant considérée comme la référence absolue pour l'évaluation de la TA (Koehn, 2010; Kit et Wong, 2015). Les études sur la corrélation entre les résultats des évaluations humaines et automatiques sont nombreuses, mais elles n'aboutissent pas forcément aux mêmes conclusions (Kit et Wong, 2015). Certaines études montrent notamment que le degré de corrélation serait dépendant de la quantité de données évaluées ainsi que du nombre de traductions de références utilisées (Kit et Wong, 2015).

Conclusion

Dans cette partie, nous avons vu qu'évaluer la TA est une tâche complexe et qu'il existe de nombreuses méthodes d'évaluation. Nous avons présenté quelques-unes des méthodes couramment utilisées pour évaluer la TA. Nous avons vu que les méthodes d'évaluation humaine tendent à donner des informations plus précises sur la qualité de l'output que les méthodes automatiques et sont souvent considérées comme plus fiables. Cependant, les méthodes automatiques offrent l'avantage d'être plus objectives, beaucoup plus rapides et bien moins couteuses. Le choix d'une ou de plusieurs métriques dépend donc de ce que l'on cherche précisément à évaluer, mais aussi du contexte dans lequel l'évaluation est conduite.

4. SYSTÈMES UTILISÉS POUR NOTRE EXPÉRIENCE

Dans la section 2.6, nous avons présenté brièvement le fonctionnement des principaux types de systèmes de TA. Nous allons maintenant présenter plus en détail le fonctionnement des systèmes que nous avons utilisés lors de notre expérience ainsi que les techniques sur lesquelles ils sont basés. Nous allons commencer par détailler le fonctionnement des systèmes statistiques avant de décrire MTH (section 4.1). Nous présenterons ensuite les principes de bases qui constituent les systèmes neuronaux et parleront de *DeepL* (section 4.2).

4.1. Les systèmes statistiques et MTH

4.1.1. Les systèmes statistiques

Parmi les systèmes statistiques non neuronaux, on trouve deux grandes générations de systèmes : les systèmes basés sur les mots et les systèmes basés sur les segments (Bouillon, 2017). Ces deux générations reposent sur un même mode de fonctionnement et ont de nombreuses caractéristiques communes. Nous allons donc présenter le mode de fonctionnement général des systèmes de TA statistiques tout en mentionnant certaines de caractéristiques qui différencient les systèmes basés sur les mots de ceux basés sur les segments.

Comme nous l'avons vu au chapitre 2 (section 2.6.2), le principe de la TA statistique est de trouver la traduction la plus probable d'un point de vue statistique. Il existe deux moyens mathématiques pour parvenir à cet objectif : le modèle ***noisy-channel*** et le modèle ***log-linear*** (Hearne et Way, 2011). Il s'agit de formules de calcul permettant d'identifier la traduction la plus probable. En fonction du modèle sur lequel il repose, un système de TAS aura différentes caractéristiques, mais nous allons tout d'abord nous concentrer sur ce qui est commun à ces deux modèles. Que l'on utilise la *noisy-channel* ou le *log-linear*, l'identification de la traduction la plus probable nécessite d'avoir deux éléments principaux : un modèle de langue et un modèle de traduction (Hearne et Way, 2011). Ces modèles sont créés lors de la phase d'entrainement du système. Un système de TA statistique fonctionne en deux étapes : l'entrainement et la traduction. La phase d'entrainement consiste à créer les modèles de langue et de traduction à partir de corpus. Durant la phase de traduction (aussi appelée **décodage**), le système utilise ces

modèles pour produire une traduction (Bouillon, 2017). Nous allons tout d'abord présenter la phase d'entrainement en détaillant la création des modèles de langue et de traduction et nous parlerons ensuite de la phase de traduction.

Entrainement

Le modèle de langue est créé à partir d'un corpus monolingue de la langue cible, c'est ce modèle qui assure la fluidité de traduction. Ce modèle va notamment faciliter le bon choix des mots et de leur ordre (Bouillon, 2017; Koehn, 2010). Le modèle de traduction quant à lui, est créé à partir d'un corpus bilingue et permet d'assurer la fidélité de la traduction (Bouillon, 2017). Nous allons présenter ces modèles plus en détail.

MODÈLE DE LANGUE

Le modèle de langue le plus simple est le modèle dit **unigramme**. Ce modèle calcule la probabilité d'apparition de chaque mot en divisant le nombre d'occurrences d'un mot par le nombre total de mots présents dans le corpus. On obtient ensuite la probabilité d'une phrase en multipliant les probabilités de chaque mot (Hearne et Way, 2011). Le problème de ce modèle est qu'il ne peut pas détecter les phrases mal formées dont l'ordre des mots n'est pas correct et qu'il tend à donner de meilleurs scores aux phrases courtes (car le nombre de probabilités à multiplier est plus faible). Un moyen de palier à cela est d'utiliser un modèle de langue qui calcule les probabilités pour des séquences de mots et non pour des mots seuls (Hearne et Way, 2011). Il existe des modèles **bigrammes**, qui prennent en compte toutes les séquences de deux mots présentes dans le corpus et qui calculent la probabilité que le premier mot soit suivi du deuxième (Hearne et Way, 2011). Prenons par exemple la séquence de deux mots « je suis », le module va calculer la probabilité que le mot « je » soit suivi par « suis ». Pour ce faire, il va compter toutes les occurrences de « je suis » dans le corpus et diviser ce nombre par le nombre total d'occurrences du mot « je ». Le modèle de langue le plus couramment utilisé est le modèle **trigramme**, qui calcule les probabilités pour une séquence de trois mots (Koehn, 2010). Pour une séquence donnée de trois mots, le modèle calcule la probabilité que les deux premiers mots soient suivis du troisième (Koehn, 2010). Si l'on reprend notre exemple et que l'on considère la séquence de trois mots « je suis ici », le module va déterminer la probabilité que le mot « ici » apparaisse après la séquence « je suis » en divisant le nombre d'occurrences de « je suis ici »

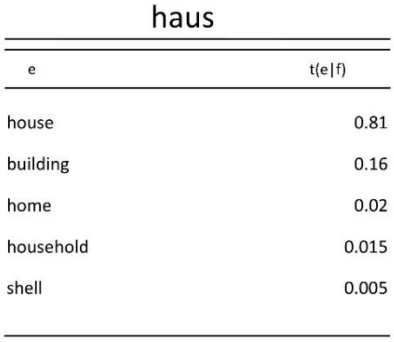
par le nombre d'occurrences de « je suis ». On peut en déduire que plus le modèle prend en compte de longs n-grammes, plus la traduction sera fluide. Cependant, plus les n-grammes sont longs, plus il y a de chances qu'ils ne figurent pas dans le corpus et donc que la phrase candidate se voit attribuer une probabilité de 0 (Hearne et Way, 2011).

Le choix d'adopter un modèle unigramme, bigramme, trigramme ou prenant en compte des n-grammes plus longs dépend notamment de la taille du corpus d'entrainement dont on dispose, car plus le corpus est grand plus il sera possible de prendre en compte des séquences de mots longues (Koehn, 2010). Mais, peu importe la taille des n-grammes pris en compte et la taille du corpus, il a toujours de grandes chances que des n-grammes absents du corpus d'entrainement apparaissent dans les phrases à traduire (Koehn, 2010). Puisque l'on obtient la probabilité d'une phrase en multipliant les probabilités de tous ses n-grammes, si un n-gramme n'est pas présent dans le corpus, alors toute la phrase de retrouvera avec une probabilité de 0 (Hearne et Way, 2011). Afin d'éviter ce genre de cas, des mécanismes de *smoothing* sont intégrés aux modèles de langues (Koehn, 2010). Il existe différents types de *smoothing,* mais ils reposent tous sur le même principe, à savoir, celui de modifier les probabilités obtenues de manière empirique afin d'avoir une réserve de probabilités à attribuer aux n-grammes absents du corpus (Koehn, 2010). Un autre moyen de palier au problème des n-grammes absents du corpus est d'utiliser l'*interpolation* ou le *back-off*. Ces deux techniques reposent sur l'utilisation de n-grammes de différentes longueurs lors de la création du modèle de langue. Dans le cas de l'interpolation, on combine simplement différents modèles de langue basés sur des n-grammes de différentes longueurs (par exemple un modèle unigramme, un modèle bigramme et un modèle trigramme). L'interpolation est dite *récursive* lorsque l'on attribue un poids plus important au modèle dont les n-grammes sont les plus longs et un poids moins important au modèle dont les n-grammes sont plus courts. Dans le cas du back-off, on utilisera en priorité le modèle de langue dont les n-grammes sont les plus longs et si ces n-grammes sont absents du corpus on utilisera alors un modèle avec des n-grammes plus courts (Koehn, 2010)

MODÈLE DE TRADUCTION

Tandis que le modèle de langue assure la fluidité de la traduction, le modèle de traduction assure sa fidélité. Le modèle de traduction est créé à partir d'un corpus bilingue, il donne la probabilité qu'un mot (ou une séquence de mots) de la langue cible soit la traduction d'un mot (ou d'une séquence de mots) de la langue source (Kenny et Doherty, 2014). Le mot

et ses traductions possibles, ainsi que les probabilités de chacune d'elles, sont présentés dans une table de traduction. Voici un exemple de table de traduction tiré de Koehn (2010) (Figure 2) :



haus

e	t(e|f)
house	0.81
building	0.16
home	0.02
household	0.015
shell	0.005

Figure 2 - *Lexical translation probability tables for four German words.*
Adaptation de Koehn (2010 : 84)

Pour les systèmes basés sur les mots, le modèle de traduction regroupe les probabilités de traduction des mots pris isolément tandis que pour les systèmes basés sur les segments, le modèle de traduction contient les probabilités de traduction de séquences de mots (Koehn, 2010).

Comme nous l'avons mentionné plus haut, le modèle de traduction donne la distribution des probabilités pour chaque traduction possible d'un mot ou d'un segment source. Or, pour calculer les probabilités des traductions, il faut tout d'abord pouvoir identifier quel mot (ou segment) cible est la traduction du mot (ou du segment) source en question. Les corpus bilingues étant alignés sur la base des phrases et non des mots, il faut trouver un moyen d'identifier les paires source-cible dans le corpus (Koehn, 2010). *L'expectation maximization algorithm* est un outil qui permet de déduire l'alignement des mots (Hearne et Way, 2011). Pour commencer, cet algorithme considère toutes les paires de mots possibles et leur assigne à toutes la même probabilité (c'est l'étape ***d'initialisation***). S'en suit l'étape dite **d'*expectation*** au cours de laquelle le modèle est appliqué aux données du corpus afin d'assigner une probabilité à chaque paire de mots en fonction de leur fréquence. Durant l'étape suivante (la ***maximisation***), les nouvelles probabilités pour chaque paire de mots sont ajoutées au modèle. Les deux dernières étapes sont ensuite réitérées plusieurs fois afin d'améliorer les estimations jusqu'à ce qu'il soit possible de déterminer les paires les plus probables pour chaque traduction. Cet algorithme permet d'obtenir un alignement des mots et donc de créer un modèle de traduction

basé sur les mots. Cet alignement des mots peut ensuite servir de base à la création d'un modèle basé sur les segments via l'application d'un algorithme d'extraction des segments (Koehn, 2010).

Nous venons de présenter les deux modèles indispensables à tout système de TA statistique. Dans les systèmes qui sont basés sur un calcul *log-linear*, il est possible de donner plus ou moins de poids au modèle de langue ou au modèle de traduction, cet ajustement de l'importance accordée à chaque modèle est appelé ***tuning*** (Koehn, 2010; Hearne et Way, 2011). La formule de calcul *log-linear* permet aussi d'inclure des composantes supplémentaires au modèle de traduction, comme des probabilités de traduction bidirectionnelles ou des pénalités pour les phrases trop courtes ou trop longues (Koehn, 2010).

Décodage

Après la phase d'entrainement vient la phase de décodage durant laquelle le système utilise les modèles décrits précédemment pour trouver la traduction la plus probable d'une phrase source (Koehn, 2010). La phrase source est tout d'abord segmentée de toutes les manières possibles, puis le système cherche les traductions des segments dans les tables de traduction du modèle de traduction (Bouillon, 2017). A ce stade, chaque traduction candidate à une certaine probabilité qui découle du modèle de traduction. Le modèle de langue assigne ensuite une probabilité à chacune d'elles et ces deux probabilités sont combinées afin d'identifier la phrase qui obtient le meilleur score (Koehn, 2010; Kenny et Doherty, 2014). Dans le cas des systèmes basés sur les segments, un modèle de réordonnancement est aussi appliqué avant le modèle de langue afin de réordonner correctement les segments (Koehn, 2010). L'étape de décodage est complexe, car le nombre de traductions possibles pour une phrase source est extrêmement grand, ce qui implique d'avoir une très grande puissance de calcul. Il existe différentes techniques pour chercher la meilleure traduction sans calculer la probabilité de l'ensemble de traductions possibles (Koehn, 2010).

Les systèmes basés sur les phrases sont actuellement les plus utilisés, car ils sont généralement plus performants que les systèmes basés sur les mots. Cela peut s'expliquer par le fait que les systèmes basés sur les mots ont du mal à traiter les cas dans lesquels un mot source est traduit par deux mots cibles (et inversement), mais aussi par le fait qu'en se basant sur des segments il est plus facile de résoudre les ambigüités.

4.1.2. *Microsoft Translator Hub*[4]

L'entreprise *Microsoft* est active dans le domaine de la recherche en TA depuis le début des années 2000. Au début, son objectif en matière de TA était de mettre au point des systèmes pour traduire les documents internes à l'entreprise, mais en 2007, *Microsoft* a lancé *Windows Live Translator*, un outil destiné au grand public. Les systèmes dont disposait alors *Microsoft* étaient spécialisés pour traduire dans des domaines très techniques et n'étaient pas encore vraiment adaptés à un usage par le grand public. Avec son premier système lancé en 2007, *Microsoft* proposait la traduction de et vers 7 langues (allemand, espagnol, français, italien, portugais, chinois et japonais). Le système supportait aussi d'autres langues, mais la traduction était alors fournie pas *Systran*.

Dans les années qui suivent, *Microsoft* développe un grand nombre d'outils de TA destinés au grand public, il intègre notamment la TA à son moteur de recherche et à son application de messagerie instantanée ainsi qu'à la suite *Office*. A partir de 2008, *Microsoft* fournit lui-même les traductions sur tous ses services (plus aucune traduction n'est fournie par *Systran*). Aujourd'hui, les différentes applications de TA proposées par *Microsoft* (*Bing, Word Translator, Tranlator Live Conversations…*) sont basées sur un même système nommé *Microsoft Translator.* Ce dernier est un système de traduction statistique, mais *Microsoft* ne donne pas d'informations précises sur son mode de fonctionnent.

En 2012, *Microsoft* lance une extension à *Microsoft Translator* : le *Microsoft Translator Hub* (MTH). Cette extension permet à l'utilisateur d'entrainer très simplement son propre système de traduction statistique, ce qui présente deux principaux avantages : tout d'abord, l'utilisateur peut adapter son système pour obtenir de meilleurs résultats dans un domaine précis, ensuite, il peut entrainer un système pour une nouvelle paire de langues qui n'est pas encore supportée par les outils de TA proposés. Ce second aspect peut permettre de redonner vie à des langues peu utilisées et peu présentes sur internet. Le fonctionnement de MTH a quelque peu évolué au cours du temps, mais nous nous contenterons de présenter ici son fonctionnement actuel.

Comme nous l'avons dit, MTH est une application qui permet d'entrainer son propre système de traduction. Une fois ce système créé, il peut être utilisé avec les différents produits de Microsoft qui utilisent le *Microsoft Translator API* ou intégré à d'autres logiciels (comme des

[4] Les sources principales de cette section sont le Microsoft Translator Blog (Microsoft 2018b), le Microsoft Translator Hub User Guide (Microsoft 2018d) et le site Microsoft Translator Hub (Microsoft 2018c).

logiciels de TAO). Comme nous l'avons vu au point précédent, l'entrainement d'un système statistique consiste à créer des modèles de langues et de traduction à partir de corpus.

MTH propose différentes options pour l'entrainement du système, l'utilisateur peut entrainer un système spécialisé en ajoutant simplement un glossaire (il n'y a pas de taille minimum requise pour le glossaire), ou en utilisant un corpus bilingue de seulement 1 000 phrases. Enfin, l'utilisateur peut réaliser un entrainement complet (full training) en utilisant des données d'entrainement, de test et de tuning. Nous allons présenter plus en détail l'entrainement complet, car c'est ce type d'entrainement que nous avons réalisé dans le cadre de notre projet. Afin d'illustrer les explications données ci-dessous, voici un aperçu de l'interface d'entrainement de MTH (Figure 3) :

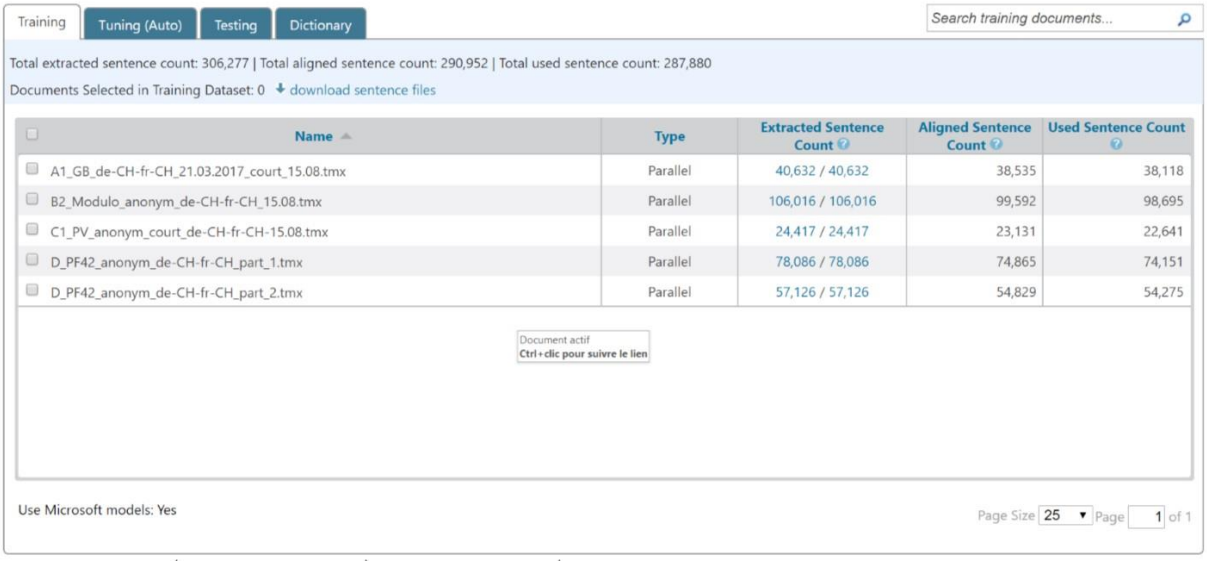


Figure 3 - *Interface d'entrainement de MTH (mars 2018)*

MTH permet d'ajouter plusieurs types de données pour la réalisation d'un entrainement complet. L'interface dispose d'un onglet pour chaque type de données : ***Training***, ***Tuning***, ***Testing*** et ***Dictionary***.

Training

Dans l'onglet *Training*, l'utilisateur peut ajouter des données d'entrainement, qui peuvent être des corpus bilingues ou monolingues en langue cible. Pour que l'entrainement fonctionne, MTH recommande d'avoir un corpus bilingue d'au moins 10 000 phrases (même si l'entrainement peut être réalisé avec seulement 1000 segments) ; concernant les données monolingues, il n'y

a pas de minimum indiqué (l'entrainement peut être réalisé sans corpus monolingue). L'onglet *Training* permet aussi à l'utilisateur de choisir s'il souhaite utiliser les modèles de Microsoft pour son système ou s'il souhaite utiliser uniquement les modèles générés par ses données.

Tuning

L'utilisateur est ensuite invité à ajouter des données dans l'onglet *Tuning*, ces données doivent être bilingues. Le tuning permet d'ajuster les paramètres du système afin d'optimiser les résultats. Selon MTH, le tuning aurait une influence importante sur la qualité de la traduction. MTH recommande d'utiliser un corpus de tuning comprenant 2 000 à 2 500 phrases. L'influence du tuning étant grande, il est conseillé de sélectionner manuellement les données de tuning afin d'avoir une sélection de phrases représentatives des contenus que le système devra traduire. Si l'utilisateur n'ajoute pas de données de tuning, MTH enlève automatiquement 2 500 phrases du corpus d'entrainement et les utilise comme corpus de tuning.

Testing

Dans l'onglet *Testing*, l'utilisateur peut ajouter un corpus de test (bilingue), le système utilisera ce corpus pour évaluer la qualité du système à la fin de l'entrainement. Pour ce faire, il calculera le score BLEU obtenu par le système pour la traduction de ce corpus de test. Si l'utilisateur n'ajoute pas de corpus de test, MTH procédera comme pour le tuning et retirera 2 500 phrases des données d'entrainement pour les utiliser comme corpus de test.

Dictionary

Enfin, l'utilisateur peut ajouter de la terminologie via l'onglet *Dictionary* (l'utilisation du dictionnaire pour l'entrainement est facultative). Il est important de noter que les mots et les séquences de mots présents dans le dictionnaire se verront toujours assigner une probabilité de 100%.

Une fois toutes les données ajoutées, l'utilisateur peut lancer l'entrainement du système. L'entrainement peut prendre plusieurs heures en fonction de la quantité de données ajoutées. Une fois l'entrainement terminé, MTH donne le BLEU score obtenu par le système sur le corpus de test, l'utilisateur peut aussi voir les traductions produites par le système sur ce corpus. S'il est satisfait par le système, il peut demander son déploiement (le système devient alors opérationnel) ou alors il peut réaliser d'autres entrainements en modifiant des paramètres afin de comparer les performances de plusieurs systèmes.

MTH à l'avantage d'offrir une certaine flexibilité à l'utilisateur, qui peut entrainer plusieurs fois un même système en modifiant certains paramètres. Il peut par exemple modifier la taille du corpus d'entrainement, utiliser ou non les modèles de Microsoft, utiliser un dictionnaire ou non et voir quelle est l'influence de chacun de ces paramètres sur la qualité de son système.

4.2. La traduction automatique neuronale et DeepL

4.2.1. La traduction automatique neuronale

Comme nous l'avons mentionné précédemment, le fonctionnement d'un système de traduction neuronal repose sur l'utilisation de corpus. Il comporte un encodeur et un décodeur qui sont composés de réseaux de neurones. Comme les systèmes statistiques, les systèmes neuronaux sont entrainés avec des corpus bilingues (des corpus monolingues peuvent aussi être utilisés). La structure d'un système neuronal est plus simple que celle d'un statistique. Le système neuronal possède un seul modèle et n'a pas de modèles séparés pour la langue, la traduction ou le réordonnancement (Koehn, 2018). Le rôle de l'encodeur est de donner une représentation de la phrase source tandis que le rôle du décodeur est de prédire la phrase cible en utilisant cette représentation (Koehn, 2017). Ce mode de fonctionnement rappelle celui des systèmes par interlangues décrits dans la section 2.6.1, puisque la traduction est réalisée en utilisant une représentation intermédiaire du sens de la phrase. Cependant, le fonctionnement des systèmes neuronaux est bien plus complexe.

L'un des éléments clés des systèmes de TA neuronale est la représentation intermédiaire générée par l'encodeur et utilisée ensuite par le décodeur, cette représentation prend la forme d'un **plongement lexical** (*word embedding* en anglais) (Koehn, 2018). Un plongement lexical est une représentation du sens des mots qui se base sur l'idée que des mots qui apparaissent dans le même contexte sont similaires (Koehn, 2017). Le plongement lexical est un espace multidimensionnel dans lequel chaque mot est représenté par un vecteur (Koehn, 2017). Les mots qui ont des propriétés communes vont se retrouver proches dans une certaine dimension (Systran, 2016). Par exemple, tous les verbes vont être proches dans une dimension donnée, mais les verbes d'état se trouveront aussi proches dans une autre dimension (et donc éloignés des verbes d'action dans cette seconde dimension). Voici une représentation d'un plongement lexical en deux dimensions, qui permet de mieux comprendre ce fonctionnement (Figure 4) :

Figure 4 - *représentation d'un plongement lexical en 2 dimensions.*
Adaptation de Koehn (2017 : 36)

Dans le plongement lexical présenté ci-dessus, on voit que les mots qui ont des similitudes sémantiques se trouvent proches les uns des autres. Le plongement lexical permet donc de regrouper des mots entre eux en fonction de leur sens et d'établir des généralités donnant ainsi au système de TA neuronale la capacité de traiter des séquences de mots nouvelles qui ne figurent pas dans le corpus d'entrainement (Koehn, 2017).

Les systèmes de TA neuronale offrent l'avantage d'être capables de prendre en compte un contexte très large lors de la traduction (bien plus large que les systèmes statistiques classiques), car ils disposent d'une certaine flexibilité lorsqu'ils rencontrent des énoncés inconnus (Koehn, 2017). Si le système est face à une séquence de 5 mots qui n'est pas présente telle qu'elle dans le corpus, par exemple : « le chat mange des croquettes », mais que la phrase « le chien mange des croquettes » est dans le corpus, le système pourra traiter cette phrase, car la proximité des mots « chat » et « chien » dans le plongement lexical lui permet de traiter cette phrase. Nous avons vu dans la section 4.1. que la prise en compte de n-grammes plus longs permet en général d'améliorer la fluidité de l'output d'un système.

Les systèmes neuronaux présentent aussi certains inconvénients. Tout d'abord, contrairement à certains systèmes statistiques comme MTH, les systèmes neuronaux accessibles en ligne ne peuvent pour le moment pas être spécialisés en fonction des données de l'utilisateur. Un autre problème majeur des systèmes neuronaux réside dans la puissance computationnelle qu'ils

requièrent. Cela limite grandement la taille du vocabulaire de ces systèmes (Koehn, 2018). De plus, l'entrainement des systèmes neuronaux nécessite d'avoir une grande quantité de données pour obtenir un résultat satisfaisant. La complexité des systèmes fait qu'il est difficile d'identifier et de corriger la source de certaines erreurs. Comme il est plus facile d'intervenir sur un SMT, il est plus facile de corriger certains aspects du système pour en améliorer l'output (Koehn, 2018).

4.2.2. DeepL[5]

DeepL est un service de TA lancé publiquement en août 2017 par l'entreprise *Linguee*, qui est aussi la société qui a créé, en 2010, le moteur de recherche de traduction du même nom. *DeepL* est un service de TA en ligne qui supporte 7 langues (français, anglais, allemand, espagnol, italien, néerlandais et polonais) et 42 combinaisons (au 1er mai 2018). Ce système de TA repose sur des réseaux de neurones, ces réseaux fonctionnent sur un superordinateur basé en Islande d'une puissance de 5,1 pétaFLOPS (à titre de comparaison, la puissance moyenne d'un ordinateur grand public est de 0,0001 pétaFLOPS, soit environ 50 000 fois moins (L'Express, 2016). Les réseaux de neurones sont entrainés avec des corpus de traduction collectés sur le web à l'aide du même algorithme que celui qui collecte les traductions pour le moteur de recherche *Linguee*. Dès son lancement, *DeepL* affirme que les résultats de son système de TA surpassent ceux des autres systèmes tels que *Google Translate* (qui fait aussi appel à des réseaux de neurones pour certaines paires de langues) ou MTH. Des évaluations humaines et automatiques conduites par *DeepL* montrent une nette préférence des traducteurs humains pour *DeepL* ainsi qu'un score BLEU nettement supérieur par rapport à d'autres systèmes. A l'heure actuelle, *DeepL* ne donne pas de détail concernant l'architecture de ses réseaux de neurones, mais explique que la qualité de ses résultats est due à un nouvel aménagement des neurones et de leurs connexions. A son lancement, *DeepL* était accessible uniquement en ligne, mais en mars 2018 la société a lancé *DeepL Pro*, un service qui permet d'intégrer *DeepL* à différentes applications et notamment à des logiciels de TAO tels que *SDL Trados Studio 2017*. *DeepL* s'adresse donc désormais aussi aux traducteurs professionnels qui souhaitent utiliser la TA.

[5] La source principale de cette section est le site de DeepL (DeepL 2018a)

Conclusion

Dans ce chapitre, nous avons expliqué le fonctionnement général des systèmes de TAS. Nous avons aussi présenté le fonctionnement et l'interface de MTH ainsi que les différentes options dont dispose l'utilisateur pour entrainer son propre système. Nous avons tenté d'expliquer, dans ses grandes lignes, le fonctionnement complexe des outils de TAN. Enfin, nous avons présenté le second outil que nous avons utilisé dans le cadre de ce projet, à savoir, le système *DeepL*.

5. ENTRAINEMENT ET ÉVALUATION DES SYSTÈMES MTH

Dans ce chapitre, nous allons décrire le travail de spécialisation et d'évaluation de MTH que nous avons réalisé dans le cadre de notre projet. Nous allons tout d'abord présenter les données d'entrainement ainsi que les corpus de test que nous avons utilisés (section 5.1.1). Nous détaillerons ensuite l'ensemble des systèmes que nous avons entrainés (section 5.1.2). Dans une deuxième partie, nous décrirons notre processus d'évaluation destiné à sélectionner le meilleur nos systèmes MTH (section 5.2.2). Nous présenterons enfin notre méthode d'évaluation pour le choix du domaine sur lequel portera la comparaison que nous allons effectuer entre MTH et *DeepL* (section 5.2.3).

5.1. Entrainement

5.1.1. Données d'entrainement

Dans le cadre de ce projet, La Poste Suisse nous a fourni certaines de ses mémoires de traduction afin que nous puissions créer des corpus d'entrainements pour MTH ainsi que des corpus de test pour évaluer nos systèmes. L'entreprise nous a aussi fourni des glossaires. Ces mémoires de traduction et ces glossaires sont issus de quatre grands domaines dans lesquels s'inscrivent les documents traduits par La Poste. La mémoire nommée GB contient les unités de traduction du rapport annuel de l'entreprise ; la mémoire Modulo contient celles de documents de formation, la mémoire PV, celles des manuels destinés aux offices de poste et enfin, la mémoire PF42, celles des documents de PostFinance. Pour chacun de ces domaines, La Poste nous a fourni un ou plusieurs glossaires. Le tableau ci-dessous (Tableau 1) récapitule l'ensemble des données mises à disposition par La Poste :

Nom de la ressource	Domaine	Type	Nbre de segments ou d'entrées	Statut
GB	Rapport annuel	MT	42 863	Public
Modulo	Formation	MT	101 781	Interne
PF42	PostFinance	MT	135 212	Interne
PV	Manuels offices de poste	MT	25 147	Interne
D1 (GB)	Rapport annuel	Glossaire	142	n/a
D2 (Modulo)	Formation	Glossaire	124	n/a
D3 (PF42)	PostFinance	Glossaire	484	n/a
D4 (PV)	Manuels offices de poste	Glossaire	1440	n/a

Tableau 1 – *Liste des ressources fournies par La Poste*

Corpus d'entrainement

La mémoire GB est une mémoire publique, c'est-à-dire qu'elle contient uniquement des données qui sont publiées par La Poste, nous pouvons donc l'utiliser telle qu'elle. En revanche, les autres mémoires de traduction contiennent des informations internes à l'entreprise et peuvent contenir des données confidentielles. Comme nous allons utiliser ces mémoires pour entrainer MTH, elles seront mises en ligne sur le serveur de Microsoft et ce dernier n'offre pas de garanties en matière de confidentialité. La Poste nous a donc demandé d'anonymiser ces mémoires en supprimant certaines informations (les noms, les prénoms, les adresses, les numéros de téléphone, les adresses e-mail, les numéros de compte, ainsi que les noms des organisations et des entreprises). Ces mémoires étant très grandes, cette tâche peut s'avérer longue et fastidieuse, nous avons donc tenté d'automatiser au maximum ce processus.

Nous avons tout d'abord utilisé l'outil d'extraction des entités nommées de Stanford (*Stanford Named Entity Recognizer* ou *Stanford NER*)[6] afin d'établir une liste de tous les noms propres présents dans les mémoires de traduction. Entrainé sur de larges corpus, cet outil repère les entités nommées et les classe en quatre catégories : les noms de lieux (LOC), les noms de

[6] Pour plus d'informations sur cet outil consulter la page https://nlp.stanford.edu/software/CRF-NER.html#Extensions

personnes (PER), les noms d'organisation (ORG) et les entités nommées non définies (MISC). Après avoir utilisé le *Stanford NER* pour extraire les entités nommées, nous avons dû nettoyer ces listes, d'une part, car l'outil fait un certain nombre d'erreurs et d'autre part, car nous devions assigner une catégorie aux entités « MISC ». Nous avons utilisé l'outil *OpenRefine*[7] pour nettoyer nos listes. Après avoir nettoyé les listes d'entités nommées, nous disposions de listes des noms de personnes, des noms d'organisation et des noms de lieux présents dans les mémoires. Nous avons utilisé un script afin de remplacer automatiquement ces entités respectivement par « NameX », « OrgX » et « AdressX » dans les mémoires PV, Modulo et PF42. L'anonymisation des numéros de téléphone, des adresses e-mail et des numéros de compte bancaire était plus simple, car ces entités sont repérables grâce à des expressions régulières. Nous avons donc utilisé un script qui repérait et remplaçait les suites de *n* chiffres ainsi que les éléments de type xxx@xxx.xx.

Nous avons ensuite vérifié que les mémoires de traduction étaient correctement anonymisées en procédant à la relecture d'un certain nombre d'échantillons de ces mémoires (nous ne pouvions bien sûr pas vérifier l'intégralité de celles-ci). Nous avons aussi demandé à La Poste de procéder à des vérifications et de nous donner son accord pour l'utilisation des mémoires.

Terminologie

Nous avons aussi effectué un nettoyage des glossaires en nous basant notamment sur les bonnes pratiques en matière de glossaires données par MTH[8]. Nous avons cherché à privilégier les noms composés et les expressions ; nous avons ôté les adjectifs, les verbes ainsi que les termes polysémiques. Le tableau ci-dessous (Tableau 2) récapitule le nombre d'entrées restantes pour chaque domaine après nettoyage des glossaires :

[7] OpenRefine est un outil gratuit de gestion de données. Pour plus d'informations, consulter http://openrefine.org/
[8] Voir *Microsoft Translator Hub User Guide* Section 3.3.2

Glossaire (domaine)	Nombre d'entrées
D1 (GB)	76
D2 (Modulo)	144
D3 (PF42)	641
D4 (PV)	1356

Tableau 2 - *Glossaires fournis par la Poste après nettoyage*

Corpus de test

Après avoir nettoyé les données d'entrainement, nous avons préparé des corpus de test qui allaient nous servir à évaluer les différents systèmes lors des entrainements de MTH. Microsoft conseille d'avoir un minimum de 2000 segments alignés pour le corpus de test, nous avons donc fait en sorte de créer des corpus de test contenant un peu plus de 2000 segments chacun (nous en avons pris plus, car il arrive que Microsoft élimine certains segments lors de l'alignement). Nous avons créé un corpus de test pour chaque domaine, ainsi qu'un corpus de test comprenant des segments issus des quatre domaines (corpus mixte). Pour créer ces corpus, nous avons tout d'abord extrait tous les segments ajoutés aux différentes mémoires à une date ultérieure à celle à laquelle les mémoires que nous avions avaient été exportées. Pour certains domaines, cela nous a permis d'avoir directement plus de 2000 segments. Lorsque le nombre de nouveaux segments était important, nous avons ajouté les nouveaux segments dont nous n'avions pas besoin dans le corpus de test au corpus d'entrainement. Pour les domaines dans lesquels il y avait moins de 2000 nouveaux segments, nous avons simplement retiré des segments des corpus d'entrainement pour les ajouter aux corpus de test. Nous avons anonymisé ces corpus de test de la même manière que les corpus d'entrainement. Les tableaux ci-dessous (Tableaux 3 et 4) indiquent le nombre de segments de chacun de nos corpus de test et de nos corpus d'entrainement :

Domaine (corpus de test)	Nombre de segments
GB	2230
Modulo	2159
PV	2237
PF42	2712
Mixte	2717

Tableau 3 - *Corpus de test et nombre de segments de chaque corpus*

Domaine (corpus d'entrainement)	Nombre de segments
GB	40 632
Modulo	106 016
PV	24 417
PF42	135 212

Tableau 4 - *Corpus d'entrainement et nombre de segments pour chaque corpus*

5.1.2. Différents systèmes entrainés

Une fois nos données d'entrainement et de test préparées, nous avons pu procéder à l'entrainement de différents systèmes MTH. Notre objectif étant de tester différents scénarios d'entrainement pour nos 4 domaines, nous avons réalisé un certain nombre d'entrainements en agissant sur les paramètres tels que l'ajout de glossaire ainsi que la quantité et la spécialisation des données d'entrainement.

Nous avons tout d'abord entrainé un système séparé pour chaque domaine sans utiliser les glossaires, puis en les utilisant. Pour ces systèmes, nous avons à chaque fois utilisé les corpus de test correspondant au domaine. Nous avons ensuite entrainé un système en utilisant tous les corpus d'entrainement (ci-après « Système Général »), nous avons cloné ce système à 8 reprises et afin de pouvoir utiliser nos 4 corpus de test avec ce système et nous avons réalisé à chaque fois un entrainement avec et sans glossaires. Pour chacun de nos entrainements, nous avons utilisé les modèles de Microsoft et nous avons laissé MTH sélectionner automatiquement le corpus de tuning. Le tableau ci-dessous (Tableau 5) présente l'ensemble des tests réalisés avec, à chaque fois, le nombre de segments utilisés par MTH pour l'entrainement et pour le test.

Système	Données d'entrainement	Segments utilisés (training)	Terminologie	Entrées utilisées	Test	Segments utilisés (testing)
Système Général	GB-PV-PF42-Modulo	288 211	/	/	GB	2053
Système Général	GB-PV-PF42-Modulo	288 211	G1	76	GB	2053
Système Général	GB-PV-PF42-Modulo	287 780	/	/	PV	2006
Système Général	GB-PV-PF42-Modulo	287 780	G3	1321	PV	2006
Système Général	GB-PV-PF42-Modulo	287 993	/	/	PF42	2500
Système Général	GB-PV-PF42-Modulo	287 976	G4	622	PF42	2500
Système Général	GB-PV-PF42-Modulo	287 894	/	/	Modulo	2019
Système Général	GB-PV-PF42-Modulo	287 894	G2	130	Modulo	2019
Système GB	GB	36 585	/	/	GB	2053
Système GB	GB	36 585	G1	76	GB	2053
Système PV	PV	21 534	/	/	PV	2002
Système PV	PV	21 534	G3	1321	PV	2002
Système PF42	PF42	126 758	/	/	PF42	2500
Système PF42	PF42	126 758	G4	622	PF42	2500
Système Modulo	Modulo	96 612	/	/	Modulo	2019
Système Modulo	Modulo	96 612	G2	130	Modulo	2019

Tableau 5 - *Liste des entrainements réalisés avec MTH et détails des données d'entrainement et de test*

Nous constatons que le nombre de segments utilisés par MTH est inférieur au nombre de segments présents dans nos corpus et qu'il varie parfois d'un système à l'autre alors que le corpus est le même. Cette variation est due à deux facteurs : d'une part, lorsque l'on ajoute les corpus sur la plateforme, MTH procède à un alignement des segments et il arrive qu'il ne parvienne pas à aligner tous les segments (cela peut notamment se produire si des segments ne

sont pas bien alignés dans le corpus)[9]. D'autre part, lors de l'entrainement MTH exclut des corpus les segments qui sont aussi présents dans les données de tuning et/ou de testing[10].

5.2. Evaluation de MTH et choix du meilleur système et du domaine

5.2.1. Méthodologie

Dans la section précédente, nous avons présenté les différents systèmes entrainés avec MTH. Ayant entrainé un grand nombre de systèmes, nous ne disposions pas du temps et des ressources nécessaires à la réalisation d'une évaluation humaine pour chaque système et chaque domaine. Nous avons donc décidé de procéder à une évaluation automatique pour choisir le meilleur système. Pour cette évaluation automatique, nous avons utilisé la métrique BLEU (Papineni et al., 2002) décrite dans la section 3.2. Nous avons calculé le score BLEU sur les corpus de test utilisés lors des entrainements de nos systèmes. Une fois l'entrainement terminé, MTH permet à l'utilisateur de télécharger la traduction du corpus produite par le système. Nous avons donc téléchargé ces traductions et calculé les scores BLEU avec la source et la référence de chaque corpus. Nous avons calculé les scores BLEU à l'aide du script *mteval-v13a.pl*[11].

5.2.2. Choix du meilleur système

Le tableau ci-dessous (Tableau 6) présente les scores BLEU (classés par ordre croissant) obtenus par chaque système et pour chaque domaine .

[9] Voir *Microsoft Translator Hub User Guide* Section 2.7
[10] Voir *Microsoft Translator Hub User Guide* Section 3.3.4

[11] Disponible sur ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl (avril 2018)

Système	Terminologie	Corpus de test	BLEU mteval
Système Général	G1	GB	41,36
Système Général		GB	41,3
Système Général	G3	PV	40,28
Système Général		PV	40,02
Système GB		GB	39,53
Système GB	G1	GB	39,39
Système PV	G3	PV	36,89
Système PV		PV	36,87
Système Général	G4	PF42	34,16
Système PF42		PF42	34,14
Système PF42	G4	PF42	34,05
Système Général		PF42	33,65
Système Général	G2	Modulo	28,64
Système Général		Modulo	28,34
Système Modulo	G2	Modulo	28,34
Système Modulo		Modulo	27,99

Tableau 6 - *Scores BLEU calculés avec mteval_v13 pour les systèmes entrainés avec MTH*

Nous voyons que le système général obtient un meilleur score BLEU que les systèmes spécifiques pour presque chaque domaine. Il n'y a que pour le domaine PF42 que le système général (sans terminologie) obtient un score BLEU inférieur à celui des systèmes spécifiques à ce domaine. Le système spécifique Modulo obtient quant à lui le même score BLEU que le système général sans terminologie pour ce même domaine. Les systèmes généraux (avec et sans terminologie) sont donc meilleurs que les systèmes spécifiques pour la plupart des domaines. En ce qui concerne l'influence de la terminologie, nous constatons que le système général avec terminologie obtient un meilleur score BLEU pour tous les domaines. L'écart de score BLEU reste cependant assez limité, ce qui pourrait s'expliquer par le fait que les glossaires que nous avons utilisés ne contiennent pas un grand nombre de termes.

D'après les scores BLEU obtenus pour chaque domaine, nous pouvons en déduire que le système général avec terminologie donne de meilleurs résultats que les systèmes spécifiques. Nous avons donc décidé d'utiliser ce corpus pour la suite de notre travail.

5.2.3. Choix du domaine

Maintenant que nous avons choisi le meilleur de nos systèmes, il nous faut sélectionner le domaine pour lequel notre système général avec terminologie est le meilleur. Pour ce faire, nous nous sommes aussi intéressé aux scores BLEU donnés par MTH, en plus de ceux que nous avions calculés avec mteval-v13. Le tableau ci-dessous (Tableau 7) reprend les scores BLEU obtenus par ce système pour chaque domaine :

Système	Terminologie	Corpus de test	BLEU mteval	BLEU MTH
Système Général	G1	GB	41,36	32,7
Système Général	G3	PV	40,28	48,39
Système Général	G4	PF42	34,16	34,26
Système Général	G2	Modulo	28,64	35,22

Tableau 7 - *Scores BLEU obtenus avec mteval-v13a et MTH pour chaque domaine*

Nous constatons que c'est pour le domaine GB que notre système général avec terminologie obtient le meilleur score BLEU avec mteval-v13. Pour le domaine PV, le système obtient un score BLEU assez proche ; en revanche, les scores BLEU sont nettement moins bons pour les domaines PF42 et Modulo. Lors de la première évaluation automatique de nos systèmes, nous avions constaté que les scores BLEU donnés par MTH avaient tendance à être plus hauts que ceux donnés par mteval-v13. Nous avions décidé de ne pas tenir compte de ces scores, car MTH ne détaille pas la manière dont son score BLEU est calculé. Il nous semblait donc plus judicieux de baser notre évaluation sur un script issu du domaine de la recherche.

Un élément a cependant attiré notre attention : alors que pour tous les systèmes et tous les domaines, le score donné par MTH est généralement plus élevé et donne lieu au même classement des systèmes que celui donné par mteval-v13, ce n'est pas le cas pour le domaine GB. Pour ce domaine, MTH attribue au système général le score BLEU le plus faible alors que mteval-v13 lui attribue le meilleur. Pour tenter de comprendre pourquoi il en était ainsi, nous avons décidé de calculer à nouveau les score BLEU en utilisant un autre outil. Nous avons opté

pour l'*Interactive BLEU score evaluator* de la plateforme Tilde Custom Machine Translation[12]. Voici les scores BLEU que nous avons obtenus (Tableau 8) :

Système	Terminologie	Corpus de test	BLEU mteval	BLEU MTH	BLEU Tilde[13]
Système Général	G1	GB	41,36	32,70	22,30
Système Général	G3	PV	40,28	48,39	40,20
Système Général	G4	PF42	34,16	34,26	28,48
Système Général	G2	Modulo	28,64	35,22	28,41

Tableau 8 - *Scores BLEU obtenus avec mteval-v13a, MTH et Tilde pour le système général dans chaque domaine*

Là encore, le système général obtient le plus mauvais score pour le domaine GB alors que le reste du classement reste le même (bien que le domaine PF42 obtienne un score sensiblement moins bon qu'avec mteval-v13). Nous avons donc cherché d'où pouvait provenir cet écart de score pour le domaine GB. L'outil de Tilde était idéal pour cela, car il permet de voir le score BLEU de chaque segment. Il présente aussi pour chaque segment la source, la référence et la TA en mettant en évidence les différences entre référence et TA. L'outil *mteval* offre aussi la possibilité de voir le score de chaque segment. Nous avons donc cherché des segments dont le score *mteval* différait fortement du score Tilde afin d'identifier la raison de cette différence. En procédant ainsi, nous avons très rapidement remarqué que les scores BLEU différaient uniquement pour les segments comportant des balises de type « <a1> » ; « </a1> » etc. *mteval* semble compter comme correctes les balises présentes dans la référence et dans la TA, et ce, même si elles sont séparées par des espaces dans la TA et pas dans la référence, alors que Tilde (et probablement MTH) compte les balises comme fausses lorsqu'elles sont séparées par une espace dans la TA et non dans la référence. Sur des segments assez courts, le fait de compter comme justes ou comme fausses une ou deux balises peut avoir un impact très important sur le score BLEU. Nous avons par exemple vu des segments qui avaient un score BLEU de 100 dans *mt-eval* et de seulement 37, 24, voire même 5 dans Tilde. Nous avons aussi remarqué que les

[12] Disponible à l'adresse https://www.letsmt.eu/Bleu.aspx (2018)
[13] Sensible à la casse

scores BLEU différaient pour les segments contenant des symboles tels que « % » ou « = » lorsque ces derniers sont suivis ou précédés d'une espace dans la référence et non dans la TA ou l'inverse.

Nous avons déduit de ces observations que la différence entre les scores BLEU de *mteval* et ceux de Tilde est due à une tokenisation[14] différente des symboles. Le script *mteval* précise en effet qu'il tokenise la ponctuation (sauf si elle est précédée et suivie de chiffres) et les symboles. Nous sommes donc arrivé à la conclusion que Tilde (et probablement MTH) ne tokenise pas les symboles tels que « % », « = » ou « > ». Voici un exemple qui illustre cette hypothèse (Figure 5) :

Sentence 434	BLEU	Length ratio	Text
Source	-	-	Die Bedürfnisse der Kundinnen und Kunden verändern sich.
Human	100.00	1.00	Les besoins des clients évoluent< / a1>< / a2>< / a3> . < / a4>
Machine	42.09	1.20	Les besoins des clients évoluent< / a1> < / a2> < / a3> . < / a4>

Figure 5 - *BLEU score donné par Tilde pour le segment 434 du corpus de test GB.*

Ici Tilde donne un score BLEU de 42,09 et considère comme fausses les unités suivantes de la TA « a1> », « < », « a2> » et « a3> » car elles ne sont pas présentes telles qu'elles dans la référence (en réalité, « a3> » devrait être correcte). Or, *mteval* attribue un score de 100 à ce segment et comptabilise 31 mots. On peut donc supposer que ce score de 100 s'explique par le fait que *mteval* a tokenisé, c'est-à-dire séparé par une espace, tous les symboles (« < », « / » et « > ») et qu'il les a donc tous considérés comme justes.

Pour résumer, Tilde à tendance à pénaliser les symboles lorsqu'ils ne sont pas reproduits dans la TA exactement comme dans la référence, ce qui entraine naturellement une pénalisation des corpus de test qui contiennent un grand nombre de symboles, tandis que *mteval* accorde moins d'importance aux symboles.

A ce stade, il nous fallait donc décider à quelle méthode de calcul de BLEU nous allions nous fier pour sélectionner le domaine sur lequel nous allions effectuer notre évaluation humaine

[14] La tokenisation consiste à isoler chaque mot du texte, dans le cas des langues latines il s'agit surtout de séparer les mots des signes de ponctuation en ajoutant des espaces (Koehn, 2010)

(mteval ou Tilde/MTH). Nous estimons qu'en TA, les mots sont plus importants que les symboles, car c'est la correction des mots qui prendra le plus de temps au post-éditeur (la correction des symboles comme <a1> pouvant être faite de manière automatique, par exemple en utilisant la fonction rechercher/remplacer d'un outil de TAO). Il nous semblait donc plus logique de nous fier aux scores donnés par *mteval*. Un autre exemple tiré de notre corpus nous a conforté dans cette idée (Figure 6) :

Sentence 442	BLEU	Length ratio	Text
Source	-	-	Um diese Investitionen tätigen zu können, müssen wir solide Gewinne erzielen.
Human	100.00	1.00	Pour pouvoir réaliser ces < / a0>investissements< / a1>< / a2> , nous devons enregistrer de solides bénéfices .
Machine	10.90	1.21	Pour ces < / a0> investissements< / a1> < / a2> de pouvoir , des , nous doivent réaliser des bénéfices solides .

Figure 6 - *Score BLEU donné par Tilde pour le segment 442 du corpus de test GB*

Pour ce segment, le score donné par Tilde est de 10,90 alors que celui donné par *mteval* est de 68,38. En plus des balises, Tilde compte comme faux le mot « investissement » alors que selon nos hypothèses, mteval le considère comme juste. Dans ce cas de figure, le post-éditeur aurait juste à supprimer une espace pour que la TA corresponde à la référence ce qui ne demande pas un gros effort de post-édition. Il nous semble donc qu'ici Tilde pénalise un peu « injustement » ce segment du fait de son mode de tokenisation. Considérant cela, nous avons décidé de nous baser sur les scores BLEU donnés par mteval pour réaliser notre évaluation.

Pour rappel, c'est pour le domaine GB que notre système général obtenait les meilleurs résultats avec *mteval*, nous avons donc sélectionné ce domaine pour procéder à l'évaluation humaine. Pour la suite de ce travail, nous évaluerons donc le système général avec terminologie pour le domaine GB.

Conclusion

Dans cette partie, nous avons procédé à une évaluation automatique de nos systèmes en deux étapes. La première étape visait à sélectionner le meilleur de nos systèmes MTH, tandis que la seconde étape avait pour but de déterminer le domaine pour lequel notre meilleur système était le plus performant.

Les systèmes généraux (avec ou sans terminologie) obtiennent presque systématiquement de meilleurs scores BLEU que les systèmes entrainés uniquement avec les données d'un domaine spécifique. La quantité de données semble donc avoir un impact sur la qualité de nos systèmes. L'ajout de glossaire a globalement un impact positif sur le score BLEU, sauf dans deux cas, où elle fait baisser le score. Dans tous les cas, la variation du score BLEU due à l'ajout de terminologie reste assez faible. On voit que le domaine du corpus de test entraine une variation bien plus importante. Tandis que la variation due à la terminologie ne dépasse jamais 1 point de BLEU, celle due au corpus de test peut atteindre jusqu'à 13 points de BLEU. Nous avons constater que le système général donne des scores nettement meilleurs pour les domaines GB et PV. Nous avons vu que ces scores peuvent largement varier en fonction de la manière dont le BLEU est calculé. Nous avons finalement décidé de nous fier au BLEU donné par *mteval-v13* et nous avons donc décidé de garder le système général et de conduire la prochaine évaluation sur le domaine GB.

6. COMPARAISON DE MTH ET DEEPL

Comme nous l'avons expliqué dans le chapitre 1, notre objectif est de comparer les performances d'un système statistique spécialisé (MTH) avec celles d'un système neuronal généraliste (*DeepL*) afin de déterminer *si les systèmes statistiques spécialisés sont en mesure de rivaliser avec les systèmes neuronaux généralistes lorsque la traduction automatique est utilisée par le traducteur professionnel comme outil d'aide à la traduction*. Pour répondre à cette première question, nous avons conduit une évaluation automatique et une évaluation humaine portant sur l'effort de PE. Nous avons aussi réalisé une évaluation humaine portant sur la qualité des traductions post-éditées de nos deux systèmes afin de voir si le système de TA utilisé avait une influence sur la qualité finale de la traduction. Nous débuterons ce chapitre en présentant notre évaluation automatique et ses résultats (section 6.1). Nous détaillerons ensuite le déroulement de nos deux évaluations humaines et nous en exposerons les résultats (section 6.2 et 6.3). Nous terminerons ce chapitre par une synthèse et une discussion des résultats de nos trois évaluations (section 6.4).

6.1. Evaluation automatique

6.1.1. Préparation de l'évaluation

Pour cette évaluation, nous avons préparé un nouveau corpus de test en exportant de nouveaux segments de la mémoire de traduction GB de La Poste. Cela nous a permis de créer un corpus de test de 1718 segments. Pour pouvoir traduire ce corpus avec nos deux systèmes de TA, nous avons d'abord dû déployer notre système général MTH avec terminologie (ci-après SG_term) et souscrire un abonnement à *DeepL Pro*. Nous avons ensuite traduit automatiquement notre corpus avec les deux systèmes en utilisant *SDL Trados 2017*. Pour MTH, nous avons utilisé le plug-in *MT Enhanced* qui nous a permis d'intégrer la TA produite par notre système à SDL. Nous avons alors utilisé l'option de pré-traduction pour traduire automatiquement notre corpus de test. Pour *DeepL*, nous avons procédé de la même manière, mais en utilisant le plug-in *DeepLMTProvider*

Il nous semble important de souligner ici que l'étape de pré-traduction en utilisant la TA prend un certain temps pour chaque système (30 à 40 minutes pour nos 1718 segments).

Nous avons ensuite généré les traductions cibles produites par MTH et par *DeepL*.

6.1.2. Evaluation

Après avoir traduit notre corpus de test avec *DeepL* et MTH, nous pouvions procéder à l'évaluation automatique. Nous avons à nouveau choisi d'utiliser la métrique BLEU (Section 3.2), que nous avons là aussi calculée à l'aide du script **mteval-v13.pl**.

Voici les résultats que nous avons obtenus (Tableau 9) :

Système	BLEU mteval
DeepL	25,23
MTH	23,46

Tableau 9 - *Résultats de l'évaluation automatique de DeepL et MTH (scores BLEU donnés par mt-eval*

En premier lieu, nous remarquons que le score BLEU obtenu par MTH pour ce corpus de test est largement inférieur à celui obtenu avec le corpus de test précédent (Section 5.2). Cet écart important nous montre que le corpus de test sélectionné influence grandement le score BLEU.

Avec cette évaluation automatique, nous remarquons que *DeepL* obtient un score BLEU légèrement meilleur que MTH, mais l'écart reste limité par rapport aux écarts de BLEU que nous avons pu constater dans la section 5.2.

Nous avons ensuite procédé à une évaluation humaine de ces deux systèmes afin de déterminer si la tendance exprimée par l'évaluation automatique est confirmée par les traducteurs.

6.2. Evaluation humaine 1 : effort de post-édition

6.2.1. Objectifs de l'évaluation humaine 1

Afin de comparer nos deux systèmes, nous avons décidé de conduire une évaluation visant à mesurer l'effort de post-édition nécessaire à la correction de la TA de chaque système. Comme nous l'avons mentionné dans la section 3.1, l'effort de PE peut être déterminé en mesurant le temps nécessaire à la PE, ainsi que le nombre de modifications effectuées par le post-éditeur. Notre choix d'opter pour ce mode d'évaluation humaine a été motivé par le contexte dans lequel le système le plus performant est destiné à être utilisé, à savoir pré-traduire les documents qui seront ensuite post-édités par les traducteurs. Nous cherchons donc à déterminer quel système est le plus utile pour le traducteur dans le contexte de la post-édition, en d'autres termes, pour quel système la post-édition demande-t-elle le moins de temps et le moins de modifications.

6.2.2. Participants à la tâche de PE

Une traductrice du Service linguistique de La Poste (ci-après **Traductrice 1**) s'est portée volontaire pour effectuer la PE. Cette tâche étant assez longue, il n'était pas possible de constituer une équipe de test plus grande au sein de La Poste. Afin de donner plus de poids à notre test, nous avons demandé à une autre traductrice (ci-après **Traductrice 2**) de participer à au test. Cette traductrice est une jeune diplômée avec la combinaison de langue DE>FR. La traductrice de La Poste avait reçu une brève formation à la PE dans le cadre d'un autre test conduit à La Poste. Nous avons fourni à nos deux traductrices un document reprenant les points essentiels de cette formation, ainsi que des consignes plus spécifiques à notre test (voir annexe B).

6.2.3. Préparation et déroulement du test de PE

Pour réaliser ce test, nous avons sélectionné de manière aléatoire 250 segments dans notre corpus de test de l'évaluation automatique (voir section 6.1). Nous avons choisi de conduire cette évaluation sur 250 segments, car nous étions soumis à une contrainte de temps qui ne nous permettait pas de réaliser une évaluation sur un plus grand nombre de segments. En outre, une étude conduite par Estrella et al. (2007) a montré qu'il est possible d'obtenir des résultats fiables avec des corpus de test relativement limités (env. 250 segments) lorsque l'on cherche à comparer des systèmes.

Nous avons décidé d'utiliser l'outil de TAO en ligne *MateCat*[15] pour conduire notre évaluation, car cet outil permet de mesurer le temps passé par le traducteur sur chaque segment et donne aussi un aperçu des modifications qu'il effectue à chaque fois. *MateCat* est un outil gratuit dans lequel l'utilisateur peut ajouter ses propres mémoires de traduction, mais aussi utiliser une grande mémoire de traduction publique, ainsi que la TA. La TA proposée par *MateCat* est fournie par *Google Translate*, *DeepL* et MTH, cependant l'utilisateur ne peut pas sélectionner un seul de ces fournisseurs et il n'a aucun moyen de savoir de quel fournisseur provient la TA qui lui est soumise. Nous ne pouvions donc pas utiliser la TA proposée par *MateCat* pour notre test. Nous avons donc procédé en créant des mémoires de traduction à partir des traductions automatiques de notre corpus de test que nous avions produites à l'aide de nos deux systèmes. Nous avons tout d'abord créé une mémoire de traduction avec comme texte source nos 250 segments sélectionnés aléatoirement et leurs traductions produites par *DeepL* (ci-après « MT_DeepL »), puis une mémoire de traduction avec la même source, mais avec comme cible la TA de MTH (ci-après « MT_MTH »). Afin de ne pas fausser notre test, nous ne souhaitions pas que les traducteurs post-éditent d'abord les segments d'un système, puis ensuite les segments de l'autre. Nous avons donc cherché un moyen de mélanger ces segments tout en étant ensuite capable de les trier à nouveau pour analyser les résultats du test. Pour ce faire nous avons procédé de la manière suivante :

- Nous avons trié par ordre alphabétique (en fonction de la source) les segments de la mémoire MT_DeepL ; nous les avons numérotés de 1 à 250 et avons extrait les segments 1 à 125 (ces segments constituent ce que nous appellerons le **corpus 1**)

[15] https://www.matecat.com/

- Nous avons effectué le même tri et la même numérotation avec la mémoire MT_MTH et nous en avons extrait les segments 125 à 250 (ces segments constituent ce que nous appellerons le **corpus 2**)
- Nous avons mélangé ces 250 segments de manière aléatoire et nous avons créé une nouvelle mémoire de traduction nommée MT_DeepL_MTH (cette mémoire contient le corpus traduit par *DeepL* et le corpus 2 traduit par MTH)
- Nous avons ensuite répété ces opérations en inversant DeepL et MTH et nous avons créé une mémoire nommée MT_MTH_DeepL (cette mémoire contient le corpus traduit par MTH et le corpus 2 traduit par *DeepL*)

Après le test, il nous suffira alors de trier à nouveau les segments par ordre alphabétique pour pouvoir identifier les traductions produites par MTH/*DeepL*.

Avec ces deux mémoires de traduction, nous pouvions alors soumettre le texte source à nos traducteurs une première fois en utilisant la mémoire MT_DeepL_MTH, puis une seconde fois en utilisant la mémoire MT_MTH_DeepL. Les traductrices post-éditeront donc deux fois 250 segments. En procédant avec des mémoires de traductions, les traducteurs verront s'afficher pour chaque segment une correspondance à 100 % (qui en est réalité une traduction automatique) qu'ils pourront post-éditer, mais ils n'auront aucun moyen de savoir de quel système provient cette TA. Nous avons donc créé deux projets de post-édition pour chaque traducteur (Projet_1 et Projet_2). Dans le projet 1, le traducteur post-éditera les 250 segments du corpus. La moitié de ces segments aura été traduite par MTH et l'autre moitié par *DeepL* (ils apparaitront dans un ordre aléatoire). Dans le projet 2, le traducteur fera la même chose, mais les systèmes de TA auront été inversés par rapport au projet 1. Pour chaque projet, nous avons sélectionné les paramètres suivants dans *MateCat* :

- Désactivation de la TA (afin d'éviter que des suggestions de TA soient proposées par *MateCat*)
- Désactivation de la mémoire de traduction publique (pour éviter que d'autres suggestions que les nôtres apparaissent)
- Désactivation de la mise à jour de nos mémoires (afin de s'assurer que notre mémoire ne soit pas modifiée).

Le tableau ci-dessous (Tableau 10) résume les données de chaque projet (les projets sont identiques pour les deux traductrices) :

Projet	Segments	Système de TA
Projet 1	1-125 (corpus 1)	MTH
	126-250 (corpus 2)	DeepL
Projet 2	1-125 (corpus 1)	DeepL
	126-250 (corpus 2)	MTH

Tableau 10 - *Récapitulatif des données utilisées pour le test de PE*

Nous avons aussi créé pour chaque traducteur un projet dit de « prise en main » contenant une vingtaine de segments issus d'un autre corpus avec une mémoire de traduction que nous avons créée avec *Google Translate*. Ce projet avait pour but de permettre aux traducteurs de se familiariser avec la PE dans *MateCat* avec de commencer le véritable test.

En procédant de cette manière, chaque traducteur donc va être amené à post-éditer deux fois des segments sources identiques traduits automatiquement par deux systèmes différents. Nous sommes conscient que cela introduit un biais dans notre étude puisqu'il est probable que les traducteurs soient influencés par le souvenir de la traduction du premier segment lors de la PE du second. Nous estimons cependant que nous aurions introduit un biais encore plus grand en choisissant de comparer MTH et *DeepL* sur des segments différents, étant donné que notre corpus de test est relativement petit. Nous avons tenté d'atténuer ce biais en divisant la PE en deux projets. Les segments identiques se trouvent dans des projets différents, ce qui permet de s'assurer que le traducteur ne tombera pas sur deux segments identiques à la suite ou très proches. Ce test étant assez long, nous avons aussi précisé aux traducteurs qu'ils pouvaient réaliser la PE en plusieurs fois et sur plusieurs jours (2 semaines en tout), ce qui permet encore d'atténuer l'éventuelle influence des traductions post-éditées précédemment. En outre, le biais est le même pour les deux systèmes, car nous avons procédé en réalisant un *crossover design*, ce qui signifie que les traductrices post-éditeront d'abord la TA de *DeepL* sur la moitié du corpus et d'abord la TA de MTH sur l'autre moitié du corpus.

Pour cette évaluation, nous avons demandé à nos post-éditrices d'effectuer la PE dans le but d'obtenir une traduction d'une qualité équivalente une traduction humaine (voir section 2.5). Pour les guider dans leur travail, nous leur avons fourni les lignes directrices en matière de PE données par TAUS[16] (2010) pour l'obtention d'un tel niveau de qualité.

[16] Ces lignes directrices sont disponibles sur : https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

6.2.4. Résultats de l'évaluation humaine 1

Analyse des résultats

Une fois que nos deux traductrices avaient terminé la tâche de PE, nous avons procédé à l'analyse des résultats. Lorsque le projet est terminé, *MateCat* permet à l'utilisateur d'ouvrir une interface (*EditingLog*) qui permet de visualiser les statistiques du projet. Voici un aperçu de cette interface (Figure 7):

Editing Log - Project: Valentine_1

Stats for translation job 1222497 - de-DE > fr-FR

Words	Total Time-to-edit	Your avg secs/word	Your avg PEE
3196	07h:21m:08s	8.3s	17.07%

Editing Details (Export detailed log as CSV)

Segment ID	Words	Time-to-edit (TTE)	Secs/Word	Post-editing effort (PEE)
681474078	17	03m:31s	12.4s	28%

Segment	Damit vereinfachen wir unseren Kunden das Handeln in einem komplexen Umfeld und verhelfen ihnen zu mehr Freiräumen.
Suggestion (TM - 100%)	Afin que nous facilitons nos clients dans un environnement complexe et leur permettons de gagner en liberté.
Translation	Ainsi, nous facilitons le commerce dans un environnement complexe à nos clients et leur permettons de gagner en liberté.
Diff View	~~Afin que~~ Ainsi, nous facilitons ~~nos clients~~ le commerce dans un environnement complexe à nos clients et leur permettons de gagner en liberté.
QA Issues	

Figure 7 - *Interface EditingLog de MateCat*

Tout en haut, *MateCat* donne les statistiques du projet entier. Il donne le nombre de mots sources, le temps total que le post-éditeur a passé sur le projet, le nombre moyen de secondes par mots, ainsi que l'effort moyen de PE en pourcentage. L'interface donne ensuite ces mêmes statistiques pour chacun des segments du projet. Elle présente aussi le segment source, la suggestion de la MT et la traduction validée par le traducteur. Enfin, la case *Diff View* affiche les modifications réalisées par le traducteur. Ces informations peuvent être téléchargées en format .csv pour être analysées plus en détail[17].

Nous avons donc téléchargé ces statistiques pour les deux projets de chacune de nos traductrices. Voici comment se présentent ces informations (Figure 8) (par souci de lisibilité, nous avons masqué les colonnes qui ne nous sont pas utiles pour notre analyse) :

[17] Pour plus d'information voir (Federico, Bertoldi, Negri, et al. 2014)

Segment ID	Words	Match perce	Time-to-edit	Post-editing	Segment	Suggestion	Translation	Statistically relevant
681474078	17	100%	211223	28%	Damit vereinfachen wir unseren Kunden das Handeln in einem komplexen Umfeld und verhelfen ihnen zu mehr Freiräumen.	Afin que nous facilitons nos clients dans un environnement complexe et leur permettons de gagner en liberté.	Ainsi, nous facilitons le commerce dans un environnement complexe à nos clients et leur permettons de gagner en liberté.	1
681474079	10	100%	70415	33%	8) Neue Berechnungsgrundlage für 2007, Werte nicht vergleichbar mit Vorjahren	(8) nouvelle base de calcul pour 2007, des valeurs pas comparable avec les exercices précédents	(8) nouvelle base de calcul pour 2007, des valeurs non comparables avec les années précédentes	1
681474080	21	100%	128286	21%	Die Schweizerische Nationalbank SNB belässt ihre Geldpolitik unverändert expansiv – mit dem Ziel, die Preisentwicklung zu stabilisieren und die Wirtschaftsaktivität zu unterstützen.	La BNS belaesst politique monétaire, sans modification expansiv - avec l' objectif de stabiliser l' évolution des prix et de soutenir l' activité économique.	La BNS laisse sa politique monétaire expansive avec l'objectif de stabiliser l'évolution des prix et de soutenir l'activité économique.	1
681474081	14	100%	259904	24%	Die Zahl der adressierten Briefe nahm im vergangenen Jahr um 3,8 Prozent ab (2015:	Le nombre des lettres adressées s' est accru à l' an dernier de 3,8 % (2015 :)	Le nombre des lettres adressées a baissé l'an dernier de 3,8 % (2015 :	1
681474082	23	100%	206429	19%	Die Post hat im vergangenen Jahr Investitionen von rund 450 Millionen Franken in die Weiterentwicklung ihrer Infrastruktur, neue Dienstleistungen und weitere Themen getätigt.	La poste a dans l' année précédente des investissements de près de 450 millions de francs dans le développement de son infrastructure, de nouvelles prestations et autres thèmes ont été effectuées.	L'année précédente, la Poste a réalisé des investissements de près de 450 millions de francs dans le développement de son infrastructure, dans de nouvelles prestations et autres thèmes.	1
681474083	4	100%	46368	0%	Strom als Wärme innerhalb	L'électricité en tant que chaleur à l'intérieur	L'électricité en tant que chaleur à l'intérieur	1
681474084	17	98%	189271	32%	Der daraus resultierende Rückgang des Erlöses um 7 Millionen Franken konnte durch andere Finanzdienstleistungen nicht kompensiert werden.	Le recul qui en résultent des recettes de 7 millions de francs par d' autres services financiers pas pu être compensé.	Le recul des recettes de 7 millions de francs qui en résulte n'a pas pu être compensé par d'autres services financiers.	1

Figure 8 - *Statistiques téléchargées en .csv via l'EditingLog de MateCat*

En exportant les statistiques, on ne dispose plus de la *Diff View* que nous avions dans l'interface, en revanche, nous disposons d'informations supplémentaires, comme une colonne *Statistically relevant.* Cette colonne comporte un 0 lorsque le segment présente un temps de PE par mot inférieur à 0,5 seconde ou supérieur à 25 secondes. *MateCat* ne prend pas en compte ces segments dans le calcul de l'effort et du temps total de PE[18]. Nous avons décidé de prendre en compte tous les segments dans nos résultats, car certains segments ont un temps de PE très faible simplement parce qu'aucune PE n'a été effectuée et d'autres ont un temps de PE très élevé, car une importante PE était nécessaire. Il nous semblait donc essentiel de considérer tous les segments pour notre test. Nous avons décidé de ne pas prendre en compte l'effort de PE donné par *MateCat* dans nos résultats, car nous ne savons pas comment celui-ci est calculé (*MateCat* ne donne aucune information à ce sujet) et nous ne savons pas non plus s'il prend en compte le temps de PE ou simplement le nombre de modifications.

Après avoir exporté les statistiques de tous les projets de notre test, nous avons effectué un tri par ordre alphabétique (en fonction de la source) afin de pouvoir identifier les segments provenant de MTH et de *DeepL*. Nous avons créé pour chaque traducteur un fichier contenant les statistiques de PE de MTH pour le premier projet et un autre pour le deuxième projet et nous avons fait la même chose pour les statistiques de PE des segments issus de *DeepL*. Nous ne voulions pas mélanger les statistiques du premier et du deuxième projet afin de voir s'il n'y avait pas de différence significative dans la vitesse de PE des traducteurs entre le premier et le deuxième projet, due par exemple, à une éventuelle lassitude face à la tâche ou une baisse de la

[18] https://www.matecat.com/support/advanced-features/editing-log/

concentration. L'une des traductrices nous a en effet fait part de son sentiment de lassitude lors du deuxième projet en raison de l'aspect un peu répétitif de la tâche et de son affinité limitée avec le domaine GB. Nous disposions pour chaque traducteur des données statistiques suivantes :

- Corpus_1_MTH_p1 (projet 1)
- Corpus_1_DL_p2 (projet 2)
- Corpus_2_MTH_p2 (projet 2)
- Corpus_2_DL_p1 (projet 1)

Le corpus_1 correspond aux 125 premiers segments (classés par ordre alphabétique) de notre corpus et le corpus_2 aux 125 derniers. P1 et p2 indiquent si le corpus a été post-édité dans le projet 1 ou dans le projet 2.

Temps de post-édition

A l'aide des informations données par *MateCat*, nous avons compilé les temps de post-édition de chaque traductrice pour chaque corpus et pour chaque système de TA. Pour avoir une idée plus précise des variations du temps de PE entre nos deux systèmes, nous avons calculé le temps de PE moyen par segment et par mot. Pour ce dernier calcul, nous nous sommes basé sur le nombre de mots de la source afin que les résultats de nos deux systèmes soient comparables. Le tableau ci-dessous présente ces temps de post-édition (Tableau 11) :

		mots	min	sec	sec/seg	sec/word
		Corpus 1				
Traductrice 1	MTH (projet 1)	1759	143,35	8841,7	70,73	5,03
	DeepL (projet 2)	1759	64,21	3852,6	30,82	2,19
		Corpus 2				
	MTH (projet 2)	1432	97,32	5839	46,71	4,07
	DeepL (projet 1)	1432	72,48	4349	34,79	3,04
		Corpus 1				
Traductrice 2	MTH (projet 1)	1759	322,1	19326,45	154,6	10,98
	DeepL (projet 2)	1759	104,35	6261,15	50,08	3,56
		Corpus 2				
	MTH (projet 2)	1432	206,63	12397,75	99,18	8,66
	DeepL (projet 1)	1432	117,83	7069	47,13	4,9

Tableau 11 - *Temps de post-édition de chaque traductrice pour chaque corpus et chaque système de TA*

Nous constatons tout d'abord que la vitesse de PE varie fortement entre nos deux traductrices. Cela s'explique en grande partie par le fait que la traductrice 1 est beaucoup plus expérimentée et connait mieux le domaine que la traductrice 2. Nous remarquons aussi que nos deux traductrices ont été plus rapides pour post-éditer le projet 2 que le projet 1 (la traductrice 1 à mis 54 minutes de moins pour le projet 2 et la traductrice 2 129,5 minutes de moins). Nous pensons que cela peut être dû deux facteurs : il est possible que les traductrices se soient senties plus à l'aise lors du projet 2 car elles avaient déjà effectué cette tâche, elles avaient alors eu le temps de bien prendre en main l'outil MateCat et de s'habituer à la tâche de PE. En outre, il se pourrait aussi que les évaluatrices aient éprouvé une certaine lassitude face à cette tâche assez longue et un peu répétitive et qu'elles aient eu tendance à se hâter de terminer la PE.

Nous constatons ensuite que pour les deux corpus, nos deux traductrices ont post-édité plus rapidement la TA issue de *DeepL* que celle de MTH. Les écarts dans le temps de PE de l'output de chaque système sont assez importants. Pour le corpus 1, la traductrice expérimentée a consacré en moyenne 40 secondes de moins par segment à la TA issue de *DeepL* tandis que notre traductrice non expérimentée y a consacré en moyenne 1 minute et 44 secondes de moins. Pour ce corpus, le temps de PE de l'output de MTH est plus de deux fois supérieur à celui de l'output de *DeepL* pour les deux traductrices. En ce qui concerne le corpus 2, la tendance est la même pour nos deux traductrices, le temps de PE de *DeepL* est toujours inférieur à celui de

MTH, mais les écarts sont un peu moins grands. Ces derniers restent cependant assez importants : près de 12 secondes par segments pour la traductrice numéro 1 et 52 secondes pour la traductrice numéro 2. Pour le corpus 1 comme pour le corpus 2, nos deux traductrices ont été plus rapides pour la PE de *DeepL*. Le tableau ci-dessous (Tableau 12) présente les temps de PE moyens de chaque évaluatrice pour le corpus entier (corpus 1 + corpus 2) :

		mots	min	sec	sec/seg	sec/word
Traductrice 1	MTH	3191	244,68	14 680,7	58,72	4,60
	DeepL	3191	136,69	8 201,6	32,81	2,57
Traductrice 2	MTH	3191	528,74	31 724,2	126,90	9,94
	DeepL	3191	222,17	13 330,15	53,32	4,18
Moyenne traductrice 1 et traductrice2	MTH	3191	386,71	23 202,45	92,81	7,27
	DL	3191	179,43	10 765,86	43,06	3,37

Tableau 12 - *Temps de post-édition des traductrices pour chaque système pour le corpus entier*

Les résultats sur le corpus entier montrent bien que la post-édition de la TA de *DeepL* est bien plus rapide que celle de MTH. La traductrice 1 et la traductrice 2 consacrent environ deux fois moins de temps (-53,6 %) par mot lors de la PE de la TA de *DeepL*.

Les statistiques de temps obtenues lors de notre test de PE nous indiquent donc clairement que pour notre corpus de test la TA issue de *DeepL* est post-éditée bien plus rapidement que celle de MTH par nos traductrices. Nous allons désormais voir si ces résultats corrèlent avec un autre aspect de l'effort de PE, à savoir, le nombre de modifications apportées à l'output.

Human-targeted Translation Edit Rate

Nous venons de voir que c'est la TA fournit par *DeepL* qui a demandé moins de temps de PE à nos traductrices. Nous allons maintenant nous intéresser au nombre de modifications que nos traductrices ont apportées à l'output de chaque système. Pour ce faire, nous avons utilisé la métrique HTER (Snover et al., 2006) (voir section 3.2.). Pour rappel, le HTER mesure

le nombre de modifications apportées à la TA par le post-éditeur pour parvenir à la nouvelle traduction de référence. Il est calculé en divisant le nombre de modifications par le nombre de mots de la référence, ce qui signifie que plus le HTER est élevé, plus le post-éditeur a effectué de modifications. Pour calculer le HTER, nous avons utilisé le script java *tercom* de Snover et al. (2007)[19] dans sa version sensible à la casse. Pour chaque corpus, nous avons soumis au script l'output de la TA de chaque système ainsi que les phrases post-éditées correspondantes.

Le tableau ci-dessous (Tableau 13) présente les scores HTER obtenus pour chaque corpus, chaque traductrice et chaque système :

		HTER
Traductrice 1	**Corpus1**	
	MTH (p1)	0,5013
	DeepL (p2)	0,1384
	Corpus 2	
	MTH (p2)	0,5084
	DeepL (p1)	0,1928
Traductrice 2	**Corpus 1**	
	MTH (p1)	0,4526
	DeepL (p2)	0,0789
	Corpus 2	
	MTH (p2)	0,4779
	DeepL (p1)	0,0770
Moyenne traductrices 1 et 2 (corpus entier)	MTH	0,4842
	DeepL	0,1204

Tableau 13 - *Scores HTER de chaque traductrice pour chaque corpus et chaque système de TA.*

Nous remarquons tout d'abord que les scores HTER sont généralement plus bas pour la traductrice numéro deux (la moins expérimentée), ce qui signifie qu'elle a effectué moins de modifications que la traductrice expérimentée. Cela peut s'expliquer par le fait que la

[19] Disponible à l'adresse http://www.cs.umd.edu/~snover/tercom/

traductrice expérimentée a probablement corrigé des erreurs que la traductrice 2 n'était pas en mesure de repérer, comme de la terminologie ou des formulations spécifiques à l'entreprise.

Nous constatons que les scores HTER de *DeepL* sont largement inférieurs à ceux de MTH, nos traductrices ont donc effectué moins de corrections sur l'output de *DeepL*. En outre, ces scores sont très réguliers entre les corpus, ce qui signifie que nos deux traductrices ont, en moyenne, apporté autant de modifications à l'output de la TA pour chaque corpus. Le HTER moyen de *DeepL* pour les deux traductrices sur le corpus entier est inférieur de 75,1 % à celui de MTH.

Le tableau ci-dessous (Tableau 14) récapitule l'ensemble des résultats de l'évaluation humaine 1 pour chaque système et pour chaque traductrice ainsi que les moyennes des résultats des deux traductrices pour chaque système :

		Temps de PE en sec/mot	HTER
MTH	Traductrice 1	4,6	0,5044
	Traductrice 2	9,94	0,4639
	Moyenne	7,27	0,4842
DeepL	Traductrice 1	2,57	0,1627
	Traductrice 2	4,18	0,0780
	Moyenne	3,38	0,1204

Tableau 14 - *Temps de PE en seconde par mot et HTER pour chaque système*

Les résultats ci-dessus indiquent clairement que la TA de *DeepL* a nécessité un effort de PE inférieur à celle de MTH, tant en termes de temps de PE, qu'en termes de modifications apportées à l'output. Les écarts de scores HTER entre les deux systèmes sont importants, mais il est difficile d'identifier ce que ces scores signifient pour le traducteur en conditions de travail réelles. Pour l'écart de temps, en revanche, on voit clairement que la TA issue de *DeepL* offre un gain de temps important par rapport à celle de MTH pour le traducteur.

Cette première évaluation humaine nous a permis d'identifier *DeepL* comme étant le meilleur de nos deux systèmes dans un contexte dans lequel la TA est destinée être post-éditée. Si l'on considère que l'effort de PE est le reflet de la qualité intrinsèque de la TA (voir section 3.1), nous pouvons en déduire que *DeepL* produit une TA de meilleure qualité que MTH pour le

domaine que nous avons évalué. Dans la section suivante, nous allons procéder à une évaluation de la qualité des traductions post-éditée afin de voir si le système de TA utilisé à une influence sur la qualité finale des traductions.

6.3. Evaluation humaine 2 : qualité des traductions

Nous avons décidé de conduire une deuxième évaluation humaine afin de voir si le système de TA utilisé avait un impact sur la qualité de la traduction post-éditée. En effet, pour être utile au traducteur, un système de TA doit permettre une post-édition rapide et nécessitant le moins de changement possible, mais il ne doit pas conduire à une altération de la qualité de la traduction finale. Nous allons donc cherché à savoir si notre système neuronal, en plus de produire une TA brute de meilleure qualité, donne aussi lieu à une traduction post-éditée de bonne qualité.

6.3.1. Déroulement de l'évaluation humaine 2

Pour conduire cette évaluation, nous avons demandé à trois étudiantes en traduction de langue maternelle française de comparer la qualité des traductions finales produites par nos post-éditrices lors de l'évaluation 1 à partir de la TA de MTH et de *DeepL*. Pour chaque segment, nous avons soumis aux étudiantes la traduction finale post-éditée à partir de la TA de MTH et celle post-éditée à partir de la TA de *DeepL* et nous leur avons demandé d'indiquer laquelle des traductions étaient, selon elles, la meilleure. Nos évaluatrices n'ayant pas toutes l'allemand dans leur combinaison de langues, nous leur avons fourni une traduction de référence (qui était la traduction présente dans la mémoire de traduction de La Poste) sur laquelle elles pouvaient s'appuyer pour juger de la qualité des traductions soumises.

Le test se présentait sous la forme d'un tableau Excel dans lequel les évaluatrices devaient indiquer si elles estimaient que la meilleure traduction était la 1 ou la 2 ou si elles les jugeaient équivalentes. Les traductions issues de *DeepL* et de MTH avaient été mélangées afin de ne pas influencer le jugement des évaluatrices. Chaque juge a évalué les deux corpus post-édités par nos deux traductrices lors de l'évaluation humaine 1 (soit 500 segments issus de *DeepL* comparé aux mêmes 500 segments issus de MTH).

Les consignes données aux participantes de l'évaluation se trouvent dans l'annexe D.

6.3.2. Résultats de l'évaluation humaine 2

Le tableau ci-dessous (Tableau 15) présente le nombre de traductions jugées meilleures par chacune des juges pour chaque système. La troisième colonne indique les traductions jugées équivalentes :

Juges	Traductions issues de *DeepL*	Traductions issues de MTH	Traductions jugées équivalentes	Total
Juge 1	221	106	173	500
Juge 2	197	206	97	500
Juge 3	221	166	113	500

Tableau 15 - *Nombre de traductions jugées meilleures par chaque juge pour chaque système*

Nous voyons que la juge 1 et la juge 3 obtiennent des résultats semblables. Sur les 500 segments qu'elles ont comparés, elles ont estimé qu'une majorité des traductions post-éditées à partir de *DeepL* étaient meilleures que celles post-éditées à partir de MTH. La juge 2, en revanche, a indiqué une majorité de traductions de MTH comme meilleures. Les trois juges ont indiqué un nombre assez important de traductions équivalentes. Afin de mesurer l'accord entre les juges, nous avons calculé le score kappa de Light (Light, 1971)[20]. Nous avons obtenu un score kappa de 0,226, ce qui indique un « accord faible » entre les juges selon la grille de lecture de Landis et Koch (1977).

Les résultats de nos trois juges étant assez partagés et l'accord entre elles étant faible, nous avons décidé de les compiler segment par segment pour départager chaque traduction à la majorité. Pour ce faire, nous avons regardé le jugement des trois évaluatrices pour chaque segment et nous avons compté, d'une part, les segments pour lesquels il y avait un jugement

[20] Le kappa de Light permet de calculer l'accord entre plus de deux juges (Light 1971).

unanime de la part de nos trois évaluatrices et, d'autre part, les segments pour lesquels il y avait un jugement majoritaire (deux juges ayant le même avis). Le tableau ci-dessous (Tableau 16) présente le nombre de segments jugés de manière unanime pour chaque système et le nombre de segments ayant été jugés comme meilleur par deux juges. La dernière colonne indique le nombre de segments pour lesquels il n'y avait aucune majorité :

Traductions DeepL jugées meilleures à l'unanimité	**Traductions DeepL jugées meilleures par 2 des 3 juges**	**Traductions MTH jugées meilleures à l'unanimité**	**Traductions MTH jugées meilleures par 2 des 3 juges**	**Traductions jugées équivalentes l'unanimité**	**Traductions jugées équivalentes par 2 des 3 juges**	**Traductions non départagées**
80 (16 %)	129 (25,8 %)	45 (9 %)	90 (18 %)	27 (5,4 %)	61 (12,2 %)	68 (13,6 %)

Tableau 16 - *Nombre de segments jugés meilleurs à l'unanimité et à la majorité (2 contre 1) pour chaque système (pourcentage du nombre total de segments (500))*

Nous voyons que 80 segments issus de *DeepL* ont été qualifiés de meilleurs à l'unanimité (soit 16 % des segments), contre seulement 45 segments (9 %) issus de MTH. Il y a peu de traductions (27 soit 5,4% des segments) qui ont été qualifiées d'équivalentes à l'unanimité. Les résultats à la majorité (deux juges contre un), penchent aussi en faveur de *DeepL* avec 129 segments (soit 25,8 % des segments) jugés meilleurs contre 90 pour MTH (18 %).

Dans le tableau ci-dessous (Tableau 17), nous avons regroupé les jugements unanimes et les jugements « deux contre un » afin d'avoir une vue d'ensemble des jugements à la majorité (au moins deux juges).

DL (au moins 2 juges)	**MTH (au moins deux juges)**	**Equivalentes (au moins deux juges)**	**Traductions non départagées**	**Total**
209 (41,80 %)	135 (27,00 %)	88 (17,60 %)	68 (13,60 %)	500 (100 %)

Tableau 17 - *Nombre de segments jugés meilleurs à la majorité (au moins 2 juges) pour chaque système (pourcentage par rapport au nombre total de segments)*

Ces résultats nous montrent que 41,80 % des segments les traductions de *DeepL* sont jugées meilleures que celles de MTH par au moins deux juges sur trois contre 27 % pour les traductions de MTH. 31,2 % des traductions sont jugées équivalentes ou n'obtiennent aucune majorité. D'après ces résultats, il semblerait que le système de TA ait une influence sur la qualité de la traduction post-éditée et dans notre cas, il semblerait que la TA de *DeepL* donne lieu, de manière générale, à des traductions de meilleure qualité que celle de MTH après post-édition. Nous pensons cependant qu'il faut prendre ces résultats avec précaution, car l'accord entre les juges est faible pour notre test et un nombre important de traductions sont équivalentes ou non départagées. L'évaluation humaine 1 a clairement indiqué *DeepL* comme étant le meilleur système pour notre tâche de PE et, dans cette seconde évaluation humaine, nous cherchions surtout nous assurer que l'utilisation de ce système n'affectait pas la qualité des traductions finales. Cette évaluation nous a montré qu'après PE, la TA de *DeepL* n'est pas moins bonne que celle de MTH, elle est même meilleure pour 41,80 % des segments.

6.4. Discussion des résultats

Les résultats de notre évaluation automatique donnent *DeepL* comme étant meilleur que notre système spécialisé MTH. Cependant l'écart de score BLEU (1,77 point) (voir section 6.2.1) reste relativement faible par rapport à certains écarts que nous avons pu obtenir lors de la première évaluation automatique qui portait sur nos différents systèmes MTH (section 5.2). La deuxième phase de notre évaluation comparative (section 6.2) corrobore cependant les résultats de notre évaluation automatique (voir section 6.1). Les résultats obtenus lors de cette évaluation humaine indiquent très clairement que la TA fournie par *DeepL* pour notre corpus de test GB a nécessité un effort de PE moins important que celle fournie par notre système spécialisé MTH. Nous avons aussi pu constater que le temps de PE est étroitement lié à la quantité de corrections apportées. Notre traductrice expérimentée et notre traductrice non expérimentée ont toutes deux mis moins de temps et effectué moins de modifications sur la TA de *DeepL* que sur celle de MTH. Notre deuxième évaluation humaine portant sur la qualité des traductions finales indique, qu'après post-édition, les traductions de *DeepL* ont tendance à être meilleures que celles de MTH.

Au regard des résultats de nos évaluations humaines et de notre évaluation automatique, il n'y a pas de doute sur le fait que *DeepL* est meilleur que notre système MTH pour le scénario et le domaine étudiés dans notre travail, à savoir, l'utilisation de la TA comme outil de pré-

traduction. La TA issue de *DeepL* obtient un meilleur score BLEU, elle est plus rapide à post-éditer, demande moins de corrections et donne lieu de manière générale à une traduction finale de meilleure qualité.

7. EVALUATION HUMAINE VS AUTOMATIQUE

Le second objectif de ce travail était de chercher à savoir si *le score BLEU est une métrique fiable pour l'évaluation des systèmes de traduction automatique neuronale* (voir chapitre 1). Nous allons tenter d'apporter des réponses à cette question dans ce chapitre. Comme nous l'avons mentionné dans la section 1.3, il existe actuellement peu d'études sur la corrélation entre le score BLEU et le jugement humain dans le domaine de la TAN. En outre, deux études conduites successivement par Shterionov et al. (2017; 2018) montrent que le score BLEU aurait tendance à sous-estimer la qualité de la TAN. Selon les auteurs de ces études, cela pourrait s'expliquer par le fait que la métrique BLEU, étant basée sur les n-grammes, est mieux adaptée à l'évaluation de la TAS, car le fonctionnement de cette dernière repose aussi sur les n-grammes. Or les systèmes de TAN fonctionnent de manière très différente et ne sont pas basés sur les n-grammes. Ils produisent donc un output qui n'a pas les mêmes caractéristiques que celui d'un système de TAS. Toujours selon les auteurs, les systèmes de TAN seraient plus enclins à produire une traduction dont la longueur, l'ordre et le choix des mots diffèrent fortement de la référence, ce qui fait naturellement baisser le score BLEU (voir section 3.2) (Shterionov et al., 2017; 2018).

Si l'on considère le fait que la métrique BLEU a été développée bien avant l'apparition de la TAN et qu'elle était donc destinée à évaluer la TA de systèmes dont le fonctionnement était très différent, il est justifié de se demander si cette métrique est adaptée à l'évaluation des systèmes de TAN. Selon Callison-Burch et al. (2006), le BLEU serait inapproprié pour comparer des systèmes dont le fonctionnement est très différent, notamment des systèmes de TAS avec des systèmes qui ne sont pas basés sur les n-grammes). C'est pour ces raisons que nous avons décidé de nous intéresser à la corrélation entre les scores BLEU que nous avons obtenus et les résultats de notre première évaluation humaine.

Nous commencerons par décrire la méthodologie adoptée pour cette évaluation (section 7.1), puis nous détaillerons les comparaisons effectuées au niveau du corpus et au niveau des segments (section 7.2 et 7.3).

7.1. Méthodologie

Pour cette évaluation, nous prenons comme point de départ l'hypothèse de Shterionov et al. (2017) selon laquelle le score BLEU tendrait à sous-estimer la qualité de l'output de la TAN. Cette hypothèse est aussi formulée par Way (2018) qui estime que les métriques basées sur les n-grammes telles que BLEU ne reflètent pas la qualité réelle de la TAN. Nous allons chercher à vérifier si cette hypothèse se confirme dans notre cas. Dans leur étude, Shterionov et al. (2017) ont compté le nombre de segments issus de la TAN qui étaient jugés comme meilleurs que ceux de la TAS par des évaluateurs humains et ils ont ensuite calculé quel pourcentage de ces segments obtenait un score BLEU inférieur à ceux de la TAS. Ils ont ainsi obtenu le pourcentage de segments sous-estimés par le score BLEU (Shterionov et al., 2017).

Nous avons décidé d'adopter une méthode similaire pour notre évaluation même si notre évaluation humaine portait sur l'effort de PE et non sur la qualité intrinsèque de l'output, car ces deux aspects sont en principe liés. En effet, de manière générale, l'effort de PE est moins important lorsque la traduction est de bonne qualité (Kit et Wong, 2015). De plus, dans leur seconde étude, Shteriovo et al. (2018) ont réalisé un test de productivité de PE et ont obtenu des résultats qui corroboraient ceux de leur évaluation humaine de la qualité intrinsèque de la TA. Nous allons donc cherché à déterminer si le score BLEU sous-estime la qualité de la TA de *DeepL* pour notre corpus.

Nous allons tout d'abord comparer les résultats de l'évaluation automatique et ceux de l'évaluation humaine 1 au niveau du corpus (section 7.2). Dans un deuxième temps, nous avons procédé à une évaluation plus précise en comparant les résultats obtenus au niveau des segments (section 7.3).

7.2. Comparaison au niveau du corpus

Afin de comparer nos résultats au niveau du corpus, nous avons dû calculer le score BLEU des corpus de TA que nous avions utilisés lors de l'évaluation humaine 1. Nous avons calculé ces scores BLEU de la même manière que lors de l'évaluation automatique (voir section 6.1).

Le tableau ci-dessous (Tableau 18) donne les scores BLEU obtenus par chacun de nos systèmes sur le corpus utilisé dans l'évaluation humaine 1 :

Système	BLEU mteval
DeepL	21,38
MTH	22,65

Tableau 18 - *Scores BLEU pour le corpus de l'évaluation humaine 1 (250 segments)*

Les scores BLEU pour notre corpus de 250 segments diffèrent légèrement de ceux obtenus lors de notre première évaluation humaine réalisée sur le corpus de test entier qui comptaient 1718 segments. Bien que ces écarts restent limités, on remarque tout de même que cette fois-ci c'est MTH qui obtient le meilleur score alors que c'était *DeepL* pour le corpus entier (voir section 6.1). Dans les deux cas, les deux systèmes obtiennent un score assez proche, on voit donc que la métrique automatique BLEU peine ici à départager nos deux systèmes de manière claire.

L'évaluation humaine 1, en revanche, a départagé très nettement nos deux systèmes en désignant *DeepL* comme étant le meilleur système. Le tableau ci-dessous (Tableau 19) rappelle les temps moyens de PE et les scores HTER moyens obtenus par chaque système lors de l'évaluation humaine 1 :

Système	Temps (s/mot)	HTER
DeepL	3,38	0,1204
MTH	7,27	0,4842

Tableau 19 - *Temps de PE et HTER moyens obtenus par chaque système lors de l'évaluation humaine 1*

Nous voyons donc qu'au niveau du corpus, les résultats de l'évaluation automatique ne corrèlent pas vraiment avec ceux de l'évaluation humaine, ce qui nous amène à remettre en question la fiabilité du score BLEU pour l'évaluation de nos systèmes.

7.3. Comparaison au niveau des segments

Dans un deuxième temps, nous avons procédé à une évaluation plus précise en comparant les résultats obtenus au niveau des segments. Notre évaluation humaine 1 nous a donné deux

mesures permettant d'évaluer les traductions de nos deux systèmes : le temps de PE et le nombre de modifications apportées à l'output de la TA (mesuré à l'aide du HTER). Pour cette évaluation, nous avons considéré qu'une traduction était donnée comme meilleure par l'évaluation humaine si son temps de PE et son HTER étaient meilleurs pour les deux traductrices. Nous n'avons donc pas pris en compte les traductions pour lesquelles l'output de *DeepL* obtenait un temps de PE meilleur que celui de MTH mais un HTER moins et inversement, car il est difficile de déterminer laquelle de ces deux métriques il faudrait prendre en compte pour départager les 2 traductions. Nous n'avons pas non plus pris en compte les segments pour lesquels il n'y avait pas d'accord entre nos deux traductrices.

Nous avons tout d'abord calculé les scores BLEU et HTER de *DeepL* et de MTH pour chacun des segments du corpus.

Le tableau ci-dessous (Tableau 20) présente le nombre de segments qualifiés de meilleurs par l'évaluation automatique pour chaque système, c'est-à-dire les segments qui ont un meilleur score BLEU. Les segments non départagés sont les segments qui obtiennent des scores identiques :

DeepL	MTH	Non départagés
117 (46,8 %)	129 (51,6 %)	4 (1,6 %)

Tableau 20 - *Résultats de l'évaluation automatique par segment (pourcentages de segments qui ont un meilleur score BLEU)*

Pour l'évaluation automatique, nous voyons qu'une majorité des segments (51,6%) de MTH obtiennent un score BLEU supérieur à ceux de *DeepL*.

Nous avons ensuite compté le nombre de segments pour lesquels la TA de *DeepL* obtient un temps de PE et un HTER meilleurs que celle de MTH pour nos deux traductrices. Le tableau ci-dessous (Tableau 21) présente le nombre de segments identifiés comme meilleurs pour nos deux traductrices lors de l'évaluation humaine pour chaque système, c'est-à-dire les segments qui ont un meilleur temps de PE et un meilleur HTER :

	DeepL	MTH	Non départagés	Total
Nombre de segments (pourcentage)	144 (57,6 %)	15 (6 %)	91 (36,4 %)	250 (100%)

Tableau 21 - *Résultats de l'évaluation humaine par segment (pourcentages de segments qualifiés de meilleurs par l'évaluation humaine)*

Ces résultats par segments montrent qu'une majorité des segments issus de *DeepL* obtiennent un temps de PE et un HTER meilleurs que ceux de MTH pour nos deux traductrices. Les segments non départagés correspondent aux segments pour lesquels le temps de PE et le HTER ne sont pas tous deux meilleurs pour les deux traductrices. Les résultats obtenus semblent donc indiquer que le score BLEU a tendance à sous-évaluer la qualité de la TA de *DeepL* puisque BLEU estime que 117 segments *DeepL* sont meilleurs contre 144 pour l'évaluation humaine.

Parmi les segments qualifiés de meilleurs par l'évaluation humaine, nous avons regardé combien d'entre eux avaient obtenu un score BLEU inférieur à celui de MTH. Nous avons fait la même chose pour les segments traduits par MTH. Pour calculer le pourcentage de sous-estimation du score BLEU, nous avons utilisé la même formule que Shterionov et al. (2017)[21] : nous avons divisé le nombre de segments désignés comme meilleurs par l'évaluation humaine et comme moins bons par le score BLEU par le nombre de segments désignés comme meilleurs par l'évaluation humaine et nous avons multiplié le tout par 100.

Dans le tableau ci-dessous (Tableau 22), présente le pourcentage de segments traduits par *DeepL* qui ont obtenu un score BLEU inférieur à ceux de MTH, mais qui ont été qualifiés de meilleurs par l'évaluation humaine :

[21] Le calcul proposé par Shterionov et al. se présente comme suit : $\frac{d_{PBSMT}^{NMT}}{d^{NMT}}$ où d^{NMT} est le nombre de traductions issues du système de TAN qui ont été qualifiées de meilleures (que celles du TAS) par l'évaluation humaine et d_{PBSMT}^{NMT} est le nombre de traductions de d^{NMT} dont le score BLEU est inférieur aux traductions correspondantes issues du système de TAN.

	Nombre de segments meilleurs selon évaluation humaine	Nombre de segments meilleurs selon l'évaluation humaine, mais pas selon le BLEU	% de segments sous-évalués par BLEU
DeepL	144	63	43,75 %
MTH	15	5	33,33 %

Tableau 22 - *Nombre et pourcentage de segments sous-évalués par BLEU pour les deux systèmes*

Ces résultats montrent que 43,75 % des segments de *DeepL* qui apparaissent comme meilleurs que ceux de MTH lors de l'évaluation humaine obtiennent des scores BLEU inférieurs à ceux des segments de MTH. De ce fait, nous pouvons dire que, dans notre cas, le score BLEU sous-évalue 43,75 % des segments traduits par *DeepL*. En ce qui concerne MTH, nos résultats indiquent que le score BLEU sous-évalue 33,33 % des segments.

Conclusion

Cette comparaison nous montre que les résultats des évaluations humaines et automatiques diffèrent sensiblement dans notre cas, tant au niveau du corpus qu'au niveau des segments. Tandis que le score BLEU ne départage pas nettement *DeepL* de MTH, l'évaluation humaine, elle, donne clairement *DeepL* comme étant le meilleur de nos deux systèmes. La métrique BLEU étant généralement considérée comme une métrique ayant une bonne corrélation avec le jugement humain pour les systèmes statistiques (voir section 3.2), nous en déduisons qu'elle sous-évalue *DeepL* dans les évaluations que nous avons conduites. Ces résultats semblent confirmer notre hypothèse selon laquelle le score BLEU aurait tendance à sous-évaluer les performances des systèmes neuronaux de manière générale.

8. CONCLUSION

Dans ce dernier chapitre, nous récapitulerons notre travail et les résultats de notre étude afin d'apporter une réponse à nos deux questions de recherche (section 8.1). Nous mentionnerons ensuite les limites de notre étude et donnerons quelques pistes pour d'éventuels travaux futurs (section 8.2). Nous terminerons en indiquant les recommandations que nous estimons utiles pour l'entreprise au regard des résultats que nous avons obtenus (section 8.3).

8.1. Synthèse et résultats de l'étude

L'étude que nous avons menée visait, d'une part, à *déterminer si les systèmes statistiques spécialisés sont encore en mesure de rivaliser avec les systèmes neuronaux généralistes lorsque la traduction automatique est utilisée par le traducteur professionnel comme outil d'aide à la traduction* et, d'autre part, à *estimer la fiabilité du score BLEU pour l'évaluation des systèmes de TAN*. Notre travail a été réalisé dans le cadre d'un projet d'implémentation de la TA au sein de La Poste Suisse, notre étude a donc été guidée par les attentes de cette entreprise et les évaluations ont été conduites sur les données qu'elle nous a fournies.

Afin d'atteindre les objectifs que nous nous étions fixés, il nous a tout d'abord fallu présenter le domaine de la traduction automatique et décrire les principales méthodes d'évaluation existantes. Nous avons aussi présenté les différents systèmes que nous avons utilisés ainsi que leur fonctionnement.

Pour répondre à notre première question, nous avons comparé un système de TAS spécialisé (*Microsoft Translator Hub*) avec un système de TAN généraliste (*DeepL*) dans un contexte d'entreprise. En premier lieu, nous avons entrainé des systèmes statistiques (avec MTH) avec les données de l'entreprise et nous avons sélectionné le meilleur d'entre eux en nous basant sur le score BLEU. Toujours en nous fiant à cette métrique, nous avons sélectionné le domaine pour lequel notre système donnait les meilleurs résultats parmi les quatre domaines de l'entreprise (GB, PV, PF et Modulo). Lors de cette évaluation, nous avons pu constater que les performances de MTH variaient considérablement d'un domaine à un autre.

Nous avons ensuite réalisé une évaluation comparative de notre meilleur système de TAS avec notre système de TAN. Cette évaluation comparative s'est déroulée en trois étapes :

- Evaluation automatique basée sur le score BLEU
- Evaluation humaine 1 portant sur l'effort de post-édition
- Evaluation humaine 2 portant sur la qualité des traductions finales

L'évaluation automatique a indiqué que la TA de *DeepL* était de meilleure qualité que celle de MTH pour notre corpus de 1718 segments issus du domaine GB (rapport de gestion de l'entreprise). Cependant, l'écart entre les scores obtenus par les deux systèmes restait assez limité (1,77 point BLEU), il n'y avait donc, selon le score BLEU, pas une différence significative de qualité entre les traductions automatiques de nos deux systèmes.

L'évaluation humaine 1 a donné des résultats nettement plus tranchés. L'effort de post-édition fournie par nos traductrices était nettement inférieur pour la TA issue de *DeepL* que pour celle issue de MTH. Nos traductrices ont toutes mis moins de temps et effectué moins de modifications lors de la post-édition des 250 segments traduits par *DeepL*. Le temps de PE moyen pour *DeepL* est inférieur de 53,6 % par rapport à MTH et le HTER est inférieur de 75,1 %. L'effort de post-édition étant généralement lié à la qualité de la traduction, cette évaluation humaine semble indiquer que *DeepL* produit de meilleures traductions que MTH pour les segments issus du domaine GB.

Nous avons ensuite conduit une seconde évaluation humaine afin d'identifier l'impact éventuel de nos systèmes de TA sur la qualité de la traduction finale (après post-édition). Cette évaluation a montré que les traductions post-éditées à partir de l'output de *DeepL* sont en général meilleures que celles post-éditées à partir de l'output de MTH.

La corrélation relativement limitée entre notre évaluation automatique et notre évaluation humaine 1 nous a amené à nous questionner sur la fiabilité du score BLEU, notamment pour l'évaluation des systèmes neuronaux. Nous avons donc comparé de manière plus précise les résultats que nous avions obtenus. La comparaison que nous avons effectuée a montré que le score BLEU avait sous-évalué une grande partie des segments issus de *DeepL* (43,75 % des segments), ce qui nous amène à penser que cette métrique a tendance à sous-évaluer la qualité de la traduction automatique neuronale.

Les résultats de notre étude nous amènent donc à penser que les performances des systèmes neuronaux, même généralistes, surpassent largement celles des systèmes statistiques

spécialisés. Ils remettent aussi en question la fiabilité du score BLEU, notamment pour l'évaluation des systèmes neuronaux.

8.2. Limites de l'étude et perspectives

L'étude que nous avons conduite dans le cadre de ce projet d'intégration de la traduction automatique à La Poste Suisse est une étude assez succincte qui connait de nombreuses limites.

Tout d'abord, nous avons comparé nos systèmes sur un domaine très restreint avec des données propres à l'entreprise avec laquelle nous avons collaboré dans le cadre de ce projet. Nos résultats peuvent donc difficilement être généralisables. Ensuite, les tests que nous avons effectués portaient sur des corpus relativement courts et les évaluations humaines ont été réalisées avec un petit nombre de juges et ne portaient que sur la paire de langues DE>FR. Les conclusions auxquelles nous sommes parvenues méritent donc d'être vérifiées par des études de plus grande ampleur.

En outre, nous nous sommes concentré sur la comparaison de la qualité de la TA de nos deux systèmes, mais nous n'avons pas confronté nos outils de TA à la traduction humaine. Nous sommes certes parvenu à déterminer lequel de nos deux systèmes donnait la meilleure TA, mais nous ne savons pas dans quelle mesure l'utilisation de ce système permettrait un gain de productivité par rapport à la traduction humaine.

Nous pensons qu'il serait aussi intéressant d'analyser la PE réalisée par nos traductrices afin de relever les types d'erreurs produites par chacun de nos systèmes. Une telle analyse permettrait aussi de comparer les corrections apportées par chacune des deux traductrices.

Les résultats que nous avons obtenus en comparant les évaluations humaines et automatiques remettent en question la fiabilité de notre évaluation automatique qui repose uniquement sur la métrique BLEU avec une référence unique. Cette évaluation automatique aurait peut-être permis d'obtenir des résultats plus fiables si nous avions utilisé plusieurs métriques automatiques ou si nous avions calculé le BLEU avec plusieurs références.

Nous n'avons pas non plus exploré les aspects pratiques concernant l'implantation d'un tel système de TA. Le système *DeepL*, par exemple, ne pourrait pas être utilisé par La Poste du fait des problèmes de confidentialité que poserait l'utilisation d'un outil en ligne.

Bien qu'ayant donné des résultats concluants, notre étude mériterait d'être approfondie en raison des limites que nous venons de mentionner. Nous pensons qu'il serait judicieux de poursuivre ces recherches, d'une part, sur des corpus plus grands et avec plus d'évaluateurs humains et, d'autre part, avec un plus grand nombre de systèmes. Il serait notamment intéressant d'intégrer des systèmes neuronaux spécialisables tels que celui proposé depuis peu par *Microsoft* (*Custom Translator*)[22].

En ce qui concerne notre deuxième question de recherche, les limites que nous avons rencontrées sont similaires. Même si les résultats que nous avons obtenus semblent indiquer que le score BLEU tend à sous-estimer la qualité de la TAN, nous pensons qu'il est indispensable de confirmer cette hypothèse en réalisant des études de plus grande ampleur. Il serait aussi intéressant de comparer différentes métriques automatiques pour l'évaluation de la TAN. Nous pensons, en outre, qu'il est légitime de s'interroger sur la nécessité de créer une nouvelle métrique qui soit spécialement adaptée à l'évaluation des systèmes neuronaux. En effet, la plupart des métriques automatiques couramment utilisées ont été développées pour évaluer des systèmes linguistiques ou statistiques, or le fonctionnement et le comportement des systèmes neuronaux sont très différents.

8.3. Recommandations pour l'entreprise

L'entreprise nous avait initialement donné pour mission d'entrainer différents systèmes statistiques à l'aide de MTH pour la paire de langues DE>FR, afin d'identifier le domaine pour lequel ce type de système donnait les meilleurs résultats, et d'évaluer la qualité de la TA produite. Nous sommes allé un peu plus loin dans nos recherches et avons comparé notre meilleur système statistique avec un système neuronal pour le domaine qui donnait les meilleurs résultats. Au regard des conclusions auxquelles nous sommes parvenus et considérant les limites de notre étude (voir section 8.2), nous adressons les recommandations suivantes :

- Les systèmes de TAS entrainés avec les données de tous les domaines confondus donnent de meilleurs résultats que les systèmes entrainés avec les données d'un seul

[22]Microsoft propose depuis mai 2018 un système neuronal personnalisable sur un modèle semblable à celui de MTH. Pour plus d'informations voir : https://blogs.msdn.microsoft.com/translation/2018/05/07/customtranslator/ (2018a)

domaine. Nous recommanderions donc d'utiliser toutes les données d'entrainement disponibles dans le cas où l'entreprise déciderait d'implémenter un système de TAS.

- Le domaine GB semble être le domaine pour lequel la TAS donne les meilleurs résultats, mais le domaine PV a lui aussi obtenu de très bons scores BLEU. Nous recommanderions donc d'évaluer aussi ce domaine plus en détail. En outre, nous recommanderions d'évaluer la performance de la TAN pour les autres domaines.
- Le système de TAN que nous avons évalué a donné de très bons résultats lors de l'évaluation humaine. Cependant, comme nous l'avons mentionné dans la section 8.2, ces résultats sont à considérer avec précaution. En outre, le système que nous avons testé ne répond pas aux exigences de l'entreprise en termes de confidentialité. Nous recommanderions donc à l'entreprise de réaliser des évaluations supplémentaires visant à estimer la qualité de différents systèmes neuronaux adaptés à ses besoins.
- N'ayant pas effectué de test de productivité, nous recommanderions à l'entreprise de mettre en place de genre de test afin de mesurer le gain de productivité que pourrait lui apporter l'implantation d'un système de traduction automatique. Nous recommandons aussi de ne pas négliger le temps nécessaire à la production de la TA, nous avons en effet constater que la pré-traduction automatique de *SDL Trados* était assez lente pour nos deux systèmes (voir section 6.1).

RÉFÉRENCES

ALLEN, J. 2003. Post-editing. Dans SOMERS, H. (ed.) *Computers and Translation. A translator's guide.* Amsterdam/Philadelphia : John Benjamin's Publishing, pp.297-318.

ALPAC 1966. *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.* Washington D.C.

ARNOLD, D., BALKAN, L., LEE HUMPHREYS, R., et al. 1992. Introduction and Overview. *Machine Translation : An Introductory Guide.* London : Blackwell-NCC.

BANERJEE, S. & LAVIE, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* (Prague), pp.65-72.

BAR-HILLEL, Y. 1951. The present state of research on mechanical translation. *Journal of the Association for Information Science and Technology,* vol. 2 (4)**,** pp.229-237.

BLANCHON, H. & BOITET, C. 2008. Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *TAL,* vol. 48 (1)**,** pp.33-65.

BOUILLON, P. 1993. Introduction et bref historique. Dans BOUILLON, P. & CLAS, A. (eds.) *La Traductique.* pp.13-20.

BOUILLON, P. 2017. *Cours de traduction automatique 2*. Université de Genève.

BOUILLON, P. & CLAS, A. 1993. *La Traductique.* Montréal.

CALLISON-BURCH, C., FORDYCE, C., KOEHN, P., et al. (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation.* (Prague), pp.136-158.

CALLISON-BURCH, C., OSBORNE, M. & KOEHN, P. Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of the*.

CASTILHO, S., MOORKENS, J., GASPARI, F., et al. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics,* vol. 108 (1)**,** pp.109-120.

COUGHLIN, D. Correlating automated and human assessments of machine translation quality. *Proceedings of the MT summit IX.* (New Orleans), pp.63-70.

DEEPL 2018a. *DeepL* [En ligne]. URL: https://www.deepl.com/translator [Consulté le 07 mai 2018].

DEEPL 2018b. *Informations presse* [En ligne]. URL: https://www.deepl.com/press.html [Consulté le 10 février 2018].

DODDINGTON, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics. *Proceedings of the second international conference on Human Language Technology Research* (San Diego) HLT, pp.138-145.

DORR, B., SNOVER, M. & MADNANI, N. 2011. Machine Translation Evaluation and Optimization. Dans OLIVE, J., et al. (eds.) *Handbook of Natural Language Processing and Machine Translation.* New York : Springer.

ESTRELLA, P., HAMON, O. & POPESCU-BELIS, A. 2007. *How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics.*

FEDERICO, M., BERTOLDI, N., NEGRI, M., et al. The MateCat Tool. *Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics.* (Dublin), pp.129-132.

GERLACH, J. 2015. Improving statistical machine translation of informal language: a rule-based pre-editing approach for French Forums. Thèse de Doctorat, Université de Genève.

HEARNE, M. & WAY, A. 2011. Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators. *Language and Linguistics Compass,* vol. 5 (5)**,** pp.205-226.

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

HUTCHINS, J. 2014. The history of machine translation in a nutshell. *Retrieved December,* vol. 20**,** pp.2009.

HUTCHINS, W. J. 1986. *Machine translation: past, present, future.* Chichester [West Sussex] : New York : Ellis Horwood ; Halsted Press.

HUTCHINS, W. J. & SOMERS, H. L. 1992. *An Introduction to Machine Translation.* London : Academic Press.

KENNY, D. & DOHERTY, S. 2014. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer,* vol. 8 (2)**,** pp.276-294.

KIT, C. & WONG, T.-M. 2015. Evaluation in machine translation and computer-aided translation. Dans CHAN, S.-W. (ed.) *The Routledge encyclopedia of translation technology.* London : Routledge.

KOEHN, P. 2010. *Statistical Machine Translation.* Cambridge : University Press.

KOEHN, P. 2017. Neural Machine Translation. *arXiv preprint arXiv:1709.07809.* [Online]. URL: https://arxiv.org/abs/1709.07809 [Consulté le 25 avril 2018].

KOEHN, P. 2018. *The State of Neural Machine Translation (NMT) by Philipp Koehn* [En ligne]. Omniscien. URL: https://omniscien.com/state-neural-machine-translation-nmt/ [Consulté le 15 mars 2018].

L'EXPRESS 2016. *Quels sont les 5 supercalculateurs les plus puissants au monde ?* [En ligne]. URL: https://www.lexpress.fr/actualite/sciences/quels-sont-les-5-supercalculateurs-les-plus-puissants-au-monde_1781992.html [Consulté le 21 mars 2018].

LANDIS, R. J. & KOCH, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics,* vol. 33 (1)**,** pp.159-174.

LIGHT, R. J. 1971. Measures of response agreement for qualitative date: some generalization and alternatives. *Psychological Bulletin,* vol. 76 (5)**,** pp.365-377.

MELBY, A. K., FIELDS, P. J. & HOUSLEY, J. 2014. Assessment of Post-Editing via Structured Translation Specifications. Dans O'BRIEN, S., et al. (eds.) *Post-editing of Machine Translation: Processes and Applications.* Newcastle upon Tyne : Cambridge Scholars Publishing, pp.274-298.

MICROSOFT 2018a. *Customize your neural translations with the new Translator custom feature* [En ligne]. URL: https://blogs.msdn.microsoft.com/translation/2018/05/07/customtranslator/ [Consulté le 18 mai 2018].

MICROSOFT 2018b. *Microsoft Translator Blog* [En ligne]. URL: https://blogs.msdn.microsoft.com/translation/ [Consulté le 01 mars 2018].

MICROSOFT 2018c. *Microsoft Translator Hub* [En ligne]. URL: https://hub.microsofttranslator.com/ [Consulté le 01 mars 2018].

MICROSOFT 2018d. *Microsoft Translator Hub User Guide* [En ligne]. URL: https://hub.microsofttranslator.com/Help/Download/Microsoft%20Translator%20Hub%20User%20Guide.pdf [Consulté le 01 mars 2018].

PAPINENI, K., ROUKOS, S., WARD, T., et al. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Associationfor Computational Linguistics.* Association for Computational Linguistics, pp.311-318.

POIBEAU, T. 2017. *Machine Translation.* Cambridge : MIT Press.

QUAH, C. K. 2006. *Translation and technology.* Houndmills ; New York : Palgrave Macmillan.

ROBERT, A.-M. 2010. La post-édition : l'avenir incontournable du traducteur. *Traduire,* vol. 222**,** pp.137-144.

ROCHARD, M. 2017. Du papier au numérique, traduction et mutations technologiques. *FORUM. Revue internationale d'interprétation et de traduction / International Journal of Interpretation and Translation,* vol. 15 (2)**,** pp.212-227.

SEEWALD-HEEG, U. 2017. Ausbildung von Posteditoren. Dans PORSIEL, J. (ed.) *Maschinelle Übersetzung.* Berlin : BDÜ Fachverlag, pp.168-175.

SHTERIONOV, D., NAGLE, P., CASANELLAS, L., et al. 2017. Empirical evaluation of NMT and PBSMT quality for large-scale translation production. *The 20th Annual Conference of the European Association for Machine Translation.* (Prague. Mai 2017), pp.74-79

SHTERIONOV, D., SUPERBO, R., NAGLE, P., et al. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation,* vol.**,** pp.1-19.

SNOVER, M., DORR, B., SCHWARTZ, R., et al. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.* (Cambridge) The Association for Machine Translation in the Americas, pp.223-231.

SNOVER, M., MADNANI, N., DORR, B. J., et al. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. *Proceedings of the Fourth Worksoshop on Statistical Machine Translation.* Association for Computational Linguistics, pp.259-268.

SNOVER, M., WANG, S. & MATSOUKAS, S. 2007. *TRANSLATION ERROR RATE (TER) 7.0*. 7.0 ed. BBN Technologies and University of Maryland.

SYSTRAN 2016. *How does Neural Machine Translation work?* [En ligne]. Systran Blog. URL: http://blog.systransoft.com/how-does-neural-machine-translation-work/ [Consulté le 15 mars 2018].

TAUS & CNGL 2010. *MT Post-editing Guidelines* [En ligne]. URL: https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines [Consulté le 02 mars 2018].

THURMAIR, G. Comparing different architectures of hybrid Machine Translation systems. *Proceedings of the MT Summit XII.* (Ottawa, Ontario, Canada), pp.340-347.

TRUJILLO, A. 1999. *Translation Engines: Techniques for Machine Translation.* London : Springer.

TUROVSKY, B. 2016. *Ten years of Google Translate* [En ligne]. Google. URL: https://www.blog.google/products/translate/ten-years-of-google-translate/ [Consulté le 20 février 2018].

WAY, A. 2018. Quality expectations of machine translation. *arXiv preprint* [Online]. URL: https://arxiv.org/abs/1803.08409 [Consulté le 21 mai 2018].

WEAVER, W. 1949. *Translation*. Carlsbad The Rockefeller Fondation.

WHITE, J. S. 2003. How to Evaluate Machine Translation. Dans SOMERS, H. (ed.) *Computers and translation : a Translator's guide.* Amsterdam : John Benjamins.

ANNEXES

A. Corpus de test (extrait) de l'évaluation humaine 1 (test de post-édition)

Source	TA MTH	TA DeepL
Dadurch konnten Abgänge von Kunden im Stückguttransport und bei den Dienstleistungen Lager und Innight, sowie tiefere Erträge im Treibstoffgeschäft überkompensiert werden.	Cela départ par les clients du transport de marchandises et les prestations de services de stockage et des services Innight, ainsi que la diminution des recettes de la vente de carburants plus que compensé.	Cela a plus que compensé les cessions de clients dans le transport de marchandises générales et dans les services d'entreposage et d'Innight, ainsi que la baisse des bénéfices dans le secteur des carburants.
Um diesen Fehlbetrag zumindest teilweise zu kompensieren, belastet PostFinance seit dem 1. Februar 2017 auf Barguthaben von Privatkunden über einem Schwellenwert von 1 Million Franken eine Guthabengebühr von 1 Prozent.	Ce montant manquant au moins partiellement pour compenser, sur PostFinance une Commission de 1 % depuis le 1er février 2017 sur les avoirs en espèces des clients privés dépassant la valeur seuil de 1 million de francs.	Afin de compenser au moins partiellement ce manque à gagner, PostFinance applique depuis le 1er février 2017 une commission de 1% sur les avoirs de la clientèle privée au-delà d'un seuil de CHF 1 million.
Unsere Gesellschaft und mit ihr die Bankenwelt werden rasant digitaler.	Notre société et du monde bancaire pleine mutation numérique.	Notre société et avec elle, le monde bancaire est en train de devenir de plus en plus numérique.
Um eine Million Franken höhere Treibstoffaufwendungen reduzierten das Betriebsergebnis zusätzlich.	D' un million de francs des charges de carburant ont en outre, le résultat d'exploitation.	Le bénéfice d'exploitation a encore été réduit d'un million de francs suisses par l'augmentation du coût du carburant.
Wegen des Kreditverbots wird es für PostFinance im aktuellen Negativzinsumfeld immer schwieriger, die Kundengelder profitabel anzulegen.	En raison de l' interdiction des, PostFinance dans l' environnement de marché actuel marqué par des taux d' intérêt négatif de plus en plus, placer soit rentable les fonds des clients.	En raison de l'interdiction de crédit, il est de plus en plus difficile pour PostFinance d'investir les fonds des clients de manière rentable dans le contexte actuel de taux d'intérêt négatifs.
Der Trend zu mehr Mobilität hält unvermindert an.	La tendance à plus de mobilité normalisera poupe.	La tendance à une plus grande mobilité se poursuit sans relâche.
Zinsergebnis von PostFinance stark unter Druck	Résultat des intérêts de PostFinance s' est fortement sous pression	Revenu net d'intérêts de PostFinance sous forte pression
Poststellen mit Bar Zahlungsverkehr	Offices de poste avec trafic des paiements en espèces	Bureaux de poste avec opérations de paiement en espèces

Die fortschreitende Digitalisierung wird weiterhin grossen Einfluss auf den Geschäftsgang der Post haben.	Subir l' essor du numérique est ont toujours de grande influence sur le chiffre d' affaires de la poste.	La numérisation progressive continuera d'avoir un impact majeur sur l'activité de La Poste Suisse.
Wenn die SNB ihre Negativzinspolitik beendet, werden wir die Guthabengebühr wieder aufheben.	Une fois que la BNS rempli sa politique des taux d' intérêt négatifs, nous la Commission sur avoirs à nouveau supprimerons.	Lorsque la BNS met fin à sa politique de taux d'intérêt négatif, nous annulons les frais de crédit.
Ich mache euch zudem aufmerksam auf den Geschäftsbericht 2016 mit vielen Hintergrundinformationen und allen Details zum Ergebnis 2016 – die Publikation erscheint ebenfalls morgen Donnerstag.	J' attire votre également l' attention sur le rapport de gestion 2016 avec de nombreuses informations de fond et tous les détails et le résultat de 2016, la publication apparaît également demain jeudi.	J'aimerais également attirer votre attention sur le rapport annuel 2016 avec beaucoup d'informations générales et tous les détails sur les résultats 2016 - la publication sera également publiée demain, jeudi.
Das Animationsvideo gibt Einblicke in das Geschäftsjahr 2016 von PostFinance.	La vidéo d' animation donne aperçu de l' exercice 2016 de PostFinance.	La vidéo d'animation donne un aperçu de l'exercice 2016 de PostFinance.
Seit der Ankündigung der Weiterentwicklung des Postnetzes bis 2020 ist die Post auf dem Weg, ein zukunftsfähiges Filialnetz mit einem breiten Angebot von physischen und digitalen Zugangsmöglichkeiten zu schaffen.	Depuis l' annonce le développement du réseau postal ici 2020, il est de la poste sur la voie à créer un réseau de filiales compatible avec le futur proposant une vaste offre de points d' accès physiques et numériques.	Depuis l'annonce de la poursuite du développement du réseau postal d'ici à 2020, La Poste Suisse s'est engagée sur la voie de la création d'un réseau d'agences durable avec un large éventail d'options d'accès physiques et numériques.
Wir können uns zwar weiterhin auf das Geschäft mit klassischen Produkten und Dienstleistungen stützen, doch der Wind ist im vergangenen Jahr deutlich rauer geworden:	Certes toujours nous pouvons nous baser sur les activités traditionnelles de produits et prestations de services, blanc d' œuf l’éolienne est l' an dernier sensiblement dégradé :	Bien que nous puissions continuer à compter sur les produits et services classiques, le vent s'est considérablement aggravé au cours de l'année écoulée :
nicht prioritäre Einzelsendungen Geschäftskunden	les envois isolés non prioritaires clients commerciaux	Expéditions individuelles non prioritaires Clients d'affaires
Der daraus resultierende Rückgang des Erlöses um 7 Millionen Franken konnte durch andere Finanzdienstleistungen nicht kompensiert werden.	Le recul qui en résultent des recettes de 7 millions de francs par d' autres services financiers pas pu être compensé.	Les autres services financiers n'ont pas été en mesure de compenser la baisse du chiffre d'affaires de 7 millions de francs suisses.
Ab 1.1.2016 wurde die Definition auf Poststellen mit Bar Zahlungsverkehr und ohne Barzahlungsverkehr angepasst.	La définition dans les offices de poste avec trafic des paiements en espèces et sans les services de paiement en espèces a été adaptée à partir de 1.1.2016.	Depuis le 1er janvier 2016, la définition des bureaux de poste avec opérations de paiement en espèces et sans opérations de paiement en espèces a été ajustée.
Das Konzerneigenkapital belief sich per 31. Dezember 2016 auf 4‘881 Millionen Franken (vor Gewinnverwendung).	S' élevant des fonds propres consolidés au 31 décembre 2016 4'881 millions de francs (avant répartition du bénéfice).	Les fonds propres du Groupe au 31 décembre 2016 s'élevaient à 4,881 millions de francs suisses (avant affectation du bénéfice).
Dies kostete die Post rund 24 Millionen Franken Negativzinsen.	Cela kostete la poste de 24 millions francs taux d' intérêt négatif.	Cela a coûté à La Poste Suisse environ 24 millions de francs d'intérêts négatifs.

B. Consignes fournies aux post-éditeurs pour l'évaluation humaine 1

La post-édition

La post-édition (PE) consiste à corriger une traduction automatique afin d'atteindre le niveau de qualité souhaité. Dans notre cas, nous souhaitons avoir une traduction d'une qualité suffisante pour la publication.

La post-édition ressemble en quelque sorte à la révision, sauf que les erreurs produites par une machine sont différentes de celles d'un traducteur humain.

Lignes directrices pour le test de PE:

- Utiliser le plus possible l'output de la TA
- S'assurer que la traduction est correcte d'un point de vue grammatical, sémantique et syntaxique
- S'assurer que les termes sont correctement traduits
- S'assurer qu'aucune information n'a été accidentellement ajoutée ou omise
- Les règles de base concernant l'orthographe et la ponctuation s'appliquent
- S'assurer que la mise en forme est correcte

Comment procéder :

- Lire le segment source et ensuite le segment cible
- Si le segment cible ne contient **pas** d'éléments utilisables → effacer la cible et retraduire le segment source
- Si le segment cible contient des éléments utilisables :
    - Corriger le segment
    - S'assurer qu'il ne manque aucune information
    - Vérifier la grammaire et la terminologie
    - Relire et valider

Il ne faut pas trop corriger (ni pas assez), mais s'assurer que la traduction est correcte.

Vous allez parfois tomber sur des segments que vous avez déjà post-édité, post-éditez les à nouveau en vous basant sur la traduction automatique fournie.

MateCat

MateCat est un outil de TAO en ligne qui permet d'utiliser la traduction automatique (TA) pour pré-traduire les documents. Pour des raisons pratiques, les segments que vous allez post-éditer proviendront d'une mémoire de traduction que nous avons créée en utilisant la TA.

MateCat fonctionne comme la plupart des outils de TAO :

- Il présente à gauche le segment source et à droite le segment cible
- Il affiche les suggestions issues de la mémoire de traduction, ainsi que le pourcentage d'analogie
- Pour enregistrer un segment, il suffit de le valider soit en cliquant sur 

, soit en utilisant le raccourci clavier Ctrl+Enter (même raccourci que dans SDL Trados)
- Le raccourci Ctrl+Shift+Enter (bouton 

vous permet de valider le segment et de passer directement au prochain segment non validé).
- Une fois qu'un segment est validé, il est toujours possible de le modifier. Pour ce faire, il suffit de se replacer dans ce segment
- Tant que vous êtes connecté à internet, le projet s'enregistre automatiquement (si votre connexion est interrompue, un message s'affiche en bas à droite)

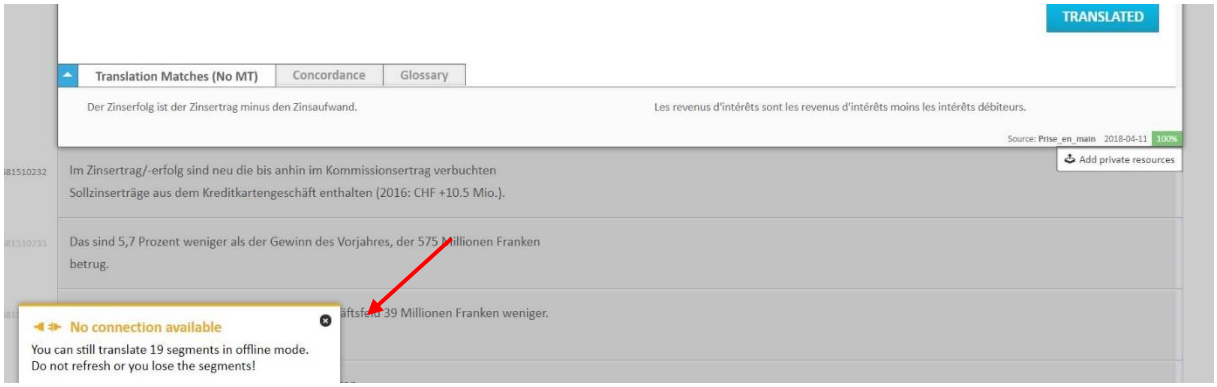
Consignes spécifiques pour notre test

Notre test vise à mesurer l'effort de post-édition afin de comparer la qualité de différents systèmes. Les traductions que vous allez post-éditer viennent de différents systèmes de TA et leur ordre est aléatoire (ce qui signifie que chaque segment est indépendant du précédent et du suivant).

Lorsque vous effectuerez la tâche, MateCat va enregistrer le temps que vous passez à corriger chaque segment et les modifications que vous y apportez. Il est donc très important de bien **fermer la page lorsque vous avez terminé ou lorsque vous interrompez votre travail**. Vous pouvez réaliser le travail de post-édition en plusieurs fois et sur plusieurs jours si vous le souhaitez, mais essayer de rester concentré sur la tâche que vous effectuez à chaque session de post-édition.

Si par mégarde vous oubliez de fermer la page alors que vous avez cessé de post-éditer, merci de me prévenir et de noter le numéro du segment (il se situe à gauche du segment) sur lequel vous vous étiez arrêté afin de ne pas fausser le test.

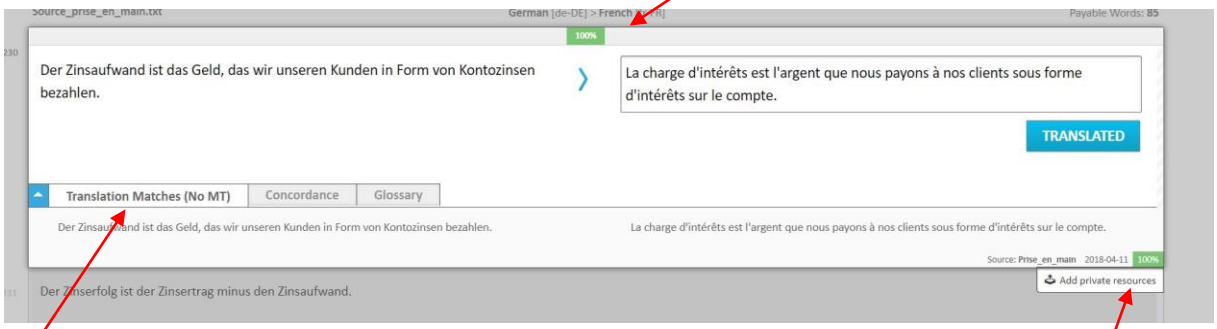
Avant de fermer la page, merci de toujours vérifier qu'aucun message indiquant une coupure d'internet ne s'affiche en bas à droite (voir capture d'écran ci-dessous), sinon votre travail ne sera pas enregistré.



Chaque participant à ce test va recevoir deux liens vers deux projets différents. Vous devez post-éditer ces deux projets, mais vous devez terminer le premier avant de commencer le second.

Comme la TA vient d'une mémoire de traduction, vous devriez toujours avoir des correspondances à 100%. Si jamais vous avez plusieurs correspondances, ne prenez en compte que celle qui à 100% (ou celle qui a le pourcentage le plus élevé). Si aucune correspondance ne s'affiche, traduisez simplement le segment. La capture d'écran ci-dessous, vous indique où trouver les pourcentages d'analogie.

Pourcentage



Permet d'afficher les suggestions de la MT

Pourcentage d'analogie de chaque suggestion

Nous avons créé un projet « prise en main » pour chacun des participants. Ce projet est un projet test qui vous permet de vous familiariser avec l'interface de MateCat avant de commencer la post-édition des deux projets qui vous sont attribués. Merci de prendre un instant pour tester MateCat avec le projet « prise en main (vous n'êtes pas obligé de post-éditer tous les segments, assurez-vous simplement que vous êtes à l'aise avec l'outil).

C. Corpus (extrait) de l’évaluation humaine 2

Référence	Traduction post-éditée à partir de DeepL	Traduction post-éditée à partir de MTH
Pour cela, elle doit continuer de transformer son réseau en s’attachant à répondre aux besoins de la clientèle et à assurer la capacité de financement de ce réseau à long terme.	La Poste doit donc poursuivre la restructuration de son réseau d'offices de poste et se concentrer sur les besoins des clients et la viabilité financière à long terme.	C'est pourquoi la Poste doit continuer de transformer son réseau d'offices de poste et de s'adapter aux besoins des clients ainsi qu'au financement à long terme.
Pour continuer à se développer, la Poste doit disposer d’une marge de manœuvre entrepreneuriale.	La Poste a besoin de toute urgence d'une certaine liberté d'action entrepreneuriale pour poursuivre son développement.	Afin que la Poste puisse continuer à se développer, elle a impérativement besoin d'une certaine latitude entrepreneuriale.
Elle dispose donc d’une solide assise financière.	Cela nous confère une base financière solide.	Notre socle financier est donc sain.
La Poste conserve ainsi une solide assise financière.	La Poste dispose ainsi d'une base financière solide.	La Poste continue donc d'afficher une bonne santé financière.
Nous facilitons les opérations de nos clients dans un environnement complexe et leur permettons de gagner en liberté.	Cela permet à nos clients d'agir plus facilement dans un environnement complexe et leur donne plus de liberté.	Ainsi, nous facilitons la vie de nos clients dans un environnement complexe et leur permettons de gagner en liberté.
Notre environnement évolue à une vitesse fulgurante et nous devons suivre le rythme. C’est pourquoi il nous faut investir dans des projets d’avenir et stabiliser notre résultat à long terme,	Pour suivre le rythme de ce développement rapide, nous devons investir dans des projets d'avant-garde et stabiliser nos revenus à long terme.	Afin de pouvoir suivre cette évolution ultrarapide, nous devons investir dans des projets d'avenir et stabiliser notre résultat à long terme.
Cela montre encore une fois que le service universel fourni par la Poste fait partie des meilleurs en Europe.	Cela montre une fois de plus que la Poste est l'un des meilleurs prestataires de services de base en Europe.	Cela montre une fois de plus que la Poste, avec son service universel, fait partie des meilleurs d' Europe.
Grâce à de bonnes performances sur le plan opérationnel et à des mesures d’amélioration de l’efficacité, PostMail a pu compenser pour une large part ce manque à gagner.	Grâce à la bonne performance de l'activité opérationnelle et aux mesures d'efficacité, PostMail a pu compenser une part considérable de la baisse du bénéfice.	Grâce à de bonnes prestations dans l'activité opérationnelle et aux mesures d'efficacité, PostMail a pu compenser le recul des produits sur une part considérable.

La vidéo d’animation offre un aperçu de l’exercice 2016 de PostFinance sous divers angles.	La vidéo donne un aperçu de l'exercice 2016 de PostFinance.	La vidéo donne un aperçu de l'exercice 2016 de PostFinance.
La vidéo d’animation donne un aperçu de quelques points essentiels du rapport de gestion 2016 de PostFinance.	La vidéo présente les différents jalons de l'exercice 2016 de PostFinance.	La vidéo illustre quelques points essentiels de l'exercice 2016 de PostFinance.
s’établissant à 704 millions de francs, le résultat d’exploitation (EBIT) a reculé de 119 millions par rapport à 2015.	Le résultat d'exploitation (EBIT) est tombé à CHF 704 millions, soit CHF 119 millions de moins que l'année précédente.	Le résultat d'exploitation (EBIT) se chiffre à 704 millions de francs, soit 119 milliards de francs de moins par rapport à l'exercice précédent.
s’élevant à 704 millions de francs, le résultat d’exploitation (EBIT) 2016	Le bénéfice d'exploitation a chuté à 704 millions de francs suisses en 2016.	Le résultat d'exploitation a reculé en 2016 à 704 millions de francs.
Le résultat d’exploitation a reculé de 823 à 704 millions de francs et le bénéfice consolidé de 645 à 558 millions de francs.	Le bénéfice d'exploitation est passé de 823 millions de francs suisses à 704 millions de francs suisses et le bénéfice net de 645 millions de francs suisses à 558 millions de francs suisses.	Le résultat d'exploitation a chuté de 823 millions à 704 millions de francs, le bénéfice consolidé de 645 millions de francs à 558 millions de francs.
Le résultat 2016 a pâti de la baisse des volumes d’envois, de la pression sur les marges et de la faiblesse des taux d’intérêt.	En 2016, les bénéfices sont marqués par la baisse des volumes de consignation, la pression sur les marges et la faiblesse des taux d'intérêt.	Le résultat de 2016 est fortement marqué par le recul des volumes d'envois, la pression sur les marges et les taux d'intérêt bas.
Notre résultat ne me surprend pas – nous avons conçu notre stratégie 2017-2020 en tenant compte de cette mutation structurelle.	Le résultat ne me surprend pas: nous avons adapté notre Stratégie 2017-2020 à ce changement structurel.	Le résultat ne me surprend pas: nous avons axé la stratégie 2017 – 2020 sur ce changement de structure.
Les activités liées aux lettres, aux journaux et aux envois publicitaires sont restées un pilier porteur pour la Poste en 2016,	En 2016, le secteur des lettres, des journaux et des envois publicitaires est resté l'un des piliers de la Poste.	En 2016 aussi, l'activité dans le domaine des lettres, des journaux et des envois publicitaires est un des piliers de la Poste.
Au 31 décembre 2016, les fonds propres consolidés s’élevaient à 4881 millions de francs (avant répartition du bénéfice).	Les fonds propres du groupe, au 31 décembre 2016, s'élevaient à 4881 millions de francs suisses (avant affectation du bénéfice).	Les fonds propres du groupe s'élevaient à 4'881 millions de francs (avant répartition du bénéfice) au 31 décembre 2016.
Le quotidien de nos clients devient plus mobile, flexible et numérique.	La vie de nos clients devient de plus en plus mobile, flexible et numérique.	La vie de nos clients devient de plus en plus mobile, numérique et flexible.

D. Consignes fournies aux évaluateurs pour l'évaluation humaine 2

Chèr(e)s évaluateurs(trices)

Tout d'abord, merci de votre participation.

L'évaluation que vous allez réaliser est une évaluation comparative dans laquelle vous devez déterminer quelle est la meilleure des deux traductions candidates qui vous sont soumises. Les traductions candidates sont des phrases traduites automatiquement puis post-éditées par des traducteurs.

Le document pour l'évaluation se présente comme suit :

- La première colonne contient un identifiant, vous ne devez pas prendre en compte cette colonne
- La deuxième colonne contient une traduction de référence (en bleu)
- La troisième et la quatrième colonnes contiennent chacune une traduction candidate, ce sont ces traductions que vous devez comparer
- Les colonnes suivantes sont nommées trad 1 / équivalentes/ trad 2 (c'est ici que vous indiquerez votre choix)

Vous devez lire les deux traductions candidates et indiquez laquelle, selon vous, est la meilleure. Pour indiquer la meilleure traduction, veuillez simplement mettre un X dans la colonne correspondante. Si vous estimez que les deux traductions sont équivalentes du point de vue de la qualité, alors mettez un X dans la colonne « équivalentes » (colonne du milieu)

La traduction de référence est donnée pour que vous puissiez comprendre le sens de la phrase, mais gardez à l'esprit que **la meilleure traduction candidate n'est pas nécessairement celle qui est la plus proche de la référence.** En effet, les traductions candidates peuvent être formulées de manière différente de la référence tout en étant de bonne qualité. Cherchez plutôt à identifier la traduction candidate qui est la meilleure du point de vue du sens, de la fluidité, de la syntaxe, de la grammaire et de l'orthographe.