

Traduction humaine et postédition : contrôle qualité en contexte académique

Perrine Schumacher

Volume 68, Number 3, December 2023

Nouvelles études en traduction économique et spécialisée
New studies in business and specialized translation

URI: <https://id.erudit.org/iderudit/1111955ar>
DOI: <https://doi.org/10.7202/1111955ar>

[See table of contents](#)

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

[Explore this journal](#)

Cite this article

Schumacher, P. (2023). Traduction humaine et postédition : contrôle qualité en contexte académique. *Meta*, 68(3), 510–536. <https://doi.org/10.7202/1111955ar>

Article abstract

Our study is a follow-up to a 2018 comparative experimental research on human translation (HT) and machine translation (MT) post-editing (PE) (English-French) that was conducted with third-year bachelor students in translation (Schumacher 2020b). In this article we provide the results of a 2nd controlled experiment conducted in 2021 with master's students. The objectives are to compare both HT and MTPE products (*DeepL* and *Google Translate*) and to contribute to the debate on adapting translator training programs. Quantitative analysis of this 2nd experiment shows that post-edited texts were judged of better overall quality than HT. The human quality assessment also shows very high success rates in PE (especially with *DeepL*) and confirmed the existence of a leveling effect in PE on the target text quality: PE—particularly with *Google Translate*—gives rise to an inverse relationship between a student's performance in HT and the quality of his/her post-edited product. In other words, students with poor-quality human-translated productions are the ones who benefit most from PE and vice versa. Finally, it appears the quality of a post-edited text depends on neural MT system since students performed better with *DeepL*.

Traduction humaine et postédition : contrôle qualité en contexte académique

PERRINE SCHUMACHER

*Université de Liège, Liège, Belgique**

p.schumacher@uliege.be

RÉSUMÉ

Notre étude s'inscrit dans le prolongement d'une recherche expérimentale comparative de traduction humaine (TH) et de postédition (PE) (anglais-français), menée en 2018 avec des étudiants de 3^e bachelier en traduction (Schumacher 2020b). Dans cet article, nous présentons les résultats d'une 2^e expérience contrôlée effectuée en 2021 avec des étudiants de master. L'objectif est de comparer les produits de la TH et de la PE de traduction automatique (TA) neuronale (*DeepL* et *Google Traduction*) et de contribuer au débat sur l'actualisation de la formation initiale des traducteurs. L'analyse quantitative de cette 2^e expérience fait ressortir que les textes postédités ont été considérés de meilleure qualité globale que les TH. Les résultats de l'évaluation humaine de la qualité ont également révélé de très hauts taux de réussite en PE (particulièrement avec *DeepL*) et ont permis de confirmer l'existence d'un effet nivelant en PE sur la qualité des textes cibles : la PE – en particulier avec *Google Traduction* – a donné lieu à une relation inverse entre le niveau d'un étudiant en TH et la qualité de sa PE. Autrement dit, plus l'étudiant est faible en TH, plus il tire avantage de la PE et vice versa. Enfin, il semblerait que la qualité d'une PE dépende du système de TA neuronale ; les étudiants ayant obtenu de meilleurs résultats avec *DeepL*.

ABSTRACT

Our study is a follow-up to a 2018 comparative experimental research on human translation (HT) and machine translation (MT) post-editing (PE) (English-French) that was conducted with third-year bachelor students in translation (Schumacher 2020b). In this article we provide the results of a 2nd controlled experiment conducted in 2021 with master's students. The objectives are to compare both HT and MTPE products (*DeepL* and *Google Translate*) and to contribute to the debate on adapting translator training programs. Quantitative analysis of this 2nd experiment shows that post-edited texts were judged of better overall quality than HT. The human quality assessment also shows very high success rates in PE (especially with *DeepL*) and confirmed the existence of a leveling effect in PE on the target text quality: PE—particularly with *Google Translate*—gives rise to an inverse relationship between a student's performance in HT and the quality of his/her post-edited product. In other words, students with poor-quality human-translated productions are the ones who benefit most from PE and vice versa. Finally, it appears the quality of a post-edited text depends on neural MT system since students performed better with *DeepL*.

RESUMEN

Este artículo es la continuación de una investigación experimental comparativa de traducción humana (TH) y posedición (PE) (inglés-francés), realizada en 2018 con estudiantes de 3.^{er} año de traducción (Schumacher 2020b). En concreto, presentamos los resultados de un segundo experimento controlado realizado en 2021 con estudiantes de máster, donde comparamos los productos de TH y de PE de traducción automática (TA) neuronal (*DeepL* y *Traductor de Google*) con el objetivo de contribuir al debate sobre la actualiza-

ción de la formación inicial de los traductores. El análisis cuantitativo de este segundo experimento revela que los textos post-editados pueden considerarse de mejor calidad general que los de TH. Los resultados de la evaluación humana de la calidad también revelan unos índices de éxito muy elevados en PE, especialmente con *DeepL*, y confirman la existencia de un efecto de nivelación en PE sobre la calidad de los textos meta: la PE (especialmente con *Traductor de Google*) da lugar a una relación inversa entre el nivel de estudiantes en TH y la calidad de su PE. En otras palabras, cuanto menos competencia tiene el estudiante en TH, más se beneficia de la PE y viceversa. Por último, parece que la calidad de PE depende del sistema de TA neuronal puesto que los estudiantes obtienen mejores resultados con *DeepL*.

MOTS CLÉS/KEYWORDS/PALABRAS CLAVE

postédition, traduction automatique neuronale, recherche expérimentale, évaluation de la qualité en traduction, étudiants en traduction

post-editing, neural machine translation, experimental research, translation quality assessment, students in translation

posedición, traducción automática neuronal, investigación experimental, evaluación de la calidad de la traducción, estudiantes en traducción

Il appartient aux traductrices et traducteurs de demain de négocier et de construire leur rôle dans ce paysage mouvant, et aux formatrices et formateurs de les y encourager.

(Rossi 2019: 103)

1. Introduction

Considérée comme l'un des plus vieux rêves de l'intelligence artificielle (Poibeau 2016), la traduction automatique est depuis quelques années l'objet de bien des éloges. Fin 2016, l'application des réseaux de neurones artificiels (*deep learning*) au domaine du traitement automatique des langues (TAL), et plus spécifiquement à la TA, a marqué un réel changement de paradigme (Dirand et Rossi 2019; Martikainen 2019) en amorçant le passage d'une approche dite « statistique »¹ à une approche neuronale. Plus personne ne nie aujourd'hui ces récentes percées technologiques dont la traduction fait en effet l'objet, comme tant d'autres métiers (Deneufbourg 2021²). Avec l'avènement de la TA neuronale, c'est l'ensemble de la profession qui connaît des transformations majeures puisque les outils de traduction assistée par ordinateur (TAO) et de TA font désormais partie intégrante du travail de traducteur (Alonso et Vieira 2020): « The technology [Neural Machine Translation] can easily be described as disruptive, since it is changing both the business model and the working processes in the translation industry » (Loock et Léchaugette 2021: 206).

Les avis dithyrambiques sur cette technologie et les prédictions alarmistes sur l'avenir du métier de traducteur sont aujourd'hui légion comme le confirment O'Brien et Ehrensberger-Dow: « There is an increasing supposition that MT is now good enough and can replace the work of professional translators » (2020: 146). Criant au « miracle de la traduction automatique neuronale » (Dirand et Rossi 2019: 71), certains vont jusqu'à prétendre que cette nouvelle approche est susceptible de sonner le glas de la TH ou, à tout le moins, capable de générer une qualité comparable à la TH (*human parity*) (Hassan, Aue *et al.* 2018³; Shoshan 2018⁴). Ce discours grandiloquent véhicule une conception agonistique qui oppose l'homme à la machine (Dirand et

Rossi 2019: 64), poussant à croire que la traduction-machine peut se substituer entièrement au traducteur humain et à envisager la traduction comme un métier en voie de disparition (Deneufbourg 2020⁵). Pourtant, tout le monde n'est pas de cet avis. De nombreux auteurs constatent et déplorent, à juste titre, une certaine exagération médiatique autour du phénomène (voir Deneufbourg 2021 ; Moorkens 2018) qui provient, entre autres, d'une surévaluation très fréquente des performances des moteurs de TA neuronale dans la presse tant généraliste que spécialisée (Loock 2019: 59). Toujours est-il que l'usage de la TA à des fins très diverses (O'Brien et Ehrensberger-Dow 2020) est de plus en plus répandu chez le grand public, mais aussi dans le monde professionnel. Ce recours exponentiel à la TA (Way 2018) et à la PE sur l'ensemble du secteur des services langagiers est d'ailleurs décrit comme une tendance qui n'est pas près de s'inverser (Deneufbourg 2019).

Pour nous, les outils de TA viennent en réalité s'ajouter à toute une palette d'outils d'aide à la traduction qui permettent au traducteur d'être plus efficace et plus productif (De Sutter, Cappelle *et al.* 2017), et ainsi de pouvoir faire face à la masse toujours plus importante et urgente de documents à traduire. En outre, Dirand et Rossi sont même convaincus qu'à l'inverse de condamner le traducteur humain, ces progrès technologiques pourraient permettre « d'envisager des interactions personne-machine plus nombreuses et plus fluides (sorties de TA intégrées à un outil de TAO, TA interactive s'ajustant à mesure que le traducteur écrit) » (Dirand et Rossi 2019: 76). Car, contrairement à ce que nous pourrions croire et comme le soulignait déjà Lavault-Olléon (2007), en dépit de tous ces développements technologiques, nous n'avons jamais eu tant besoin de traducteurs spécialisés qu'aujourd'hui. La chercheuse remarque toutefois que

la traduction a perdu son autonomie [car] elle n'a jamais été aussi dépendante de facteurs autres que linguistiques – principalement de l'évolution des outils et de la conjoncture économique [...]. D'où le défi que représente l'offre d'une formation efficace et adaptée à ces nouveaux métiers, et la réflexion didactique qu'elle doit engendrer. (Lavault-Olléon 2007: x)

D'ailleurs, la nécessité de repenser et d'adapter les programmes de formation proposés aux futurs traducteurs en tenant compte des évolutions technologiques de la profession s'impose dans la littérature scientifique (voir notamment De Faria Pires 2018 ; Guerberof Arenas et Moorkens 2019 ; Loock et Léchaugette 2021). Et depuis 2017, l'interaction avec la TA a fait son apparition dans le référentiel de compétences du réseau européen des masters en traduction (EMT): « la capacité d'interagir avec la traduction automatique (TA) dans le processus de traduction fait maintenant partie intégrante de la compétence de traduction professionnelle » (Groupe d'experts 2017)⁶. Aussi, l'étude qui fait l'objet de cet article s'inscrit dans le cadre d'une thèse qui entend contribuer à cette réflexion sur l'intégration des outils de TA à la formation initiale des traducteurs. Nous espérons que nos résultats serviront à faire valoir la nécessité de revoir et d'adapter les cursus des étudiants.

Par ailleurs, s'il est indubitable que la technologie qui sous-tend la TA a considérablement évolué et que les sorties des moteurs de TA sont aujourd'hui sans cesse plus exploitables (Deneufbourg 2021), il est légitime de s'interroger sur la qualité du produit fini. C'est la raison pour laquelle nous nous sommes attachée dans ce travail à comparer le produit de la TH au produit de la PE en contexte académique. En 2018, nous avons mené une première expérience comparative (prétest) avec des étudiants

de 3^e bachelier en traduction pour la combinaison de langues anglais-français (Schumacher 2020b). Ce prétest a été envisagé comme une étape méthodologique en vue de l'élaboration d'une deuxième expérience dont il est question dans cet article. De cette manière, nous avons pu tester et affiner les différents paramètres. Par ailleurs, les résultats obtenus dans cette première expérience nous ont permis de reconsidérer et de reformuler certaines de nos hypothèses de travail. Dans ce prétest, nous avons notamment pu observer un meilleur taux de réussite en PE de TA neuronale par rapport aux productions humaines, ainsi qu'un effet nivelant en PE. Cet effet s'est traduit par un resserrement des notes autour de la moyenne et par l'établissement d'une relation inverse entre le niveau d'un étudiant en TH et la qualité de sa PE, autrement dit, plus l'étudiant est faible en TH, plus il tire avantage de la PE et à l'inverse, plus il est compétent en TH, plus la PE entraîne une baisse de qualité de sa production. Dans cet article, nous présentons les résultats de l'analyse quantitative de notre deuxième recherche expérimentale menée en 2021 selon la même méthodologie qu'en 2018, mais cette fois auprès d'étudiants plus expérimentés et au seuil de leur vie professionnelle puisqu'il s'agit d'étudiants de master. Nous verrons que ces résultats confirment les résultats du prétest.

2. Objectifs et hypothèses de recherche

La présente étude est consacrée à la comparaison de deux modes de traduction : la TH et la PE de TA en contexte académique (étudiants de 2^e master) pour la combinaison de langues anglais-français. Notre objectif principal consiste à explorer les effets de ces deux processus en mettant en lumière les potentielles différences de qualité de produit fini. Comme nous l'avons déjà évoqué dans l'introduction, notre volonté est aussi de contribuer au débat sur l'actualisation des formations proposées aux futurs traducteurs.

En nous fondant sur les travaux de Daems, Vandepitte *et al.* (2017), Depraetere, De Sutter *et al.* (2014) et Jia, Carl *et al.* (2019), ainsi que sur les résultats issus de notre prétest, nous formulons l'hypothèse que la PE de TA neuronale (anglais-français) effectuée en contexte académique n'a pas d'influence négative sur la qualité globale du produit fini comparativement à la TH. Nous nous attendons même à ce que les textes postédités par les étudiants contiennent, en moyenne, moins d'erreurs que les productions humaines (Martikainen et Mestivier 2020⁷) et nous pensons observer un meilleur taux de réussite en TA neuronale qu'en TH (Schumacher 2020b). En outre, nous cherchons à vérifier l'existence d'un *effet nivelant* (*leveling effect*) en TA neuronale (Garcia 2011 ; Killman 2018) chez les étudiants plus expérimentés (2^e master). Nous tenons également à vérifier si la qualité d'un texte postédité dépend du système de TA neuronale (*Google Traduction* ou *DeepL*⁸). Bien que les performances d'un moteur de TA varient fortement en fonction de divers facteurs tels que le type de texte, les attentes en matière de qualité et la combinaison de langues (Poibeau 2022), *DeepL* se révélerait particulièrement performant par rapport aux concurrents pour certaines combinaisons de langues dont la traduction anglais-français, c'est du moins ce qu'attestent les résultats de l'étude menée en janvier 2020 par la société (*DeepL GmbH*, 2020⁹). C'est pourquoi, en nous appuyant sur le discours véhiculé dans les médias et sur les prétentions de la société *DeepL GmbH*, nous pensons que *DeepL* est capable de fournir de meilleures performances que *Google Traduction*.

3. Méthodologie

Pour mettre à l'épreuve ces hypothèses, nous avons à nouveau mené une expérience contrôlée en suivant la même méthodologie qu'en 2018, mais cette fois auprès d'un échantillon d'étudiants de master en traduction de l'Université de Liège (ULiège) en Belgique.

3.1. Participants

Cette recherche expérimentale a été menée avec 24 étudiants de 2^e année de master en traduction de l'ULiège dont la langue maternelle est le français. Sans être des postéditeurs aguerris, ces étudiants en fin de parcours universitaire ont déjà tous eu l'occasion de faire de la PE au cours de leur cursus (que ce soit dans le cadre d'un cours, d'un séminaire ou de leur propre initiative) sans pour autant avoir reçu de formation spécifique. Avant de prendre part à cette expérience, chaque participant a été invité à prendre connaissance d'un formulaire de consentement qui reprend les modalités du projet d'étude, ainsi que les exigences et dispositions du Règlement Général sur la Protection des Données (RGPD) de l'ULiège et à marquer son accord en signant ce document.

3.2. Corpus

Le corpus de départ est composé de trois textes sources (TS) en anglais (entre 300 et 340 mots) identiques à ceux du prétest (voir annexe 1). Dans cette expérience, nous avons cherché à reproduire un environnement qui soit relativement proche du contexte pédagogique dans lequel évoluent les étudiants en traduction à l'ULiège. C'est pourquoi nous avons demandé à trois enseignantes de la filière traduction-interprétation de choisir chacune un TS susceptible d'être donné en préparation dans leurs cours respectifs. Les textes sélectionnés sont des textes informatifs qui relèvent de trois thématiques distinctes, car nous avions l'intention de vérifier si ce facteur pouvait être considéré comme une variable dans cette étude; il s'agit d'un article de presse tiré de *The Guardian*, d'un article économique qui provient du site de *PricewaterhouseCoopers* (PWC), ainsi que d'un article de vulgarisation scientifique publié dans la revue américaine *Scientific American*.

3.3. Déroulement

Cette expérience contrôlée s'est déroulée en juin 2021 à l'ULiège. En raison de la capacité du local dont nous disposions et des exigences sanitaires en vigueur à ce moment-là, l'expérience a dû être divisée en deux sessions de 4 heures. Dans le but de comparer les produits de la TH aux produits de la PE de TA neuronale, les étudiants ont été invités à traduire vers le français les trois TS selon trois modes de traduction différents. Chaque étudiant a ainsi produit :

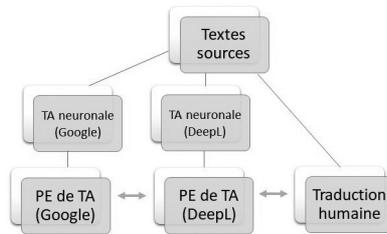
- une traduction humaine;
- une postédition de traduction automatique neuronale (*Google Traduction*);
- une postédition de traduction automatique neuronale (*DeepL*).

Nous avons travaillé sur un échantillon de 24 étudiants et, en suivant les recommandations de Saldanha et O'Brien (2014), nous avons veillé à disposer systématiquement de plusieurs versions du même TS produites par plusieurs étudiants. Pour éviter tout biais, tout effet d'ordre et pour que l'échantillon de productions soit le plus représentatif possible des trois modes de traduction, nous avons réparti les tâches entre les étudiants selon un plan croisé (principe du carré latin) en fonction des variables suivantes : TS, modes de traduction et ordre dans lequel les tâches devaient être accomplies.

À la suite de l'expérience, notre corpus final se compose des trois TS, des six TA brutes, de 24 PE de *Google Traduction*, de 24 PE de *DeepL* et de 24 TH. La composition du corpus final est représentée sur la figure 1.

FIGURE 1

Composition du corpus final



3.4. Tâches

Pour la tâche de TH, l'étudiant pouvait disposer de toutes les ressources qu'il jugeait utiles (dictionnaires explicatifs et traductifs, Internet, etc.). Il lui était toutefois interdit d'utiliser tout moteur de TA et tout logiciel de TAO. Pour nous assurer que les étudiants ne recourent pas à ces outils lors de la tâche de TH, l'expérience a eu lieu en présentiel, dans une salle informatique et, grâce au logiciel *AVIDAnet®Live*¹⁰, nous pouvions surveiller l'écran de chaque étudiant (voir figure 2) tout au long de la session.

FIGURE 2

AVIDAnet®Live (Televic Education 2018¹¹)



Pour la PE, les trois TS ont été traduits préalablement par un moteur de TA et nous avons demandé aux participants de postéditer ces sorties de TA sur *MateCat*¹² (Federico, Bertoldi *et al.* 2014). Étant donné que l'une de nos hypothèses postule que la qualité des textes postédités dépend du système de TA neuronale, nous avons comparé indirectement les performances de deux outils de TA génériques disponibles en ligne. Nous avons choisi *Google Traduction* et *DeepL*, les deux moteurs de TA grand public les plus connus aujourd'hui et, sans surprise, les moteurs les plus fréquemment utilisés chez les étudiants (Loock et Léchauguette 2021). Précisons que pour *DeepL*, nous avons eu recours à la version disponible en ligne gratuitement et non à la version professionnelle *DeepL Pro*. Concernant la qualité attendue en PE, il a été demandé aux étudiants de fournir une production de qualité comparable ou égale à la qualité d'une TH. Nous leur avons fourni à cet effet une liste de consignes (voir annexe 2).

3.5. Évaluation humaine

Une fois récoltées, les productions des étudiants ont été soumises à une évaluation humaine. Cinq enseignantes ayant plusieurs années d'expérience dans l'enseignement de la traduction ont accepté d'évaluer, à l'aveugle, la qualité des TH et des textes postédités, à raison de deux enseignantes par TS. Comme méthode d'évaluation, nous avons eu recours à une typologie détaillée d'erreurs, élaborée dans le cadre de cette étude, qui a permis de classer les erreurs par catégories, de les comptabiliser, de calculer la moyenne des erreurs entre les annotations des évaluateuses et enfin d'attribuer une note sur 20 à chaque production selon une pondération prédéfinie (tableau 1).

Cette évaluation de la qualité était divisée en deux étapes. La première étape consistait à évaluer *l'acceptabilité* du texte cible, entendu comme le respect des normes et usages de la langue cible (figure 3). Il s'agissait donc essentiellement de repérer les erreurs de langue ainsi que les problèmes de fluidité et de cohérence en langue cible. Pour cette étape, l'évaluatrice disposait uniquement du texte cible en français. Tandis que la deuxième étape permettait de juger de *l'adéquation* du texte cible par rapport au TS (figure 4). Il s'agissait de repérer les erreurs de sens/de traduction en disposant, cette fois, du TS. Pour une explication détaillée de la méthode d'évaluation, voir Schumacher (2019).

TABLEAU 1

Pondération des erreurs pour la notation sur 20

Acceptabilité (40%)	Catégorie	Grammaire/ syntaxe	Orthographe	Typographie/ ponctuation	Lexique	Incompréhension	Style	Cohérence
	Pénalité		1,0	1,0	0,2	0,3	1,3	0,3
Adéquation (60%)	Catégorie	Contresens/ non-sens	Glissement de sens	Omission	Ajout	Calque	Vocabulaire	Autre
	Pénalité		1,3	0,7	0,7	0,7	0,5	0,5

FIGURE 3
Critères Acceptabilité

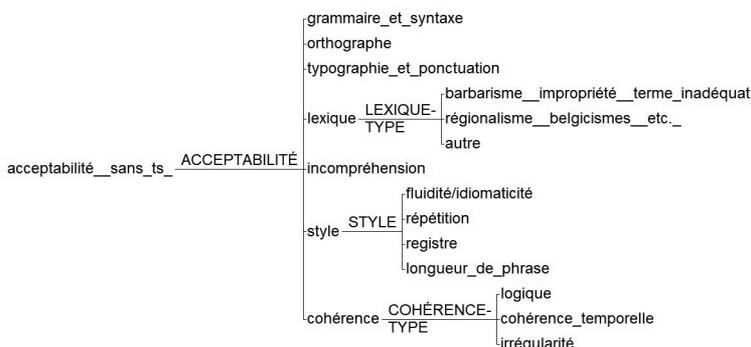
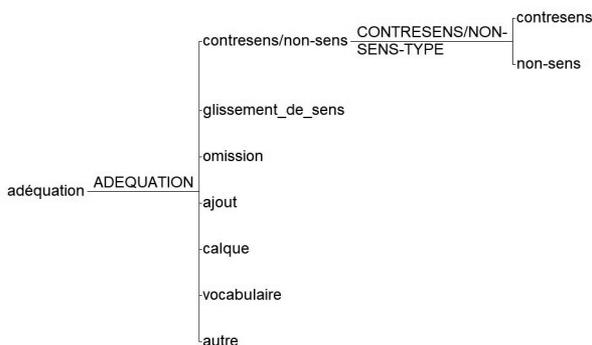


FIGURE 4
Critères Adéquation



4. Résultats

Dans cet article, nous nous limitons à la présentation des résultats issus de l’analyse statistique descriptive de notre corpus constitué des productions d’apprenants en traduction après évaluation humaine.

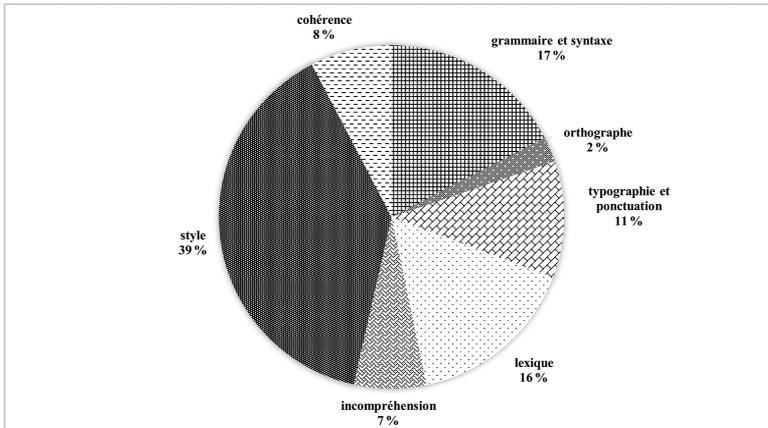
4.1. Distribution des erreurs par catégories

Tout d’abord, les figures 5 et 6 nous permettent de visualiser la distribution des erreurs par catégories dans l’ensemble de notre corpus, indépendamment du mode de traduction et du TS.

La première phase d’évaluation consistait à déterminer l’acceptabilité du texte cible. Pour rappel, dans cette phase, l’évaluateur humain ne disposait pas du TS. On peut tout d’abord observer une nette prépondérance des erreurs de *style* (fluidité, idiomaticité, répétition, longueur de phrase, registre inadéquat, autre) qui représentent 39% de l’ensemble des erreurs d’acceptabilité commises par les étudiants. Les

FIGURE 5

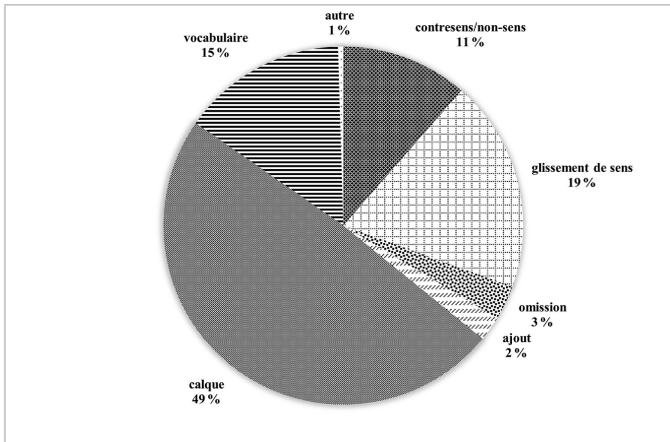
Distribution des erreurs en phase d'acceptabilité



catégories qui arrivent ensuite sont *grammaire et syntaxe* et *lexique* avec respectivement 17 % et 16 %, suivies des catégories *typographie et ponctuation* (11 %), *cohérence* (8 %), *incompréhension* (7 %) et enfin *orthographe* (2 %).

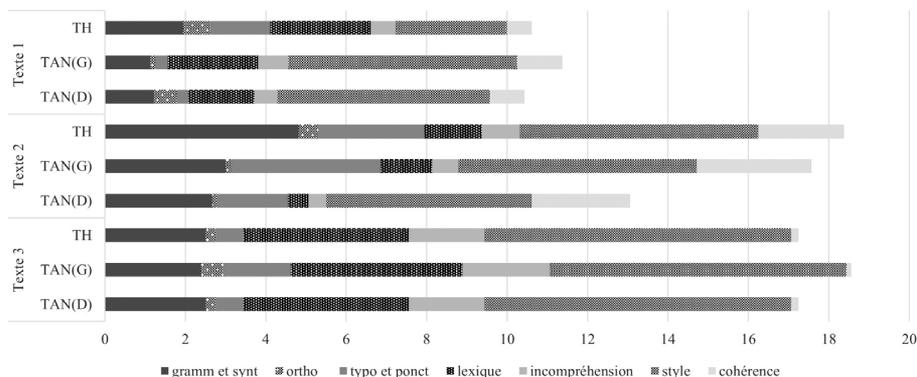
FIGURE 6

Distribution des erreurs en phase d'adéquation



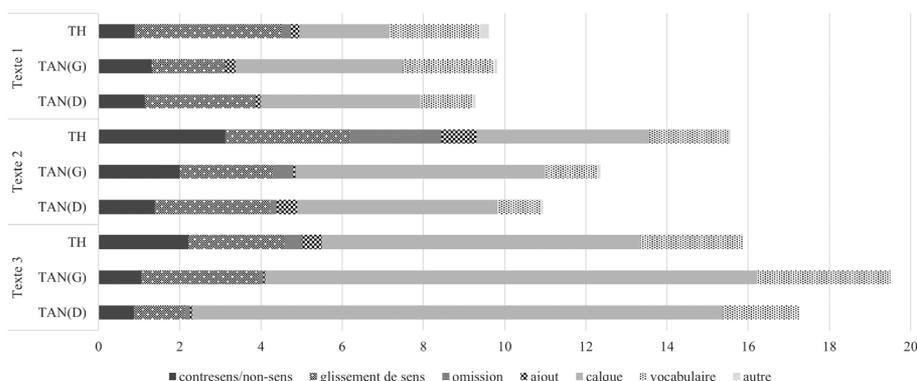
La deuxième phase d'évaluation consistait à déterminer l'adéquation du texte cible par rapport au TS. Dans cette phase, les calques¹³ représentent près d'une erreur sur deux (49 %). La deuxième catégorie la plus représentée est celle des *glissements de sens*, comptant pour 19 % de l'ensemble des erreurs, arrivent ensuite les catégories *vocabulaire* (15 %), *contresens/non-sens*¹⁴ (11 %), *omission* (3 %), *ajout* (2 %) et *autre* (1 %). La distribution des erreurs en fonction du mode de traduction est reprise dans les figures 7 et 8.

FIGURE 7
Distribution des erreurs en fonction du mode de traduction (Acceptabilité)



En analysant la figure 7, nous pouvons affirmer deux choses : premièrement que le texte 1 (article de presse) a généré moins d’erreurs par rapport aux deux autres TS quel que soit le mode de traduction, et deuxièmement, que pour le texte 2 (texte économique), les étudiants ont, en moyenne, commis sensiblement moins d’erreurs d’acceptabilité en PE de TA neuronale générée par *DeepL* par rapport aux deux autres modes de traduction. Étant donné que nous croyons que plus l’étudiant intervient sur la TA, plus il est susceptible d’introduire des erreurs d’acceptabilité, notre hypothèse interprétative est que la TA brute *DeepL* du texte 2 est davantage exploitable et qu’elle a nécessité moins de postéditions que la TA de *Google Traduction*.

FIGURE 8
Distribution des erreurs en fonction du mode de traduction (Adéquation)



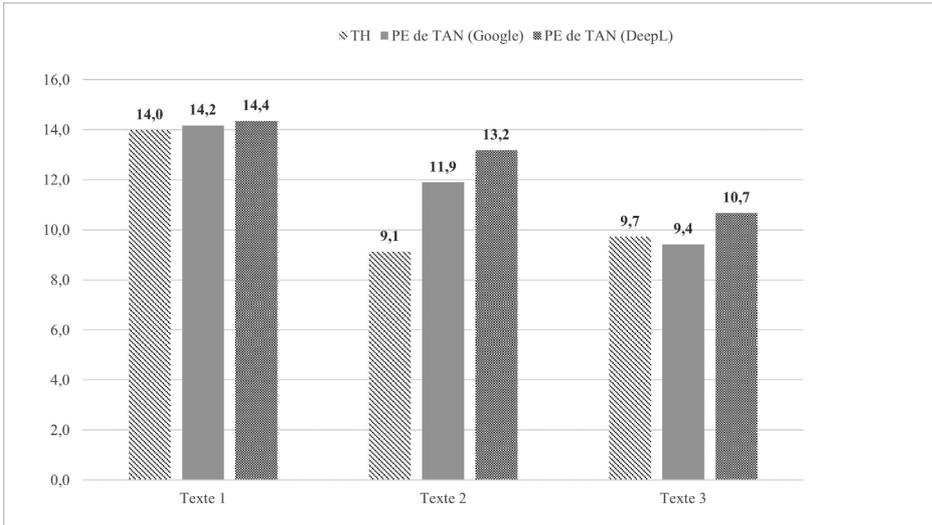
En analysant la figure 8, il apparaît que la TH est le mode de traduction qui a généré le moins de calques, quel que soit le TS; nous reviendrons sur cette surreprésentation de calques en TA au point 4.7. La PE de TA générée par *DeepL* a entraîné, en moyenne, moins d’erreurs de vocabulaire par rapport aux deux autres modes de traduction. Enfin, c’est en PE de TA générée par *Google Traduction* que les étudiants ont commis le plus d’erreurs en moyenne, quel que soit le TS.

4.2. Comparaison de la qualité des productions par mode de traduction

Nous avons calculé, pour chaque production, une note globale selon la pondération que nous avons prédéfinie pour chaque catégorie d'erreurs (tableau 1). Cette note globale combine la note obtenue en phase d'acceptabilité (à raison de 40 %) et la note obtenue en phase d'adéquation (à raison de 60 %).

FIGURE 9

Moyenne des notes

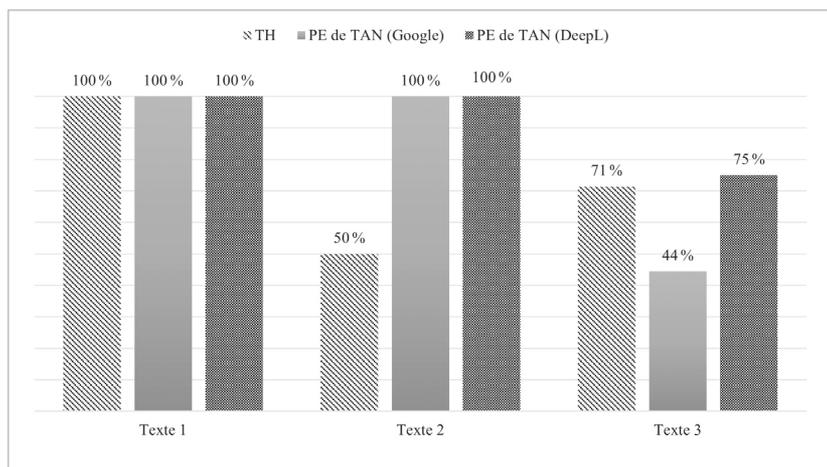


La figure 9 nous permet de visualiser la moyenne générale des notes en fonction du mode de traduction et en fonction du TS. Nous pouvons observer que la TH est le mode de traduction pour lequel la moyenne générale des notes est systématiquement la plus basse par rapport aux deux autres modes, quel que soit le TS, tandis que la PE de TA neuronale générée par *DeepL* est le mode de traduction pour lequel la moyenne des notes est toujours la plus haute, même si la différence est parfois très faible (voir texte 1).

4.3. Taux de réussite

Précisons tout d'abord que dans l'enseignement supérieur en Fédération Wallonie-Bruxelles de Belgique, le seuil de réussite est fixé à 10 sur 20 (Fédération Wallonie-Bruxelles¹⁵). Sur la figure 10, il apparaît que la PE de TA neuronale (*Google Traduction* ou *DeepL*) a permis d'obtenir de très bons taux de réussite pour les trois TS, sauf dans le cas du texte 3 en PE de TA générée par *Google* avec un taux de réussite de 44 %. Il est surprenant de noter que pour le texte 2 (texte économique), la TH n'obtient que 50 % de taux de réussite alors que tous les étudiants ayant postédité ce texte ont obtenu une note de réussite. La TA brute leur aurait-elle permis d'avoir un meilleur « accès » au sens du TS? Nous avons déjà évoqué cette possibilité dans notre 1^{re} expérience; nous pensons que la TA neuronale peut, dans certains cas, pallier la compréhension lacunaire ou partielle des étudiants en clarifiant certains passages complexes du TS, ce qui

FIGURE 10

Taux de réussite selon le mode de traduction

permet d'améliorer la compréhension générale qu'ils ont du TS (Schumacher 2020b). Cette explication trouve notamment écho dans les avis recueillis par González Pastor auprès d'étudiants en traduction de l'Université de Valence : « A large number of respondents (16) reported to resort to MT as a problem-solving tool when faced with text fragments containing especially difficult sentences or complex syntactic structures » (2021 : 56). Sans vouloir tirer de conclusions hâtives, nous pouvons affirmer que, dans cette étude, si l'on considère les trois TS, la PE de TA neuronale produite par *DeepL* est le mode de traduction le plus à même de permettre à un étudiant d'obtenir une note supérieure ou égale au seuil de réussite.

Les résultats représentés sur les figures 9 et 10 suscitent également plusieurs interrogations : peut-on en conclure que le texte 1 (article de presse) était (trop) simple à traduire/postéditer pour les étudiants de 2^e master puisque tous ont obtenu une note égale ou supérieure au seuil de réussite ? Et, à l'inverse, que le texte 3 (texte scientifique) était d'une trop grande complexité à traduire/postéditer ou encore que les évaluatrices ont été (trop) clémentes dans leur évaluation des productions du texte 1 et (trop) sévères pour les productions du texte 3 ? Il s'agit là d'hypothèses explicatives qu'il nous est impossible de confirmer à ce stade.

4.4. Moyenne des notes et écart-type

Nous avons repris sur cette figure la moyenne des notes obtenues en fonction du mode de traduction et du TS ainsi que l'écart-type qui, rappelons-le, est un « indice de dispersion autour de la moyenne exprimant la plus ou moins grande hétérogénéité des données » (Gana et Broc 2018). D'emblée, nous constatons une réduction de l'écart-type en PE de TA neuronale (*Google Traduction* et *DeepL*) par rapport à l'écart-type calculé en TH, quel que soit le TS. Ces résultats reflètent un resserrement des notes autour de la moyenne en PE, et ce, pour les trois textes. Ce resserrement traduit une répartition moins étalée des notes en PE ; les textes postédités ont récolté moins de (très) mauvaises notes et moins de (très) bonnes notes par rapport aux TH.

FIGURE 11
Moyenne des notes et écart-type

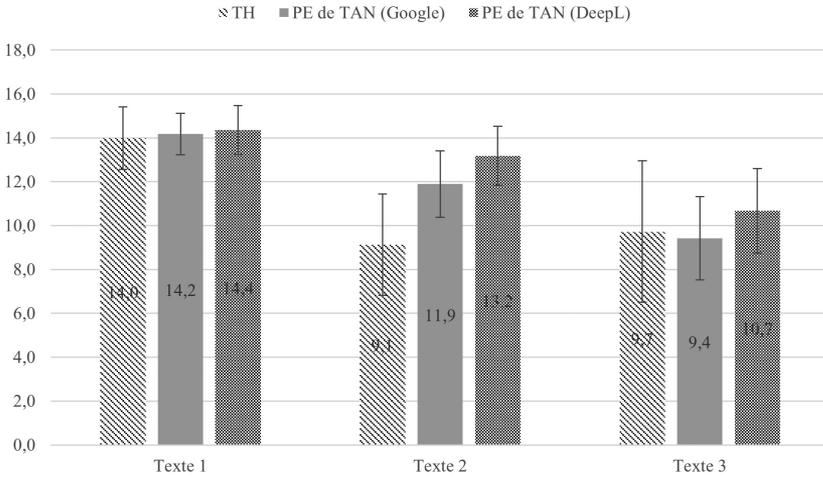
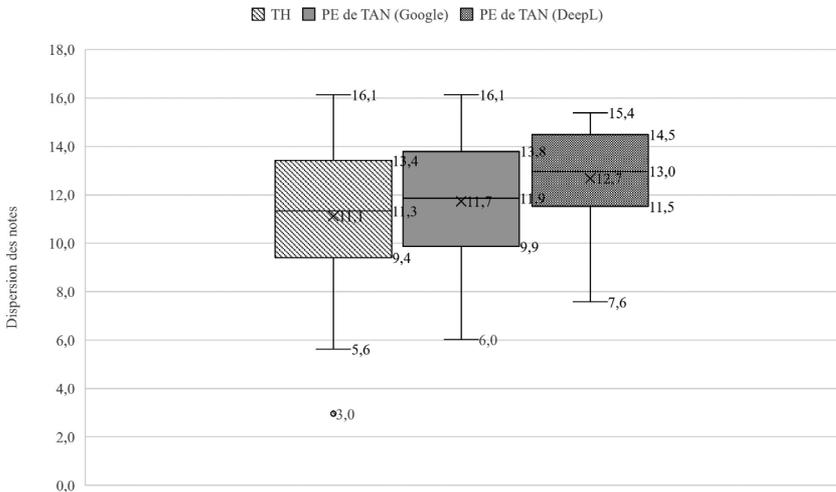


FIGURE 12
Dispersion des notes autour de la médiane



Sur la figure 12 en boîte à moustaches, nous pouvons observer la dispersion des notes autour de la médiane selon le mode de traduction sans tenir compte du TS. Pour chaque mode de traduction, la moustache supérieure et la moustache inférieure indiquent respectivement la note la plus élevée et la note la plus basse de la série de notes; la boîte délimitée par le 1^{er} quartile (Q1) et le 3^e quartile (Q3) représente l'intervalle interquartile qui contient au moins 50 % des notes; le marqueur x représente la moyenne des notes et la ligne horizontale représente la médiane. Cette figure nous permet tout d'abord de constater à nouveau un resserrement des notes, cette fois autour de la médiane. Ensuite, nous relevons également que non seulement

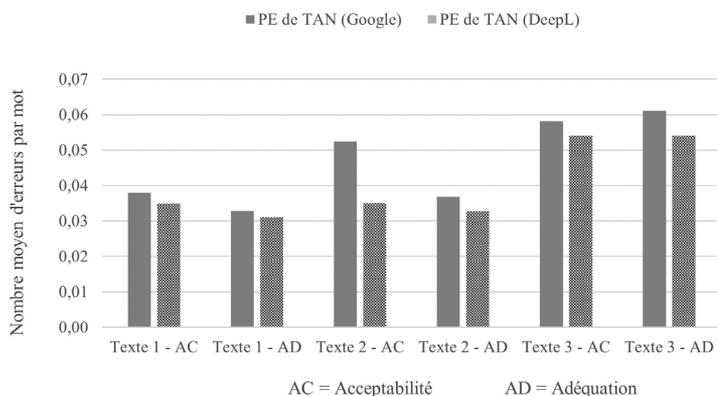
la moyenne, mais aussi la note médiane sont plus élevées en PE de TA générée par *DeepL* par rapport à la PE de TA générée par *Google Traduction*, mais davantage encore par rapport à la TH. Le resserrement des notes se traduit par une réduction de l'écart interquartile (Q3-Q1): 4,02 en TH; 3,92 en PE de TAN (*Google*) et 2,97 en PE de TAN (*DeepL*). Enfin, notons la présence d'une valeur extrêmement basse (3,0) en TH qui se trouve représentée par un point en dessous de la moustache inférieure; il s'agit d'une valeur extrême dite « aberrante »¹⁶. Cela nous amène à formuler l'hypothèse interprétative suivante: le recours à la PE de TA neuronale – particulièrement la PE de *DeepL* – chez les étudiants a tendance à homogénéiser l'ensemble des notes.

4.5. *Google Traduction vs DeepL*

Nous avons voulu vérifier notre deuxième hypothèse selon laquelle la qualité d'un texte postédité serait influencée par le moteur de TA neuronale (*Google Traduction* ou *DeepL*). Pour ce faire, nous avons comparé les productions obtenues en PE de TA générée avec *Google Traduction* aux productions obtenues en PE de TA générée avec *DeepL* selon le niveau moyen d'erreurs comptabilisées par TS.

FIGURE 13

Google Traduction vs DeepL



Sur la figure 13, nous retrouvons, pour chaque texte, le nombre moyen d'erreurs relevées en phase d'acceptabilité (AC) et le nombre moyen d'erreurs en phase d'adéquation (AD). Afin d'établir une base de comparaison valable, nous avons normalisé nos résultats en prenant comme indicateur le nombre moyen d'erreurs par mot du TS. Les résultats représentés sur cette figure révèlent que la PE de TA générée avec *DeepL* a permis d'aboutir à des productions de meilleure qualité, puisqu'elles contiennent, en moyenne, moins d'erreurs que les PE de TA générée avec *Google Traduction*. Cette conclusion vaut tant pour la phase d'acceptabilité que pour la phase d'adéquation, et quel que soit le TS. Ces résultats confirment ainsi notre hypothèse de départ.

Dans les tableaux 2 et 3, nous avons repris deux exemples tirés de notre corpus pour lesquels la TA brute de *DeepL* était de meilleure qualité que la TA brute de *Google Traduction*.

TABLEAU 2

Comparaison *Google Traduction* et *DeepL* (1)

Texte source 3	Of the various manifestations of synesthesia, the most common involves seeing monochromatic letters, digits and words in unique colors.	
TA brute	<i>Google</i> 2020	... des lettres, des chiffres et des mots monochromatiques ...
	<i>DeepL</i> 2020	... des lettres, des chiffres et des mots monochromes ...
Traduction correcte	... des lettres, des chiffres et des mots monochromes ...	
Traduction fautive	... monochromatiques ...	
Occurrence de l'erreur	TH: 29% (2 sur 7) PE de <i>Google</i> : 78% (7 sur 9) PE de <i>DeepL</i> : 12,5% (1 sur 8)	

Dans cet exemple, l'adjectif *monochromatic* a été traduit de manière calquée par *Google Traduction*, mais pas par *DeepL*. L'adjectif *monochromatique* existe bel et bien, mais il s'agit d'un terme vieilli (le Petit Robert, s.d.¹⁷) auquel on préférera l'emploi de l'adjectif *monochrome*. Ce calque se retrouve dans sept PE de *Google*, dans une PE de *DeepL*, ainsi que dans plusieurs TH.

TABLEAU 3

Comparaison *Google Traduction* et *DeepL* (2)

Texte source 3	But synesthetes also report making computational errors because 6 and 8 have the same color.	
TA brute	<i>Google</i> 2020	Mais les synesthetes rapportent aussi avoir fait des erreurs de calcul...
	<i>DeepL</i> 2020	Mais les synesthésiens déclarent également faire des erreurs de calcul...
Traduction fautive	... avoir fait des erreurs de calcul...	
Traduction correcte	... déclarent qu' ils font des erreurs de calcul/... déclarent faire des erreurs de calcul...	
Occurrence de l'erreur	PE de <i>Google</i> : 44% (4 sur 9) PE de <i>DeepL</i> : 0% (0 sur 8) TH: 0% (0 sur 7)	

Dans ce tableau 3, le gérondif *making* a été rendu par *Google Traduction* par une construction infinitive au passé, *avoir fait*, marquant une relation temporelle d'antériorité alors que pour des raisons de cohérence, la temporalité qui aurait dû être privilégiée est celle du présent, ce qui est le cas dans la TA brute de *DeepL*. Cette erreur s'est retrouvée dans plusieurs PE de *Google*, mais dans aucune PE de *DeepL*.

4.6. Résultats par étudiant

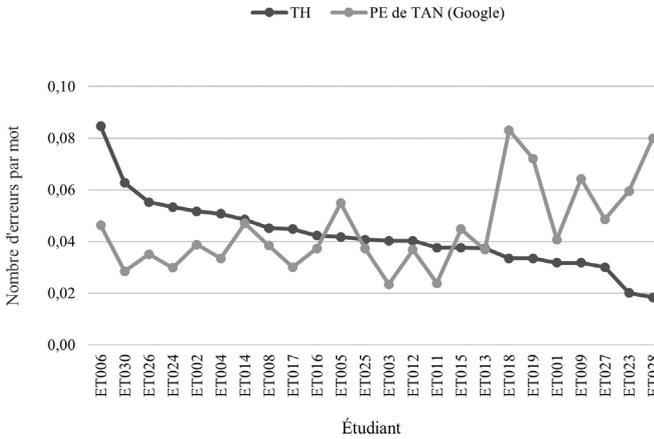
Nous avons voulu approfondir ces observations en analysant les résultats par étudiant.

4.6.1. Nombre moyen d'erreurs

Les données ont à nouveau été normalisées en prenant comme référence le nombre d'erreurs par mot du TS. Sur les figures 14 et 15, les traductions humaines (TH) sont classées de manière croissante: de la production qui contient le plus grand nombre

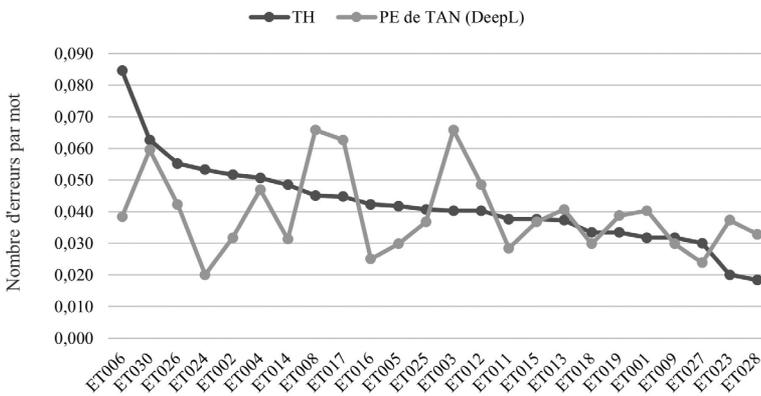
d'erreurs par mot à celle qui en contient le moins, autrement dit, de la moins bonne TH à la meilleure, indépendamment du TS.

FIGURE 14
Comparaison du nombre d'erreurs par étudiant (Adéquation) Google



Sur cette figure 14, la TH de l'étudiant 25 (ET025) se situe à la médiane (0,041). Nous constatons d'emblée que les étudiants qui commettent en moyenne le plus grand nombre d'erreurs d'adéquation en TH commettent systématiquement moins d'erreurs dans leur PE. En effet, pour tous les étudiants dont la TH se situe au-dessus de cette médiane ($> 0,041$) (sur la gauche du graphique), il semblerait que le recours à la PE de TA neuronale ait été bénéfique et leur ait permis d'aboutir à un texte cible final contenant moins d'erreurs en moyenne que leur TH, à l'exception de l'étudiant ET005. Et l'inverse se vérifie également : la majorité des étudiants dont la TH se situe en dessous de cette médiane ($< 0,041$) ont produit une PE qui contient (nettement) plus d'erreurs par mot du TS que n'en contient leur production humaine (sur la droite du graphique).

FIGURE 15
Comparaison du nombre d'erreurs par étudiant (Adéquation) DeepL



Les trois figures suivantes, 15, 16 et 17, nous permettent d’observer ce phénomène en phase d’adéquation avec les PE de TA générée avec *DeepL* (figure 15), ainsi qu’en phase d’acceptabilité avec les PE de TA produite par *Google* (figure 16) et par *DeepL* (figure 17). La même tendance que sur la figure 14 semble se dégager, quoique beaucoup moins nettement.

FIGURE 16
Comparaison du nombre d’erreurs par étudiant (Acceptabilité) Google

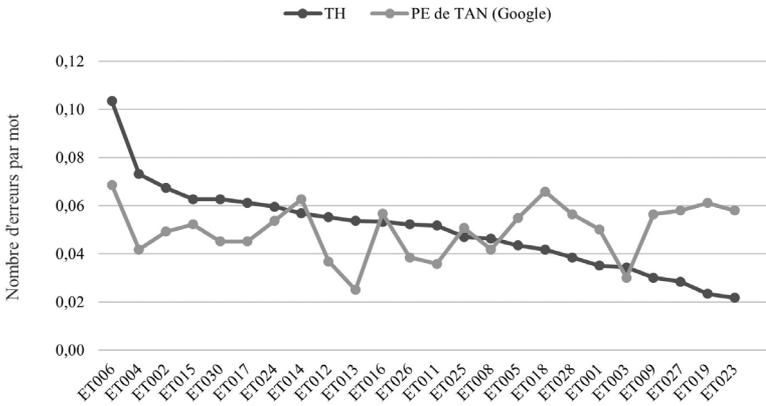
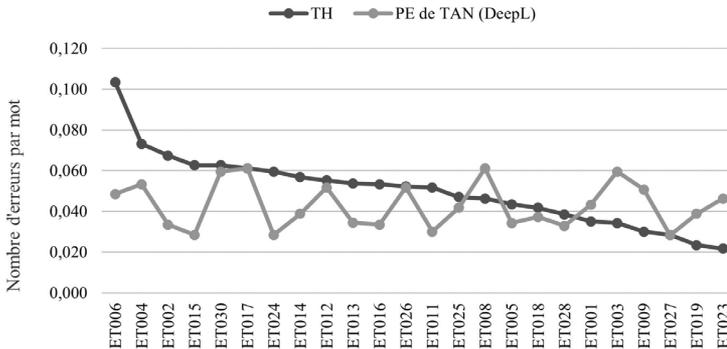


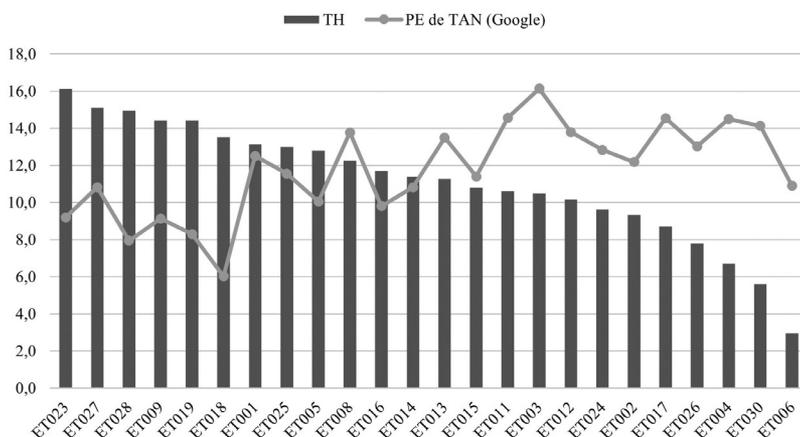
FIGURE 17
Comparaison du nombre d’erreurs par étudiant (Acceptabilité) DeepL



4.6.2. Notation

Nous avons souhaité vérifier si cette tendance se dégageait également dans la notation. Les figures 18 et 19 ont été obtenues en comparant, pour un même étudiant, les notes attribuées en TH à celles attribuées en PE de TA (*Google Traduction* pour la figure 18 et *DeepL* pour la figure 19). Sur ces figures, les TH sont classées à présent de manière décroissante: de la meilleure production à la moins bonne.

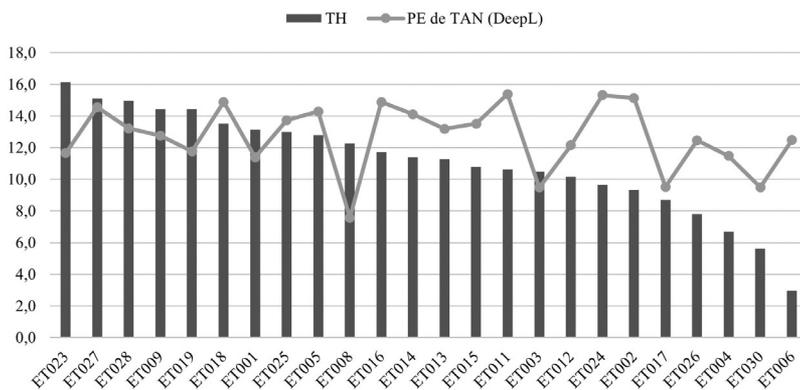
FIGURE 18
Effet nivelant en PE de TAN (Google)



Sur la figure 18, nous voyons que les étudiants dont la note en TH est située en deçà de la médiane (11,3) – ET013 à ET006 inclus – obtiennent systématiquement une meilleure note en PE de TA générée avec *Google Traduction*. De plus, les étudiants que nous considérons comme les plus faibles, à savoir ceux dont la note arrondie en TH n’atteint pas le seuil de réussite de 10/20 – ET002 à ET006 inclus –, ont obtenu une note nettement plus élevée en PE. Nous en concluons que ces étudiants bénéficient particulièrement de la TA neuronale. À l’inverse, les étudiants dont les productions humaines sont situées au-delà de la valeur médiane – ET023 à ET014 inclus – ont tous obtenu une note plus basse en PE de TA neuronale, à l’exception de l’étudiant ET008. Cette figure nous permet de mettre au jour un « effet nivelant » en PE de TA neuronale (voir notamment Killman 2018¹⁸).

Sur la figure 19, nous observons également ce phénomène en PE de TA générée avec *DeepL*, toutefois de manière moins marquée.

FIGURE 19
Effet nivelant en PE de TAN (DeepL)



Les résultats que nous venons de présenter et particulièrement les figures 12, 14 et 18 nous amènent à tirer la conclusion suivante : la PE de TA neuronale (en particulier *Google Traduction*) chez les étudiants aurait un effet nivelant sur la qualité des textes cibles. En comparaison de la TH, ce mode de traduction permettrait aux étudiants faibles de s'en sortir en réduisant le nombre moyen d'erreurs et en leur évitant d'obtenir de (très) mauvaises notes, mais ce processus aurait également pour effet de diminuer les performances des bons éléments. Nos observations révèlent que la (PE de) TA neuronale a eu un effet d'homogénéisation notable sur la qualité des textes cibles. Ces observations nécessitent bien évidemment d'être étayées par d'autres travaux.

4.7. Qualité des TA brutes

Nous avons décidé d'intégrer, pour chaque TS, les deux TA brutes (*Google Traduction* et *DeepL*) générées en septembre 2020 au corpus d'évaluation humaine à l'insu des évaluateurices. Chaque sortie brute a donc été évaluée par deux évaluateurices, au même titre que les productions des étudiants, ce qui nous a permis de leur attribuer une note sur 20 (tableau 4). Pour rappel, les trois TS n'ont pas été évalués par les mêmes binômes d'évaluateurices. Dans un souci de pertinence et de validité des résultats, ces six TA brutes (deux par TS) ont bien entendu été exclues des figures que nous venons de présenter.

TABLEAU 4
Notes attribuées aux TA brutes

Note générale /20		
Texte	<i>Google Traduction</i>	<i>DeepL</i>
Texte 1 (presse)	10,9	10,7
Texte 2 (économique)	10,9	13,4
Texte 3 (vulg. scientifique)	4,3	6,8

Sur le tableau 4, il est très interpellant de constater que quatre sorties de TA sur six ont obtenu une note supérieure au seuil de réussite fixé à 10 sur 20, même s'il faut bien noter que trois d'entre elles sont inférieures à la note de 11 sur 20. C'est la TA produite par *DeepL* pour le texte économique qui reçoit la note la plus élevée : 13,4 sur 20. En comparaison, les TA du texte de vulgarisation scientifique ont récolté de très mauvaises notes (4,3 et 6,8 sur 20). On pourrait donc légitimement se demander si le texte 3 ne se prêtait pas moins bien à une traduction machine en raison de plusieurs facteurs (thématique ; complexité syntaxique, terminologique...) ou si le binôme d'évaluateurices du texte 3 n'a pas été plus sévère dans son évaluation de la qualité par rapport aux deux autres textes. Si l'on compare à présent les moyennes générales (figure 9) aux notes attribuées par les évaluateurices humaines aux TA brutes (tableau 4), il apparaît que les textes postédités ont été jugés, en moyenne, de meilleure qualité que les TA brutes. La PE a donc permis de faire systématiquement remonter la moyenne des notes à une exception : la moyenne générale en PE de TA générée par *DeepL* pour le texte 2 est de 13,2 sur 20, soit une valeur plus basse que la note qui a été attribuée par les évaluateurices à la TA brute (13,4). Cette note est également supérieure aux valeurs médianes obtenues en TH, en PE de TA (*Google*) et en PE de TA (*DeepL*) pour le texte 2.

TABLEAU 5

Distribution des erreurs en TA brute (Acceptabilité et Adéquation)

Erreurs		Texte 1		Texte 2		Texte 3	
		Google	DeepL	Google	DeepL	Google	DeepL
Acceptabilité	Grammaire/syntaxe	1,5	2	2	2	4	4
	Orthographe	0	1	0	0	0	0
	Typographie/ponctuation	0	0,5	3	2	2,5	0
	Lexique	5	3,5	1,5	0,5	5,5	7,5
	Incompréhension	1	1	1,5	0	2,5	2
	Style	12	10	9,5	6	12,5	11,5
	Cohérence	2	0,5	3	1,5	0	0
Adéquation	Contresens/non-sens	1	2	1	1	1,5	1
	Glissement de sens	2	1	3	2,5	2,5	0
	Omission	0	0	0	0	0	0
	Ajout	0	0	0	0	0	0
	Calque	10	12	12	9	26	23
	Vocabulaire	3,5	1	0	0,5	1,5	2
	Autre	0	0	0	0	0	0

En observant le tableau 5, nous constatons que les catégories d'erreurs les plus représentées sont celles du *style* et du *lexique* en phase d'acceptabilité. Tandis qu'en phase d'adéquation, les évaluateuses ont jugé que les TA brutes contenaient principalement des calques fautifs, en particulier le texte 3, ce qui corrobore les résultats issus de nos précédentes expériences également menées en contexte pédagogique (Schumacher 2019 et 2020a). Cette tendance marquée de la TA à procéder par calque a, par ailleurs, déjà été relevée à maintes reprises dans d'autres études (voir notamment Depraetere 2010; Martikainen et Kübler 2016; Martikainen et Mestivier 2020). Aussi, il est bon de veiller à ce que les apprenants soient particulièrement attentifs à ce type d'erreurs récurrentes en TA.

5. Conclusion

La présente étude s'inscrit dans le cadre d'une thèse de doctorat et dans la continuité d'un prétest mené en 2018 (Schumacher 2020b). Dans cet article, nous avons présenté les résultats de l'analyse quantitative d'une deuxième recherche expérimentale comparative de TH et de PE (anglais-français) effectuée en 2021 avec des étudiants de 2^e année du master en traduction à finalité spécialisée de l'ULiège. Les productions des apprenants ont été soumises à évaluation humaine à l'aveugle.

Les résultats présentés dans cet article rejoignent les principaux résultats de notre prétest 2018. Notre première hypothèse a été confirmée puisque nous pouvons affirmer que, comparativement à la TH, la PE de TA neuronale n'a pas eu d'influence négative sur la qualité des productions des étudiants. En effet, il ressort de l'évaluation humaine que les textes postédités n'ont pas été jugés de moindre qualité que les TH. Nous avons même relevé une moyenne générale des notes en TH systématiquement plus basse par rapport aux deux autres modes de traduction; cette observation

vaut pour les trois TS (figure 9). Notre analyse montre que, globalement, les textes postédités ont été considérés de meilleure qualité que les TH. Preuve en est que si l'on considère l'ensemble des productions (tous TS confondus), la moyenne des notes et la valeur médiane obtenues en PE de TA (que ce soit *Google Traduction* ou *DeepL*) sont supérieures aux valeurs mesurées en TH (figure 12). De plus, l'évaluation humaine de la qualité a mis en évidence de très hauts taux de réussite en PE de TA neuronale, particulièrement avec le moteur *DeepL*. Nous pouvons d'ailleurs affirmer que, dans cette étude, la probabilité qu'une production soit jugée de bonne qualité et que celle-ci obtienne une note de réussite (supérieure ou égale à 10/20), tous TS confondus, est plus élevée lorsque l'étudiant postédite de la TA neuronale générée par *DeepL* que lorsqu'il traduit humainement ou encore que lorsqu'il postédite de la TA neuronale générée par *Google Traduction*. En outre, nous avons constaté un resserrement des notes autour de la moyenne en PE de TA neuronale (*Google Traduction* et *DeepL*) (figure 11), quel que soit le TS, ainsi qu'un resserrement des notes autour de la médiane (figure 12). Selon nous, ces resserrements traduisent un effet nivelant en PE sur la qualité des textes cibles en comparaison de la TH. D'ailleurs, l'analyse des notes par étudiant (figures 18 et 19) a permis de confirmer l'existence d'un tel effet. La PE de TA neuronale – en particulier *Google Traduction* – a donné lieu à une relation inverse entre le niveau d'un étudiant en TH et la qualité de sa PE (figure 18) : ce mode de traduction permettrait aux étudiants faibles de s'en sortir en réduisant le nombre moyen d'erreurs et en leur évitant d'obtenir de (très) mauvaises notes, mais il aurait également pour effet de diminuer les performances des bons éléments. Ceci corrobore les résultats d'autres études ayant déjà mis au jour cet effet nivelant en PE chez les étudiants (voir notamment Garcia 2011 et Killman 2018). Il apparaît également que la qualité d'un texte postédité dépend bel et bien du système de TA neuronale employé. Tel que nous l'avions postulé, *DeepL* a été plus performant que *Google Traduction* dans cette expérience. Effectivement, l'analyse du corpus a révélé qu'avec *DeepL*, les étudiants ont produit de meilleurs résultats, puisque leurs productions contiennent, en moyenne, moins d'erreurs que les PE de TA générée par *Google Traduction*. C'est le cas pour les deux phases d'évaluation (acceptabilité et adéquation) et quel que soit le TS (figure 13).

Nous tenons à souligner que cette expérience contrôlée a été menée avec des apprenants en traduction pour la combinaison de langues anglais-français. Il va de soi que ces résultats ne valent que dans le cadre spécifique dans lequel notre recherche s'est tenue. Nous n'avons effectivement aucune prétention de généralisation, car nous sommes consciente que toute modification des conditions expérimentales (population observée, choix du TS, critères d'évaluation de la qualité, moteur de TA employé, etc.) pourrait aboutir à d'autres résultats. De futurs travaux de recherche sont par conséquent nécessaires pour pouvoir confirmer et généraliser nos observations.

Pour conclure, si ces résultats tendent à plaider en faveur de l'usage de la TA neuronale en contexte académique et alors que l'enseignement des outils de TA est de plus en plus présent dans les formations en traduction (O'Brien et Ehrensberger-Dow 2020), nous tenons à insister sur la nécessité de sensibiliser les étudiants aux limites, toujours existantes, de cette technologie ainsi qu'aux enjeux, y compris éthiques, qu'elle pose (Loock 2020). Pour ce faire, de plus en plus de voix s'élèvent en faveur d'une *Machine Translation Literacy* (Bowker et Buitrago-Ciro 2019). Tandis que les deux chercheurs à l'origine de ce concept ont envisagé l'usage de la TA uniquement dans le cadre

d'activités de recherche scientifique et de publication par des non-professionnels, O'Brien et Ehrensberger-Dow suggèrent d'en adopter une définition plus large : « MT literacy means knowing how MT works, how it can be useful in a particular context, and what the implications are of using MT for specific communicative needs » (2020 : 146). Loock propose de traduire ce concept en « utilisation raisonnée des outils de TA » (2020 : 270) et exhorte le traducteur humain à « définir sa plus-value, sa valeur ajoutée par rapport à une traduction obtenue de façon automatique » (Loock 2018 : 787). Aussi, nous sommes convaincue que l'un des défis majeurs dans l'enseignement de la TA est de parvenir à faire prendre conscience aux futurs traducteurs de leur valeur ajoutée par rapport à une machine qui traduit automatiquement. Comme le rappelle si bien O'Brien : « Machine Translation (MT) is an imperfect technology » (2022 : 105). Il n'existe, à l'heure actuelle, aucun outil technologique capable de remplacer (totalement) un traducteur humain pour traduire un texte de qualité publiable. Les systèmes de traduction en ligne, devenus aujourd'hui très performants et à la portée de tous, n'ont pas tué le métier de traducteur, mais ils l'ont transformé (Moorkens 2022). Ainsi, en dépit des nombreux progrès technologiques, c'est bien le traducteur humain qui demeure au cœur du processus de traduction, et il est urgent d'en prendre conscience et d'en faire prendre conscience.

NOTES

- * Affiliation : Université de Liège et Université de Genève ; unité de recherche « CIRTI » (Centre Interdisciplinaire de Recherches en Traduction et en Interprétation) de l'ULiège, Belgique.
1. Si l'on a aujourd'hui abandonné l'adjectif qualificatif « statistique », il n'en demeure pas moins que ces deux approches sont des approches statistiques comme le rappelle très justement Poibeau : « les récents systèmes de traduction automatique reposent sur une base essentiellement statistique, notamment quand il est question de réseaux de neurones » (Poibeau 2019 : 137).
 2. DENEUFBOURG, Guillaume (2021) : Traduction automatique : la dangereuse « sagesse des foules ». *La Libre*. 17 octobre 2021.
 3. « We find that our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations » (Hassan, Aue, Chen *et al.* 2018 : 1).
 4. SHOSHAN, Ofer (2018) : Make Neural Machine Translation Better, Faster – A New Way to Measure NMT Quality. *Slator*. Consulté le 28 février 2022, <<https://slator.com/make-neural-machine-translation-better-faster-a-new-way-to-measure-nmt-quality/>>.
 5. DENEUFBOURG, Guillaume (2020) : La traduction : un métier en voie de disparition ? Journée d'études : #TQ2020 | Traduction & Qualité : biotraduction et traduction automatique, Université de Lille, Lille, 31 janvier 2020.
 6. Groupe d'experts EMT (2017) : Référentiel de compétences de l'EMT. Consulté le 22 septembre 2020, <https://commission.europa.eu/system/files/2018-02/emt_competence_fw_k_2017_en_web.pdf>.
 7. MARTIKAINEN, Hanna et MESTIVIER, Alexandra (2020) : Les outils de traduction nouvelle génération : quel effet sur la qualité des textes traduits ? Journée d'études : #TQ2020 | Traduction & Qualité : biotraduction et traduction automatique, Université de Lille, Lille, 31 janvier 2020.
 8. DEEPL GMBH : DeepL Translator. Cologne : DeepL GmbH. Consulté le 22 janvier 2020, <<https://www.deepl.com/>>.
 9. DEEPL GMBH (2020) : Comparaison de DeepL Traducteur avec ses concurrents. Consulté le 26 janvier 2023, <<https://www.deepl.com/fr/quality.html>>.
 10. AVIDANet® Live est une suite logicielle de contrôle de classe permettant à un enseignant de donner des cours dans un environnement enrichi par la technologie (Televic Education 2018). Ce logiciel est proposé par Televic Education qui est l'une des quatre divisions du groupe Televic.
 11. TELEVIC EDUCATION (2018) : Manuel d'utilisation v4.1 AVIDANet® Live. Consulté le 14 février 2022, <<https://support.televic-education.com/hc/en-us/articles/213859829--FR-Manuel-d-utilisation-AVIDANet-Live>>.

12. Cet outil de traduction assistée par ordinateur (TAO) est accessible gratuitement en ligne à l'adresse: <<https://www.matecat.com>>.
13. « Le calque consiste à utiliser des éléments lexicaux qui existent dans une langue donnée avec la construction ou le sens qu'ont ces éléments dans l'autre langue » (Chuquet et Paillard 1989: 223).
14. Un non-sens est une « faute de traduction qui consiste à attribuer à un segment du texte de départ un sens erroné qui a pour effet d'introduire dans le texte d'arrivée une formulation absurde » (Delisle 2013: 649).
15. FÉDÉRATION WALLONIE-BRUXELLES (2021): Enseignement supérieur — conditions de réussite et voies de recours. Portail officiel de la Fédération Wallonie-Bruxelles. Consulté le 25 janvier 2022, <<http://www.enseignement.be/index.php?page=23812>>.
16. « Une valeur qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel il apparaît » (Grubbs 1969: 1).
17. LE PETIT ROBERT DE LA LANGUE FRANÇAISE (s.d.): monochromatique. *Dictionnaire le Petit Robert de la langue française*. <<https://petitrobert.lerobert.com/robert.asp>>.
18. « Both they [translators well above the median] and the translators at and just above the median all experienced a reduction in performance levels when post-editing, while those below it all experienced a rise, giving way to a leveling effect » (Killman 2018: 137).

RÉFÉRENCES

- ALONSO, Elisa et VIEIRA, Lucas Nunes (2020): The impact of technology on the changing role of the translator in globalized translation workflows. In: Esperança BIELSA et Dionysios KAPSASKIS, dir. *Routledge Handbook of Translation and Globalization*. Londres: Routledge, 391-405.
- BOWKER, Lynne et BUITRAGO CIRO, Jairo (2019): *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.
- CHUQUET, Hélène et PAILLARD, Michel (1989): *Approche linguistique des problèmes de traduction anglais-français*. Paris: Ophrys.
- DAEMS, Joke, VANDEPITTE, Sonia, HARTSUIKER, Robert et MACKEN, Lieve (2017): Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta*. 62(2):245-270.
- DE FARIA PIRES, Loïc (2018): Intégration de la traduction automatique neuronale à la formation universitaire des futurs traducteurs: pistes d'exploration. *Myriades*. 4:50-61.
- DE SUTTER, Gert, CAPPELLE, Bert, DE CLERCQ, Orphée *et al.* (2017): Towards a Corpus-based, Statistical Approach to Translation Quality: Measuring and Visualizing Linguistic Deviance in Student Translations. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*. 16. <<https://doi.org/10.52034/lanstts.v16i0.440>>.
- DELISLE, Jean (2013): *La Traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. (3^e éd.). Ottawa: Presses de l'Université d'Ottawa.
- DENEUFBOURG, Guillaume (2019): Post-édition de traduction automatique: se méfier des apparences. Consulté le 22 janvier 2022, <<https://www.ata-divisions.org/FLD/index.php/tag/guillaume-deneufbourg/>>.
- DEPRAETERE, Ilse (2010): What counts as useful advice in a university post-editing training context? Report on a case study. In: Viggo HANSEN et François YVON, dir. *Proceedings of the 14th annual conference of the European Association for Machine Translation*. 14th Annual Conference of the European Association for Machine Translation, Palais des Congrès de Saint-Raphaël, Saint-Raphaël, 27-28 mai 2010. Saint-Raphaël: European Association for Machine Translation.
- DEPRAETERE, Ilse, DE SUTTER, Nathalie et TEZCAN, Arda (2014): Post-edited quality, postediting behaviour and human evaluation: a case study. In: Sharon O'BRIEN, Laura WINTHER-BALLING, Michael CARL *et al.*, dir. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, 78-108.
- DIRAND, Baptiste et ROSSI, Caroline (2019): Biotraducteur et traducteur automatique. L'homme outillé et la machine-homme. *Des mots aux actes*. 8:63-78.

- FEDERICO, Marcello, BERTOLDI, Nicola, CETTOLO, Mauro *et al.* (2014) : The Matecat Tool. In: Lamie TOUNSI et Rafal RAK, dir. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. COLING 2014, Ollscoil Chathair Bhaile Átha Cliath, Dublin, 23-29 août 2014. Dublin: Dublin City University et Association for Computational Linguistics, 129-132.
- GAN, Kamel et BROC, Guillaume (2018) : *Introduction à la modélisation par équations structurales: manuel pratique avec lavaan*. Londres: ISTE Editions.
- GARCIA, Ignacio (2011) : Translating by post-editing: is it the way forward? *Machine Translation*. 25(3):217-237.
- GONZÁLEZ PASTOR, Diana (2021) : Introducing Machine Translation in the Translation Classroom: A Survey on Students' Attitudes and Perceptions. *Revista Tradumàtica: tecnologies de la traducció*. 19:47-65.
- GRUBBS, Frank E. (1969) : Procedures for detecting outlying observations in samples. *Technometrics*. 11(1):1-21.
- GUERBEROF ARENAS, Ana et MOORKENS, Joss (2019) : Machine translation and post-editing training as part of a master's programme. *JoSTrans*. 31:217-238.
- HASSAN, Hany, AUE, Anthony, CHEN, Chang *et al.* (2018) : Achieving human parity on automatic Chinese to English news translation. *arXiv preprint*. <<https://doi.org/10.48550/arXiv.1803.05567>>.
- JIA, Yanfang, CARL, Michael et WANG, Xiangling (2019) : How does the post-editing of Neural Machine Translation compare with from-scratch translation? A product and process study. *JoSTrans*. 31:60-86.
- KILLMAN, Jeffrey (2018) : Translating the same text twice: An English-Spanish comparative product study of post-edited translations vs. human translations. *The Journal of Internationalization and Localization*. 5(2):114-141.
- LAVALT-OLLÉON, Élisabeth (2007) : Avant-propos. In: Élisabeth LAVALT-OLLÉON, dir. *Traduction spécialisée: pratiques, théories, formations*. Berne: Peter Lang, ix-xii.
- LOOK, Rudy (2018) : Traduction automatique et usage linguistique: une analyse de traductions anglais-français réunies en corpus. *Meta*. 63(3):786-806.
- LOOK, Rudy (2019) : La plus-value de la biotraduction face à la machine. *Traduire*. 241:54-65.
- LOOK, Rudy (2020) : No more rage against the machine: how the corpus-based identification of machine-translations can lead to student empowerment. *JoSTrans*. 34:150-170.
- LOOK, Rudy et LÉCHAUGUETTE, Sophie (2021) : Machine translation literacy and undergraduate students in applied languages: report on an exploratory study. *Revista Tradumàtica: tecnologies de la traducció*. 19:204-225.
- MARTIKAINEN, Hanna (2019) : Post-Editing Neural MT in Medical LSP: Lexico-Grammatical Patterns and Distortion in the Communication of Specialized Knowledge. *Informatics*. 6(3):26.
- MARTIKAINEN, Hanna et KÜBLER, Nathalie (2016) : Ergonomie cognitive de la post-édition de traduction automatique: enjeux pour la qualité des traductions. *ILCEA*. 27. <<http://journals.openedition.org/ilcea/3863>>.
- MOORKENS, Joss (2018) : What to expect from Neural Machine Translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*. 12(4):375-387.
- MOORKENS, Joss (2022) : Le traducteur, une espèce en voie de disparition?. *Le Courrier de l'UNESCO*. avril-juin 2022. <<https://courrier.unesco.org/fr/articles/le-traducteur-une-espece-en-voie-de-disparition>>.
- O'BRIEN, Sharon (2022) : How to deal with errors in machine translation: Postediting. In: Dorothy KENNY, dir. *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, 105-120.
- O'BRIEN, Sharon et EHRENSBERGER-DOW, Maureen (2020) : MT Literacy: A cognitive view. *Translation, Cognition & Behavior*. 3(2):145-164.
- POIBEAU, Thierry (2016) : Traduire sans comprendre? La place de la sémantique en traduction automatique. *Langages*. 201(1):77-90.

- POIBEAU, Thierry (2019): *Babel 2.0: Où va la traduction automatique?*. Paris: Odile Jacob.
- POIBEAU, Thierry (2022): On “Human Parity” and “Super Human Performance” in Machine Translation Evaluation. In: Nicoletta CALZOLARI, Frédéric BÉCHET, Philippe BLACHE *et al.*, dir. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. LREC 2022, Le Palais du Pharo, Marseille, 20-25 juin 2022. Marseille: European Language Resources Association, 6018-6023.
- ROSSI, Caroline (2019): L'apprenti traducteur et la machine: des connaissances aux perceptions de la traduction automatique. *Des mots aux actes*. 8:93-105.
- SALDANHA, Gabriela et O'BRIEN, Sharon (2014): *Research methodologies in translation studies*. Londres, New York: Routledge.
- SCHUMACHER, Perrine (2019): Avantages et limites de la post-édition. *Traduire*. 241:108-123.
- SCHUMACHER, Perrine (2020a): La traduction automatique neuronale: technologie révolutionnaire ou poudre de perlimpinpin? Compte-rendu d'une expérience pédagogique. *Al-Kimiya*. 18:67-89.
- SCHUMACHER, Perrine (2020b): Post-édition et traduction humaine en contexte académique: une étude empirique. *Transletters*, 4:239-274.
- WAY, Andy (2018): Quality expectations of machine translation. In: Joss MOORKENS, Sheila CASTILHO, Federico GASPARI *et al.*, dir. *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer, 159-178.

ANNEXES

Annexe 1

Texte source 1 – Article de presse

Can Theresa May and her government survive? Our writers' verdicts

With two resignations in one week, four Westminster watchers discuss whether the Conservative leader and her party can cling on to power

Andrew Gimson: There's no replacement for May nor appetite for election

Theresa May is in a stronger position than the press is willing to admit. Before the election, her frailties were ignored. Since that time, a narrative of extreme vulnerability has taken hold. Yet she remains in Downing Street, and the loss of two ministers who were found in different ways to have misbehaved does not change the powerful reasons for keeping her there. This summer she offered the British people the chance to turn her into an elected dictator, and they decided in their wisdom they would rather keep her as a prime minister who must operate with a degree of tact.

There is no popular demand for another election – the last one was an election too many – and no popular call for some particular individual to replace her. Nor does the Conservative party have a replacement for her in mind. On the great issue of the day, which is Brexit, the party is split. It recognises that the referendum decision must be implemented, but also that to implement it with gung-ho gusto, of the kind that Boris Johnson could provide, would be perilously divisive.

The time may well come when in order to achieve Brexit, a Gordian knot needs to be cut. But neither the country nor the party is ready for that. The clear preference is for May to continue with the difficult, perhaps impossible, task of disentangling the knot. Her duty is to carry on with this unenviable task. And since she is a dutiful woman, that is probably what she will do.

Source: <https://www.theguardian.com/commentisfree/2017/nov/09/can-theresa-may-conservative-government-survive-writers-verdict> (consulté le 23/01/2018)

Texte source 2 – Article économique

How will the global economic order change by 2050?

Key findings

- We project that the world economy could more than double in size by 2050, assuming broadly growthfriendly policies (including no sustained long-term retreat into protectionism) and no major global civilisation-threatening catastrophes.
- Emerging markets will continue to be the growth engine of the global economy. By 2050, the E7 economies could have increased their share of world GDP from around 35% to almost 50%. China could be the largest economy in the world, accounting for around 20% of world GDP in 2050, with India in second place and Indonesia in fourth place (based on GDP at PPPs).
- A number of other emerging markets will also take centre stage – Mexico could be larger than the UK and Germany by 2050 in PPP terms and six of the seven largest economies in the world could be emerging markets by that time.
- Meanwhile, the EU27 share of world GDP could be down to less than 10% by 2050, smaller than India.
- We project Vietnam, India and Bangladesh to be three of the world's fastest growing economies over this period. UK growth has the potential to outpace the average rate in the EU27 after the transitional impact of Brexit has passed, although we project the fastest growing large EU economy to be Poland.
- Today's advanced economies will continue to have higher average incomes, but emerging economies should make good progress towards closing this gap by 2050. This will open up great opportunities for businesses prepared to make long-term investments in these markets. But this will require patience to ride out the storms we have seen recently in economies like, for example, Brazil, Nigeria and Turkey, all of which still have considerable long-term economic potential based on our analysis.
- To realise this growth potential, emerging market governments need to implement structural reforms to improve macroeconomic stability, diversify their economies away from undue reliance on natural resources (where this is currently the case), and develop more effective political and legal institutions.

Source: <https://www.pwc.com/gx/en/world-2050/assets/pwc-the-world-in-2050-full-report-feb-2017.pdf>
(consulté le 12/01/2018)

Texte source 3 – vulgarisation scientifique

What is synesthesia?

When you eat chicken, does it feel pointy or round? Is a week shaped like a tipped-over D with the days arranged counterclockwise? Does the note B taste like horseradish? Do you get confused about appointments because Tuesday and Thursday have the same color? [...] When you read a newspaper or listen to someone speaking do you see a rainbow of colors? If so, you might have synesthesia.

Synesthesia is an anomalous blending of the senses in which the stimulation of one modality simultaneously produces sensation in a different modality. Synesthetes hear colors, feel sounds and taste shapes. What makes synesthesia different from drug-induced hallucinations is that synesthetic sensations are highly consistent: for particular synesthetes, the note F is always a reddish shade of rust, a 3 is always pink or truck is always blue.

The estimated occurrence of synesthesia ranges from rarer than one in 20,000 to as prevalent as one in 200. Of the various manifestations of synesthesia, the most common involves seeing monochromatic letters, digits and words in unique colors. This is called grapheme-color synesthesia.

Synesthetes report having unusually good memory for things such as phone numbers, security codes and polysyllabic anatomical terminology because digits, letters and syllables take on such a unique panoply of colors. But synesthetes also report making computational errors because 6 and 8 have the same color and claim to prejudge couples they meet because the colors of their first names clash so hideously.

For too long, synesthetes were dismissed as having overactive imaginations, confusing memories for perceptions or taking metaphorical speech far too literally. Recent research, however, has documented the reality of synesthesia and is beginning to make headway into understanding what might cause such unusual perceptions [...]

The causes of synesthesia also remain unknown. [...] Whatever its etiology, synesthesia provides cognitive neuroscientists with a unique opportunity to learn more about how the brain creates our perceptual reality.

Source: <https://www.scientificamerican.com/article/what-is-synesthesia/> (consulté le 23/01/2018)

Annexe 2

Consignes de postédition pour obtenir une qualité comparable ou égale à une traduction humaine

Ce niveau de qualité est en général défini comme compréhensible (un lecteur comprend parfaitement le contenu du message), exact (la traduction a le même sens que le texte source) et stylistiquement correct [...]. La syntaxe est normale, la grammaire et la ponctuation sont correctes.

- Viser une traduction correcte sur le plan grammatical, syntaxique et sémantique
- Vérifier que la terminologie importante est correctement traduite [...]
- Vérifier qu'aucune information n'a été accidentellement ajoutée ou oubliée
- Réviser le contenu injurieux, inapproprié ou culturellement inacceptable
- Exploiter au maximum le résultat brut de la traduction automatique
- Appliquer les règles fondamentales d'orthographe, de ponctuation et de coupure des mots
- Vérifier que la mise en forme est correcte

Source: <https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines-french>