Évaluation de la traduction automatique neuronale de textes littéraires de l'italien vers le français

Utushkina, Anastasia

ANASTASIA UTUSHKINA

TRADUCTION AUTOMATIQUE NEURONALE DE LA LITTÉRATURE

Évaluation de la traduction automatique neuronale de textes littéraires de l'italien vers le français

Directrice : Pierrette Bouillon

Jurée : Annarita Felici

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

Université de Genève

Janvier 2023

Table des matières

Remerciements

Je voudrais dédier quelques mots de remerciement aux personnes qui m'ont accompagné dans la rédaction de ce mémoire.

Mes remerciements vont tout d'abord à la directrice de mon mémoire, Mme Pierrette Bouillon, pour sa patience et pour tous les conseils nécessaires pour réaliser ce mémoire dans la meilleure manière possible.

Je tiens également à remercier Mme Annarita Felici pour sa disponibilité, pour avoir gentiment accepté d'être ma jurée et pour avoir consacré son temps à la lecture de ce mémoire, ainsi que pour ses commentaires et conseils.

Mes remerciements vont également aux étudiants de traduction et aux participants non-traducteurs anonymes qui ont participé à ce mémoire et qui ont pris le temps d'évaluer les traductions automatiques.

Je remercie toutes ces personnes, sans lesquelles ce mémoire n'aurait pu être réalisé.

1. Introduction

L'objectif de ce travail de recherche est d'évaluer la qualité de la traduction automatique neuronale lorsqu'elle est appliquée à des textes littéraires, en comparant deux systèmes de traduction automatique neuronale, ceux de DeepL et de Google Translate. Étant donné que les deux systèmes de TA utilisent la méthode d'apprentissage profond (appelée « deep learning » en anglais), il est intéressant de faire une comparaison pour voir les différences et les difficultés de chacun. Notre question de recherche est donc la suivante : « Les systèmes de traduction automatique neuronale sont-ils capables de rivaliser avec une traduction effectuée par des traducteurs professionnels dans le cadre de la traduction d'œuvres littéraires contenant des comparaisons ? ».

La motivation qui nous a poussé à mener cette recherche est le progrès constant dans le domaine de la traduction automatique, qui est appliquée dans des domaines de plus en plus variés, y compris celui de la littérature. C'est pourquoi nous avons décidé de combiner une technologie moderne, à savoir la traduction automatique avec un roman du XIXe siècle écrit par Alessandro Manzoni, « I Promessi sposi » (« Les Fiancés » en traduction française), ce qui ne facilite pas nécessairement le processus de traduction étant donné le langage plutôt archaïque de l'œuvre italienne. De plus, un autre aspect qui rend cette étude particulièrement intéressante est le fait qu'elle se concentre sur la traduction des figures de style, plus précisément des comparaisons. La combinaison linguistique, où l'italien est la langue source et le français la langue cible, constitue également une particularité importante : il s'agit d’une paire de langue moins dotée où la qualité est moins bonne que pour d’autres langues.

Pour réaliser cette étude, nous avons demandé la participation de 18 étudiants en traduction et de 18 personnes de langue maternelle française (non-traducteurs), qui ont évalué la qualité et l'efficacité de la traduction des comparaisons à partir de 120 phrases extraites de « I Promessi sposi », traduites respectivement avec DeepL et Google Translate, puis insérées dans deux types de questionnaires différents. L'objectif de la participation d'évaluateurs humains, en particulier de non-traducteurs, est de répondre à la question « Les traductions automatiques neuronales de comparaisons contenues dans des textes littéraires parviennent-elles à transmettre l'émotion initialement souhaitée par l'auteur ? ».

En outre, afin de disposer d'une évaluation aussi complète que possible, nous avons également procédé à une évaluation automatique des traductions produites par les deux systèmes de traduction automatique, en utilisant le score BLEU (Papineni et al., 2002).

Comme l'affirment plusieurs auteurs, la traduction automatique d'œuvres littéraires est bien plus complexe que les autres types de traduction, puisque l'objectif premier est celui de pouvoir offrir au lecteur une expérience qui soit égale à celle de l'original, en dehors bien sûr de fournir une traduction fidèle du texte (Toral et Way, 2015b).

En dépit des progrès accomplis par la traduction automatique au cours des dernières années, notamment depuis l'apparition de la traduction automatique neuronale, nous pensons qu'elle ne peut pas rivaliser avec la traduction humaine, surtout lorsqu'il s'agit de traduire des œuvres littéraires riches en figures de style, qui impliquent des sentiments et des images qui, à notre avis, sont difficiles à transmettre pour les machines.

Afin de répondre à nos questions de recherche, nous commencerons par examiner l'histoire et la théorie de la traduction automatique en général (Section 2), puis nous aborderons la traduction automatique appliquée à la littérature (Section 3), les méthodes d'évaluation (Section 4) et la méthodologie (Section 5), avant de passer à l'analyse des résultats (Section 6) et à la conclusion (Section 7).

2. La traduction automatique

2.1 Introduction

La traduction a toujours été l’une des activités humaines à la base du progrès, du commerce et des échanges culturels, comme le dit Cronin (2013), chercheur spécialisé dans les études de traduction, qui définit le paradigme des 3T (Technology, Trade, Translation). L'auteur souligne également que l'avènement de l'internet a facilité l'accès à une multitude de pratiques, suscitant ainsi l'attente d'une réponse ou d'une action immédiate. De même, la quantité d'informations et de services a augmenté de manière exponentielle, ce qui a créé le besoin de développer de nouvelles technologies. La globalisation croissante conduit à la nécessité de disposer de

matériel multilingue, d'où une augmentation des activités de traduction et de localisation pour adapter le contenu linguistique et culturel aux nouveaux marchés.

Afin de faciliter et d'optimiser le travail des traducteurs, des outils de traduction assistée ont été développés, notamment des outils de TAO (« CAT tools » en anglais), tels que des mémoires de traduction, des bases terminologiques, des logiciels d'alignement et de concordance. Aujourd'hui, les traducteurs professionnels ne peuvent pas se passer des technologies de traduction, comme le souligne Sin-Wai dans sa publication en 2014 :

« Translation technology is widely used by translation companies as an indispensable tool to conduct their business with high productivity and efficiency, by international corporations as a foundation for their global language solutions, by professional translators as a core component of their personal workstations, and by occasional users as an important means of multilingual information mining. The advent of translation technology has totally globalized translation and drastically changed the way we process, teach, and study translation. » (Chan, 2014, « The Routledge encyclopedia of translation technology », p. xxvii)

Dans les sections suivantes de ce chapitre, nous allons présenter les étapes les plus significatives de l'histoire et de l'évolution de la traduction automatique (Section 2.2), ainsi que ses objectifs et son utilisation actuelle (Section 2.3).

2.2 L’évolution de la traduction automatique

L'idée d'utiliser des machines pour traduire une langue remonte à une époque assez lointaine : les premières idées à ce sujet datent du XVIIe siècle, lorsque les concepts de langue universelle et de dictionnaire mécanique se sont répandus (Hutchins, 2007). L'origine de la traduction automatique telle que nous la connaissons aujourd'hui remonte au XXe siècle, lorsque le français George Artsrouni et le russe Petr Smirnov-Trojanskij ont déposé leurs brevets en 1933. L'invention d'Artsrouni consistait en une sorte de dictionnaire mécanique multilingue, tandis que celle de Trojanskij était un dispositif de traduction multilingue qui utilisait une méthode d'encodage et de décodage des fonctions grammaticales basée sur l'espéranto (Hutchins, 2007).

En 1946, le cristallographe britannique Andrew Booth et le mathématicien américain Warren Weaver ont envisagé la possibilité d'utiliser l'ordinateur, qui était encore une invention récente à l'époque, pour la traduction des langues (Hutchins, 2007). Toutefois, il faut attendre 1949 pour que la traduction automatique fasse un pas en avant et attire l'attention de plusieurs chercheurs, notamment à la suite de la publication du Mémorandum de Warren Weaver. Dans son mémorandum, Weaver s'est appuyé sur ses diverses connaissances en cryptographie, statistiques, théorie de l'information, logique et linguistique pour formuler des suggestions spécifiques, notamment pour résoudre le problème de l'ambiguïté du langage (Hutchins, 2007) et effectuer la traduction.

En 1952, Yehoshua Bar-Hillel a organisé la première conférence sur la traduction automatique à l'Institut de technologie du Massachusetts (MIT), à laquelle ont participé presque tous les chercheurs travaillant dans ce domaine. La démonstration du premier système en 1954 à l'université de Georgetown a donné une impulsion à la recherche : avec un vocabulaire de seulement 250 mots et 6 règles grammaticales, un échantillon sélectionné de 49 phrases a été traduit du russe en anglais (Hutchins, 2007). Cette démonstration a eu un si grand impact sur l'opinion publique qu'elle a stimulé le financement de la recherche sur la traduction automatique aux États-Unis et a conduit au lancement de projets similaires dans d'autres pays, dont l'URSS (Hutchins, 2007).

Les années suivantes ont été marquées par la naissance de plusieurs groupes de recherche en TA dans différentes parties du monde : aux États-Unis, les principales langues de recherche dans le domaine de la traduction automatique étaient l'anglais et le russe, pour des raisons politiques et militaires. Ce n'est que dans un deuxième temps que la recherche s'est étendue à d'autres combinaisons linguistiques (Hutchins, 2007). Dans l'ensemble, les années 1950 ont été caractérisées par un fort optimisme à l'égard de la recherche en TA, de sorte qu'au milieu des années 1960, il existait de nombreux groupes de recherche dans différents pays. Comme le précise Hutchins, en raison des progrès effectués dans le domaine du traitement des données et de la linguistique formelle, pour de nombreux chercheurs de l'époque, la création de systèmes de traduction entièrement automatiques semblait être envisageable (Hutchins, 2007). Cependant, les problèmes linguistiques se sont rapidement révélés de plus en plus complexes et le sentiment d'enthousiasme s'est transformé en une déception croissante (Hutchins, 2007).

En 1964, le gouvernement américain, qui était le principal sponsor de la recherche en traduction automatique, a fait appel à la Fondation nationale des sciences pour créer l'ALPAC (Automatic Language Processing Advisory Committee), qui a élaboré le Rapport ALPAC en 1966 (Hutchins, 2007). Dans ce rapport, le comité ALPAC a décrit la traduction automatique comme étant plus lente, plus coûteuse et moins précise que la traduction humaine, en suggérant que la recherche devrait se concentrer sur le développement de dispositifs d'aide aux traducteurs et que le financement de la recherche en traduction automatique, qui ne semblait pas être une technologie si prometteuse, devrait être arrêté (Hutchins, 2007). Le rapport ALPAC a eu un impact non négligeable sur le monde de la recherche et a causé un ralentissement important de la recherche en matière de la TA.

Parmi les expériences réussies au cours de la décennie qui a suivi l'idéation d'ALPAC, il convient de citer le système Météo canadien, qui a été introduit en 1976 pour traduire les bulletins météorologiques entre l'anglais et le français et qui a été spécifiquement conçu pour traiter le vocabulaire et la syntaxe typiques de ce domaine (Hutchins, 2007). La seconde partie des années 1970 est également caractérisée par la diffusion de systèmes opérationnels et commerciaux, dont par exemple Systran, Logos et METAL, tous fondés sur des règles. Le système Systran notamment a été acheté par la Commission des Communautés européennes en 1976 et a également été installé dans de nombreuses organisations gouvernementales, dont l'OTAN et l'Agence internationale de l'énergie atomique. Ensuite, dans les années 1970 et 1980, des systèmes destinés à des secteurs ou domaines spécifiques ont également été développés (Hutchins, 2007).

Les années 1980 ont été marquées par la prédominance de l'approche par transfert, suivie par le modèle interlangue, et par une intensification remarquable de la recherche en matière de TA au Japon, où un grand nombre d'entreprises du secteur informatique (Fujitsu, Hitachi, NEC, Sharp, Toshiba) a commencé à investir dans cette technologie aussi, ainsi que dans d'autres pays d'Asie, qui ne s'étaient pas encore intéressés à la traduction automatique (Hutchins, 2007).

Depuis 1989, la recherche s'est concentrée sur de nouvelles méthodes de traduction automatique basées sur des corpus, soit de grandes quantités de textes en version numérique (Hutchins, 2007). Ces méthodes comprenaient la traduction automatique basée sur des exemples (EBMT) et la traduction automatique statistique (SMT). La recherche s'est également poursuivie dans le

paradigme des systèmes à base de règles, avec des approches à la fois de transfert et d'interlangue (Hutchins, 2007).

Les années 1980 et 1990 ont également été marquées par le développement et la diffusion de divers outils destinés aux traducteurs, tels que les logiciels de gestion terminologique, les outils de recherche de concordance et les programmes de gestion de mémoires de traduction (Hutchins, 2007). Dans les années 1990, la recherche a également été lancée dans le domaine de l'évaluation de la TA, qui était initialement basée exclusivement sur l'évaluation subjective de divers facteurs, dont la fidélité, la fluidité et la compréhensibilité. Depuis les années 2000, cette évaluation est également réalisée avec des méthodes automatiques ou semi-automatiques, telles que la méthode BLEU (Papineni et al., 2002). Au cours de cette décennie, l'utilisation des systèmes de traduction automatique a également augmenté, notamment de la part des entreprises multinationales et des agences gouvernementales, également grâce à la diffusion progressive des ordinateurs (Hutchins, 2007).

Au cours des années 1990, le marché de la traduction automatique a été influencé par le développement d'Internet, qui a généré une nouvelle demande de traduction. Les services de traduction automatique en ligne, tels que Babel Fish, lancé en 1997 par le moteur de recherche AltaVista, ont fait leur première apparition (Hutchins, 2007). Ce service, basé sur un système Systran, a donné à la traduction automatique une visibilité mondiale, ce qui a permis de la rendre facilement accessible à tous les utilisateurs du réseau. Le lancement de Babel Fish a été suivi par celui de nombreux autres services de traduction automatique en ligne. À l'époque, la qualité des traductions proposées par ce type de plateforme en ligne était plutôt médiocre, mais l'avantage était l'incroyable rapidité du service. Néanmoins, la qualité de la traduction automatique en ligne a laissé une image négative (Hutchins, 2007).

En ce qui concerne l'architecture des systèmes de traduction automatique, depuis la fin des années 1990, l'idée d'un système hybride fait son chemin. Il s'agit d'un système qui combine les méthodes statistiques de la SMT ou de l'EBMT avec des approches basées sur des règles linguistiques (Hutchins, 2007).

La traduction automatique neuronale (NMT), une approche d'apprentissage automatique basée sur les réseaux neuronaux artificiels, a fait son apparition ces dernières années. En effet, en 2016, Google, Systran et Microsoft ont annoncé la naissance de nouveaux systèmes de

traduction automatique. De nos jours, de nombreux services de traduction automatique en ligne sont disponibles, tels que Google, DeepL, Microsoft, Facebook, Amazon, SDL, Yandex et bien d'autres.

Dans la section 2.3, nous procéderons à une description plus approfondie des différents systèmes de traduction automatique.

2.3 Systèmes de traduction automatique

Dans cette section, nous présentons les différents systèmes de traduction automatique en commençant par les systèmes basés sur des règles (Section 2.3.1) puis les systèmes basés sur des corpus (Section 2.3.2).

2.3.1 Systèmes à base de règles

Comme évoqué dans la section 2.2, consacrée à l'évolution de la traduction automatique, les systèmes à base de règles ont été les premiers à être créés et mis en œuvre. La domination de ces systèmes a duré jusqu'aux années 1980, période qui correspond à l'arrivée des systèmes basés sur les corpus (Hutchins, 2007), dont nous parlerons plus en détail dans la section 3.2. Le fonctionnement des systèmes de traduction automatique qui reposent sur une approche fondée sur des règles se fait généralement phrase par phrase, avec l'analyse de chaque mot, en essayant d'identifier la partie du discours du mot et les diverses significations et traductions possibles (Somers, 2011). Dans ces types de systèmes, il existe trois approches différentes : directe, interlangue et transfert.

2.3.1.1 Approche directe

L'approche directe est l'une des premières approches utilisées par les systèmes de traduction automatique à base de règles, conçus pour traduire deux langues. Les systèmes directs sont considérés comme relativement simples (Somers, 2011). Plus précisément, ces systèmes fonctionnent en utilisant un dictionnaire bilingue pour trouver la traduction équivalente d'un mot de la langue source dans le dictionnaire de la langue cible, sans passer par une représentation intermédiaire (Quah, 2006).

La simplicité de ces systèmes est par ailleurs justifiée par le fait qu'ils ne disposent pas d'une connaissance approfondie de la grammaire de la langue vers laquelle ils effectuent la traduction, ni de la structure syntaxique ou des relations sémantiques du texte original (Arnold et al., 1994). Ce que font ces systèmes, consiste à attribuer une catégorie grammaticale à chaque mot, en identifiant les lemmes, grâce à l'utilisation d'un dictionnaire monolingue de la langue source (Hutchins et Somers, 1992). Ensuite, les éléments identifiés sont traduits à l'aide d'un dictionnaire bilingue et réarrangés pour former une phrase dans la langue cible (Hutchins et Somers, 1992).

Parmi les exemples de systèmes de traduction directe citons Systran, METEO et CULT (Quah, 2006).

2.3.1.2 Approche indirecte

Comparée à l'approche directe (section 3.1.1), l'approche indirecte présente une différence principale, représentée par le fait qu'elle crée une représentation intermédiaire, au lieu de traduire directement d'une langue à l'autre (Quah, 2006). Il existe deux types de systèmes qui utilisent l'approche indirecte pour effectuer la traduction et qui diffèrent par le type de représentation intermédiaire qu'ils génèrent : le transfert et l'interlangue.

Traduction automatique par transfert

Les systèmes de ce type utilisent des représentations propres à la langue source, qu'ils transforment ensuite en une représentation dans la langue cible (Trujillo, 1999). En fait, ces systèmes fonctionnent en trois étapes : il y a d'abord l'analyse de la structure de la phrase en langue source, puis le transfert vers la langue cible et enfin la génération de la phrase traduite à partir de l'analyse effectuée lors de la première étape (Jurafsky et Martin, 2009). Des dictionnaires spécifiques sont utilisés à chaque étape, notamment un dictionnaire de la langue source dans l'étape d'analyse, un dictionnaire bilingue dans l'étape de transfert et un dictionnaire de la langue cible dans l'étape de génération (Jurafsky et Martin, 2009).

La distinction entre les connaissances monolingues pendant la phase d'analyse et les connaissances bilingues pendant la phase de transfert permet aux systèmes basés sur le transfert d'être plus faciles à adapter et à maintenir, car un nombre important d'informations peut être réutilisé si la langue source est la même. Cependant, il faut préciser que ce type de système comporte des inconvénients, tels que la nécessité de définir des règles de transfert pour chaque paire de langues (Jurafsky et Martin, 2009).

Traduction automatique interlangue

En ce qui concerne l'approches interlangue, le processus de traduction est effectué en deux étapes (Jurafsky et Martin, 2009). Plus précisément, le texte de la langue source est analysé dans une représentation abstraite, appelée interlangue, représentant le sens et ensuite le texte est généré dans la langue cible à partir de cette représentation interlangue (Jurafsky et Martin, 2009).

Il convient de souligner que la représentation interlangue est indépendante de la langue source ou de la langue cible, elle est donc commune à toutes les langues et permet par conséquent de générer des phrases sources dans plusieurs langues différentes, ce qui facilite la création de systèmes multilingues et l'ajout de plusieurs nouvelles langues (Trujillo, 1999).

Les systèmes interlangues ont besoin d'un niveau d'analyse très élevé en ce qui concerne la sémantique d'un domaine et la formalisation des concepts dans une ontologie. Ils sont généralement utilisés dans des domaines peu complexes, comme les réservations d'un service par exemple, où les définitions contenues dans la base de données indiquant les possibles entités et leurs relations (Jurafsky et Martin, 2009, p. 910).

Le principal problème à surmonter pour un système interlangue consiste à définir une représentation universelle qui puisse inclure toutes les langues et tous les sens d'un mot (Hutchins et Somers, 1992).

Une façon de représenter ces trois approches est le triangle de Vauquois (Figure 1), qui montre la complexité croissante de l'analyse requise quand on passe de l'approche directe au transfert vers l'approche interlangue. Il montre également la diminution de connaissances de transfert

requises à mesure qu'on monte dans le triangle (Jurafsky et Martin, 2009). Le réseau encodeur-décodeur des approches NMT peut être considéré comme une approche interlangue qui permet au décodeur d'accéder directement aux mots ou à leurs représentations (Jurafsky et Martin, 2009).

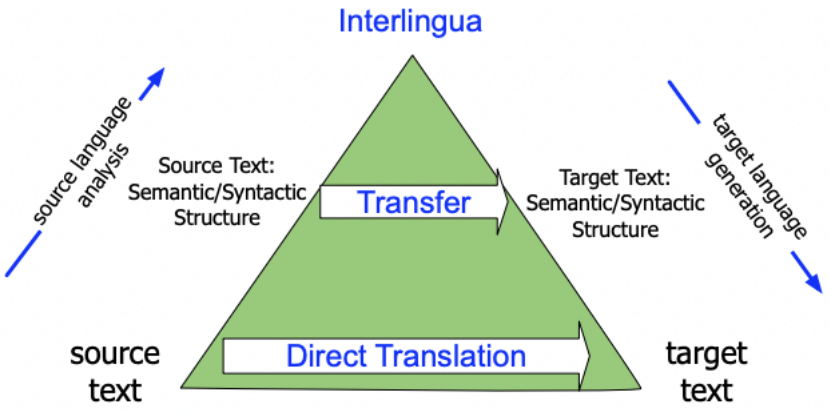

Figure 1 : Triangle de Vauquois (Jurafsky et Martin, 2009, p.27)

2.3.2 Systèmes à base de corpus

Au début des années 1990, la recherche dans le domaine de la traduction automatique a connu une nouvelle étape et s'est intéressée à un autre type de systèmes basés sur des corpus parallèles, qui utilisent des exemples de textes déjà traduits par des humains et à partir desquels de nouvelles traductions pouvaient être réalisées (Somers et Diaz, 2004). Il y a eu une réflexion sur le fait qu'au lieu de formuler manuellement des règles linguistiques avec les exceptions correspondantes, il serait plus facile de donner à la machine des traductions déjà existantes (Koehn, 2020).

Les approches de la traduction automatique qui font partie des systèmes basés sur des corpus sont les suivantes : traduction basée sur des exemples, traduction statistique, traduction hybride et traduction neuronale.

Traduction automatique basée sur les exemples

L'idée centrale de cette approche est de réunir des corpus bilingues de paires de traductions et d'utiliser ensuite un algorithme pour trouver l'exemple le plus proche de la phrase source traitée pour la traduire (Arnold et al, 1994).

Cette approche a été proposée par Nagao qui, en 1984, a utilisé la combinaison linguistique du japonais et de l'anglais et a identifié en même temps les trois éléments principaux de cette méthode de traduction : la mise en correspondance des segments dans une base de données d'exemples réels, l'identification des traductions correspondantes et la recomposition de ces éléments dans le résultat final de la traduction.

Cependant, ce type de système a aussi des inconvénients, comme le fait de dépendre des exemples disponibles. Ainsi, si les phrases de la langue source sont très complexes, il faut ajouter des règles pour générer des phrases correctes (Carl et Way, 2003).

Traduction automatique statistique

Ce type d'approche de traduction automatique génère une traduction basée sur des modèles probabilistes, où les corpus sont utilisés comme source pour trouver la phrase qui est la plus probable d'être la bonne traduction de la phrase concernée (Hearne et Way, 2011).

La traduction automatique statistique utilise deux processus différents, notamment l'apprentissage et le décodage : la phase d'apprentissage consiste à extraire un modèle statistique de traduction à partir d'un corpus parallèle et un modèle statistique de la langue cible à partir d'un corpus monolingue (Brown et al., 1993).

Plus en détail, la traduction automatique statistique utilise un corpus bilingue pour créer un modèle de traduction, qui est une sorte de dictionnaire bilingue contenant toutes les traductions possibles, auxquelles sont attribuées une probabilité en fonction de leur fréquence d'occurrence dans le corpus (Hearne et Way, 2011). En outre, le système crée un modèle de langue qui attribue des probabilités à des séquences de mots de la langue cible, appelées n-grammes, en utilisant un corpus monolingue dans la langue cible (Hearne et Way, 2011). Enfin, le système tente de trouver la traduction qui a la plus forte probabilité selon le modèle de traduction et le modèle de langue parmi toutes les possibilités disponibles (Hearne et Way, 2011).

Traduction automatique hybride

Les systèmes de traduction hybrides ont été conçus dans le but de combiner les avantages des systèmes basés sur les règles et des systèmes statistiques, tout en cherchant à éviter leurs problèmes respectifs (Thurmair 2009).

L'approche hybride peut être utilisée de différentes manières : dans certains cas, les traductions sont effectuées dans la première phase en utilisant une approche basée sur des règles, suivie d'un ajustement ou d'une correction du résultat en utilisant des informations statistiques (Okpor, 2014). Dans d'autres cas, les règles sont utilisées pour traiter préalablement les données d'entrée et entraîner le système avec des données plus riches. Cette méthode est meilleure que la précédente et offre plus de puissance, de flexibilité et de contrôle dans la traduction (Okpor, 2014).

Un exemple de ce type de système hybride est le système de traduction automatique SYSTRAN. Il permet de surmonter les difficultés liées à l'utilisation de corpus parallèles en utilisant un dictionnaire bilingue, similaire à celui utilisé dans la plupart des systèmes de traduction automatique basée sur les règles, et un corpus monolingue de la langue cible (Dirix et al., 2005).

Traduction automatique neuronale

Les systèmes de traduction automatique les plus récents sont les systèmes neuronaux, qui apprennent de manière similaire à un être humain et améliorent leur rendement grâce à l'entraînement (Singh et al., 2017, p. 162).

Un réseau neuronal est une méthode d'apprentissage automatique qui utilise un ensemble d'entrées pour inférer des résultats (Koehn, 2017). Les composants principaux de la traduction automatique neuronale sont l'encodeur, qui traite les mots en entrée et les convertit en une séquence de représentations contextualisées, et le décodeur, qui génère une séquence de mots en sortie (Koehn, 2017). Aujourd'hui, il est fréquent d'entendre parler de ce qui est appelé en anglais « deep learning », ou apprentissage profond, qui met l'accent sur le fait que de meilleures

performances peuvent être souvent obtenues en superposant des couches cachées (Koehn, 2017).

Parmi les éléments principaux des systèmes de traduction automatique neuronale, il y a le word embedding, ou plongement lexical en français, qui constitue la représentation intermédiaire générée par l'encodeur, puis utilisée par le décodeur (Koehn, 2017). Lorsque l'on parle de plongement lexical, on parle d'un espace multidimensionnel dans lequel chaque mot est représenté par un vecteur (Koehn, 2017). Plus en détail, le word embedding correspond à la représentation du sens des mots, qui repose sur le concept selon lequel les mots qui apparaissent dans le même contexte sont similaires (Koehn, 2017). Le plongement lexical rend possible le regroupement des mots sur la base de leur signification, puis des généralisations, ce qui permet au système de TA neuronal de traiter de nouvelles séquences de mots qui n'étaient pas initialement incluses dans le corpus d'entraînement (Koehn, 2017).

- *but the cute dog jumped*
- *but the cute cat jumped*
- *child hugged the cat tightly*
- *child hugged the dog tightly*
- *like to watch cat videos*
- *like to watch dog videos*

Figure 2 : Un exemple illustrant comment « chien » et « chat » peuvent être utilisés dans des contextes similaires (Koehn, 2017, p. 33)

Les exemples qui montrent comment les termes « chien » et « chat » peuvent être utilisés dans des contextes similaires (Figure 2), illustrent comment les deux termes deviennent en quelque sorte interchangeables pour le système de traduction. De cette manière, les plongements lexicaux permettent la généralisation entre les mots (clustering) et donc la prédiction du mot qui pourrait se trouver dans une phrase, même dans des contextes jamais vus auparavant (Koehn, 2017).

Le fait qu'ils soient capables de prendre en compte un contexte très vaste pendant la traduction par rapport aux systèmes statistiques traditionnels, puisqu'ils disposent d'une certaine flexibilité

lorsqu'ils rencontrent des termes inconnus, est l'un des principaux avantages des systèmes de traduction automatique neuronale (Koehn, 2017).

Cependant, par rapport à certains systèmes di traduction automatique statistique, les systèmes neuronaux disponibles aujourd'hui nécessitent une grande puissance de calcul ce qui limite de manière significative la dimension du vocabulaire employé par ces systèmes (Koehn, 2017). En outre, afin d'obtenir un résultat de qualité satisfaisante, l'entraînement des systèmes neuronaux exige une quantité importante de données, tandis que la complexité des systèmes rend complexe l'identification et la correction de certaines erreurs (Koehn, 2017).

La traduction automatique neuronale est devenue l'approche de traduction automatique la plus prospère ces dernières années, enregistrant une adoption rapide dans les implémentations par exemple de Google (Wu et al., 2016).

Le sujet des systèmes de traduction neuronaux de Google Translate et DeepL sera traité plus en détail prochainement, dans les sections consacrées aux systèmes de traduction automatique utilisés dans le cadre de cette recherche (sections 5.2.1 et 5.2.2).

3. Traduction automatique de la littérature

3.1 Introduction

Dans le domaine de la traduction automatique, il y a encore relativement peu de recherches menées sur la traduction d'œuvres littéraires, étant donné l'intérêt récent pour ce type d'étude, comme l'indique par exemple Toral (2020) :

« Most previous studies about MT of literary texts are quite recent (from 2012 onwards). An important fact that has affected the conclusions from the first works to the most recent is the emergence of a relatively new paradigm of MT: Neural Machine Translation (NMT). The great increase in quality of NMT systems, compared to their predecessor, Statistical Machine Translation (SMT), has boosted the interest in MT for literary texts. »

Néanmoins, il faut souligner que, surtout au cours de la dernière décennie, l'intérêt pour la traduction automatique de la littérature a considérablement augmenté, tout comme les études dans ce domaine, surtout depuis le séminaire Computational Linguistics for Literature[1], qui s'est tenu pour la première fois en 2012, et le séminaire plus récent Qualities of Literary Machine Translation[2], qui a eu lieu pour la première fois en 2019.

Dans ce chapitre, nous traitons du développement et de la croissance de l'intérêt pour la traduction automatique appliquée à la littérature. Dans la section 4.2.1, nous présenterons les études qui traitent de la possibilité d'appliquer la traduction automatique aux œuvres littéraires, en examinant si elle a atteint un degré de maturité suffisant et dans quelle mesure elle peut être utile à la traduction littéraire. Dans la section 4.2.2, nous évoquerons les recherches qui se concentrent sur l'étude de la traduction automatique d'un phénomène particulier des œuvres littéraires, comme par exemple la métaphore (Shutova, 2013) ou les locutions (Li et Sporleder, 2009).

3.2 Traduction automatique appliquée à la littérature

Toral, Way (2014, 2015)

En 2014, Toral et Way ont soulevé la question suivante : la traduction automatique est-elle effectivement prête pour la traduction d'œuvres littéraires ? Ils ont donc décidé de comparer le degré de liberté et les limites de la traduction de textes littéraires par rapport aux textes portant sur des éléments techniques ou d'actualité. Un roman traduit de l'espagnol au catalan a été utilisé comme cas d'usage pour cette recherche.

Il a été possible d'observer que le niveau atteint par la traduction automatique entre deux langues proches l'une de l'autre est satisfaisant et que le produit final peut être facilement post-édité : dans 20 % des cas, le résultat des traductions était identique à celui produit par des traducteurs humains spécialisés, tandis que 10 % des cas ne nécessitaient pas plus de 5 modifications. D'autre part, grâce à l'évaluation humaine, il a été constaté que plus de 60% des traductions étaient considérées de qualité équivalente à la TH par les locuteurs natifs. De plus, les auteurs

[1] Source : https://aclanthology.org/N12-1.pdf
[2] Source : https://aclanthology.org/W19-73.pdf

de la recherche ont effectué une analyse qualitative permettant de repérer les types d'erreurs de traduction les plus fréquentes.

En 2015, Toral et Way se sont intéressés à la transmission de l'expérience de lecture dans la version traduite d'une œuvre littéraire, un facteur qui dépend beaucoup des choix de traduction. Dans cette deuxième recherche, le cas d'usage utilisé a été L'Étranger de Camus, traduit en anglais et en italien par Google Translate, qui à l'époque était encore un système statistique. Comme indiqué précédemment, le meilleur résultat et donc le plus approprié pour être post-édité est celui traduit dans une langue plus proche du français, dans ce cas l'italien. L'analyse a été réalisée à deux niveaux, qualitatif et quantitatif.

Comme l'indiquent les deux auteurs, la plus grande difficulté pour les systèmes de traduction automatique est le fait qu'il n'existe pas une seule traduction correcte.

Toral, Wieling and Way (2018)

En 2018, Toral, Wieling et Way ont décidé de traduire automatiquement le premier chapitre d'un roman, Warbreaker, de l'anglais vers le catalan et de le faire post-éditer par six traducteurs spécialisés dans la traduction littéraire parce qu'ils pensaient que la traduction automatique neuronale pouvait traduire des textes riches du point de vue lexical, comme les textes littéraires, et produire des résultats de meilleure qualité que les systèmes de traduction automatique statistique.

Lors de l'analyse de l'effort de post-édition, trois dimensions ont été prises en compte : temporelle, technique et cognitive. La traduction a été effectuée dans trois conditions différentes, c'est-à-dire : sans traduction automatique, par post-édition avec la traduction automatique statistique et par post-édition avec traduction automatique neuronale.

Le résultat de la recherche a permis de déterminer que les deux systèmes ont augmenté la productivité de la traduction (de 18% pour la traduction automatique statistique et de 36% pour la traduction automatique neuronale). Concernant le facteur technique, le nombre de frappes utilisées avec la traduction automatique neuronale a diminué considérablement (de 23% par rapport à 9% avec la traduction automatique statistique). En ce qui concerne le facteur cognitif, il a été constaté que les deux systèmes de traduction ont réduit de manière significative le

nombre de pauses (de 42% dans la traduction automatique neuronale et de 29% dans la traduction automatique statistique).

L'une des principales limites de la traduction automatique dans le domaine littéraire, tant statistique que neuronale, est certainement celle de la transmission de l'expérience de lecture, dont une part importante est assurée par les figures de style et les éléments culturels.

Matusov (2019)

Dans son étude de 2019, Matusov a adapté un système de traduction neuronal pour traduire des œuvres littéraires de l'anglais vers le russe et de l'allemand vers l'anglais. Il faut souligner que les systèmes de traduction adaptés produisent de meilleurs résultats que ceux qui ne sont pas adaptés avec les corpus du domaine. Dans cette étude, Matusov a également proposé un système de classification des erreurs pour les traductions automatiques d'œuvres littéraires.

Avec la participation d'évaluateurs bilingues, il a été constaté que 30% des résultats des traductions automatiques étaient acceptables. En outre, dans le reste des traductions, peu d'erreurs syntaxiques ont été relevés mais les mots ambigus posent de nombreuses difficultés (Figure 3).

German source (F. Kafka)	Human Reference	Google's online MT
Er lag auf seinem panzerartig harten Rücken und sah, wenn er den Kopf ein wenig hob, seinen gewölbten, braunen, von bogenförmigen Versteifungen geteilten Bauch, auf dessen Höhe sich die Bettdecke, zum gänzlichen Niedergleiten bereit, kaum noch erhalten konnte.	He lay on his armour-like back, and if he lifted his head a little he could see his brown belly, slightly domed and divided by arches into stiff sections. The bedding was hardly able to cover it and seemed ready to slide off any moment.	He lay on his panzerartig hard back and saw, if he raised his head a little, his arched, brown, divided by arc-shaped stiffened stomach on the height of the blanket, ready for total descent, could barely maintain.

Figure 3 : Exemple de qualité de la TAN de l'allemand vers l'anglais. En rouge sont marquées les erreurs majeures de la TA (Matusov, 2019).

Encore une fois, comme dans certaines des études illustrées ci-dessus, la traduction faite entre deux langues proches, dans ce cas de l'allemand vers l'anglais, était meilleure que la traduction faite de l'anglais vers le russe.

Fonteyne et al. (2020)

Parmi les études les plus récentes, citons celle réalisée par Fonteyne et al. en 2020, où le roman d'Agatha Christie, The Mysterious Affair at Styles, a été traduit de l'anglais vers le néerlandais à l'aide de Google Translate. Dans le cadre de cette étude, une attention particulière a été accordée aux erreurs de fluidité et de précision, qui sont évaluées à l'aide de la taxonomie des erreurs SCATE MT (2019) élaborée par Tezcan et al. (Section 5.1).

Cette recherche a révélé que 44% des phrases ne comportaient pas d'erreurs, tandis que les problèmes les plus courants étaient dus à des erreurs de traduction, des problèmes de cohérence, des problèmes de style et des problèmes de registre.

3.2.2 Traduction automatique des figures de style

Sporleder, Li (2009)

La recherche réalisée par Sporleder et Li en 2009 est l'un des exemples d'études portant sur le traitement automatique de textes littéraires, où l'accent est mis sur la recherche d'éléments idiomatiques typiques d'une langue particulière.

Dans le cadre de cette recherche, il est proposé une méthode de classification non supervisée qui permet de distinguer l'utilisation littérale et non littérale des expressions idiomatiques. La méthode utilisée est une méthode non supervisée qui détermine les liens cohésifs entre les mots qui composent l'expression cible et son contexte, parvenant à prédire les usages littéraux lorsqu'il y a des liens forts. Il faut également noter que les auteurs sont parmi les premiers à proposer une approche basée sur la cohésion pour détecter les utilisations non littérales (« token-based idiom classification »), alors que la plupart des études de ce genre se concentrent sur la classification « type-based ». L'intérêt d'une approche token-based réside dans le fait que les approches type-based ne sont pas adaptées aux expressions qui peuvent être utilisées aussi bien au sens figuré que littéral, car elles doivent être désambiguïsées en contexte, ce que vise à faire la classification token-based (Li et Sporleder, 2010). Le corpus de langue anglaise de cette recherche a été constitué à partir d'expressions idiomatiques extraites de l'Oxford Dictionary of Idiomatic English (Cowie et al., 1997) et de listes tirées d'Internet.

Grâce à cette étude, il a été possible d'observer que dans le cas de l'utilisation littérale des expressions idiomatiques, il existe un lien cohésif fort avec le discours qui les entoure, contrairement aux utilisations non littérales. En outre, il a été constaté que cette méthode convient également aux expressions ayant une forme canonique, c'est-à-dire une forme fixe (ou un ensemble de formes fixes) correspondant au schéma syntaxique dans lequel l'idiome apparaît normalement (Riehemann, 2001), comme par exemple la forme infinitive d'un verbe qui figure habituellement dans les dictionnaires.

Voigt, Jurafsky (2012)

Dans leur étude de 2012, Voigt et Jurafsky se demandent comment la cohésion référentielle, c'est-à-dire l'ensemble de liens cohérents sur le plan conceptuel qui permettent d'établir des relations entre les éléments de référence dans une phrase, est exprimée dans les textes littéraires et non littéraires et comment cette cohésion influence la traduction.

L'analyse se fonde sur un corpus en langue anglaise, dont la version originale est en chinois. Compte tenu du fait que, par rapport aux textes à vocation informative, les textes littéraires utilisent des chaînes de référence plus complexes pour exprimer une plus grande cohésion référentielle, l'étude a montré que Google Translate ne parvient pas à exprimer la cohésion littéraire aussi bien que la traduction effectuée par un traducteur professionnel. L'étude a révélé que la traduction automatique présente des difficultés à transmettre de manière adéquate la cohésion des textes littéraires, soulignant la nécessité de poursuivre les recherches à ce sujet.

Shutova (2013, 2015)

Comme le souligne Shutova dans ses deux études, les figures de style, notamment les métaphores, sont un élément important d'une langue. Pour cette raison, il devient de plus en plus indispensable d'étudier ce phénomène également avec des techniques du traitement automatique des langues. Dans son étude de 2013, Shutova illustre une nouvelle méthode d'identification des métaphores grâce à leur interprétation, qui consiste à trouver une paraphrase littérale pour exprimer le même concept que celui exprimé avec la métaphore (Figure 4).

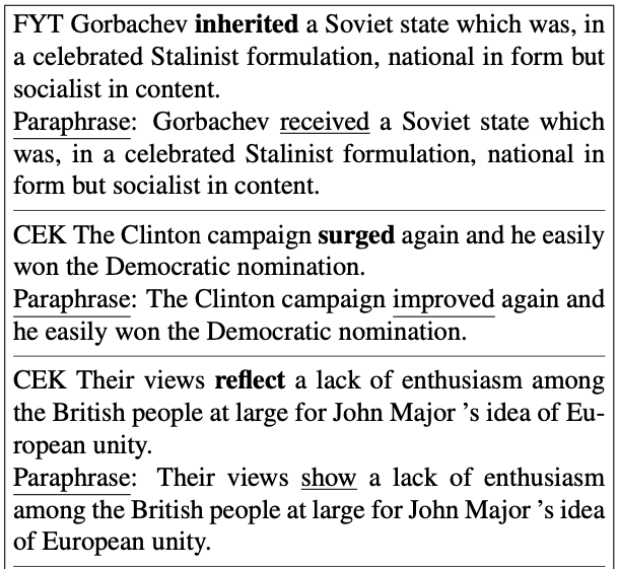

Figure 4 : Exemples de métaphores repérées par le système (en gras) et leurs paraphrases (Shutova, 2013).

Il s'agit ainsi du premier système à combiner les deux processus (identification et interprétation). Il faut noter qu'il s'agit d'une méthode qui ne nécessite pas d'étiquetage manuel des données. Les verbes qui peuvent être utilisés métaphoriquement sont paraphrasés à l'aide de la méthode de paraphrase littérale basée sur le contexte de Shutova (2010). Alors que dans l'étude précédente, Shutova n'a utilisé la méthode que pour paraphraser les métaphores annotées manuellement, dans cette recherche, la méthode a été étendue et appliquée à la paraphrase des termes utilisés littéralement et à l'identification des métaphores, éliminant ainsi la nécessité d'une annotation manuelle des expressions métaphoriques (Shutova, 2013).

Shutova mesure le potentiel que possède un mot à être utilisé de manière métaphorique en fonction de sa force de préférence sélective (SPS), ou « Selectional Preference Strength » en anglais (Resnik, 1993). L'idée principale du filtrage SPS est que tous les verbes n'ont pas le même potentiel à être une métaphore (Shutova, 2013).

Après avoir obtenu la liste initiale des substituts possibles du verbe, le système filtre les termes dont le sens ne partage aucune propriété commune avec celui du verbe. Cette superposition de propriétés est identifiée à l'aide de WordNet (Shutova, 2013).

Le système conçu par Shutova (2013) identifie les métaphores avec une précision de 68 % et un rappel de 66 %, ce qui est un résultat très encourageant (Shutova, 2013).

Comme les résultats de cette première recherche se sont révélés positifs, Shutova a décidé, dans son étude de 2015, d'approfondir le sujet du traitement des métaphores, en étudiant leurs principales caractéristiques et stratégies d'évaluation.

Almahasees (2017)

L'étude réalisée par Almahasees en 2017 porte sur la qualité de la traduction automatique du livre Le Prophète de Gibran traduit de l'arabe vers l'anglais à l'aide de deux systèmes de traduction automatique, Google Translate et Microsoft Bing. La recherche se concentre en particulier sur la façon dont les métaphores et les références culturelles sont traduites, en évaluant les résultats avec le score BLEU.

En raison de la complexité et de la différence des deux langues, ainsi que des différentes manières d'exprimer les concepts, les systèmes ont obtenu de mauvais résultats au niveau des collocations, mais de bons résultats au niveau de la traduction des mots, si considérés sans tenir compte des collocations. Il a été constaté que les deux systèmes donnent souvent la même traduction pour certaines phrases. La recherche de l'auteur a également révélé la présence d'imprécisions sémantiques, ainsi que certaines erreurs lexicales et syntaxiques dans les traductions automatiques. L'auteur de la recherche suggère également l'utilisation d'une méthode d'évaluation humaine, en partant du principe que les résultats pourraient être différents dans ce cas.

Kumari et al. (2021)

Les auteurs suggèrent que souvent les systèmes de TA ne parviennent pas à préserver des informations importantes telles que le sentiment ou l'émotion contenus dans le texte source : l'une des causes est que les expressions qui expriment un sentiment sont traduites par des expressions neutres ou ne figurent pas dans la traduction. Cette recherche vise à concilier le sentiment et la sémantique en utilisant une TAN basée sur l'attention globale, un mécanisme qui se concentre de manière sélective sur des parties de la phrase source pendant la traduction au cours de laquelle tous les mots sources sont concernés (Luong et al., 2015), afin de générer

des traductions qui expriment le sentiment de la phrase source tout en préservant son contenu sémantique.

Les auteurs utilisent une méthode d'apprentissage par renforcement, à savoir une méthode pour l'apprentissage automatique basée sur la récompense des comportements souhaités et la pénalisation des comportements non souhaités (Nguyen et al., 2017), pour affiner le système de traduction automatique afin qu'il apprenne à générer les traductions les plus appropriées pour la tâche.

Pour l'étude, ils ont créé un Sentiment Corpus, réalisé à partir du domaine du tourisme pour la combinaison hindi-anglais, avec des données collectées sur Tripadvisor, tandis qu'ils ont pris des tweets pour la combinaison italien-anglais. Pour évaluer les performances du modèle, les auteurs ont utilisé deux métriques d'évaluation automatique courantes : le score F1 (Chincor, 1992), un indicateur combinant la précision et le rappel, qui mesure si les textes traduits et les textes sources sont cohérents en termes de sentiment, et la métrique BLEU (Papineni et al., 2002), utilisée pour évaluer le maintien du contenu ou le transfert sémantique de la source vers la cible. Afin d'effectuer une analyse des sentiments dans la traduction automatique, ils ont décidé de créer des classes positives, négatives et neutres pour observer les tendances générées par la traduction automatique envers une classe de sentiments. Des traducteurs ont participé à la traduction des phrases, puis des évaluateurs ont évalué la fluidité et l'adéquation des phrases, avant de qualifier les sentiments (en positifs, négatifs, neutres ou autres).

Les résultats pour la paire de langues hindi-anglais ont montré que la méthode proposée améliore de manière significative la performance du système TAN, en comparaison avec la combinaison italien-anglais. Le résultat pour la paire hindi-anglais suggère que l'analyse de sentiments des textes traduits en utilisant un système d'analyse de sentiments en langue anglaise est un choix plus efficace que de faire l'analyse de sentiments du texte source hindi directement. Cependant, pour la classification des tweets italiens-anglais, le score du classificateur spécifique à la langue suggère que le score F1 de la langue cible (anglais) n'a pas dépassé le score du classificateur source. Une raison possible pourrait être la nature des données (tweets avec du bruit) pour l'italien-anglais. En conséquence, la source elle-même est bruyante, ce qui pourrait avoir un impact négatif sur la qualité de la traduction.

Les auteurs ont démontré que leur stratégie d'apprentissage permet d'améliorer le système TAN, ce qui augmente les performances du classificateur de sentiments. En particulier, pour l'hindi-anglais, cela a amélioré les performances du classificateur de sentiments de 3,48 points en termes de score F1 et de 4,57 points en BLEU par rapport au modèle générique. Cependant, pour la traduction italien-anglais, le classificateur spécifique à la langue est plus performant.

Avec ce travail, les auteurs soulignent que lors du développement d'un système de TA, l'accent doit être mis non seulement sur une meilleure traduction du texte source mais aussi sur une meilleure association, en préservant leur relation en termes de sentiment.

3.3 Conclusion

Comme on peut le constater, il existe encore peu d'études traitant de l'application de la traduction automatique aux textes littéraires. Pour cette raison, il a également été difficile de trouver des études traitant de sujets plus spécifiques dans ce domaine, tels que la traduction automatique des figures de style. Néanmoins, ce domaine suscite un intérêt croissant, en raison de l'évolution de plus en plus rapide de la traduction automatique et des ressources technologiques en général. Ce facteur rend notre recherche encore plus intéressante, car elle sera probablement l'une des premières à traiter du phénomène des comparaisons, notamment dans un texte datant du milieu du XIXe siècle et qui représente une importance culturelle majeure pour la littérature et la culture italiennes.

Afin de réaliser cette recherche, nous utiliserons deux systèmes de traduction automatique neuronale, Deepl et Google Translate, dont les résultats seront ensuite évalués automatiquement et également avec la participation d'évaluateurs humains. Nous en parlerons davantage dans les sections qui suivent (Section 5 et 6).

4. Évaluation de la traduction automatique

Dans cette section, nous allons examiner différentes méthodes d'évaluation utilisées pour évaluer la traduction automatique. Il faut souligner qu'il n'y a pas qu'une seule façon d'évaluer la traduction automatique, mais plusieurs, en fonction du sujet et des objectifs de la recherche. L'évaluation de la traduction est une étape très importante, mais aussi très difficile, notamment

parce qu'il existe de nombreuses façons de traduire une phrase (Toral et Way, 2015). Il est souvent utile de combiner plusieurs méthodes d'évaluation.

Dans cette section, nous allons expliquer les principales méthodes d'évaluation, à savoir l'évaluation humaine (Section 5.1) et l'évaluation automatique (Section 5.2), les deux méthodes d'évaluation utilisées dans notre étude sur la traduction automatique des comparaisons contenues dans I Promessi Sposi.

4.1 Évaluation humaine

Ce type d'évaluation, comme son nom le suggère, est réalisé avec la participation de juges humains. On peut classer les principales méthodes d'évaluation humaine en trois catégories, à savoir le jugement intuitif, l'analyse d'erreurs et l'évaluation comparative (Koehn, 2009).

En ce qui concerne la méthode d'évaluation basée sur le jugement intuitif, les deux critères à prendre en compte pour la qualité sont la fidélité et la fluidité. La fidélité se rapporte au sens de la phrase, au fait qu'elle soit correctement exprimée et transmise par la traduction. Dans ce cas, on fournit généralement à la fois la traduction et la phrase dans la langue d'origine. La fluidité, par contre, se rapporte au fait que la phrase traduite soit compréhensible et facile à lire. Toutefois, dans ce cas, seule la phrase traduite est fournie pour l'évaluation et pas la source (Koehn, 2020).

Pour mettre en place ces méthodes, différents aspects doivent être pris en compte. L'un d'entre eux concerne les juges, plus précisément le choix des personnes chargées d'évaluer la traduction. Le choix des juges dépend de l'aspect que nous voulons évaluer : dans le cas de l'évaluation de la fidélité, nous avons besoin de juges bilingues qui auront accès à la version originale de la traduction. Il est aussi possible de faire appel à des juges monolingues qui devront recevoir une traduction humaine à titre de référence (Kit et Wong, 2015), ce qui est fortement déconseillé (Toral et Way, 2015).

Si on décide d'utiliser la méthode d'analyse des erreurs, il faudra que les juges comptent les erreurs à corriger et les classent. Afin de classer ce type d'erreur, il existe des métriques spécifiques, en fonction du domaine d'utilisation, comme le modèle LISA QA dans le domaine

de la localisation (Mateo, 2014). Dans le domaine de la recherche sur la traduction automatique d'œuvres littéraires, il a été mis au point un système spécifique de classification des erreurs détectées. Il s'agit de la taxonomie des erreurs SCATE (Smart Computer-aided Translation Environment), qui concerne deux aspects liés à la qualité de la traduction automatique, soit la fluidité et la fidélité (Tezcan et al., 2017). Ces métriques permettent de classer chaque erreur de manière plus précise et objective, en donnant à chacune un poids différent. Un aspect à prendre en compte est l'interprétation et la tolérance différentes qu'un juge peut donner à l'erreur. Il faut également prendre en compte le fait que certaines erreurs peuvent appartenir à plus d'une catégorie et que le type de catégorie auquel chaque erreur est attribuée peut varier d'une personne à l'autre de manière subjective (Kit et Wong, 2015).

Évaluer une traduction automatique au moyen d'une méthode de comparaison peut parfois sembler plus facile que d'évaluer la fluidité ou la fidélité (Koehn, 2009). Cette méthode d'évaluation consiste à mettre les phrases traduites dans un ordre de préférence, en choisissant la traduction qui est préférée. Il a été constaté qu'en utilisant ce type d'évaluation, il y a un plus grand accord entre les juges (Koehn, 2009). Cependant, il y a aussi un aspect problématique, car l'évaluation comparative nous permet de comprendre quelle traduction est meilleure que l'autre, mais elle ne nous aide pas à comprendre les forces et les faiblesses d'un système.

Les méthodes d'évaluation automatique sont une alternative si l'on ne veut pas recourir à l'évaluation d'une traduction automatique qui nécessite la participation de personnes.

4.2 Évaluation automatique

Dans cette section, nous allons illustrer les différentes méthodes d'évaluation automatique des traductions, dont le principe de base est toujours le même pour les différentes approches, notamment celui de comparer une traduction automatique avec une ou plusieurs traductions de référence créées par un humain.

Parmi les approches d'évaluation automatique, il y a la précision et le rappel. Pour calculer la précision, il faut diviser le nombre de mots corrects dans la référence par le nombre total de mots générés par la traduction automatique, tandis que le rappel est calculé en divisant le nombre de mots corrects par le nombre de mots de la traduction de référence (Koehn, 2009).

$$\text{precision}(C|R) = \frac{\text{MMS}(C,R)}{|C|}$$
$$\text{recall}(C|R) = \frac{\text{MMS}(C,R)}{|R|}$$

Figure 5 : Formules de calcul de la précision et du rappel (Turian et al., 2006).

Dans la formule utilisée pour calculer la précision et le rappel (Figure 5), « C » correspond au texte candidat, « R » au texte de référence, tandis que « MMS » indique le nombre des mots en commun (« Maximum Match Size » en anglais).

Il est ainsi possible d'observer le bruit, soit les mots erronés, ainsi que le silence, c'est-à-dire les mots qui n'ont pas été générés par le système de TA. Il convient également de souligner que ces métriques d'évaluation ne tiennent pas compte de l'ordre des mots, ce qui constitue un inconvénient qui les rend inadaptées à l'évaluation des traductions automatiques, car elles peuvent qualifier une traduction correcte mais différente comme étant incorrecte.

Une autre méthode d'évaluation automatique est le Word Error Rate (WER), qui est aussi basé sur les mots, mais qui, contrairement à la méthode précédente, prend également en compte l'ordre des mots d'une traduction. Le WER calcule le nombre minimum de modifications nécessaires à apporter à une phrase issue d'un système de traduction automatique pour qu'elle devienne identique à la traduction de référence (Koehn, 2009). Le nombre de modifications est ensuite divisé par le nombre de mots de la traduction de référence.

$$WER = \frac{S+I+D}{N}$$

Figure 6 : Formules de calcul du WER (Dorr et al., 2009)

Dans la formule de calcul du WER (figure 6), « S » représente le nombre de substitutions effectuées, « D » le nombre de suppressions, « I » le nombre de mots ajoutés et « N » le nombre de mots dans la référence.

En ce qui concerne l'interprétation du résultat, plus il est petit, meilleure est la traduction car il y a moins de changements à mettre en œuvre. L'un des problèmes les plus significatifs de cette approche est le fait que le WER a très souvent tendance à méconnaître des phrases qui sont en réalité correctes, mais qui sont traduites d'une manière différente de celle proposée par la traduction de référence (Gerlach, 2015).

Une méthode d'évaluation similaire au WER décrit ci-dessus est le Translation Edit Rate (TER), qui consiste à comparer le résultat de la traduction automatique avec la traduction de référence et à calculer le nombre de modifications nécessaires pour obtenir une traduction égale à la traduction humaine (Dorr et al., 2009).

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

Figure 7 : Formules de calcul du TER (Dorr et al., 2009)

Plus précisément, pour calculer le TER (Figure 7), il est nécessaire d'appliquer la formule montrée plus en haut, qui divise le nombre d'éditions par le nombre total de mots contenus dans la référence.

Un aspect qui distingue le TER du WER est le fait qu'il prend également en compte les déplacements de séquences de mots, auxquels accorde un poids égal aux autres éditions (Snover et al., 2006). C'est précisément pour cette raison que le score TER peut différer du score WER, en comptant le changement d'une séquence de mots comme s'il s'agissait d'un seul changement, alors que le WER aurait compté un changement par mot.

Nous concluons la description des différentes méthodes d'évaluation automatique par la métrique BLEU, à savoir Bilingual Evaluation Understudy (Papineni et al., 2002), une approche largement utilisée dans le domaine de l'évaluation de la traduction automatique, que nous avons également utilisée nous-mêmes dans le cadre de ce travail de recherche.

BLEU est une métrique qui compare une traduction automatique avec une référence, mais contrairement aux méthodes vues précédemment, elle ne prend pas en compte les mots comme unités de mesure, mais les N-grammes (Papineni et al., 2002). L'idée principale de cette

métrique est basée sur la similarité entre la traduction automatique et celle produite par un traducteur professionnel, ce qui signifie que plus la première est similaire à la seconde, meilleure est la traduction (Papineni et al., 2002). L'avantage du BLEU est qu'il ne vérifie pas seulement si un mot est présent dans la traduction de référence, mais qu'il prend aussi en considération le nombre maximum de fois qu'un mot y est présent (Papineni et al., 2002). Cette même précision modifiée pénalise également les phrases lorsqu'elles sont plus courtes que la référence, par le biais de la brevity penalty (Koehn, 2009).

Un autre aspect à souligner est que le BLEU peut être calculé en utilisant plus d'une traduction de référence, ce qui augmente le niveau de tolérance du calcul du score, puisque la traduction est évaluée en la confrontant à plusieurs versions (Papineni et al., 2002).

La valeur produite par BLEU est toujours un nombre compris entre 0 et 1, une valeur qui sert à indiquer le degré de similarité entre la traduction automatique et la traduction de référence (Papineni et al., 2002). Pour calculer le score BLEU, le nombre de N-grammes de différents types dans la traduction de référence est divisé par le nombre total de N-grammes de chaque type dans la traduction automatique (Koehn, 2009).

$$BP = \begin{cases} 1, & if\ c > r; \\ e^{\left(1-\frac{r}{c}\right)}, & if c < r. \end{cases}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

Figure 8 : Formules de calcul de BLEU (Papineni et al., 2002).

Pour calculer le score BLEU (Figure 8), en premier lieu il faut donc calculer la pénalité de brièveté, « BP », pour pouvoir ensuite calculer le score BLEU. Ici, « r » est la longueur effective du corpus de référence, « c » la longueur totale du corpus de traduction, « Pn » est la moyenne géométrique de la précision modifiée des n-grammes et « N » la longueur des n-grammes utilisés pour calculer « Pn ».

Pour comprendre comment interpréter le score BLEU, il suffit de regarder son score : plus le score est élevé, plus le système est bon par rapport à un autre.

La métrique d'évaluation BLEU présente toutefois aussi des inconvénients. Il ignore le degré de pertinence des différents mots : en effet, certains mots ont plus d'importance que d'autres. L'un des exemples les plus marquants est le mot « not », qui entraîne des traductions erronées s'il est omis (Koehn, 2009). De plus, le score BLEU intervient uniquement à un niveau local et ne tient pas compte de la cohérence grammaticale globale. Le résultat du système peut sembler satisfaisant sur une base de n-grammes, mais reste très imprécis par la suite (Koehn, 2009). Le score BLEU en soi ne représente rien, car il varie en fonction de nombreux facteurs, tels que le nombre de traductions de référence, la paire de langues, le domaine et même le système de tokenisation utilisé (Koehn, 2009). Des études ont permis de calculer les scores BLEU obtenus par des personnes, c'est-à-dire qu'une traduction de référence humaine a été évaluée par rapport à d'autres traductions de référence humaines. Ces scores BLEU humains sont presque aussi élevés que les scores BLEU calculés pour les traductions automatiques, même si les traductions humaines sont nettement de meilleure qualité (Koehn, 2009).

À cause de ces aspects négatifs, des variantes et des extensions du BLEU ont été proposées (Koehn, 2009). Un de ces exemples est METEOR (section 3.1), qui considère aussi les variantes morphologiques des mots, déterminés avec WordNet, une ontologie de mots anglais qui existe aussi pour d'autres langues (Koehn, 2009). Au contraire de BLEU, METEOR utilise les lemmes et les synonymes prenant en considération les correspondances entre les mots qui sont des variantes morphologiques avec le même lemme et les mots qui sont des synonymes (Koehn, 2009). Le principal inconvénient de METEOR est que sa méthode et sa formule pour calculer un score sont beaucoup plus complexes que celles de BLEU. Le processus de mise en correspondance implique un alignement des mots coûteux en termes de calcul. Il existe bien d'autres paramètres, comme le poids relatif du rappel et de la précision, la pondération des correspondances de déformation ou des synonymes, qui doivent être réglés (Koehn, 2009). Malgré les différents points faibles mis en évidence, le score BLEU est l'une des méthodes d'évaluation de la traduction automatique les plus utilisées (Koehn, 2009).

4.3 Conclusion

Dans les sections précédentes, nous avons illustré les méthodes d'évaluation humaine et automatique appliquées à la traduction automatique. Il ne s'agit pas d'une liste exhaustive, mais seulement de quelques exemples des méthodes les plus utilisées dans ce domaine. Il faut

souligner que, pour avoir une évaluation plus appropriée d'une traduction automatique, il est préférable d'utiliser les deux méthodes d'évaluation. Comme précisé par Lee et al. dans une étude publiée en 2021 (Vol. 67), les deux méthodes d'évaluation se complètent :

« Most of the widely-used automatic metrics (BLEU, ROUGE, BertScore, etc.) aim to measure humanlikeness. When the goal is to assess real-world NLG applications, evaluations should ultimately address the "usefulness" of the system. This is something that humanlikeness metrics are typically incapable of, and for which human evaluation remains the gold standard ».

Si l'évaluation humaine est plus précise et parvient à saisir des éléments difficiles à cerner par l'évaluation automatique, cette dernière est plus rapide et moins coûteuse en termes de coût et de ressources humaines. En fait, nous utiliserons les deux méthodes dans le cadre de notre travail de recherche.

5. Méthodologie

5.1 Objectifs de la recherche

L'objectif de ce travail de Master est de comparer la traduction automatique neuronale d'un texte littéraire de l'italien vers le français, réalisée par deux systèmes différents, DeepL et Google Translate, afin de comprendre dans quelle mesure ces systèmes de traduction sont capables de rivaliser avec la traduction humaine. Plus précisément, la recherche se concentrera sur les figures de style présentes dans le texte, les comparaisons, en accordant une attention particulière à leur efficacité à transmettre le message et l'émotion voulus par l'auteur. Nous avons donc formulé la question de recherche comme suit :

Les systèmes de traduction automatique neuronale sont-ils capables de rivaliser avec une traduction effectuée par des traducteurs professionnels dans le cadre de la traduction d'œuvres littéraires contenant des comparaisons ?

En outre, nous avons également formulé une sous-question pour notre étude, en nous intéressant à l'intensité émotionnelle d'une traduction littéraire, qui a commencé à susciter l'intérêt des études assez récemment (Toral, 2020) :

Les traductions automatiques neuronales de comparaisons contenues dans des textes littéraires parviennent-elles à transmettre l'émotion initialement souhaité par l'auteur ?

En outre, nous voulons également savoir si les comparaisons utilisées dans le texte original ont été traduites littéralement ou si elles ont été adaptées avec d'autres expressions, et si elles sont encore utilisées aujourd'hui ou si elles sont devenues obsolètes.

Pour conduire cette étude, nous avons réalisé deux questionnaires en ligne contenant des extraits du roman I Promessi Sposi d'Alessandro Manzoni, destinés à des étudiants en traduction. Les deux questionnaires utilisent des échelles de Likert (Likert, 1932). Dans l'un des questionnaires, il est demandé aux étudiants d'évaluer avec une approche professionnelle les traductions réalisées par DeepL et Google Translate en les comparant à la version originale. Dans le second questionnaire destiné à des francophones, les participants sont invités à ne prendre en compte que la phrase traduite contenant la comparaison, dans le but d'évaluer l'efficacité de la figure de style sur le plan émotionnel. En outre, nous avons également voulu utiliser la méthode d'évaluation automatique BLEU (Papineni et al., 2002), afin d'avoir un type d'évaluation différent de l'évaluation humaine et ainsi un point de vue différent.

Les évaluations humaine et automatique seront effectuées à partir de la traduction réalisée par le marquis de Montgrand en 1877 et publiée par l'éditeur Garnier frères, intitulée Les Fiancés dans sa version française.

Dans les chapitres suivants, nous examinerons les systèmes utilisés pour cette recherche (Section 6.2), puis nous fournirons des informations contextuelles sur le roman utilisé pour cette recherche et la raison de ce choix (Section 6.3), ensuite nous expliquerons comment les extraits contenant les comparaisons ont été sélectionnés (Section 6.4), nous discuterons de la manière dont la traduction automatique des comparaisons a été effectuée (Section 6.5) et de la procédure d'évaluation humaine menée par les participants au moyen de questionnaires (Section 6.6).

5.2 Systèmes utilisés pour cette recherche

Tandis que dans le chapitre 3 nous avons présenté les types de systèmes de traduction automatique, dans cette partie nous présenterons les deux systèmes de traduction automatique neuronaux utilisés dans le cadre de cette recherche, en mettant l'accent sur les détails nécessaires pour bien comprendre leur fonctionnement et l'intérêt qu'ils représentent pour cette recherche. Nous aborderons tout d'abord le système de traduction automatique DeepL (section 6.2.1), puis nous passerons au système Google Translate (section 6.2.2).

5.2.1 DeepL

Deepl Translate est un service de traduction automatique allemand de la société DeepL, fondée en 2009 sous le nom de Linguee et basée à Cologne. Depuis août 2017, ce système de traduction automatique neuronale (Section 3.2) est disponible en ligne et peut être utilisé gratuitement. En mars 2018, la version payante DeepL Pro a été lancée, apportant différentes améliorations notamment la possibilité d'intégration dans des outils dédiés à la traduction professionnelle, comme par exemple SDL Trados Studio et d'autres outils de TAO. Le traducteur est disponible dans 26 langues actuellement (février 2022), et la langue source est détectée automatiquement.

DeepL Translate a été entraîné avec la mémoire de traduction de Linguee, un service qui est toujours disponible aujourd'hui. La société a développé un système de calcul à haute performance de 5,1 pétaflops sur le campus du centre de données de Verne Global en Islande. Selon une étude[3] réalisée par DeepL en 2020, le système a battu la performance de ses principaux concurrents sur le marché, tels que Google, Amazon et Microsoft.

5.2.2 Google Translate

Google Translate est un service de traduction automatique neuronale, même si à l'origine, il s'agissait d'un système de traduction automatique de type statistique (Section 3.1). Le service de traduction a été mis à disposition pour la première fois en 2006, pour ensuite évoluer vers un système avec une approche neuronale basée sur le deep learning en 2016[4], qui supporte aujourd'hui 109 langues.

[3] Source : https://www.deepl.com/en/quality.html
[4] Source : https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

5.3 Informations contextuelles sur le roman analysé

I Promessi Sposi (en français Les Fiancés) est un roman historique italien d'Alessandro Manzoni, publié dans sa première version en 1827, puis révisé et édité pour être publié dans sa version finale entre 1840 et 1842.

La narration se déroule dans la région de la Lombardie, en Italie, en 1628, sous le règne espagnol. Toute l'intrigue tourne autour de l'histoire d'amour entre Renzo et Lucia, qui tentent de se marier malgré les difficultés qu'ils rencontrent tout au long de l'histoire.

Ce roman, très populaire dans la littérature italienne, aborde de nombreux thèmes, souvent avec un ton ironique. Comme l'écrit Riccardi en 2011, on dit souvent que Les Fiancés sont un élément central de la culture italienne, car cette œuvre a joué un rôle décisif dans la consolidation et la diffusion de la langue italienne. Manzoni a choisi de recourir à une langue parlée non seulement par l'aristocratie érudite, mais aussi par les gens les plus pauvres et les plus humbles. Ce roman italien peut aussi être qualifié comme une œuvre de grande importance en Europe.

5.4 Sélection des comparaisons à traduire

Dans le cadre de cette recherche, nous avons choisi d'analyser non seulement une œuvre littéraire du XIXe siècle, ce qui pose déjà de plus grandes difficultés qu'une œuvre littéraire moderne du point de vue de la langue, mais nous avons décidé de nous concentrer sur la traduction des comparaisons. Comme le roman est très connu et étudié en Italie, également pour sa richesse en figures de style, nous avons choisi de recourir aux documents qui en traitent pour sélectionner les comparaisons utilisées dans notre recherche : se sont révélés notamment très utiles les études en italien par Maria Gabriella Riccobono en 2015 et 2017, où l'on trouve listées dans un recueil les comparaisons contenues dans le roman, dans deux volumes différents. Grâce à cette démarche, il a été possible de simplifier le processus de constitution du corpus, en évitant de devoir effectuer une recherche manuelle dans le roman et en permettant de prendre en compte des comparaisons issues de plusieurs chapitres différents du roman. Nous avons sélectionné les segments à l'aide de marqueurs, en choisissant ceux contenant le mot « come » qui introduit la comparaison. Dans le cadre de cette étude, nous avons sélectionné un total de 120 phrases qui contiennent des comparaisons. La figure 9 montre trois exemples issus des 120 phrases

sélectionnées dans le cadre de notre étude, avec leur numéro d'identification et la comparaison en question soulignée en gras.

12. Dopo la voltata, la strada correva diritta, forse un sessanta passi, e poi si divideva in due viottole, **a foggia d'un ipsilon**.
68. Ma più sconcia era la figura della donna: un pancione smisurato, che pareva tenuto a fatica da due braccia piegate: **come una pentolaccia a due manichi**.
85. Il vicario scendeva le scale, mezzo strascicato e mezzo portato da altri suoi servitori, bianco **come un panno lavato**.

Figure 9 : Quelques exemples de phrases sélectionnées.

5.5 Traduction automatique et évaluation des comparaisons

Dans l'étape suivante, une fois les 120 phrases traduites dans les systèmes de traduction automatique respectifs de DeepL et Google Translate (tâche effectuée les 12 et 13 février 2022), elles ont également été placées dans deux fichiers distincts, formant ainsi nos deux corpus en langue française.

Cette section explique en détail la méthode utilisée pour évaluer les phrases traduites par les deux systèmes de traduction, méthode que nous avons différenciée ici en deux sections, dont l'évaluation humaine réalisée avec la collaboration des participants (Section 5.5.1) et l'évaluation automatique réalisée au moyen du score BLEU (Section 5.5.2).

5.5.1 Évaluation humaine

Pour procéder à l'évaluation, nous avons créé deux types de questionnaires avec Microsoft Forms, le premier destiné aux étudiants en traduction maîtrisant aussi bien l'italien que le français, et le second conçu pour les participants de langue maternelle française, qui ne devaient pas nécessairement être spécialisés dans la traduction. En outre, pour chaque traduction, la version originale italienne de la phrase traduite en français était également incluse. Contrairement au premier questionnaire, celui destiné aux participants francophones n'incluait pas la version italienne originale de la phrase traduite, mais se composait uniquement des phrases traduites par DeepL et Google Translate, où la comparaison prise en compte est soulignée en gras. Le temps de réponse à chaque questionnaire était d'environ 30 minutes.

Lors de notre recherche, 36 personnes au total ont participé à la traduction des comparaisons sélectionnées (Figure 10). Sachant que pour avoir une prépondérance d'opinion et observer une majorité, il faut un minimum d'au moins 3 personnes par phrase et considérant qu'il y avait 240 phrases à évaluer au total, afin de les répartir, nous avons créé 6 questionnaires comprenant chacun 20 phrases. Chaque questionnaire de 20 phrases par système a été rempli par 3 personnes, soit un total de 18 personnes pour la totalité des questionnaires destinés au public bilingue et 18 personnes pour les questionnaires créés pour les personnes monolingues françaises.

	Quantité des questionnaires	Phrases DeepL (par questionnaire)	Phrases Google (par questionnaire)	Participants (par questionnaire)	Total des phrases	Total des participants
Questionnaire bilingue	6	20	20	3	240	36
Questionnaire monolingue	6	20	20	3	240	36

Figure 10 : Tableau récapitulatif des participants et des phrases contenues dans les questionnaires.

En ce qui concerne la recherche des participants, elle a été principalement menée par le biais de plateformes sociales telles que LinkedIn, Instagram, Facebook et WhatsApp. Par ailleurs, dans le cadre de la recherche de participants pour le questionnaire bilingue, un contact direct a également été établi avec certains étudiants de bachelor et de master de la Faculté d'Interprétation et de Traduction de l'Université de Genève.

Afin de ne pas confondre ou induire en erreur les participants, nous avons attribué des numéros d'identification aux questions, sans préciser par quel système chaque phrase avait été traduite. De plus, toujours dans le but de ne pas permettre de distinguer un système de traduction ou un autre, les phrases ont été distribuées dans le questionnaire dans un ordre aléatoire.

Dans le but d'essayer de focaliser l'attention des évaluateurs de manière ciblée sur la traduction de la figure de style et non sur le reste du segment, nous avons décidé de mettre en évidence la comparaison contenue dans chaque phrase en la soulignant.

12A. Après le virage, la route était droite, peut-être à soixante pas, puis se divisait en deux voies en forme d'ipsilon.
68B. Mais plus obscène était la figure de la femme : un ventre énorme, qui semblait difficilement retenu par deux bras croisés : comme une pentolaccia à deux anses.
85A. Le vicaire descendit les escaliers, moitié traîné, moitié porté par ses autres serviteurs, blanc comme un linge lavé.

Figure 11 : Exemples de phrases traduites par des systèmes de traduction automatique contenues dans les questionnaires, où la comparaison est soulignée.

Dans les deux questionnaires, nous avons utilisé la méthode d'évaluation de Likert, qui consiste à définir un certain nombre d'affirmations exprimant un avis positif ou négatif à l'égard d'un énoncé. Pour chaque affirmation, il existe une échelle permettant au participant d'exprimer son accord ou son désaccord, en utilisant une série de notes allant de 1 à 5. Les deux questionnaires étant destinés à deux publics différents, donc les affirmations à noter étaient également différentes.

Questionnaire pour les étudiants bilingues spécialisés en traduction
1. La comparaison est compréhensible dans le contexte de la phrase.
2. Vous traduiriez la comparaison de la même manière.
3. Le registre est adéquat.
4. La traduction choisie pour la comparaison transmet le message de l'auteur.
5. La traduction est fluide.

Figure 12 : Questions du questionnaire destiné aux étudiants bilingues spécialisés en traduction

Questionnaire pour les participants de langue maternelle française
1. La comparaison semble être traduite par un traducteur professionnel.
2. Je comprends le sens général de la phrase.
3. La comparaison parvient à éveiller mon imagination.
4. La comparaison me semble appropriée dans le contexte.
5. L'image véhiculée par la comparaison est appropriée et explicite l'idée.

Figure 13 : Questions du questionnaire destiné aux participants de langue maternelle française

Les figure 12 et 13 ci-dessus reprennent les énoncés utilisés pour chaque questionnaire, tandis que des copies des questionnaires sont disponibles dans les annexes de ce document de recherche. En outre, chaque questionnaire contient comme premier élément, à titre d'information, un formulaire de consentement avec une description de la recherche, les instructions relatives à la tâche à accomplir par les participants, les informations liées à la confidentialité et à l'anonymat des donnés, ainsi qu'une adresse électronique de contact pour les éventuelles questions. Précisons que, afin de prendre part au questionnaire, chaque participant a dû confirmer qu'il était informé des conditions de la recherche et les accepter.

5.5.2 Évaluation automatique

En ce qui concerne l'évaluation automatique des 120 segments, nous avons choisi d'utiliser le score BLEU (Section 5.2) comme méthode de calcul. Pour cela, il a fallu créer un quatrième corpus, contenant les segments traduits en français à partir d'une traduction officielle du roman. La source de la traduction française est la version traduite par le Marquis de Montgrand en 1877, publiée par l'éditeur Garnier frères, disponible en ligne au format PDF.

Après la création des quatre corpus, respectivement un pour la version originale en italien, un pour la traduction humaine en français et un pour chacun des deux systèmes de traduction, nous avons procédé au calcul du score BLEU. Les corpus ont été créés à l'aide de Notepad++, où nous avons écrit un segment par ligne et tenu compte, lors de leur enregistrement, du fait que le fichier était enregistré avec le code UTF-8 et au format txt (Figure 14).

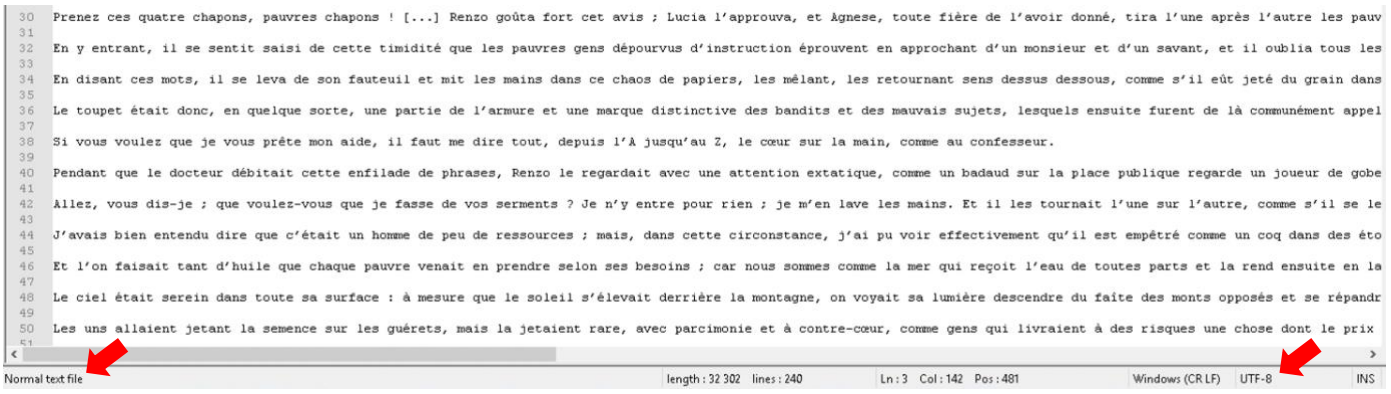

Figure 14 : Exemple de quelques phrases du corpus des traductions humaines, avec les options d'encodage en format UTF-8 et txt mises en évidence.

Une fois que nous avons obtenu les quatre corpus, nous les avons téléchargés sur Tilde (voir le site https://www.letsmt.eu/Bleu.aspx), plus précisément sur l'évaluateur interactif de score BLEU. Une fois les fichiers importés et les critères de calcul définis, à savoir « Lowercase » et « Tokenized », nous avons procédé au calcul du score BLEU. Les résultats obtenus montrent le score atteint par DeepL et Google Translate, en nous donnant à la fois une vue graphique et numérique.

Les résultats de ces deux méthodes d'évaluation, humaine et automatique, seront discutés dans la section dédiée (section 7).

6. Analyse des résultats

Après avoir traité en détail la méthodologie (Chapitre 5) utilisée dans notre étude, ce chapitre présentera et analysera les résultats des évaluations humaine et automatique des traductions automatiques des deux systèmes de traduction de Google et DeepL. Dans un premier temps, nous aborderons les résultats de l'évaluation humaine (Section 6.1). Nous verrons d'abord ceux obtenus pour le questionnaire destiné aux traducteurs (Section 6.1.1), puis ceux pour le questionnaire des non-traducteurs (Section 6.1.2). Enfin nous procéderons aux résultats de l'évaluation automatique (Section 6.2) et ensuite nous ferons une corrélation entre le score manuel, le score automatique et les scores des traducteurs humains (Section 6.3).

6.1 Évaluation humaine

6.1.1 Questionnaire complété par les traducteurs

Cette section présente les résultats des évaluations obtenus à partir du questionnaire complété par des étudiants en traduction bilingues italiens et français et visant à évaluer la qualité des traductions des systèmes de DeepL et Google Translate (Section 5).

La figure 15 montre les deux moyennes globales pour tous les participants et toutes les questions. Pour arriver à ce résultat, nous avons d'abord calculé la moyenne obtenue pour chaque participant pour l'ensemble du questionnaire. Ensuite, en utilisant les 18 moyennes obtenues pour tous les questionnaires (les moyennes des 18 participants traducteurs), nous

avons procédé au calcul de la moyenne globale pour chaque système. Compte tenu du fait que l'échelle Likert utilisée dans les questionnaires suit la logique où 1 correspond à *Tout à fait d'accord* et 5 à *Pas du tout d'accord*, cela signifie que plus le score obtenu est élevé, plus la phrase a été considérée comme problématique par l'évaluateur. D'après la figure 15, nous pouvons donc dire que les étudiants en traduction ont jugé DeepL globalement meilleur que Google Translate, même si la différence entre les deux scores n'est pas très importante (0.08 points).

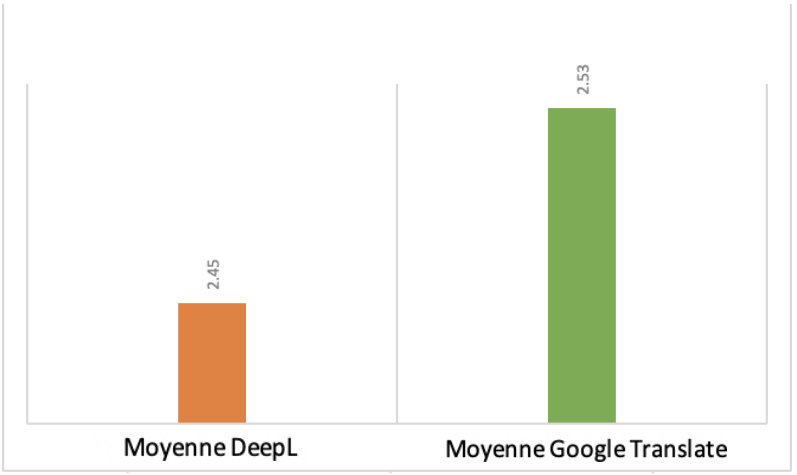

Figure 15 : Moyenne globale de chaque système de traduction pour le questionnaire des traducteurs (Moyenne des notes de 18 personnes, données sur une échelle de 1 à 5, où 1 est la meilleure note possible et 5 la plus mauvaise).

Pour mesurer la dispersion statistique, nous avons également calculé pour chaque système l'écart type. Celle-ci estime la variabilité d'un ensemble de données calculée en prenant la racine carrée de la variance. Comme le montre la figure 16, l'écart type calculé à partir de la moyenne obtenue par les participants pour le questionnaire n'est pas élevé, bien qu'il soit plus élevé dans Google Translate (0.54) que dans DeepL (0.51). Selon ces données aussi, les notes données pour les segments traduits par DeepL sont ainsi plus proches entre elles que celles de Google Translate.

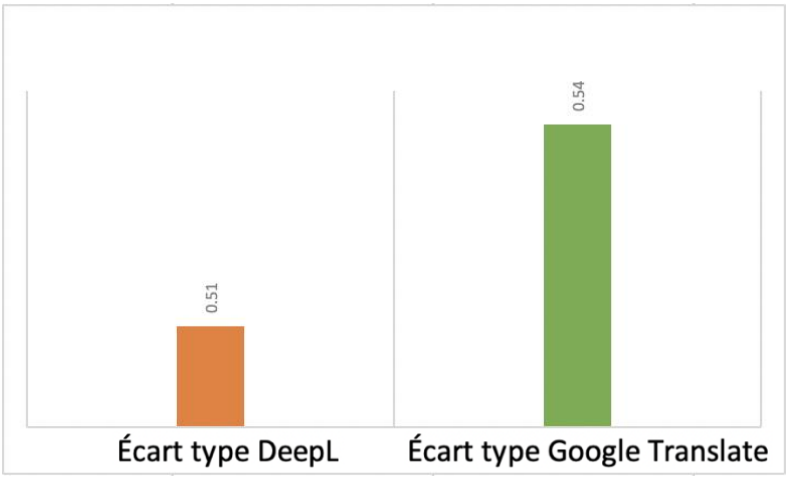

Figure 16 : Écart type calculé à partir de la moyenne obtenue par les participants pour le questionnaire.

Afin d'évaluer les résultats des réponses à chacune des cinq questions du questionnaire, nous avons calculé la moyenne (Figure 17) et l'écart type (Figure 18) pour chacune des questions pour DeepL et Google Translate.

Dans la figure 17, nous voyons que la moyenne obtenue pour les cinq questions des segments traduits par DeepL est toujours inférieure à celle obtenue à partir des segments traduits par Google Translate, ce qui nous amène à la conclusion qu'en moyenne les traductions effectuées par le système DeepL sont meilleures et présentent moins de problèmes que celles traduites par Google Translate.

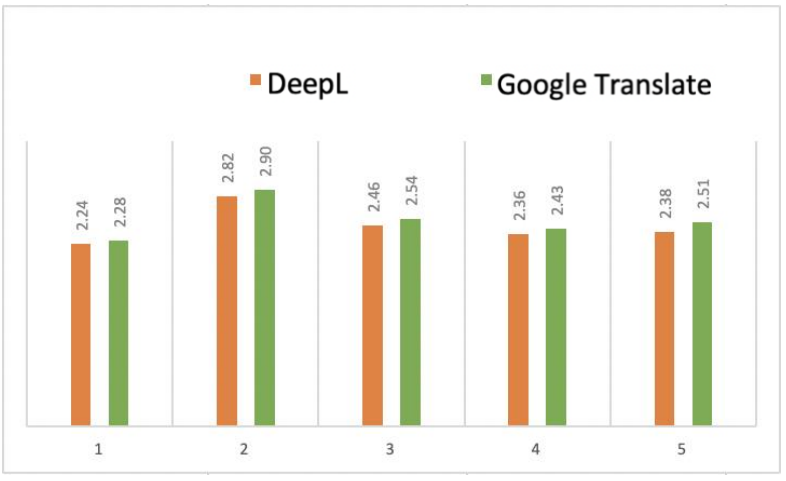

Figure 17 : Moyenne des réponses aux cinq questions de chaque segment de DeepL et Google Translate.

Nous reprenons dans la suite des exemples des trois phrases qui ont obtenu les meilleurs et les moins bons scores lors de cette évaluation.

Parmi les segments qui ont reçu une note moyenne proche de 5 (la pire note possible) avec l'évaluation humaine, on trouve les segments 67, 72 et 75. Il est intéressant de noter que le segment 67 obtient l'un des scores les plus bas à la fois pour DeepL (Figure 21), avec 4.5 points, et pour Google Translate (Figure 22), avec 4.67 points. Nous pouvons voir que dans ce cas, la comparaison est traduite sans interprétation par les systèmes, elle est presque une traduction littérale de la version italienne « *come se gli fossero state peste l'ossa* ». Le segment 72 (Figure 23) est l'un des segments ayant reçu les notes les plus basses pour DeepL, avec un score moyen de 5 points. Dans ce cas, le système s'est trompé de genre en mettant le féminin au lieu du masculin, bien que le contexte nécessaire ait été fourni. Nous pensons que cet aspect a fait baisser le score de manière significative. Le segment 75 (Figure 24) est également parmi les plus mauvais, avec un score de 4.87 pour Google Translate. Dans ce cas également, nous pensons que la difficulté est due à la traduction littérale de la comparaison, qu'il faudrait interpréter pour mieux en rendre le sens.

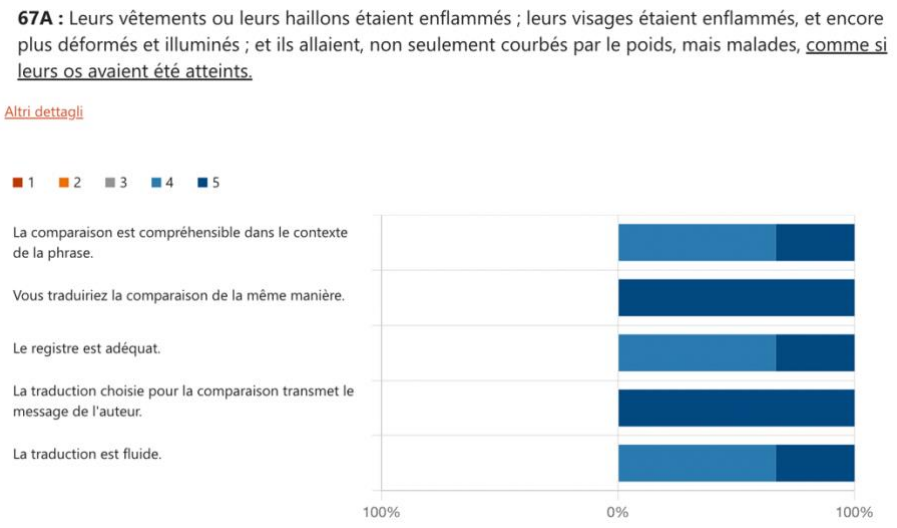

Figure 21 : Segment 67 traduit par DeepL, avec les notes données par les trois évaluateurs.

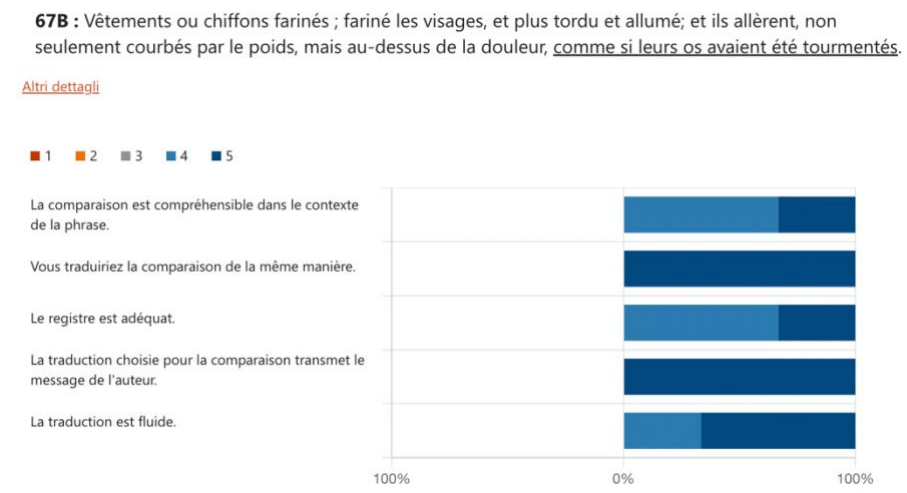

Figure 22 : Segment 67 traduit par Google Translate, avec les notes données par les trois évaluateurs.

Figure 23 : Segment 72 traduit par DeepL, avec les notes données par les trois évaluateurs.

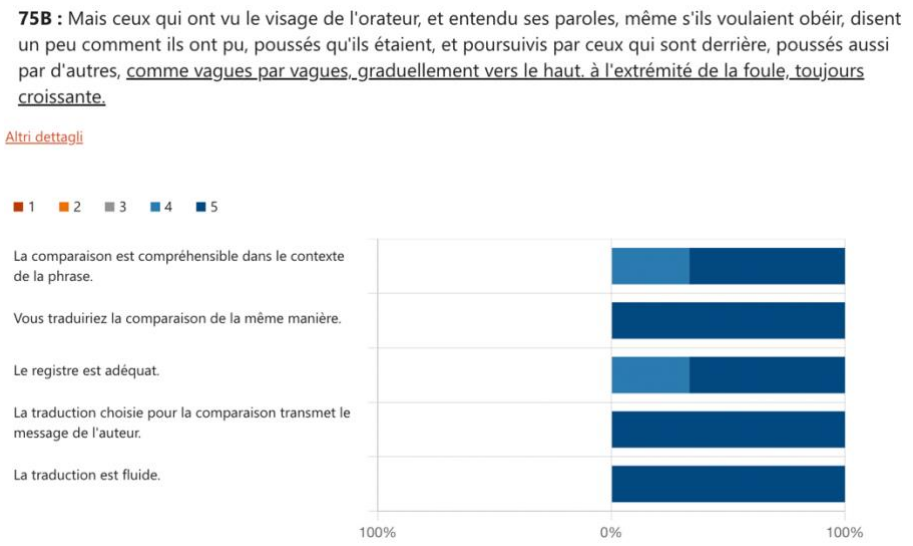

Figure 24 : Segment 75 traduit par Google Translate, avec les notes données par les trois évaluateurs.

En ce qui concerne les segments ayant obtenu les meilleurs scores, prenons comme exemple les segments 70, 76 et 77. Considérons le segment 70 (Figure 25), qui a obtenu une note moyenne de 1 avec DeepL, soit la note la plus haute possible. Dans ce cas, nous pouvons observer que la comparaison entre la farine et la neige a été très efficace. Le même constat s'applique au segment 77 (Figure 26), qui a obtenu un score de 1 avec DeepL, ce qui montre l'efficacité de la comparaison entre les pierres qui tombent et la grêle. En ce qui concerne les segments traduits par Google Translate, l'un des segments ayant obtenu le score le plus élevé de 1 est le segment numéro 76 (Figure 27), qui utilise les mouches comme comparaison pour les personnes qu'on veut faire mourir.

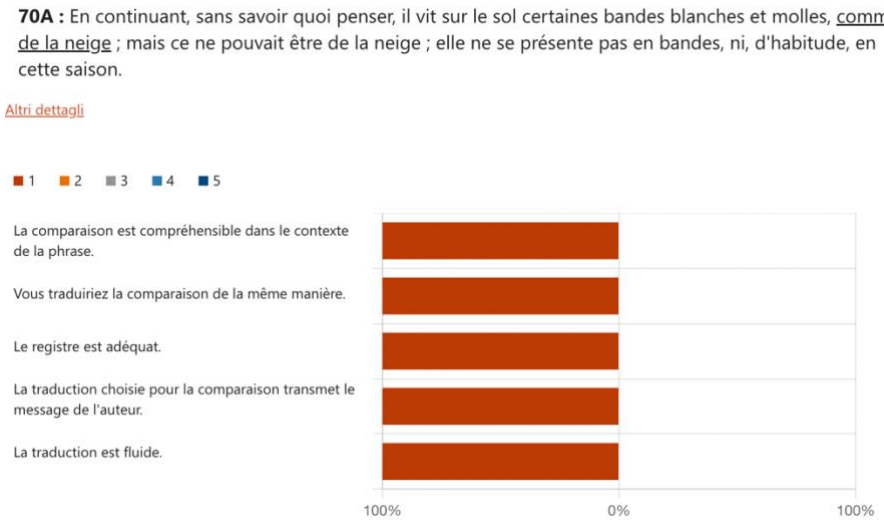

Figure 25 : Segment 70 traduit par DeepL, avec les notes données par les trois évaluateurs.

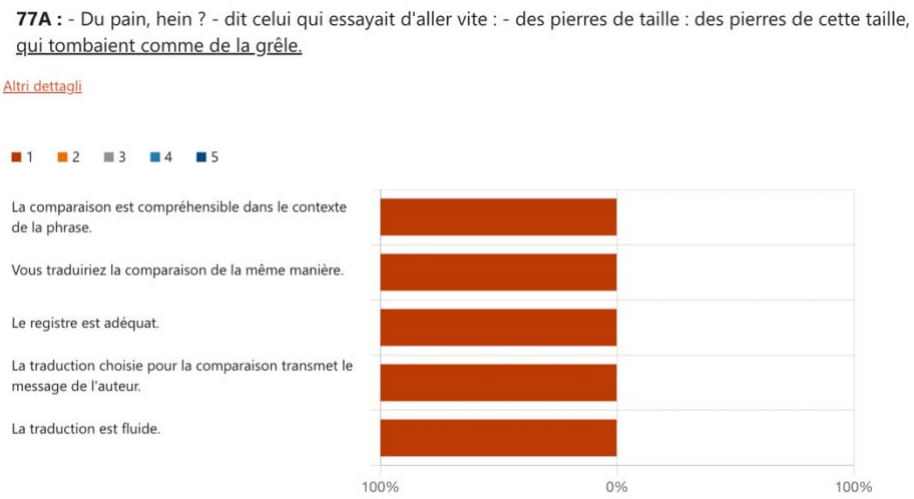

Figure 26 : Segment 77 traduit par DeepL, avec les notes données par les trois évaluateurs.

Figure 27 : Segment 76 traduit par Google Translate, avec les notes données par les trois évaluateurs.

Comme on peut le constater, dans ces exemples, la traduction littérale est possible, donc c'est aussi plus facile à traduire pour un système de traduction automatique. Nous n'observons pas la même facilité de traduction pour les segments qui nécessitent d'une adaptation. Par exemple comme dans le cas du segment 67 (Figure 21), qui obtient l'un des scores les plus bas pour DeepL et pour Google Translate (Figure 22).

La figure 18 montre le calcul de l'écart type pour chacune des cinq questions du questionnaire. Nous constatons ici que l'écart type est plus élevé pour le système de DeepL dans les questions 1 (*La comparaison est compréhensible dans le contexte de la phrase*), 2 (*Vous traduisez la comparaison de la même manière*), 3 (*Le registre est adéquat*) et 4 (*La traduction choisie pour la comparaison transmet le message de l'auteur*), alors que pour la question 5 (*La traduction est fluide*), l'écart type est plus élevé pour Google Translate. On peut également remarquer que l'écart type est le même pour les questions 1 et 3 à la fois pour DeepL et Google, alors qu'il est plus élevé pour les questions 2, 4 et 5 pour les deux systèmes de traduction (Figure 4). Nous pensons que cette différence entre les réponses est due à la subjectivité qui joue un rôle important lorsque des juges humains sont impliqués.

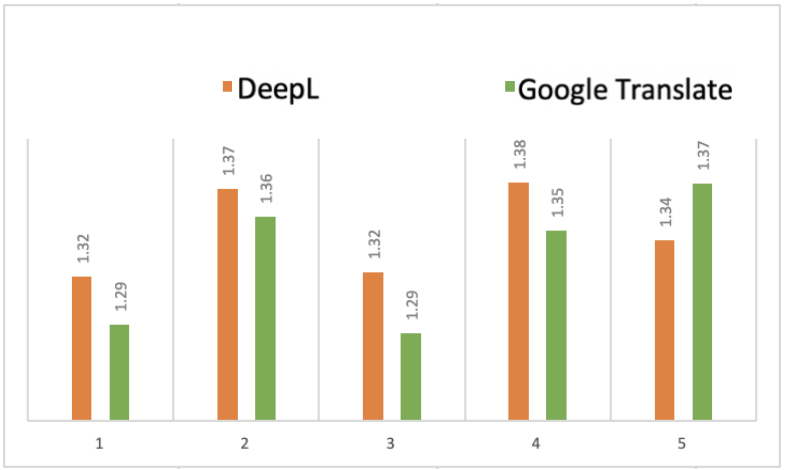

Figure 18 : L'écart type des réponses aux cinq questions de chaque segment de DeepL et Google Translate.

Pour calculer l'inter-accord entre les juges, nous avons mesuré le Kappa de Fleiss (Figure 19).

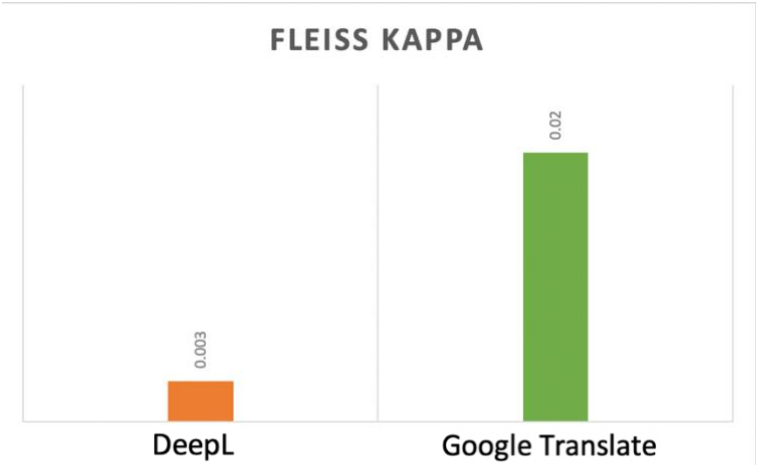

Figure 19 : Le Kappa de Fleiss de DeepL et Google Translate.

Pour pouvoir interpréter les résultats du Kappa de Fleiss, il faut se référer au tableau de Landis et Koch (1977) présenté ci-dessous (Figure 20) :

κ	Interprétation
< 0	Pauvre concordance
0.01 – 0.20	Faible concordance
0.21 – 0.40	Légère concordance
0.41 – 0.60	Concordance moyenne
0.61 – 0.80	Concordance importante
0.81 – 1.00	Concordance presque parfaite

Figure 20 : Tableau pour l'interprétation du score Kappa de Fleiss (Landis et Koch, 1977).

Le score obtenu par le Kappa de Fleiss (Figure 19) est de 0.003 pour DeepL et de 0.02 pour Google, ce qui indique que le niveau d'accord entre les participants qui ont évalué les segments est très faible, en particulier dans le cas de DeepL. Selon le tableau de référence de Landis et Koch (Figure 20) utilisé pour interpréter le Kappa de Fleiss, le niveau de concordance atteint pour DeepL est considéré comme pauvre, tandis que celui atteint par Google Translate est considéré comme faible. Étant donné que les deux résultats sont faibles, cela montre qu'il y a beaucoup de subjectivité dans l'évaluation, comme cela est souvent le cas dans les évaluations faites par des humains. Il faut également prendre en compte le fait que l'évaluation a été faite par plusieurs juges différents, chacun ayant sa propre opinion subjective.

Dans la section suivante (Section 6.1.2) nous discutons les résultats obtenus grâce à la participation d'évaluateurs de langue maternelle française, non spécialisés en traduction.

6.1.2 Questionnaire complété par les non-traducteurs

Cette section présentera les résultats des évaluations du questionnaire complété par les participants de langue maternelle française et visant à évaluer la qualité des traductions des

systèmes DeepL et Google Translate, notamment la transmission de l'émotion au moyen de métaphores.

Dans la figure 28, nous pouvons voir les deux moyennes globales calculées à partir des résultats des questionnaires réalisés par les participants de langue maternelle française. Comme dans le cas du questionnaire destiné aux étudiants en traduction, afin de calculer la moyenne globale pour chaque système, nous avons d'abord calculé la moyenne pour chaque participant par questionnaire. Ensuite, en utilisant les moyennes obtenues par les participants, nous avons calculé la moyenne pour chaque système. Dans ce cas également, 1 correspond à Tout à fait d'accord et 5 à Pas du tout d'accord, ce qui signifie que plus le score obtenu est élevé, plus la phrase a été considérée comme problématique par l'évaluateur. D'après la figure 28, il est possible d'affirmer que les participants de langue maternelle française ont jugé DeepL meilleur que Google Translate. Bien que les points de différence ne soient pas très importants (0.25 points), ils sont néanmoins supérieurs de 0.08 points à ceux obtenus dans le questionnaire réalisé par les étudiants en traduction pour la moyenne globale entre les deux systèmes.

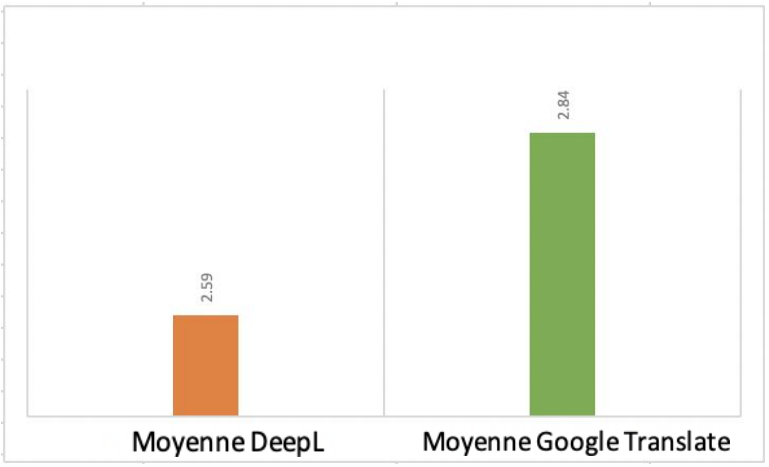

Figure 28 : Moyenne globale de chaque système de traduction pour le questionnaire destiné aux étudiants non-traducteurs de langue maternelle française (évaluation monolingue). (Moyenne des notes de 18 personnes, données sur une échelle de 1 à 5, où 1 est la meilleure note possible et 5 la plus mauvaise).

Pour ce questionnaire également, nous avons calculé l'écart type pour chaque système : comme on peut le voir sur la figure 29, l'écart type calculé à partir de la moyenne de chaque questionnaire n'est pas élevé, bien qu'il soit plus élevé pour DeepL (0.54) que pour Google Translate (0.37).

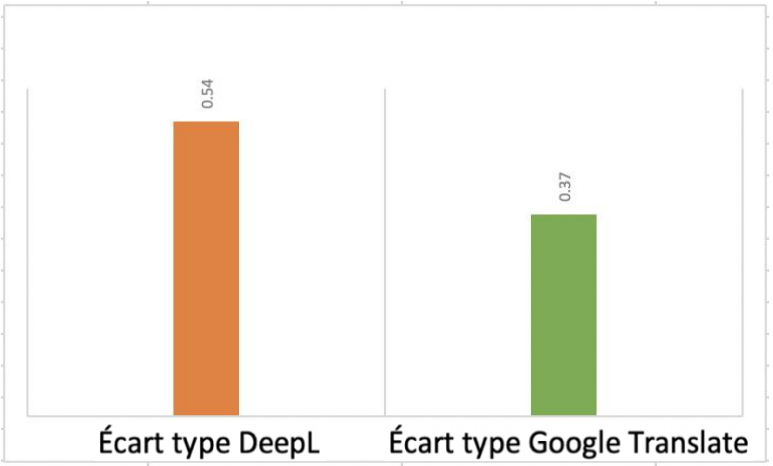

Figure 29 : Calcul de l'écart type à partir de la moyenne de chaque questionnaire réalisé par les participants de langue maternelle française.

Concernant l'évaluation des résultats des réponses à chacune des cinq questions du questionnaire, la moyenne et l'écart type ont été calculés pour chacune des questions se référant à DeepL et Google Translate (Figure 30).

Dans la figure 30, nous voyons que la moyenne obtenue pour l'ensemble des cinq questions dans les segments traduits par DeepL est toujours inférieure à celle obtenue pour les segments traduits par Google Translate : de cette façon, nous pouvons observer qu'en moyenne, les traductions effectuées par DeepL sont meilleures et présentent moins de difficultés que celles traduites par Google Translate, bien que la différence de points ne soit jamais significative.

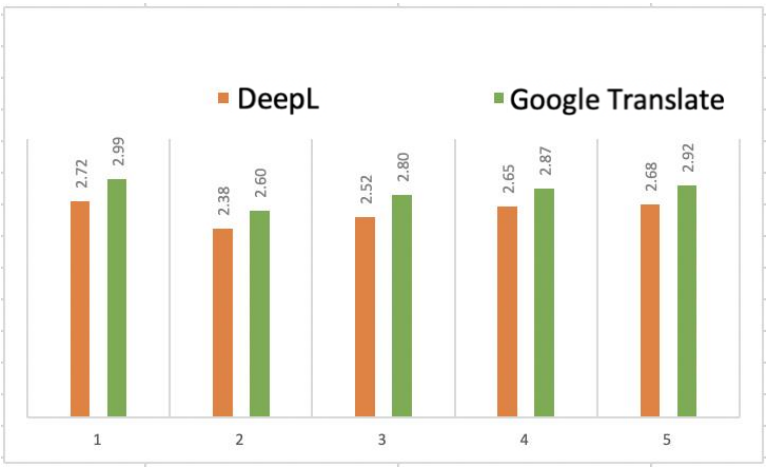

Figure 30 : Moyenne des réponses par question pour DeepL et Google Translate.

Nous reprenons dans la suite des exemples des trois phrases qui ont obtenus les meilleurs et les moins bons scores lors de cette évaluation.

Comme exemple de segments ayant reçu la plus mauvaise note, c'est-à-dire la plus proche de 5, nous présentons ci-dessous les segments 68, 72 et 78. Nous remarquons que les segments 68 (Figure 33) et 78 (Figure 34), tous les deux traduits par Google Translate, ont obtenu le même score par les évaluateurs, avec une note moyenne de 4.7. Dans le cas du segment 68, le problème est dû à une traduction partielle de la comparaison, qui laisse « pentolaccia » en italien, alors que, par exemple, DeepL le traduit par « pot ». Dans le cas du segment 78, le terme « stagnation » semble ne pas être assez efficace pour transmettre le message souhaité : nous voyons que dans la traduction humaine l'expression est traduite par « être suspendu ». Un exemple de segment qui a obtenu un score extrêmement bas avec DeepL, égal à 4.93, est le segment 72 (Figure 35) : nous constatons que le sujet masculin, indiqué dans la phrase qui précède celle de la comparaison, est repris par un sujet féminin. En outre, la comparaison elle-même n'est pas suffisamment efficace. La version humaine, par exemple, la traduit par « il fit comme une femme sur le retour, qui croirait se rajeunir en altérant son extrait de baptême ».

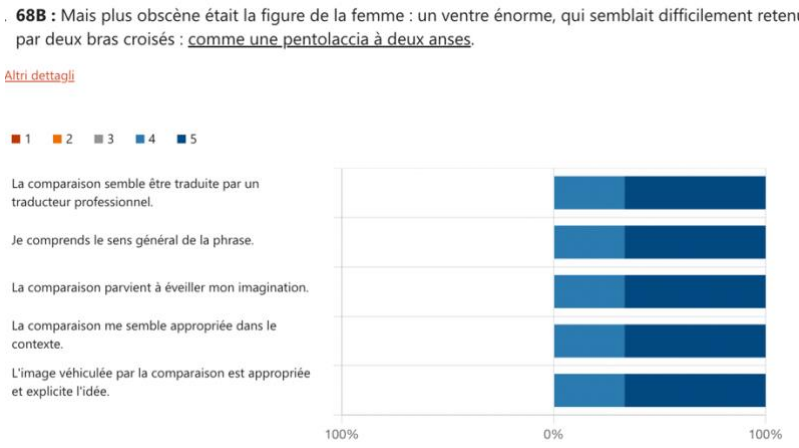

Figure 33 : Segment 68 traduit par Google Translate, avec les notes données par les trois évaluateurs.

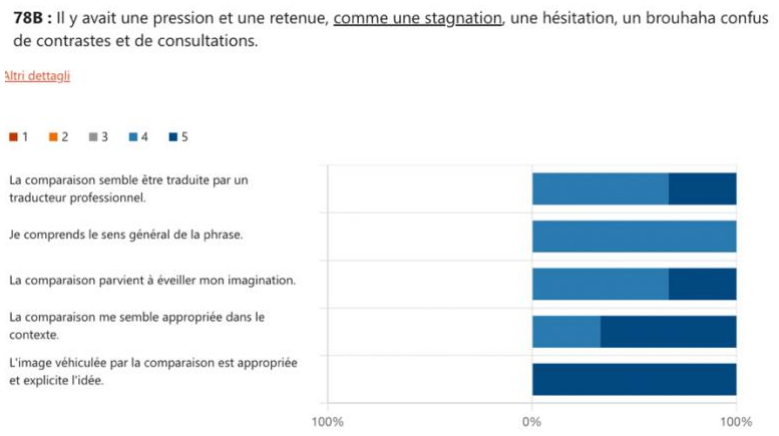

Figure 34 : Segment 78 traduit par Google Translate, avec les notes données par les trois évaluateurs.

Figure 35 : Segment 72 traduit par DeepL, avec les notes données par les trois évaluateurs.

Parmi les segments ayant obtenu les meilleurs résultats, figurent les segments 17, 61 et 70. Le segment 17 (Figure 36) a obtenu le meilleur score avec DeepL, égal à 1.2 points, ce qui nous indique que la traduction de la comparaison a été efficace. Un autre segment traduit par DeepL qui a obtenu un score élevé est le numéro 61 (Figure 37), avec une moyenne de 1.4 points : ici la comparaison du lièvre a été utilisée pour parler de la vivacité du petit prince. Un segment traduit par Google Translate qui a obtenu l'un des meilleurs scores est le segment 70 (Figure 38), qui a obtenu un score moyen de 1.2 points. Dans ce cas également, la comparaison entre la farine et la neige s'est avérée efficace.

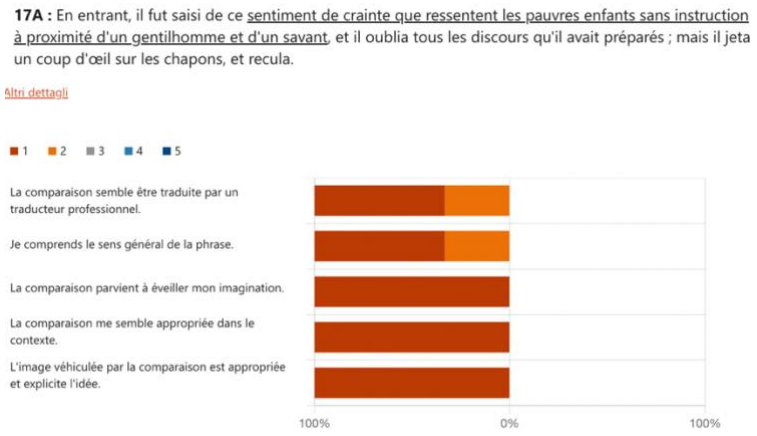

Figure 36 : Segment 17 traduit par DeepL, avec les notes données par les trois évaluateurs.

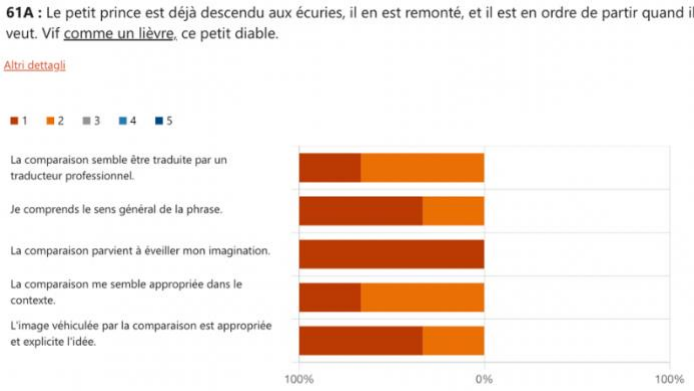

Figure 37 : Segment 61 traduit par DeepL, avec les notes données par les trois évaluateurs.

Figure 38 : Segment 70 traduit par Google Translate, avec les notes données par les trois évaluateurs.

Dans la figure 31, l'écart-type est calculé pour chacun des cinq questions : nous voyons que l'écart type est légèrement plus élevé pour le système de DeepL dans les questions 4 (*La comparaison me semble appropriée dans le contexte*) et 5 (*L'image véhiculée par la comparaison est appropriée et explicite l'idée*), alors que dans les questions 1 (*La comparaison semble être traduite par un traducteur professionnel*), 2 (*Je comprends le sens général de la phrase*) et 3 (*La comparaison parvient à éveiller mon imagination*) l'écart type est plus important pour Google Translate. Donc les valeurs sont moins dispersées autour de la moyenne pour la question 2 et 4, et cela pour les deux systèmes de traduction.

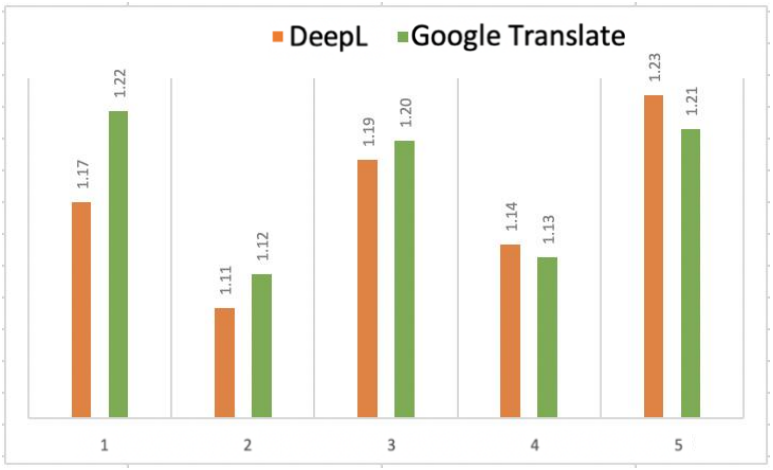

Figure 31 : Écart type par question pour DeepL et Google Translate.

Finalement, nous avons calculé le Kappa de Fleiss pour ce questionnaire également (Figure 32) afin de connaître le degré d'accord entre les juges.

Selon le tableau (Figure 20) de Landis et Koch (1977) utilisé pour évaluer le Kappa de Fleiss, le score obtenu pour le Kappa de Fleiss de DeepL, égal à 0.018, et celui obtenu par Google Translate, égal à 0.017, nous pouvons affirmer que la concordance entre les juges à l'égard de chaque système de traduction est faible, car elle se situe dans la marge de score entre 0.01 et 0.20. On observe donc que le degré d'accord entre les juges non-traducteurs est plus fort pour DeepL que celui des traducteurs. Le degré d'accord pour Google Translate est par contre plus faible pour les juges non-traducteurs par rapport aux juges spécialisés en traduction.

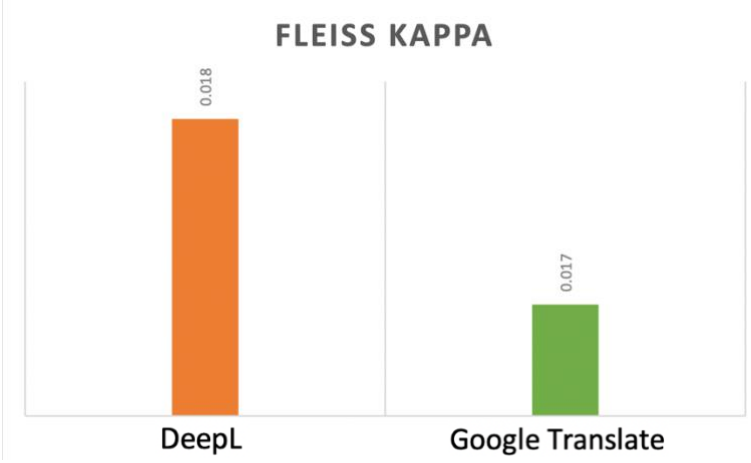

Figure 32 : Kappa de Fleiss de DeepL et Google Translate pour les non-traducteurs.

Dans la section suivante (Section 6.2), nous procéderons à l'analyse des résultats de l'évaluation automatique (Section 4.2).

6.2 Évaluation automatique

Dans cette section, nous aborderons les résultats de l'évaluation automatique des traductions effectuées par DeepL et Google Translate obtenus par la méthode d'évaluation BLEU (Papineni et al., 2002), traitée dans la section 4.2.

L'évaluation automatique des segments a été effectuée en ligne à l'aide de Tilde MT, qui est permet de calculer le score BLEU (Section 4.2). En ce qui concerne les paramètres utilisés lors de l'évaluation, nous avons laissé les paramètres par défaut (Figure 39), car aucune explication précise n'a été fournie sur le site web au sujet de ces options.

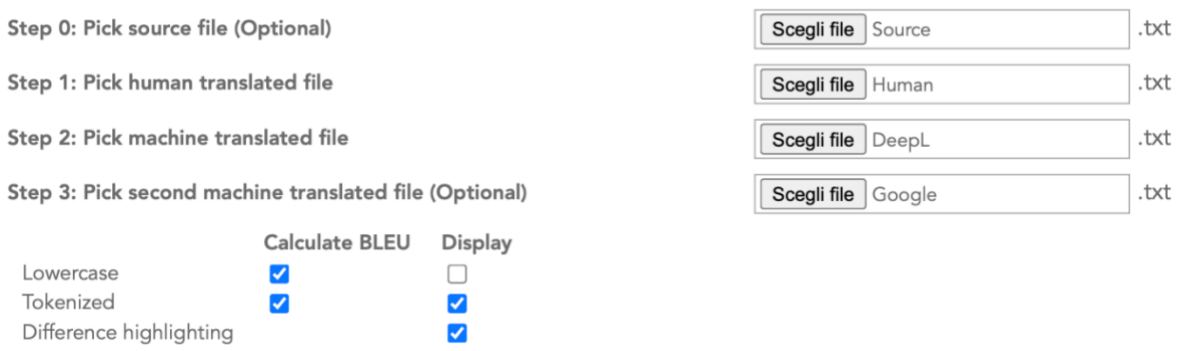

Figure 39 : Paramétrage de Tilde MT pour le calcul du score BLEU.

Grâce à Tilde MT, il est possible de comparer la source, la traduction humaine, la traduction de DeepL et de Google Translate, en voyant en détail le score obtenu par chaque segment et le score global obtenu par les deux systèmes de traduction comparés. Dans le cadre de notre recherche, nous nous concentrerons davantage sur le score global obtenu par les systèmes de traduction, en donnant quelques exemples de segments qui se distinguent des autres en termes de score obtenu.

En ce qui concerne le score BLEU global (Figure 40) obtenu par les deux systèmes au moyen de Tilde MT, il n'y a pas de grande différence puisque DeepL a obtenu 25.10, tandis que Google Translate a obtenu 25.02 points. On observe que, même si la différence est moins importante entre les deux scores automatiques par rapport à ceux humains, on observe que la tendance est d'aller dans a même direction. Il convient de noter que la note est donnée sur un total maximum possible de 100 points.

	DeepL	Google Translate
Score BLEU	25.10	25.02

Figure 40 : Score BLEU global calculé par Tilde MT.

En ce qui concerne les segments qui se sont distingués, nous avons décidé de sélectionner les trois segments ayant obtenu les meilleurs scores pour DeepL (Figure 41) et Google Translate (Figure 42).

Segment	DeepL
76	70.01
112	56.91
84	52.04

Figure 41 : Les trois scores BLEU les plus élevés de DeepL.

Segment	Google Translate
94	54.00
17	49.26
92	47.17

Figure 42 : Les trois scores BLEU les plus élevés de Google Translate.

Comme nous pouvons le voir dans la figure 43, DeepL est le système qui a obtenu le meilleur score, atteignant 70.01 points avec le segment 76, tandis que Google Translate a obtenu 43.43 points pour le même segment (Figure 43). Dans ce cas, la comparaison entre les personnes empoisonnées et les mouches a été rendue de la même manière dans les deux systèmes de traduction. La traduction de DeepL, en particulier, est très proche de la traduction humaine : le seul élément qui différencie la traduction humaine de la traduction automatique est l'article « des ». Le même segment a également été jugé meilleur dans la version traduite par DeepL par les participants de langue maternelle française (avec un score moyen de 1.87 points), tandis que les étudiants en traduction ont considéré que le segment traduit par Google (avec un score moyen de 1 point) était meilleur.

Sentence 76	BLEU	Length ratio	Text
Source	-	-	Il pane verrà a buon mercato, ma ci metteranno il veleno, per far morir la povera gente, come mosche.
Human	100.00	1.00	Le pain sera à bon marché ; mais ils y mettront du poison , pour que les pauvres gens meurent comme mouches .
Machine	70.01	1.00	Le pain sera bon marché , mais ils y mettront du poison , pour que les pauvres gens meurent comme des mouches .
Machine	43.43	1.00	Le pain sera bon marché , mais ils y mettront du poison , pour faire mourir les pauvres , comme des mouches .

Figure 43 : Segment 76 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

La figure 44 montre que le segment 94 a obtenu le plus grand nombre de points dans la traduction effectuée par Google Translate, soit un total de 54.00 points. En effet, le mot tenaille a été traduit par « pince », tandis que le mot « cavadenti », qui désigne la personne qui extrayait les dents malades, a été traduit par « cure-dents », qui désigne le petit bâton de bois utilisé pour nettoyer les dents. La même préférence pour le segment traduit par Google est également confirmée par le jugement des évaluateurs spécialisés en traduction (où le segment obtient un score de 2.8 points) et par les non-traducteurs (obtenant un score de 2.4 points).

Sentence 94	BLEU	Length ratio	Text
Source	-	-	A quel nuovo scongiuro, don Abbondio, col volto, e con lo sguardo di chi ha in bocca le tanaglie del cavadenti, proferì [...].
Human	100.00	1.00	À cette nouvelle adjuration , don Abbondio , avec le visage et le regard de celui qui a dans sa bouche les tenailles de l ' arracheur de dents , prononça [...] .
Machine	36.04	0.94	A cette nouvelle supplique , don Abbondio , avec la figure et le regard de quelqu ' un qui a dans la bouche les tanaglie del cavadenti , prononça [...] .
Machine	54.00	0.94	A cette nouvelle conjuration , Don Abbondio , avec le visage et le regard de celui qui a la pince du cure - dents dans la bouche , prononça [...] .

Figure 44 : Segment 94 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

En observant les évaluations humaines, nous constatons une congruence entre les évaluations humaines et les évaluations automatiques (sections 6.1.1 et 6.1.2). En fait, les segments 76 et 17 sont parmi les meilleurs dans les évaluations automatiques et humaines. Alors que nous avons déjà mentionné le segment 76, dans le cas du segment 17, le meilleur score BLEU se trouve dans la traduction de Google, tandis que selon les étudiants en traduction et les évaluateurs non spécialisés en traduction, le segment a été mieux traduit par DeepL (avec une moyenne de 1.7 et 1.2 points respectivement).

Les trois segments ayant le score BLEU le plus bas pour DeepL (Figure 45) et Google Translate (Figure 46) sont présentés ci-dessous.

Segment	DeepL
93	2.95
74	4.30
67	4.51

Figura 45 : Les trois scores BLEU les plus bas de DeepL.

Segment	Google Translate
26	2.62
67	3.06
74	3.66

Figura 46 : Les trois scores BLEU les plus bas de Google Translate.

En ce qui concerne DeepL, le segment 93 (Figure 47) a obtenu le plus petit nombre de points (2.95), tandis que Google Translate est le système qui a obtenu le plus petit nombre de points avec le segment 26 (Figure 48), soit 2.62 points. Le score bas du segment 93 est principalement dû au fait que DeepL n'a pas traduit l'ensemble du segment, omettant la partie initiale. Quant à la comparaison elle-même, elle a été évaluée avec un bon score (moyenne de 2.2 points) par les participants de langue maternelle française, tandis qu'elle a été évaluée avec un score plus faible mais suffisamment bon par les participants spécialisés dans la traduction (avec un score moyen de 2.47 points). Le segment 26 a reçu un meilleur score, même si de seulement 0.13 point, de la part des participants non spécialisés en traduction lorsqu'il a été traduit par Google (avec 2.87 points), tandis que les étudiants en traduction ont préféré la traduction de DeepL (avec 2.27 points), confirmant le résultat BLEU.

Sentence 93	BLEU	Length ratio	Text
Source	-	-	Eh! le schioppettate non si dànno via come confetti: e guai se questi cani dovessero mordere tutte le volte che abbaiano!
Human	100.00	1.00	Bah ! les coups de fusil ne se donnent pas comme des prunes : et où en serions - nous si tous ces chiens - là mordaient toutes les fois qu ' ils aboient ?
Machine	2.95	0.49	Et malheur à ces chiens s ' ils devaient mordre chaque fois qu ' ils aboient !
Machine	5.33	0.83	Eh ! les plombs ne sont pas donnés comme des confettis : et malheur à ces chiens s ' ils mordent à chaque fois qu ' ils aboient !

Figure 47 : Segment 93 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Sentence 26	BLEU	Length ratio	Text
Source	-	-	Alcuni andavan gettando le lor semente, rade, con risparmio, e a malincuore, come chi arrischia cosa che troppo gli preme; altri spingevan la vanga come a stento, e rovesciavano svogliatamente la zolla.
Human	100.00	1.00	Les uns allaient jetant la semence sur les guérets , mais la jetaient rare , avec parcimonie et à contre - cœur , comme gens qui livraient à des risques une chose dont le prix était grand pour eux ; d ' autres semblaient faire effort pour enfoncer la pioche en terre , et retournaient la motte d ' un air d ' abattement .
Machine	5.09	0.62	Certains ont semé avec parcimonie et à contrecœur , comme celui qui risque quelque chose qui lui est trop cher ; d ' autres ont poussé leur bêche avec difficulté , et ont retourné le gazon sans se presser .
Machine	2.62	0.62	Certains semaient leurs graines , clairsemés , parcimonieusement et à contrecœur , comme ceux qui risquent quelque chose qui tient trop à cœur ; d ' autres poussaient la bêche comme avec peine , et renversaient la motte nonchalamment .

Figure 48 : Segment 26 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Il est intéressant de constater que les deux segments avec le score le plus bas coïncident pour les deux systèmes de traduction : en effet, le segment 74 (Figure 49) est le deuxième segment avec le nombre de points le plus bas pour DeepL, soit 4.30 points, et le troisième segment avec

le score le plus bas pour Google Translate, avec 3.66 points. En ce qui concerne l'évaluation humaine, les participants, à la fois spécialistes de la traduction et non-traducteurs, préfèrent le segment traduit par DeepL, lui attribuant respectivement une moyenne de 2.4 et 3.73 points. La traduction de Google obtient des points que nous pouvons définir comme faibles, égaux à 4.07 pour les étudiants en traduction et 4.27 pour ceux qui ne sont pas spécialisés dans la traduction. Dans ce cas, la comparaison qui, en italien, est rendue par la chute d'un « salterello » n'est pas traduite de manière compréhensible, car « pistolet allumé » et « saut allumé » ne rendent pas l'idée de quelque chose qui bouge de manière sinueuse.

Sentence 74	BLEU	Length ratio	Text
Source	-	-	Il primo comparire d'uno di que' malcapitati ragazzi dov'era un crocchio di gente, fu come il cadere d'un salterello acceso in una polveriera.
Human	100.00	1.00	Le premier de ces malencontreux garçons qui parut devant un groupe fit l ' effet d ' un serpenteau qui tombe dans une poudrière .
Machine	4.30	1.20	La première apparition d ' un de ces malheureux garçons dans une foule de gens était comme la chute d ' un pistolet allumé dans un baril de poudre .
Machine	3.66	1.40	La première apparition d ' un de ces malheureux garçons , où il y avait un groupe de personnes , était comme la chute d ' un saut allumé dans un baril de poudre .

Figure 49 : Segment 74 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Le segment 67 (Figure 50) fait également partie des trois segments qui ont obtenu les scores les plus bas pour les deux systèmes de traduction. En fait, il s'agit de la troisième note la plus mauvaise pour DeepL, avec 4.51 points, et de la deuxième note la plus mauvaise pour Google Translate, avec 3.06 points. La même tendance est confirmée dans l'évaluation humaine : le segment est considéré comme meilleur lorsqu'il est traduit par DeepL, bien qu'il n'obtienne un score significativement élevé ni selon les étudiants en traduction (4.6 points) ni selon les participants de langue maternelle française (3.3 points). Nous remarquons également que pour la version traduite par Google, les scores sont mauvais tant pour les participants spécialisés dans la traduction (4.7 points) que pour les autres (4.47 points).

Sentence 67	BLEU	Length ratio	Text
Source	-	-	I vestiti o gli stracci infarinati; infarinati i visi, e di più stravolti e accesi; e andavano, non solo curvi, per il peso, ma sopra doglia, come se gli fossero state peste l'ossa.
Human	100.00	1.00	Ils étaient blancs de farine sur leurs vêtements ou les haillons qui leur en tenaient lieu , blancs de farine sur leurs visages , dont les traits bouleversés marquaient une vive agitation : et ils marchaient , non - seulement courbés sous leur fardeau , mais d ' un air de souffrance , comme s ' ils avaient été foulés dans tous leurs membres .
Machine	4.51	0.65	Leurs vêtements ou leurs haillons étaient enflammés ; leurs visages étaient enflammés , et encore plus déformés et illuminés ; et ils allaient , non seulement courbés par le poids , mais malades , comme si leurs os avaient été atteints .
Machine	3.06	0.65	Vêtements ou chiffons farinés ; fariné les visages , et plus tordu et allumé ; et ils allèrent , non seulement courbés par le poids , mais au - dessus de la douleur , comme si leurs os avaient été tourmentés .

Figure 50 : Segment 67 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

En outre, trois segments se distinguent par les points égaux ou presque égaux obtenus par les deux systèmes de traduction : les segments 23, 78 et 116. Le segment 23 (Figure 51) a obtenu un score de 4.70 points avec DeepL et de 4.77 points avec Google Translate, soit une différence de seulement 0.07 points. La traduction s'est donc révélée être presque également problématique pour les deux systèmes. La traduction a également été jugée médiocre par les évaluateurs humains : la traduction de DeepL a été estimée meilleure dans les deux questionnaires (avec une moyenne de 3 points pour les étudiants en traduction et de 3.3 points pour les non-traducteurs), tandis que la traduction de Google a obtenu une note de 3.7 par les étudiants en traduction et de 4.13 par les non-traducteurs. En ce qui concerne la comparaison, si DeepL a été capable de traduire poussin mais pas étoupe, Google n'a été capable de traduire aucun des deux termes.

Le segment 78 (Figure 52) a obtenu le même nombre de points pour les deux systèmes, à savoir 8.10, se révélant ainsi tout aussi complexe à traduire. Le cas est différent pour l'évaluation humaine, car les étudiants en traduction et les participants qui ne sont pas spécialisés dans la traduction jugent la traduction de DeepL meilleure, lui donnant une moyenne de 2.07 et 3.47 points respectivement, tandis que le segment de Google obtient une moyenne de 2.27 et 4.3 points respectivement. Nous voyons que la comparaison est traduite par le terme « stagnation » dans les deux systèmes, ce qui ne rend pas l'idée voulue par l'auteur. En effet, nous constatons que la traduction humaine explique le terme italien par « être suspendu ».

Pour le segment 116 (Figure 53), le score obtenu par DeepL est de 28.85 points, tandis que Google Translate a obtenu 28.80 points, soit une différence de 0.05 points. La qualité du

segment 116 est donc également acceptable pour les deux systèmes, si l'on tient compte du fait que tous deux dépassaient le seuil de 20 points, considéré comme le seuil minimum par Toral et Way (2015). Selon les deux groupes de participants humains, la traduction de DeepL est également la meilleure : en effet, les étudiants en traduction lui donnent une moyenne de 2.8 points, tandis que les participants de langue maternelle française lui donnent une moyenne de 1.7 points. Google, en revanche, obtient un score de 3.87 points pour le premier et de 3 points pour le second. Bien que le verbe utilisé dans la comparaison par les deux systèmes (« s'épanouir ») soit différent de celui de la traduction humaine (« éclore »), il peut être considéré comme un synonyme, car il rend tout aussi bien l'idée d'une fleur qui s'ouvre.

Sentence 23	BLEU	Length ratio	Text
Source	-	-	Già l'avevo sentito dire ch'era un uomo da poco; ma in quest'occasione, ho dovuto proprio vedere che è più impicciato che un pulcin nella stoppa.
Human	100.00	1.00	J ' avais bien entendu dire que c ' était un homme de peu de ressources ; mais , dans cette circonstance , j ' ai pu voir effectivement qu ' il est empêtré comme un coq dans des étoupes .
Machine	4.70	0.95	J ' avais déjà entendu dire qu ' il était un petit homme ; mais à cette occasion , j ' ai vraiment dû constater qu ' il est plus emmêlé qu ' un poussin dans la remorque .
Machine	4.77	0.98	Je l ' avais déjà entendu dire qu ' il était un petit homme ; mais à cette occasion , j ' ai vraiment dû voir qu ' il est plus empêtré qu ' un pulcin dans la remorque .

Figure 51 : Segment 23 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Sentence 78	BLEU	Length ratio	Text
Source	-	-	C'era un incalzare e un rattenere, come un ristagno, una titubazione, un ronzìo confuso di contrasti e di consulte.
Human	100.00	1.00	On poussait et on retenait ; la marche était suspendue , il y avait indécision , et un bruit confus régnait de délibérations et de débats .
Machine	8.10	0.93	Il y avait une pression et une retenue , comme une stagnation , une hésitation , un bourdonnement confus de contrastes et de consultations .
Machine	8.10	0.93	Il y avait une pression et une retenue , comme une stagnation , une hésitation , un brouhaha confus de contrastes et de consultations .

Figure 52 : Segment 78 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Sentence 116	BLEU	Length ratio	Text
Source	-	-	Vi son de' momenti in cui l'animo, particolarmente de' giovani, è disposto in maniera che ogni poco d'istanza basta a ottenerne ogni cosa che abbia un'apparenza di bene e di sacrifizio: come un fiore appena sbocciato, s'abbandona mollemente sul suo fragile stelo, pronto a concedere le sue fragranze alla prim'aria che gli aliti punto d'intorno.
Human	100.00	1.00	Il est des moments où l ' âme , surtout dans le jeune âge , est disposée de telle sorte que la moindre instance suffit pour en obtenir tout ce qui peut avoir une apparence de bien et de sacrifice ; comme la fleur qui vient d ' éclore s ' abandonne mollement sur sa tige fragile , prête à livrer ses parfums au premier souffle de l ' air qui se fasse sentir à l ' entour .
Machine	28.85	0.94	Il y a des moments où l ' âme , surtout celle des jeunes , est disposée de telle sorte que chaque petit effort suffit pour obtenir tout ce qui a une apparence de bonté et de sacrifice : comme une fleur qui vient de s ' épanouir , elle se laisse doucement tomber sur sa tige fragile , prête à donner son parfum au premier air qui souffle autour d ' elle .
Machine	28.80	0.87	Il y a des moments où l ' âme , particulièrement des jeunes , est disposée de telle manière que chaque petit effort suffit pour obtenir tout ce qui a une apparence de bien et de sacrifice : comme une fleur qui vient de s ' épanouir , oui . sur sa tige fragile , prêt à accorder ses parfums au premier air qui souffle autour de lui .

Figure 53 : Segment 116 avec le score BLEU, la source, la traduction humaine et la traduction des deux systèmes de traduction (DeepL est indiqué en BLEU, Google Translate en vert).

Dans la section suivante (Section 6.3), nous mettrons en corrélation les résultats du score manuel (celui donné par les étudiants en traduction et les non-traducteurs) et du score automatique (score BLEU).

6.3 Corrélation entre score manuel et automatique

Dans cette section, nous ferons une corrélation entre le score obtenu à partir de l'évaluation faite par les participants humains, notamment les étudiants en traduction et les participants de langue maternelle française non-spécialisés en traduction, et le score obtenu à partir de l'évaluation automatique, soit le score BLEU. Pour établir la corrélation, nous avons décidé d'utiliser le coefficient relationnel de Pearson, qui est le plus fréquemment utilisé pour établir des corrélations entre deux méthodes d'évaluation (Koehn, 2010, p. 229).

BLEU & étudiants DeepL	BLEU & Non-étudiants DeepL
-0.09	-0.12

Figure 54 : Le coefficient de Pearson entre le score BLEU et le score attribué par les participants humains à chacun des segments traduits par DeepL.

BLEU & étudiants Google	BLEU & Non-étudiants Google
-0.17	-0.25

Figure 55 : Le coefficient de Pearson entre le score BLEU et le score attribué par les participants humains à chacun des segments traduits par Google Translate.

Comme nous pouvons le constater dans la figure 54, il existe une corrélation de -0,09 entre le score BLEU et le score des étudiants en traduction et de -0,12 entre le score BLEU et le score des non-traducteurs. Cela nous conduit à conclure que l'association entre le score humain et le score automatique est faible pour les segments traduits par DeepL. Dans la figure 55, on observe une corrélation de -0,17 entre le score BLEU et le score des étudiants en traduction, et de -0,25 entre le score BLEU et le score des non-traducteurs. Il est possible de voir que pour les segments traduits par Google Translate, la corrélation est plus forte, mais elle est encore faible pour dire qu'elle confirme le score de l'évaluation humaine. Nous pouvons également observer la corrélation de manière visuelle dans les figures 56 et 57 ci-dessous.

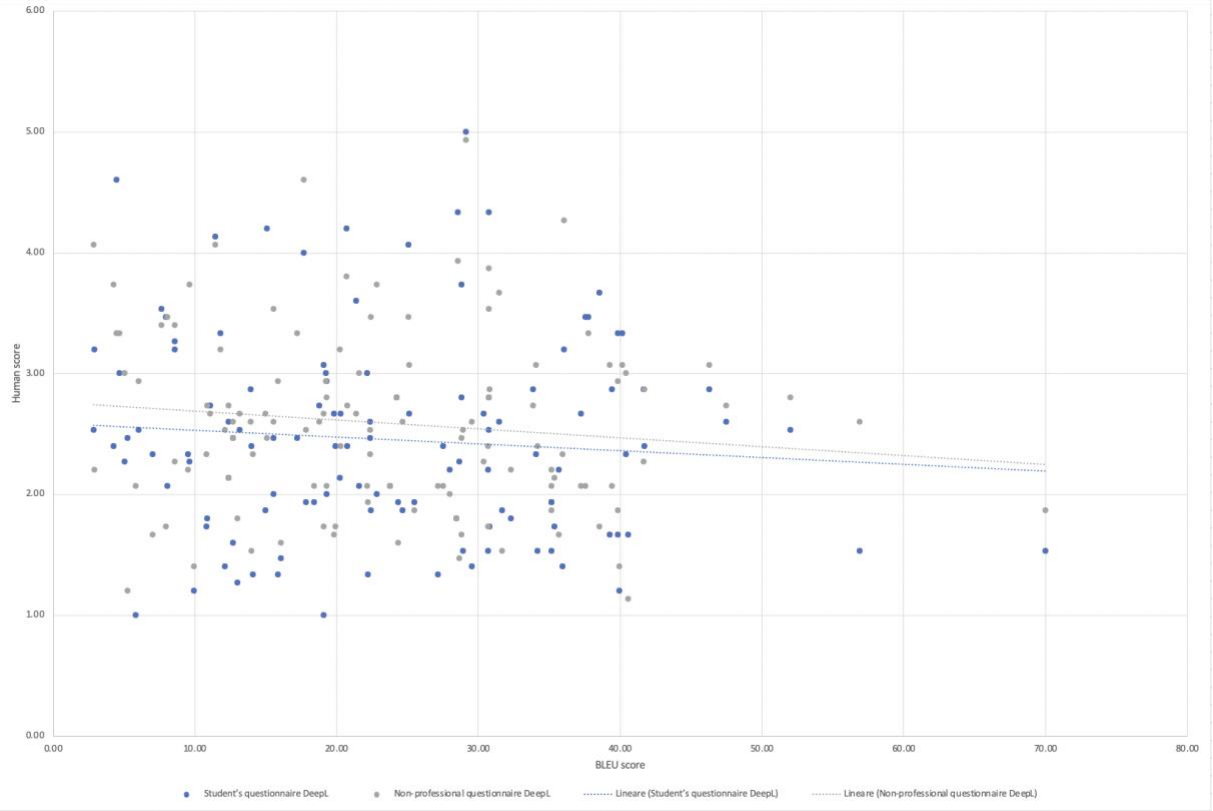

Figure 56 : Association entre le score BLEU et le score donné par les participants humains aux segments traduits par DeepL.

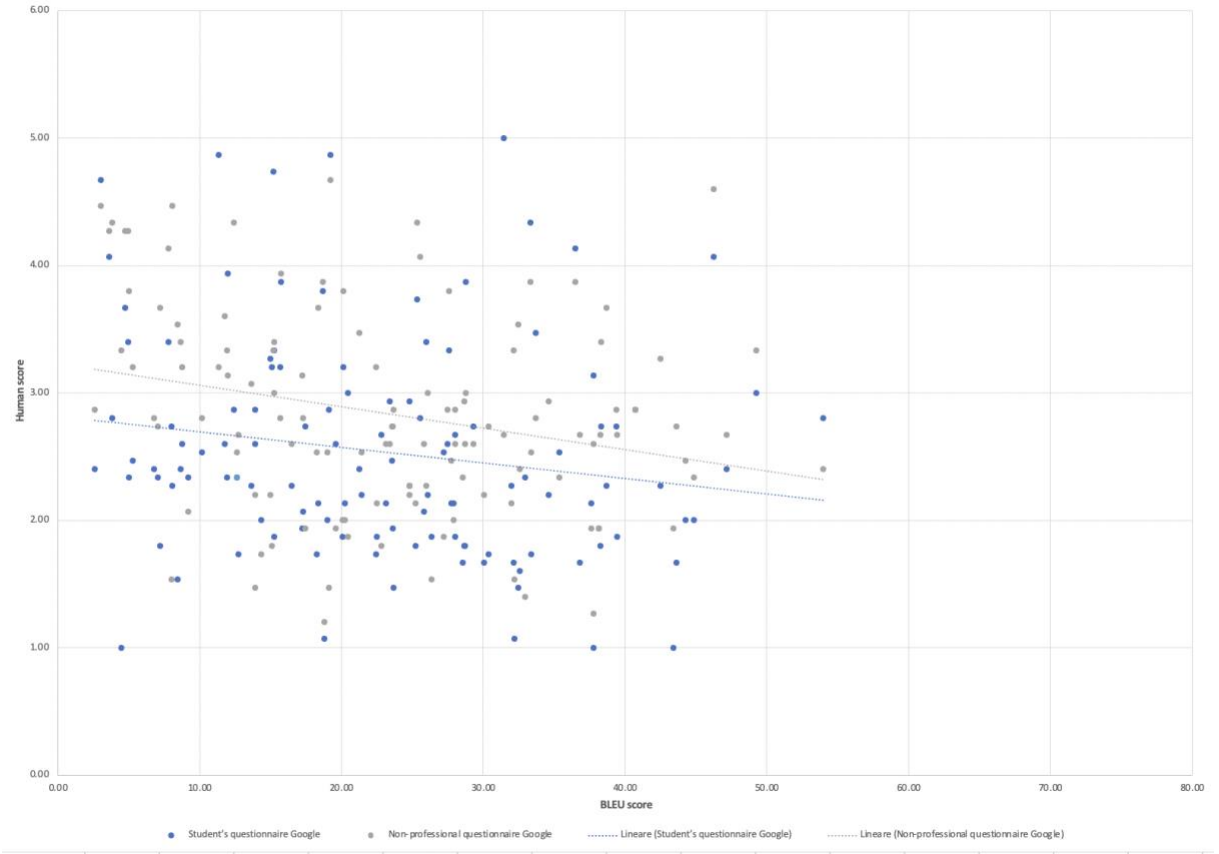

Figure 57 : Association entre le score BLEU et le score donné par les participants humains aux segments traduits par Google Translate.

Ces résultats montrent que l'évaluation automatique de BLEU n'est pas toujours fiable, surtout lorsqu'il s'agit de traductions portant sur des figures de style, dans notre cas des comparaisons. C'est pourquoi les évaluations fournies par les participants humains ont une valeur plus importante.

En ce qui concerne le score obtenu par Google Translate, il n'est pas très différent de celui de DeepL : les étudiants en traduction lui ont attribué une moyenne de 2.53 points, tandis que les participants non spécialisés en traduction lui ont donné 2.84 points. Le score de l'évaluation automatique n'est pas non plus très différent de celui de DeepL, car Google obtient un score BLEU de 25.02 points.

De suite, nous procédons à la comparaison entre les scores donnés par les participants humains, traducteurs et non-traducteurs. Nous pouvons constater dans la figure 58 que le score obtenu pour le coefficient relationnel de Pearson est assez élévé, ce que veut dire que les juges étaient

assez d'accord entre eux dans l'évaluation des traductions. Le score est majeur dans le cas de DeepL, mais la différence est très faible (de 0.08 points).

Non-étudiants & étudiants DeepL	Non-étudiants & étudiants Google
0.51	0.43

Figure 58 : Le coefficient de Pearson entre le score attribué par les juges traducteurs et non-traducteurs à chacun des segments traduits par DeepL et Google Translate.

Comme on l'observe dans la figure 59, la majorité des scores donnés pour DeepL par les deux groupes de juges se concentre autour des évaluations entre 2 et 3, c'est-à-dire une note moyenne.

En ce qui concerne les scores donnés par les deux groupes des traducteurs à Google Translate (Figure 60), les scores sont plus dispersés mais dans ce cas aussi, la plupart est concentrée autour des notes moyennes.

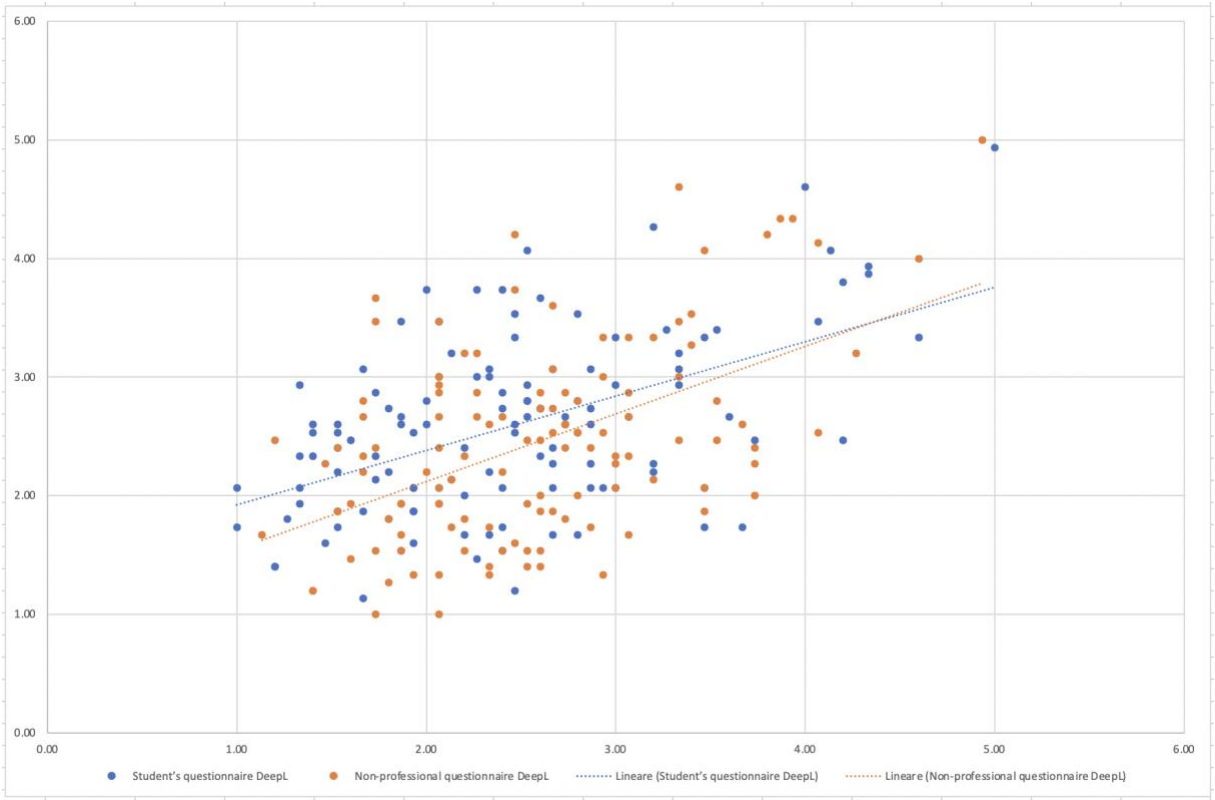

Figure 59 : Association entre le score donné par les participants humains traducteurs et non-traducteurs aux segments traduits par DeepL.

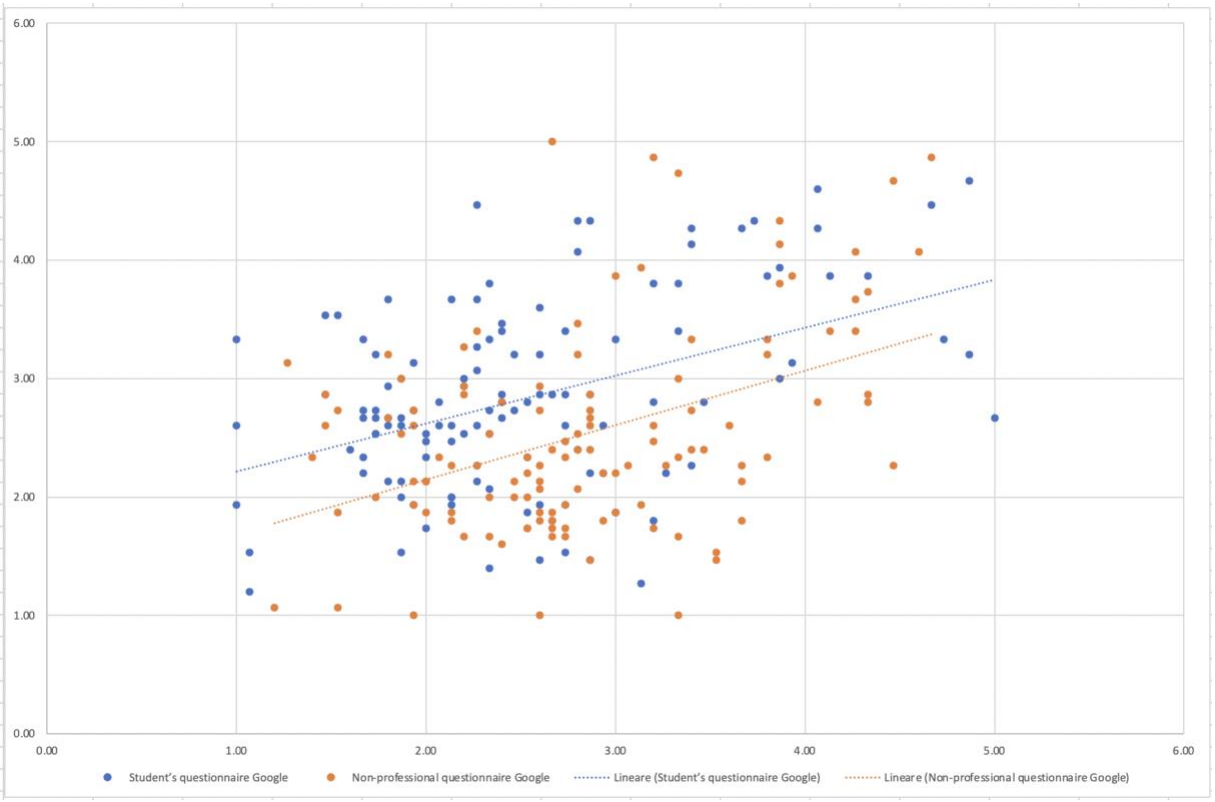

Figure 60 : Association entre le score donné par les participants humains traducteurs et non-traducteurs aux segments traduits par Google Translate.

Il est donc possible de voir qu'il y a une corrélation plus forte entre les juges humains qu'entre l'évaluation automatique et celle humaine.

Dans la section suivante (Section 6.4), nous aborderons les difficultés rencontrées dans la traduction des figures de style.

6.4 Traduction des comparaisons

Dans cette section, nous allons nous concentrer sur les principales difficultés rencontrées par les systèmes de traduction dans cette tâche.

Sur la base des données recueillies et d'une évaluation globale, il est possible d'affirmer qu'en général, il n'y a pas de scores très élevés ou très bas pour les questions des questionnaires, mais plutôt des scores moyens. En moyenne, le système qui obtient le meilleur score selon l'évaluation humaine est DeepL : il totalise 2.45 points selon la moyenne calculée à partir des résultats du questionnaire rempli par des étudiants en traduction et 2.59 points selon la moyenne

des résultats du questionnaire réalisé par des participants de langue maternelle française non spécialisés en traduction. Il convient de noter que plus le score est élevé, plus les problèmes rencontrés par les évaluateurs sont nombreux. Selon les résultats de l'évaluation automatique BLEU, DeepL est également meilleur que Google, car il obtient un score plus élevé de 25.10 points. Dans ce cas, plus le score est élevé, meilleur est le système.

En ce qui concerne les questionnaires réalisés par les étudiants en traduction, la question 1 (*La comparaison est compréhensible dans le contexte de la phrase.*), obtient le meilleur score avec les deux systèmes et atteint 2.24 points pour DeepL et 2.28 points pour Google Translate. Cela nous montre que, dans l'ensemble, les segments sont généralement compréhensibles lorsqu'ils sont traduits par les deux systèmes de traduction, présentant une difficulté à peine plus grande pour les segments traduits par Google. La question ayant obtenu le plus mauvais score, en revanche, est la question numéro 2 (*Vous traduisez la comparaison de la même manière*) pour les deux systèmes, obtenant 2.82 points pour DeepL et 2.90 points pour Google, tout en recevant un score moyen suffisant. Cela nous indique que même si les traducteurs auraient traduit certains segments différemment, dans l'ensemble la traduction était satisfaisante pour les deux systèmes. En ce qui concerne la question 4, qui traite de l'efficacité de la comparaison (La traduction choisie pour la comparaison transmet le message de l'auteur), une réponse moyenne de 2.36 est constatée pour DeepL et 2.43 pour Google Translate. En même temps, cependant, on remarque un écart type assez élevé par rapport au reste des questions, égal à 1.38 pour DeepL et 1.35 pour Google Translate, ce qui est dû à la subjectivité de la réponse des participants.

Concernant les résultats des questionnaires remplis par les participants non-traducteurs de langue maternelle française, nous pouvons affirmer que dans ce cas également, les résultats sont moyens. On observe donc que pour les deux groupes d'évaluateurs, qu'ils soient des traducteurs professionnels ou non, les phrases traduites automatiquement sont généralement compréhensibles, même si on reconnait que les traductions ne sont pas faites par un être humain et qu'il y a un aspect artificiel.

Dans les questionnaires remplis par les non-traducteurs, la question 2 *(Je comprends le sens général de la phrase*) obtient le meilleur score pour les deux systèmes, avec 2.38 pour DeepL et 2.60 pour Google. Cela montre que le sens de la phrase est suffisamment compréhensible lorsqu'elle est traduite par les deux systèmes. En revanche, le plus mauvais score moyen est obtenu par la question numéro 1 (*La comparaison semble être traduite par un traducteur*

professionnel) pour les deux systèmes, ce qui nous indique que la traduction automatique est quand même reconnaissable et diffère de la traduction humaine. Les deux questions visant à évaluer l'efficacité de la comparaison, la numéro 3 (*La comparaison parvient à éveiller mon imagination*) et la numéro 5 (*L'image véhiculée par la comparaison est appropriée et explicite l'idée*), ont reçu un meilleur score moyen avec DeepL (2.52 et 2.68 respectivement). Avec Google, en revanche, la question 3 a reçu un score de 2.80 points, tandis que la question 5 a obtenu 2.92 points. D'après ces résultats, nous pouvons dire que DeepL transmet plus efficacement le sens premier d'une comparaison, même s'il ne peut pas rivaliser avec une traduction humaine.

Quant aux données obtenues par le calcul du score BLEU, l'un des exemples les plus significatifs est le segment 74 (Figure 49), qui obtient l'un des scores le plus bas pour les deux systèmes de traduction. Dans cet exemple en particulier, la comparaison « *come il cadere d'un salterello acceso in una polveriera* » n'a pas atteint son objectif premier, consistant à transmettre une image et un message efficaces, puisqu'il a été traduit par « *comme la chute d'un pistolet allumé dans un baril de poudre* » par DeepL et par « *comme la chute d'un saut allumé dans un baril de poudre* » par Google Translate, alors que la traduction humaine est « *fit l'effet d'un serpenteau qui tombe dans une poudrière* », ce qui évoque mieux l'idée de quelque chose qui tombe et bouge de manière sinueuse, dans ce cas comme un serpent. Un autre exemple qui a reçu les notes les plus basses avec le score BLEU et aussi avec le questionnaire rempli par les étudiants de traduction pour les deux systèmes est le segment 67 (Figure 50), contenant la comparaison « *come se gli fossero state peste l'ossa* », qui a été traduit par « *comme si leurs os avaient été atteints* » par DeepL et par « *comme si leurs os avaient été tourmentés* » par Google Translate, alors que la traduction humaine est « *comme s'ils avaient été foulés dans tous leurs membres* ». Comme dans l'exemple précédent, la traduction humaine parvient à mieux illustrer l'image initialement voulue par l'auteur du texte original, tandis que la traduction automatique des deux systèmes est jugée insuffisante tant par l'évaluation automatique qu'humaine. Cela confirme une fois de plus le fait que la traduction automatique ne peut aujourd'hui rivaliser avec la traduction humaine.

Dans la section suivante (Section 7), nous formulerons des conclusions, en résumant les résultats de la recherche (Section 7.1), les limites de l'étude (Section 7.2) et les possibilités de développement pour des études futures (Section 7.3).7.

7 Conclusion

Dans cette section, nous ferons un résumé des résultats et conclusions obtenus par cette recherche par rapport aux questions de recherche (Section 5.1). Nous aborderons également les limites de cette étude (Section 7.2) et ses possibles développements futures (Section 7.3).

7.1 Résumé et résultats de l'étude

Un des objectifs de cette étude (Section 5.1) était celui de comparer la traduction automatique neuronale d'un texte littéraire de l'italien vers le français, réalisée par deux systèmes de traduction différents, DeepL et Google Translate, pour comprendre dans quelle mesure ils sont capables de rivaliser avec la traduction humaine. La recherche se concentre sur les comparaisons, en portant l'attention à leur efficacité à transmettre le message et l'émotion voulus par l'auteur : nous voulons savoir si la traduction automatique arrive à transmettre cette émotion. En outre, nous voulons également savoir si les comparaisons utilisées dans le texte original ont été traduites littéralement ou si elles ont été adaptées avec d'autres expressions.

Afin de répondre à la première question de recherche, nous avons appliqué plusieurs méthodes d'évaluation. Dans la section 5.5.1 nous avons évalué les segments grâce à la participation de juges humains divisés en étudiants de traduction et non-traducteurs. Nous observons que selon les juges humains, DeepL performe mieux, en obtenant un score moyen de 2.45 par les étudiants en traduction et 2.59 par les non-traducteurs. Dans la section 5.5.2 nous avons fait une évaluation automatique en utilisant le score BLEU. Selon les résultats de l'évaluation automatique, DeepL obtient également un résultat meilleur par rapport à Google Translate, en totalisant 25.10 points. Même si les scores obtenus par DeepL sont meilleurs de ceux obtenus par Google, le score reste moyen (le meilleur score possible est 1 et le pire possible est 5). Donc on peut répondre à la première question de recherche en disant que les systèmes automatiques ne peuvent pas rivaliser les traducteurs humains professionnels à ce stade de développement de la technologie et de fournir une traduction optimale et naturelle.

En ce qui concerne la deuxième question qui porte sur la transmission de l'émotion des comparaisons dans les traductions automatiques, nous constatons que les questions portantes sur cet aspect contenues dans le questionnaire complété par les non-traducteurs ont obtenu en

moyenne des résultats meilleurs avec DeepL (Figure 30). Nous voyons aussi que la dispersion entre les réponses des juges (Figure 31) est importante pour les questions 3 (*La comparaison parvient à éveiller mon imagination*) et 5 (*L'image véhiculée par la comparaison est appropriée et explicite l'idée*), ce qui peut être expliqué par le caractère subjectif des questions. Néanmoins, le score reste moyen, la note ne s'approche pas du 1 (la note la plus haute possible) et donc n'arrive pas au même niveau d'une traduction humaine.

Nous avons également procédé à une corrélation entre les scores humains et automatiques en utilisant le coefficient relationnel de Pearson (Section 6.3) pour voir jusqu'à quel point les scores étaient en relation les uns des autres. Nous avons observé que la corrélation est faible entre le score automatique BLEU et les scores humains. Par contre, nous avons constaté que la corrélation était plus importante entre les juges humains quand il s'agissait du même système, avec un score légèrement majeur pour DeepL. Cela nous aide à comprendre qu'il y avait un accord majeur entre les humains qu'entre les scores automatiques et humains.

Concernant la troisième question de recherche, celle qui porte sur le type des comparaisons utilisées par les systèmes (littérales ou adaptations), nous avons observé que les systèmes performent mieux, aussi selon les évaluations humaines, quand il y a des traductions littérales.

Par contre, un grand problème se présente pour la traduction automatique quand il s'agit d'adapter une comparaison qui ne transmet pas le message si traduit littéralement. Nous pouvons prendre le segment 67 comme exemple (Figure 50), qui a obtenu un score très bas pour DeepL (Figure 21), de 4.5 points, et pour Google Translate (Figure 22), de 4.67 points. Dans ce cas, la comparaison est traduite sans adaptation par les systèmes.

7.2 Limites de l'étude et possibilités de développement

Pour réaliser cette étude, nous avons fait appel à des étudiants en traduction et à des personnes de langue maternelle française n'ayant aucun lien avec la traduction. Les participants n'ont bénéficié d'aucun avantage pour leur contribution à l'étude. Nous pensons que certains pourraient avoir effectué l'évaluation des traductions avec plus ou moins d'attention et de rigueur que les autres. En fait, cela pourrait expliquer certaines divergences, comme par exemple l'écart type parfois important dans l'évaluation humaine.

Nous pensons donc qu'il serait intéressant de mener une étude similaire, en faisant appel à des traducteurs professionnels et en offrant une récompense aux participants, surtout si l'on considère le temps nécessaire pour effectuer une évaluation.

Un autre aspect intéressant serait d'appliquer cette recherche à une œuvre littéraire plus récente, car les systèmes de traduction automatique sont davantage entraînés à travailler avec ce type de texte.

Il serait également intéressant d'effectuer un type de recherche similaire sur d'autres figures de style et de se concentrer encore plus sur leur efficacité à transmettre des sentiments, puisque dans ce cas, nous nous sommes concentrés uniquement sur les similitudes.

Bibliographie

Almahasees, Z. M. (2017). *Machine Translation Quality of Khalil Gibran's the Prophet* (SSRN Scholarly Paper No ID 3068518). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3068518

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S., & Sadler, L. (1994). *Machine translation: An introductory guide*. London: NCC Blackwell.

Carl, M., & Way, A. (Éds.). (2003). *Recent Advances in Example-Based Machine Translation* (Vol. 21). Springer Netherlands. https://doi.org/10.1007/978-94-010-0181-6

Chan, S. (Éd.). (2015). *Routledge encyclopedia of translation technology*. Routledge, Taylor & Francis Group.

Chinchor, N. (1992). *MUC-4 evaluation metrics in Proc. of the Fourth Message Understanding Conference 22–29.*

Cronin, M. (2013). *Translation in the digital age*. Routledge. New York.

Dirix, P., Schuurman, I., & Vandeghinste, V. (2005). *METIS-II: example-based machine translation using monolingual corpora—System description*. In Workshop on example-based machine translation (pp. 43-50).

Dorr, B., Snover, M. G., & Madnani, N. (2010). *Part 5: Machine Translation Evaluation.* Chapter 5.1 Introduction. https://www.semanticscholar.org/paper/Part-5%3A-Machine-Translation-Evaluation-Chapter-5.1-Dorr-Snover/22b6bd1980ea40d0f095a151101ea880fae33049

Fonteyne, M., Tezcan, A., & Macken, L. (2020). *Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level*. In 12th Language Resources and Evaluation Conference (pp. 3783-3791). European Language Resources Association (ELRA).

Gerlach, J. (2015). *Improving statistical machine translation of informal language: A rule-based pre-editing approach for French Forums*. (Doctoral dissertation, University of Geneva). https://doi.org/10.13097/ARCHIVE-OUVERTE/UNIGE:73226

Greene, L. (2016, février 2). Everything you ever wanted to know about Google Translate, and finally got the chance to ask. *TAUS - The Language Data Network*. https://www.taus.net/insights/reports/everything-you-ever-wanted-to-know-about-google-translate-and-finally-got-the-chance-to-ask

Hadley, J., Popović, M., Afli, H., & Way, A. (Éds.). (2019). *Proceedings of the Qualities of Literary Machine Translation*. European Association for Machine Translation. https://aclanthology.org/W19-7300

Han, L., Jones, G. J. F., & Smeaton, A. F. (2021). Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. *ArXiv:2105.03311 [Cs]*. http://arxiv.org/abs/2105.03311

Hansen, D. (2021). *Les lettres et la machine : Un état de l'art en traduction littéraire automatique*. In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles (pp. 61-78). ATALA.

Hearne, M., & Way, A. (2011). Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators. *Language and Linguistics Compass*, *5*(5), 205-226. https://doi.org/10.1111/j.1749-818X.2011.00274.x

van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer4, E. (2021) *Human evaluation of automatically generated text: Current trends and best practice guidelines*. Computer Speech & Language, 67, 101151.

Hutchins, J. (2007). *Machine translation: A concise history*. Computer aided translation: Theory and practice, 13 (29-70), 11.

Jones, R., & Irvine, A. (2013). *The (Un)faithful Machine Translator*. Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 96-101. https://aclanthology.org/W13-2713

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, P. (2017). Neural Machine Translation. *ArXiv:1709.07809 [Cs]*. http://arxiv.org/abs/1709.07809

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lavie, A. (2013). *Automated Metrics for MT Evaluation*. Machine Translation, 11, 731.

Li, L., & Sporleder, C. (2010). *Using Gaussian Mixture Models to Detect Figurative Language in Context*. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 297-300).

Likert, R. (1932). *A technique for the measurement of attitudes*. Archives of psychology.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *ArXiv:1508.04025 [Cs]*. http://arxiv.org/abs/1508.04025

Manzoni, A. (1877). Les Fiancés. Garnier Frères, Paris. https://fr.wikisource.org/wiki/Les_Fianc%C3%A9s_(Manzoni_1840)/Texte_entier

Manzoni, A. (1985). I Promessi sposi. Mondadori, Milano. http://www.letteraturaitaliana.net/pdf/Volume_8/t337.pdf

Martínez, R. (2014). *A deeper look into metrics for translation quality assessment (TQA): case study*. Miscelánea: A Journal of English and American Studies, 49.

Matusov, E. (2019). *The Challenges of Using Neural Machine Translation for Literature*. Proceedings of the Qualities of Literary Machine Translation, 10-19. aclweb.org/anthology/W19-7302

Miller, D. R., & Monti, E. (2014). *Tradurre Figure / Translating Figurative Language*.

Moss, S. (2017, septembre 21). DeepL deploys 5.1 petaflop supercomputer on Verne Global campus. *Datacenterdynamics.Com*. https://www.datacenterdynamics.com/en/news/deepl-deploys-51-petaflop-supercomputer-on-verne-global-campus/

Nguyen, K., Daumé III, H., & Boyd-Graber, J. (2017). Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1464-1474. https://doi.org/10.18653/v1/D17-1153

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318. https://doi.org/ 10.3115/1073083.1073135

Poibeau, T. (2017). *Machine Translation*. http://archive.org/details/ThierryPoibeauMachineTranslation

Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. University of Pennsylvania.

Reyes Pérez, A. (2012). *Linguistic-based Patterns for Figurative Language Processing: The Case of Humor Recognition and Irony Detection*. Universitat Politècnica de València. https://doi.org/10.4995/Thesis/10251/16692

Riccardi, A. (2011) «I promessi sposi»: Cuore dell'Italia unita. *Corriere della Sera*. Corriere.it. https://www.corriere.it/unita-italia-150/recensioni/11_gennaio_17/riccardi-prefazione-i-promessi-sposi_dfe43410-2252-11e0-83ff-00144f02aabc.shtml

Riccobono, M. G. (2015). Le similitudini nei Promessi Sposi (Quarantana). Regesto (Introduzione e capp. I-XII), in A. VV., Le radici della razionalità critica : Saperi, pratiche, teleologie. Studi offerti a Fabio Minazzi, a cura di Dario Generali, Milano, Mimesis. https://www.academia.edu/19505408/Le_similitudini_nei_Promessi_Sposi_Quarantana_Rege

sto_Introduzione_e_capp_I_XII_in_A_VV_Le_radici_della_razionalit%C3%A0_critica_saperi_pratiche_teleologie_Studi_offerti_a_Fabio_Minazzi_a_cura_di_Dario_Generali_Milano_Mimesis_2015

Riccobono, M.G. (2017). « Le Similitudini Nei Promessi Sposi (quarantana): Regesto (XIII-XXXVIII). » CONSONANZE 8: pp. 513-517. https://air.unimi.it/retrieve/handle/2434/558422/981480/2.Regesto.similitudini.Pr.Sposi.2a.parte.pdf

SAE Vehicle E/E Systems Diagnostic Standards Committee. (2001). *Translation Quality Metric*.

Shutova, E. (2010). *Automatic Metaphor Interpretation as a Paraphrasing Task*. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics. https://aclanthology.org/N10-1147.pdf

Shutova, E. (2013). *Metaphor Identification as Interpretation*. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, 276-285. aclanthology.org/S13-1040

Shutova, E. (2015). Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics, 41(4):579–623*. https://aclanthology.org/J15-4002.pdf

Singh, S. P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., & Jain, S. (2017). Machine translation using deep learning: An overview. *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, 162-167. https://doi.org/10.1109/COMPTELIX.2017.8003957

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231).

Snover, M., Madnani, N., Dorr, B. J., & Schwartz, R. (2009). Fluency, adequacy, or HTER? : Exploring different human judgments with a tunable MT metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09*, 259. https://doi.org/10.3115/1626431.1626480

Somers, H., Diaz, G. F., & de Sevilla, U. (2004). *Translation Memory vs. Example-based MT – What's the difference?* International Journal of Translation, 16 (2).

Somers, H. (2011). *Machine Translation: History, Development, and Limitations*. In The Oxford Handbook of Translation Studies (ch. 28).

Sommerlad, J. (2021, mars 24). This is how Google Translate actually works. *The Independent*. https://www.independent.co.uk/tech/how-does-google-translate-work-b1821775.html

Sporleder, C., & Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 754-762. https://doi.org/10.3115/1609067.1609151

Thurmair, G. (2009). *Comparing different architectures of hybrid machine translation systems*. In Proceedings of Machine Translation Summit XII: Posters.

Tomás, J., Mas, J. À., & Casacuberta, F. (2003). *A Quantitative Method for Machine Translation Evaluation*. In Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? (pp. 27-34).

Toral, A., & Way, A. (2014). *Is machine translation ready for literature?* In Proceedings of Translating and the Computer 36, London, UK. AsLing. https://aclanthology.org/2014.tc-1.23.pdf

Toral, A., & Way, A. (2015a). Translating Literary Text between Related Languages using SMT. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 123-132. https://doi.org/10.3115/v1/W15-0714

Toral, A., & Way, A. (2015b). Machine-assisted translation of literary text: A case study. *Translation Spaces*, *4*(2), 240-267. https://doi.org/10.1075/ts.4.2.04tor

Toral, A., Wieling, M., & Way, A. (2018). Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*. https://doi.org/10.3389/fdigh.2018.00009

Toral, A., Oliver, A., & Ballestín, P. R. (2020). *Machine translation of novels in the age of transformer*. arXiv preprint arXiv:2011.14979.

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation*. Springer Science & Business Media. https://books.google.ch/books?hl=en&lr=&id=kZ0lc3QD_3wC&oi=fnd&pg=PA1&dq=Trujillo+Translation+Engines:+Techniques+for+Machine+Translation&ots=cp0bMxGTlv&sig=blo974eGPBStc6ZO4tjG2T_0HJc&redir_esc=y#v=onepage&q=Trujillo%20Translation%20Engines%3A%20Techniqu

Turian, J. P., Shea, L., & Melamed, I. D. (2006). *Evaluation of Machine Translation and its Evaluation:* Defense Technical Information Center. https://doi.org/10.21236/ADA453509

Turovsky, B. (2016, novembre 15). Found in translation: More accurate, fluent sentences in Google Translate. *Google*. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

van Cranenburgh, A., & Bod, R. (2017). *A Data-Oriented Model of Literary Language*. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/E17-1115.pdf

Voigt, R., & Jurafsky, D. (2012). *Towards a Literary Machine Translation : The Role of Referential Cohesion*. In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, pages 18–25, Montréal, Canada. Association for Computational Linguistics. https://aclanthology.org/W12-2503.pdf

Reitter, D., & Levy, R. (2012). Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012). Association for Computational Linguistics, Montréal, Canada. https://aclanthology.org/W12-1700.pdf

Singh, S.P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., Jain, S. (2017). Machine Translation using Deep Learning: An Overview. *2017 International Conference on Computer, Communications and Electronics (Comptelix)*. pp. 162-167, doi: 10.1109/COMPTELIX.2017.8003957. https://ieeexplore.ieee.org/document/8003957

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Dean, J. (2016). Google's Neural Machine Translation System : Bridging the Gap between Human and Machine Translation. *ArXiv:1609.08144 [Cs]*. http://arxiv.org/abs/1609.08144

Figures

Annexes

Annexe 1 : Tableau résumant les résultats des réponses des étudiants en traduction pour DeepL.

	Note moyenne de chaque segment de chaque questionnaire					
	Quest. 1/6	Quest. 2/6	Quest. 3/6	Quest. 4/6	Quest. 5/6	Quest. 6/6
Seg. 1	2.467	4.333	2.133	1.200	2.133	2.600
Seg. 2	1.267	1.867	1.600	4.200	1.333	1.800
Seg. 3	1.200	3.000	2.933	2.667	2.533	1.867
Seg. 4	2.067	1.933	2.600	2.733	2.533	2.000
Seg. 5	1.733	2.000	3.333	3.000	2.067	1.333
Seg. 6	3.000	2.267	2.600	1.533	1.667	2.867
Seg. 7	1.933	2.867	2.800	4.600	2.733	3.200
Seg. 8	2.333	3.333	2.400	1.467	1.733	1.533
Seg. 9	2.667	2.333	2.867	1.333	2.467	2.667
Seg. 10	3.467	2.200	4.067	1.000	3.267	3.667
Seg. 11	2.267	3.533	1.400	1.867	2.400	1.333
Seg. 12	2.467	1.933	2.533	5.000	3.333	1.533
Seg. 13	2.000	3.733	1.867	2.400	3.200	1.933
Seg. 14	2.200	2.200	2.867	2.400	3.200	3.067
Seg. 15	3.467	2.667	3.600	4.200	2.533	1.933
Seg. 16	1.733	1.933	1.533	1.533	2.800	2.800
Seg. 17	1.667	2.867	1.400	1.000	2.667	2.533
Seg. 18	4.000	1.800	2.467	2.067	2.267	2.400
Seg. 19	2.333	1.667	2.400	1.800	2.600	4.133
Seg. 20	3.467	4.333	1.533	2.467	2.333	1.400

Annexe 2 : Tableau résumant les résultats des réponses des étudiants en traduction pour Google Translate.

	Note moyenne de chaque segment de chaque questionnaire					
	Quest. 1/6	Quest. 2/6	Quest. 3/6	Quest. 4/6	Quest. 5/6	Quest. 6/6
Seg. 1	2.333	3.733	2.267	1.067	1.933	2.733
Seg. 2	2.000	3.267	1.467	4.733	1.533	1.667
Seg. 3	1.867	3.667	3.333	5.000	2.133	1.867
Seg. 4	2.667	1.667	2.133	2.600	2.867	2.200
Seg. 5	2.400	2.733	3.467	2.933	1.733	1.800
Seg. 6	2.600	2.400	2.733	1.933	2.133	2.533
Seg. 7	2.133	2.267	4.333	4.667	1.733	1.933
Seg. 8	2.267	3.400	2.533	4.867	1.800	1.467
Seg. 9	2.867	2.333	1.667	2.933	1.867	1.733
Seg. 10	3.133	2.333	1.667	1.067	2.600	3.867
Seg. 11	2.267	1.800	1.733	1.867	1.733	2.000
Seg. 12	2.333	3.400	2.333	2.600	2.400	2.000

Seg. 13	2.333	4.067	3.933	2.667	2.467	2.200
Seg. 14	3.333	1.800	3.200	4.067	2.800	2.067
Seg. 15	2.867	3.200	3.800	4.867	2.400	3.200
Seg. 16	2.133	2.867	1.867	1.000	2.133	3.867
Seg. 17	3.000	2.800	1.000	1.000	2.467	3.400
Seg. 18	1.867	2.267	2.533	2.267	2.600	2.067
Seg. 19	2.400	3.000	2.267	2.200	2.000	2.800
Seg. 20	1.800	4.133	1.667	2.733	2.733	1.600

Annexe 3 : Tableau résumant les résultats des réponses des étudiants en traduction concernant chaque question du questionnaire pour DeepL et Google Translate.

DeepL		**Google**	
Question	Moyenne	Question	Moyenne
1	2.24	1	2.28
2	2.82	2	2.90
3	2.46	3	2.54
4	2.36	4	2.43
5	2.38	5	2.51

Annexe 4 : Tableau résumant les résultats des réponses des non-traducteurs pour DeepL.

	Note moyenne de chaque segment de chaque questionnaire					
	Quest. 1/6	Quest. 2/6	Quest. 3/6	Quest. 4/6	Quest. 5/6	Quest. 6/6
Seg. 1	1.200	3.933	2.133	1.400	3.200	3.667
Seg. 2	1.800	3.467	2.467	2.467	2.933	2.200
Seg. 3	1.400	3.333	2.067	2.067	2.800	2.600
Seg. 4	3.000	2.067	2.733	2.667	2.800	2.600
Seg. 5	2.333	3.733	2.933	2.067	2.067	2.067
Seg. 6	2.933	3.000	2.333	2.200	3.067	2.067
Seg. 7	1.867	2.733	3.533	3.333	2.600	2.267
Seg. 8	3.000	3.200	2.733	1.600	2.867	2.533
Seg. 9	1.667	2.200	2.600	1.933	2.533	2.400
Seg. 10	2.067	2.400	3.467	1.733	3.400	1.733
Seg. 11	1.467	3.400	2.600	1.533	2.867	2.333
Seg. 12	3.533	2.533	2.933	4.933	3.067	2.600
Seg. 13	2.800	2.467	2.667	1.733	2.200	1.867
Seg. 14	1.667	2.000	2.267	3.733	4.267	2.667
Seg. 15	1.733	3.067	2.667	3.800	2.667	1.600
Seg. 16	2.133	2.067	1.733	1.867	2.800	1.667
Seg. 17	1.133	3.067	2.333	2.067	2.267	4.067
Seg. 18	4.600	2.733	2.600	3.467	3.733	2.067
Seg. 19	1.667	1.867	1.533	1.800	2.733	4.067
Seg. 20	3.333	3.867	2.400	3.333	3.067	2.533

Annexe 5 : Tableau résumant les résultats des réponses des non-traducteurs pour Google Translate.

	Note moyenne de chaque segment de chaque questionnaire					
	Quest. 1/6	Quest. 2/6	Quest. 3/6	Quest. 4/6	Quest. 5/6	Quest. 6/6
Seg. 1	2.533	4.333	2.133	1.533	3.133	2.867
Seg. 2	1.733	2.200	3.533	3.333	3.533	3.333
Seg. 3	2.000	4.267	3.400	2.667	3.667	3.000
Seg. 4	2.867	2.667	2.000	1.467	2.867	2.533
Seg. 5	3.467	3.400	2.800	2.200	2.667	2.600
Seg. 6	1.933	2.867	1.933	1.933	2.467	2.333
Seg. 7	1.933	3.667	3.867	4.467	3.200	2.733
Seg. 8	3.267	4.133	1.867	4.667	2.667	2.867
Seg. 9	1.467	2.067	2.733	2.600	2.600	2.533
Seg. 10	1.267	1.400	2.333	1.200	3.600	3.933
Seg. 11	2.267	3.667	2.733	1.533	2.533	2.533
Seg. 12	3.800	2.267	2.733	2.867	2.667	2.467
Seg. 13	3.333	4.600	3.133	1.800	3.200	2.933
Seg. 14	3.800	2.933	2.800	4.267	2.400	2.800
Seg. 15	4.333	3.800	3.867	3.200	3.400	1.800
Seg. 16	2.000	2.200	2.133	1.933	2.600	3.000
Seg. 17	3.333	4.067	2.600	3.333	2.733	4.267
Seg. 18	2.667	2.600	2.800	4.467	3.200	2.600
Seg. 19	2.800	1.867	3.067	3.000	2.333	4.333
Seg. 20	2.133	3.867	2.200	1.533	2.600	2.400

Annexe 6 : Tableau résumant les résultats des réponses des non-traducteurs concernant chaque question du questionnaire pour DeepL et Google Translate.

DeepL		**Google**	
Question	Moyenne	Question	Moyenne
1	2.72	1	2.99
2	2.38	2	2.60
3	2.52	3	2.80
4	2.65	4	2.87
5	2.68	5	2.92

Annexe 7 : Corpus contenant les segments source en italien

La costiera, formata dal deposito di tre grossi torrenti, scende appoggiata a due monti contigui, l'uno detto di san Martino, l'altro, con voce lombarda, il Resegone, dai molti suoi cocuzzoli in fila, che invero lo fanno somigliare a una sega.
Avevano entrambi intorno al capo una reticella verde, che cadeva sull'omero sinistro, terminata in una gran nappa, e dalla quale usciva sulla fronte un enorme ciuffo: due lunghi mustacchi arricciati in punta: una cintura lucida di cuoio, e a quella attaccate due pistole: un piccol corno ripieno di polvere, cascante sul petto, come una collana.
Egli, tenendosi sempre il breviario aperto dinanzi, come se leggesse, spingeva lo sguardo in su, per ispiar le mosse di coloro; e, vedendoseli venir proprio incontro, fu assalito a un tratto da mille pensieri.
Cosa comanda? - rispose subito don Abbondio, alzando i suoi dal libro, che gli restò spalancato nelle mani, come sur un leggìo.
Signor curato, l'illustrissimo signor don Rodrigo nostro padrone la riverisce caramente.
Questo nome fu, nella mente di don Abbondio, come, nel forte d'un temporale notturno, un lampo che illumina momentaneamente e in confuso gli oggetti, e accresce il terrore.
Il povero don Abbondio rimase un momento a bocca aperta, come incantato; poi prese quella delle due stradette che conduceva a casa sua, mettendo innanzi a stento una gamba dopo l'altra, che parevano aggranchiate.
Gli uomini poi incaricati dell'esecuzione immediata, quando fossero stati intraprendenti come eroi, ubbidienti come monaci, e pronti a sacrificarsi come martiri, non avrebber però potuto venirne alla fine, inferiori com'eran di numero a quelli che si trattava di sottomettere, e con una gran probabilità d'essere abbandonati da chi, in astratto e, per così dire, in teoria, imponeva loro di operare.
Il nostro Abbondio non nobile, non ricco, coraggioso ancor meno, s'era dunque accorto, prima quasi di toccar gli anni della discrezione, d'essere, in quella società, come un vaso di terra cotta, costretto a viaggiare in compagnia di molti vasi di ferro.
E lei mi vorrà sostenere che non ha niente! - disse Perpetua, empiendo il bicchiere, e tenendolo poi in mano, come se non volesse darlo che in premio della confidenza che si faceva tanto aspettare.
Vuol dunque ch'io sia costretta di domandar qua e là cosa sia accaduto al mio padrone? - disse Perpetua, ritta dinanzi a lui, con le mani arrovesciate sui fianchi, e le gomita appuntate davanti, guardandolo fisso, quasi volesse succhiargli dagli occhi il segreto.
Poi alzava il viso, e, girati oziosamente gli occhi all'intorno, li fissava alla parte d'un monte, dove la luce del sole già scomparso, scappando per i fessi del monte opposto, si dipingeva qua e là sui massi sporgenti, come a larghe e inuguali pezze di porpora.
Dopo la voltata, la strada correva diritta, forse un sessanta passi, e poi si divideva in due viottole, a foggia d'un ipsilon.
Date qui, date qui, - disse don Abbondio, prendendole il bicchiere, con la mano non ben ferma, e votandolo poi in fretta, come se fosse una medicina.
L'accoglienza fredda e impicciata di don Abbondio, quel suo parlare stentato insieme e impaziente, que' due occhi grigi che, mentre parlava, eran sempre andati scappando qua e là, come se avesser avuto paura d'incontrarsi con le parole che gli uscivan di bocca.
Che? che? che? - balbettò il povero sorpreso, con un volto fatto in un istante bianco e floscio, come un cencio che esca del bucato.
Pigliate quei quattro capponi, poveretti! [...] Renzo abbracciò molto volentieri questo parere; Lucia l'approvò; e Agnese, superba d'averlo dato, levò, a una a una, le povere bestie dalla stìa,

riunì le loro otto gambe, come se facesse un mazzetto di fiori, le avvolse e le strinse con uno spago.
All'entrare, si sentì preso da quella suggezione che i poverelli illetterati provano in vicinanza d'un signore e d'un dotto, e dimenticò tutti i discorsi che aveva preparati; ma diede un'occhiata ai capponi, e si rincorò.
Così dicendo, s'alzò dal suo seggiolone, e cacciò le mani in quel caos di carte, rimescolandole dal sotto in su, come se mettesse grano in uno staio.
Il ciuffo era dunque quasi una parte dell'armatura, e un distintivo de' bravacci e degli scapestrati; i quali poi da ciò vennero comunemente chiamati ciuffi.
Se volete ch'io v'aiuti, bisogna dirmi tutto, dall'a fino alla zeta, col cuore in mano, come al confessore.
Mentre il dottore mandava fuori tutte queste parole, Renzo lo stava guardando con un'attenzione estatica, come un materialone sta sulla piazza guardando al giocator di bussolotti, che, dopo essersi cacciata in bocca stoppa e stoppa e stoppa, ne cava nastro e nastro e nastro, che non finisce mai.
Andate, vi dico: che volete ch'io faccia de' vostri giuramenti? Io non c'entro: me ne lavo le mani -. E se le andava stropicciando, come se le lavasse davvero.
Già l'avevo sentito dire ch'era un uomo da poco; ma in quest'occasione, ho dovuto proprio vedere che è più impicciato che un pulcin nella stoppa.
E si faceva tant'olio, che ogni povero veniva a prenderne, secondo il suo bisogno; perché noi siam come il mare, che riceve acqua da tutte le parti, e la torna a distribuire a tutti i fiumi.
Il cielo era tutto sereno: di mano in mano che il sole s'alzava dietro il monte, si vedeva la sua luce, dalle sommità de' monti opposti, scendere, come spiegandosi rapidamente, giù per i pendìi, e nella valle.
Alcuni andavan gettando le lor semente, rade, con risparmio, e a malincuore, come chi arrischia cosa che troppo gli preme; altri spingevan la vanga come a stento, e rovesciavano svogliatamente la zolla.
Ma perché si prendeva tanto pensiero di Lucia? E perché, al primo avviso, s'era mosso con tanta sollecitudine, come a una chiamata del padre provinciale? E chi era questo padre Cristoforo?
Due occhi incavati eran per lo più chinati a terra, ma talvolta sfolgoravano, con vivacità repentina; come due cavalli bizzarri, condotti a mano da un cocchiere, col quale sanno, per esperienza, che non si può vincerla, pure fanno, di tempo in tempo, qualche sgambetto, che scontan subito, con una buona tirata di morso.
Ma il fondaco, le balle, il libro, il braccio, gli comparivan sempre nella memoria, come l'ombra di Banco a Macbeth, anche tra la pompa delle mense, e il sorriso de' parassiti.
Per acquietare, o per esercitare tutte queste passioni in una volta, prendeva volentieri le parti d'un debole sopraffatto, si piccava di farci stare un soverchiatore, s'intrometteva in una briga, se ne tirava addosso un'altra; tanto che, a poco a poco, venne a costituirsi come un protettor degli oppressi, e un vendicatore de' torti.
Que' due si venivano incontro, ristretti alla muraglia, come due figure di basso rilievo ambulanti.
A quella vista, Lodovico, come fuor di sé, cacciò la sua nel ventre del feritore, il quale cadde moribondo, quasi a un punto col povero Cristoforo.
Allora, fatto venire un notaro, dettò una donazione di tutto ciò che gli rimaneva (ch'era tuttavia un bel patrimonio) alla famiglia di Cristoforo: una somma alla vedova, come se le costituisse una contraddote, e il resto a otto figliuoli che Cristoforo aveva lasciati.
Fermandosi, all'ora della refezione, presso un benefattore, mangiò, con una specie di voluttà, del pane del perdono: ma ne serbò un pezzo, e lo ripose nella sporta, per tenerlo, come un ricordo perpetuo.

Sa il cielo quando il podestà avrebbe preso terra; ma don Rodrigo, stimolato anche da' versacci che faceva il cugino, si voltò all'improvviso, come se gli venisse un'ispirazione, a un servitore, e gli accennò che portasse un certo fiasco.
Lo prometto. Lucia fece un gran respiro, come se le avesser levato un peso d'addosso.
Il conte tacque, e il podestà, come un bastimento disimbrogliato da una secca, continuò, a vele gonfie, il corso della sua eloquenza.
Onde, ritirata placidamente la mano dagli artigli del gentiluomo, abbassò il capo, e rimase immobile, come, al cader del vento, nel forte della burrasca, un albero agitato ricompone naturalmente i suoi rami, e riceve la grandine come il ciel la manda.
Sono nelle vostre mani; vi considero come se foste proprio mia madre.
La legge l'hanno fatta loro, come gli è piaciuto; e noi poverelli non possiamo capir tutto. E poi quante cose... Ecco; è come lasciar andare un pugno a un cristiano. Non istà bene; ma, dato che gliel abbiate, né anche il papa non glielo può levare.
E i testimoni? Trovar due che vogliano, e che intanto sappiano stare zitti! E poter cogliere il signor curato che, da due giorni, se ne sta rintanato in casa? E farlo star lì? ché, benché sia pesante di sua natura, vi so dir io che, al vedervi comparire in quella conformità, diventerà lesto come un gatto.
E i testimoni? Trovar due che vogliano, e che intanto sappiano stare zitti! E poter cogliere il signor curato che, da due giorni, se ne sta rintanato in casa? E farlo star lì? ché, benché sia pesante di sua natura, vi so dir io che, al vedervi comparire in quella conformità, [...] scapperà come il diavolo dall'acqua santa.
Lucia tentennava mollemente il capo; ma i due infervorati le badavan poco, come si suol fare con un fanciullo, al quale non si spera di far intendere tutta la ragione d'una cosa, e che s'indurrà poi, con le preghiere e con l'autorità, a ciò che si vuol da lui.
Questa parola fece venir le fiamme sul viso del frate: il quale però, col sembiante di chi inghiottisce una medicina molto amara, riprese [...].
Tonio, allungando la mano per prender la carta, si ritirò da una parte; Gervaso, a un suo cenno, dall'altra; e, nel mezzo, come al dividersi d'una scena, apparvero Renzo e Lucia.
Ci volle tutta la superiorità del Griso a tenerli insieme, tanto che fosse ritirata e non fuga.
Come il cane che scorta una mandra di porci, corre or qua or là a quei che si sbandano; ne addenta uno per un orecchio, e lo tira in ischiera; ne spinge un altro col muso; abbaia a un altro che esce di fila in quel momento.
Ci fu allora di quelli che, alzando la voce, proposero d'inseguire i rapitori: che era un'infamità; e sarebbe una vergogna per il paese, se ogni birbone potesse a man salva venire a portar via le donne, come il nibbio i pulcini da un'aia deserta.
Addio, monti sorgenti dall'acque, ed elevati al cielo; cime inuguali, note a chi è cresciuto tra voi, e impresse nella sua mente, non meno che lo sia l'aspetto de' suoi più familiari; torrenti, de' quali distingue lo scroscio, come il suono delle voci domestiche; ville sparse e biancheggianti sul pendìo, come branchi di pecore pascenti.
L'urtar che fece la barca contro la proda, scosse Lucia, la quale, dopo aver asciugate in segreto le lacrime, alzò la testa, come se si svegliasse.
Renzo uscì il primo, e diede la mano ad Agnese, la quale, uscita pure, la diede alla figlia; e tutt'e tre resero tristamente grazie al barcaiolo. - Di che cosa? - rispose quello: - siam quaggiù per aiutarci l'uno con l'altro, - e ritirò la mano, quasi con ribrezzo, come se gli fosse proposto di rubare.
Ma quella fronte si raggrinzava spesso, come per una contrazione dolorosa; e allora due sopraccigli neri si ravvicinavano, con un rapido movimento.
Due occhi, neri neri anch'essi, si fissavano talora in viso alle persone, con un'investigazione superba; talora si chinavano in fretta, come per cercare un nascondiglio.

Illustrissima signora, - disse, - io posso far testimonianza che questa mia figlia aveva in odio quel cavaliere, come il diavolo l'acqua santa: voglio dire, il diavolo era lui; ma mi perdonerà se parlo male, perché noi siam gente alla buona.
Quelle parole frizzavano sull'animo della poveretta, come lo scorrere d'una mano ruvida sur una ferita.
All'immagine del principino impaziente, tutti gli altri pensieri che s'erano affollati alla mente risvegliata di Gertrude, si levaron subito, come uno stormo di passere all'apparir del nibbio.
Tutti quegli occhi addosso alla poveretta l'obbligavano a studiar continuamente il suo contegno: ma più di tutti quelli insieme, la tenevano in suggezione i due del padre, a' quali essa, quantunque ne avesse così gran paura, non poteva lasciar di rivolgere i suoi, ogni momento. E quegli occhi governavano le sue mosse e il suo volto, come per mezzo di redini invisibili.
Ma qui, vedendo che Gertrude era diventata scarlatta, che le si gonfiavan gli occhi, e il viso si contraeva, come le foglie d'un fiore, nell'afa che precede la burrasca, troncò quel discorso, e, con aria serena, riprese: - via, via, tutto dipende da voi, dal vostro buon giudizio.
Talvolta anche, il pensiero di dover abbandonare per sempre que' godimenti, gliene rendeva arnaro e penoso quel piccol saggio; come l'infermo assetato guarda con rabbia, e quasi rispinge con dispetto il cucchiaio d'acqua che il medico gli concede a fatica.
Il cuore, trovandosene così poco appagato, avrebbe voluto di quando in quando aggiungervi, e goder con esse le consolazioni della religione; ma queste non vengono se non a chi trascura quell'altre: come il naufrago, se vuole afferrar la tavola che può condurlo in salvo sulla riva, deve pure allargare il pugno, e abbandonar l'alghe, che aveva prese, per una rabbia d'istinto.
Allora il principe si diffuse a spiegar ciò che farebbe per render lieta e splendida la sorte della figlia. Parlò delle distinzioni di cui goderebbe nel monastero e nel paese; che, là sarebbe come una principessa, come la rappresentante della famiglia; che, appena l'età l'avrebbe permesso, sarebbe innalzata alla prima dignità; e, intanto, non sarebbe soggetta che di nome.
Il signor principino è già sceso alle scuderie, poi è tornato su, ed è all'ordine per partire quando si sia. Vispo come una lepre, quel diavoletto.
Barattate queste poche parole, i due interlocutori s'inchinarono vicendevolmente, e si separarono, come se a tutt'e due pesasse di rimaner lì testa testa; e andarono a riunirsi ciascuno alla sua compagnia, l'uno fuori, l'altra dentro la soglia claustrale.
Nel vòto uggioso dell'animo suo s'era venuta a infondere un'occupazione forte, continua e, direi quasi, una vita potente; ma quella contentezza era simile alla bevanda ristorativa che la crudeltà ingegnosa degli antichi mesceva al condannato, per dargli forza a sostenere i tormenti.
Come un branco di segugi, dopo aver inseguita invano una lepre, tornano mortificati verso il padrone, co' musi bassi, e con le code ciondoloni, così, in quella scompigliata notte, tornavano i bravi al palazzotto di don Rodrigo.
Certo è che un così gran segreto stava nel cuore della povera donna, come, in una botte vecchia e mal cerchiata, un vino molto giovine, che grilla e gorgoglia e ribolle, e, se non manda il tappo per aria, gli geme all'intorno, e vien fuori in ischiuma, e trapela tra doga e doga, e gocciola di qua e di là, tanto che uno può assaggiarlo, e dire a un di presso che vino è.
Il Griso prese i due compagni, e partì con faccia allegra e baldanzosa, ma bestemmiando in cuor suo Monza e le taglie e le donne e i capricci de' padroni; e camminava come il lupo, che spinto dalla fame, col ventre raggrinzato, e con le costole che gli si potrebber contare, scende da' suoi monti, dove non c'è che neve, s'avanza sospettosamente nel piano, si ferma ogni tanto, con una zampa sospesa, dimenando la coda spelacchiata.
I vestiti o gli stracci infarinati; infarinati i visi, e di più stravolti e accesi; e andavano, non solo curvi, per il peso, ma sopra doglia, come se gli fossero state peste l'ossa.

Ma più sconcia era la figura della donna: un pancione smisurato, che pareva tenuto a fatica da due braccia piegate: come una pentolaccia a due manichi.
Renzo, salito per un di que' valichi sul terreno più elevato, vide quella gran macchina del duomo sola sul piano, come se, non di mezzo a una città, ma sorgesse in un deserto.
Andando avanti, senza saper cosa si pensare, vide per terra certe strisce bianche e soffici, come di neve; ma neve non poteva essere; che non viene a strisce, né, per il solito, in quella stagione.
[...] le insopportabili gravezze, imposte con una cupidigia e con un'insensatezza del pari sterminate, [...] andavano già da qualche tempo operando lentamente quel tristo effetto in tutto il milanese: le circostanze particolari di cui ora parliamo, erano come una repentina esacerbazione d'un mal cronico.
Nell'assenza del governatore don Gonzalo Fernandez de Cordova, che comandava l'assedio di Casale del Monferrato, faceva le sue veci in Milano il gran cancelliere Antonio Ferrer, pure spagnolo [...]. Fissò la meta, fissò la meta del pane al prezzo che sarebbe stato il giusto, se il grano si fosse comunemente venduto trentatre lire il moggio: e si vendeva fino a ottanta. Fece come una donna stata giovine, che pensasse di ringiovinire, alterando la sua fede di battesimo.
La sera avanti questo giorno in cui Renzo arrivò in Milano, le strade e le piazze brulicavano d'uomini, che trasportati da una rabbia comune, predominati da un pensiero comune, conoscenti o estranei, si riunivano in crocchi, senza essersi dati l'intesa, quasi senza avvedersene, come gocciole sparse sullo stesso pendìo.
Il primo comparire d'uno di que' malcapitati ragazzi dov'era un crocchio di gente, fu come il cadere d'un salterello acceso in una polveriera.
Ma quelli che vedevan la faccia del dicitore, e sentivan le sue parole, quand'anche avessero voluto ubbidire, dite un poco in che maniera avrebber potuto, spinti com'erano, e incalzati da quelli di dietro, spinti anch'essi da altri, come flutti da flutti, via via fino al l'estremità della folla, che andava sempre crescendo.
Il pane verrà a buon mercato, ma ci metteranno il veleno, per far morir la povera gente, come mosche.
Pane eh? - diceva uno che cercava d'andar in fretta: - sassate di libbra: pietre di questa fatta, che venivan giù come la grandine.
C'era un incalzare e un rattenere, come un ristagno, una titubazione, un ronzìo confuso di contrasti e di consulte.
Mentre ascoltan l'avviso, vedon comparire la vanguardia: in fretta e in furia, si porta l'avviso al padrone. [...] I servitori ne hanno appena tanto che basti per chiuder la porta. Metton la stanga, metton puntelli, corrono a chiuder le finestre, come quando si vede venire avanti un tempo nero, e s'aspetta la grandine, da un momento all'altro.
L'urlìo crescente, scendendo dall'alto come un tuono, rimbomba nel vòto cortile; ogni buco della casa ne rintrona: e di mezzo al vasto e confuso strepito, si senton forti e fitti colpi di pietre alla porta.
I portatori, all'una e all'altra cima, e di qua e di là della macchina, urtati, scompigliati, divisi dalla calca, andavano a onde: uno, con la testa tra due scalini, e gli staggi sulle spalle, oppresso come sotto un giogo scosso, mugghiava.
Ogni tanto però, qualche parola, anche qualche frase, ripetuta da un crocchio nel suo passaggio, gli si faceva sentire, come lo scoppio d'un razzo più forte si fa sentire nell'immenso scoppiettìo d'un fuoco artifiziale.
Chiudete ora: no; eh! eh! la toga! la toga! - Sarebbe in fatti rimasta presa tra i battenti, se Ferrer non n'avesse ritirato con molta disinvoltura lo strascico, che disparve come la coda d'una serpe, che si rimbuca inseguita.
La porta s'apre; Ferrer esce il primo; l'altro dietro, rannicchiato, attaccato, incollato alla toga salvatrice, come un bambino alla sottana della mamma.

Il vicario scendeva le scale, mezzo strascicato e mezzo portato da altri suoi servitori, bianco come un panno lavato.
Ma il voto pubblico era abbastanza spiegato per lasciar andare in prigione il vicario; e nel tempo della fermata, molti di quelli che avevano agevolato l'arrivo di Ferrer, s'eran tanto ingegnati a preparare e a mantener come una corsìa nel mezzo della folla, che la carrozza poté, questa seconda volta, andare un po' più lesta, e di seguito.
Ma tutte le strade del contorno erano seminate di crocchi. [...] là tutto un crocchio si moveva insieme: era come quella nuvolaglia che talvolta rimane sparsa, e gira per l'azzurro del cielo, dopo una burrasca; e fa dire a chi guarda in su: questo tempo non è rimesso bene.
Ora, andate a dire ai dottori, scribi e farisei, che vi facciano far giustizia, secondo che canta la grida: vi dànno retta come il papa ai furfanti: cose da far girare il cervello a qualunque galantuomo.
Però, di queste riflessioni nulla trasparve sulla faccia dell'oste, la quale stava immobile come un ritratto: una faccia pienotta e lucente, con una barbetta folta, rossiccia, e due occhietti chiari e fissi.
Un manico di coltellaccio che spuntava fuori d'un taschino degli ampi e gonfi calzoni: uno spadone, con una gran guardia traforata a lamine d'ottone, congegnate come in cifra, forbite e lucenti: a prima vista si davano a conoscere per individui della specie de' bravi.
[Il nuovo governatore], informato [...] che... ogni dì più [...] altro si sente che ferite appostatamente date, omicidii e ruberie et ogni altra qualità di delitti,[...] prescrive di nuovo gli stessi rimedi, accrescendo la dose, come s'usa nelle malattie ostinate.
Il povero curato non c'entra: fanno i loro pasticci tra loro, e poi... e poi, vengon da noi, come s'anderebbe a un banco a riscotere; e noi... noi siamo i servitori del comune.
Eh! le schioppettate non si dànno via come confetti: e guai se questi cani dovessero mordere tutte le volte che abbaiano!
A quel nuovo scongiuro, don Abbondio, col volto, e con lo sguardo di chi ha in bocca le tanaglie del cavadenti, proferì [...].
I neri e giovanili capelli, spartiti sopra la fronte, con una bianca e sottile dirizzatura, si ravvolgevan, dietro il capo, in cerchi moltiplici di trecce, trapassate da lunghi spilli d'argento, che si dividevano all'intorno, quasi a guisa de' raggi d'un'aureola, come ancora usano le contadine nel Milanese.
Era questo uno stanzone, su tre pareti del quale eran distribuiti i ritratti de' dodici Cesari; la quarta, coperta da un grande scaffale di libri vecchi e polverosi: [...] una spalliera alta e quadrata, terminata agli angoli da due ornamenti di legno, che s'alzavano a foggia di corna [...].
[...] i bravi di mestiere, e i facinorosi d'ogni genere, usavan portare un lungo ciuffo, che si tiravan poi sul volto, come una visiera, all'atto d'affrontar qualcheduno [...].
Il palazzotto di don Rodrigo sorgeva isolato, a somiglianza d'una bicocca, sulla cima d'uno de' poggi ond'è sparsa e rilevata quella costiera.
Appiè del poggio, […], e verso il lago, giaceva un mucchietto di casupole, abitate da contadini di don Rodrigo; ed era come la piccola capitale del suo piccol regno.
Il conte tacque, e il podestà, come un bastimento disimbrogliato da una secca, continuò, a vele gonfi e, il corso della sua eloquenza.
Pigliarne tre o quattro o cinque o sei, di quelli che, per voce pubblica, son conosciuti come i più ricchi e i più cani, e impiccarli.
Tonio scodellò la polenta sulla tafferìa di faggio, che stava apparecchiata a riceverla: e parve una piccola luna, in un gran cerchio di vapori.
E Perpetua? non avete pensato a Perpetua. Tonio e suo fratello, li lascerà entrare; ma voi! voi due! pensate! avrà ordine di tenervi lontani, più che un ragazzo da un pero che ha le frutte mature.

La disputa durava tuttavia, e non pareva vicina a finire, quando un calpestìo affrettato di sandali, e un rumore di tonaca sbattuta, somigliante a quello che fanno in una vela allentata i soffi ripetuti del vento, annunziarono il padre Cristoforo.
Il padre Cristoforo arrivava nell'attitudine d'un buon capitano che, perduta, senza sua colpa, una battaglia importante, afflitto ma non scoraggito, sopra pensiero ma non sbalordito, di corsa e non in fuga, si porta dove il bisogno lo chiede, a premunire i luoghi minacciati, a raccoglier le truppe, a dar nuovi ordini.
Sapete che diavoli d'occhi ha il padre: mi leggerebbe in viso, come sur un libro, che c'è qualcosa per aria.
A questo punto, Agnese si staccò dai promessi, e, detto sottovoce a Lucia: - coraggio; è un momento; è come farsi cavar un dente, - si riunì ai due fratelli, davanti all'uscio.
Due folte ciocche di capelli, che gli scappavano fuor della papalina, due folti sopraccigli, due folti baffi, un folto pizzo, tutti canuti, e sparsi su quella faccia bruna e rugosa, potevano assomigliarsi a cespugli coperti di neve, sporgenti da un dirupo, al chiaro di luna.
Il lucignolo, che moriva sul pavimento, mandava una luce languida e saltellante sopra Lucia, la quale, affatto smarrita, non tentava neppure di svolgersi, e poteva parere una statua abbozzata in creta, sulla quale l'artefice ha gettato un umido panno.
Dà di piglio alle brache, che teneva sul letto; se le caccia sotto il braccio, come un cappello di gala, e giù balzelloni per una scaletta di legno.
Il garzoncello trema come una foglia, e non tenta neppur di gridare; ma, tutt'a un tratto, in vece di lui, e con ben altro tono, si fa sentir quel primo tocco di campana così fatto, e dietro una tempesta di rintocchi in fila.
Ora, svanito così dolorosamente quel sogno, si pentiva d'essere andata troppo avanti, e, tra tante cagioni di tremare, tremava anche per quel pudore che non nasce dalla trista scienza del male, per quel pudore che ignora se stesso, somigliante alla paura del fanciullo, che trema nelle tenebre, senza saper di che.
Il palazzotto di don Rodrigo, con la sua torre piatta, elevato sopra le casucce ammucchiate alla falda del promontorio, pareva un feroce che, ritto nelle tenebre, in mezzo a una compagnia d'addormentati, vegliasse, meditando un delitto.
Se quel buon religioso lì, ottiene di mettervi nelle sue mani, e che lei v'accetti, vi posso dire che sarete sicure come sull'altare.
Queste immagini cagionarono nel cervello di Gertrude quel movimento, quel brulichìo che produrrebbe un gran paniere di fiori appena colti, messo davanti a un alveare.
Vi son de’ momenti in cui l’animo, particolarmente de’ giovani, è disposto in maniera che ogni poco d’istanza basta a ottenerne ogni cosa che abbia un’apparenza di bene e di sacrifizio: come un fiore appena sbocciato, s’abbandona mollemente sul suo fragile stelo, pronto a concedere le sue fragranze alla prim’aria che gli aliti punto d’intorno.
Uscivano, sul far del giorno, dalle botteghe de' fornai i garzoni che, con una gerla carica di pane, andavano a portarne alle solite case. Il primo comparire d'uno di que' malcapitati ragazzi dov'era un crocchio di gente, fu come il cadere d'un salterello acceso in una polveriera.
Quelli di dentro, vedendo la mala parata, scapparono in soffitta: il capitano, gli alabardieri, e alcuni della casa stettero lì rannicchiati ne' cantucci; altri, uscendo per gli abbaini, andavano su pe' tetti, come i gatti.
Quel po' di senno che gli tornò, gli fece in certo modo capire che il più se n'era andato: a un di presso come l'ultimo moccolo rimasto acceso d'un'illuminazione, fa vedere gli altri spenti.
Guardò amorevolmente l'oste, con due occhietti che ora scintillavan più che mai, ora s'eclissavano, come due lucciole.