# La reconnaissance vocale : un atout pour la post-édition ? : Étude de l'influence de cette technologie sur la productivité d'un traducteur italophone

Grasso, Silvia

*Ai miei angeli custodi*

*Marisa, Luciano e Milena*

UNIVERSITÉ DE GENÈVE
FACULTÉ DE TRADUCTION ET D'INTERPRÉTATION

# La reconnaissance vocale :

# un atout pour la post-édition ?

## Étude de l'influence de cette technologie sur la productivité d'un traducteur italophone

Directrice de mémoire : Marianne Starlander

Jurée : Johanna Gerlach

Mémoire présenté à la Faculté de traduction et d'interprétation (Département TIM, Unité d'italien) pour l'obtention de la Maîtrise universitaire en traduction, mention technologies de la traduction

Université de Genève

Étudiante : Silvia Grasso

N° de matricule : 14 309 322

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

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

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

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

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

Silvia Grasso

Genève, le 17/02/2020

# REMERCIEMENTS

Un grand merci à ma directrice du mémoire, Marianne Starlander, qui m'a suivi tout au long de ce projet, en me donnant toujours des conseils et étant toujours disponible et à l'écoute. Merci aussi pour m'avoir donné la possibilité de donner vie à cette idée et de m'avoir toujours soutenu dans la réalisation de ce mémoire. Merci pour toutes les discussions constructives et la confrontation presque quotidienne. Merci pour m'avoir transmis depuis la première année de Bachelor toutes vos connaissances et la passion pour les technologies de la traduction.

Merci à Johanna Gerlach, pour avoir accepté d'être ma jurée et pour l'aide technique et informatique lors du calcul des scores. Merci pour votre patience et votre disponibilité.

Merci à mes participants, qui ont accepté de participer à mon évaluation et qui se sont prêtés à effectuer différentes tâches et questionnaires. Merci aussi à mes évaluateurs humains pour le temps qui ont dédié à l'évaluation humaine.

Merci à Sabrina, toujours disponible et prête à m'aider quand j'étais perdue et à discuter avec moi pour me faire retrouver le bon chemin.

Merci à Sonia, pour m'avoir toujours motivée et pour avoir partagé avec moi ta positivité et tes conseils expérimentés. Merci pour avoir parcouru une partie de ce chemin avec moi. À toi aussi, merci pour ton encouragement.

Merci à Sandra, pour avoir été toujours disponible à corriger mon français pendant la rédaction et pour avoir pris le temps pour la correction et relecture finale. Merci pour le travail immense que tu as fait pour moi. Merci pour n'avoir jamais été irritable avec moi, malgré mes demandes insistantes et quotidiennes. Merci pour la compagnie et pour tes blagues. Merci pour ton encouragement continu.

Grazie a Daniela, la mia guida da sempre, per avermi insegnato tutto quello che so oggi e per aver sempre creduto in me. Grazie per il tempo dedicato alla rilettura e alla correzione di questo lavoro. Grazie per l'incoraggiamento e la presenza costante, nonostante la distanza. La distanza è spesso motivo di allontanamento, ma ho sempre saputo di averti al mio fianco.

Grazie ai miei genitori che sono stati presenti nonostante la distanza.

Grazie a mio fratello Stefano, per il modo tutto tuo di incoraggiarmi e motivarmi. Grazie per avermi aiutato ogni volta che te l'ho chiesto. Sei sempre stato la mia fonte di ispirazione.

Grazie a Esra, per aver condiviso con me questi sei anni, con tutte le sfumature di emozioni che abbiamo attraversato in questo periodo. Grazie per la tua ammirazione e per il tuo incoraggiamento. Grazie per avermi sempre ascoltata e capita, e per avermi sempre offerto il tuo aiuto. Grazie per aver condiviso con me la motivazione, per i tuoi consigli e i ragionamenti fatti insieme. Grazie per il tuo tempo e per la tua amicizia, te ne sarò sempre grata.

Infine, grazie a Barbara, per avermi sempre supportato e sopportato nei momenti peggiori e quando ero insopportabile. Grazie per le tue parole di motivazione, per la tua ammirazione e la fierezza che hai nei miei riguardi: tutto questo mi ha sempre dato la forza di andare avanti. Grazie per aver sempre creduto in me dall'inizio, per avermi sempre incoraggiato a osare e non mollare: come dice qualcuno *Riesco a non arrendermi, se ci sei tu a difendermi*. Grazie per il tempo che hai passato ad ascoltare le mie parole infinite. Grazie per aver riflettuto con me e avermi aiutato a prendere decisioni importanti. Grazie per la tua presenza quotidiana, te ne sarò sempre riconoscente.

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# LISTE DES ABRÉVIATIONS

| | |
|---|---|
| CAT | Computer aided translation |
| DNS | Dragon Naturally Speaking |
| HMM | Hidden Markov Model (modèle de Markov caché) |
| IA | Intelligence artificielle |
| PE | Post-édition |
| RAP | Reconnaissance automatique de la parole |
| RV | Reconnaissance vocale |
| TA | Traduction automatique |
| TALN | Traitement automatique de la langue naturelle |
| TAN | Traduction automatique neuronale |
| TAO | Traduction assistée par ordinateur |
| TAS | traduction automatique statistique |
| TD | Traduction dictée |

# 1. Introduction

Les technologies de la traduction ont connu un véritable essor ces dernières années. Les progrès de la traduction automatique sont nombreux : intégration de cette dernière à des logiciels de traduction assistée ainsi qu'à des outils et des programmes en ligne ; l'avènement de la post-édition comme tâche et création du nouveau métier de post-éditeur ; utilisation de la reconnaissance automatique de la parole (RAP) non seulement en traduction, mais également dans d'autres domaines. En effet, de nombreuses technologies sont apparues sur le marché, telles que Google Assistant par Google, Alexa par Amazon, Siri par Apple, ou encore Swisscom Box par Swisscom. Toutes ces enceintes se basent sur un système de reconnaissance vocale très développé, qui permet une nouvelle gestion de la vie quotidienne. Sachant cela, notre travail se concentre plutôt sur la reconnaissance vocale employée comme outil d'aide à la traduction.

Malgré cet essor dans le domaine, les traducteurs plus sceptiques ont des doutes par rapport à l'avenir de leur métier. Ces craintes sont compréhensibles au vu de l'expansion de la traduction automatique ces dernières années, notamment depuis l'arrivée des systèmes neuronaux sur le marché. En effet, l'industrie de la traduction automatique générait plus de 100 000 millions de mots par jour en 2017 déjà (Nolla & Abril, 2017). Par ailleurs, les clients s'intéressent vivement à cette évolution et sont de plus en plus nombreux à demander aux traducteurs de corriger les textes produits par des systèmes de traduction automatique, soit d'effectuer des tâches de « post-édition ».

Cependant, le but de ces technologies n'est pas de remplacer les traducteurs, mais, au contraire, de leur fournir une aide importante dans leur travail, d'où le nom de « traduction assistée par ordinateur ». Il est également important de souligner que ces nouveaux outils, tels que la traduction automatique, les logiciels de traduction assistée ainsi que la reconnaissance vocale, requièrent une formation et des connaissances préalables, afin de bien les utiliser et d'en tirer profit.

Puisque la reconnaissance vocale (RV) fait désormais partie intégrante de nos vies, et constitue une nouvelle manière d'écrire des textes de toute longueur, il est nécessaire de mieux cerner les avantages et ses limites. De plus, bien que beaucoup de tâches soient plus faciles à réaliser à l'aide d'un clavier, l'utilisation de la parole permet une meilleure interface pour les tâches pour lesquelles le clavier n'est pas adapté ou pour celles où la communication constitue un élément important (Dan Jurafsky & Martin, 2009).

## 1.1 Objectifs

L'objectif principal de ce mémoire est d'étudier l'impact de la reconnaissance automatique de la parole sur l'activité d'un traducteur (ou post-éditeur) italophone. Nous avons choisi de nous intéresser à la langue italienne, en particulier, comme c'est notre langue maternelle. De ce fait, notre connaissance de cette langue est plus profonde et elle nous permettra de tester plus efficacement le logiciel de RV.

Puisque la reconnaissance vocale occupe une place prépondérante dans nos vies, et est également utilisée dans le travail du traducteur, nous estimons nécessaire d'analyser en détail son fonctionnement et ses apports aux métiers de la traduction. Par ailleurs, l'introduction des réseaux neuronaux a grandement amélioré la qualité de la RV et pourrait également influencer la productivité, que ce soit du point de vue quantitatif (temps) et du point de vue qualitatif (qualité de l'output). Néanmoins, il est important d'étudier tous ces aspects à travers une expérience menée dans le domaine afin de vérifier ce constat.

Nous avons choisi de nous pencher sur ce sujet car la reconnaissance vocale intégrée à la post-édition est un sujet encore très peu étudié. Puisque les deux technologies sont très récentes, leur utilisation combinée n'a pas encore été considérée comme un véritable atout dans ce métier. Nous tenterons donc de déterminer si l'utilisation combinée de ces deux technologies exerce une réelle influence sur le métier du traducteur.

## 1.2 Plan

Après avoir exposé nos objectifs pour ce mémoire, nous présenterons brièvement l'historique de la reconnaissance vocale au chapitre 2, pour mieux comprendre l'évolution de cette technologie encore assez récente.

Le chapitre 3 sera dédié à l'état de l'art. Nous citerons certaines études conduites dans ce domaine encore peu étudié et mentionnerons les études principales qui constituent la base de notre recherche.

Au chapitre 4, nous présenterons la traduction automatique neuronale et expliquerons en détail le fonctionnement des réseaux neuronaux. Nous définirons ensuite la post-édition et exposerons les typologies de cette dernière et le nouveau métier de post-éditeur.

Le chapitre 5 concernera le sujet au cœur de notre recherche : la reconnaissance vocale. Nous présenterons d'abord le fonctionnement d'un système de reconnaissance vocale, pour ensuite

spécifier les deux modèles de langage statistique et neuronal. Nous terminerons par une brève présentation du logiciel que nous avons utilisé.

Le chapitre 6 présentera la méthodologie EAGLES/ISO utilisée et expliquera quelles sont les caractéristiques des standards ISO utilisées et comment celles-ci seront mesurées.

Le chapitre 7 définit le protocole de l'expérience effectuée, avec une présentation de tous les logiciels utilisés lors de cette dernière ainsi qu'une introduction sur le profil de nos participants. Enfin, nous détaillerons le déroulement de l'expérience.

Les résultats et notre analyse sont présentés dans le chapitre 8. Nous les présenterons caractéristique par caractéristique en incluant les réponses fournies par les participants au questionnaire de post-évaluation.

Nous terminerons par une conclusion au chapitre 9. Nous présenterons les principaux résultats de notre expérience, nous évoquerons les limites de notre travail et nous proposerons différentes pistes de recherches qui pourront être conduites à l'avenir.

# 2 Historique

Dans le contexte de la traduction, à partir des années soixante et septante, les dictaphones ont été les premiers outils à être utilisés par les organisations internationales. À cette époque, les traducteurs collaboraient avec les dactylographes qui dictaient leurs traductions. Ensuite, à partir de la fin du XIX$^{e}$ siècle, beaucoup de progrès ont été faits dans l'exploration des objectifs et des enjeux de la reconnaissance automatique de la parole (RAP). Le but principal consistait à comprendre et à étudier une possible réduction du taux d'erreur combinant la reconnaissance vocale (RV) avec la traduction automatique (TA) (Vidal et al., 2006). Récemment, des progrès importants ont été enregistrés dans les outils professionnels de RV pour la traduction, mais aussi pour la vie quotidienne dans les environnements mobiles, tels que les smartphones. En outre, un autre champ de recherche très actuel concerne l'utilisation de la reconnaissance vocale pour la post-édition : la dictée pourrait considérablement améliorer la rapidité et, par conséquent, augmenter la productivité du processus de traduction (Liyanapathirana et al., 2019). D'après Garcia-Martinez et al. (2014) « *voice input is more interesting than the keyboard alone for post-editing* »; Mesa-Lao (2014) demontre effectivement dans son experience que « *12 out of 15 translators would welcome the integration of voice as one of the possible input modes for performing PE tasks* ».

Nous comprenons, donc, que la dictée de la parole n'est pas une nouveauté dans le domaine de la traduction. Aujourd'hui, on utilise surtout la reconnaissance automatique de la parole afin de dicter des phrases entières ou des mots à corriger comme dans le cas de la post-édition. « L'un des premiers appareils d'enregistrement utilisés en TD [traduction dictée] a été le magnétophone […] auquel ont succédé les dictaphones à cassette, […] et aujourd'hui numériques. » (Zapata & Quirion, 2016). Dans notre domaine, comme le décrit Jiménez Ivars (1998), les traducteurs utilisent la dictée pour différentes applications : ils dictent une première version brute pour, ensuite, taper la traduction finale ; ou ils enregistrent certaines phrases pour les réécouter et contrôler la qualité et la fluidité, ou encore ils utilisent les enregistrements pour les soumettre à des experts afin de vérifier la précision de la terminologie. Néanmoins, son usage principal reste la traduction, où la dictée constituait initialement une tâche à effectuer par un copiste, par le client même ou par le langagier. Nous remarquons donc un intérêt constant pour la dictée, qui a été ranimé par la reconnaissance vocale et qui a apporté « une dimension nouvelle à la TD conventionnelle au dictaphone » (Ibidem).

Effectivement, s'il y a une cinquantaine d'années les traducteurs dictaient leurs textes à travers un magnétophone ou dictaphone et un copiste transcrivait l'enregistrement par la suite, « à l'ère du numérique, il est désormais possible d'aller au-delà de la TD conventionnelle. » (Zapata & Quirion, 2016). Gouadec (2009) affirme que le recours à la reconnaissance vocale est comme « la revanche du dictaphone, qui avant l'arrivée du traitement de texte, constituait le seul moyen de gains de productivité. » et que l'on peut « décrire les systèmes de [RV] comme des dictaphones qui saisissent le texte ».

L'élément positif de la traduction dictée interactive (TDI) est le fait que le traducteur puisse non seulement dicter les phrases, mais donner aussi des commandes. Ces manipulations sont aujourd'hui possibles grâce à des outils tels que Dragon Naturally Speaking. Ce dernier est l'un des meilleurs logiciels sur le marché et se démarque des autres systèmes de reconnaissance automatique de la parole qui prévoient seulement la dictée et non pas les commandes vocales (*voice tags*). De plus, la reconnaissance vocale utilisée dans ce domaine constitue un avantage non négligeable, étant donné l'état du « marché actuel où la rémunération est fortement liée au nombre de mots traduits en un laps de temps donné », ce qui en fait « un atout économique potentiel. » (Zapata & Quirion, 2016).

# 3 État de l'art

De nombreux systèmes de reconnaissance vocale sont apparus sur le marché tels que Siri, introduit par Apple en 2011 (Izbassarova et al., 2020), Google Home et Alexa Amazon, pour citer les plus célèbres au niveau mondial. Des systèmes similaires ont également été introduits sur des territoires plus restreints comme la Suisse, tels que la Swisscom Box de Swisscom, dont le fonctionnement se rapproche de ceux cités plus haut, même si ce système est plus focalisé sur la télévision[1]. Nous remarquons, donc, un véritable intérêt pour tout ce qui est lié à la reconnaissance vocale et aux commandes vocales dans la vie quotidienne : dicter des messages sur son smartphone, allumer ou éteindre les lumières, changer les chaînes à la télévision ou démarrer de la musique, etc. Les emplois sont nombreux et très variés ; cependant, dans le cadre de notre recherche, nous nous intéresserons à l'utilisation de la reconnaissance automatique de la parole dans le domaine de la traduction et de l'environnement de ce métier.

La reconnaissance automatique de la parole peut s'avérer utile pour des scénarios très actuels qui requièrent une transcription des discours, comme des interventions parlementaires, des

[1] https://www.swisscom.ch/it/clienti-privati/abbonamenti-tariffe/inone-home/swisscom-tv.html consulté le 06/01/2020

interviews, ou des entretiens. Elle réduit le travail de l'humain concernant la correction des erreurs de reconnaissance, de ponctuation ou de formatage (Salimbajevs & Ikauniece, 2017). C'est pour cela que, récemment, de plus en plus d'éditeurs de transcription et d'outils de reconnaissance vocale sont apparus sur le marché. Toutefois, si les outils pour l'exécution de commandes vocales cités auparavant sont très nombreux, nous ne pouvons pas affirmer la même chose pour les logiciels professionnels dans ce domaine : le plus performant, ainsi que le leader sur le marché, est Dragon Naturally Speaking de *Nuance*, dont nous nous sommes servie pour notre expérience. Néanmoins, si l'on cherche des outils du même niveau dans l'industriel, on éprouvera des difficultés à en trouver qui proposent les mêmes fonctionnalités ainsi que la même qualité de reconnaissance, de transcription et d'exécution des commandes vocales. Différents outils d'aide à la traduction offrent désormais l'intégration de la RV : Matecat, bien qu'il ne permet que la dictée des phrases et ne comprend aucune commande ni aucun signe de ponctuation, ce qui le rend très peu performant et peu utile ; MemoQ, dans sa version 8.7 qui, depuis décembre 2018, a aussi intégré la fonctionnalité de reconnaissance vocale, grâce à la sortie de « *Hey MemoQ* » (Lossner, 2018). Cependant, pour pouvoir se servir de cet outil, il faut avoir un dispositif avec système d'exploitation iOS et dicter à travers celui-ci[2]. L'avantage de *Hey MemoQ* par rapport à Dragon Naturally Speaking réside dans le nombre de langues supportées, comme le montre aussi Lossner dans son article « *Integrated iOS speech recognition in memoQ 8.7* »[3] sur son blog.

En outre, Microsoft a également développé son propre système de RV. Bien que l'API (SAPI) de Microsoft avait déjà été publié pour Windows 98, c'est la dernière version, Microsoft Speech API 5.4, à être utilisée depuis quelques années (Duarte et al., 2014). Anticipé par Cortana, développée à partir de 2009 et intégrée pour la première fois en 2014 dans le Windows Phone 8.1, Windows perfectionne ensuite son propre système, intégré pour la première fois à Windows 7[4], qui est désormais inclus au système d'exploitation Windows 10. Malheureusement, ce système de RV prévoit seulement six langues : anglais américain et britannique, français, espagnol, chinois simplifié et traditionnel, allemand et japonais[5] (Ciobanu, 2014). L'italien n'est pas inclus, raison pour laquelle nous n'avons pas pu nous en servir pour une comparaison

[2] https://www.memoq.com/products/hey-memoq consulté le 08/01/2020
https://blog.memoq.com/hey-memoq-frequently-asked-questions consulté le 08/01/2020
[3] https://www.translationtribulations.com/2018/12/integrated-ios-speech-recognition-in.html consulté le 08/01/2020
[4] https://support.microsoft.com/en-us/help/14213/windows-how-to-use-speech-recognition consulté le 06/01/2020
[5] https://answers.microsoft.com/en-us/windows/forum/all/what-languages-does-microsoft-speech-recognition/cb5eaf9d-7391-4ab5-8ce9-f1e44096c853 consulté le 06/01/2020

dans le cadre de notre expérience. En revanche, *Nuance* a ajouté l'italien, le néerlandais et le japonais à Dragon Naturally Speaking, mais il ne supporte pas le chinois (Ibidem). Depuis 2016, Microsoft a aussi intégré la reconnaissance vocale à Microsoft Word[6] et, les langues supportées sont les mêmes sauf le japonais, qui est remplacé par l'italien. À ce propos, un article paru sur INTERSPEECH 2017 par Salimbajevs & Ikauniece (2017) présente la transcription à travers la reconnaissance vocale directement sur Microsoft Word, car il est de facto l'outil standard le plus utilisé pour l'élaboration de documents. Il serait donc très pratique et approprié d'effectuer le processus de transcription entier dans un seul logiciel. Néanmoins, le système de Microsoft n'est pas le seul à proposer ce type d'intégration. En effet, Dragon s'intègre parfaitement au système d'exploitation et à tout type de logiciel sans causer de problèmes ni engendrer des difficultés d'utilisation, comme nous avons pu le remarquer lors de notre expérience.

Pour ce travail, nous nous appuierons majoritairement sur les études conduites par Liyanapathirana et al. (2019) et Mesa-Lao (2014), qui présentent des expériences avec beaucoup de points en commun avec notre recherche. La post-édition est un sujet très actuel, que ce soit du point de vue professionnel ou de la recherche. Toutefois, comme l'indiquent aussi Liyanapathirana et al. (2019), la reconnaissance vocale intégrée dans la post-édition demeure un sujet nouveau et très peu étudié. C'est pour cela que les études conduites dans ce domaine sont encore peu nombreuses, bien qu'il y ait beaucoup d'études sur les deux sujets analysés indépendamment.

La première trace d'une étude conduite dans ce domaine remonte à un sondage pionnier effectué au Canada en 1978 auprès de 44 traducteurs. Il ressort de cette étude que 49 % des personnes interviewées travaillaient avec le stylo, 33 % tapaient à la machine à écrire et seulement 18 % dictaient les textes. De plus, pendant cette période, plusieurs études se sont concentrées sur les outils utilisés dans la traduction dictée, ainsi que sur l'évaluation de ces produits et de leur qualité. Trois ans après, Laroque Divirgilio présente une expérience conduite sur 43 étudiants de la Faculté de Traduction à l'Université de Montréal. Le but de cette expérience était de simuler l'utilisation de la traduction dictée en milieu professionnel. Il « a révélé l'enthousiasme des participants envers cette technique et a conclu que la TD conduit bel et bien à des gains en productivité, sans que la qualité ne soit affectée. » (Zapata & Quirion, 2016). Du point de vue

[6] https://support.microsoft.com/it-it/help/14198/windows-7-dictate-text-using-speech-recognition consulté le 06/01/2020

des chiffres strictement, cette étude montre une économie de 24 minutes en moyenne pour environ 260 mots, ce qui correspond à un gain de productivité de 20 %.

La reconnaissance vocale semble donc notamment être utilisée pour la dictée de textes entiers, tandis qu'au sein de la post-édition, elle n'est pas encore complétement intégrée. Puisque la place de la post-édition progresse rapidement dans le domaine de la traduction, certaines études ont été conduites sur le sujet et ont progressivement intégré la reconnaissance vocale dans le processus de post-édition des textes. Le but premier était d'analyser ainsi si une telle intégration pourrait constituer un gain au niveau de la productivité. De manière générale, pour la dactylographie, la dictée permet d'écrire 160 mots à la minute au lieu de 70, soit plus du double (Bouillon, 2017a). Ciobanu (2014) révèle dans son étude qu'en termes de productivité, un traducteur n'utilisant pas la RV a un rendement d'environ 3000 mots/jour, tandis que si la RV est utilisée, cette moyenne monte à 3100 mots/jour. Le gain le plus important est constaté lorsque le traducteur combine les deux approches : lorsqu'il tape et dicte, sa moyenne journalière s'élève à environ 3500 mots/jour. La productivité maximale serait atteinte si les enregistrements étaient envoyés à un dactylographe, avec un rendement de 4000 mots/jour. Néanmoins, il est fondamental d'observer si le gain du temps n'affecte pas la qualité du texte. De plus, cela pourrait également avoir des avantages notamment au niveau du confort (ergonomie physique) : beaucoup de traducteurs se plaignent effectivement de fortes tendinites à cause de la position du poignet lors de l'utilisation constante de la souris et du clavier durant le travail (Bouillon, 2017a) ou encore de mal au cou ou au dos, ainsi que d'une fatigue des yeux (Ciobanu, 2014). D'après Pym et al. (2013), les traducteurs professionnels passent plus de quatre heures par jour devant l'écran de leur ordinateur, ce qui augmente sans doute la fatigue des yeux.

Comme mentionné plus haut, Mesa-Lao (2014, p. 99) a conduit une étude sur la reconnaissance automatique de la parole utilisée dans la tâche de post-édition.

> « *ASR systems have the potential to improve the productivity and comfort of performing computer-based tasks for a wide variety of users, allowing them to enter both text and commands into the computer using just their voice. However, further studies need to be conducted to build up new knowledge about the way in which state-of-the-art ASR software can be applied to one of the most common tasks translators face nowadays, i.e. post-editing of MT outputs.* »

Cette étude a deux buts principaux : étudier la satisfaction des traducteurs sur la base du sondage effectué après avoir montré l'utilisation de la RV dans la PE et, en s'appuyant sur les retours donnés par les participants, évaluer l'éventuel changement du ressenti des utilisateurs. Cette

technique paraît étonnante pour les traducteurs moins habitués à son utilisation dans la vie professionnelle, qui voient la dictée comme un divertissement plutôt qu'un véritable outil d'aide à la traduction. L'étude vise à analyser le potentiel de l'union entre MemoQ et Dragon Naturally Speaking (Mesa-Lao, 2014). Les participants, étant tous des traducteurs, étaient des utilisateurs habituels des systèmes de mémoire de traduction (SMT) tels que MemoQ ou SDL Trados Studio, et avaient aussi des connaissances en post-édition. Dans une première phase, ils ont dû répondre à un questionnaire en ligne. Dans la deuxième phase, ils ont passé une véritable expérience avec deux conditions différentes : la modalité « non-ASR », donc avec le clavier et la souris, et la modalité « ASR *input modality* » où la reconnaissance vocale a été combinée avec les méthodes traditionnelles d'entrée. Selon les résultats de l'expérience, les avantages de l'utilisation de la reconnaissance vocale seraient surtout la réduction de la fatigue, la rapidité ainsi que la facilité d'utilisation. Par contre, la principale raison pour laquelle les participants ne l'utiliseraient pas est que les méthodes traditionnelles d'entrée sont plus faciles à utiliser, notamment grâce aussi aux raccourcis, considérés comme plus rapides et plus simples d'utilisation. Une autre cause serait aussi le temps d'investissement employé pour entraîner le système afin d'améliorer la précision. Le plus grand doute reste tout de même lié à la précision (*accuracy*) : le questionnaire de satisfaction auquel les participants ont répondu montre un écart assez conséquent (52,7 % pour la RV contre 85,3 % en faveur des méthodes traditionnelles).

Néanmoins, il est possible que ces doutes soient en grande partie dus à une mauvaise connaissance des logiciels ainsi qu'à un manque d'expérience. Comme l'écrit Mesa-Lao (2014), les participants sont globalement d'accord pour affirmer qu'il est plus facile d'utiliser la RV pour la post-édition que ce qu'ils pensaient en réalité. La même impression positive concerne la reconnaissance de commandes vocales telles que « *select* » ou « *scratch* » (Ibidem). Les résultats sont donc positifs par rapport à la reconnaissance vocale utilisée pour la tâche de post-édition, ce qui donne de l'espoir dans l'intégration de cette technologie, comme une « aide » aux traducteurs et non pas comme un « remplaçant ».

D'après leur étude, Liyanapathirana et al. (2019) affirment que six traducteurs, parmi les participants, déclarent d'utiliser les méthodes d'entrée vocale quotidiennement, en particulier pour dicter des messages vocaux sur leurs smartphones ou pour donner des commandes du type Google Assistant ou Alexa. Toutefois, lorsqu'on leur demande d'utiliser des logiciels de reconnaissance de la parole, tels que Dragon Naturally Speaking, ils ne sont pas complètement d'accord pour différentes raisons : ils jugent que la dictée est « à la mode » mais pas efficace, et ils estiment que la perturbation causée aux collègues dans des bureaux en open-space

constitue un problème important. Toutefois, bien que la plupart des traducteurs (73 %) estiment que la reconnaissance automatique de la parole peut être une méthode d'entrée plus rapide, 79 % ne croient pas qu'elle soit plus précise que l'écriture (Ibidem). Il faut souligner que l'étude conduite par Liyanapathirana cible les organisations internationales : elle prend en considération des traducteurs travaillant à l'Organisation Mondiale du Commerce (OMC) et montre que sept personnes sur huit parmi celles qui jugent la reconnaissance vocale utile l'utilisent déjà dans leur travail. Le problème semblerait résider dans le manque de volonté de mélanger la reconnaissance vocale à la post-édition. D'après Liyanapathirana et al. (2019, p. 153) « *8 out of 17 translators were open to the idea of speech-based post-editing for translation and only 2 out of 17 assumed that mixing speech and post-editing together would be confusing.* », ce qui montre l'ouverture pour les technologies, sous réserve que celles-là ne soient pas mélangées, afin d'éviter toute confusion.

Une autre étude a été menée sur la possibilité d'intégrer la RV à la post-édition. (Garcia-Martinez et al. (2014) ont publié un article relatif à une enquête pré-pilote et pilote où la première phase comprenait les différentes tâches suivantes : les deux participants devaient traduire de zéro en n'utilisant d'abord que le clavier, ensuite en utilisant seulement la technologie de la reconnaissance vocale. Ensuite, le même processus était appliqué à la post-édition. Enfin, ils devaient effectuer la traduction et la post-édition avec les deux modalités combinées. À noter qu'un des participants était traducteur, tandis que l'autre n'avait aucune connaissance préalable, les résultats mettent en évidence que la combinaison de taper et dicter peut enregistrer une productivité similaire. Les chercheurs ont obtenu un meilleur résultat dans l'étude pilote : les participants étaient tous des traducteurs habitués aux outils d'aide à la traduction. Ils ont employé moins de temps dans la phase de post-édition à « modalité combinée » que dans la phase où seulement l'interaction avec le clavier était possible. Il faut souligner que, s'ils avaient des connaissances concernant les outils d'aide à la traduction et à la post-édition, ils n'en avaient aucune relative à la reconnaissance vocale. De plus, les participants ont estimé que la RV semble être prometteuse et en pleine croissance avec ses fonctionnalités dans le cadre d'un travail à l'aide des outils d'aide à la traduction. Ils estiment néanmoins qu'une préparation adéquate serait nécessaire afin d'acquérir davantage de connaissances de cette technologie, spécialement dans un contexte de post-édition.

Comme nous l'avons déjà évoqué auparavant, ce sujet est encore très peu étudié actuellement : parmi les études conduites sur la post-édition effectuée par le biais de la reconnaissance vocale nous ne comptons que les trois recherches que nous avons présentées. Pour cette raison, nous

allons présenter brièvement certaines études qui analysent la reconnaissance vocale pour la traduction et mettent en avant ses avantages et ses inconvénients.

Une étude à grande échelle dans le domaine a été conduite par Dragsted et al. (2011). Leur étude présente certaines différences, telles que la direction de la traduction (les participants ont traduit vers leur L2) et, surtout elle n'inclut pas la post-édition. Malgré cela, l'étude montre de nombreuses similarités avec celles citées ci-dessus, ainsi qu'avec notre recherche. Les chercheurs s'interrogent, précisément, sur la possibilité d'un gain de temps pour une traduction à vue effectuée à l'aide de la reconnaissance vocale. D'après une autre étude précédente conduite à petite échelle, on avait démontré un gain de temps remarquable à l'oral en comparaison avec la traduction écrite, sans toutefois affecter la qualité. Cela étant, la traduction demeure écrite à l'écran, ce qui amène à une meilleure qualité de l'output de la RV, par rapport à la qualité d'un potentiel texte traduit à vue. Une des raisons pour lesquelles ils ont décidé de développer ce projet concerne la maîtrise des technologies de la traduction de plus en plus présente dans la vie professionnelle des traducteurs (Ibidem). Il est important de remarquer que dans l'étude conduite à petite échelle, seulement les utilisateurs expérimentés en reconnaissance vocale ont obtenu un gain de temps important avec le logiciel de RV.

Bassil et Alwani (2012) montrent dans leur recherche que les systèmes de reconnaissance automatique de la parole sont encore susceptibles de faire des erreurs et imprécis, notamment si utilisés dans des milieux inadéquats. Les principales erreurs de reconnaissance vocale concernent généralement les fautes d'orthographe et la cause principale de ces erreurs reste un bruit excessif dans l'environnement de travail, ainsi que la qualité du discours et de son énonciation. Nous observons donc des similarités avec les autres études, bien que leur objectif soit d'analyser une approche de correction sur les technologies de suggestion de résultats de Bing, afin de détecter des erreurs linguistiques ou lexicales engendrées par la RV. Si nous nous concentrons sur l'utilisation de la post-édition à travers la reconnaissance vocale, les chercheurs souhaitent plutôt observer si cette technologie est performante ou si elle produit encore beaucoup d'erreurs.

La reconnaissance vocale est utilisée de manière différente dans les systèmes de traduction assistée : on pense qu'il est préférable de les utiliser pour dicter une traduction entière, mais elle est employée aussi pour déterminer l'acceptabilité de certaines parties de la traduction. Cependant, d'après Vidal et al. (2006) la reconnaissance vocale actuelle n'est pas encore en mesure de permettre une précision suffisamment élevée et, pour cela, elle requiert encore une

intervention humaine. À noter que cette affirmation date de 2006 : entretemps la RV s'est améliorée notablement, comme nous pouvons le constater avec Dragon ainsi qu'avec les autres systèmes cités. Néanmoins, cette étude corrobore celle présentée ci-dessus, dont l'objectif est d'analyser l'efficience et l'efficacité de la simple reconnaissance vocale et d'observer le degré d'intervention humaine nécessaire. Vidal et al. (2006) estiment que dans un système de traduction assistée, la productivité de la traduction peut augmenter ultérieurement si la « parole » est utilisée. Toutefois, ils affirment que les traducteurs humains peuvent commettre des erreurs, notamment si la traduction est dictée au lieu d'être tapée. Un avantage majeur par rapport à « *the text-only version of CAT* » est que le traducteur peut maintenant choisir s'il souhaite taper ou dicter :

> « *The key point is that speech should be encouraged only if low-perplexity recognition is possible, while typing should be preferred when the results of speech recognition are expected to be poor.* » (Vidal et al., 2006, p. 944)

Enfin, Ciobanu (2014) a conduit une recherche en 2014 à la *University of Leeds Centre for Translation Studies* (CTS) afin d'observer la pratique des traducteurs professionnels utilisant les systèmes de reconnaissance vocale durant leur travail. En effet, Ciobanu s'intéresse à la reconnaissance vocale dans sa globalité, soit un des sujets les plus étudiés ces dix dernières années (« Introducing Translatotron », 2019[7]) ; en outre, il a également donné un cours démonstratif de l'intégration de DNS à un logiciel de TAO, notamment MemoQ, en montrant que, globalement, le système fonctionne bien et qu'il serait donc possible de l'intégrer au travail du traducteur[8]. Le but de sa recherche est d'observer les avantages et les désavantages d'un logiciel de reconnaissance automatique de la parole dans le travail du traducteur. Pour ce faire, il a mené sa recherche sur des personnes utilisant régulièrement la RV : il en ressort que le groupe dont l'expérience sur le terrain est entre 11 et 25 ans semble être celui qui utilise le plus la RV et, à l'intérieur de ce groupe, seule une personne est freelance avec moins d'une année d'expérience. En examinant le profil des professionnels, on remarque que seulement une personne en situation de handicap a requis cette technologie. Le reste du groupe l'a choisie pour des raisons variées telles qu'une meilleure qualité de la traduction, pour la productivité ou un meilleur environnement de travail. En ce qui concerne l'environnement de travail, plus de 95 % des participants affirment travailler dans leur propre bureau, bien que deux personnes travaillent

[7] Consultable à l'adresse https://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html consulté le 08/01/2020

[8] https://amara.org/it/videos/J3p2jHYnOyoN/info/demonstration-of-multilingual-tts-memoq-and-dns-integration/ consulté le 08/01/2020

dans un espace partagé. En outre, le système d'exploitation préféré est Windows (Ciobanu, 2014). L'élément central de la recherche reste toutefois le logiciel de reconnaissance vocale privilégié : Dragon Naturally Speaking est, en effet, préféré par tous les participants. Ce résultat corrobore donc notre étude ainsi que la position de Dragon comme leader sur le marché. Un problème à ne pas négliger, évoqué par Ciobanu (Ibidem), réside dans les difficultés propres à chaque langue : le français est un bon exemple avec ses terminaisons muettes, ses formes différentes entre singulier et pluriel et sa grande variété de conjugaison. Dans ces cas, un système de reconnaissance vocale, bien que très efficace et performant comme DNS, est induit en erreur, notamment à cause des fréquents cas d'homophonie et d'ambiguïtés ; cela mène donc à un important processus de révision, ce qui est moins rentable au niveau de la production et de l'utilisation de la reconnaissance vocale.

# 4 Post-édition

D'importants progrès ont été accomplis dans le domaine des technologies de la traduction ces dernières années, ce qui a permis d'intégrer ces technologies au workflow des différentes organisations internationales, grandes entreprises et agences de traduction, etc. Puisque le volume de textes à traduire est de plus en plus élevé et les délais accordés sont très brefs, la traduction automatique s'est progressivement imposée afin d'alléger le travail du traducteur. Toutefois, même si cette dernière s'est nettement améliorée récemment, notamment grâce aux réseaux neuronaux, elle n'est pas en mesure de fournir les mêmes performances qu'un humain. Afin de garantir un niveau de qualité équivalent à la traduction humaine, une étape supplémentaire de correction effectuée par un humain est nécessaire : la post-édition.

Dans ce chapitre, nous présenterons une courte introduction relative à la traduction automatique neuronale, nécessaire pour comprendre l'utilité et le fonctionnement de la post-édition.

## 4.1 Traduction automatique

Selon Bouillon, « *[l]a traduction automatique* (TA) se définit comme l'application de l'informatique à la traduction des textes d'une langue naturelle de départ (ou langue source LS) dans une langue d'arrivée (ou langue cible LC). » (1993, p. 15). En se basant sur cette définition, il est possible de regrouper différents types et différents degrés d'automatisation de la traduction tels que la traduction automatique assistée par l'homme (TAAH) ou la traduction entièrement automatique de haute qualité (TEAHQ). Pour mieux comprendre ce classement, Hutchins et Somers (1997) ont proposé une classification qui place les différentes « typologies » sur un axe (voir *Figure 1*) :

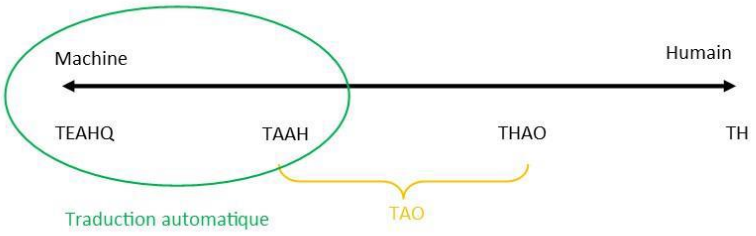



*Figure 1: Classification des typologies de traduction de (Hutchins & Somers, 1997; adapté par Volkart, 2018, p. 7)*

L'extrémité gauche de l'axe indique une intervention importante de la machine ; l'extrémité droite signale une intervention importante de l'humain. La traduction automatique comprend,

donc, tous les types de traduction compris entre la TEAHQ et la TAAH, tandis que la TAO (traduction assistée par ordinateur) regroupe les systèmes dont l'automatisation « s'applique à des parties du processus ou à une tâche précise qui lui est associée » (L'Homme, 2008, p. 11), soit les systèmes compris entre la TAAH et la THAO (traduction humaine assistée par ordinateur). Il existe trois principaux types de TA : linguistique, statistique et neuronale.

Puisque l'expérience de notre travail était basée sur un texte issu de la traduction automatique neuronale (TAN) (traduit par DeepL), nous expliquerons le fonctionnement des réseaux neuronaux à la section suivante.

### 4.1.1 Réseaux neuronaux

L'explication que nous donnons ci-dessous des réseaux neuronaux pour la traduction nous servira également ultérieurement pour expliquer le fonctionnement des systèmes de reconnaissance automatique de la parole de type neuronal à la section 5.3.

D'après Forcada (2017), « *[n]eural machine translation [(NMT)] is corpus-based machine translation* ». En effet, les systèmes de TAN sont entraînés à l'aide de corpus conséquents, composés de paires de segments source – cible et de leurs traductions. De ce point de vue, les systèmes statistiques, fonctionnent de la même façon, mais ne recourent pas à des réseaux neuronaux.

Les réseaux neuronaux sont un outil computationnel pour le traitement du langage. On parle de cette technologie depuis longtemps : en effet elle a été décrite à partir des années 80 déjà par Waibel et al. (1988), bien qu'il n'y avait pas la puissance computationnelle nécessaire à cette époque, et ensuite, dès la fin des années 90, notamment dans l'article de « *McCulloch-Pitts neuron* » (McCulloch & Pitts, 1990). Dans leur article, McCulloch et Pitt décrivent un modèle simplifié de neurone humain comme élément computationnel qui peut être décrit en termes de logique propositionnelle. Quand on parle de réseaux neuronaux, on associe souvent ce nom à « apprentissage profond », en raison de la « profondeur » de ces réseaux : cette profondeur est représentée par les différentes couches (*layers*) sur lesquels se base le fonctionnement (voir *Figure 2*) (Jurafsky, 2009, Chap. 7). La structure des modèles neuronaux est plus simple que celle des modèles linguistiques :

« […] *There is no separate language model, translation model, and reordering model, but just a single sequence model that predicts one word at a time.* »[9].

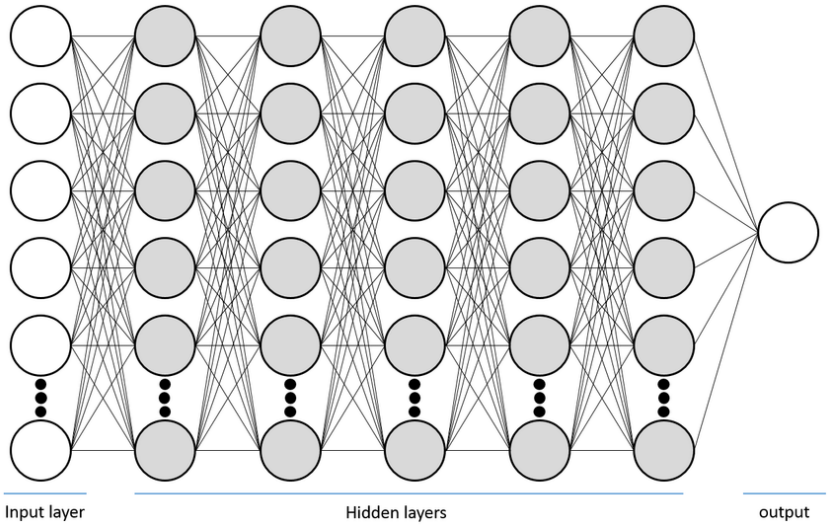



*Figure 2: Structure de réseaux neuronaux montrant les différentes couches (layers) (image tirée de (Kamal et al., 2019))*

Pour comprendre le fonctionnement de ces réseaux, il faut s'imaginer que les concepts (mots, phrases, etc.) sont placés sous forme de vecteurs dans l'espace. Les concepts similaires sont situés proches les uns des autres et ont des coordonnés analogues. Les concepts très différents sont plus éloignés et auront des coordonnées diverses. Il s'agit donc d'un système en trois dimensions. Or, trois dimensions ne sont pas suffisantes pour représenter la richesse de la langue, raison pour laquelle les codages des mots et les représentations des phrases nécessitent beaucoup plus de dimensions afin de pouvoir accueillir tous ces concepts ainsi que leurs relations mutuelles (Forcada, 2017).

L'élément de base d'un réseau neuronal est une simple unité computationnelle. Elle prend un ensemble de nombres réels comme entrée et, ensuite, les élabore en produisant une sortie. Contrairement à ce que on pense souvent, les réseaux neuronaux, qui devraient être appelés plus correctement « réseaux neuronaux artificiels » (*artificial neural network)*, n'ont pas grand-chose à voir avec le concept des neurones humains : l'appellation dérive des milliers d'unités artificielles sur lesquelles la traduction automatique neuronale se base. Ces éléments, selon les degrés de stimulation (excitation ou inhibition) réceptionnée, transmettent ces mêmes stimulations (Forcada, 2017).

Les systèmes de TAN fonctionnent en quatre étapes principales : l'entraînement (*training*), la « prédiction du mot successif », l'encodage (*encoding*) et le décodage (*decoding*).

[9] https://omniscien.com/state-neural-machine-translation-nmt/ consulté le 15/01/2020

La première phase d'entraînement fonctionne à priori comme l'entraînement des systèmes statistiques : il se base, effectivement, sur des corpus afin de déterminer le poids ou la force des connexions parmi les neurones. La différence avec les systèmes statistiques réside dans la puissance computationnelle requise : la TAN requiert souvent entre un jour et plusieurs mois de plus pour être entraînée. Durant cette phase, les poids sont modifiés afin de déterminer une valeur qui décrit la similarité des sorties de la traduction de référence pour qu'elle soit le plus réduite possible (Forcada, 2017). Ensuite, la particularité de ces systèmes est la capacité de prédire le mot suivant en se basant sur le contexte. De ce point de vue également, le fonctionnement se rapproche de celui de la TAS (traduction automatique statistique), le mot le plus probable pour chaque position sera sélectionné, en se basant sur des phrases déjà traduites. Puis vient l'étape de l'encodage de la phrase source : durant cette phase, un encodage préexistant pour une phrase vide est combiné avec le plongement du premier mot. Ensuite, ce processus est répété pour encoder chaque mot l'un après l'autre afin d'obtenir la phrase complète.

Afin d'illustrer notre propos, nous reprenons l'exemple donné par (Forcada, 2017) pour les langues anglais – espagnol. D'après lui, si l'on veut traduire la phrase anglaise « *My flight is delayed* » (« mon vol est retardé ») en espagnol, il faudrait penser à une représentation de la phrase comme des plongements de vecteurs pour chaque mot (« *the vector embeddings of individual words* »). Cela signifie que chaque mot correspond à un vecteur à visualiser comme un point dans l'espace. L'on parle de « plongement » car ces mots sont « plongés » dans l'espace. Il faut donc se représenter chaque mot séparément : e('my'), e('flight'), e('is'), e('delayed') et e('.'). À noter que « e(…) » a été utilisé comme encodage du vecteur qui peut avoir une centaine de composantes.(Ibidem).

Ensuite, l'encodage préexistant combine le premier espace vide E('') avec le plongement du premier mot e('My'), ce qui donne un résultat tel que E('My'). Par la suite, l'encodeur combine la représentation de E('My') avec le plongement de e('flight') afin de produire l'encodage

E(‘My flight’). Ce processus se répète jusqu’à obtenir la représentation de la phrase complète E(‘My flight is delayed’) (voir *Figure 3*).

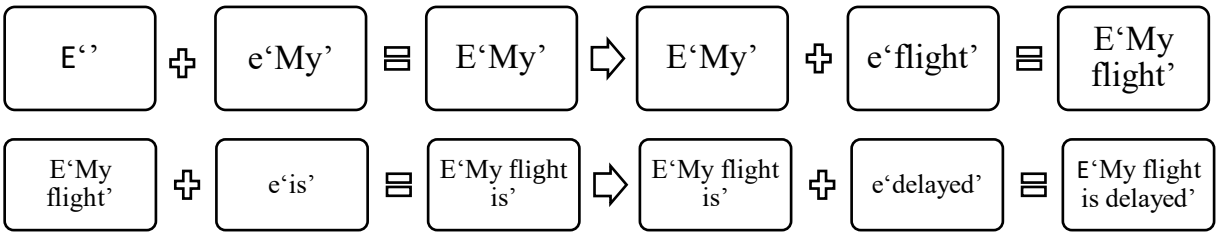



*Figure 3: Schéma de l’encodage*

Dans la phase de décodage, à partir de l’encodage de la phrase complète E(‘My flight is delayed’), le décodeur produit deux vecteurs : le premier correspond à la phrase source plus un espace vide pour la séquence des mots faisant partie de la phrase cible D(‘My flight is delayed’,‘’), le deuxième représente toutes les possibilités de tous les mots output possibles dans la première position de la phrase cible p(x|‘My flight is delayed’,‘’). Un décodeur bien entraîné sera en mesure de sélectionner le mot espagnol $x_{=}$‘Mi’, comme la traduction la plus probable pour ‘My’. Ensuite, le décodeur lit D(‘My flight is delayed’,‘’) ainsi que le mot ‘Mi’ et produit, à nouveau, deux vecteurs : le premier D(‘My flight is delayed’,‘Mi’) est un vecteur de probabilités de tous les outputs possibles avec le mot *x* à la deuxième place de la phrase p(x|‘My flight is delayed’,‘Mi’). Comme dans la phase précédente, un décodeur bien entraîné sera en mesure de sélectionner le mot espagnol ‘vuelo’ comme la traduction la plus probable de ‘vol’. Le fonctionnement se poursuivra pour compléter la phrase, jusqu’à l’obtention de la phrase cible complète ‘Mi vuelo lleva retraso’ (voir *Figure 4*) (Forcada, 2017).

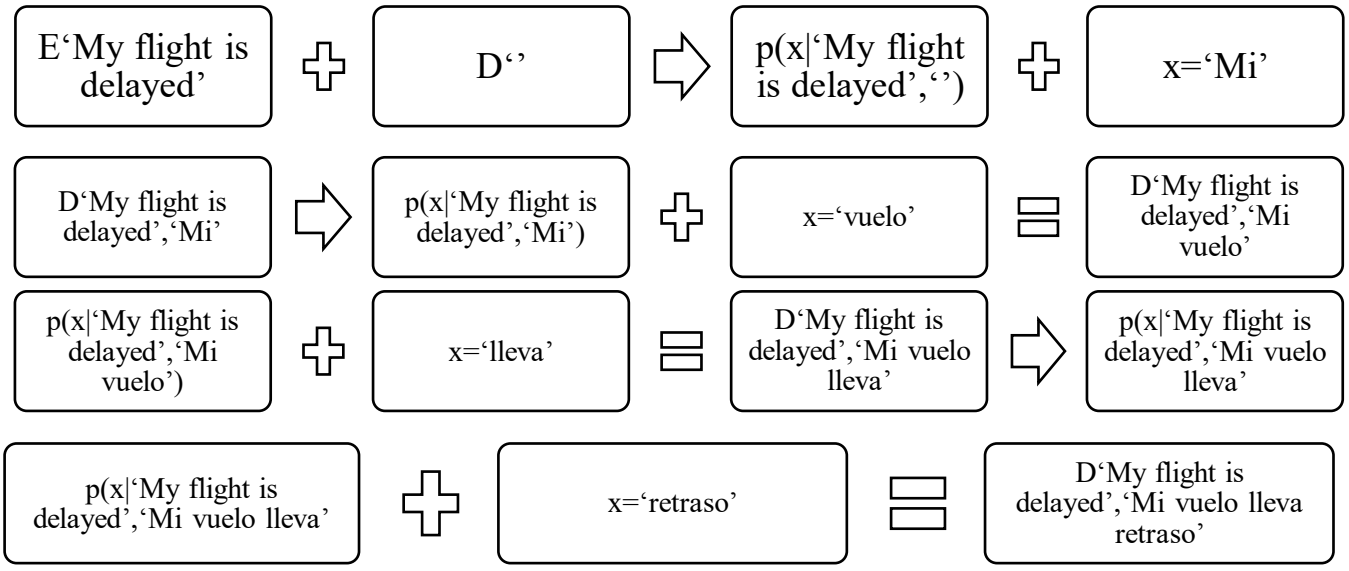



*Figure 4: Schéma du décodage*

Comme nous l'avons expliqué, nous nous sommes servie de cet exemple seulement afin de comprendre le fonctionnement des plongements dans le contexte des réseaux neuronaux, étant donné qu'il s'agit du même fonctionnement pour le modèle de langage neuronal.

## 4.2 Définition et typologies de la post-édition

Après cette introduction à la traduction automatique neuronale, passons maintenant à la post-édition. D'après L'Homme (2008) :

> « *La post-édition consiste à corriger la traduction [automatique] brute afin de la rendre acceptable pour diffusion.* » (L'Homme, 2008, p. 264)

Il s'agit donc d'une sorte de révision d'un texte issu d'une traduction automatique produite par une machine. La post-édition est devenue nécessaire à cause de la qualité relativement mauvaise de la traduction automatique à ses débuts. En effet, les textes traduits automatiquement, présentaient beaucoup d'erreurs qui ne permettaient pas l'utilisation de ces textes dans un cadre professionnel. Encore aujourd'hui, malgré les énormes progrès effectués dans ce domaine, les systèmes de TA commettent toujours de nombreuses erreurs, qui ne peuvent être corrigées que par un humain. Pour cette raison, si l'on veut publier un texte issu de la traduction automatique, il est nécessaire d'effectuer une tâche de post-édition. Il faut distinguer deux types de post-édition :

- **Post-édition minimale** (ou *light post-editing*) : les directives du TAUS[10] indiquent que le produit de ce type de post-édition doit être compréhensible, le contenu principal du message est saisi, et exact, soit que sa signification est identique à celle du texte source ; mais « sa forme stylistique est discutable »[11]. Le texte traduit peut donc sembler un peu artificiel, notamment en raison des éventuelles erreurs de syntaxe et de grammaire. Afin de mener bien sa tâche, le post-éditeur doit viser les objectifs principaux suivants :
    - S'assurer que toutes les informations sont transmises, sans qu'aucune d'entre elle ne soit ni oubliée ni ajoutée.
    - Exploiter au maximum le résultat brut de la traduction automatique.
    - Appliquer les règles d'orthographe fondamentales.
    - Ne pas effectuer de correction de style.

[10] https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines-french consulté le 27/01/2020
[11] Ibidem

- Ne pas modifier la structure du texte dans le seul but d'améliorer la fluidité du texte.

Ce type de post-édition ne convient pas à tous les types de besoins, mais plutôt à une demande urgente et à une circulation restreinte (Saint-André, 2015, p. 30). En outre, ce type de post-édition peut se révéler aussi utile pour ceux qui souhaiteraient simplement comprendre le message général d'un texte source dans une langue complètement inconnue (Screen, 2019).

- **Post-édition complète ou maximale** (ou *full post-editing*) : toujours selon les directives duTAUS[12], « ce niveau de qualité est en général défini comme compréhensible, […], exact […] et stylistiquement correct, même si le style n'est pas nécessairement aussi bon que celui obtenu par un traducteur humain dont la langue maternelle est la langue cible. »[13]. Les objectifs principaux des post-éditeurs sont les suivants :
    - Corriger la traduction du point de vue de la grammaire, de la syntaxe et de la sémantique.
    - Vérifier que la terminologie importante est correctement traduite, selon les directives et les consignes du client.
    - Vérifier que toutes les informations sont transmises, sans qu'aucune d'entre elle ne soit ni oubliée ni ajoutée.
    - Exploiter au maximum le résultat brut de la traduction automatique.
    - Appliquer les règles d'orthographe fondamentales, ainsi que de la ponctuation et de césure.
    - Vérifier que le formatage est correct.

Ce type de post-édition est recommandé pour tout texte officiel qui vise à être publié (Screen, 2019). La qualité de la post-édition complète doit, effectivement, être très élevée (plus élevée que celle d'une PE minimale) (Saint-André, 2015) et la traduction doit pouvoir être comparée à une traduction humaine.

Nous pouvons donc observer que les objectifs des deux types de post-édition sont relativement semblables, notamment en ce qui concerne le maintien maximal de la TA brute. Cet aspect est très important parce que la post-édition est bel et bien une tâche de « révision » d'un texte issu de la traduction d'une machine et non pas une « (re)traduction ». Cette idée s'applique

[12] https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines-french consulté le 27/01/2020
[13] Ibidem

également à la révision traditionnelle. En outre, la transmission du sens du texte source est aussi un élément central dans les deux cas : la seule différence réside dans le fait que pour la PE complète, toutes les règles de la langue cible (syntaxe, grammaire, ponctuation, orthographe, etc.) doivent être respectées. Le style, quant à lui, reste un élément à ne pas modifier dans la post-édition minimale, tandis que dans la PE complète, le TAUS ne donne aucune directive à ce sujet, car ce dernier sera probablement traité par les réviseurs afin d'obtenir une qualité comparable à celle d'une traduction humaine.

## 4.3 Un nouveau métier : le post-éditeur

Le post-éditeur est, à priori, comparable à la figure traditionnelle du réviseur, avec la seule différence qu'il travaille et « révise » des traductions issues de la traduction automatique. Si l'on reprend la définition de la PE de O'Brien :

> *« [p]ost-editing is the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria. »* (2011, p. 197)

Nous pouvons observer que l'idée de ce nouveau métier apparaît déjà dans cette définition, où l'on précise que la tâche doit être effectuée selon des directives spécifiques. À ce sujet, il est important de souligner la nécessité d'une formation dans le domaine de la post-édition, afin que cette dernière soit accomplie efficacement et qu'elle respecte les exigences du client.

Le post-éditeur doit avoir acquis différentes compétences. D'abord, comme le traducteur, il a besoin de compétences linguistiques, parce que la maîtrise des deux langues de travail est fondamentale. En outre, les erreurs commises par la traduction automatique pourraient concerner la terminologie ou la linguistique, raison pour laquelle des notions dans ces domaines sont souhaitables. Ensuite, des compétences informatiques sont également nécessaires, car désormais, l'utilisation de tout type de logiciel de traduction requiert des connaissances informatiques de base. De plus, les tâches de post-édition sont souvent effectuées dans les outils de TAO. Par ailleurs, des connaissances relatives aux systèmes de traduction automatique peuvent également se révéler utiles lors du repérage des éventuelles erreurs commises par la TA : si un post-éditeur connaît les erreurs plus fréquentes effectuées par ces systèmes, il lui sera plus facile de les repérer. Enfin, les post-éditeurs doivent également faire preuve d'une excellente culture générale, ainsi que d'une certaine responsabilité en ce qui concerne le respect des délais, souvent urgents, des mandataires. De plus, ceux qui démontrent une attitude positive envers l'utilisation de la TA semblent être plus performants que ceux forcés à l'utiliser (O'Brien, 2011).

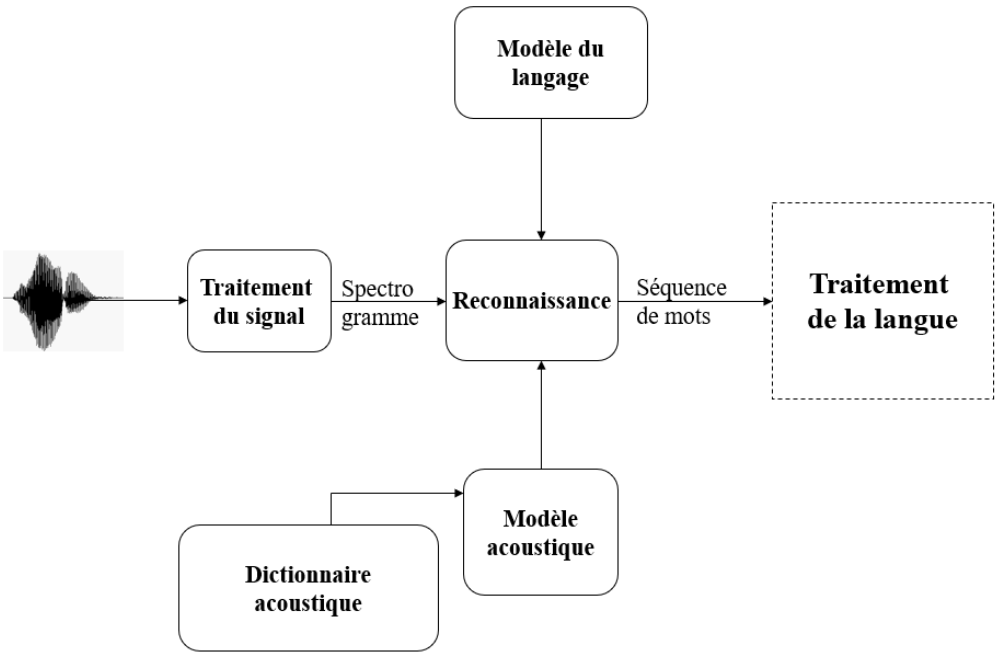
Toutefois, la post-édition peut ne pas donner le résultat souhaité. Deux scénarios sont possibles : le premier concerne le cas où le post-éditeur passe beaucoup de temps à effectuer les corrections (*over-correcting*) ; le deuxième concerne la situation inverse (*under-correcting*), soit de ne pas corriger suffisamment le texte, en laissant ainsi des erreurs graves dans l'output final (Somers, 2003, p. 305). Tous ces aspects soulignent l'importance du rôle du post-éditeur, ainsi que de ses compétences, dans le workflow d'une traduction issue de la traduction automatique.

# 5 Reconnaissance vocale

La reconnaissance vocale est définie comme une « [t]echnologie ayant pour but de permettre à un ordinateur de reconnaître les signaux émis par la voix humaine en vue de les transformer en données numériques » (Zapata & Quirion, 2016, p. 532). Au cours des dernières années, d'importants progrès ont été accomplis en reconnaissance vocale statistique, basée donc sur des probabilités : dans d'autres mots, sur la vraisemblance que la suite de sons prononcés forme tel énoncé plutôt que tel autre. Ces progrès ont vu le jour grâce aux percées en traitement automatique de la langue naturelle (TALN), couplées à la capacité d'analyse multipliée des ordinateurs (Ibidem).

## 5.1 Fonctionnement

Dans ce chapitre, nous expliquerons le fonctionnement relativement complexe ainsi que les différentes étapes de la reconnaissance vocale (voir *Figure 5*), qui peuvent être schématisés




dans la manière suivante.

*Figure 5: Différentes étapes et fonctionnement de la reconnaissance vocale (schéma tiré et adapté de (Rayner, 2006)*

La reconnaissance se fait donc de cette manière : d'abord, le système traite le son entrant, capté par le microphone et, à travers la première étape du traitement du signal, l'onde sonore est transformée en diagramme qui représente le son (spectrogramme) en 3D (Dan Jurafsky & Martin, 2009). Dans cette première étape, afin d'obtenir le spectre du signal, l'onde acoustique est découpée en tranches de 10 à 20 millisecondes. Ensuite, le système transforme le

spectrogramme en mots, grâce au modèle acoustique qui détermine les phonèmes composant le signal afin de reconnaître chaque son qui le compose (Ibidem). Enfin, le dictionnaire acoustique ainsi que le modèle du langage permettent de déterminer les séquences de mots avec la plus forte probabilité d'apparition en fonction du signal acoustique enregistré (Ibidem).

Il est important de souligner que ce processus est à la base des systèmes statistiques de reconnaissance automatique de la parole, basés donc sur des probabilités et des corpus, comme nous verrons par la suite (voir sections suivantes).

Nous pouvons donc observer que dans un premier temps, le signal sonore (onde sonore) est transformé en spectrogramme durant la phase de traitement du signal. Haton et al. (2006, p. 7) décrivent cette étape comme suit :

> « *La transformation de Fourier permet d'obtenir le spectre d'un signal, en particulier son spectre fréquentiel, c'est-à-dire sa représentation amplitude-fréquence. La parole étant un phénomène non stationnaire, il importe de faire intervenir le temps comme troisième variable dans la représentation. La juxtaposition des spectres obtenus pour des tranches successives permet d'approcher l'évolution du signal au cours du temps sous la forme d'un spectrogramme.* »

Tous les systèmes de reconnaissance vocale sont composés de trois éléments de base : un **modèle acoustique**, un **dictionnaire acoustique** et un **modèle de langage** (Bouillon, Cervini, Rayner, 2016). Parmi ces trois éléments, le modèle acoustique est la composante essentielle : il fait le lien entre le spectrogramme, donc les phonèmes, et la parole. Comme le dit Haton (2006, p. 71), il permet « le passage de l'onde sonore semi-continue à une suite discrète d'unités phonétiques ou lexicales ». Toutefois, notamment pour ce qui est de la langue française, un phonème peut se prononcer différemment selon celui qui le précède ou qui le suit. De ce fait, la transformation du spectrogramme en phonèmes demeure particulièrement complexe et les correspondances dépendront surtout du contexte (Bouillon, Cervini, Rayner, 2016). Cela étant, il est presque impossible d'écrire toutes les règles linguistiques qui se rapportent aux correspondances entre phonèmes et parole. Par conséquent, il est nécessaire que le système « apprenne » à partir d'autres textes : cette phase est communément appelée « entraînement du modèle acoustique » (Ibidem).

Le dictionnaire acoustique, quant à lui, établit la correspondance entre les séquences de phonèmes et les mots ayant pu être prononcés par l'utilisateur (Haton et al., 2006). Il peut être de deux types :

- Soit il contient une quantité limitée de mots, ce qui est le cas pour les logiciels avec un but spécifique ainsi que des applications limitées, dont l'objectif principal est de reconnaître des sons émis par différents locuteurs (comme les services vocaux par téléphone). Dans ce cas, le système n'a pas besoin d'être entraîné parce qu'il est adapté à capter différentes voix et intonations.
- Soit il contient une quantité importante de mots, comme dans le cas des systèmes monolocuteurs, nommés généralement « systèmes à grand vocabulaire ».

Généralement, la deuxième typologie de système est utilisée pour la dictée automatique, la traduction, ainsi que le sous-titrage (Haton et al., 2006). Dragon Naturally Speaking s'inscrit dans cette typologie qui, comme nous le verrons par la suite, s'adapte à la voix de l'utilisateur, améliorant ainsi les résultats de la reconnaissance vocale, notamment grâce à l'entraînement du vocabulaire. Malheureusement, le dictionnaire ne sera tout de même jamais complet à priori, étant donné qu'il ne contient que les mots qui y sont inscrits et qu'il ne reconnaît donc des mots spécifiques, tels que les néologismes, les noms propres, les abréviations, les termes techniques ou encore les mots d'origine étrangère. Malgré cet inconvénient, ce logiciel est capable, grâce à l'entraînement du système et à une spécialisation, de mieux comprendre les différentes prononciations, ainsi que d'augmenter la taille du dictionnaire afin d'ajouter des mots précédemment inconnus, tels que des termes ou des mots d'origine étrangère. En effet, plus une prononciation est fréquente, mieux elle sera reconnue. Une autre possibilité est d'ajouter des prononciations spécifiques à travers une liste (voir section 6) afin que ces dernières soient liées à des mots spécifiques, augmentant ainsi la précision de la reconnaissance vocale. Cet aspect sera approfondi à la section suivante.

Enfin, le modèle de langage sélectionne la séquence de mots qui a la plus grande probabilité d'apparition parmi toutes les possibilités suggérées par le modèle acoustique (Bouillon, Cervini, Rayner, 2016). Ce choix change en fonction des systèmes : pour les systèmes linguistiques, il sera basé sur les grammaires intégrées, tandis que pour les systèmes statistiques se basera sur les données statistiques. Il existe désormais trois types de modèles de langage : linguistique, statistique et neuronal. Le premier se base sur des grammaires CFG (*context free grammar* ou grammaire non contextuelle), qui décrivent les séquences correctes avec des règles (appelés aussi *rule-based* pour cette raison) ; le deuxième se base sur des séquences de N-grammes. Les modèles N-grammes sont donc des modèles statistiques (*probabilistic models*) qui prédisent les mots qui suivent (Dan Jurafsky & Martin, 2009). Si la prédiction s'effectue plutôt sur des plongements, on parle alors de modèle neuronal.

D'habitude, les trois éléments de base de la RV sont entraînés séparément. Cependant, puisque l'interaction entre eux s'avère compliquée, Araki et al. (2015) ont proposé un modèle unifié, comme le montre la *Figure 6*. Ils ont démontré, aussi, que ce type de modèle unifié permet d'obtenir une reconnaissance vocale très précise.

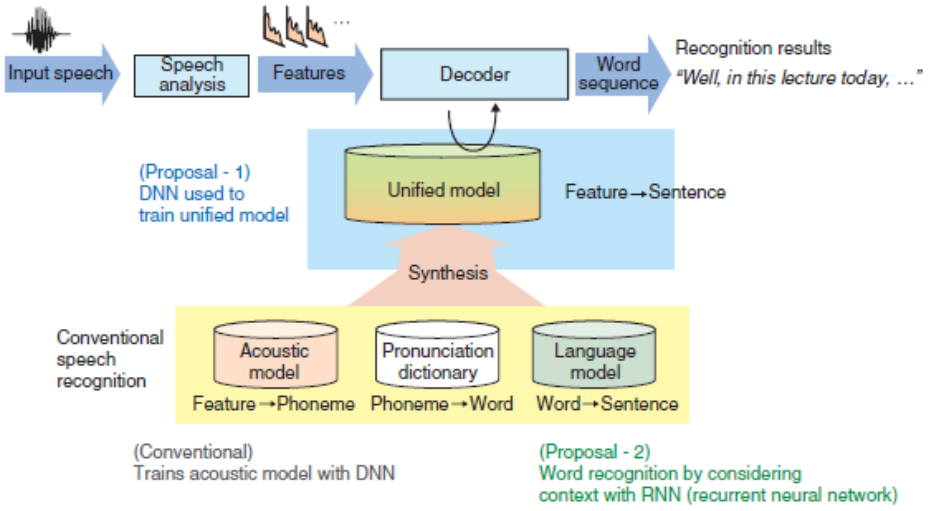



*Figure 6: Processus de la reconnaissance vocale dans les systèmes neuronaux ou à modèle unique (image tirée de Araki et al. 2015)*

En outre, ils ont montré qu'un réseau neuronal récurrent (*recurrent neural network*, RNN), qui se base aussi sur la technologie de l'apprentissage profond, contribue notablement à la performance du modèle de langage. D'après Araki (2015), un modèle de langage basé sur un RNN est très efficace pour la reconnaissance de la parole spontanée parce qu'il est en mesure de capter les mots et les reconnaître en considérant aussi un contexte plus vaste. La performance du modèle de langage RNN obtient les meilleurs résultats de performance par rapport à un modèle acoustique DNN (*deep neural networks*) conventionnel (Ibidem).

Comme le montre la *Figure 7* : le système transforme, en effet, le spectrogramme en caractères. On ne parle donc plus d'un véritable modèle de langage, car la transcription est produite caractère par caractère (Bouillon, 2019b). Plus récemment, ces réseaux neuronaux ont été proposés afin de remplacer le modèle Gaussien (*Gaussian mixture model*, GMM) comme base pour les modèles acoustiques pour les systèmes de reconnaissance vocale HMM (*Hidden Markov Model*, voir section 5.2). D'après Abdel-Hamid et al. (2012), il a été prouvé que les modèles acoustiques basés sur les réseaux neuronaux peuvent obtenir des performances beaucoup plus élevées, et cela grâce à leur fonctionnement « distribué » sur plusieurs couches (voir section 4.1.1).

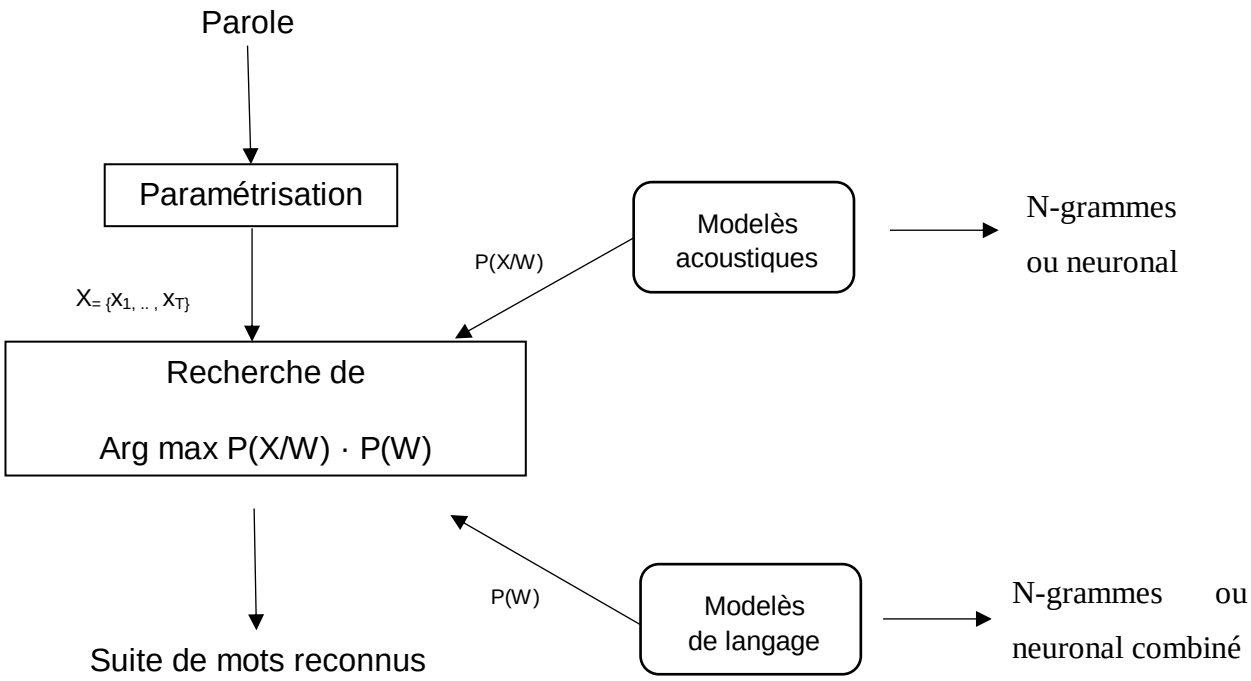



*Figure 7: Principe de la reconnaissance de forme bayésienne (image tirée de Bouillon, 2019, Haton et al. 2006, p. 11)*

Par la suite, nous nous concentrerons en particulier sur le modèle neuronal, étant donné que le logiciel utilisé durant notre expérience fonctionne selon ce modèle (voir section 7). Nous avons intégré l'explication concernant les systèmes statistiques, à cause des similarités de fonctionnement avec les systèmes neuronaux. En revanche, nous n'aborderons pas les systèmes linguistiques, du fait qu'ils ne sont pas utilisés pour la langue générale et beaucoup moins efficaces.

Depuis la dernière version en 2016, Dragon est le premier outil à introduire l'apprentissage profond sur les données personnelles de chaque utilisateur afin d'améliorer le modèle de langage et acoustique. La technologie basée sur l'apprentissage profond, qui est au cœur du module de reconnaissance vocale, permet à DNS d'apprendre les accents et les particularités de la voix, ainsi que de s'adapter à l'acoustique des différents environnements[14]. Dragon utilise, effectivement, les réseaux neuronaux *end to end*, au niveau du modèle de langage, en capturant la fréquence des mots et la combinaison dans laquelle ils apparaissent, et au niveau du modèle acoustique, en déchiffrant les phonèmes d'une langue[15]. De plus, grâce à un apprentissage

[14] https://www.speechtechmag.com/Articles/Editorial/FYI/Nuance-Adds-Deep-Learning-to-Dragon-114607.aspx consulté le 15/01/2020

[15] Ibidem

constant des mots de l'utilisateur, les réseaux neuronaux permettent une augmentation de la précision jusqu'à 25 %[16].

## 5.2 Modèle de langage statistique

Les modèles de langage statistiques, comme le dit leur nom, se basent sur des connaissances statistiques, soit mathématiques, afin de déterminer les possibles phrases énoncées par l'utilisateur. Pour ce faire, ils utilisent des approximations stochastiques telles que les modèles n-grammes ou les modèles de Markov (Haton et al., 2006).

D'après Jurafsky et Martin (2009, p. 211), les modèles de Markov cachés (*Hidden Markov Model*, HMM) « *allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model*. » Ils sont la principale technique de modélisation statistique utilisée dans les systèmes de RV, bien que, maintenant, ils aient été remplacés par les modèles neuronaux, basés sur les réseaux neuronaux, (voir section 4.1.1). Le principe de base de ce modèle est la recherche de la séquence la plus probable de mots, sur la base de corpus oraux annotés. Ils calculent, donc, la probabilité de séquences de mots contenues dans la phrase, à partir du corpus de référence. Ces approximations permettent ainsi de calculer la probabilité qu'un mot suive un autre à partir du nombre *n* de mots : on parlera alors de modèles n-grammes. Les modèles N-grammes sont donc des modèles statistiques (*probabilistic models*) qui prédisent les mots qui suivent. Jurafsky et Martin définissent un modèle N-grammes comme suit :

> « *We formalize this idea of word prediction with probabilistic models called N-gram models, which predict the next word from the previous N – 1 words. An N-gram is an N-token sequence of words […]* » (2009, p. 117)

Le modèle unigramme calcule la probabilité d'un seul mot : par exemple dans le cas d'une phrase telle que « Le chat court », le modèle essaiera d'établir la probabilité de l'occurrence de « chat » ou « shah » (Bouillon, 2017a) ; le modèle bigramme, quant à lui, calcule la probabilité que deux mots se suivent (dans la même phrase, il s'agira de calculer la probabilité pour « chat court » et « chat cour », par exemple) (Ibidem) ; le modèle trigramme calculera, en revanche, la probabilité de trois mots qui se suivent. Le modèle quadrigramme est le plus utilisé par les systèmes de reconnaissance vocale, en raison des meilleurs résultats qu'une séquence de quatre

[16] https://techcrunch.com/2016/08/16/dragon-15/ consulté le 15/01/2020 et https://www.nuance.com/fr-fr/about-us/newsroom/press-releases/Nuance-annonce-la-nouvelle-version-de-Dragon-pour-Windows-qui-la-technologie-de-Deep-Learning-de-Nuance.html consulté le 28/01/2020

mots peut donner par rapport à un modèle bigramme (Ibidem). Jurafsky et Martin (2009), expliquent que l'on parle de bigramme (*bigram*) lorsqu'il s'agit d'une séquence de deux mots comme « *please turn* », et de trigramme (*trigram*) si la séquence de mots prévoit trois mots, comme « *please turn your* ». Ce type de modèles est aussi appelé modèle de langage (*language models*). Les n-grammes sont fondamentaux dans toute tâche dans laquelle il faut identifier des mots dans un discours. Dans la reconnaissance vocale, aussi, les *input speech sounds* sont très imprévisibles et, notamment dans certaines langues, la prononciation de beaucoup de mots se ressemble (Dan Jurafsky & Martin, 2009).

## 5.3 Modèle de langage neuronal

Le modèle de langage neuronal se base sur le fonctionnement des réseaux neuronaux expliqué à la section 4.1.1. Par la suite, nous nous pencherons sur le fonctionnement des réseaux neuronaux dans le cadre des systèmes de reconnaissance vocale avec des modèles de langage et acoustique neuronaux.

Bien que dans les systèmes neuronaux, la distinction entre modèle de langage et le modèle acoustique ne soit plus si nette qu'auparavant, nous utiliserons l'appellation « modèle de langage neuronal » pour nous référer à ce type de système.

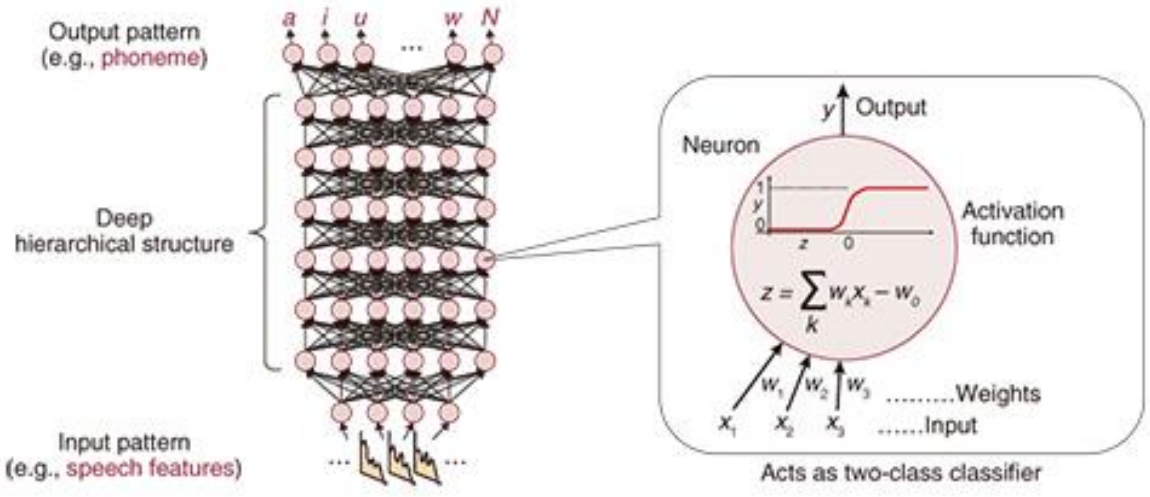



*Figure 8: Modèle acoustique avec réseaux neuronaux (image tirée de (Araki et al., 2015))*

Comme le montre la *Figure 8* ci-dessus, en premier lieu, les sons sont prétraités avant leur entrée dans le réseau neuronal. Les signaux de la parole sont captés par le microphone avec un taux de 16 000 mots à la seconde. Ensuite, une séquence d'éléments est créée et environ une vingtaine de fragments à chaque fois sont envoyés au modèle acoustique. En d'autres termes, le modèle acoustique est le réseau neuronal qui est entraîné et qui attribue les probabilités de chaque phonème. Après cela, le détecteur calcule le score, en évaluant la possibilité de

similarité que la phrase prononcée corresponde à la phrase que l'on cherche et si ce score est plus élevé que la valeur minimale acceptée, le logiciel sélectionne la phrase ou la commande souhaitée (Izbassarova et al., 2020).

Ces systèmes présentent plusieurs avantages par rapport aux systèmes statistiques : d'abord, les systèmes neuronaux n'ont pas besoin de simplifications et peuvent traiter aussi de longues phrases (au moins jusqu'à 30 mots). De plus, ils sont capables de généraliser sur des contextes similaires de mots, grâce à leur fonctionnement par contexte, expliqué auparavant (Forcada, 2017; Jurafsky, 2009, Chap. 7). Comme le montre la *Figure 9* ci-dessous, il faut supposer que ces systèmes augmentent la possibilité du mot suivant (nommé $c_k$) selon la similarité du contexte avec le mot précédent (nommé $w_j$), tandis qu'ils réduisent la probabilité d'une similarité avec un mot improbable comme dans le cas de « *aardvark* » (désigné comme $c_n$) (Bouillon, 2019b).

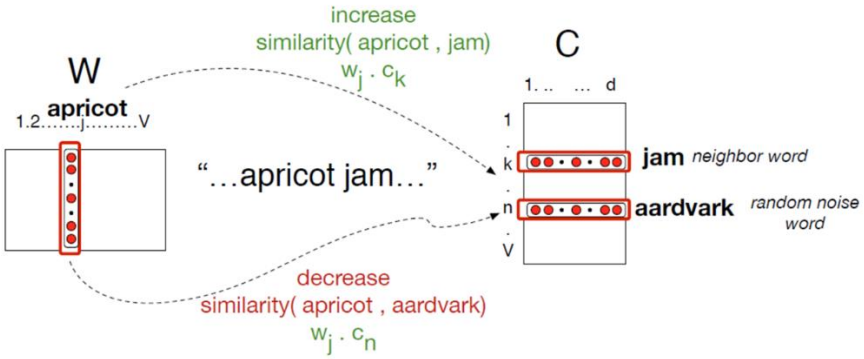



*Figure 9 : Fonctionnement de la prédiction des mots (image tirée de (Bouillon, 2019b))*

À la base, les modèles de langage neuronaux étant basés sur les réseaux neuronaux, ils fonctionnent de la même manière que ceux-ci. Cela dit, leur entraînement demeure, un travail long, bien que la manière d'opérer soit la même, soit se baser sur les probabilités d'un certain mot dans un certain contexte. Toutefois, le contexte de départ des modèles de langage neuronaux est représenté par les plongements (*embeddings*) : cela permet à ces modèles de pouvoir mieux généraliser les données par rapport aux modèles n-grammes. En effet, par exemple, dans une phrase telle que « *I have to make sure when I get home to feed the cat* », ces derniers sont capables de prédire le mot « *cat* », mais pas « *dog* ». En revanche, un modèle neuronal sera capable de prédire aussi le mot « *dog* », en se basant sur le fait que les deux mots

ont des plongements similaires ; il assignera donc la même probabilité pour « *cat* » que pour « *dog* », tout simplement car ils ont des vecteurs similaires (Jurafsky, 2009, Chap. 7).

Essentiellement, ce qui rend les systèmes de TAN meilleurs par rapport à ceux de TAS est l'apprentissage du sens depuis le corpus de traduction. En effet, si les deux systèmes apprennent depuis un corpus, les systèmes statistiques se limitent à assimiler des règles de traduction depuis un corpus aligné, tandis que les systèmes neuronaux vont au-delà, en se basant sur les différentes couches (*layers*) et sur les coordonnées des vecteurs-mots, qui se rapprochent selon le sens et le contexte. En outre, lorsque l'on dicte à un smartphone par exemple, on remarque que le système de RV de ces appareils a la tendance à fausser la proposition en devinant le mot suivant, mais, une fois à la fin de la phrase ou de la question, le système apprend de lui-même. Par ailleurs, les modèles neuronaux ne se limitent pas qu'aux n-grammes, soit aux quatre ou cinq mots qui précèdent, mais ils se basent sur la phrase en entière, et prennent en compte les concepts, ainsi que leurs relations, outre les informations syntactiques. Pour toutes ces raisons, bien que l'entraînement des deux systèmes soit similaire, l'output d'un système neural demeure plus fluide[17]. Le taux d'erreur est aussi notablement réduit, d'après Nassif et al. (2019). En 2012, après avoir rendu public la dernière version de son système de reconnaissance vocale basé sur l'apprentissage profond, Microsoft a montré que le WER était réduit d'environ 30 % par rapport aux modèles Gaussiens (Xie et al., 2018). Ce constat, prouve une nouvelle fois la meilleure qualité des systèmes neuronaux.

## 5.4 Dragon Naturally Speaking

Dans ce chapitre, nous décrirons le logiciel de reconnaissance vocale leader sur le marché, Dragon Naturally Speaking, dont nous nous sommes servie tout au long de notre expérience.

Dragon Naturally Speaking (DNS) est apparu pour la première fois sur le marché en 1997, pour remplacer son prédécesseur Dragon Dictate, paru en 1990[18]. DNS a été le premier système de RV à permettre une dictée continue, au contraire de Dragon Dictate qui nécessitait une pause après chaque mot. Selon le vice-président senior et manager général de Dragon, avant cela la reconnaissance vocale, ainsi que les produits qui s'en servaient étaient très limités. Grâce à l'ingénierie pionnière dans ce domaine, Dragon a permis l'utilisation continue de la

[17] https://info.sdl.com/rs/689-LLU-525/images/NMT_for_Studio_Dec2019.mp4?mkt_tok=eyJpIjoiTkdNM09HUmpZekkxWmpSayIsInQiOiJHYzFjc1dJMDNlRWdjRWlNZHdCdW5EOHBsZmNtWDM3N2dTZjdsS0l0Sm1IQnhUNG1YUjh0Skkzd0JIWlY2ZGJGN3l4SFFla1lMV1hyOWpMSThqTnRwdUxNbnpuK3dVc1F2VmJxZm5haURzTWdBb2RwbmNIVVJRZHJTVEMyYlhPSiJ9 vidéo visionnée le 23/01/2020

[18] https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen consulté le 28/01/2020

reconnaissance vocale pour la création d'un document (soit sans effectuer de pause après chaque mot). En effet, ce logiciel permet de reconnaître environ 100 mots à la minute[19]. À ce stade, nous pourrions donc conclure que la reconnaissance vocale est très proche de l'être humain. La seule différence majeure réside dans ce que l'on appelle l'« effet cocktail party » : ce dernier se réfère aux difficultés rencontrées par un logiciel de RV, tel que Dragon, à capter les sons et à les reconnaître dans un environnement bruyant. Afin que le système fonctionne bien, il est important de parler très proche de microphone de manière claire et d'être dans un espace au calme[20].

Dragon Naturally Speaking est le logiciel leader sur le marché, non seulement grâce à son fonctionnement irréprochable, mais aussi à ses nombreuses fonctionnalités, notamment :

- Dictée vocale ;
- Rédaction et modification plus rapide que la vitesse de mots tapés en moyenne (environ le double) (Bouillon, 2017a) ;
- Possibilité d'utiliser les *voices tags*, afin de contrôler Windows sans utiliser le clavier et la souris pour donner des commandes vocales telles qu'ouvrir des fenêtres, des documents, envoyer des mails, naviguer sur Internet, etc. ;
- Compatibilité avec la plupart des logiciels (Word, Excel, Outlook, Trados, Google Chrome, etc.).

De plus, la création d'un ou plusieurs profils utilisateurs permet à Dragon de spécialiser la reconnaissance vocale à la voix de chaque utilisateur, apprenant ainsi des éventuels accents, des différentes prononciations, ou encore des noms propres spécifiques.

Parmi les différents avantages de Dragon, la personnalisation du dictionnaire demeure l'un des plus important. Dans l'éditeur du vocabulaire, il est effectivement possible d'ajouter ou de supprimer des mots. Il est possible d'ajouter des mots tels que des noms propres, des termes spécifiques à un domaine, mais aussi de supprimer des mots jamais utilisés et qui pourraient engendrer une confusion durant la dictée. Lors de la dictée, l'utilisateur peut également ajouter des mots dans l'éditeur. À noter que si l'on veut ajouter une liste importante de mots, il est possible d'importer une liste de mots : Dragon permet, pour chaque mot ou expression de préciser la prononciation afin de faciliter la reconnaissance. Cette « préférence » peut être très

[19] https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen consulté le 28/01/2020
[20] Ibidem

différente du mot, car Dragon sera en mesure d'associer telle prononciation à tel mot (Bouillon, 2017a).

En outre, Nuance a annoncé la dernière version de Dragon en février 2017, basée sur la technologie d'apprentissage profond. Dès lors la productivité a progressée, même si elle était déjà importante[21]. En effet, cette technologie permet une précision plus élevée, ainsi qu'une rapidité et une personnalisation de haut niveau. Nuance affirme que grâce à cette technologie « Dragon s'ouvre à toujours plus d'utilisateurs par sa capacité à apprendre avec une précision accrue les accents et les habitudes de langage des individus, et à s'adapter savamment aux conditions acoustiques des environnements ouverts ou mobiles. »[22]. De plus, cette version est la première qui permet l'intégration de l'apprentissage profond au système personnel de l'utilisateur, afin que ce système puisse optimiser les capacités acoustiques et linguistiques à partir des données vocales[23]. D'après Vlad Sejnoha, directeur technique de Nuance Communications, l'entraînement des modèles neuronaux nécessite de gros volumes de données ainsi qu'un environnement de calcul de haute performance. Toutes ces fonctionnalités permettent un gain de précision jusqu'à 24 %[24].

[21]https://www.nuance.com/fr-fr/about-us/newsroom/press-releases/Nuance-annonce-la-nouvelle-version-de-Dragon-pour-Windows-qui-la-technologie-de-Deep-Learning-de-Nuance.html consulté le 28/01/2020
[22]Ibidem
[23]Ibidem
[24]https://www.nuance.com/fr-fr/about-us/newsroom/press-releases/Nuance-annonce-la-nouvelle-version-de-Dragon-pour-Windows-qui-la-technologie-de-Deep-Learning-de-Nuance.html consulté le 28/01/2020

# 6 Méthodologie

Comme annoncé à la section 3, nous nous sommes appuyée en particulier sur l'étude conduite par Mesa-Lao (2014). Par la suite, nous procéderons à la description détaillée de l'évaluation en expliquant les caractéristiques choisies pour l'évaluation, la méthodologie adoptée et nous présenterons le profil des participants.

Cette étude vise à observer si l'utilisation de la reconnaissance vocale a un impact sur la productivité d'un traducteur italophone lors d'une tâche de post-édition. Puisque la reconnaissance vocale est une technologie de plus en plus fréquemment utilisée, nous nous intéressons depuis plusieurs années à cette nouvelle méthode d'interaction, notamment en lien avec la traduction. Tout au long de notre parcours académique, nous avons eu l'occasion d'aborder ce sujet, qui a tout de suite suscité notre intérêt. C'est pour cette raison que nous avons choisi d'étudier plus en détails le fonctionnement de ce moyen d'interaction et son impact sur le travail du traducteur.

Pour que notre évaluation soit complète et que le score soit objectif, nous avons d'abord défini un modèle de qualité, présenté à la section 6.1, et ensuite le modèle des tâches, que nous verrons plus en détails aux sections 6.2.1 et 6.2.2.

## 6.1 Méthodes EAGLES et ISO

Notre évaluation s'appuie sur certaines caractéristiques des normes ISO, en particulier sur les normes 25000 (*ISO/IEC 25010:2011(en), Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models*, s. d.), avec une attention particulière aux normes des divisions ISO/IEC 25 010 :2011 2501n et 2502n, qui correspondent à la section du modèle de qualité et la section sur la mesure

de la qualité, comme le montre la *Figure 10*. Ces deux sections présentent des modèles détaillés pour évaluer la qualité à l'usage.

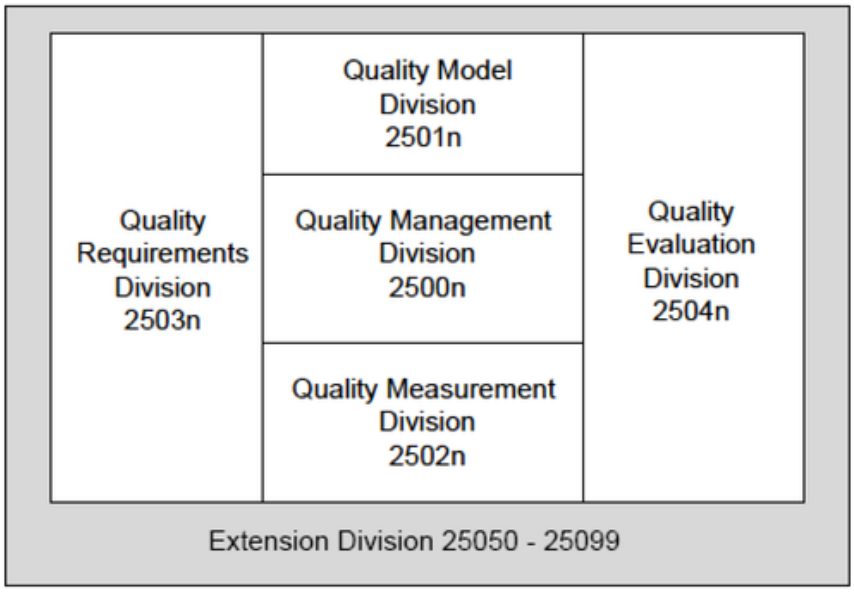



*Figure 10: Organisation SQuaRE series of International Standards*[25]

Selon la série des normes *series of Internation Standards (SQaRe)*, la qualité d'un système est le niveau auquel ledit système satisfait les besoins (implicites et explicites) des utilisateurs. Ces besoins sont représentés dans *SQaRE* à travers différents modèles qui décomposent la qualité du produit en différentes caractéristiques et sous-caractéristiques. Ces normes permettent d'étudier de manière systématique si le logiciel analysé répond aux besoins de l'utilisateur dans un contexte donné et d'établir un modèle de qualité afin de trouver des moyens efficaces pour mesurer la qualité du logiciel (Starlander, 2016).

Au-delà de la définition du modèle de qualité, le groupe de travail européen EAGLES[26] propose un guide pour une évaluation à l'aide d'une méthode en sept étapes :

| N° | Description étape EAGLES |
|---|---|
| 1 | **Définition du but de l'évaluation :**<br>Quel est l'objectif de l'évaluation ? Est-ce le système en entier ou une composante du système qui est évalué ? |
| 2 | **Définition du modèle des tâches :**<br>Identifier qui va utiliser le système et pour quelle utilisation.<br>Quel est le profil de l'utilisateur ? |
| 3 | **Définition des caractéristiques de haute qualité :**<br>Identifier les caractéristiques à évaluer dans le contexte d'utilisation donné. |
| 4 | **Définition du modèle de qualité hiérarchique :**<br>Choisir les caractéristiques et sous-caractéristiques. Le modèle de qualité doit aboutir à une subdivision finale appelée attribut lequel pourra être quantifié par une mesure. |

[25] Image tirée de de https://www.iso.org/obp/ui/es/#iso:std:iso-iec:25041:ed-1:v1:en:fig:1
[26] Site du projet EAGLES, https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html consulté le 14/10/19

| | |
|---|---|
| 5 | **Spécification des mesures à appliquer :**<br>Comment mesurer les attributs identifiés à l'étape précédente ? Définition des mesures et de l'interprétation des échelles. |
| 6 | **Préparation de l'évaluation :**<br>Mise en place de l'évaluation. Définition du protocole d'évaluation. |
| 7 | **Exécution de l'évaluation et rédaction du rapport.** |

*Tableau 1: Principales étapes de l'évaluation selon EAGLES*[27]

Cette méthode propose une marche à suivre à l'aide d'exemples pratiques. Les deux étapes principales sont la création d'un modèle des tâches (étape 2) et d'un modèle de qualité (étapes 3 et 4). Le modèle des tâches consiste à définir les tâches à évaluer à l'aide des caractéristiques présentées dans le modèle de qualité. Le but est d'effectuer une évaluation sur la base des caractéristiques de qualité importantes dans un contexte d'utilisation donné. Par la suite, nous verrons en détail les tâches spécifiques de notre évaluation.

Dans notre modèle des tâches, nous avons défini deux objectifs principaux avec deux modalités différentes, qui incluent les différentes tâches à compléter. Chacune est à compléter à la fois avec la reconnaissance vocale et avec les méthodes traditionnelles d'entrée (clavier et souris). Les deux objectifs principaux sont :

1) Post-éditer les deux textes
    - Avec la reconnaissance vocale
    - Avec les méthodes traditionnelles d'entrée
2) Fournir une traduction finale pouvant être considérée comme de bon niveau
    - À l'aide de la RV
    - À l'aide des méthodes traditionnelles d'entrée

Pour que les deux objectifs soient remplis, il faut effectuer correctement une post-édition ainsi que fournir une traduction de haute qualité.

Au sein de ces deux objectifs principaux, nous avons défini cinq tâches à effectuer et à compléter afin d'obtenir le résultat prévu.

a) Dicter correctement les phrases (tous les mots sont bien compris par le micro et écrits correctement)
b) Exécuter correctement les commandes vocales (la commande souhaitée est exécutée correctement)

[27] Ibidem

c) Corriger correctement à l'aide de la reconnaissance vocale les segments, en particulier sans rajouter d'erreurs
d) Corriger de manière autonome d'éventuelles erreurs de dictée ou de commandes vocales
e) Corriger correctement les segments à l'aide des méthodes traditionnelles d'entrée, en particulier sans rajouter d'erreurs

Il faut préciser ce que nous entendons par « dictée » et « commande vocale ». En effet, dans le cas de notre travail, nous parlons de « dictée » lorsque nous nous référons à la dictée d'une phrase ou d'au moins un mot ; tandis que « commande vocale » se réfère à une commande comme « *seleziona* » (sélectionner*)* ou « *correggi* » (corriger), etc. Il est important de faire cette distinction parce qu'il est possible que le logiciel comprenne bien la phrase, donc les mots prononcés par l'interlocuteur, mais pas forcément la commande vocale et donc qu'il les écrive au mauvais endroit, par exemple.

Le modèle de qualité permet, quant à lui, de décomposer la qualité d'un logiciel en caractéristiques (étape 3 EAGLES) et sous-caractéristiques (étape 4). Ainsi, il est possible d'avoir des moyens efficaces et objectifs de mesurer la qualité du logiciel. Notre attention s'est portée particulièrement sur le modèle de qualité à l'usage, qui correspond aux normes ISO/IEC 25010 :2011 et ISO/IEC 25022 :2016.

La qualité à l'usage d'un système définit l'impact que le produit a sur ses usagers. Elle est déterminée par la qualité du software, du hardware et du système d'exploitation, ainsi que les caractéristiques à l'usage (*ISO/IEC 25010:2011(en), Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models*, s. d.).

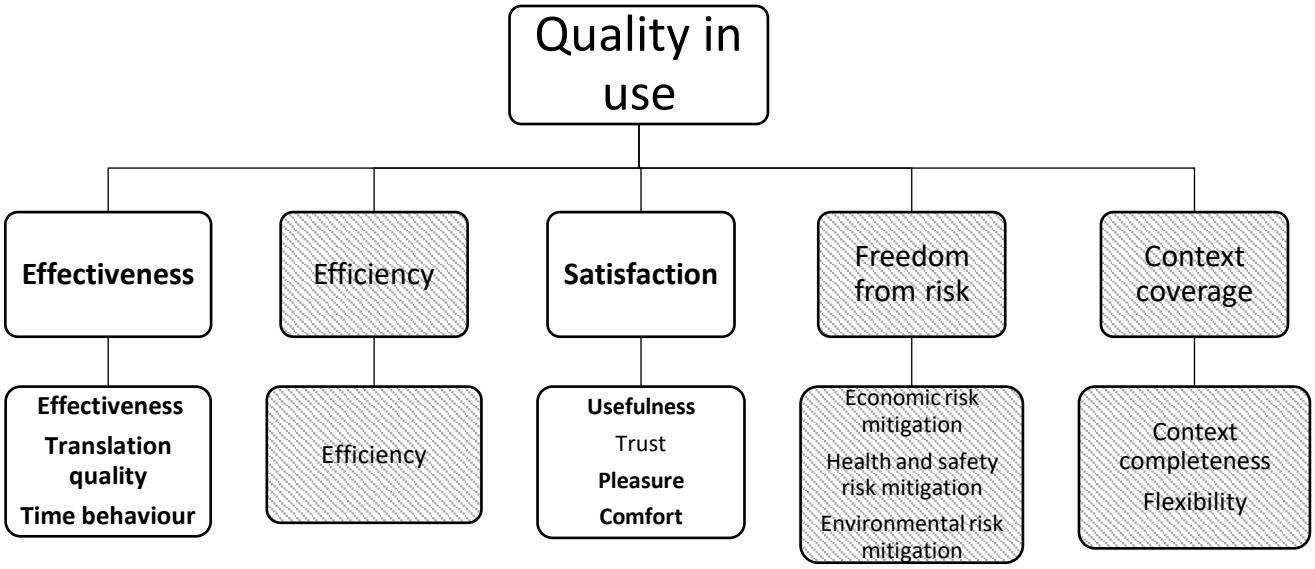



*Figure 11 : Modèle de qualité à l'usage*

La qualité à l'usage se mesure selon cinq caractéristiques : efficacité, efficience, satisfaction, absence de risque et conformité au contexte d'utilisation. Nous nous intéressons à deux d'entre elles : **l'efficacité** (*effectiveness*) et **la satisfaction** (*satisfaction*), car ce sont celles qui peuvent être considérées comme les plus pertinentes dans le cadre de notre recherche. En effet, l'efficacité, conjuguée à la satisfaction, permet d'obtenir une vision globale et complète en ce qui concerne l'utilisation de cette technologie en plein essor. Au sein de ces deux caractéristiques, nous avons défini des sous-caractéristiques, comme le montre la *Figure 11* ci-dessus. En effet, la décomposition et la sélection des sous-caractéristiques permettent une analyse complète de chaque caractéristique : pour l'efficacité, l'étude du temps employé et de la qualité de la traduction, en plus de l'efficacité elle-même, rend l'analyse certainement meilleure avec une perspective à plusieurs niveaux. Ainsi, la satisfaction, concerne plusieurs sous-caractéristiques, telles que l'utilité, le plaisir et le confort, soit des éléments qui examinés séparément ne pourraient pas donner une idée précise et complète de la satisfaction ressentie par les utilisateurs durant toute l'expérience. Dans les sections suivantes, nous verrons en détails ces caractéristiques et leurs sous-caractéristiques.

## 6.2 Efficacité

L'efficacité est définie, d'après la norme ISO/IEC 25010 :2011, comme « [the] *accuracy and completness with which users achieve specified goals* » (« la précision et l'exhaustivité avec laquelle l'utilisateur atteint les objectifs fixés »).

Pour évaluer cette caractéristique, au-delà des mesures proposées par la norme, telles que le nombre des tâches et d'objectifs complétés, le calcul du nombre d'erreurs par tâche et des tâches qui ont présenté des erreurs, nous avons établi des sous-caractéristiques, notamment par rapport à la reconnaissance vocale, qui peuvent répondre aux questions suivantes :

- Le logiciel écrit-il correctement tous les mots dictés ?
- Tous les mots sont-ils compris par le logiciel ?
- Le logiciel exécute-t-il correctement les commandes vocales ?
- Le logiciel reconnaît-il les termes du domaine (issus de la spécialisation) ?
- La correction est-elle effectuée de manière simple et rapide ?

Les réponses à ces questions nous permettront d'évaluer certains attributs et d'obtenir, par conséquent, un score objectif afin de déterminer si la reconnaissance vocale peut s'avérer efficace.

Par la suite, nous expliquerons en détails chaque point à évaluer, en illustrant les tâches que nous avons prises en considération ainsi que les points que nous avons attribué pour le calcul final.

### 6.2.1 Tâches complétées

Selon la norme ISO, un autre aspect à prendre en considération pour calculer l'efficacité est le nombre de tâches complétées. Il s'agit de la proportion des tâches complétées correctement sans aide. Pour le calculer, nous nous sommes basée sur les tâches présentées à la section 6.1. Pour chacune d'elles, nous avons attribué les points de la façon suivante :

- Dix points pour tâche complétée totalement ;
- Cinq points si la tâche est partiellement complétée, par exemple si le logiciel produit une phrase, mais certains mots sont incorrects ;
- Zéro point si le participant a manifesté des difficultés à plusieurs reprises et n'a pas achevé la tâche.

#### *6.2.1.1 Objectifs complétés*

Il s'agit de la proportion des objectifs complétés sans aide. Notre modèle de tâche comporte deux objectifs principaux, énumérés à la section 6.1. Nous avons donc attribué un point pour chaque objectif principal complété et zéro point si l'objectif n'a pas été atteint. Étant donné qu'ils incluent les tâches à compléter, les points sont cohérents avec ceux relatifs aux tâches. Nous avons décidé de ne pas insérer le demi-point dans notre échelle parce qu'il est important

pour nous que l'objectif soit complétement atteint. En outre, il est difficile de définir, le cas échéant, s'il n'est complété que partiellement.

#### *6.2.1.2 Erreurs dans une tâche*

Il s'agit de calculer le nombre d'erreurs effectuées pendant chaque tâche, sachant que ces erreurs peuvent être liées au nombre d'actions à mener pour effectuer chaque tâche. En effet, plus il y a d'actions, plus le risque d'erreurs est élevé. Nous avons classé les erreurs en cinq types :

a. Erreur mineure de commande vocale : l'utilisateur se trompe avec une commande, mais il la corrige ou trouve une solution acceptable en peu de tentatives
b. Erreur de terminologie : l'utilisateur ne corrige pas certains termes relatifs du domaine
c. Erreur de style : l'utilisateur n'améliore pas le style, issu de la traduction automatique, en laissant une langue cible peu naturelle
d. Erreur majeure d'usage de la RV : l'utilisateur ne réussit pas à saisir le texte souhaité à l'aide de la reconnaissance vocale uniquement et est obligé d'utiliser la souris ou le clavier
e. Erreur d'orthographe : la dictée ne donne pas de bons résultats (mots écrits incorrectement) et l'utilisateur ne les corrige pas ou ajoute lui-même des fautes en dictant

Nous avons attribué un demi-point pour les trois premiers types (*a, b, c)*, un point pour le type *d* et deux points pour le type *e*. Chaque tâche vaut 10 points, nous arrivons donc à un total de 50 points, auxquels nous avons soustrait le(s) point(s) d'erreur. Cette répartition des points se justifie par le fait que les erreurs *a, b, c* sont à considérer comme moins graves que les *d* et *e* qui, au contraire, influencent beaucoup l'efficacité du système de RV.

#### *6.2.1.3 Tâches avec erreurs*

Il s'agit de la proportion de tâches au cours desquelles des erreurs ont été commises par l'utilisateur. Selon la norme ISO/IEC 25 022 :2016, cette proportion X correspond à A/B, soit le nombre de tâches au cours desquelles des erreurs ont été faites (A), sur le nombre total de tâches effectuées (B). Nous obtenons ainsi un score entre A et B. Puisque nous avons un total de cinq tâches, B est égal à 5 pour tous les participants et A correspond au nombre de tâches qui ont présenté une ou plusieurs erreurs. Nous avons considéré comme suffisant un score

supérieur ou égal à 3/5, ce qui signifie qu'il est acceptable d'avoir un maximum de deux tâches qui comprennent au moins une erreur.

### 6.2.2 Qualité de la traduction

Pour pouvoir effectuer une évaluation complète et objective de la qualité finale de la traduction, nous avons utilisé du côté quantitatif les scores BLEU (*Bilingual Evaluation Understudy*) et TER (*Translation Edit Rate*). L'évaluation automatique sert comme point de vue pratique pour voir l'interaction entre le système de TA et l'usager. De plus, une évaluation automatique est plus rapide et moins coûteuse par rapport à l'évaluation humaine : elle requiert moins de temps ainsi que moins de ressources (Specia et al., 2010). D'après Snover et al. (2006), la qualité de ce type de mesures automatiques peut uniquement être déterminée en comparaison aux évaluations humaines. Raison pour laquelle nous avons intégré une évaluation de type humain, et cela du point de vue qualitatif : un humain peut reconnaître facilement une traduction mal faite, en juger la fluidité et évaluer la qualité en se basant notamment sur le texte source (Specia et al., 2010).

Nous expliquerons les deux scores que nous avons utilisés aux sections suivantes, ainsi que la manière dont nous nous sommes servie de l'évaluation humaine.

#### *6.2.2.1 Bilingual Evaluation Understudy (BLEU)*

D'après Snover et al. (2006, p. 224)) « *BLEU calculates the score of a translation by measuring the number of n-grams, of varying length, of the system output that occur within the set of references.* ».

D'après KIT & WONG (2014) la logique derrière le calcul du score BLEU est « *the closer a machine translation is to a professional human translation, the better it is*». Cette mesure calcule le nombre de n-grammes (qui correspondent à des séquences de *n* mots) présents dans la cible et dans une ou plusieurs traductions de référence (Papineni et al., 2002). Logiquement, plus une phrase sera brève, plus le score obtenu sera élevé, ce qui ne permet pas d'obtenir un score objectif. Afin de ne pas biaiser l'évaluation, les chercheurs ont introduit une pénalité par la brièveté (*brevity penalty*), qui va réduire le score dans le cas où la traduction cible est trop courte (Koehn, 2010). Cette méthode présente un double aspect innovant ; si elle permet de prendre en compte plusieurs références en même temps, elle prend également en compte l'ordre des mots, ce qui assure une meilleure objectivité dans le calcul de la précision. Étant donné qu'il s'agit d'une méthode automatique, elle reste fidèle aux évaluations humaines, vu qu'elle mesure la fluidité et la fidélité au lieu de se baser sur la métrique des mots (Bouillon, 2017b).

Pour le calculer, nous avons utilisé l'outil en ligne https://www.letsmt.eu/Bleu.aspx, qui appartient à l'entreprise Tilde, leader européen des technologies langagières. Cette dernière offre différents types de service, allant de la traduction automatique à la localisation en passant par la terminologie.

#### *6.2.2.2 Translation Edit Rate (TER)*

D'après Snover et al. (2006, p. 223)), « *Translation Edit Rate (TER) measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation* ». Concrètement, TER se calcule comme suit (Ibidem) :

$$\text{TER} = \frac{\#\ d'éditions}{\#\ moyenne\ de\ mots\ dans\ la\ référence}$$

TER est fondé sur la quantification de la distance d'édition entre deux séquences de mots, ce qui correspond au nombre minimal d'opérations nécessaires pour transformer une séquence de mots en une autre (KIT & WONG, 2014). Parmi ces opérations d'édition sont incluses les insertions, les suppressions et les substitutions d'un seul mot, ainsi que le déplacement des séquences des mots (Snover et al., 2006). Contrairement à BLEU, TER est moins sensible au nombre de références humaines ; toutefois, TER montre une bonne corrélation avec les mesures humaines parce qu'il s'agit d'un algorithme plus complet, raison pour laquelle est considérée une mesure automatique de deuxième génération (Starlander, 2016). La grande différence par rapport aux méthodes précédentes réside dans un changement de perspective : avant, la supposition à la base de toute mesure était qu'une traduction automatique contenait inévitablement des défauts ; par conséquent, l'objectif primaire était de réduire le taux d'erreur le plus possible afin de parvenir à une traduction impeccable. En effet, la question sous-jacente à la création des mesures était *How good is machine translation ?* (Koehn, 2010). Mais, d'après Koehn, au lieu de juger la qualité de la traduction produite en termes absolus, il faudrait plutôt l'évaluer en fonction de la finalité à du texte. La question qu'il faudrait se poser serait donc plutôt la suivante : *la traduction automatique remplit-elle le but auquel elle est destinée ?*

#### *6.2.2.3 Évaluation humaine*

D'après Kohen (2010), l'évaluation humaine se fonde sur les jugements exprimés par des êtres humains (pas forcément des traducteurs) sur la qualité d'une traduction, en vertu des connaissances et de la faculté de discernement des évaluateurs. Elle prévoit la présentation d'un certain nombre de phrases traduites par un logiciel de traduction automatique à un juge humain,

qui doit se prononcer sur la précision de la cible. Ce type d'évaluation peut être effectué de deux manières : avec ou sans phrase source. Dans le premier cas, on parlera d'évaluation bilingue, dans le deuxième, d'évaluation monolingue. Un juge bilingue, connaissant la langue source et la langue cible, est considéré comme le plus qualifié pour mener à bien cette tâche. Cependant, les évaluateurs bilingues sont parfois très difficiles à trouver, et sont très coûteux par rapport aux monolingues (Ibidem). Pour éviter ce problème, il est aussi possible que l'évaluation soit effectuée par des évaluateurs monolingues qui comprennent la langue cible et qui ont une traduction de référence à disposition.

Normalement, les évaluateurs jugent selon une échelle de points donnée (avec des valeurs de 1 à 5 généralement) afin de mesurer la qualité de chaque phrase. L'évaluation humaine est effectuée à l'aide de deux critères : l'adéquation (*adequacy* ou *fidelity*) et la fluidité (*fluency*). L'adéquation mesure le degré auquel la traduction transmet et conserve le sens de la phrase source, si une partie du message a été ajoutée, déformée ou n'a pas été transmise. Elle est effectuée en version bilingue, avec la phrase source et la cible. En revanche, l'évaluation de la fluidité peut être effectuée en mode monolingue, uniquement avec la cible. Elle mesure le degré de « fluidité » de la traduction, soit si elle semble « naturelle » ou idiomatique à un natif de cette langue.

Chaque valeur peut donc être associée à un jugement, comme expliqué dans les *Tableaux 2* et *3* suivants (Koehn, 2010) :

| **Adéquation** | |
|---|---|
| 5 | Totalité du sens |
| 4 | La plupart du sens |
| 3 | Beaucoup de sens |
| 2 | Très peu de sens |
| 1 | Aucun sens transmis |

*Tableau 2: Échelle d'évaluation - adéquation*

| **Fluidité** | |
|---|---|
| 5 | Phrase parfaite |
| 4 | Phrase de bon niveau |
| 3 | Phrase non-natif[28] |
| 2 | Phrase non fluide |
| 1 | Phrase incompréhensible |

*Tableau 3: Échelle d'évaluation - fluidité*

L'un des inconvénients de cette approche est le manque de précision de ces descriptifs. En effet, il est difficile (voire impossible) d'établir une échelle « globale » et parfaitement objective, capable d'exprimer exactement d'une façon universelle les signifiés associés aux scores. C'est

[28] Dans le cas où la phrase reflète la langue produite par un non-natif

pourquoi il est difficile pour les annotateurs d'assurer la cohérence dans leurs jugements. Une autre difficulté relève des annotateurs eux-mêmes. En effet, un nombre illimité de facteurs peut déterminer le jugement : chacun s'exprime selon sa perception et son degré de connaissance de la langue source ou cible. Ensuite, il ne faut pas négliger deux autres inconvénients, le temps et les coûts. L'évaluation humaine est en effet très coûteuse (Coughlin, 2003), ce qui va à l'encontre des intérêts des chercheurs, qui souhaitent obtenir des résultats dans les meilleurs délais, et parfois exécuter plusieurs évaluations par jour. En outre, les annotateurs doivent percevoir une récompense pour leur travail ; rémunération qui sera plus élevée pour les évaluateurs bilingues. Cette longue série d'inconvénients a encouragé les chercheurs à élaborer d'autres systèmes plus rentables pour estimer la qualité de la traduction d'un système de traduction automatique.

### 6.2.3 Temps

Le temps constitue également un élément important à prendre en compte. Il permet de donner un jugement complet et objectif, étant donné qu'il peut être enregistré et peut mesurer de facto l'activité des participants. En outre, dans le cadre de notre question de recherche, il nous a permis d'observer s'il y a véritablement un gain temporel, sans affecter la qualité du travail pour autant. Par le biais des enregistrements effectués à l'aide de *BB Flashback Express*, nous avons pu surveiller le temps employé par chaque participant dans les différentes tâches, et observer ainsi une corrélation entre certains facteurs tels que : la relation entre des connaissances préalables des logiciels et le temps employé à effectuer l'expérience, ou la connaissance préalable des directives de post-édition et le temps employé. Dans ces cas, il est normal d'observer une durée inférieure dans l'accomplissement de l'expérience, car certains participants étaient déjà familiers avec l'utilisation du logiciel ou le fonctionnement de la post-édition. D'autre part, les enregistrements permettent aussi d'observer le temps employé en détails, autrement dit le « temps de focalisation » sur une phrase en particulier ou sur une commande spécifique (Guzmán et al., 2015).

## 6.3 Satisfaction

La satisfaction est une des caractéristiques principales de la qualité à l'usage. Elle est divisée en quatre sous-caractéristiques : utilité (*usefulness*), confiance (*trust*), plaisir (*pleasure*) et confort (*comfort*). D'après la norme ISO 25022, la satisfaction mesure « [the] *degree to which user needs are satisfied when a product or system is used in a specified context of use* » (« La capacité du logiciel de satisfaire les besoins des utilisateurs dans un contexte d'utilisation

donné »). À noter que, dans le cadre de notre travail, nous n'avons tenu compte que de l'utilité, du plaisir et du confort. Contrairement à l'efficacité, qui se base sur une évaluation objective, la détermination de la satisfaction est réalisée au moyen d'un questionnaire de satisfaction soumis aux participants après l'accomplissement de l'expérience (Starlander, 2016). L'utilité a été évaluée selon une échelle composée de quatre valeurs auxquelles nous avons attribué des points selon le *Tableau 4* suivant :

| | |
|---|---|
| 0 | Pas du tout satisfait |
| 1 | Peu satisfait |
| 2 | Satisfait |
| 3 | Très satisfait |

*Tableau 4: Échelle de valeurs mesurant la satisfaction*

Le confort a également été évalué selon une échelle composée de quatre valeurs et nous avons attribué des points à chacune d'entre elles selon le *Tableau 5* suivant :

| | |
|---|---|
| 0 | Pas du tout confortable |
| 1 | Peu confortable |
| 2 | Confortable |
| 3 | Très confortable |

*Tableau 5: Échelle de valeurs mesurant le confort*

L'avantage de ces deux échelles est qu'il est impossible que les participants aient tendance à choisir les valeurs moyennes, grâce au nombre pair de réponses possibles.

Enfin, nous avons calculé la satisfaction sur la base des réponses au questionnaire et les évaluations attribuées selon ces échelles en faisant la somme totale. Les résultats seront présentés à la section 8.2.

# 7 Expérience

Les deux phases de post-édition se sont déroulées dans le logiciel utilisé pour l'expérience : *Microsoft Word*, avec l'intégration de *Dragon Naturally Speaking (Version Premium)* pour la reconnaissance vocale.

Nous avons soumis aux participants deux parties du même texte, à peu près de même longueur. Il s'agit du mode d'emploi d'une Smartwatch : un texte technique, mais adressé au grand public.

La terminologie n'est donc pas spécialisée, ce qui rend le contenu compréhensible pour un jeune traducteur peu habitué au domaine. Néanmoins, il a été nécessaire de spécialiser *Dragon Naturally Speaking*. Pour améliorer le modèle de langage, une phase de spécialisation a été effectuée en ajoutant un corpus (en format .pdf ou .docx) du domaine (Ciobanu, 2014). Effectivement, le modèle de langage représente le lien entre le son et la combinatoire des mots. Un modèle de langage statistique se base sur des probabilités de séquences des mots, ou n-grammes, extraites du corpus. Ce modèle consiste à prédire le mot suivant ; de cette façon, les textes ajoutés facilitent la reconnaissance des mots parce que le logiciel peut s'appuyer sur un plus grand nombre de séquences de mots (Bouillon, 2017a). Cependant, grâce à l'essor de l'apprentissage profond et à l'utilisation de celui-ci dans plusieurs applications, la version que nous utilisons de *Dragon Naturally Speaking* fonctionne selon un système neuronal : le modèle de langage utilisé est donc un modèle de langage neuronal. Comme nous avons vu à la section 5.3, un modèle de langage neuronal ressemble grandement à un modèle statistique, mais il se base sur des réseaux neuronaux et donc sur des couches (*layers*).

Une autre manière de spécialiser le système consiste à saisir une liste de mots en format .txt : ladite liste aura un format de ce type « nom\nom à dicter », ainsi il suffit de dicter le nom choisi pour indiquer un terme qui peut être compliqué à prononcer ou un nom complet (prénom et nom) par exemple. Ce processus rend la dictée plus facile. Toutefois, nous n'avons pas utilisé cette méthode parce qu'elle reste propre à chaque utilisateur et qu'il n'y avait pas tant de mots « difficiles » à dicter.

Le texte a été traduit du français à l'italien à l'aide du système de traduction automatique *DeepL*[29]. La traduction a donc été saisie comme mémoire de traduction au sein du logiciel *SDL Trados Studio*, comme s'il agissait d'une traduction déjà effectuée. Ainsi, nous avons pu réaliser la tâche de post-édition.

Tous les participants ont travaillé les deux textes, d'abord sans la reconnaissance vocale, puis avec. Les données ont été collectées selon une méthode croisée, appelée *cross-case analysis* (Saldanha & O'Brien, 2013). En effet, étudier plus d'un cas à la fois se révèle compliqué en raison des nombreux facteurs à prendre en considération et de la difficulté à établir des comparaisons raisonnées. Pour cette raison, les analyses croisées donnent un scénario particulièrement riche (Ibidem). Nous avons donc croisé les participants ainsi que l'ordre des textes à effectuer, de façon à ce que les étudiants n'interagissent pas et à rendre l'évaluation

[29]Traduction effectué au mois de juin 2019.

objective. Pendant la première phase, les participants ont travaillé de manière autonome et avec les méthodes de saisie standard. Dans la phase avec la reconnaissance vocale, ils ont également eu à disposition les méthodes traditionnelles d'entrée (clavier et souris), mais ils ont été fortement encouragés à ne les utiliser qu'en cas de réelle nécessité. Après avoir abordé le milieu expérimental, passons à la présentation des logiciels utilisés pendant l'expérience.

## 7.1 Les logiciels

Nous avons utilisé *Dragon Naturally Speaking* (DNS) au sein de *Microsoft Word*. DNS est un logiciel de reconnaissance vocale qui permet de dicter n'importe quel type de texte et également de donner des commandes vocales (*voice tags*) dans un éditeur, un traitement de texte ou un champ destiné à recevoir du contenu textuel. Nous avons utilisé la dernière version (*Dragon Professional Individual, v15*) qui se base sur la nouvelle génération de *speech engine* et donc sur l'apprentissage profond. De ce fait, la qualité de la dictée s'avère meilleure que celle de la version précédente, notamment pour ce qui est de l'adaptation aux variations environnementales et aux accents étrangers. De plus, un usage constant de l'outil permet que le logiciel s'habitue à la voix du traducteur et soit donc plus performant[30]. Néanmoins, avant de procéder à l'expérience, nous avons soumis un texte d'entraînement aux participants, afin d'entraîner le système à leur voix.

Enfin, nous avons enregistré les choix et le comportement des participants grâce à *BB Flashback* Express. Il s'agit d'un logiciel d'enregistrement d'écran, qui permet de retenir toute l'activité effectuée sur l'ordinateur après son démarrage. Nous avons ainsi pu voir les hésitations et les commandes vocales données par chaque participant[31], ainsi qu'avoir une trace effective de la durée de chaque activité.

De cette manière, nos participants ont pu travailler dans un environnement habituel et réel, avec la traduction divisée en segments, et intervenir sur le texte comme ils le feraient pour un travail réel. Nous avons souhaité faire travailler nos participants dans le tableau de révision car dans un contexte professionnel certaines corrections peuvent être saisies après la traduction et doivent ensuite être intégrées dans la mémoire de traduction, afin que cette dernière soit à jour et qu'elle inclut les corrections voulues par le client. Dans ces cas, les traductions sont souvent

[30] https://www.nuance.com/dragon/business-solutions/dragon-professional-individual.html#standardpage-mainpar_backgroundimage, consulté le 08/05/2019
[31] https://www.flashbackrecorder.com/it/express/ consulté le 12/05/19

ajoutées dans le fichier exporté et ensuite, grâce à un travail de comparaison, corrigées directement dans le logiciel.

## 7.2 Le profil des participants

Les participants étaient cinq étudiants en Maîtrise en Traduction, âgés entre 23 et 30 ans, désormais jeunes traducteurs professionnels. Tous étaient de langue maternelle italienne et avaient le français parmi leurs langues passives.

Selon un petit sondage préliminaire (voir Annexe 1), tous les étudiants avaient des connaissances préalables du logiciel *SDL Trados Studio*, ainsi que la certification de niveau intermédiaire ; aussi la post-édition était un domaine connu par tous. Quant à la reconnaissance vocale, deux participants possèdent un niveau de base et trois n'ont jamais utilisé un tel système. C'est pourquoi tous les participants ont suivi le petit tutoriel proposé par DNS et ont ensuite effectué une courte phase d'entraînement avant de commencer l'expérience.

Nous avons sélectionné ces participants en particulier, car nous étions intéressée à évaluer le comportement d'un jeune traducteur, avec un niveau d'expérience professionnelle limité dans le domaine de la traduction, donc vraisemblablement plus ouvert aux nouvelles technologies et aux possibilités offertes par la reconnaissance vocale.

De plus, nous voulions simuler un contexte de travail probable : le cas d'un traducteur indépendant qui fait son entrée dans la vie professionnelle qui traduit des textes techniques et qui se tourne vers cette technologie désormais mieux développée grâce à l'*apprentissage profond*. Ce jeune professionnel décide d'essayer d'autres possibilités offertes par le domaine des technologies de la traduction pour améliorer sa productivité, optimiser son travail et évaluer un éventuel investissement dans un logiciel comme *Dragon Naturally Speaking*. Après avoir présenté le profil des participants, passons maintenant à la description de la méthodologie.

## 7.3 Déroulement

D'abord, nous avons cherché un texte technique qui se prêtait bien à une traduction automatique qui pouvait ensuite être post-édité. Notre choix s'est porté sur le mode d'emploi d'une Smartwatch. Ensuite, nous avons sélectionné deux parties équivalentes du texte à soumettre aux participants, pour avoir deux textes traitant du même sujet. Dans *SDL Trados Studio*, nous avons créé une mémoire de traduction, issue de la traduction automatique effectuée par *DeepL*. Naturellement, les participants ont été informés que la mémoire de traduction était en réalité une traduction automatique. Ensuite, nous avons créé un projet de traduction, choisi les langues

de travail (du français à l'italien) et traduit les deux fichiers. Par la suite, nous avons exporté les deux fichiers en format .docx pour la révision et cela pour chaque participant.

Nous avons utilisé *Dragon Naturally Speaking* pour la tâche de reconnaissance vocale. Premièrement, nous avons dû créer un profil pour chaque utilisateur : il a fallu sélectionner la langue, le pays et la région. Puis s'est déroulée la phase d'adaptation à la voix, qui consiste à dicter un court texte. Le logiciel a ensuite proposé un *didacticiel interactif* qui permet de prendre ses marques avec le système et les commandes. Les participants ont suivi ce tutoriel utile à notre expérience, qui comprenait surtout des exercices concernant les commandes de base, la dictée, le microphone, la correction et la modification. Comme nous l'avons expliqué à la section 7, pour chaque profil, il a fallu ajouter le corpus avec la terminologie spécifique, soit le mode d'emploi officiel en langue italienne du produit.

Ensuite, chaque participant s'est entraîné avec DNS sur un texte conçu spécifiquement à cet effet, afin qu'il maîtrise les commandes vocales et se comporte de manière assez naturelle lors de l'expérience. Pour la phase d'entraînement, une autre partie du même texte a été sélectionnée. Cette portion de texte était, en revanche, plus courte et l'attention portait sur l'utilisation des commandes vocales. Après avoir terminé l'entraînement avec Dragon, nous avons expliqué aux participants le fonctionnement de *BB Flashback Express* et leur avons demandé de le démarrer juste avant de débuter l'expérience.

Puis l'expérience a commencé. Les participants ont démarré l'enregistrement de l'écran afin d'enregistrer toute leur activité et ont commencé à post-éditer le texte. Puisque la reconnaissance vocale était déjà active, ils ont tout simplement allumé le microphone au moment voulu. La tâche s'effectuait au sein de Microsoft Word, et l'interface se présentait de la façon suivante (voir *Tableau 6*).

| **Source segment** | **Target segment** |
|---|---|
| Entraînement avec Fitbit Coach | Allenamento con Fitbit Coach |
| L'application Fitbit Coach offre des séances d'entraînement de poids corporel sur votre poignet pour vous aider à rester en forme n'importe où. | L'applicazione Fitbit Coach offre sessioni di allenamento al polso per mantenersi in forma ovunque. |
| Pour commencer une séance d'entraînement : | Per avviare un esercizio: 1. |
| Sur votre Versa, appuyez sur l'application Fitbit Coach. | Su Versa, tocca l'app Fitbit Coach . 2. |
| Faites défiler la liste des séances d'entraînement. | Scorri l'elenco degli esercizi. 3. |

| | |
|---|---|
| Appuyez sur une séance d'entraînement, puis appuyez sur le bouton de lecture pour commencer. | Tocca un esercizio e premi il pulsante di riproduzione per iniziare. |

*Tableau 6 : Bi-texte pour la post-édition*

Après avoir accompli la tâche principale, les participants ont dû remplir un questionnaire (Annexe 2) pour nous aider à mieux comprendre leur avis et leurs impressions sur l'expérience. Ce questionnaire portait en particulier sur la reconnaissance vocale, soit sur la technologie elle-même soit en lien avec la post-édition. Les questions posées étaient ouvertes afin de laisser à chaque participant la possibilité d'exprimer sa propre opinion et sa perception de l'évaluation effectuée ; le questionnaire ne comptait que deux questions fermées, soit deux tableaux avec des questions à choix multiple (à choisir dans un menu déroulant).

Dans la section suivante, nous présenterons les résultats de l'analyse.

# 8 Résultats

Dans notre analyse, nous prendrons en compte les résultats de chaque caractéristique, en analysant les données de manière parallèle à leur présentation à la section 6 et nous les expliquerons afin de mettre en évidence le calcul effectué pour obtenir le score final.

Il faut préciser que nos participants n'étaient pas des experts en matière de PE, mais, à l'aide des directives pour la post-édition, ils ont tous bien compris l'activité à effectuer et ont réussi à la compléter de manière pertinente. Comme nous l'avons appris grâce au sondage préliminaire, ils avaient tous des connaissances de base sur le sujet, acquises dans le cadre d'un cours universitaire. Toutefois, ils n'avaient mis à l'épreuve leurs connaissances qu'une fois ou deux maximum.

*Questionnaire de post-évaluation*

Notre questionnaire de post-évaluation était composé de 15 questions, dont 13 ouvertes et 2 fermées sous forme de tableau qui contenait les réponses à choisir dans un menu déroulant. Les impressions et les opinions de nos participants étaient très similaires, toutefois, leur avis sur l'expérience effectuée ne correspondait pas toujours à la réalité, ce qui a pu être vérifié non seulement pendant l'expérience, mais également grâce aux enregistrements de *BB Flashback Express*.

Les questions portaient spécialement sur la reconnaissance vocale ; seule la première avait comme sujet la traduction automatique (TA) issue de DeepL. Puisque les participants étaient

déjà au courant que la traduction était le résultat de la TA de DeepL, cette question a été posée pour comprendre si cela les a influencés. Apparemment, cela n'a posé aucun problème et n'a pas exercé d'influence particulière, sauf dans un cas où le participant, confiant de la traduction automatique, a révélé avoir été moins critique.

Ensuite, le but des questions restantes était de pouvoir observer la perception des participants par rapport à ce qui avait été enregistré lors de l'expérience même. Nous pouvons regrouper les questions en deux « catégories » :

- Opinion sur la tâche effectuée
- Opinion sur *Dragon Naturally Speaking*

Nous avons pu constater que, en ce qui concerne la première catégorie, les réponses ne correspondent pas dans la totalité à ce qui a pu être observé lors de l'expérience. Nous présenterons les réponses données dans les sous-chapitres correspondants.

## 8.1 Efficacité

L'efficacité est l'une des caractéristiques que nous avons choisies pour notre évaluation. Comme expliqué à la section 6.2, elle a été calculée à l'aide de différentes mesures. Nous expliquerons plus en détails l'attribution des points, en montrant spécifiquement pour chacun de nos cinq participants leurs scores et points.

Le choix de se baser sur ces mesures se justifie par le fait que dans leur totalité, elles peuvent donner un jugement complet et exhaustif de l'efficacité du système analysé, aussi parce que seulement l'une parmi elles ne serait pas suffisante pour une évaluation précise. Comme décrit dans les normes ISO : « *If tasks can be partially completed the objectives achieved mesure is more appropriate. »* (2016).

Nous commencerons par les tâches complétées, que nous avons énumérées auparavant, ensuite nous passerons aux autres mesures et, enfin, nous détaillerons la qualité de la traduction et le temps, soit deux aspects qui jouent bel et bien un rôle dans l'efficacité d'un système.

### 8.1.1 Tâches complétées

Nous avons listé cinq tâches à compléter, afin que les objectifs soient atteints dans leur totalité. Si le tâches étaient complétées entièrement, les participants recevaient 10 points ; si elles étaient partiellement complétées, ils recevaient 5 points, et si elles n'étaient pas complétées, ils ne recevaient aucun point. Dans le *Tableau 7*, nous pouvons observer que deux tâches n'ont pas

été achevées dans deux cas uniquement. Il s'agit de deux participants différents ainsi que de deux tâches distinctes. Les deux tâches en question sont en fait la *b* (exécuter correctement les commandes vocales) et la *c* (corriger correctement à l'aide de la reconnaissance vocale les segments, en particulier sans rajouter d'erreurs autre part) (voir section 6.1). Cela signifie que, dans un cas, les commandes vocales n'ont pas été exécutées correctement même après plusieurs tentatives. Effectivement, il faut préciser que pour les trois tâches concernant la dictée (dictée et commandes vocales), nous n'avons pas considéré la tâche comme « non complétée » ou « partiellement complétée » après une seule tentative infructueuse, mais lorsque le participant n'a pas réussi à dicter la phrase ou à exécuter la commande souhaitée après plusieurs tentatives. Il se peut, donc, que le participant 4 ait fini pour effectuer l'opération souhaitée à l'aide de la souris ou du clavier.

En revanche, en ce qui concerne la tâche *c* du participant 5, nous pouvons constater que la correction des segments n'a pas été précise, à cause de certaines erreurs qui ont été ajoutées. Le contributeur, qui s'est focalisé sur l'utilisation de la reconnaissance vocale plutôt que sur la correction du texte, peut avoir négligé le texte et la cohérence des corrections. Nous le verrons également plus en détails dans la section suivante.

Enfin, nous pouvons observer que, de manière générale, toutes les tâches ont été accomplies, sauf dans sept cas où elles ont été achevées partiellement. Ces résultats ne sont que peu surprenants si l'on considère l'inexpérience dans le domaine de la reconnaissance vocale et dans l'utilisation du logiciel de nos participants. Cependant, en moyenne, nous constatons un bon niveau d'efficacité : aucun participant n'a totalisé moins de la moitié des points pouvant être additionnés, comme nous pouvons l'observer dans la dernière ligne du tableau ci-dessous.

| Tâche | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| **a)** | 10 | 10 | 10 | 10 | 10 |
| **b)** | 5 | 5 | 10 | 0 | 5 |
| **c)** | 10 | 10 | 5 | 10 | 0 |
| **d)** | 5 | 10 | 10 | 5 | 5 |
| **e)** | 10 | 10 | 10 | 10 | 10 |
| **Total** | 40 | 45 | 45 | 35 | 30 |
| | **0,8** | **0,9** | **0,9** | **0,7** | **0,6** |
| | 80% | 90% | 90% | 70% | 60% |

*Tableau 7 : Récapitulatif des scores pour les tâches effectuées*

#### *8.1.1.1 Objectifs complétés*

Les objectifs complétés ont été atteints dans quatre cas sur cinq. Le seul participant qui a obtenu un score de 0,75 n'a pas effectué une traduction optimale en utilisant la reconnaissance vocale, car il a ajouté des erreurs pendant la correction, comme le montre le *Tableau 8*, ce qui ne permet pas de considérer la traduction comme acceptable dans un contexte professionnel. En effet, dans ce cas, la reconnaissance vocale a piégé le participant, qui s'est concentré sur l'utilisation de cette technologie plutôt que sur le texte ainsi que sa qualité. Il est vrai aussi que le mode « suivi des modifications » étant activé, il aurait pu engendrer une confusion pour le participant dans la visualisation des phrases.

| **Source segment** | **Target segment** |
|---|---|
| Lorsque vous acceptez un appel, c'est votre téléphone à proximité qui décroche. | Quando si accetta una **chiamata,è** il telefono più vicino a rispondere. I. |
| Pour rejeter un appel, appuyez sur l'icône symbolisant un téléphone rouge pour rediriger votre interlocuteur vers la messagerie vocale. | Per rifiutare una chiamata, premere l'icona rossa del telefono per reindirizzare **il interlocutore** alla segreteria telefonica. |
| Le nom de votre interlocuteur s'affiche s'il est répertorié dans vos contacts. | Il nome **del interlocutore** viene visualizzato se è presente nell'elenco dei contatti. |

*Tableau 8: Erreurs ajoutées par un des participants*

| **Objectif** | **P1** | **P2** | **P3** | **P4** | **P5** |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 1 | 1 | 1 |
| **2** | 1 | 1 | 1 | 1 | 1 |
| **3** | 1 | 1 | 1 | 1 | 0 |
| **4** | 1 | 1 | 1 | 1 | 1 |
| **Total** | 4 | 4 | 4 | 4 | 3 |
| | **1** | **1** | **1** | **1** | **0,75** |
| | **100%** | **100%** | **100%** | **100%** | **75%** |

*Tableau 9 : Scores relatifs aux objectifs complétés*

Tous les autres participants ont atteint tous les objectifs, comme nous pouvons voir dans la le *Tableau 9* résumant les scores concernant cet aspect.

#### *8.1.1.2 Erreurs dans une tâche*

Pour calculer le poids des erreurs commises dans chaque tâche, nous nous sommes basée sur les catégories d'erreurs données à la section 6.2.1.2 ; nous avons donc additionné toutes les erreurs et les avons ensuite soustraites à 50 (le maximum de points pour l'ensemble des tâches). Dans le *Tableau 10* ci-dessous, nous pouvons observer le total des points des erreurs, ensuite le total déjà soustrait des points pour chaque tâche et les points attribués sur le total d'un point.

Nous pouvons observer que les erreurs les plus fréquentes concernent les commandes vocales, ce qui n'est pas surprenant étant donné que nous connaissions déjà un manque de formation sur ce sujet. Le style a également posé quelques problèmes, sûrement à cause de l'influence de la RV et de la distraction que celle-ci a pu apporter. Le même constat peut être dressé pour toutes les erreurs d'orthographe, causées sans doute par la concentration maximale portée sur la reconnaissance vocale et non sur l'output final de la traduction. En revanche, notons que la terminologie n'a pas été source de problème, même s'il s'agissait d'un texte technique.

| | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| **a)** | 1,5 | 1 | 1 | 1 | 0,5 |
| **b)** | 0 | 0 | 0 | 0 | 0 |
| **c)** | 1 | 1 | 1 | 1 | 0,5 |
| **d)** | 1 | 0 | 0 | 3 | 0 |
| **e)** | 2 | 2 | 0 | 0 | 4 |
| **Total** | **5,5** | **4** | **2** | **5** | **5** |
| **Points** | 44,5 | 46 | 48 | 45 | 45 |
| | 0,89 | 0,92 | 0,96 | 0,9 | 0,9 |
| | **89%** | **92%** | **96%** | **90%** | **90%** |

*Tableau 10 : Résultats des erreurs dans les différentes tâches*

#### *8.1.1.3 Tâches comportant des erreurs*

Pour dénombrer les tâches qui comportaient des erreurs, nous avons calculé combien d'entre elles comptaient au moins une erreur. Sur un total de cinq tâches, trois participants ont accompli quatre tâches sur cinq sans erreurs, et deux ont fait au moins une faute dans deux tâches. Cela nous donne un score de 0,8 pour les premiers et 0,6 pour les seconds, comme le montre le *Tableau 11*. Nous pouvons observer qu'aucun participant n'a pu compléter la tâche *b* (exécuter correctement les commandes vocales) sans commettre au moins une erreur : il s'agit effectivement du problème le plus récurrent. Cela s'explique aussi en partie par la nouveauté du logiciel pour la plupart des participants, qui ont donc dû prendre du temps pour maîtriser cette technologie.

| Tâche | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| a) | 1 | 1 | 1 | 1 | 1 |
| b) | 0 | 0 | 0 | 0 | 0 |
| c) | 0 | 1 | 1 | 1 | 0 |
| d) | 1 | 1 | 1 | 1 | 1 |
| e) | 1 | 1 | 1 | 1 | 1 |
| Total | 3 | 4 | 4 | 4 | 3 |
| Points | 0,6 | 0,8 | 0,8 | 0,8 | 0,6 |
| | 60% | 80% | 80% | 80% | 60% |

*Tableau 11 : Tâches incluant au moins une erreur*

Nous avons inclus un tableau (voir *Tableau 12*) complet de toutes les mesures appliquées et utilisées pour le calcul de l'efficacité et nous avons effectué la somme afin d'avoir un score final pour chaque participant pour cette caractéristique. Dans la dernière colonne « Total » nous avons calculé la somme de chaque résultat, c'est pourquoi il est calculé sur un total de 4.

| | Tâches complétées | Objectifs complétés | Erreurs dans tâche | Tâches avec erreurs | Total | Total en pourcentage |
|---|---|---|---|---|---|---|
| **Participant 1** | 0,8/1 | 1/1 | 0,89/1 | 0,6/1 | 3,29/4 | 82,25% |
| **Participant 2** | 0,9/1 | 1/1 | 0,92/1 | 0,8/1 | 3,62/4 | 90,5% |
| **Participant 3** | 0,9/1 | 1/1 | 0,96/1 | 0,8/1 | 3,66/4 | 91,5% |
| **Participant 4** | 0,7/1 | 1/1 | 0,9/1 | 0,8/1 | 3,4/4 | 85% |
| **Participant 5** | 0,6/1 | 0,75/1 | 0,9/1 | 0,6/1 | 2,5/4 | 62,5% |

*Tableau 12: Total des scores relatifs à l'efficacité*

Les résultats obtenus permettent de répondre à notre question de recherche, soit que la reconnaissance vocale joue un rôle important au niveau de l'efficacité lors de la tâche de post-édition, sans impacter la qualité pour autant, comme nous le verrons à la section suivante.

Passons maintenant aux réponses du questionnaire de post-évaluation. Nous avons demandé aux participants si l'utilisation de la reconnaissance vocale avait influencé le processus de correction en les incitant à effectuer moins de modifications par rapport aux méthodes traditionnelles d'entrée. Quatre participants sur cinq estiment qu'ils ont corrigé de la même manière, voire un peu plus, et que la RV n'a posé aucun problème de ce point de vue. Un participant, en revanche, affirme que la correction à l'aide des méthodes d'entrée traditionnelles permet de corriger plus librement, avec la possibilité de se déplacer rapidement dans plusieurs

parties du texte. Tandis que la RV, si elle permet la même liberté, requiert plus de temps, ce qui demande une réflexion plus importante avant d'effectuer des corrections.

Ensuite, nous leur avons demandé si la correction a été effectuée de la même manière pour les deux méthodes (surtout en ce qui concerne le nombre de corrections) : trois participants sur cinq croient avoir apporté les mêmes corrections ; un affirme avoir corrigé plus judicieusement avec la reconnaissance vocale, bien qu'il ait ajouté des erreurs lors de la phase de correction avec la reconnaissance vocale. Enfin, un pense avoir apporté plus de corrections avec la RV.

Nous avons également demandé aux participants si, à la fin de l'expérience, les difficultés avec l'utilisation de la RV étaient les mêmes. Tous ont affirmé que le fonctionnement est assez intuitif, ce qui facilite l'apprentissage ainsi que l'utilisation.

Aucun participant n'estime que la reconnaissance vocale pourrait se révéler meilleure au niveau de la productivité, même avec une meilleure connaissance que celle-ci, surtout pour une tâche de post-édition ; l'un d'eux estime que la correction avec des fonctionnalités telles que « rechercher et remplacer » ou des expressions régulières pourrait être plus productive. Un participant, en revanche, pense que cette possibilité serait utile du point de vue de l'accessibilité. Toutefois, dans le cadre de notre étude nous ne nous intéressons pas à ce domaine.

Enfin, tous se disent satisfaits du point de vue de l'efficacité du fonctionnement des commandes vocales et de la dictée. Les seuls cas qui ont posé problème sont probablement liés à un manque de maîtrise des participants, comme pour le déplacement du curseur à travers la commande « griglia del mouse » (voir section 8.2).

### 8.1.2 Qualité de la traduction

Comme expliqué au chapitre 6.2.2, nous avons évalué la qualité de la traduction à l'aide d'évaluations automatique et humaine. En ce qui concerne l'évaluation automatique, nous nous sommes appuyée sur les scores BLEU et TER. Nous avons choisi ces deux mesures, car la première est effectuée automatiquement et nous a permis de faire un tri pour la sélection des phrases à analyser dans notre évaluation humaine. Néanmoins, nous sommes consciente que ce type de score a ses limites, étant donné qu'il ne se base que sur la correspondance des mots, sans tenir compte du sens. Pour cette raison, la reformulation est pénalisée lorsque l'on utilise ce score. Nous avons ensuite ajouté TER, parce qu'il permet également une comparaison avec un texte de référence, mais calcule tout de même le nombre d'éditions apportées. Certes, en plus de ces méthodes, il a été nécessaire d'analyser les résultats de l'évaluation humaine.

Les quatre *Tableaux 13, 14, 15* et *16* résument les résultats obtenus pour chaque score, ainsi que l'écart. Cet écart a toujours été calculé de la manière suivante : scores des méthodes traditionnelles d'entrée (dans les tableaux ci-dessous « ET ») moins les scores de la reconnaissance vocale[32]. En effet, nous partons du principe que la tâche effectuée à l'aide des méthodes traditionnelles d'entrée requiert moins de temps, car les participants ont l'habitude de taper leurs textes et corrections, par rapport à l'utilisation de la reconnaissance vocale pour les mêmes activités. Nous reprendrons les résultats par mesures dans les sections suivantes.

| **Texte 1** | **BLEU (ET)** | **BLEU (RV)** | **ÉCART (ET-RV)** |
|---|---|---|---|
| P1 | 52,38 | 54,14 | -1,76 |
| P2 | 51,91 | 52,31 | -0,4 |
| P3 | 51,27 | 53,05 | -1,78 |
| P4 | 53,47 | 53,47 | 0 |
| P5 | 50,31 | 52,01 | -1,7 |
| **Moyenne** | **51,868** | **53,00** | **-1,132** |

*Tableau 13: Récapitulatif des scores BLEU pour le texte 1 et l'écart entre les méthodes de RV et d'entrée traditionnelles*

| **Texte 2** | **BLEU (ET)** | **BLEU (RV)** | **ÉCART (ET-RV)** |
|---|---|---|---|
| P1 | 80,66 | 82,27 | -1,61 |
| P2 | 89,57 | 89,57 | 0 |
| P3 | 89,35 | 87,24 | 2,11 |
| P4 | 89,37 | 86,10 | 3,27 |
| P5 | 81,99 | 84,13 | -2,14 |
| **Moyenne** | **86,188** | **85,862** | **0,326** |

*Tableau 14: Récapitulatif des scores BLEU pour le texte 2 et l'écart entre les méthodes de RV et d'entrée traditionnelles*

| **Texte 1** | **TER (ET)** | **TER (RV)** | **ÉCART (ET-RV)** |
|---|---|---|---|
| P1 | 0,190 | 0,145 | 0,045 |
| P2 | 0,103 | 0,077 | 0,026 |
| P3 | 0,138 | 0,057 | 0,081 |
| P4 | 0,088 | 0,048 | 0,04 |
| P5 | 0,120 | 0,107 | 0,013 |
| **Moyenne** | **0,128** | **0,09** | **0,038** |

*Tableau 15: Récapitulatif des scores TER pour le texte 1 et l'écart entre les méthodes de RV et d'entrée traditionnelles*

| **Texte 2** | **TER (ET)** | **TER (RV)** | **ÉCART (ET-RV)** |
|---|---|---|---|
| P1 | 0,245 | 0,171 | 0,074 |
| P2 | 0,063 | 0,040 | 0,023 |
| P3 | 0,063 | 0,049 | 0,014 |
| P4 | 0,053 | 0,100 | -0,047 |
| P5 | 0,100 | 0,081 | 0,019 |
| **Moyenne** | **0,105** | **0,088** | **0,017** |

*Tableau 16 Récapitulatif des scores TER pour le texte 2 et l'écart entre les méthodes de RV et d'entrée traditionnelles*

[32] Formule à retrouver dans les tableaux ci-dessous : ET-RV

#### *8.1.2.1 BLEU*

BLEU a été calculé grâce à l'outil en ligne https://www.letsmt.eu/Bleu.aspx. Nous avons utilisé la version italienne du manuel comme référence humaine, étant donné qu'il a été publié sur le site officiel, et nous avons calculé le score pour les deux textes post-édités par les 5 participants. L'outil offre une visualisation claire du score, en montrant d'abord le score cumulatif, correspondant aux 4-grammes ; ensuite il donne le score individuel, correspondant aux 1-grammes, et cumulatif pour chaque typologie de n-gramme.

Pour les deux textes, nous nous sommes basée sur les bigrammes, utilisant ainsi une version lissée de BLEU, à cause du score plus bas observé dans certaines phrases très courtes et, par conséquent, un score de corrélation global plus faible (Specia et al., 2010).

En moyenne, les participants ont obtenu un score relativement bas pour le premier texte avec une moyenne de 53. Le second texte, au contraire, se révèle de loin meilleur, avec un score moyen de 85,8. Ces moyennes sont issues de tous les scores BLEU des textes post-édités par nos participants (texte 1 et texte 2).

Comme le montre le *Tableau 13* récapitulatif ci-dessus, pour le texte 1 les participants ont obtenu un score BLEU meilleur avec la RV dans quatre cas sur cinq. Dans la colonne « écart », nous pouvons constater des valeurs négatives, qui correspondent donc à un meilleur résultat, étant donné que nous avons toujours soustrait le score de la RV du score « ET » et que plus BLEU élevé, plus le score est bon. Bien que les écarts ne soient pas si marqués, nous pouvons remarquer un écart moyen de 1,132. Seulement dans un cas, BLEU a été le même pour les deux méthodes d'entrée (P4).

Pour le texte 2, par contre, il n'y a que deux cas où BLEU est meilleur lors de l'utilisation de la RV (P1 et P5) : si nous nous focalisons sur l'écart, nous pouvons constater que, pour le P5, il demeure assez important. Ici aussi, dans un cas, BLEU a été le même (P2). Dans deux cas (P3 et P4), les méthodes traditionnelles d'entrée ont donné un meilleur résultat, confirmé aussi par un écart assez important (voir *Tableau 14*).

Bien que BLEU soit considéré comme un bon moyen pour l'évaluation automatique, il ne faut pas négliger les limites liées à cette méthode : il ne pénalise pas le mélange parmi n-grammes, et ne reconnaît pas les synonymes, qui peuvent donc être pénalisés incorrectement. De plus, en ne se basant que sur la correspondance des n-grammes, il ne tient pas compte des phrases qui ont un sens différent : il se peut que les n-grammes correspondent, mais pas la signification.

Comme il n'indique pas s'il manque une information de la « traduction candidate », il se limite à examiner la correspondance de la traduction parmi les n-grammes, sans se focaliser sur les mots manquants par rapport à la référence (Cer et al., 2010; Lin & Och, 2004).

#### *8.1.2.2 TER*

Nous avons sélectionné TER, parce qu'il ne nécessite pas un grand nombre de phrases de référence afin de corréler à un jugement humain (O'Brien, 2011). Il a été calculé à l'aide du script *tercom* qui fonctionne avec Java[33]. Comme déjà expliqué au chapitre 6.2.2.2, il mesure le minimum d'éditions nécessaires entre deux textes. Dans notre cas, nous avons calculé le nombre d'éditions entre le texte issu de la traduction automatique et soumis aux participants et celui post-édité par ces derniers (Snover et al., 2006).

Étant donné que TER mesure des modifications, 0 est le meilleur score possible, parce qu'il signifie qu'il n'y a eu besoin d'aucun changement pour que le texte corresponde à la référence. Bien que TER soit utilisé pour mesurer la qualité de la traduction automatique, nous l'avons utilisé surtout pour observer s'il existe une différence de TER entre les méthodes traditionnelles d'entrée et la reconnaissance vocale. De cette manière, il a été donc possible de voir si les modifications était plus important à l'aide de la RV ou à l'aide des méthodes traditionnelles d'entrée : il se peut que l'habitude à taper les corrections produise un nombre majeur d'interventions par rapport aux corrections effectuées à l'aide de la RV.

Comme le montrent les *Tableaux 15* et *16* récapitulatifs, pour le texte 1, TER a été meilleur pour tous les participants, comme l'indique l'écart positif dans la colonne de droite, en restant toujours en dessous de 0,1. La même analyse peut être associée au texte 2, où la reconnaissance vocale a obtenu un meilleur score dans quatre cas sur cinq, malgré une différence assez réduite (et toujours en dessous de 0,1). Cependant, pour ce texte, le P4 a fourni un output apparemment meilleur à l'aide des méthodes traditionnelles d'entrée.

*Reconnaissance vocale*

Pour les textes post-édités à l'aide de la reconnaissance vocale, le P1 a obtenu le meilleur score BLEU, dans le premier texte. Par contre, il a obtenu le score le plus bas pour ce qui est de TER. Toutefois, les autres résultats sont presque tous similaires, puisqu'il n'y a pas une large différence entre les points obtenus pour BLEU. TER, au contraire, montre une différence plus marquée entre le P3 et le P1 par exemple. Il ne faut pas négliger, cependant, que pour calculer

[33] Lien vers le script: http://www.cs.umd.edu/~snover/tercom/

BLEU, nous avons utilisé le texte officiel comme référence ; tandis que pour obtenir TER, les textes utilisés ont été la traduction automatique et sa version post-éditée à l'aide de la reconnaissance vocale.

Pour le texte 2, les scores sont plus similaires : le P2 a obtenu, en effet, le meilleur score BLEU et TER et les autres participants montrent aussi des résultats comparables. Le P3 a également obtenu de bons scores BLEU et TER, tandis que le P1, qui a obtenu les points les plus bas pour BLEU, a aussi enregistré un mauvais score pour TER par rapport aux autres participants.

Certes, ce score a aussi des limites : certains participants n'ont peut-être jamais pratiqué à un niveau plus professionnel la post-édition et sont donc influencés par leur expérience limitée dans le domaine. La post-édition, sauf en cas d'erreurs graves, reste assez subjective et il se peut que certains n'apportent pas beaucoup de corrections (O'Brien, 2011).

*Méthodes traditionnelles d'entrée*

Pour la post-édition avec les méthodes traditionnelles d'entrée, nous pouvons remarquer que le premier texte a obtenu des scores plus bas que ceux de la reconnaissance vocale, en ce qui concerne BLEU. En effet, comme le montrent les *Tableaux 13* et *14*récapitulatifs, le meilleur score BLEU est celui du P4 (53,47), contre 54,14 pour le P1. Certes, l'écart n'est pas majeur, mais en observant également les résultats globaux, nous observons une meilleure moyenne pour l'utilisation de la RV. En revanche, dans le texte numéro 2, le score BLEU a été meilleur pour la post-édition avec les méthodes traditionnelles d'entrée de manière générale, mais seul le P2 obtient le meilleur score dans les deux cas, avec 89,57. Ce que nous pouvons constater c'est qu'il n'y a pas de différences notables entre les deux méthodes de post-édition, ce qui corrobore notre hypothèse.

Passons maintenant au TER : pour les deux textes, les participants ont obtenu des scores moins élevés dans cette phase plutôt que dans celle de la RV. Pour le premier texte, l'écart est tout de même légèrement plus important car les meilleurs scores sont de 0,048 pour la RV et de 0,088 pour les méthodes traditionnelles d'entrée, ce qui correspond presque au double. Ensuite, pour le texte 2, la différence n'est pas si marquée. Mais dans ce cas aussi, le meilleur score a été obtenu dans la phase d'utilisation de la reconnaissance vocale (0,040), tandis qu'avec les méthodes traditionnelles le meilleur score correspond à 0,053.

Si pour BLEU nous pouvons penser que le petit écart confirme notre hypothèse, dans le cas de TER, nous ne devons pas négliger que ce dernier correspond au nombre de modifications

effectuées. Il est donc possible que nos participants aient effectué moins de modifications dans la phase avec la reconnaissance vocale, parce qu'ils ne se sentaient pas totalement à l'aise avec cette dernière et qu'ils ont donc corrigé le minimum nécessaire. Alors que pour les méthodes traditionnelles d'entrée, ils ont l'habitude d'utiliser ces outils, comme ils l'ont également mentionné dans le questionnaire.

#### *8.1.2.3 Évaluation humaine*

Bien que l'évaluation humaine des textes soit souvent perçue et décrite comme subjective et incohérente, elle reste toujours le dernier « gold standard » qui ne peut être dépassé par aucune mesure automatique (KIT & WONG, 2014). Le désavantage principal est qu'elle requiert beaucoup de temps et qu'elle est très coûteuse si elle doit être effectuée régulièrement (Koehn & Monz, 2006).

Toutefois, nous avons fait recours aux deux critères expliqués à la section 6.2.2.3 : l'adéquation (*adequacy*) et la fluidité (*fluency*). Pour ce faire, nous avons soumis à sept évaluateurs humains un certain nombre de phrases issues de la post-édition, avec et sans reconnaissance vocale. Le critère de ce choix a été le score BLEU : sur la base de ce score, nous avons choisi les phrases à sélectionner. Nous nous sommes concentrée sur les segments qui ont posé le plus de problèmes et qui ont obtenu un score plus bas par rapport au texte de référence, soit le texte officiel qui a été finalement publié par l'entreprise productrice de la Smartwatch.

Nous avons procédé de la même façon pour la post-édition effectuée des deux manières. Pour le premier texte, nous avons choisi une limite égale au score 50 : toutes les phrases avec un score en dessous de ces points ont donc été sélectionnées. Pour le deuxième texte, nous avons été obligée de prendre une limite plus haute, parce que le texte a donné des résultats meilleurs par rapport au premier, il n'aurait donc pas été intéressant d'analyser des phrases sans erreurs. La limite choisie a donc été celle de 80 : les segments sélectionnés ont été relativement moins nombreux par rapport au texte 1. Nous avons donc choisi 12 segments pour le premier texte post-édité avec les méthodes traditionnelles d'entrée et 6 pour le deuxième texte ; tandis que pour la post-édition à l'aide de la RV nous avons identifié 11 segments pour le texte 1 et 5 pour le texte 2.

Les *Tableaux 17, 18, 19* et *20* montrent les moyennes des résultats de la fluidité et de l'adéquation évaluées par les sept évaluateurs humains. Comme le montrent les tableaux les résultats ne montrent pas de grands écarts, tout de même, pour ce qui est de la fluidité, nos

évaluateurs estiment que le texte 1 est plus fluide lorsqu'il est post-édité à l'aide de la reconnaissance vocale, tandis que le texte 2 semble être plus fluide lorsqu'il a été post-édité à l'aide des méthodes traditionnelles d'entrée. En ce qui concerne l'adéquation, donc lorsque les évaluateurs ont dû donner un jugement en tenant compte aussi du texte source, les deux méthodes d'entrée sont très proches pour les deux textes. Ces résultats sont donc globalement positifs pour la reconnaissance vocale utilisée pendant la phase de post-édition puisqu'ils montrent que la qualité n'est pas affectée de manière négative lors de l'utilisation d'un système de RV tel que Dragon.

| **FLUIDITÉ ET** | | |
|---|---|---|
| | **TEXTE 1** | **TEXTE 2** |
| EVALUAT. 1 | 3,64 | 3,1 |
| EVALUAT. 2 | 4,03 | 3,46 |
| EVALUAT. 3 | 4,33 | 4,53 |
| EVALUAT. 4 | 3,85 | 4,06 |
| EVALUAT. 5 | 3,94 | 3,62 |
| EVALUAT. 6 | 3,7 | 3,56 |
| EVALUAT. 7 | 3,33 | 3,5 |

*Tableau 17: Récapitulatif des moyennes de scores de la fluidité pour les méthodes traditionnelles d'entrée*

| **ADÉQUATION ET** | | |
|---|---|---|
| | **TEXTE 1** | **TEXTE 2** |
| EVALUAT. 1 | 3,61 | 3 |
| EVALUAT. 2 | 4,96 | 4,8 |
| EVALUAT. 3 | 4,75 | 4,56 |
| EVALUAT. 4 | 4,8 | 4,5 |
| EVALUAT. 5 | 4,5 | 3,5 |
| EVALUAT. 6 | 4,23 | 3,76 |
| EVALUAT. 7 | 3,52 | 3,26 |

*Tableau 18: Récapitulatif de moyennes de scores de l'adéquation pour les méthodes traditionnelles d'entrée*

| **FLUIDITÉ RV** | | |
|---|---|---|
| | **TEXTE 1** | **TEXTE 2** |
| EVALUAT. 1 | 3,63 | 3,2 |
| EVALUAT. 2 | 4,1 | 3,16 |
| EVALUAT. 3 | 4,5 | 4,44 |
| EVALUAT. 4 | 3,89 | 4,32 |
| EVALUAT. 5 | 4,01 | 3,3 |
| EVALUAT. 6 | 4,12 | 3,6 |
| EVALUAT. 7 | 3,34 | 3,08 |

*Tableau 19: Récapitulatif de moyennes de scores de la fluidité pour la reconnaissance vocale*

| **ADÉQUATION RV** | | |
|---|---|---|
| | **TEXTE 1** | **TEXTE 2** |
| EVALUAT. 1 | 3,7 | 3,24 |
| EVALUAT. 2 | 4,98 | 4,8 |
| EVALUAT. 3 | 4,92 | 4,64 |
| EVALUAT. 4 | 4,74 | 4,4 |
| EVALUAT. 5 | 4,03 | 3,32 |
| EVALUAT. 6 | 4,27 | 3,76 |
| EVALUAT. 7 | 3,2 | 3,12 |

*Tableau 20: Récapitulatif des moyennes de scores de l'adéquation pour la reconnaissance vocale*

*Kappa*

Si les mesures humaines sont très utiles et permettent d'obtenir des résultats pertinents, elles présentent toutefois un inconvénient : selon diverses études, il est très rare que les juges soient cohérents, raison pour laquelle l'accord entre évaluateurs demeure un problème majeur (Starlander, 2016). Néanmoins, plus le nombre d'évaluateurs est élevé, plus l'évaluation est fiable : dans notre cas nous avons sélectionné sept évaluateurs. Afin de déterminer l'accord entre les évaluateurs, il est indispensable de calculer le degré de cohérence entre ceux-ci (*inter-rater agreement*). À cet effet, nous avons eu recours au *kappa*, défini par (Hallgren, 2012, p. 26) comme suit :

> « *Kappa statistics measure the observed level of agreement between coders for a set of nominal ratings and corrects for agreement that would be expected by chance, providing a standardized index of IRR that can be generalized across studies.* »

Il s'agit donc d'une mesure capable de quantifier la cohérence entre plusieurs évaluateurs. Initialement proposée par Cohen en 1960 pour le domaine médical, elle a été adaptée à la linguistique computationnelle par Jean Carletta en 1996.

Pour calculer le *kappa*, nous avons utilisé le logiciel R, qui fonctionne à peu près comme un langage de programmation avec des variables. Il existe différents *kappa*, selon les méthodes d'annotations : on parle de *Light's κ* lorsque les évaluateurs évaluent l'ensemble de données, ce qui correspond à un design complètement croisé (*fully-crossed design)*. Autrement, on parle de *Fleiss's κ* lorsque les annotateurs, toujours en nombre égal ou supérieur à deux, ont annoté de manière casuelle (l'on parle de « *randomly sampled* ») (Bouillon, 2019a). Puisque nos évaluateurs ont travaillé sur les mêmes phrases, nous avons utilisé le *Light's κ*.

Les mesures de kappa varient de 0 à 1, 1 correspondant à un accord parfait. L'interprétation des résultats demeure une difficulté majeure. Landis et Koch (1977) ont donné une échelle afin d'interpréter les valeurs qui est la plus utilisée dans le domaine du traitement automatique de la langue (TAL) :

| **Échelle** | **Interprétation** |
|---|---|
| **< 0,00** | Poor (Désaccord) |
| **0,00-0,20** | Slight (Accord très faible) |
| **0,21-0.40** | Fair (Accord faible) |

| | |
|---|---|
| **0,41-0,60** | Moderate (Accord modéré) |
| **0,61-0,80** | Substantial (Accord important) |
| **0,81-1,00** | Almost Perfect (Accord presque parfait) |

*Tableau 21: Échelle d'interprétation de Landis et Koch*

À noter que deux effets peuvent générer une déformation de l'IRR (i*nter-rater reliability)* d'une mesure. D'après Starlander (2016) si le résultat de l'évaluation est très homogène (les mêmes catégories sont choisies par les évaluateurs), les données présenteront une prévalence qui a tendance à baisser artificiellement les estimations de fiabilité des évaluateurs. Hallgren le définit comme suit :

> « *The first effect appears when the marginal distributions of observed ratings fall under one category of ratings at a much higher rate over another, called the prevalence problem, which typically causes kappa estimates to be unrepresentatively low. Prevalence problems may exist within a set of ratings due to the nature of the coding system used in a study, the tendency for coders to identify one or more categories of behavior codes more often than others, or due to truly unequal frequencies of events occurring within the population under study*. » (Hallgren, 2012, p. 27)

Les résultats obtenus sur la base de nos sept évaluateurs ne montrent pas un accord très élevé : tous les scores recueillis restent en dessous de 0,20, ce qui donne un accord très faible. Ce phénomène peut s'expliquer par le nombre d'évaluateurs assez élevé. En effet, ils ne sont généralement que deux ou trois pour une même évaluation. Nous en avons choisi sept, ce qui montre donc qu'avec autant d'évaluateurs, les résultats sont tout de même acceptables.

Le *Tableau* 18 résume les résultats issus de kappa.

| **Critère et méthode d'entrée** | **Kappa** | **Interprétation selon (Landis & Koch, 1977)** |
|---|---|---|
| Fluidité des méthodes tradition. d'entrée | 0,112 | Slight (Accord très faible) |
| Fluidité de la reconnaissance vocale | 0,056 | Slight (Accord très faible) |
| Adéquation méthodes tradition. d'entrée | 0,065 | Slight (Accord très faible) |
| Adéquation de la reconnaissance vocale | 0,089 | Slight (Accord très faible) |

*Tableau 22: Récapitulatif des résultats du kappa*

L'accord le plus important concerne la fluidité des phrases post-éditées avec les méthodes traditionnelles d'entrée. Nous devons souligner que nous avons rencontré le problème classique du manque d'accord entre évaluateurs, si l'on observe les très faibles scores obtenus pour le kappa. Nous pouvons donc conclure que le kappa n'est sans doute pas le meilleur indice de cohérence entre évaluateurs (Starlander, 2016).

Dans notre questionnaire de post-évaluation, nous avons demandé aux participants si, selon eux, la qualité de la traduction subit un impact lors de l'utilisation de la reconnaissance vocale : deux participants estiment que « non », un suppose que cela varie selon le post-éditeur et ses préférences (soit un facteur personnel) et deux croient que le fait de dicter un texte le rend plus « définitif » et qu'avec une bonne connaissance et un peu d'expérience avec le logiciel la qualité ne devrait pas être pénalisée. Nous avons également demandé à nos participants s'ils avaient eu l'impression que la reconnaissance vocale les avait aidés lors de la PE : tous estiment que non, en raison de leur manque de pratique.

### 8.1.3 Temps

La mesure du temps permet de nous renseigner sur productivité. Comme présenté à la section 6.2.3, nous avons évalué cette caractéristique grâce aux enregistrements d'écrans effectués. À noter que les deux phases du test ont été effectuées sur les mêmes textes, ce qui a sans doute influé sur le temps aussi. Lors de la deuxième phase, en effet, les participants connaissaient déjà les textes et savaient déjà où se trouvaient les parties à corriger, même si les participants n'ont pas effectué les mêmes corrections dans certains cas. Cette situation a réduit légèrement le temps utilisé pour cette tâche.

Les cinq participants ont tous employé un temps similaire : 45 minutes au total. Il leur a fallu environ une dizaine de minutes pour la phase d'entraînement, et ensuite environ une quinzaine de minutes en moyenne par activité.

Pendant la phase avec méthodes traditionnelles d'entrée, les participants ont pris le temps de lire attentivement chaque segment et de vérifier s'il y avait des corrections à effectuer. Ensuite, la correction a été très rapide, même s'il y avait quelques hésitations, plutôt d'ordre linguistique. Pendant la deuxième tâche, au contraire l'activité qui a pris le plus de temps a été le positionnement dans le texte ou la sélection de la phrase à corriger pour ensuite exécuter la bonne commande vocale.

Le *Tableau 19* résume le temps employé pour chaque phase :

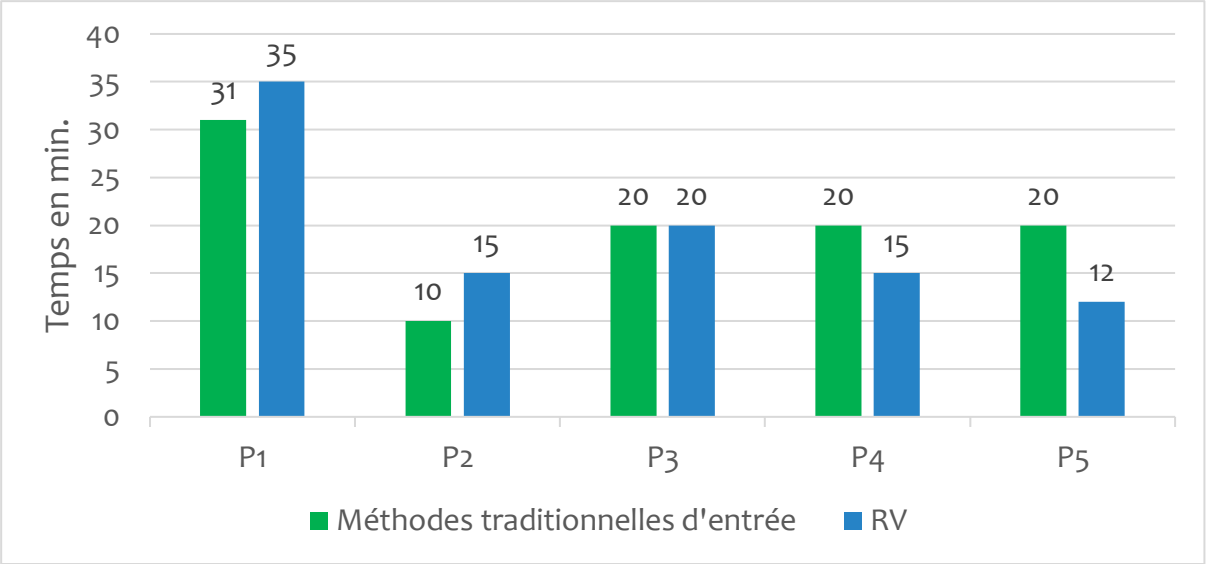



*Tableau 23 : Graphique des temps de chaque participant*

Nous constatons qu'il y a un gain de temps dans trois cas sur cinq, bien que dans les deux cas restants, la perte de temps n'est pas considérable (4 et 5 minutes de plus). Grâce au sondage préliminaire, nous avons remarqué que le participant 5, qui a employé le moins de temps dans la post-édition avec la reconnaissance vocale, avait déjà des connaissances préalables de *Dragon Naturally Speaking*, ce qui peut expliquer le gain de temps majeur et aussi une certaine familiarité déjà acquise. Le gain de temps du participant 4 a été de cinq minutes, bien qu'il n'ait pas de connaissances préalables dans le logiciel, mais des connaissances dans le domaine des technologies de la traduction, ce qui peut aussi avoir un impact. Le participant 3, quant à lui, a employé le même temps dans les deux phases, malgré aucune connaissance déclarée lors du sondage préliminaire en ce qui concerne DNS. Certes, les variables d'intuitivité et de prédisposition aux technologies peuvent aussi influencer le temps employé.

Ces résultats nous permettent donc de confirmer notre hypothèse selon laquelle un traducteur qui possède une bonne connaissance et une bonne maîtrise du logiciel, gagnera du temps grâce à la reconnaissance vocale. Certes, l'apprentissage demandera un certain investissement, mais cette phase a été très rapide dans le cadre de notre étude.

Toutefois, le ressenti des participants ne correspond pas aux résultats et aux données issues des enregistrements. En effet, les participants s'accordent à dire qu'ils ont employé moins de temps pour la correction à l'aide des méthodes traditionnelles d'entrée. Deux d'entre eux estiment que cela est dû à une mauvaise connaissance de la reconnaissance vocale qui, avec un bon degré de connaissance, pourrait se révéler utile et rendre le travail plus rapide. De plus, tous pensent s'être habitués aux commandes vocales, bien qu'un entraînement et un apprentissage plus long

et plus ciblé auraient été souhaitables. L'un d'eux estime en outre que cette technologie ne permet pas un gain de temps majeur par rapport à des fonctionnalités telles que « rechercher et remplacer » ou des expressions régulières (voir section 8.1.1).

## 8.2 Satisfaction

Nous avons calculé la satisfaction sur la base des réponses données par les participants au questionnaire de satisfaction. Nous avons décomposé la caractéristique de la satisfaction en sous-caractéristiques, en nous appuyant sur les normes ISO/IEC 25010 :2011 et nous avons sélectionné l'utilité, le plaisir et le confort.

L'utilité a été évaluée par rapport aux fonctionnalités pouvant être exploitées durant l'expérience, telle que la dictée, l'exécution des commandes vocales ou la correction. Les questions concernant le confort étaient plutôt liées à l'utilisation de la reconnaissance vocale et des méthodes traditionnelles d'entrée. Enfin, nous avons demandé aux participants de donner un score au divertissement ressenti lors de l'expérience d'évaluer le plaisir d'utilisation.

Comme expliqué à la section 6.3, nous avons attribué les points selon les schémas précédents pour l'utilité et le confort. Nous avons également demandé aux participants d'évaluer le plaisir sur une échelle de 0 à 3. Dans le *Tableau 20* suivant, nous avons résumé les moyennes pour chaque sous-caractéristique selon les tâches, et ce en fonction de l'utilisation soit de la reconnaissance vocale, soit des méthodes traditionnelles d'entrée.

| **Caractéristique** | **Sous-caractéristique** | **Tâche/fonctionnalité** | **RV (Moyenne de 5 participants)** | **Entrée trad (Moyenne de 5 participants)** |
|---|---|---|---|---|
| | Utilité | **Dicter correctement** | **2/3** | |
| | | **Exécuter commandes correctement** | **1,5/3** | |
| | | **Corriger facilement** | **1,5/3** | **3/3** |
| | | **Correction autonome** | **2,5/3** | **3/3** |
| | Confort | **Ergonomie de la dictée** | **3/3** | |

| | | | | |
|---|---|---|---|---|
| | | **Ergonomie des inputs traditionnels** | | **3/3** |
| | | **Ergonomie des commandes vocales** | **2/3** | |
| | Plaisir | **Divertissement dans l'utilisation de la RV durant l'expérience** | **1,8/3** | |
| | Moyenne totale | | 2,04/3 | 3/3 |
| | **Moyenne en pourcentage** | | **68%** | **100%** |

*Tableau 24 : Moyennes des scores du questionnaire de satisfaction*

On constate que la satisfaction liée à la reconnaissance vocale est légèrement inférieure à celle des méthodes traditionnelles d'entrée. Notamment en ce qui concerne l'exécution des commandes et la facilité de la correction, nous remarquons une légère insatisfaction générale (1,5 de moyenne), ainsi que pour l'exécution des commandes vocales. Néanmoins, il faut souligner que l'usage et l'expérience sont deux facteurs qui facilitent l'exécution en peu de tentatives des commandes vocales ou de la dictée, éléments qui faisaient défaut à nos participants. Aussi, les résultats peuvent être considérés comme acceptables, avec une marge de progression importante.

En revanche, l'ergonomie de la dictée et des commandes vocales s'est avérée globalement satisfaisante, en comparaison avec les méthodes traditionnelles d'entrée. Nous pouvons constater que ces deux éléments ont obtenu un jugement « confortable ». La moyenne totale est de 3 sur 3 pour la satisfaction liée aux méthodes traditionnelles d'entrée. Il faut toutefois souligner que cette moyenne n'est calculée qu'à l'aide de trois données, facteur qui pourrait biaiser légèrement la moyenne. Peut-on parler de moyenne indicative lorsqu'elle est basée sur trois valeurs ? Cependant, il est certain aussi que notre intérêt se base plutôt sur la reconnaissance vocale, raison pour laquelle les données relatives à celle-ci sont plus nombreuses ainsi que les valeurs employées pour calculer la moyenne, ce qui la rend probablement plus indicative.

Nous pouvons donc conclure que, contrairement à l'expérience effectuée par Mesa-Lao (2014) mentionnée dans les sections précédentes, nos participants n'ont pas éprouvé beaucoup de plaisir à effectuer la tâche, et ils ne l'ont pas trouvée particulièrement utile. À noter que

l'efficacité a été bien évaluée, ainsi que la satisfaction des participants, ce qui confirme notre hypothèse : la reconnaissance vocale peut être un atout si elle est utilisée consciemment et avec une formation préalable adéquate. Enfin, la professionnalité des systèmes, en particulier de *Dragon Naturally Speaking*, a été impeccable selon nos participants. Cet aspect est certainement dû aux nouvelles technologies d'apprentissage profond qui sont à la base de la version 15 utilisée lors de notre étude.

Par ailleurs, l'éventuelle distraction que peut susciter la reconnaissance vocale durant la correction représente également un élément important. Nous avons demandé aux participants s'ils pensaient avoir été distraits par la RV. Trois d'entre eux estiment ne pas avoir été distraits par cette technologie. Toutefois l'un d'eux a ajouté des erreurs à cause de l'utilisation de la dictée, et ne s'est pas s'aperçu des espaces en trop ou du changement d'articles des substantifs corrigés. Deux estiment que la RV les a distraits parce qu'elle demandait une concentration élevée sur les commandes vocales. Les commandes vocales telles que « *seleziona* ***questo*** » (« sélectionner ceci ») ou « *annulla* ***questo*** » (« annule ceci ») ont particulièrement posé problème aux participants. Effectivement, le fait de devoir spécifier « questo » (« cela » en français) reste un peu artificiel dans l'oralité de la langue italienne, mais cette formule est nécessaire pour que logiciel puisse distinguer la commande vocale souhaitée. Tous les participants s'accordent à dire que cette méthode peut poser certains problèmes surtout si l'on oublie de dire « questo ».

Nous souhaitions également recueillir l'opinion de nos participants en ce qui concerne l'utilisation de cette technologie dans un contexte professionnel : deux affirment qu'ils seraient prêts à l'utiliser, mais seulement pour la rédaction d'un texte entier (et pas pour apporter des modifications ou des corrections) ; l'un d'eux estime que dans un contexte d'entreprise, cela pourrait s'avérer problématique pour les collègues, mais que pour un freelance, cette technologie pourrait être prise en considération ; deux estiment qu'ils n'excluraient pas son utilisation, mais seulement après une préparation adéquate.

Enfin, nous les avons questionnés au sujet de *Dragon Naturally Speaking*, qui s'est révélé un excellent système de reconnaissance vocale. Les participants n'ont jamais dû utiliser le clavier pour écrire des mots qui n'ont pas été reconnus, même dans le cas de mots étrangers et de termes du domaine (entre autres grâce à la spécialisation effectuée). Les participants ont répondu que le système a presque toujours identifié les commandes vocales et toujours les mots dictés. Cette performance peut être considérée comme un excellent résultat, probablement due à

l’implémentation de l’apprentissage profond dans la version du logiciel utilisée pour l’expérience. Les cas qui ont posé problème peuvent aussi être dû au manque d’expérience des participants. Ainsi, le problème le plus commun, soit déplacer le curseur de la souris, avec la commande « griglia del mouse », peut s’expliquer par le manque de pratique des participants et par la nature de la commande qui est intuitive et pratique. En outre, les participants ne se montrent pas satisfaits de l’aide apportée par la reconnaissance vocale pour la post-édition.

# 9 Conclusion

Dans le cadre de ce mémoire, nous visons à observer l'impact de la reconnaissance automatique de la parole sur la tâche de post-édition effectuée par un traducteur italophone.

Pour ce faire, nous avons passé en revue des études conduites sur le même sujet, malgré leur faible nombre. À partir de ces études, nous avons choisi d'effectuer une expérience directe afin d'observer le véritable impact de cette technologie. Il en est ressorti que la reconnaissance vocale pourrait avoir un grand potentiel, notamment grâce à l'efficacité et à la performance des systèmes actuels de reconnaissance vocale, majoritairement de type neuronal, spécialement si les utilisateurs possèdent les compétences adéquates.

Notre expérience comptait cinq participants, jeunes traducteurs avec la même formation. D'abord nous leur avons soumis un questionnaire préliminaire afin de déterminer leur profil.

Les traducteurs étaient tous de langue maternelle italienne et avaient le français parmi leurs langues passives. Ils avaient tous des connaissances préalables concernant les outils de traduction assistée (en l'espèce SDL Trados Studio), mais seuls deux d'entre eux avaient déjà utilisé la reconnaissance vocale lors d'un cours universitaire.

L'expérience, quant à elle, consistait à post-éditer deux textes techniques issus du même mode d'emploi en deux phases. Une phase devait s'effectuer à l'aide des méthodes traditionnelles d'entrée, tandis que l'autre requérait l'utilisation du logiciel de reconnaissance vocale Dragon Naturally Speaking. Leur activité complète a été enregistrée de sorte à garder une trace du temps employé et à pouvoir visionner les enregistrements afin de comprendre les éventuelles erreurs ou attitudes particulières.

Dans le *Tableau 21,* nous avons résumé les résultats globaux, concernant l'efficacité, le temps et la satisfaction relatifs à l'utilisation avec la reconnaissance vocale. La qualité, étant difficile à résumer avec un pourcentage, sera prise en compte séparément.

| EFFICACITÉ | MOYENNE TEMPS (min.) | SATISFACTION |
|---|---|---|
| 81,1% | 19,4 | 68% |

*Tableau 25: Récapitulatif global des moyennes des résultats relatifs à la RV*

Il est important de spécifier que nous avons calculé le pourcentage, extrait des résultats déjà évoqués dans les sections respectives, pour l'efficacité et la satisfaction. Pour le temps, nous avons pris en considération la moyenne. Les moyennes présentées concernent la reconnaissance

vocale, car il s'agit de notre sujet d'étude principal et que, notamment pour l'efficacité et la satisfaction, les résultats obtenus pour les méthodes traditionnelles d'entrée s'avèrent presque excellents. Cette différence de résultats pour ces deux caractéristiques était prévisible puisque nos participants n'avaient presque aucune connaissance dans le domaine de la reconnaissance vocale et ont l'habitude d'écrire leurs corrections. Ce facteur exerce une forte influence sur l'efficacité : il se base notamment sur les tâches à compléter et les erreurs présentes, ce qui demeure plus simple à effectuer à l'aide d'une méthode déjà connue et habituelle. Le même constat peut être dressé pour la satisfaction, puisqu'il est normal que les participants soient plus satisfaits par une manière de travailler qu'ils connaissent et qu'ils maîtrisent. Il faut souligner que les résultats concernant le temps indiquent bel et bien un gain même si ce dernier est léger (19,4 minutes avec la RV contre 20,2 minutes avec les méthodes traditionnelles d'entrée, voir *Tableau* 19, page 6565). Dans ce cas aussi, il faut prendre en compte qu'il s'agissait d'une nouveauté pour la plupart des participants. Il est donc probable que ce gain aurait pu être plus important si ces derniers possédaient des connaissances préalables ou plus détaillées dans le domaine.

Les *Tableaux 26* et *27* montrent les scores comparés relatifs au temps et à la satisfaction. Le temps est donné par sa moyenne, tandis que la satisfaction est montrée en pourcentage.

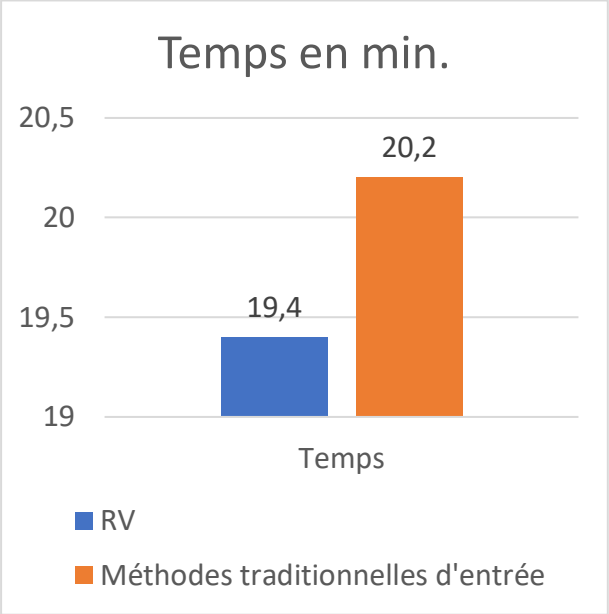



*Tableau 26: Moyenne du temps de deux méthodes d'entrée*

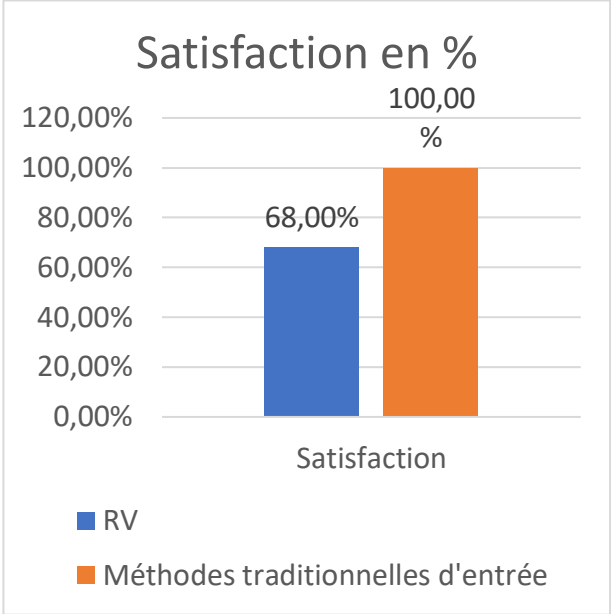



*Tableau 27: Moyenne en pourcentage de la satisfaction de deux méthodes d'entrée*

Enfin, en ce qui concerne la qualité de la traduction, elle ne semble pas être affectée de manière négative : si l'on se base sur l'accord kappa, donc sur une évaluation humaine de la qualité,

nous pouvons remarquer un accord plus important pour les textes post-édités à l'aide de la reconnaissance vocale. L'évaluation automatique, elle aussi, fournit de meilleurs résultats dans trois cas sur quatre (exception faite pour BLEU pour le texte 2). En observant l'écart issu des deux scores pour les deux textes, nous remarquons un écart avantageux pour BLEU pour ce qui est du texte 1 post-édité à l'aide de la reconnaissance vocale (1,132), bien que cette différence ne soit pas si marquée. Le même s'applique pour TER des deux textes avec un écart de 0,038 et 0,017 ; dans les deux cas, il s'agit d'une différence minime. Nous ne pouvons en dire autant de BLEU du texte 2, où l'écart est légèrement plus grand (0,326). Certes, nous sommes consciente que ces différences sont relativement faible, mais elles représentent un bon point de départ pour de nouvelles améliorations.

Il ressort différents avantages et aspects positifs de notre expérience. D'abord, prononcer une phrase à voix haute pourrait se révéler plus efficace que de la taper, puisqu'il est généralement plus simple de repérer les erreurs ou les répétitions oralement. En outre, dicter une phrase pourrait aussi impliquer une réflexion plus aboutie : le traducteur aura terminé son raisonnement avant de prononcer la traduction. À noter que cet avantage concerne la traduction plutôt que la révision. Cette technique pourrait donc impacter de manière positive la qualité de la traduction finale, malgré peut-être un emploi plus consistant du temps afin de développer un raisonnement plus précis.

Enfin, comme l'expérience et l'opinion des participants l'ont démontré, les difficultés présentes au début de l'expérience se sont évaporées. Nous pouvons donc conclure qu'avec un entraînement plus long et approprié, un usage quotidien, ainsi qu'une formation et des connaissances plus profondes, un système de reconnaissance vocale peut s'avérer utile et efficace pour la productivité au travail. Ce système permet de gagner du temps, sans que la qualité soit forcément affectée.

Les seuls aspects négatifs, évoqués aussi par les participants dans le questionnaire de post-évaluation, seraient la difficulté de l'usage d'un système de reconnaissance vocale dans un contexte d'entreprise où il pourrait déranger les autres employés, notamment dans des environnements « open-space ». Et ce même s'il est assez fréquent de faire et de recevoir des appels internes à l'entreprise, sans que cela ne dérange considérablement les autres travailleurs. Ce problème, ne concernerait pas les traducteurs indépendants, qui pourraient donc davantage en profiter davantage. L'autre note négative évoquée concerne l'utilisation de la RV pour la

post-édition. Les participants estiment qu'il serait plus judicieux d'y avoir recours pour la rédaction d'un texte en entier.

## 9.1 Perspectives et limites

Les données collectées dans le cadre de cette étude sont limitées par le nombre réduit de participants (cinq au total) ainsi que l'échantillonnage des textes courts pour éviter de rallonger l'expérience. Ce dernier point a probablement eu un impact sur la correction faite à l'aide de la reconnaissance vocale, étant donné que les traducteurs venaient de corriger les mêmes textes avec les méthodes traditionnelles d'entrée. Malgré l'utilisation de la *cross-case analysis* le nombre de texte était réduit, ainsi que celui des phrases à corriger, ce qui peut avoir biaisé les corrections des participants. Le temps aussi était assez réduit : ils ont effectué les tâches en seulement une heure à peu près. De plus, l'entraînement à l'aide Dragon Naturally Speaking a été très bref (environ 5-10 minutes), ce qui n'a permis qu'une connaissance très basique de l'utilisation du microphone pour dicter et exécuter les commandes vocales. Ainsi, cette situation a certainement influencé le temps employé, étant donné que les participants n'avaient pas l'habitude de se déplacer ou d'effectuer les tâches et d'écrire à l'aide de leur voix, sans utiliser la souris et le clavier.

Par conséquent, nous estimons que cette même étude mériterait davantage de ressources, ce qui permettrait d'obtenir des résultats importants qui contribueraient au développement des technologies de la traduction comme la RAP. Le sujet pourrait être étudié à plus grande échelle afin d'avoir des résultats plus clairs, basés notamment sur des échantillons plus vastes qui pourraient donner une perspective plus grande et peut-être plus réalistique de la situation actuelle dans ce domaine.

Néanmoins, nous ne prétendions ni obtenir des résultats parfaits ni proposer des conclusions qui servent de références. Nous souhaitions nous pencher sur la question de l'intégration de la reconnaissance vocale dans le travail habituel du traducteur, en l'espèce lors de la post-édition.

Nous sommes tout de même satisfaite de constater que, avec les bonnes connaissances, la reconnaissance vocale peut se révéler un véritable atout dans le travail du traducteur, permettant de gagner du temps et donc d'améliorer sa productivité.

# 10 Bibliographie


Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4277-4280. https://doi.org/10.1109/ICASSP.2012.6288864

Araki, S., Fujimoto, M., Yoshioka, T., Delcroix, M., Espi, M., & Nakatani, T. (2015). *Deep Learning Based Distant-talking Speech Processing in Real-world Sound Environments*. *13*(11), 6.

Bassil, Y., & Alwani, M. (2012). Post-Editing Error Correction Algorithm For Speech Recognition using Bing Spelling Suggestion. *International Journal of Advanced Computer Science and Applications*, *3*(2). https://doi.org/10.14569/IJACSA.2012.030217

Bouillon, Cervini, Rayner. (2016). Translation and technology. The case of translation games for language learning. In *The routledge handbook of language learning and technology* (p. 536-549). Routledge. https://archive-ouverte.unige.ch/unige:82377

Bouillon, P. (Éd.). (1993). *La traductique : Études et recherches de traduction par ordinateur*. Les Presses de l'Université de Montréal.

Bouillon, P. (2017a). *Notes de cours Ingénierie linguistique*.

Bouillon, P. (2017b). *Notes de cours Traduction automatique*. Université de Genève.

Bouillon, P. (2019a). *Note du cours Traduction automatique 2*.

Bouillon, P. (2019b). *Notes de cours Ingénierie Linguistique*.

Cer, D., Manning, C. D., & Jurafsky, D. (2010). *The best lexical metric for phrase-based statistical MT system optimization*. 9.

Ciobanu, D. (2014). Of Dragons and Speech Recognition Wizards and Apprentices. *Tradumàtica: Tecnologies de La Traducció*, *12*, 524. https://doi.org/10.5565/rev/tradumatica.71

Coughlin, D. (2003). *Correlating Automated and Human Assessments of Machine Translation Quality*. 8.

Dragsted, B., Mees, I. M., & Hansen, I. G. (2011). Speaking your translation : Students' first encounter with speech rcognition technology. *Translation & Interpreting*, *3*(1), 10-43-43. https://doi.org/10.12807/t&i.v3i1.115

Duarte, T., Prikladnicki, R., Calefato, F., & Lanubile, F. (2014). Speech Recognition for Voice-Based Machine Translation. *IEEE Software*, *31*(1), 26-31. https://doi.org/10.1109/MS.2014.14

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291-309. https://doi.org/10.1075/ts.6.2.06for

Garcia-Martinez, M., Singla, K., Tammewar, A., Mesa-Lao, B., Thakur, A., Anusuya, M. A., Banglore, S., & Carl, M. (2014). *SEECAT: Speech Eye- tracking Enabled Computer Assisted Translation*. 9.

Gouadec, D. (2009). *Profession : Traducteur : APE748F alias ingénieur en communication multilingue (et) multilmédia* ([2e édition].). La Maison du dictionnaire.

Guzmán, F., Abdelali, A., Temnikova, I., Sajjad, H., & Vogel, S. (2015). How do Humans Evaluate Machine Translation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 457–466. https://doi.org/10.18653/v1/W15-3059

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data : An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Haton, J.-P., Cerisara, C., Fohr, D., Laprie, Y., & Smaïli, K. (2006). *Reconnaissance automatique de la parole : Du signal à son interprétation*. Dunod.

Hutchins, W. J., & Somers, H. L. (1997). *An introduction to machine translation* (2. printing). Academic Press.

Introducing Translatotron : An End-to-End Speech-to-Speech Translation Model. (2019, mai). *Google AI Blog*. http://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html

*ISO/IEC 25010:2011(en), Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models*. (s. d.). Consulté 19 août 2019, à l'adresse https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en

Izbassarova, A., Duisembay, A., & James, A. P. (2020). Speech Recognition Application Using Deep Learning Neural Network. In A. P. James (Éd.), *Deep Learning Classifiers with Memristive Networks* (Vol. 14, p. 69-79). Springer International Publishing. https://doi.org/10.1007/978-3-030-14524-8_5

Jiménez Ivars, M. A. (1998). *La traducción a la vista. Un análisis descriptivo*.

Jurafsky, Dan, & Martin, J. H. (2009). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (2. ed., Pearson internat. ed). Prentice Hall, Pearson Education Internat.

Jurafsky, Daniel. (2009). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed., Pearson international edition.). Prentice Hall.

Kamal, I., Bae, H., Sunghyun, S., Kim, H., Kim, D., Choi, Y., & Yun, H. (2019). Forecasting High-dimensional Multivariate Regression of Baltic Dry Index (BDI) Using Deep

Neural Networks (DNN). *ICIC Express Letters*, *13*, 427-434. https://doi.org/10.24507/icicel.13.05.427

KIT, C. Y., & WONG, B. T.-M. (2014). Evaluation in machine translation and computer-aided translation. In Sin-Wai Chan (Éd.), *Routledge Encyclopedia of Translation Technology* (p. 213–236). Routledge.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Koehn, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. *Proceedings on the Workshop on Statistical Machine Translation*, 102–121. https://www.aclweb.org/anthology/W06-3114

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

L'Homme, M.-C. (2008). *Initiation à la traductique* (2e éd. rev. et augm). Linguatech.

Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 605-es. https://doi.org/10.3115/1218955.1219032

Liyanapathirana, J., Bouillon, P., & Mesa-Lao, B. (2019). Surveying the potential of using speech technologies for post-editing purposes in the context of international organizations : What do professional translators think? *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, 149–158. https://www.aclweb.org/anthology/W19-6728

Lossner, K. (2018, décembre). Integrated iOS speech recognition in memoQ 8.7. *Translation Tribulations*. https://www.translationtribulations.com/2018/12/integrated-ios-speech-recognition-in.html

McCulloch, W. S., & Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*(1), 99-115. https://doi.org/10.1007/BF02459570

Mesa-Lao, B. (2014). Speech-Enabled Computer-Aided Translation : A Satisfaction Survey with Post-Editor Trainees. *Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation*, 99-103. https://doi.org/10.3115/v1/W14-0315

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks : A Systematic Review. *IEEE Access*, *7*, 19143-19165. https://doi.org/10.1109/ACCESS.2019.2896880

Nolla, F. C., & Abril, Á. P. (2017). Traducció automàtica neuronal. *Revista Tradumàtica: tecnologies de la traducció*, *0*(15), 66-74. https://doi.org/10.5565/rev/tradumatica.203

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, *25*(3), 197-215. JSTOR. www.jstor.org/stable/41487494

Organisation internationale de normalisation, & Commission électrotechnique internationale. (2016). *Systems and software engineering—Systems and software quality requirements and evaluation (SQuaRE)—Measurement of system and software product quality = Ingénierie des systèmes et du logiciel—Exigences de qualité et évaluation des systèmes et du logiciel (SQuaRE)—Mesurage de la qualité du produit logiciel et du système* (First edition 2016-06-15.). ISO/IEC. http://data.rero.ch/01-R008630338/html?view=GE_V1

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

Pym, A., Grin, F., Sfreddo, C., & Chan, A. L. J. (2013). *The Status of the Translation Profession in the European Union*. Anthem Press.

Rayner, E. (2006). *Putting Linguistics into Speech Recognition*. CSLI; Archive ouverte UNIGE. http://archive-ouverte.unige.ch/unige:3482

Saint-André, L. (2015). *Quelle formation donner aux traducteurs-postéditeurs de demain?* 236.

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies*. St. Jerome Publishing.

Salimbajevs, A., & Ikauniece, I. (2017). *System for Speech Transcription and Post-Editing in Microsoft Word*. 2.

Screen, B. (2019). JOSTRANS : JOURNAL OF SPECIALISED TRANSLATION - 1740-357X | MIAR 2019 live. Information Matrix for the Analysis of Journals. *What Effect Does Post-Editing Have on the Translation Product from an End-User's Perspective?*, *31*, 133-157. http://miar.ub.edu/issn/1740-357X

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. 9.

Somers, H. L. (Éd.). (2003). *Computers and translation : A translator's guide*. John Benjamins Pub. Co.

Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, *24*(1), 39-50. https://doi.org/10.1007/s10590-010-9077-2

Starlander, M. (2016). *Méta-évaluation de la traduction automatique de la parole (TAP) dans le domaine médical*.

*The EAGLES 7-step recipe*. (s. d.). Consulté 14 octobre 2019, à l'adresse https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html

Vidal, E., Casacuberta, F., Rodriguez, L., Civera, J., & Hinarejos, C. D. M. (2006). Computer-assisted translation using speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(3), 941-951. https://doi.org/10.1109/TSA.2005.857788

Volkart, L. (2018). *Traduction automatique statistique vs. neuronale : Comparaison de MTH et DeepL à La Poste Suisse* [University of Geneva]. https://archive-ouverte.unige.ch/unige:113749

Waibel, A., Hanazawa, T., Hinton, G., & Ic, I. S. (1988). Phoneme Recognition : Neural Networks vs. Hidden Markov models. *IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, 107110.

Xie, Y., Le, L., Zhou, Y., & Raghavan, V. V. (2018). Chapter 10—Deep Learning for Natural Language Processing. In V. N. Gudivada & C. R. Rao (Éd.), *Handbook of Statistics* (Vol. 38, p. 317-328). Elsevier. https://doi.org/10.1016/bs.host.2018.05.001

Zapata, J., & Quirion, J. (2016). La traduction dictée interactive et sa nécessaire intégration à la formation des traducteurs. *Babel*, *62*(4), 531-551. https://doi.org/10.1075/babel.62.4.01zap

# 11 Webographie

« Hey MemoQ » dictations support for MemoQ users :https://www.memoq.com/products/hey-memoq consulté le 08/01/2020

Demonstration of multilingual TTS, memoQ and DNS integration : https://amara.org/it/videos/J3p2jHYnOyoN/info/demonstration-of-multilingual-tts-memoq-and-dns-integration/ consulté le 08/01/2020

Dragon Professional Individual : https://www.nuance.com/dragon/business-solutions/dragon-professional-individual.html#standardpage-mainpar_backgroundimage, consulté le : 08/05/2019

FAQ « Hey MemoQ » : https://blog.memoq.com/hey-memoq-frequently-asked-questions consulté le 08/01/2020

Flashback Express : https://www.flashbackrecorder.com/it/express/, consulté le : 12/05/2019

Forum Microsoft : https://answers.microsoft.com/en-us/windows/forum/all/what-languages-does-microsoft-speech-recognition/cb5eaf9d-7391-4ab5-8ce9-f1e44096c853 consulté le 06/01/2020

Interactive BLEU score evaluator : https://www.letsmt.eu/Bleu.aspx, consulté le 14/11/2019

Introducing Translatotron: An End-to-End Speech-to-Speech Translation Model: https://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html consulté le 08/01/2020

Omniscien: « The State of Neural Machine Translation (NMT) » by Philipp Koehn: https://omniscien.com/state-neural-machine-translation-nmt/ consulté le 15/01/2020

Site du projet EAGLES : https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html consulté le 14/10/19

Speech Technology « Nuance Adds Deep Learning to Dragon » :https://www.speechtechmag.com/Articles/Editorial/FYI/Nuance-Adds-Deep-Learning-to-Dragon-114607.aspx consulté le 15/01/2020

Support Microsoft « Dettare testo con il riconoscimento vocale »: https://support.microsoft.com/it-it/help/14198/windows-7-dictate-text-using-speech-recognition consulté le 06/01/2020

Support Microsoft « How to use Speech Recognition in Windows »: https://support.microsoft.com/en-us/help/14213/windows-how-to-use-speech-recognition consulté le 06/01/2020

Swisscom, Swisscom TV Voice assistant « Hey Swisscom » :https://www.swisscom.ch/it/clienti-privati/abbonamenti-tariffe/inone-home/swisscom-tv.html consulté le 06/01/2020

Translation tribulations « Integrated iOS speech recognition in MemoQ 8.7 »:https://www.translationtribulations.com/2018/12/integrated-ios-speech-recognition-in.html consulté le 08/01/2020

# Annexes

## 1) Sondage préliminaire

### Sondage préliminaire

1. Conosci SDL Trados Studio?
2. Se sì, lo hai mai usato? In che occasione?
3. Come valuteresti le tue conoscenze di questo CAT-tool? (Es: Ottime, buone, discrete, sufficienti, di base…)
4. Conosci Dragon Naturally Speaking?
5. Se sì, lo hai mai usato? In che occasione?
6. Come valuteresti le tue conoscenze di questo programma? (Es: Ottime, buone, discrete, sufficienti, di base…)
7. Ti è già capitato di usare sistemi di riconoscimento vocale?
8. Se sì quali e con che frequenza?
9. Hai mai sentito parlare del Post-editing?
10. Lo hai mai messo in pratica? Se sì in che occasione?
11. Cosa pensi del post-editing? Lo ritieni utile per l'attività del traduttore?

## 2) Questionnaire de post-évaluation

**Questionario post valutazione**

Lo scopo di tale questionario è comprendere meglio i risultati derivanti dalla valutazione effettuata e poter avere un riscontro riguardo all'opinione di ogni partecipante. A questo proposito, ti invito a rispondere nel modo più esaustivo possibile. Non ci sono risposte giuste né sbagliate, pertanto sentiti libero/a di esprimere tranquillamente la tua opinione personale. Grazie.

1. Sapere che il testo target fosse frutto della traduzione automatica di *DeepL* ha influenzato il tuo modo di post-editare?
2. Utilizzare il riconoscimento vocale (RV) ha influenzato le tue correzioni? Tendevi a correggere meno rispetto alla fase con i sistemi di input tradizionali per evitare di dover usare i comandi vocali?
3. Ritieni di aver apportato tutte le dovute correzioni a entrambi i testi e in egual modo?
4. Come descriveresti la qualità della traduzione finale da te effettuata con il RV? In generale e rispetto a quella svolta con gli input standard.
5. Alla fine della valutazione avevi le stesse difficoltà rispetto all'inizio?
6. Hai avuto l'impressione di aver risparmiato tempo usando il RV?
7. Ritieni che il RV ti "abbia distratto" in qualche modo dal concentrarti sul testo da correggere? (Per es. dopo aver modificato/inserito correttamente la parola desiderata eri soddisfatto/a per essere riuscito/a a correggerla utilizzando il riconoscimento vocale e hai "perso di vista" il resto del testo?)
8. Hai avuto difficoltà a usare alcuni comandi vocali? (Es. "seleziona questo", "annulla questo", "correggi", "digita", "sposta di *x* caratteri a sinistra/destra", ecc.)
9. Ritieni di aver acquisito dimestichezza con tali comandi?
10. Ritieni che con una maggiore pratica sarebbe più produttivo utilizzare i comandi vocali invece degli input tradizionali? Perché?
11. Ritieni che la qualità della traduzione subisca un impatto durante l'utilizzo del riconoscimento vocale? Se sì, di che tipo?
12. Pensi che il riconoscimento vocale ti abbia aiutato durante l'attività di post-editing? Se sì, in cosa? E se no, perché?
13. Utilizzeresti il riconoscimento vocale in un contesto professionale?

14. Come giudicheresti il sistema di riconoscimento vocale da te utilizzato? Nella tabella sottostante, seleziona un avverbio tra quelli proposti. (Per poter inserire il valore nella seconda colonna cliccare due volte sulla cella, verrà visualizzata la schermata di Excel verde in alto e solo allora potrai inserire il valore desiderato)

| | | |
|---|---|---|
| Riconosce correttamente le parole dettate | | |
| La correzione avviene in modo semplice e immediato | | |
| Riconosce i comandi vocali | | |
| Riconosce le parole straniere | | |
| Riconosce termini inerenti al testo (Smartwatch, Fitbit, Today, Thumbnail, ecc.) | | |
| È sensibile ai rumori circostanti | | |
| Il sistema non ha riconosciuto il comando vocale da me dettato e ho dovuto usare il mouse e/o la tastiera | | |
| Il sistema non ha riconosciuto la parola da me dettata e ho dovuto scriverla usando la tastiera | | |
| Non sono riuscito/a a posizionare il cursore dove volevo con il RV e ho dovuto usare il mouse | | |
| Sono riuscito/a a trovare soluzioni alternative quando mi si è presentato un "ostacolo" con il sistema di RV (Ad es. comandi non scritti nella tabella riassuntiva per spostare il cursore o inserire parole) | | |

15. Nella tabella sottostante, indica su una scala di valori (da un minimo di 1 a un massimo di 5) che valore attribuiresti ai seguenti indicatori. (Per poter inserire il valore nella seconda colonna cliccare due volte sulla cella, verrà visualizzata l'intestazione di Excel verde in alto e solo allora potrai inserire il valore desiderato)

| Descrizione | Valutazione | |
|---|---|---|
| Divertimento | | |
| Utilità | | |
| Efficacia | | |
| Soddisfazione | | |
| Qualità (della traduzione) | | |
| Professionalità (degli strumenti utilizzati) | | |

## 3) Questionnaire de satisfaction

### Questionnaire de satisfaction

La satisfaction par rapport aux tâches a été évalué selon une échelle composée de 5 valeurs : tâche impossible, pas du tout satisfait, peu satisfait, satisfait, très satisfait. Nous avons attribué des points à chaque valeur selon le schéma suivant :

- pas du tout satisfait : 0
- peu satisfait : 1
- satisfait : 2
- très satisfait : 3

Nous avons considéré toutes les tâches comme étant également importantes.

En revanche, le confort a été évalué selon une échelle composée de 4 valeurs (pas du tout confortable, peu confortable, confortable, très confortable) et nous avons attribué des points à chaque valeur selon le schéma suivant :

- pas du tout confortable : 0
- peu confortable : 1
- confortable : 2
- très confortable : 3

| **Caractéristique** | **Sous-caractéristique** | **Tâche/fonctionnalité** | **RV** | **Input trad** |
|---|---|---|---|---|
| Satisfaction | Satisfaction par rapport aux tâches | **Dicter correctement** (*la phrase est écrite correctement*) | | |
| | | **Exécuter commandes correctement** (*la commande est exécutée en max. 2 tentatives*) | | |
| | | **Corriger facilement** (*la correction est* | | |

| | | | | |
|---|---|---|---|---|
| | | *effectuée en max 2-3 tentatives*) | | |
| | | **Correction autonome** (*en regardant le manuel ou en essayant commandes pas écrites*) | | |
| | | | | |
| | Confort | **Ergonomie de la dictée** | | |
| | | **Ergonomie des inputs traditionnels** | | |
| | | **Ergonomie des commandes vocales** | | |

## 4) Texte pour l'entraînement

| Segment ID | Source segment | Target segment |
|---|---|---|
| 1 | Ajouter de la musique et des podcasts avec votre Mac | Aggiungi musica e podcast con il tuo Mac |
| 2 | Télécharger des listes de lecture de votre musique personnelle et des podcasts depuis votre bibliothèque iTunes sur votre Versa. | Scarica le playlist della tua musica personale e dei podcast dalla tua libreria iTunes al tuo Versa. |
| 3 | Créer une liste de lecture | Creare una playlist |
| 4 | Dans iTunes, créez au moins 1 liste de lecture de chansons ou de podcasts à télécharger sur votre montre. | In iTunes, crea almeno 1 playlist di canzoni o podcast da scaricare sul tuo orologio. |
| 5 | Assurez-vous d'autoriser l'application iTunes à partager des listes de lecture avec votre montre : | Assicurarsi di consentire all'applicazione iTunes di condividere le playlist con l'orologio: |
| 6 | Ouvrez iTunes sur votre ordinateur > Modifier > Préférences > Avancé > Partager la bibliothèque iTunes XML avec d'autres applications > OK. | Apri iTunes sul tuo computer > Modifica > Modifica > Preferenze > Avanzate > Condividi la libreria iTunes XML con altre applicazioni > OK. |
| 7 | Connexion au Wi-Fi | Connessione Wi-Fi |
| 8 | Assurez-vous que votre Versa et votre Mac peuvent se connecter au même réseau Wi-Fi : | Assicurati che Versa e Mac possano connettersi alla stessa rete Wi-Fi: |
| 9 | 1. | 1. |
| 10 | Dans le tableau de bord de l'application Fitbit, appuyez ou cliquez sur l'icône Compte > vignette Versa. | Nel cruscotto dell'applicazione Fitbit, premere o fare clic sull'icona Account > Versa thumbnail. |
| 11 | 2. | 2. |
| 12 | Appuyez ou cliquez sur Paramètres du Wi-Fi. | Toccare o fare clic su Impostazioni Wi-Fi. |
| 13 | 3. | 3. |

| | | |
|---|---|---|
| 14 | Tapez ou cliquez sur Ajouter un réseau et suivez les instructions qui s'affichent à l'écran pour ajouter votre réseau Wi-Fi ou vérifiez la liste des réseaux pour vous assurer qu'il y figure déjà. | Digitare o fare clic su Aggiungi rete e seguire le istruzioni sullo schermo per aggiungere la rete Wi-Fi o controllare l'elenco delle reti per assicurarsi che sia già presente nell'elenco. |
| 15 | 4. | 4. |
| 16 | Appuyez sur le nom du réseau > Se connecter. | Toccare il nome della rete > Connetti. |
| 17 | 5. | 5. |
| 18 | Pour afficher le réseau auquel votre ordinateur est connecté, appuyez ou cliquez sur le symbole du Wi-Fi sur votre écran. | Per visualizzare la rete a cui è collegato il computer, premere o fare clic sul simbolo Wi-Fi sullo schermo. |
| 19 | Connectez-vous au même réseau Wi-Fi que votre montre. | Connettersi alla stessa rete Wi-Fi dell'orologio. |
| 20 | Remarque : si votre réseau Wi-Fi nécessite que vous vous connectiez par le biais d'un navigateur, cela n'est pas pris en charge sur Versa. | Nota: se la rete Wi-Fi richiede la connessione tramite browser, ciò non è supportato da Versa. |
| 21 | Pour en savoir plus, rendez-vous sur help.fitbit.com. | Per ulteriori informazioni, visita il sito help.fitbit.com. |
| 22 | Installer Fitbit Connect | Installare Fitbit Connect |

## 5) Texte 1 à post-éditer

| Segment ID | Source segment | Target segment |
|---|---|---|
| 1 | Accepter et rejeter des appels téléphoniques | Accettare e rifiutare le chiamate telefoniche |
| 2 | S'il est appairé à un iPhone ou à un téléphone Android (8,0+), Versa vous permet d'accepter ou de rejeter des appels téléphoniques entrants. | Se abbinato a un iPhone o a un telefono Android (8.0+), Versa consente di accettare o rifiutare le chiamate in arrivo. |
| 3 | Si votre téléphone utilise une version plus ancienne du système d'exploitation Android, vous pouvez rejeter, mais pas accepter des appels depuis votre montre. | Se il telefono utilizza una versione precedente del sistema operativo Android, è possibile rifiutare, ma non accettare, le chiamate dall'orologio. |
| 4 | Pour accepter un appel, appuyez sur l'icône symbolisant un téléphone vert sur l'écran de votre montre. | Pour accepter un appel, appuyez sur l'icône symbolisant un téléphone vert sur l'écran de votre montre. |
| 5 | Remarque : vous ne pouvez pas parler dans la montre. | Nota: non è possibile parlare nell'orologio. |
| 6 | Lorsque vous acceptez un appel, c'est votre téléphone à proximité qui décroche. | Quando si accetta una chiamata, il telefono vicino risponde. |
| 7 | Pour rejeter un appel, appuyez sur l'icône symbolisant un téléphone rouge pour rediriger votre interlocuteur vers la messagerie vocale. | Per rifiutare una chiamata, premere l'icona rossa del telefono per reindirizzare il chiamante alla segreteria telefonica. |
| 8 | Le nom de votre interlocuteur s'affiche s'il est répertorié dans vos contacts. | Il nome della persona di contatto viene visualizzato se è presente nell'elenco dei contatti. |
| 9 | Sinon, l'écran indique simplement le numéro de téléphone. | In caso contrario, il display visualizza semplicemente il numero di telefono. |
| 10 | Répondre aux messages | Rispondere ai messaggi |

| | | |
|---|---|---|
| 11 | Répondez directement aux SMS et aux notifications de certaines applications de votre montre grâce à des réponses rapides préformatées. | Rispondere direttamente ai messaggi SMS e alle notifiche provenienti da applicazioni specifiche dell'orologio con risposte rapide preformattate. |
| 12 | Cette fonctionnalité est actuellement disponible sur les montres appairées à un téléphone Android. | Questa funzione è attualmente disponibile sugli orologi abbinati a un telefono Android. |
| 13 | Pour utiliser les réponses rapides : | Per usare le risposte rapide: |
| 14 | 1. | 1. |
| 15 | Appuyez sur la notification sur votre montre. | Toccare la notifica sull'orologio. |
| 16 | Pour voir des messages récents, faites glisser votre doigt depuis l'affichage de l'horloge. | Per visualizzare i messaggi recenti, far scorrere il dito dal display dell'orologio. |
| 17 | 2. | 2. |
| 18 | Appuyez sur Répondre. | Premere Rispondi. |
| 19 | Si vous ne voyez pas d'option pour répondre au message, cela signifie que les réponses rapides ne sont pas disponibles sur l'application qui a envoyé la notification. | Se non viene visualizzata un'opzione per rispondere al messaggio, significa che le risposte rapide non sono disponibili sull'applicazione che ha inviato la notifica. |
| 20 | 3. | 3. |
| 21 | Choisissez une réponse de texte dans la liste des réponses rapides ou appuyez sur l'icône d'emoji pour choisir un emoji. | Scegliere una risposta testuale dall'elenco delle risposte rapide o premere l'icona emoji per scegliere un emoji. |
| 22 | Vous pouvez également personnaliser les réponses rapides. | È inoltre possibile personalizzare le risposte rapide. |
| 23 | Pour en savoir plus, rendez-vous sur help.fitbit.com. | Per ulteriori informazioni, visita il sito help.fitbit.com. |

6) Texte 2 à post-éditer

| Segment ID | Source segment | Target segment |
|---|---|---|
| 1 | Entraînement avec Fitbit Coach | Allenamento con Fitbit Coach |
| 2 | L'application Fitbit Coach offre des séances d'entraînement de poids corporel sur votre poignet pour vous aider à rester en forme n'importe où. | L'applicazione Fitbit Coach offre sessioni di allenamento al polso per mantenersi in forma ovunque. |
| 3 | Pour commencer une séance d'entraînement : | Per avviare un esercizio: 1. |
| 4 | Sur votre Versa, appuyez sur l'application Fitbit Coach. | Su Versa, tocca l'app Fitbit Coach . 2. |
| 5 | Faites défiler la liste des séances d'entraînement. | Scorri l'elenco degli esercizi. 3. |
| 6 | Appuyez sur une séance d'entraînement, puis appuyez sur le bouton de lecture pour commencer. | Tocca un esercizio e premi il pulsante di riproduzione per iniziare. |
| 7 | Pour voir un aperçu de la séance d'entraînement au préalable, appuyez sur l'icône de menu en haut à droite. | Per visualizzare l'anteprima dell'esercizio, tocca l'icona di menu in alto a destra. |
| 8 | Pour en savoir plus, rendez-vous sur help.fitbit.com. | Per ulteriori informazioni, visita il sito help.fitbit.com. |
| 9 | Pendant une séance d'entraînement, vous pouvez écouter de la musique via les applications Musique, Deezer ou Pandora sur votre montre, ou contrôler la musique qui est jouée sur votre téléphone. | Durante un esercizio, puoi riprodurre musica tramite l'app Musica , l'app Pandora o l'app Deezer sul tuo smartwatch o controllare la riproduzione della musica sul tuo smartphone. |
| 10 | Pour écouter de la musique stockée dans votre montre, ouvrez la musique, l'application Pandora ou Deezer, et choisissez une liste de lecture. | Per riprodurre musica memorizzata sul tuo smartwatch, apri prima l'app Musica, Pandora o Deezer e scegli una playlist. |

| | | |
|---|---|---|
| 11 | Revenez ensuite dans l'application Fitbit Coach et démarrez une séance d'entraînement. | Quindi, torna indietro a Fitbit Coach e avvia un esercizio. |
| 12 | Remarque : vous devez appairer un appareil audio Bluetooth, comme des écouteurs ou un haut-parleur, à votre Versa pour écouter la musique enregistrée sur votre montre. | Tieni presente che dovrai associare un dispositivo audio Bluetooth, come gli auricolari o un altoparlante, a Versa per ascoltare la musica memorizzata sul tuo smartwatch. |
| 13 | Pour plus d'informations, consultez « Musique et podcasts », à la page 38. | Per ulteriori informazioni, vedi "Musica e podcast" a pagina 38. |
| 14 | Partage de votre activité | Condivisione dell'attività |
| 15 | Une fois que vous avez terminé une séance d'entraînement, synchronisez votre montre avec l'application Fitbit pour partager vos statistiques avec vos amis et votre famille. | Dopo aver completato un esercizio, sincronizza il tuo smartwatch con l'app Fitbit per condividere le tue statistiche con amici e familiari. |
| 16 | Pour en savoir plus, rendez-vous sur help.fitbit.com. | Per ulteriori informazioni, visita il sito help.fitbit.com. |
| 17 | Suivi de votre score de forme cardio | Rilevamento del punteggio di stato di forma |
| 18 | Suivez votre forme cardiovasculaire globale dans Fitbit Today ou dans l'application Fitbit. | Registra il tuo allenamento cardiovascolare generale in Fitbit Oggi o nell'app Fitbit. |
| 19 | Affichez votre score de forme cardio, ainsi que votre niveau de forme cardio qui vous compare aux autres utilisateurs. | Controlla il tuo punteggio e livello di attività cardio per confrontare i tuoi risultati con quelli dei tuoi colleghi. |

7) Liste officielle en langue italienne de commandes vocales Dragon Naturally Speaking

**Riconoscimento vocale Dragon**
Nuance Dragon NaturallySpeaking 13

**Riferimento rapido per i comandi**

Nuance®
Dragon®
NaturallySpeaking

**Note preliminari**
- Per fare clic su un pulsante o su un altro elemento dell'interfaccia, pronunciarne il nome preceduto da "clic su" (vedere la scheda Comandi della finestra di dialogo Opzioni).
- Fare una pausa prima e dopo l'enunciazione dei comandi, ma non durante.

**Comandi per il microfono**
A riposo | Interrompi l'ascolto
Spegni microfono

**Guida**
Apri la Guida
Cosa posso dire
Guida centro di apprendimento Dragon
Cerca nella Guida di Dragon...

**Ricerche nel computer**
Cerca nel computer...
Cerca nei documenti...
Cerca nella posta elettronica...

**Uso di Internet**
(occorre attivare le estensioni web di Dragon in Internet Explorer, Google Chrome o Firefox)
Vai a barra degli indirizzi, Premi Alt I
Vai lì, Premi Invio
Aggiorna pagina, Premi F5
Apri nuova scheda, Premi Control T
Trova in questa pagina, Premi Control F
Clic su <nome collegamento>
Clic su collegamento
Clic su campo di testo o clic su casella di modifica
Clic su pulsante
Clic su casella di controllo
Clic su immagine
Clic su lista, quindi mostra scelte
- In caso di più corrispondenze: scegli <n> o nascondi numeri o annulla

**Selezione del testo**
Seleziona tutto
Seleziona <xyz>
Seleziona successive <n> parole
Seleziona da <inizio> fino a <fine>
Seleziona paragrafo precedente
Seleziona documento
Annulla selezione

**Correzione degli errori di Dragon**
Correggi <xyz>
Correggi questo

**Modifica del testo**
Riprendi da <xyz>
Elimina riga
Elimina ultime <n> parole
Cancella questo <n> volte
Indietro <n>
Annulla questo
Taglia questo
Incolla questo
Apri la finestra di dettatura

**Compitazione**
Scrivi lettere
Scrivi lettere <maiuscolo b trattino 5>
Scrivi lettere <spazio como roma domodossola>
Passa a modalità compitazione

**Spostamento del cursore**
Inserisci prima di <xyz>
Vai indietro
Vai all'inizio / Vai alla fine
Vai <n> righe in basso
Vai a fine riga
Spostarsi a sinistra <n> caratteri
Pagina in alto / Pagina in basso

**Aggiungi righe e spazi**
Nuova riga
Nuovo paragrafo
Premi Invio
Premi tabulazione
Tasto tabulazione <n> volte

**Spostamento in un elenco**
Vai <n> in basso
Vai alla fine / Vai all'inizio
Premi Invio
Premi freccia destra

**Formattazione**
Imposta selezione come elenco puntato,
Imposta questo senza elenco puntato
La precedente linea in grassetto
Sottolinea <xyz>, Maiuscola iniziale <xyz>
Questo in maiuscolo, Rendi il testo minuscolo
Attiva/Disattiva tutto maiuscolo

**Operazioni sulle finestre**
Passa a <nome finestra>
Riduci a icona la finestra
Mostra desktop
Ripristina finestra
Elenca tutte le finestre
Elenca finestre relative a <programma>

**Apertura e chiusura di programmi**
(vedere le opzioni per il menu Start e il desktop)
Fai clic nel menu Start
Avvia DragonPad
Avvia <nome elemento>
Avvia Microsoft Word
Avvia posta
Avvia Internet Explorer
Apri Pannello di controllo
Chiudi finestra, Premi Alt F4

**Spostamento del mouse**
Mouse in alto
Mouse a destra
Mouse in basso più lentamente
Termina

**Posizionamento del mouse**
Griglia del mouse
Griglia del mouse sulla finestra
Griglia del mouse <da 1 a 9><da 1 a 9>
Annulla

**Pulsanti del mouse**
Mouse clic
Mouse doppio clic
Mouse clic destro

**Trascinamento del mouse**
Trascina il mouse in basso più velocemente
Trascina il mouse in basso a destra veramente veloce
Trascina il mouse in alto veramente veloce

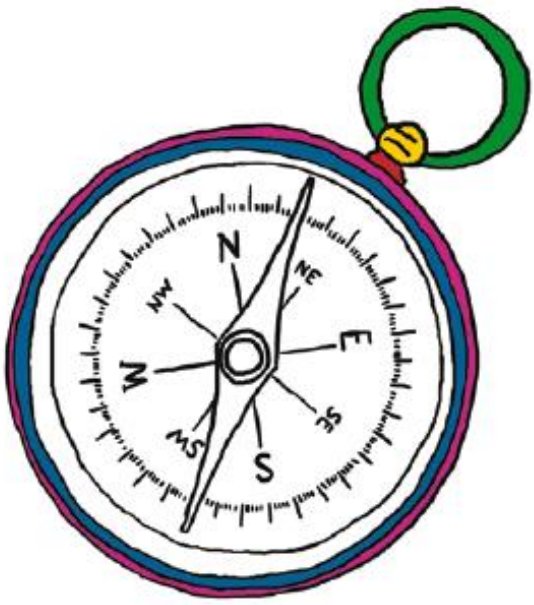
| Tipo di comando | Esempi di frasi | Azioni eseguite |
|---|---|---|
| Cerca in Internet (motore di ricerca predefinito) | “Cerca in Internet ristoranti arabi a Bolzano” | Il browser predefinito mostra i risultati del motore di ricerca predefinito per le parole chiave specificate. |
| Cerca in siti web specifici | “Cerca in eBay vestiti premaman” | Il browser predefinito mostra i risultati della ricerca in eBay delle parole chiave specificate. Dragon 13 supporta questa funzione per molti siti web inclusi i seguenti:<br>eBay, MSN, Twitter, Wikipedia |
| Apri il primo risultato della ricerca “Mi sento fortunato” di Google per determinate parole chiave | “Apri sito principale per previsioni del tempo locali” | Viene aperto il browser predefinito con la prima pagina proposta da Google per le parole chiave specificate. |
| Cerca in Internet con un motore di ricerca specifico | “Cerca su Google 53 diviso 12” | Viene aperto il browser predefinito con i risultati della ricerca per le parole chiave specificate. Dragon 13 supporta questa funzione per i motori di ricerca seguenti:<br>Bing, Google, Yahoo! |
| Cerca in Internet un tipo di informazioni specifico | “Cerca sul Web il video sul discorso d’insediamento di Napolitano" | Viene aperto il browser predefinito con i risultati della ricerca di determinate parole chiave nella categoria video del vostro motore di ricerca predefinito. Le categorie valide sono:<br>– Cerca sul Web (notizie \| eventi) relativi a...<br>– Cerca sul Web (prodotti \| negozi) per...<br>– Cerca (mappe \| luoghi) di…<br>– Cerca sul Web (video \| filmati) per…<br>– Cerca sul Web (immagini \| foto) di… |
| **Pubblica su Facebook | “Pubblica su Facebook ‘Non vedo l’ora di giocare a poker stasera’” | Dragon mostra una casella. Inoltre potete dapprima dettare, quindi pronunciare il comando “Pubblica questo su Facebook”. |

8) Taus Guidelines

# Machine Translation Postediting Guidelines

TAUS **GUIDELINES**

*In partnership with CNGL (Centre for Next Generation Localisation)*




TAUS

# MACHINE TRANSLATION POSTEDITING GUIDELINES

## Objectives and Scope

These guidelines are aimed at helping customers and service providers set clear expectations and can be used as a basis on which to instruct post-editors.

Each company's postediting guidelines are likely to vary depending on a range of parameters. It is not practical to present a set of guidelines that will cover all scenarios. We expect that organisations will use these baseline guidelines and will tailor them as required for their own purposes. Generally, these guidelines assume bi-lingual postediting (not monolingual) that is ideally carried out by a paid translator but that might in some scenarios be carried out by bilingual domain experts or volunteers. The guidelines are not system or language-specific.

## Recommendations

To reduce the level of postediting required (regardless of language pair, direction, system type or domain), we recommend the following:

- Tune your system appropriately, i.e. ensure high level dictionary and linguistic coding for RBMT systems, or training with ***clean, high-quality, domain-specific data*** for data-driven or hybrid systems.
- Ensure the ***source text*** is written well (i.e. correct spelling, punctuation, unambiguous) and, if possible, tuned for translation by MT (i.e. by using specific authoring rules that suit the MT system in question).
- Integrate ***terminology management*** across source text authoring, MT and TM systems.
- ***Train*** post-editors in advance.
- Examine the ***raw MT output quality*** before negotiating throughput and price and set reasonable expectations.
- Agree a definition for the ***final quality*** of the information to be post-edited, based on user type and levels of acceptance.
- Pay post-editors to give ***structured feedback*** on common MT errors (and, if necessary, guide them in how to do this) so the system can be improved over time.

## Postediting Guidelines

Assuming the recommendations above are implemented, we suggest some basic guidelines for postediting. The effort involved in postediting will be determined by two main criteria:

1. The quality of the MT raw output.
2. The expected end quality of the content.

To reach quality similar to "high-quality human translation and revision" (a.k.a. "publishable quality"), full postediting is usually recommended. For quality of a lower standard, often referred to as "good enough" or "fit for purpose", light postediting is usually recommended. However, light postediting of really poor MT output may not bring the output up to publishable quality standards. On the other hand, if the raw MT output is of good quality, then perhaps all that is needed is a light, not a full, post-edit to achieve publishable quality. So, instead of differentiating between guidelines for light and full-postediting, we will differentiate here between two levels of expected quality. Other levels could be defined, but we will stick to two here to keep things simple. The diagram that follows attempts to illustrate what is meant by different levels of postediting to achieve different levels of quality and how this might alter depending on the general quality of the raw MT output. The set of guidelines proposed below are conceptualised as a group of guidelines where individual guidelines can be selected, depending on the needs of the customer and the raw MT quality.

## Guidelines for achieving "good enough" quality

"Good enough" is defined as comprehensible (i.e. you can understand the main content of the message), accurate (i.e. it communicates the same meaning as the source text), but as not being stylistically compelling. The text may sound like it was generated by a computer, syntax might be somewhat unusual, grammar may not be perfect but the message is accurate.

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling apply.
- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text.
- Guidelines for achieving quality similar or equal to human translation:

This level of quality is generally defined as being comprehensible (i.e. an end user perfectly understands the content of the message), accurate (i.e. it communicates the same meaning as the source text), stylistically fine, though the style may not be as good as that achieved by a native-speaker human translator. Syntax is normal, grammar and punctuation are correct.

- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms".
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling, punctuation and hyphenation apply.
- Ensure that formatting is correct.

## 9) Instructions pour évaluation humaine

# Istruzioni “Évaluation humaine”

1. Cos’è l’évaluation humaine ?

L’évaluation humaine se fonde sur les jugements exprimés relativement à la qualité d’une traduction de la part des êtres humains (pas forcément des traducteurs) en vertu de leurs connaissances et de leur faculté de discernement. Elle prévoit la présentation d’un certain nombre de phrases traduites par un logiciel de traduction automatique à un juge humain, auquel on demande de s’exprimer sur la précision de la cible. Ce type d’évaluation peut être effectué de deux manière : avec ou sans phrase source. Dans le premier cas, l’on parlera d’évaluation bilingue, dans le deuxième d’évaluation monolingue. Un juge bilingue, connaissant la langue source et la langue cible, est considéré le plus qualifié pour donner ce jugement.

Normalement, il est fréquent de fournir aux évaluateurs une échelle de points (normalement, avec des valeurs de 1 à 5) afin de mesurer la qualité de chaque phrase. Dans l’approche humaine, l’évaluation est effectuée à l’aide de deux critères : l’adéquation (*adequacy* ou *fidelity*) et la fluidité (*fluency*). L’adéquation mesure le degré auquel la traduction transmet et conserve le sens de la phrase source, si une partie du message a été ajoutée, déformée ou n’a pas été transmise. Elle est effectuée en version bilingue, avec la phrase source et la cible. En revanche, l’évaluation de la fluidité peut être effectuée en mode monolingue, c’est-dire, seulement avec la cible. Elle mesure le degré de « fluidité » de la traduction, c’est-à-dire, si elle semble « naturelle » ou idiomatique à un natif de cette langue.

2. Cosa bisogna fare ?

Avete ricevuto due file Excel in ognuno ci sono 4 fogli (in basso potete vederli e scorrere tra essi). Due fogli si riferiscono al testo 1 (T1) e gli altri due al testo 2 (T2). Per entrambi dovrete valutare adequacy e fluency dando un voto a ogni frase secondo la tabella esplicativa presente in alto di ogni foglio. Alcune frasi saranno simili tra loro ma non è importante, voi dovete dare un vostro giudizio personale basato sulla lingua italiana. Per poter dare il voto dovrete scegliere un valore da 1 a 5 nel menù a tendina presente per ogni cella “score”.

**IMPORTANTE: gli spazi in più tra i segni di punteggiatura o gli errori nell’andare a capo <u>NON</u> sono da considerarsi errore per la scelta del voto.**

Vi prego di fare attenzione a **<u>non dimenticare</u>** nessun foglio in nessun file: sono quindi **<u>8 fogli in totale (4 fogli x 2 file)</u>**.

Due file:

1) FLU+ADEQ_ET
    a. Fluency T1
    b. Adequacy T1
    c. Fluency T2
    d. Adequacy T2
2) FLU+ADEQ_RV
    a. Fluency T1
    b. Adequacy T1
    c. Fluency T2
    d. Adequacy T2

Per qualsiasi domanda o dubbio resto ovviamente a disposizione. Non vi dovrebbe prendere più di 30 minuti e vi ringrazio già da ora per il tempo che mi dedicherete.

10) Protocole pour expérience

# Protocollo per valutazione

Per **Dragon** (distribuire lista comandi Dragon e riassunto)

1. Creare nuovo profilo utente (da fare sul momento)
2. Far allenare l'utente a utilizzare comandi in modo che il microfono si alleni sulla voce e l'utente familiarizzi con i comandi
3. Aggiungere che il corpus sia presente per ogni utente

Per **Word**

1. Per l'allenamento far allenare l'utente a post editare utilizzando il riconoscimento vocale
2. quindi avviare BB flash express in modo da registrare l'attività del partecipante.
3. dopo che l'utente ha finito l'allenamento sia per il riconoscimento vocale sia per il post editing far iniziare la valutazione vera e propria

**Valutazione in WORD**

1. allenamento con RV
2. testo 1 con RV
3. testo 2 con RV
4. testo 1 standard
5. testo 2 standard