Université Paris Cité

CLILLAC arp

Atrium Humanités et Sciences Sociales

Mémoire de Master 2 Recherche

Articles scientifiques en TAL : évaluation de traductions automatiques et de post-éditions produites par des spécialistes du domaine et des linguistes

présenté par

MINDER Joachim

Natalie Kübler
directrice de mémoire

Alexandra Mestivier
co-directrice de mémoire

Christopher Gledhill
membre du jury

Université Paris-Cité

Master 2 Langues de spécialité, corpus et traductologie (LSCT)
UFR-EILA (Études Interculturelles de Langues Appliquées)
CLILLAC-ARP

Année 2023-2024

Cette recherche constitue une contribution au projet de recherche ANR MaTOS (https://anr-matos.github.io/)

Remerciements

Je tiens tout d'abord à remercier ma directrice de mémoire, Natalie Kübler, pour ses conseils pratiques, méthodologiques, ainsi que pour son aide lors de l'élaboration de la typologie d'erreurs et de l'annotation.

Je remercie également Alexandra Mestivier, co-directrice de mémoire, pour son aide concernant la typologie d'erreurs et l'utilisation des outils d'annotation.

Je tiens aussi à remercier Maëva Gougeaud et Michèle Colin, étudiantes en M2 LSCT, pour leur aide dans l'élaboration du manuel d'annotation.

Je souhaite également remercier Maud Bénard et Lichao Zhu pour leur aide lors des séances d'annotation collectives.

D'un point de vue plus personnel, je remercie Antoine Hilgers pour son soutien tout au long de ce travail et pour ses relectures.

Ce projet a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'avenir portant la référence « ANR-21-EXES-0002 »

Table des matières

Liste des abréviations

EBMT	traduction automatique à base d'exemples (*example-based machine translation*)
FPE	post-édition complète (*full post-editing*)
IA	intelligence artificielle
HMTE	évaluation humaine de la traduction automatique
LPE	post-édition légère (*light post-editing*)
PE	post-édition
RBMT	traduction automatique à base de règles (*rule-based machine translation*)
SMT	traduction automatique statistique (*statistical machine translation*)
TA	traduction automatique
TAL	traitement automatique des langues
TAO	traduction assistée par ordinateur
TAN	traduction automatique neuronale
TH	traduction humaine
THR	traduction humaine de référence
TQA	évaluation de la qualité de la traduction (*translation quality assessment*)
TS	texte source

1. Introduction

1.1. Contextualisation

Chaque jour, de nouvelles avancées technologiques révolutionnent notre quotidien : les smartphones, l'Internet, l'intelligence artificielle, la réalité virtuelle, l'impression 3D et la robotisation modifient profondément notre manière de vivre. Ces innovations n'épargnent aucun secteur, et les métiers de la traduction font partie des domaines de première ligne face à ces transformations rapides et révolutionnaires.

Le développement rapide de ces nouvelles technologies, autant l'intelligence artificielle que la traduction automatique ou les outils d'aide à la traduction, constitue un intérêt central en traductologie. La traduction automatique, par exemple, est de plus en plus utilisée dans le monde professionnel et dans les contextes de recherche. En effet, alors que la science et les connaissances se diffusent en grande partie en anglais (voir, par exemple, Kamadjeu 2019 ; Swales 1997), la traduction automatique apparaît comme une technologie précieuse pour rendre la science plus accessible à ceux dont la connaissance de l'anglais est limitée. Par ailleurs, de plus en plus de formations en traduction incluent désormais dans leur programme des cours ou des projets visant à sensibiliser les apprenants aux outils de TA et, dès lors, à développer chez les étudiants de nouvelles compétences (voir, par exemple, Gledhill & Zimina-Poirot 2019 ; Gledhill et al. 2023).

Il y a une dizaine d'années, Robert expliquait que la TA n'était qu'une sorte de brouillon servant à faciliter par la suite le réel processus de traduction :

> Pour l'heure, il n'est pas encore véritablement question de traduction automatique en tant que telle, mais plutôt de prétraduction automatique dans le sens de « premier jet effectué par une machine ». (Robert, 2010, p. 137)

Cette affirmation n'est plus tout à fait réaliste de nos jours, alors que l'utilisation des sorties brutes de TA comme traduction finale est de plus en plus fréquente. La conciliation de ce premier jet et d'une révision humaine porte le nom de post-édition, tâche lors de laquelle le post-éditeur a pour mission de rendre le produit final « humainement intelligible » (*loc. cit.*). La post-édition est dès lors une tâche qui permet d'assurer la qualité de ces traductions. Dans cette tendance à l'augmentation de la demande en traductions, la post-édition ne cesse de gagner de l'importance. En effet, les échanges de plus en plus mondialisés et le développement de

l'Internet, du *e-commerce* et des sites Web, entre autres, ne font qu'augmenter le volume de traductions, partagé entre traduction humaine, traduction automatique et post-édition (*ibid.*, p. 138). Différents publics, entre autres, sont susceptibles d'effectuer ces post-éditions : d'un côté, des spécialistes du domaine scientifique, parfois les auteurs mêmes des articles eux-mêmes ; de l'autre, des traducteurs. *A priori*, chacun de ces deux publics est bien qualifié pour prendre en charge la post-édition de ces traductions automatiques. Les spécialistes sont familiers avec le domaine qu'ils côtoient au quotidien, dont ils connaissent et maîtrisent les principes, le fonctionnement et la terminologie. Par leur formation (spécialisée ou non), les traducteurs acquièrent des compétences en transfert interlinguistique et interculturel, mais aussi en utilisation d'outils appropriés pour la traduction spécialisée, comme les corpus. Toutefois, il est pertinent de se demander si un des deux publics est plus apte que l'autre pour cette tâche de post-édition. L'un des objectifs de ce travail est de tenter de répondre à cette question.

1.2. Objectifs du travail

Ce travail vise à évaluer la qualité des traductions automatiques (TA) et des post-éditions (PE) — effectuées par deux publics différents — de résumés d'articles scientifiques portant sur le domaine du traitement automatique des langues (TAL) de l'anglais vers le français. Les TA ont été générées par trois systèmes de traduction automatique commerciaux différents : DeepL, eTranslation et Systran. Les deux publics qui ont effectué ces post-éditions sont, d'un côté, des membres de la communauté du TAL (les auteurs des articles eux-mêmes) et, de l'autre côté, des membres de la communauté linguistique (enseignants en traduction, étudiants en traduction, traducteurs et chercheurs en traduction). L'évaluation de la qualité de ces TA et de ces PE se fait par une annotation sur la base d'une typologie d'erreurs élaborée dans le cadre de ce projet. Cette évaluation humaine manuelle s'inscrit dans le cadre du projet MaTOS (Machine Translation for Open Science)[1], mené conjointement par l'équipe MLIA de l'ISIR, l'Inria, le CLILLAC-ARP et l'Inist. Dans le cadre de ce projet, des centaines de résumés d'articles scientifiques portant sur le TAL ont été traduits automatiquement et post-édités. Toutefois, seulement une vingtaine de ces articles ont été post-édités à la fois par les membres de la communauté du TAL et les membres de la communauté linguistique. Ce sont ces articles qui seront évalués dans le cadre de ce travail. Ce travail tentera de répondre à deux questions

[1] https://anr-matos.github.io/index.html.

principales : les systèmes de TA génèrent-ils des traductions de qualité comparable ? Les deux publics (spécialistes vs linguistes) ont-ils des profils de post-éditeurs différents ?

1.3. Qu'entend-on par « qualité » ?

Pour pouvoir se pencher sur la question de l'évaluation, il faut d'abord évoquer la question de la *qualité*, qui a été pendant longtemps un défi majeur en traductologie. C'est dans les années 1990 qu'on a commencé à développer les premiers cadres d'évaluation de la qualité des traductions, qui servaient à identifier et à quantifier les erreurs pour produire un score de qualité. Cet intérêt a émergé dans le domaine de la localisation logicielle. Une fois que ces cadres d'évaluation sont apparus dans cette branche de la traduction, ils étaient utilisés dans d'autres domaines. Par conséquent, la question de la qualité était vue sous un angle universel. Quelques années plus tard, les pressions qu'a commencé a connaître l'industrie de la traduction en ce qui concerne les tarifs, les délais de livraison et les volumes à traiter, mais avant tout le développement des systèmes commerciaux de TA, ont remis en question cette notion universelle et absolue de la qualité (Burchardt, 2013).

À la lumière de ces bouleversements, plusieurs chercheurs en traductologie ont proposé une définition, selon eux universelle, de la qualité :

> A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs. (Koby et al., 2014)

Alors que la notion de la qualité fait désormais consensus, la question de l'évaluation — qui sera traitée en détail dans ce travail — reste un sujet épineux.

1.4. Hypothèses

Grâce à la littérature scientifique disponible en lien avec ce travail et aux connaissances établies, il est possible de formuler différentes hypothèses en amont de ce travail.

Pour la qualité des traductions automatiques, au vu des avancées et évolutions rapides des systèmes de TA, il est pertinent d'avancer que les trois systèmes de TA seront performants, du moins d'un point de vue linguistique et stylistique ; en effet, étant donné que le domaine traité dans ce travail est spécialisé, la terminologie peut constituer un frein à la qualité des TA. Toutefois, étant donné la surreprésentation et l'exploitation massive de DeepL dans la littérature

scientifique et les évaluations hautement favorables de ce système (voir, par exemple, Hidalgo-Ternero, 2021 ; Plenter, 2023 ; Yulianto & Supriatnaningsih, 2021), on peut supposer qu'il générera des traductions automatiques de qualité supérieure aux deux autres.

En ce qui concerne la qualité des post-éditions, plusieurs hypothèses peuvent être formulées. D'un côté, on peut s'attendre à ce que les spécialistes du TAL produisent des post-éditions plus correctes sur les plans terminologique et phraséologique. De l'autre, les traducteurs, les linguistes et les enseignants/étudiants en traduction étant des (futurs) spécialistes de la langue et du transfert interlinguistique, leurs post-éditions devraient être plus correctes d'un point de vue linguistique et stylistique. Enfin, quant au transfert de contenu et à la fidélité, les deux publics devraient *a priori* produire des post-éditions de qualité comparable, étant donné que les membres de la communauté du TAL maîtrisent le domaine et que les membres de la communauté linguistique disposent d'excellentes compétences en compréhension et en transfert des langues.

1.5. Structure du travail

Ce travail est divisé en plusieurs parties distinctes. Tout d'abord sera dressé un état de l'art portant sur l'histoire et le développement de la traduction automatique, la post-édition et les différentes méthodes d'évaluation (automatiques et humaines). Ensuite, j'exposerai ma méthodologie, de la définition du corpus source à la génération des traductions automatiques et des post-éditions. Sera également présenté le processus d'annotation du corpus parallèle, ainsi que le programme et la typologie d'erreurs utilisés. La partie principale de ce travail est la présentation des statistiques et l'analyse des résultats, tant des métriques automatiques réalisées par l'Inria que de l'évaluation humaine qui fait l'objet de ce travail. Par la suite, il convient de répondre, entre autres, aux deux questions principales formulées ci-dessus. En annexe se trouve le manuel d'annotation élaboré parallèlement à ce travail. Il sert de guide à l'annotation d'erreurs dans la cadre de l'évaluation de traductions humaines, automatiques ou de post-éditions. C'est également dans ce manuel qu'est expliquée en détail la typologie d'erreurs développée aux fins de ce projet.

2. État de l'art

Avant de passer à l'annotation et à l'analyse du corpus à proprement parler, il convient de définir deux notions importantes — celles de la traduction automatique neuronale (désormais TAN) et de la post-édition (désormais PE) — et d'évoquer leur enjeu pour la science ouverte. De plus, il est également essentiel de dresser un état de l'art sur les différentes méthodes existantes d'évaluation des traductions.

2.1. La traduction automatique : une histoire météorique

Depuis plusieurs années, plus spécifiquement depuis les derniers progrès technologiques, l'automatisation occupe une place de plus en plus prépondérante dans de nombreux secteurs, et la traduction et ses métiers connexes n'y font pas exception. Par ailleurs, la course à la productivité a engendré le développement d'une « nouvelle » technologie, la traduction automatique neuronale (TAN) (Ragni & Nunes Vieira, 2022, p. 137). Il ne s'agit pas réellement d'une nouvelle technologie, mais plutôt d'un « nouveau paradigme de traduction automatique » (Forcada, 2017, p. 291). En effet, la TAN est une amélioration de modèles de TA précédents.

C'est à la moitié du XX^{e} siècle qu'a débuté la traduction automatique, avec la traduction automatique à base de règles (*rule-based machine translation*, RBMT) (Schmidhofer & Mair, 2018, p. 165). Dans ce contexte de guerre froide entre l'URSS et les États-Unis et de la course à l'espace, c'est l'IBM (*International Business Machines Corporation*) et l'Université de Georgetown qui créent ce premier paradigme de TA, fonctionnant alors avec les langues russe et anglaise (Loffler-Laurian, 1996, pp. 35–36). La RBMT est un paradigme dans le cadre duquel la TA repose sur des entrées de dictionnaires et des règles linguistiques, codées par des linguistes, ces connaissances guidant le modèle au long de la traduction (*loc. cit.* ; Shiwen & Xiaojing, 2015, p. 186).

Les règles linguistiques évoluant sans cesse et les langues n'obéissant pas toutes aux mêmes règles, ce paradigme a laissé place à la traduction automatique à base d'exemples (*example-based machine translation*, EBMT) (Wang et al., 2022, p. 143). Ce premier paradigme de TA dit *corpus-based* est créé au milieu des années 1980 (Hutchins, 1995). Dans ce modèle, l'intervention humaine, c'est-à-dire l'élaboration de règles par l'humain, n'est plus requise, car la machine exécute toutes les tâches. En effet, dans l'EBMT, la machine extrait « des connaissances de traductions existantes (exemples) en vue de faciliter la traduction de

nouveaux énoncés » (Wong Tak-ming & Webster, 2015, p. 137). Ce paradigme alimenté par des corpus parallèles peut dès lors être comparé aux outils de traduction assistée par ordinateur, les mémoires de traductions utilisées sur ces outils de TAO étant des sortes de corpus parallèles (*ibid.* 137-138). Toutefois, l'EBMT pose certains problèmes : la qualité de la TA dépend fortement du degré de similarité entre l'énoncé qui est retrouvé dans le corpus d'apprentissage et le nouvel énoncé ; dès lors, la TA est plus acceptable lorsque le programme retrouve des segments comparables et similaires (Wang et al., 2022, p. 144).

Par conséquent, l'apparition d'un nouveau paradigme dans les années 1990 — la traduction automatique statistique (*statistical machine translation*, SMT) — était inévitable pour améliorer le paradigme précédent (Wang et al., 2022, p. 144). En effet, les corpus parallèles sont ici exploités d'une manière différente :

> Dans ces corpus parallèles, différentes unités sont mises en relation, du paragraphe au mot, en passant par l'expression et la phrase. Sont ensuite créées des statistiques de fréquence : celles-ci permettent de déterminer à quelle fréquence des mots ou expressions retrouvés dans le corpus se succèdent. (Minder, 2023, p. 5)

En 2003, la TA statistique connaît une grande avancée : l'apparition des modèles de *phrase-based SMT*. Dans ces modèles sont observés des segments entiers constituant des unités sémantiques. Ils seront plus tard adoptés par Google pour créer une plateforme de traduction automatique en ligne, puis par Microsoft et Baidu (Wang et al., 2022, p. 144).

Enfin, depuis 2014 se développe la traduction automatique neuronale (TAN), qui a une histoire « courte, mais météorique » (Ragni & Nunes Vieira, 2022, p. 137). Ce paradigme est considéré comme une extension des deux précédents, à savoir l'EBMT et la SMT, car il fonctionne également sur la base de corpus parallèles volumineux, mais il se différencie de ses prédécesseurs par l'intégration d'une nouvelle fonctionnalité : le réseau de neurones, une technologie de l'apprentissage profond (*deep learning*). Tous les éléments constituant ce système — les neurones — sont interconnectés, et leur activation ainsi que leur production dépendent largement de leur stimulus. En effet, cela s'apparente à la façon dont les statistiques de fréquence étaient exploitées dans le précédent modèle, c'est-à-dire que les connexions entre les neurones n'ont pas toutes le même poids : les exemples trouvés dans le corpus influencent la force des liens entre chaque neurone. En fin de compte, la traduction résulte de la somme de toutes ces connexions qui se sont établies entre les neurones (Forcada, 2017, pp. 292–295). La

TAN fonctionne en trois étapes successives. La première est celle de l'entraînement, étape durant laquelle l'« on détermine le poids et la force des connexions entre chaque neurone afin d'obtenir le résultat attendu » (*ibid.*, p. 295). Cette étape se fait sur la base d'un corpus d'entraînement très volumineux, sa taille influençant la durée de son entraînement, qui peut durer entre quelques jours et plusieurs mois (*loc. cit.*). L'étape suivante est celle de l'encodage (*encoding*) de la phrase à traduire, qui se fait grâce à la recherche de représentations (E). La dernière étape est celle du décodage (*decoding*), c'est-à-dire celle de la traduction. Lors de cette étape, le système génère deux vecteurs : un vecteur initial D et un vecteur de probabilités P (*ibid.*, pp. 297–298). Ces trois étapes font partie du modèle de base de la traduction automatique neuronale, à savoir le modèle *seq2seq* (*sequence to sequence*) ou l'architecture encodeur-décodeur, dont voici une image illustrant les étapes de l'encodage et du décodage.

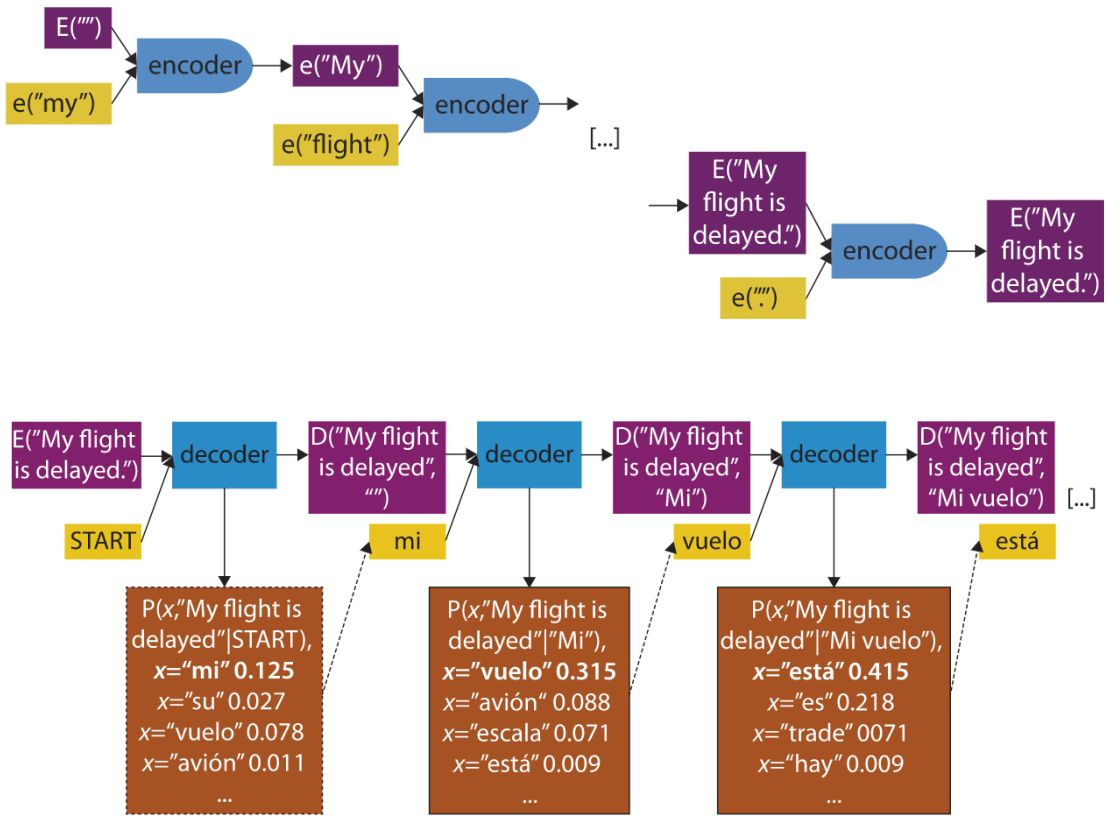

Figure 1 : schéma de l'encodage et du décodage par Forcada, 2017, pp. 298–299

Ce modèle *seq2seq* est généralement renforcé par certains mécanismes assurant une traduction de meilleure qualité, dont le mécanisme d'attention, la convolution (*ibid.*, pp. 299-300) ou encore l'encodage bidirectionnel (Wang et al., 2022, p. 146). L'attention occupe une place

prépondérante, étant donné que c'est le mécanisme qui est utilisé dans les architectures de type « Transformer » utilisées aujourd'hui. L'attention — très utile pour la traduction de phrases plus longues — permet aux neurones de garder chacune des sorties produites en mémoire, ce qui améliore la sortie finale (Vaswani et al., 2017). La convolution se distingue quant à elle des architectures encodeur-décodeur récurrentes (FIGURE 1) — où l'encodage se fait sur les plongements des mots sources un à un — de la manière suivante :

> Instead of producing an encoding of the whole source sentence by recursively ingesting the embeddings of source words one by one, their decoder produces representations of each word by taking into account a few words (let's say 2) to the left and to the right of it. (Forcada, 2017, p. 298)

Enfin, l'encodage bidirectionnel, comme son nom l'indique, permet à l'encodeur de faire des statistiques sur les couches cachées à la fois à de gauche à droite et de droite à gauche, c'est-à-dire vers les phrases passées et futures du texte source. (Wang et al., 2022, p. 145). Ce mécanisme n'est pas le même que l'attention, qui elle fonctionne sur le texte produit (cible).

2.2. Introduction à la post-édition

Étant donné que ce travail vise à étudier des post-éditions produites par différents publics, il convient de définir de manière précise la post-édition, ses différents niveaux ainsi que la qualité attendue de la PE et ses avantages et limites.

2.2.1. Tentative de définition

Comme expliqué dans l'introduction, il y a une dizaine d'années, Robert (2010) définissait la TA comme une sorte de brouillon et de prétraduction permettant de faciliter par la suite le vrai processus de traduction. Alors qu'il arrive que certaines personnes utilisent les sorties brutes de la TA comme version finale, cette définition n'est pas tout à fait vraie.

La révision humaine de la traduction automatique porte le nom de post-édition, qui est une tâche lors de laquelle un post-éditeur est chargé de rendre la traduction « humainement intelligible » (*loc. cit.*), mais pas seulement. L'on peut trouver un grand nombre de définitions, bonnes et moins bonnes, de la post-édition. Robert (*loc. cit.*), par exemple, définit la post-édition comme « l'activité qui consiste à repasser derrière un texte prétraduit automatiquement pour le rendre humainement intelligible ». Cependant, les définitions de ce terme omettent

souvent des notions essentielles et introduisent fréquemment une confusion entre *post-édition* et *révision*. Dès lors, la définition que je propose pour ce terme est la suivante : tâche lors de laquelle un individu a pour mission de corriger la sortie brute de traduction automatique en vue de la rendre au mieux bonne, dans le cas d'une post-édition complète, ou au moins acceptable, dans le cas d'une post-édition légère. Évidemment, la PE n'est considérée ni comme une traduction — étant donné qu'un premier jet (la TA) sert de base — ni comme une révision — la révision étant souvent l'étape finale, soit la relecture du produit fini (*ibid.*, p. 141). Les concepts de post-édition complète et légère seront définis plus en détail au point suivant.

2.2.2. Différents niveaux de post-édition

En 2010, le TAUS[2], un groupe de réflexion sur l'automatisation et l'innovation en traduction réunissant des utilisateurs et des fournisseurs de technologies et de services de traduction, se penche de manière plus précise sur le processus de post-édition en formulant des consignes. Deux niveaux de post-édition sont alors distingués : d'un côté, la post-édition de qualité « acceptable » ; de l'autre, la post-édition de qualité « comparable ou égale à une traduction humaine » (Taus, 2010, pp. 3–4). Parmi les professionnels de la traduction, cette distinction est faite également, mais les deux niveaux de PE portent des noms spécifiques, respectivement la LPE (*light post-editing*, ou post-édition légère) et la FPE (*full post-editing*, post-édition complète) (Deneufbourg, 2019). Pour une LPE, l'idéal est de viser un produit final correct sur le plan sémantique, exempt d'ajouts et d'omissions, de propos injurieux et (culturellement) inappropriés, d'utiliser au maximum la sortie brute de la TA et de veiller au respect de l'orthographe et de la grammaire ; pour la LPE, il n'est pas nécessaire d'apporter des corrections purement stylistiques ou « de modifier la structure des phrases dans le seul but d'améliorer la fluidité du texte » (Taus, 2010, pp. 3–4). Pour la FPE, la liste des consignes est plus stricte. Pour cette tâche, il faut :

- « Viser une traduction correcte au niveau grammatical, syntaxique et sémantique
- Vérifier que la terminologie importante est correctement traduite et que les termes non traduits font partie de la liste des termes à ne pas traduire du client
- Vérifier qu'aucune information n'a été accidentellement ajoutée ou oubliée

[2] Translation Automation User Society

- 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 » (*ibid.*, p. 4).

En effet, la FPE visant à obtenir un produit final de qualité comparable à une traduction humaine (TH), les modifications à apporter sont plus nombreuses.

2.2.3. Une qualité comparable à la traduction humaine ?

L'idée d'obtenir, par le biais de la post-édition, un texte final comparable à une traduction humaine est-elle un fantasme ? Différentes études récentes portant sur la qualité des post-éditions réalisées par différents publics semblent révéler des résultats divergents, parfois contradictoires, qui ne pourraient *a priori* pas s'expliquer par les profils différents des participants de ces études.

Une étude menée par Deneufbourg (2019) révèle des statistiques intéressantes. Selon son étude, les étudiants en traduction seraient *a priori* de meilleurs candidats pour une post-édition légère (LPE, *light post-editing*), alors que les professionnels — traducteurs — semblent l'être pour la *full post-editing* (FPE). En effet, les professionnels semblent apporter trop de modifications dans le cadre de la LPE. À l'inverse, les étudiants semblent quant à eux apporter trop peu de modifications pour la FPE. Un autre problème s'est posé :

> [Les étudiants et les professionnels] semblent en outre induits en erreur par le pouvoir persuasif de la NMT [TAN] et se laissent plus facilement tromper par l'apparente fluidité des traductions, même lorsque la machine commet de lourdes erreurs de sens. (*Ibid.*, p. 5)

Deneufbourg observe également, en règle générale, un rapport inversement proportionnel entre la sévérité des erreurs commises par le programme de TA et l'effort cognitif nécessaire à leur correction :

> Ainsi, pour prendre un exemple volontairement caricatural, un contresens (gravité élevée) qui serait lié à l'oubli d'une négation pourra se corriger en quelques secondes en ajoutant les mots « ne… pas » à la phrase (effort faible). En revanche, des phrases bancales sur le plan stylistique (gravité faible) nécessiteront le plus

> souvent une reformulation de l'ensemble de la phrase (effort élevé). Cette tendance pourrait selon nous encourager les post-éditeurs à laisser les phrases en l'état lorsqu'elles sont *borderline*. (*Loc. cit.*)

Dans une étude comparative de traduction spécialisée impliquant des étudiants de master (anglais-français), Martikainen & Mestivier (2020) ont observé une amélioration de la qualité en post-édition de TAN avec une réduction du nombre d'erreurs de moitié (en moyenne) en comparaison avec la traduction humaine. Elles ont toutefois constaté la présence d'un *Post-editese* (langage de post-édition), ce que Schumacher & Sutera (2022, p. 3) définissent comme « la reprise des structures syntaxiques du texte de départ, par le calque et par l'uniformisation des solutions de traduction ». Ce phénomène d'interférences entre le texte source et le texte cible plus marqué avec la post-édition qu'avec la traduction humaine est également confirmé dans une étude menée par Čulo et al. (2014).

Dans le cadre d'une étude réalisée avec une trentaine d'étudiants de master en traduction, Jia et al. (2019) sont, quant à eux, arrivés à la conclusion suivante : la post-édition de sorties de TAN avec la paire de langues anglais-chinois permet *a priori* d'obtenir des produits finaux tout aussi fidèles[3] (*accuracy*) et fluides (*fluency*) que les traductions humaines. Toutefois, un étude menée par Lacruz et al. (2014) montre que les erreurs de langue (appartenant à la catégorie *fluency*) sont plus faciles à détecter et à corriger que les erreurs de transfert de contenu.

Dans la même veine, une étude impliquant des étudiants et des traducteurs professionnels menée par Daems et al. (2018) (avec la paire de langues anglais-néerlandais) arrive à la conclusion suivante : il n'existe pas de *Post-editese*, et il serait dès lors impossible de distinguer une FPE d'une traduction humaine, tant sur les plans lexicaux et sémantiques que syntaxiques.

Enfin, une étude de Screen (2019) — dans le cadre de laquelle ont été utilisés d'une part un système d'*eye-tracking* et d'autre part un système de notation par les participants — démontre que la FPE n'affecte en aucun cas la lisibilité et la compréhension par les participants du texte post-édité : « The implications of this for the translation industry is that post-edited

[3] La notion de fidélité ayant de multiples acceptions, elle est ici utilisée comme une référence à la restitution du *sens* du texte source.

texts, given these results, are perceived by end users to be just as readable and comprehensible as [human-]translated ones. » (P. 147)

Ces différentes études montrent à quel point les résultats divergent, et qu'il est difficile de parler d'une PE de qualité comparable à la traduction humaine, même si certaines études démontrent que cet « idéal » est possible. Toutefois, ces observations dépendent de plusieurs facteurs, notamment de la qualité de la traduction automatique initiale, de la paire de langues concernée, mais aussi (et surtout) du profil des post-éditeurs (étudiants, professionnels de la traduction, etc.).

2.2.4. Bénéfices et limites de la post-édition

Bien qu'elle ne permette pas (toujours) d'aboutir à des résultats comparables à une traduction humaine, la post-édition permet d'éviter certaines erreurs. Par exemple, Schumacher (2019, p. 112) démontre que la post-édition de TAN permet aux étudiants d'éviter des calques fautifs que l'on retrouve dans les traductions humaines. Dans cette même étude, elle démontre que la PE évite également l'emploi de régionalismes, qu'elle retrouve dans son corpus de traductions humaines. Enfin, la TAN ayant « intégré l'ensemble des règles grammaticales […] de la langue française » (*ibid.*, p. 114), la présence d'erreurs grammaticales dans les PE est très rare, voire inexistante, toujours selon Schumacher (2019).

Ces effets positifs peuvent être nuancés, voire totalement contredits par d'autres chercheurs. Par exemple, certains semblent montrer que la PE favorise le recours à des équivalences formelles et littérales (voir par exemple Depraetere (2010) et Martikainen & Kübler (2016)). La PE présente également un risque accru d'omissions, de problèmes de référents, de biais de genre (pour plus d'informations sur les biais de genre en TA, cf. Wisniewski et al., 2021), d'incohérences temporelles ou encore d'irrégularités terminologiques (pour plus d'informations à ce sujet, voir par exemple Schumacher 2019 ; Bénard et al. 2022).

2.3. L'enjeu de la TAN et de la PE pour la science ouverte

Dans de nombreux domaines de la vie, l'anglais s'est imposé comme la langue de communication — la *lingua franca* —, créant ainsi un monolinguisme et une sorte de domination linguistique de l'anglais (cf. Swales, 1997), et les domaines scientifiques n'y font pas exception (Gordin, 2015). Pour favoriser l'accessibilité des ressources scientifiques par le plus large public possible, la traduction automatique constitue une passerelle entre ces

ressources et les différents publics non anglophones (voir par exemple Fiorini et al. (2020) et Bénard et al. (2023)). De nombreux projets de traduction automatique pour des documents scientifiques ont déjà vu le jour, notamment le projet ANR/COSMAT (à ce sujet, voir Lambert et al. (2012)) ou encore *Health In My Language*[4], un projet européen. Plus récemment, OPERAS a lancé le projet « Translations and Open Science : Study on machine translation evaluation in the context of scholarly communication »[5]. Enfin, le projet ANR MaTOS[6] (Machine Translation for Open Science), conduit par l'Inria, l'INIST, l'ISIR et le CLILLAC-ARP a pour objectif de développer des nouvelles méthodes de TA et des métriques automatiques pour l'évaluation de la qualité des traductions. Ce travail s'inscrit dans la problématique du projet MaTOS et est par ailleurs réalisé dans le cadre de cette ANR.

Utiliser les résultats bruts de TA peut se révéler risqué au vu des limites formulées ci-dessus, d'autant plus que certains chercheurs se sont positionnés sur la question de savoir si les contenus scientifiques traduits automatiquement avaient leur place dans les espaces de dépôt. Selon Dony et al. (2023), cette pratique soulèverait plusieurs questions. Par exemple, les outils d'IA ne sont pas censés être listés comme auteurs d'un contenu, étant donné qu'ils ne peuvent pas s'en porter responsables :

> [T]here are issues of accountability related to the issue of authorship. In cases where a mistranslated document is used, who is accountable? Transferring accountability to the person depositing an exclusively MTed resource into a repository may be problematic or difficult to explain or enforce. (*Loc. cit.*)

Dès lors, il est intéressant de se pencher sur la post-édition de TA et d'observer son potentiel pour la science ouverte. La question de la qualité des PE a déjà été abordée précédemment, mais il convient également d'établir un état de l'art, permettant d'avoir une vue d'ensemble sur les différentes méthodes d'évaluation (automatiques et humaines) utilisées pour évaluer la qualité de traductions automatiques et de post-éditions.

---

[4] https://www.himl.eu/

[5] https://operas.hypotheses.org/5630

[6] https://anr-matos.github.io/index.html

2.4. L'évaluation de la qualité : une question épineuse

La question de l'évaluation de la qualité des traductions (TQA) soulève quelques débats. Il existe deux grandes méthodes servant à observer la qualité des traductions automatiques et des post-éditions. La première consiste en une analyse automatique et automatisée de TA et de PE ; la seconde, en une analyse humaine ou manuelle. Il convient de noter que ces deux méthodes ne s'excluent pas mutuellement ; au contraire, elles sont souvent combinées en vue de permettre des résultats plus précis, combinant d'une part des statistiques et des métriques et d'autre part des observations et analyses humaines et manuelles. Avant d'exposer ces deux méthodes, il convient d'exposer en quoi la question de l'évaluation de la qualité est complexe.

La qualité et son évaluation sont en réalité des questions qui se confondent. Alors que l'évaluation de la qualité est centrale dans le domaine de la traduction, « aucune méthode d'évaluation ne fait consensus entre les acteurs, […] dont les approches diffèrent parfois fortement » (Bénard, 2020, p. 20). Les diverses méthodes d'évaluation étant profondément liées aux différentes théories de traduction existantes, il existe de nombreuses perceptions de la qualité (*loc. cit.*). On peut distinguer deux familles d'approches : les approches psychologistes et les approches textuelles ou discursives, incluant les approches descriptives et linguistiques (House, 2015, p. 10-14). Pour les approches psychologistes, c'est la fonction communicative du texte de départ qui est centrale (*ibid.*, p. 11) :

> [Les approches psychologistes] reposent essentiellement sur la réponse du lecteur au texte cible en cherchant à déterminer si elle est équivalente à celle d'un lecteur du texte source ou cohérente avec la fonction du texte cible (compréhension et intelligibilité du texte). Il est cependant difficile d'évaluer la perception et la réception du texte par un lecteur […]. (Bénard, 2020, p. 20)

Quant aux approches discursives ou textuelles, il s'agit plutôt de comparaisons du texte source et du texte cible en vue de dégager des régularités sur les plans sémantique, syntaxique, pragmatique ou stylistique (House, 2015, p. 13).

En outre, les typologies d'erreurs qui sont utilisées pour ces évaluations diffèrent en fonction du contexte : l'enseignement, la recherche en TA ou encore le monde professionnel ne se servent pas forcément des mêmes typologies d'erreurs. Par ailleurs, les méthodes d'évaluation de la qualité des traductions humaines et automatiques diffèrent également (Bénard, 2020, p. 21).

Enfin, le recours accru aux outils de traduction assistée par ordinateur ainsi que l'essor de la post-édition soulèvent une autre question : où se situe la frontière entre l'évaluation des systèmes de TA et celle des biotraductions (Moorkens et al., 2018, p. 27) ? En d'autres termes, convient-il, et si oui, par quels moyens, de distinguer l'évaluation des traductions automatiques de l'évaluation des traductions humaines ? Par extension, quelle est la typologie des erreurs qu'il convient d'utiliser pour la post-édition ?

2.4.1. Évaluation automatique de TA et de PE

Si l'on peut analyser la qualité des traductions (automatiques) ou même des post-éditions de manière manuelle et humaine, pourquoi mettre au point des systèmes automatiques et automatisés ? Les analyses humaines ont un coût, tant financier que temporel. En effet, les méthodes d'évaluation humaines peuvent durer de quelques semaines à quelques mois. De plus, la main-d'œuvre humaine ne peut pas être réutilisée (Papineni et al., 2001), c'est-à-dire que les évaluations humaines ne peuvent pas être remises en place de manière systématique et automatique, contrairement aux métriques automatisées. Dès lors, l'évaluation humaine n'est pas la solution optimale pour suivre l'évolution quotidienne des systèmes de traduction automatique, qui connaissent une évolution spectaculaire (*loc. cit.*). Ces constats sont à l'origine du développement de méthodes d'évaluation automatiques, qui n'ont cessé et ne cessent de voir le jour.

2.4.1.1. Les systèmes de première génération

Le premier système de TQA (*translation quality assessment*) largement utilisé à être mis au point est BLEU, un score qui se veut rapide à calculer, peu onéreux, indépendant de la paire de langues et doté d'une grande corrélation avec l'évaluation humaine. En effet, l'idée principale derrière cette technologie est que plus la TA est proche d'une traduction humaine professionnelle, plus sa qualité est bonne. Dès lors, pour évaluer sa qualité, BLEU utilise une ou plusieurs traductions humaines de référence. Par conséquent, pour calculer le score BLEU, deux ingrédients sont nécessaires, à savoir une métrique calculant la similarité (précision) entre la TA et la biotraduction de référence ainsi qu'un corpus de traductions humaines de bonne qualité (Papineni et al., 2001, p. 311).

La pierre angulaire de la métrique BLEU est la mesure de précision. Pour calculer cette donnée, la machine compte le nombre de mots candidats dans la traduction (« candidate

translation words ») (*loc. cit.*) se trouvant dans une traduction humaine de référence (THR), puis divise ce nombre par le nombre total de mots dans la traduction candidate (« candidate translation ») (*loc. cit.*). Pour éviter que la présence en masse de certains mots ne fausse le résultat, est mis en place le principe de *modified unigram precision*, calculé comme suit :

> [O]ne first counts the maximum number of times a word occurs in any single reference translation. Next, one clips the total count of each candidate word by its maximum reference count, adds these clipped counts up, and divides by the total (unclipped) number of candidate words. (*Ibid.*, p. 312)

Ce calcul est effectué pour chaque n-gramme ; c'est-à-dire que l'on divise le nombre de n-grammes candidats par la valeur de référence maximum, puis ces deux valeurs sont additionnées, et enfin, cette somme est divisée par le nombre de n-grammes candidats. Cette métrique de *modifier n-gram precision* est censée mesurer la fidélité et la fluidité. En effet, une traduction candidate dans laquelle sont utilisés les mêmes mots (1-grammes) que dans la THR assure la fidélité, alors que les correspondances de n-grammes plus longues garantissent la fluidité (*ibid.*, p. 313). Cette métrique, à elle seule, permettrait déjà de distinguer les TA des TH. En effet, le graphique ci-dessous (cf. Figure 2) aide à visualiser la différence de précision entre les traductions humaines (bleu foncé), plus précises, et les traductions automatiques (bleu clair), moins précises, en fonction de la longueur des n-grammes.

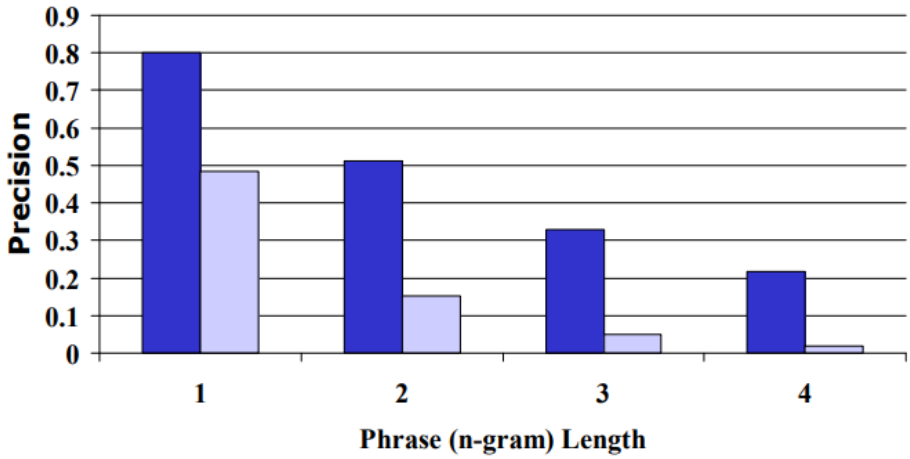

Figure 2 : tableau proposé par Papineni et al., 2001, p. 313

La deuxième métrique derrière le score BLEU est celle relative à la longueur des phrases candidates. La métrique de *n-gram precision* permet déjà, dans une certaine mesure, de

pénaliser les TA quand certains mots ne sont pas retrouvés dans les THR ou lorsqu'un mot est retrouvé plus souvent dans la traduction candidate que dans les THR. Traditionnellement, l'on utilisait le *recall* (rappel) pour cette métrique, mais étant donné que BLEU inclut plusieurs traductions de référence, qui peuvent chacune utiliser un mot différent pour traduire le même mot source, cette métrique ne serait pas optimale. Pour pallier ce manque, BLEU utilise la *sentence brevity penalty*. La traduction candidate doit dès lors correspondre à la THR en termes de longueur de phrase. Quand il y a plusieurs références, ne sera considérée que la phrase de référence qui correspond le plus à la longueur de la phrase candidate.

Les deux métriques de précision et de longueur sont ensuite combinées pour donner un chiffre entre 0 et 1, 0 correspondant à une mauvaise traduction et 1 à une traduction parfaite (*ibid.*, pp. 314-315).

Le score BLEU est dès lors calculé sur des quantités et sans prise en compte du texte source, ce qui peut soulever des questions, notamment la question suivante : peut-on vraiment évaluer la qualité d'une TA en ne tenant compte que de textes de référence et sans prise en considération du texte de départ ?

En 2005 est présentée la métrique METEOR, qui calcule un score sur la base de correspondances entre la traduction candidate et les textes de référence également. Cette métrique a été mise au point pour résoudre les problèmes liés au score BLEU, en l'occurrence l'absence de la métrique de rappel, la prise en compte de n-grammes plus longs, la non-prise en compte de mots concurrents ou synonymes ou des variantes morphologiques d'un mot ou encore le fait d'attribuer un score de 0 à toute une phrase comportant un n-gramme avec un score de 0, ce qui fausse le score de toute la phrase (Banerjee & Lavie, 2005, p. 65-67).

Pour le calcul du score METEOR sont utilisées deux mesures : la première est celle de la précision (P) et la seconde, celle du rappel (*recall*, R). La précision est calculée comme suit :

$$P = \frac{m}{w_t}$$

où *m* correspond au nombre d'unigrammes de la traduction candidate retrouvés dans la THR et w_t correspond au nombre d'unigrammes dans la traduction candidate.

Le rappel est calculé comme suit :

$$R = \frac{m}{w_r}$$

où *m* correspond au nombre d'unigrammes de la traduction candidate retrouvés dans la THR et w_r correspond au nombre d'unigrammes dans la THR.

Ensuite est calculé le score *Fmean* en combinant le score de précision et le score de rappel de la manière suivante :

$$Fmean = \frac{10PR}{R + 9P}$$

Fmean est donc un score qui donne plus de poids au rappel qu'à la précision.

Cependant, les trois mesures (*R*, *P*, et *Fmean*) ne se basent que sur les correspondances d'unigrammes. Pour prendre en compte les correspondances plus longues est calculée une *chunk penalty* (une pénalité relative aux syntagmes) (*p*), c'est-à-dire que le système regroupe les unigrammes (les mots isolés) en n-grammes (*chunks*, c'est-à-dire en syntagmes) retrouvés tels quels dans la traduction de référence, et plus il y aura de correspondances, moins la pénalité sera élevée :

$$p = 0{,}5 \left(\frac{c}{u_m}\right)^3$$

où *c* correspond au nombre de syntagmes dans la traduction candidate et u_m au nombre d'unigrammes dans la traduction candidate. Enfin, le score METEOR — entre 0 et 1 — est calculé comme suit :

$$Score = Fmean\ (1 - pénalité)$$

(Banerjee & Lavie, 2005, p. 67-69)

En comparaison avec d'autres métriques de TQA, METEOR enregistre de meilleurs scores de correspondances avec les évaluations humaines. À titre d'exemple, METEOR correspond aux TQA humaines avec un score de 0,964, alors que BLEU obtient un score de 0,817. Les meilleurs résultats de METEOR peuvent s'expliquer par l'utilisation de la métrique du rappel (*ibid.*, p. 69).

2.4.1.2. Les systèmes plus récents

Toutefois, les deux approches BLEU et METEOR ne se concentrent que sur des phénomènes au niveau du lexique. Pour pallier le manque de modèles plus dynamiques, quinze ans plus tard

est développé un système de TQA neuronal, COMET, qui atteint une correspondance sans précédent avec les scores d'évaluation humaine.

> Modern neural approaches to MT result in much higher quality of translation that often deviates from monotonic lexical transfer between languages. For this reason, it has become increasingly evident that we can no longer rely on metrics such as BLEU to provide an accurate estimate of the quality of MT. (Rei et al., 2020, p. 1)

COMET est un système qui sert à entraîner des modèles d'évaluation de TA multilingues et adaptables. Ce système génère des estimations prévisionnelles d'évaluations humaines, c'est-à-dire adaptées et optimisées pour différents types d'évaluations humaines de la qualité de TA, telles que *Direct Assessments* (DA), *Human-mediated Translation Edit Rate* (HTER) ou encore d'autres métriques respectant le cadre de *Multidimensional Quality Metric* (MQM). C'est le premier système qui ne s'appuie pas uniquement sur des traductions de référence, mais aussi sur le texte source, et il prend également en compte directement des prévisions d'évaluation humaine (Rei et al., 2020). *Direct Assessments* (DA) sont des évaluations humaines pures. La mesure HTER, une variante de la mesure TER, quant à elle, sert à calculer le nombre minimum de modifications à apporter à l'hypothèse (la TA) de sorte qu'elle corresponde à l'une des traductions de référence (Snover et al., 2006, p. 225). Enfin, MQM sera définie au point 2.4.2.4.

Enfin, en 2021 est développée une autre métrique dépassant les performances de 16 anciennes métriques sur 22 : le score BART. Ce système utilise une hypothèse (par exemple la traduction automatique), le texte source ainsi qu'une ou plusieurs traductions humaines de référence. Le score BART entend prendre en compte les facteurs traditionnels utilisés pour l'évaluation humaine de traductions, à savoir l'informativité (*informativeness*), la pertinence (*relevance*), la fluidité (*fluency*), la cohérence (*coherence*), la factualité (*factuality*), la couverture sémantique (*semantic coverage*) et l'adéquation (*adequacy*) (Yuan et al., 2021, p. 1-2).

Plus concrètement, le score BART sert à évaluer, sur la base d'un modèle préentraîné de type seq2seq, quatre aspects. Le premier est celui de la fidélité (*faithfulness*), qui est calculée du texte source vers l'hypothèse (TA). Ce premier aspect sert à mesurer le niveau de probabilité que l'hypothèse ait été générée sur la base du texte source. Avec cette mesure sont évaluées la factualité, la pertinence, mais aussi la cohérence et la fluidité. Ensuite est mesurée la précision (*precision*) du texte de référence vers la traduction automatique. Cela permet d'évaluer si

l'hypothèse a été générée sur la base d'un texte de référence et si celle-ci est précise. Par la suite est utilisé le rappel (*recall*) de l'hypothèse vers le texte de référence, de la même manière que dans le cadre du score METEOR. Le rappel permet dès lors d'évaluer la couverture sémantique. Enfin est calculé le *F score*, qui est une moyenne arithmétique de la précision et du rappel :

> [*F score*] [c]onsider[s] both directions and use[s] the arithmetic average of Precision and Recall ones. This version can be broadly used to evaluate the semantic overlap (informativeness, adequacy […]) between reference texts and generated texts. (*Ibid.*, p. 5)

Par conséquent, le score BART est effectivement censé couvrir les sept critères mentionnés ci-dessus, ce qui en ferait un score fiable. En effet, les comparaisons des performances de différents scores ont permis de démontrer que, dans la plupart des cas, le score BART obtient de meilleurs résultats (*ibid.*, pp. 5-7).

Comme expliqué sous ce point, de nombreux auteurs s'accordent sur le fait que l'évaluation automatique n'est pas aussi précise que l'évaluation humaine, mais qu'elle est plus efficace d'un point de vue financier et moins chronophage. En effet, l'évaluation humaine nécessite plus de temps et plus de main-d'œuvre, et est dès lors plus onéreuse. Toutefois, l'idée que l'évaluation humaine est plus précise fait consensus. Il existe diverses méthodes d'évaluation humaine, qui seront en partie couvertes par le point suivant.

2.4.2. Évaluation humaine de TA et de PE

Les méthodes d'évaluation humaines sont les plus anciennes, mais aussi les plus précises. Comme expliqué, elles sont souvent remplacées par des métriques automatiques (ou semi-automatiques) à cause de leur caractère chronophage et de leur non-répétitivité. Les méthodes d'évaluation humaines permettent d'identifier, de manière précise, les forces et les faiblesses d'une TH, d'une TA ou d'une traduction hybride, à savoir une PE. Ces méthodes s'inscrivent par conséquent dans les approches typologiques et linguistiques évoquées ci-dessus (Bénard, 2020).

La littérature scientifique sur la question de l'évaluation humaine de la traduction automatique suggère que plusieurs méthodes d'évaluation humaine de la traduction automatique (HMTE, *human machine translation evaluation*) co-existent. Traditionnellement, les méthodes de HMTE impliquent des annotateurs (humains) :

> Human MT evaluation approaches employ the (often tacit) knowledge of human annotators to assess the quality of automatically produced translations along the two axes of target language correctness and semantic fidelity. (Vela et al., 2014)

Lorsque plusieurs annotateurs participant à une même évaluation, il peut être intéressant d'utiliser le coefficient *kappa*, qui est un score inter-annotateurs, à savoir un score qui calcule le taux d'accord entre les différents annotateurs (cf. Fleiss, 1971).

La méthode d'évaluation humaine qui semble la plus simple est celle qui consiste en l'établissement de classements entre différentes hypothèses de traduction automatique (Vela et al., 2014). Ces classements peuvent être effectués par le biais de nombreux outils. Par exemple, Federmann (2012) a mené une expérimentation d'évaluation humaine de TA par le biais de classements. Pour ce faire, il a utilisé Appraise, une boîte à outils open source développée dans le cadre du projet EuroMatrixPlus[7]. Dans le cadre de cette expérimentation, la première tâche demandée aux participants était de classer différentes hypothèses de TA (en allemand) d'une même source (en anglais). Les traductions évaluées étaient des TA d'un système personnalisé hybride par rapport à une traduction de référence. Dès lors, l'objectif de cette expérimentation était d'observer si le système qu'ils avaient développé permettait d'améliorer la qualité des traductions.

Il existe également des méthodes indirectes de HMTE. Par exemple, des tests de compréhension à la lecture peuvent servir à cette fin. En effet, les métriques mesurant uniquement la similarité entre une TA et une traduction de référence, elles ne permettent pas d'évaluer directement la compréhension et la lisibilité des TA. Pour pouvoir observer la compréhensibilité d'une traduction, la participation d'humains est indispensable. Par conséquent, Weiss & Ahrenberg (2012) ont mené une expérimentation consistant en des tests de compréhension. Les textes choisis pour cette expérimentation étaient des sites web en polonais traduits en anglais par trois systèmes de TA : Google Translate, Bing Translator et Systran. Avant les tests de compréhension, les organisateurs ont annoté les TA avec six types d'erreurs, à savoir les omissions, les ajouts, les non-traductions, les mauvaises traductions, les formes incorrectes et l'ordre des mots. Les participants aux tests ne pouvaient pas voir ces annotations. Pour le test de compréhension, ils ont sélectionné deux textes sources, chacun avec

[7] https://www.euromatrixplus.net/.

sa meilleure et sa moins bonne TA en fonction de l'annotation d'erreurs. Les 20 participants (dont dix avaient l'anglais comme langue maternelle et les dix autres avaient au minimum un niveau C1) devaient ensuite répondre à des questions de contenu avec différents niveaux de difficulté (trois niveaux), sans savoir quel système de TA était à l'origine de la traduction. Les questions étaient formulées sur la base des textes sources. Les auteurs ont observé que ce n'était pas le nombre d'erreurs qui influençaient la compréhension, mais les types d'erreurs et leurs gravités. En effet, les erreurs de mauvaise traduction et de non-traduction sont les plus problématiques. En revanche, les erreurs de forme incorrecte n'ont presque aucun impact sur la compréhension. Dans le cadre d'une étude similaire, Maney et al. (2012) ont quant à eux observé que les erreurs les plus critiques pour la compréhension étaient les erreurs d'omission et de modification de l'ordre des mots. Enfin, d'autres études (cf. Condon et al. 2010, Vilar et al. 2006, Popović & Ney 2007) démontrent que les erreurs de mauvais choix lexicaux et les omissions ont l'influence la plus négative sur la compréhension.

Des métriques automatiques peuvent également être utilisées dans le cadre de l'évaluation humaine. Par exemple, la métrique HTER (voir 2.4.1.2.) peut être calculée sur la base de post-éditions. Est alors calculé le nombre de modifications apportées dans la post-édition pour corriger la traduction automatique correspondante (cf. Vela et al., 2014 ; Snover et al., 2006). Snover et al. (2006) montrent qu'il y a une corrélation forte entre le score HTER et les évaluations humaines traditionnelles de l'adéquation et de la fluidité.

La méthode d'annotation par types d'erreurs étant retenue pour ce travail, nous établissons un état de l'art sur les différentes méthodes d'annotation et les différentes typologies existantes. Souvent, les évaluateurs catégorisent les erreurs par types d'erreurs. Pour ce faire sont utilisées différentes typologies d'erreurs. Les différentes typologies sont souvent des sujets de débats, et aucune ne semble faire l'unanimité. Par ailleurs, il existe à la fois des typologies générales, mais aussi des typologies spécialisées, c'est-à-dire basées sur des difficultés et des catégories linguistiques précises ou ayant une portée ou une finalité spécifique.

2.4.2.1. Des typologies générales

En 2006, Vilar et al. ont proposé une typologie d'erreurs pour plusieurs paires de langues, à savoir chinois-anglais, espagnol-anglais et anglais-espagnol. Cette typologie est souvent utilisée pour des annotations peu granulaires, étant donné qu'elle contient assez peu de catégories d'erreurs. D'ailleurs, l'expérimentation de compréhension à la lecture de Weiss & Ahrenberg (2012) mentionnée au point précédent utilise la typologie de Vilar et al. (2006) pour l'annotation des erreurs dans les traductions automatiques. Dans cette typologie à trois niveaux sont distinguées cinq catégories (voir Figure 3). Cette typologie est cependant quelque peu problématique, puisqu'elle « regroupe […] sous un même type d'erreur des occurrences lexicales, morphologiques et syntaxiques très différentes » (Bénard, 2020, p. 28).

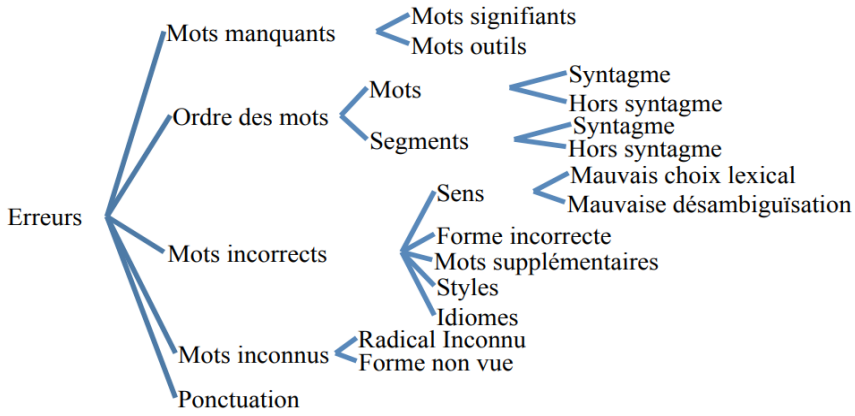

Figure 3 : typologie d'erreurs de Vilar et al. (2006), traduite et proposée par Esperança-Rodier & Becker, 2018, p. 7

En 2018, Popović résume les différentes typologies d'erreurs développées depuis celle de Vilar et al. La première est celle proposée par Farrús, Costa-Jussà, Mariño et Fonollosa en 2010, qui est une typologie simple en un seul niveau, développée à la base pour une évaluation bidirectionnelle espagnol < > catalan. Leur typologie comprend simplement les erreurs morphologiques, les erreurs lexicales, les erreurs orthographiques, les erreurs syntaxiques et les erreurs sémantiques (Farrús et al., 2010).

Par la suite, en 2012, Kirchhoff, Capurro et Turner ont développé une typologie sous forme de schéma d'annotation ayant pour objectif l'analyse des préférences des utilisateurs par rapport à la catégorisation des erreurs de traduction. Cette typologie (cf. Figure 4) est plus complète que celle proposée par Farrús et al. en 2010.

Level 1	Level 2
Missing words	Content words
	Function words
Extra words	Content words
	Function words
Word order	Local range
	Long range
Morphology	Verbal
	Nominal
Word sense error	
Punctuation	
Spelling	
Capitalisation	
Untranslated	Medical term
	Proper name
Pragmatics	
Diacritics	
Other	

Figure 4 : typologie de Kirchhoff et al. (2012) par Popović, 2018, p. 135

En 2012 également, Stymne et Ahrenberg proposent une typologie (cf. Figure 5) prenant en compte le taux d'accord entre les différents annotateurs ainsi que la gravité des erreurs. Cette catégorisation a été testée avec des traductions anglais-suédois ; le premier test s'est déroulé sans exemples fournis avec la typologie, et dans le second test étaient fournis des exemples et des consignes.

Level 1	Level 2
Error rates	Missing words
	Extra words
	Wrong word
	Word order
Linguistic	Orthography
	Semantics
	Syntax
GF	Grammatical words
	Function words
Form	Morphological categories
POS+	Part of speech
	Punctuation
FA	Fluency
	Adequacy
	Neither
	Both
Reo (cause of reordering)	
Index (position of an error)	
Other (other categories)	
Ser (seriousness of an error	

L'étude a montré que dans le premier cas, le taux d'accord interannotateur n'était que de 25 %, tandis que dans le second cas l'accord montait à 40 %. Cet accord était encore meilleur dans le cadre de l'utilisation d'une typologie simplifiée avec un taux de 65 % en absence d'exemples et de consignes et un taux de 80 % avec ces éléments. (Bénard, 2020, p. 29)

Figure 5 : typologie de Stymne & Ahrenberg (2012) par Popović, 2018, p. 135

Deux ans plus tard, Federico, Negri, Bentivogli et Turchi ont proposé une typologie comparable à celle de Farrús et al. Cette typologie d'erreurs a été développée pour l'évaluation de traduction de l'anglais vers le russe, le chinois ou l'arabe. Cette typologie comprend sept catégories : les erreurs morphologiques, les erreurs de choix lexical, les ajouts, les omissions, les erreurs de casse et de ponctuation, les erreurs d'ordre des mots et enfin une catégorie intitulée « trop d'erreurs » (Federico et al., 2014).

2.4.2.2. Des typologies spécialisées

En 2016, Comelles et al. ont proposé une typologie d'erreurs axée autour de problématiques linguistiques précises et spécifiques (cf. Figure 6). Ces catégories ont dès lors un ancrage plus linguistique que les approches généralistes citées ci-dessus. Leur typologie associe des connaissances linguistiques sur différents plans, à savoir les plans orthographique, lexical, morphologique, syntaxique et sémantique.

Orthography	Capitalisation
	Punctuation
	Date, time, money
Lexical error	Multi-word expressions
	Acronyms and abbreviations
	Untranslated source words
	Omissions
	Proper nouns
Morphology	Inflectional
	Derivational
	Compounding
	Morpho-syntax
Syntax	Syntactic structure
	Word order
	Prepositions
	Relative clauses
	Ungrammatical chunks
Semantics	Lexical semantic relations (synonymy, homonymy, etc.)
	Sentence semantics

Figure 6 : typologie de Comelles et al. (2016) par Popović (2018)

En 2016, Comparin partage la typologie d'erreurs utilisée au sein de l'entreprise Unbabel, une typologie comprenant sept catégories, sur trois niveaux (cf. Figure 7).

ACCURACY	**Mistranslation**	Overly literal
		False friend
		Should not have been translated
		Lexical selection
	Omission	
	Untranslated	
	Addition	
FLUENCY	**Inconsistency**	Word selection
		Tense selection
	Coherence	
	Duplication	
	Spelling	Orthography
		Capitalization
		Diacritics
	Typography	Punctuation
		Unpaired quote marks/brackets
		Whitespace
		Inconsistency in character use
	Grammar	Propositions
		Conjunctions
		Determiners
		Part-of-speech
		Agreement
		Tense/mood/aspect
		Word order
		Sentence structure
STYLE	**Register**	
	Inconsistent register	
	Repetitive style	
	Awkward style	
TERMINOLOGY	**Noncompliance with client/company style guide**	
	Noncompliance with glossary/vocabulary	
WRONG LANGUAGE VARIETY		
NAMED ENTITIES	**Person**	
	Organization	
	Location	
	Function	
	Product	
	Amount	
	Time	
FORMATTING AND ENCODING		

Figure 7 : typologie Unbabel synthétisée par Comparin (2016)

De la même manière, mais avec seulement trois catégories linguistiques (morpho-syntaxique, lexico-syntaxique et syntaxique) — sans la catégorie sémantique —, Isabelle et al. (2017) proposent une approche typologique sur la base d'un jeu de tests (cf. Figure 8). En effet, ils ont compilé un jeu de test de 100 phrases comprenant chacune une illustration d'un phénomène linguistique spécifique pouvant poser des problèmes ou difficultés dans la traduction. C'est ensuite qu'ils ont classé ces erreurs sous trois catégories.

Category	Subcategory
Morpho-syntactic	Agreement across distractors
	through control verbs
	with coordinated target
	with coordinated source
	of past participles
	Subjunctive mood
Lexico-syntactic	Argument switch
	Double-object verbs
	Fail-to
	Manner-of-movement verbs
	Overlapping subcat frames
	NP-to-VP
	Factitives
	Noun compounds
	Common idioms
	Syntactically flexible idioms
Syntactic	Yes-no question syntax
	Tag questions
	Stranded preps
	Adv-triggered inversion
	Middle voice
	Fronted should
	Clitic pronouns
	Ordinal placement
	Inalienable possession
	Zero REL PRO

Figure 8 : typologie d'erreurs proposée par Isabelle et al. (2017)

2.4.2.3. *La typologie MeLLANGE pour l'enseignement*

Cette typologie relativement exhaustive a été mise au point dans le cadre du projet MeLLANGE, un projet européen (cf. Kübler, 2008 ; Castagnoli et al., 2011). Celle-ci entend prendre en compte les différents types d'erreurs qui peuvent être commises par les apprenants en traduction dans six différentes langues. La typologie MeLLANGE comprend neuf catégories, composées au total de 38 sous-catégories (cf. Figure 9).

Dans cette typologie d'erreurs, on remarque deux grands types d'erreurs : *content transfer* (transfert du contenu et du sens, souvent appelée *adequacy*) et *language* (langue, souvent appelée *fluency*). Cette typologique, conçue à l'origine pour l'annotation de traductions humaines, a été adaptée pour l'annotation de traductions automatiques et de post-éditions (Kübler et al., 2022).

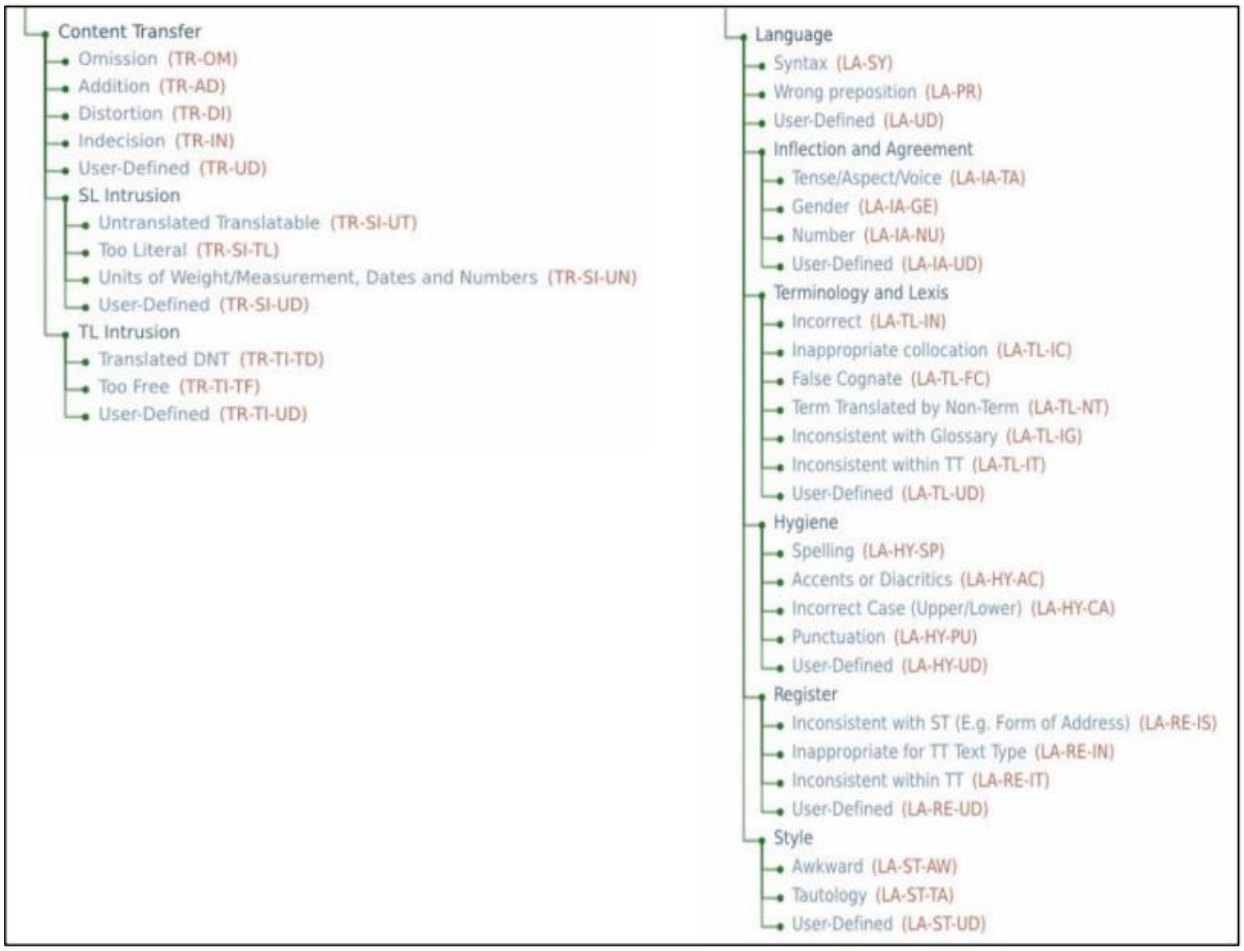

Figure 9 : typologie d'erreurs MeLLANGE par Castagnoli et al. (2011)

2.4.2.4. La typologie MQM, à portée universelle

La typologie MQM (*Multidimensional Quality Metrics*) est une typologie complète couvrant au total huit catégories de potentielles erreurs : *accuracy* (exactitude ou précision), *design* (mise en page et formatage), *fluency* (fluidité), *internationalization* (internationalisation), *locale convention* (localisation des conventions), *style* (style), *terminology* (terminologie) et *verity* (vérité), chacune divisée en sous-catégories. Cette typologie a été proposée pour l'évaluation de traductions humaines dans un contexte professionnel, mais aussi pour l'évaluation de traductions automatiques (Castilho et al., 2018, p. 16-17). Les illustrations qui suivent représentent chacune une catégorie de la typologie MQM.

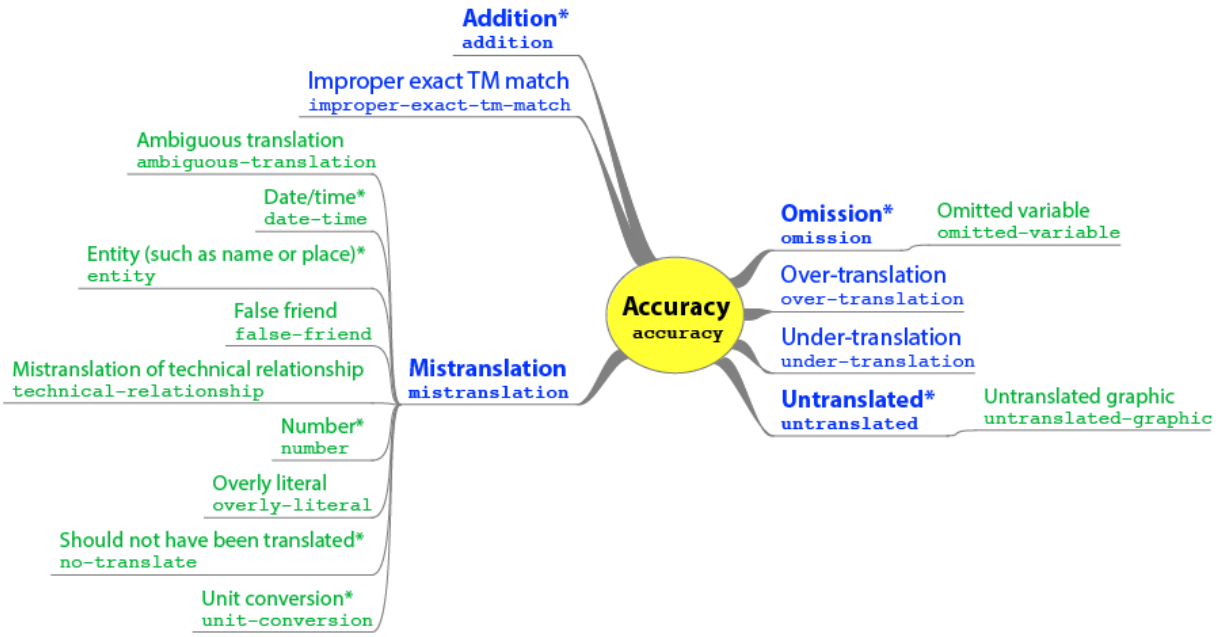

Figure 10 : typologie MQM, catégorie Accuracy[8]

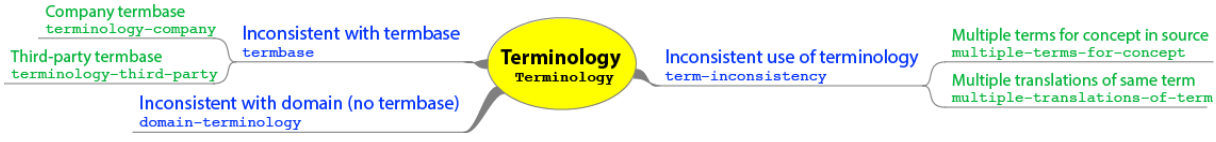

Figure 11 : typologie MQM, catégorie Terminology[9]

[8] https://web.archive.org/web/20211216145215/http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html.

[9] *Idem*.

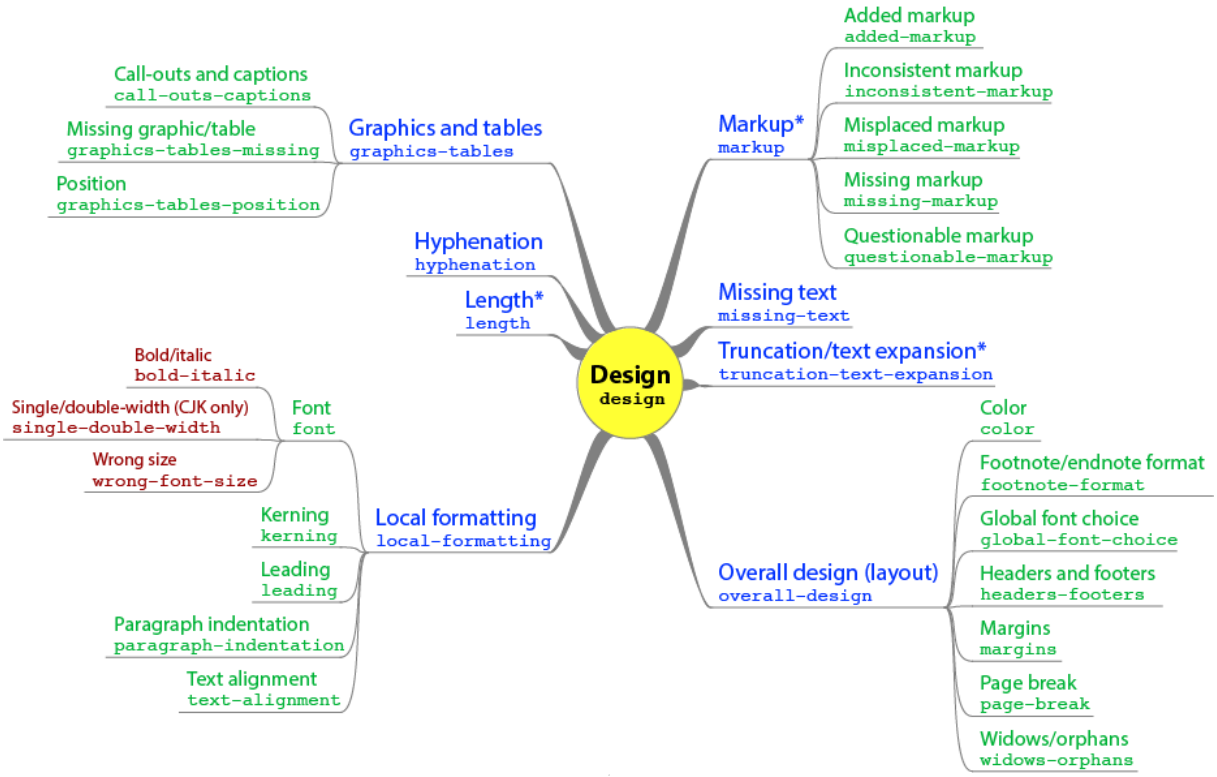

Figure 12 : typologie MQM, catégorie Design[10]

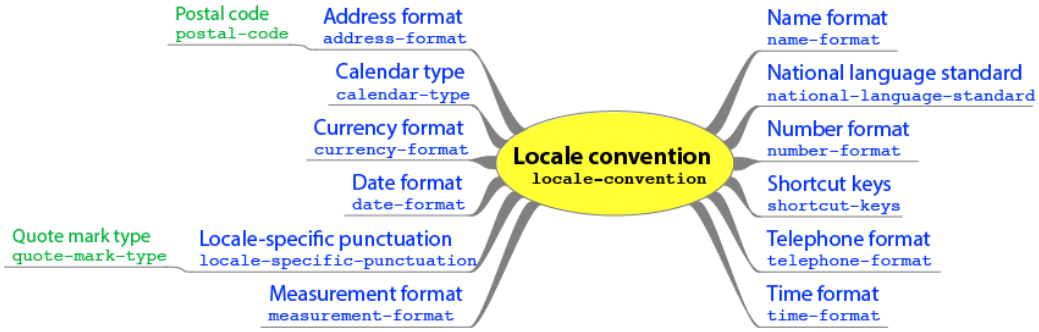

Figure 13 : typologie MQM, catégorie Locale convention[11]

Figure 14 : typologie MQM, catégorie Verity[12]

[10] *Idem.*

[11] *Idem.*

[12] *Idem.*

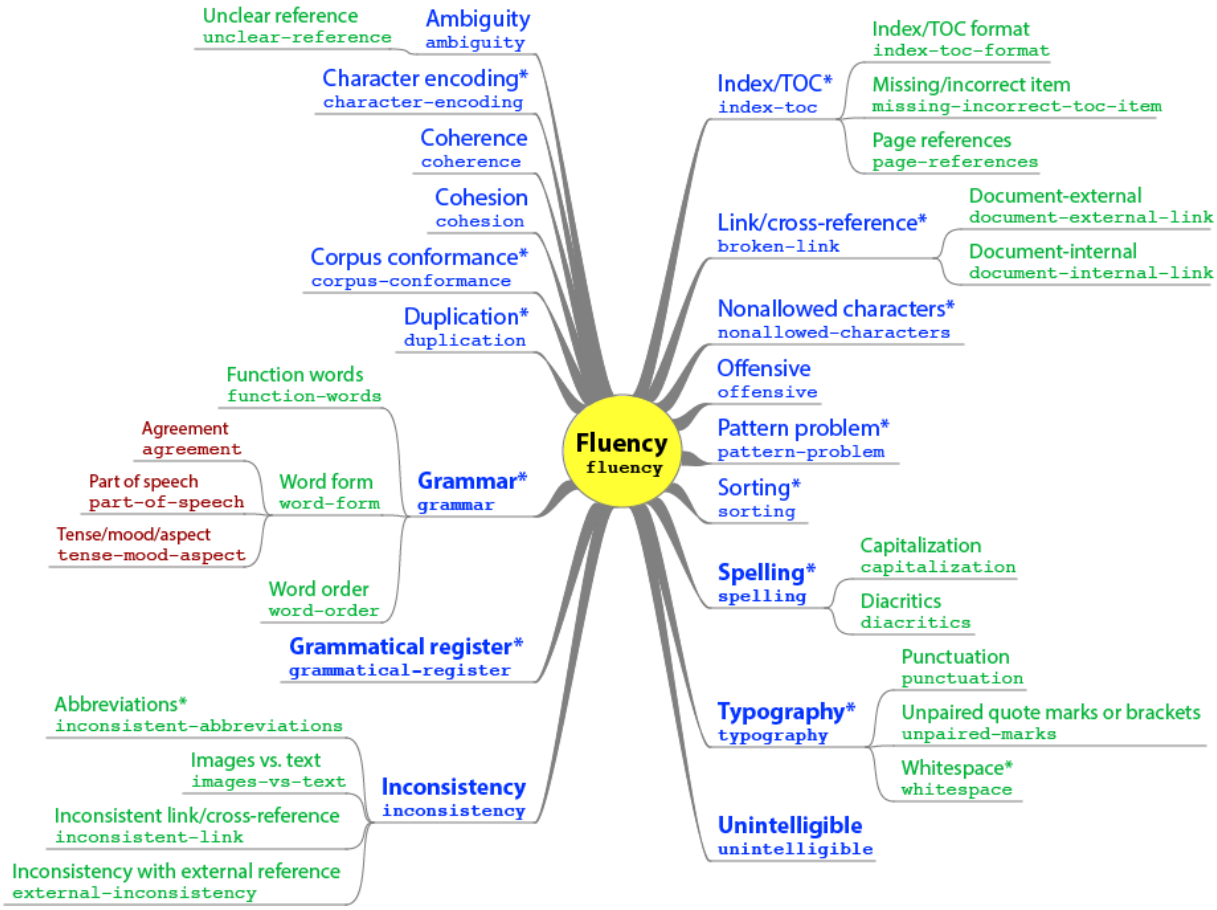

Figure 15 : typologie MQM, catégorie Fluency[13]

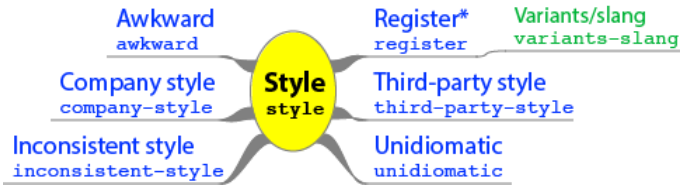

Figure 16 : typologie MQM, catégorie Style[14]

[13] *Idem.*

[14] *Idem.*

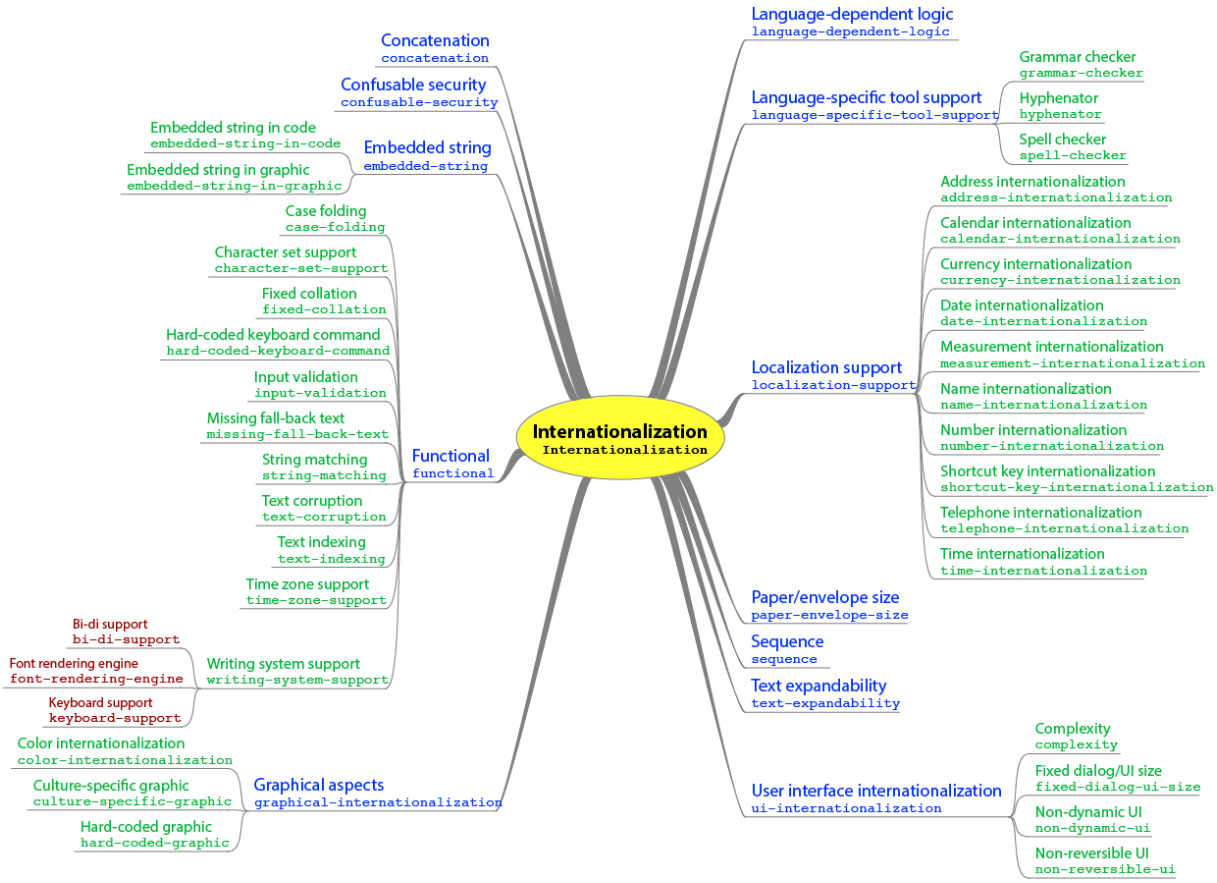

Figure 17 : typologie MQM, catégorie Internationalization[15]

Grâce à ces nombreux exemples de typologies — qui ont l'air distinctes —, on remarque en réalité qu'elles utilisent les mêmes catégories, voire sous-catégories, organisées ou nommées de manière différente. Par exemple, l'erreur « omission » est presque toujours répertoriée dans les typologies, tantôt sous l'appellation « omission(s) », tantôt sous l'appellation « missing words » ou « mots manquants ». Les erreurs terminologiques se retrouvent elles aussi dans chaque typologie, sous différentes appellations, mais elles sont organisées différemment. Parfois, elles sont retrouvées dans une catégorie « terminologie » (voir, par exemple, la typologie Unbabel, Figure 7), mais aussi dans une catégorie « langue » (voir, par exemple, la catégorie MeLLANGE, Figure 9). Certaines typologies sont plus granulaires que d'autres. MQM, par exemple, comporte au total plus de 100 types d'erreurs, alors que la typologie de Federico et al. (2014) ne regroupe que 7 types d'erreurs, organisés en grandes catégories non granulaires.

[15] *Idem.*

La question de la granularité est pertinente ; il est important de conserver un équilibre entre une typologie *trop* granulaire et une typologie *trop peu* granulaire. Une typologie comme MQM, qui comporte plus de 100 types d'erreurs, peut être plus compliquée à manipuler qu'une typologie très condensée et moins précise. Toutefois, une typologie moins granulaire permet plus difficilement d'identifier précisément les types d'erreurs commises et de faire des statistiques sur ces erreurs.

2.4.3. Évaluer une post-édition : comment procéder ?

Il n'existe pas qu'une seule bonne manière d'évaluer une post-édition. Néanmoins, la méthode la plus couramment utilisée consiste à utiliser un système d'annotation.

2.4.3.1. Des annotations basées sur une typologie d'erreurs

Très souvent, dans le cadre d'annotations, les chercheurs, les évaluateurs ou les enseignants s'appuient sur une typologie d'erreurs, comme celles développées au point précédent.

Par exemple, Comparin & Mendes (2017) ont mené une expérience d'annotation humaine d'un corpus de post-éditions de l'anglais vers l'italien. Pour ce faire, elles se sont appuyées sur la typologie d'erreurs utilisée chez Unbabel (cf. Figure 7). Elles ont procédé en plusieurs étapes. Tout d'abord, elles ont annoté la traduction automatique. Ensuite, elles ont annoté la post-édition. Cette démarche leur a permis de calculer le nombre d'erreurs corrigées dans la post-édition, mais aussi les erreurs que les traducteurs n'ont pas remarquées ou pas corrigées. L'interface qu'elles ont utilisée pour l'annotation permettait d'afficher les trois sources de textes en même temps : le TS, la TA et la PE. Les catégories d'erreurs, préenregistrées sur l'interface, s'affichaient automatiquement lorsqu'un mot ou une suite de mots était sélectionné. Par ailleurs, l'annotateur pouvait également donner une note de 0 à 5 pour évaluer la fluidité du texte. De cette manière, elles ont tiré les conclusions suivantes. Entre la traduction automatique et la post-édition, le nombre d'erreurs diminue de 85 %, les erreurs de fluidité étant corrigées à environ 90 % et celles de précision/exactitude à environ 75 %. Par ailleurs, les erreurs les plus observées dans la traduction automatique sont les erreurs de déterminants, les erreurs de choix lexicaux, les erreurs d'accord, les erreurs de temps/modes/aspects et les erreurs d'ordre des mots (*ibid.*, pp. 2-4).

	MT		First edition	
Accuracy errors	abs. freq.	rel. freq.	abs. freq.	rel. freq.
Mistranslation				
Overly literal	9	0.01	4	0.02
False friend	0	0	0	0
Should not have been translated	18	0.02	3	0.02
Lexical selection	165	0.15	37	0.22
Omission	6	0.01	0	0
Untranslated	27	0.02	9	0.05
Addition	11	0.01	2	0.01
Total	236	0.21	55	0.32

Figure 18 : tableau indiquant les réductions d'erreurs dans la catégorie Accuracy (Comparin & Mendes, 2017)

	MT		First edition	
Fluency errors	abs. freq.	rel. freq.	abs. freq.	rel. freq.
Inconsistency				
Word selection	1	0	1	0.01
Tense selection	0	0	0	0
Coherence	2	0	1	0.01
Duplication	0	0	0	0
Spelling				
Orthography	1	0	1	0.01
Capitalization	52	0.05	19	0.11
Diacritics	0	0	0	0
Typography				
Punctuation	9	0.01	4	0.02
Unpaired quote marks and brackets	1	0	0	0
Whitespace	17	0.02	5	0.03
Inconsistency in character use	0	0	0	0
Grammar				
Function words				
Prepositions	70	0.06	10	0.06
Conjunctions	12	0.01	1	0.01
Determiners	237	0.21	19	0.11
Word form				
Part-of-speech	30	0.03	1	0.01
Agreement	159	0.14	13	0.08
Tense/mood/aspect	101	0.09	3	0.02
Word order	106	0.10	4	0.02
Sentence structure	50	0.05	1	0.01
Total	848	0.77	83	0.49

Figure 19 : tableau indiquant les réductions d'erreurs dans la catégorie Fluency (Comparin & Mendes 2017)

Comme le montrent ces tableaux, l'annotation permet d'extraire des statistiques sur les types d'erreurs commis par les traducteurs ou post-éditeurs, sur leur fréquence, la diminution de ces erreurs entre la TA et la PE, et plus encore.

2.4.3.2. Des annotations basées sur une typologie de modifications

Lefer et al. (2022) ont proposé la taxonomie MTPEAS (*Machine Translation Post-Editing Annotation System*). Il s'agit d'un cadre élaboré en vue d'accompagner l'apprentissage, l'enseignement et l'évaluation de la post-édition dans la formation en traduction. Cette taxonomie prévoit qu'avant de confier une tâche de post-édition aux étudiants, l'enseignant identifie les segments erronés dans la TA à corriger dans la PE. Cette annotation préalable n'est pas fournie aux étudiants. Cette annotation n'influence que peu le processus final d'évaluation :

> La correction d'une PE implique de se concentrer sur les segments étiquetés de la TA pour examiner leur traitement par l'étudiant·e dans la PE. Toutefois, il se peut que des segments non étiquetés soient modifiés par l'étudiant·e, soit parce qu'une amélioration semble opportune à ses yeux, soit parce qu'une erreur de la TA est passée inaperçue lors de l'annotation préalable par l'enseignant·e. (*loc. cit.*, p. 5)

Le processus d'évaluation est déterminé par un arbre décisionnel.

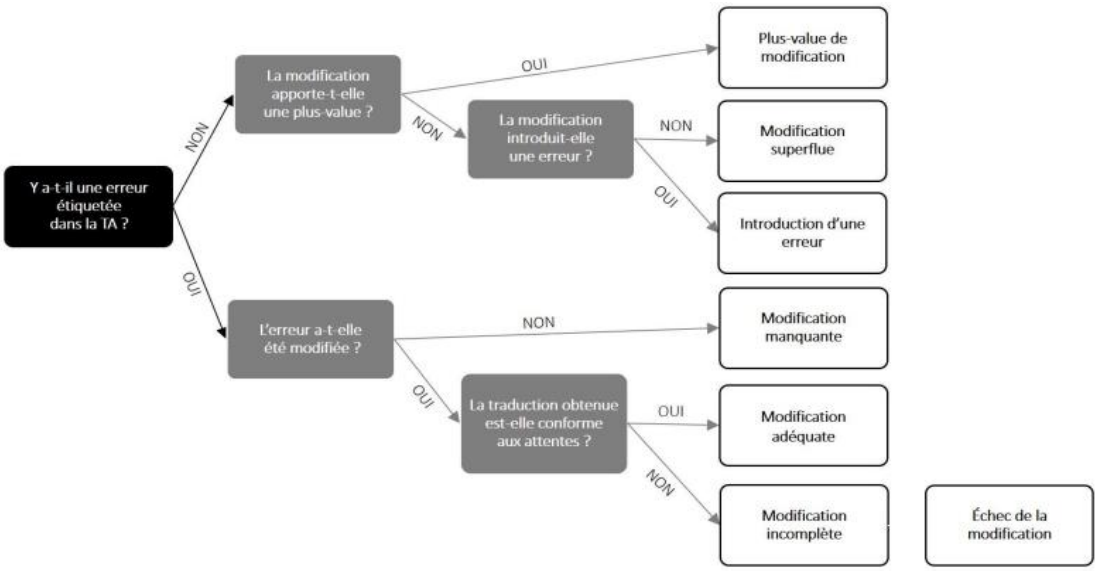

Figure 20 : arbre décisionnel élaboré par Lefer et al. (2022, p. 6)

Les cases blanches, produits de la modification, ont chacune un effet sur la qualité de la post-édition. Ont un effet positif sur la post-édition la plus-value de modification ainsi que la modification adéquate. La modification superflue, quant à elle, n'apporte et n'enlève rien à la PE. La modification incomplète n'améliore que partiellement la PE et a dès lors un effet légèrement négatif. L'échec de la modification ainsi que l'introduction d'une erreur ont, par définition, un effet négatif sur la post-édition. Enfin, la modification manquante a elle aussi un effet négatif sur la qualité de la post-édition (*ibid.*, p. 7).

Cette typologie ne s'appuie donc pas sur des catégories linguistiques d'erreurs, mais plutôt sur des catégories de types de modifications. Toutefois, ces deux types d'annotations sont susceptibles de faire intervenir la subjectivité de la personne qui évalue la PE.

Au vu de cet état de l'art, qui expose les nombreuses divergences et manières de procéder en matière d'évaluation de la qualité, ici de la post-édition, il convient d'exposer la méthodologie adoptée dans ce travail.

3. Méthodologie

Ce travail visant à évaluer de façon humaine des post-éditions produites par deux publics différents de résumés d'articles scientifiques portant sur le traitement automatique des langues (désormais TAL), une méthodologie rigoureuse a été mise en place à cette fin.

3.1. Corpus source

Le corpus de textes sources est composé de titres et des résumés de publications scientifiques portant sur le TAL. Ces publications sont en grande partie des articles de conférence, mais aussi de revues scientifiques, de livres, de prépublications, de rapports ou de chapitres de livres. Toutes ces données sont issues de l'archive ouverte HAL et ont subi un filtrage qualitatif. Tout d'abord, seulement les publications portant sur le domaine général de l'informatique ont été retenues, puis un deuxième filtrage a eu lieu pour ne garder que les publications portant sur le TAL (vérification des mots-clés). Enfin, les publications dotées d'un résumé traduit en français ont été éliminées, de même que les publications qui ne sont pas en open-source.

3.2. Traduction automatique

Ces titres et résumés ont ensuite été traduits par le biais de trois systèmes de traduction automatique commerciaux, à savoir DeepL Pro (version 7.5), Systran Translation Pro et eTranslation (version 12.3), qui sont tous des systèmes de traduction automatique neuronale.

3.3. Post-édition

Les post-éditions de ces résumés d'articles scientifiques ont été menées par deux publics différents : d'un côté, la « communauté » (c'est-à-dire des membres de la communauté du TAL qui ont été encouragés à post-éditer leur(s) propre(s) publication(s)) ; de l'autre, les « traducteurs » (c'est-à-dire des traducteurs, des linguistes, des étudiants et des enseignants du master en traduction à l'Université Paris Cité). Chaque résumé pouvait être post-édité au maximum trois fois, soit une post-édition pour chaque système de TAN, mais celui-ci n'était pas communiqué aux post-éditeurs.

Les deux publics ont effectué ces post-éditions sur une interface spécifique. La consigne donnée à tous les post-éditeurs était la suivante :

> Modifiez le texte (titre et résumé) pour qu'il soit clair, compréhensible et acceptable, comme vous le feriez pour une publication dans un journal en français (p. ex. la revue TAL). Pour ce faire, merci de ne pas vous servir d'outils de traduction automatique. Dans la mesure du possible, merci de faire cette révision sans vous interrompre pour que la durée enregistrée corresponde au temps effectif de post-édition.

Comme indiqué dans la consigne, les post-éditeurs étaient chronométrés, et ces données peuvent être utilisées également pour obtenir des statistiques sur la qualité de la TA et sur les efforts nécessaires. En revanche, cette donnée peut être faussée si les post-éditeurs s'interrompent dans leur tâche, bien que cela soit expressément déconseillé. Ces consignes ne permettent en revanche pas de dire s'il s'agit d'une FPE ou d'une LPE. Les consignes précisent cependant que la qualité doit être celle d'une revue scientifique en français, ce qui exige une traduction de bonne qualité, plutôt qu'une qualité uniquement « acceptable ».

Par ailleurs, il a également été demandé aux post-éditeurs d'évaluer de manière simple le résultat de la TA. Pour ce faire, il leur a été posé différentes questions. Pour les membres de la communauté, la question était : « Quelle importance donneriez-vous aux problèmes de traduction constatés ? » ; ils avaient la possibilité de choisir entre quatre réponses : aucun problème, peu grave (orthographe, ponctuation, etc.), moyennement grave (ne gênent pas la compréhension, mais linguistiquement ou stylistiquement inacceptables) et grave (gênent la compréhension, manquent de fidélité au contenu d'origine). Les traducteurs devaient quant à eux répondre à des questions plus élaborées, c'est-à-dire qu'ils devaient évaluer la gravité des erreurs des catégories suivantes : fidélité, grammaire, terminologie, orthographe et ponctuation, style et cohérence textuelle. Pour ces évaluations, ils bénéficiaient des quatre mêmes niveaux de gravité. Enfin, tant les membres de la communauté que les traducteurs avaient la possibilité de laisser un commentaire libre.

3.4. Annotation du corpus parallèle

Après récupération des données de post-éditions, nous obtenons un corpus parallèle composé du texte source anglais, de la TA en français et de la PE. Chaque entrée est accompagnée de métadonnées, les plus importantes aux fins de cette étude étant le code (3, 4 ou 6) associé au système de TAN utilisé pour la TA et l'identifiant de l'utilisateur (permettant d'identifier s'il s'agit d'un membre de la communauté ou d'un traducteur). Dans notre corpus, nous avons trois

types de post-éditions différents : les post-éditions effectuées par les membres de la communauté du TAL, les post-éditions effectuées par les « traducteurs », et enfin les textes pour lesquelles nous avons à la fois une post-édition de la communauté et une post-édition des traducteurs. Aux fins de cette étude, ce sont les post-éditions qui ont été effectuées par les deux publics que nous avons annotées en priorité.

3.4.1. Programme utilisé

Pour l'annotation de ces corpus, nous avons utilisé *brat rapid annotation tool*[16], qui permet aux annotateurs d'ajouter des étiquettes prédéfinies sur des segments du corpus (voir Figure 21).

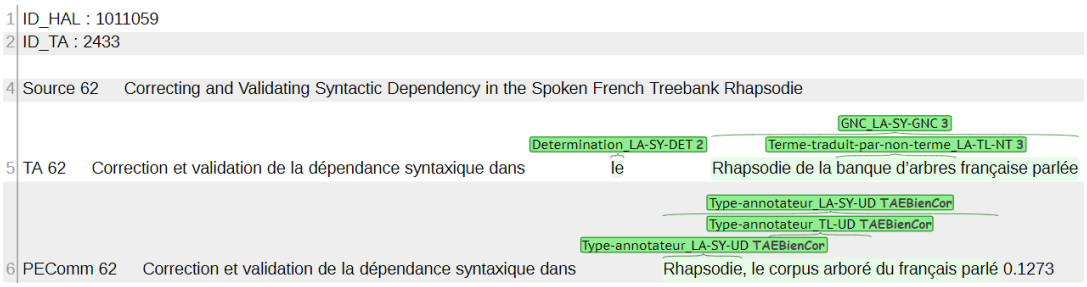

Figure 21 : aperçu du programme brat

Comme on peut le voir ci-dessus, le corpus est aligné au niveau des phrases : la ligne 4 correspond au texte source, la ligne 5 correspond à la TA, la ligne 6 correspond à la post-édition effectuée par la « communauté » et la ligne 7 correspond à la post-édition effectuée par les « traducteurs ».

3.4.2. Schéma d'annotation

Aux fins de cette évaluation, il a fallu élaborer une typologie d'erreurs. D'une part, l'équipe du CLILLAC-ARP avait l'habitude de travailler avec la typologie MeLLANGE (voir Figure 9), qui est destinée à l'évaluation de traductions dans un cadre d'enseignement (voir, par exemple Kübler et al. 2016, 2017, 2021). D'autre part, MQM est une typologie très largement utilisée pour l'évaluation de la qualité de traductions, à la fois humaines et automatiques, voire hybrides (PE). Dès lors, ayant constaté des manques dans la typologie MeLLANGE et des pistes d'améliorations dans MQM, nous avons décidé d'enrichir la typologie MeLLANGE (déjà modifiée par Kübler et al. 2022) avec certains types d'erreurs retrouvés dans MQM.

[16] https://brat.nlplab.org/index.html.

3.4.2.1. *Présentation de la typologie d'erreurs*

- [entities]
- Transfert-contenu
  - Omission_TR-OM
  - Rajout_TR-AD
  - Distorsion_TR-DI
  - Indecision_TR-IN
  - Type-annotateur_TR-UD
  - Intrusion-langue-source
    - Non-traduit-traduisible_TR-SI-UT
    - Trop-litterale_TR-SI-TL
    - Unites-mesure-dates-nombres_TR-SI-UN
    - Type-annotateur_TR-SI-UD
  - Intrusion-langue-cible
    - Traduction-unites-intraduisibles_TR-TI-TD
    - Trop-libre_TR-TI-TF
    - Type-annotateur_TR-TI-UD
- Langue
  - Syntaxe_LA-SY
    - Determination_LA-SY-DET
    - Mauvaise-preposition_LA-SY-PR
    - GNC_LA-SY-GNC
    - Type-annotateur_LA-SY-UD
  - Flexion-accord
    - Temps-aspect_LA-IA-TA
    - Genre_LA-IA-GE
    - Nombre_LA-IA-NU
  - Typographie
    - Orthographie_LA-HY-SP
    - Accent-diacritiques_LA-HY-AC
    - Mauvaise-casse_LA-HY-CA
    - Ponctuation_LA-HY-PU
    - Type-annotateur_LA-HY-UD
  - Registre
    - Incompatible-texte-source_LA-RE-IS
    - Inadapte-au-type-texte-cible_LA-RE-IT
    - Type-annotateur_LA-RE-UD
  - Style
    - Formulation-maladroite_LA-ST-AW
    - Tautologie_LA-ST-TA
    - Style-titre_LA-ST-TS
    - Type-annotateur_LA-ST-UD
  - Reference-pas-claire_LA-UR
  - Conventions-textuelles
    - Coherence_LA-TC-CE
    - Cohesion_LA-TC-CN
  - Terminologie-lexique
    - Choix-incorrect-Termino_LA-TL-INS
    - Choix-incorrect-Langue-Generale_LA-TL-ING
    - Mauvais-acronyme-abreviation_LA-TL-MAA
    - Faux-amis_LA-TL-FC
    - Terme-traduit-par-non-terme_LA-TL-NT
    - Collocation-incorrecte-Specialise_LA-TL-ICS
    - Collocation-incorrecte-Langue-Generale_LA-TL-ICG
    - Choix-incompatible-avec-texte-cible_LA-TL-IT
    - Incoherence-terminologique
      - Differents-termes-traduction_LA-TL-TI-DT
      - Differentes-abbreviations-traduction_LA-TL-TI-DA
    - Type-annotateur_TL-UD
- Outils
  - Hallucination_OU-TAH
  - Conformite-corpus_OU-CC
  - Duplication_OU-DU
  - Choix-incompatible-glossaire_OU-GC

Figure 22 : typologie utilisée

Les catégories d'erreurs présentes dans la typologie MQM que nous avons décidé d'intégrer à la typologie MeLLANGE sont les suivantes :

Mots-outils	Détermination
	Mauvaise préposition
Style du titre	
Référence pas claire	
Cohésion	
Cohérence	
Incohérence terminologique	Différents termes dans la traduction pour le même terme dans le texte source
	Différentes abréviations dans la traduction
Mauvais acronyme/abréviation	
Utilisation des outils	Hallucination de la TA
	Conformité au corpus
	Duplication

Les autres types d'erreurs que l'on retrouve dans notre typologie étaient déjà présents dans la typologie MeLLANGE (Figure 9).

La typologie d'erreurs utilisée dans le cadre de ce projet est divisée en trois grandes catégories :

- Erreurs de transfert de contenu,
- Erreurs de langue,
- Erreurs liées aux outils.

La catégorie « transfert de contenu » regroupe les erreurs dites de traduction, à savoir les différentes erreurs qui altèrent le sens et le contenu du texte source ou qui impactent le transfert et la compréhension du message. Dans cette grande catégorie d'erreurs, on retrouve les omissions, les ajouts, les distorsions du contenu, les indécisions, les intrusions du texte source, ainsi que les intrusions dans la langue cible.

Ensuite, la catégorie « langue » comprennent les erreurs linguistiques. Ici, on peut retrouver les erreurs de syntaxe, de flexion et d'accord, de typographie, de registre, de style, de référence, de conventions textuelles, ainsi qu'une large gamme d'erreurs terminologiques (langue de spécialité) et lexicales (langue générale).

Enfin, la catégorie « outils », qui a été créée aux fins de ce projet, regroupe les erreurs liées aux outils ou à la maîtrise de ces derniers. On y retrouve dès lors les hallucinations de la

TA, le non-respect du corpus ou du glossaire fourni – le cas échéant –, ainsi que les erreurs de duplication.

Tous les types et sous-types d'erreurs sont définis en détail et illustrés à l'aide d'exemples du corpus dans le manuel d'annotation en Annexe 1.

3.4.2.2. Attributs

Nous avons intégré à cette typologie les attributs ajoutés par Kübler et al. (2021) (cf. Figure 23).

```
[attributes]
TA_Correct      Arg:<ENTITY>, Value:TACorBienCorr|TACorMalCorr
TA_Erronee      Arg:<ENTITY>, Value:TAEBienCor|TAEMalCor|TAENonCor
Score_Grav      Arg:<ENTITY>, Value:0|1|2|3
```

Figure 23 : attributs de l'annotation

Ces attributs sont à ajouter lorsqu'on annote la post-édition, et non la traduction automatique. Quand la TA est correcte, deux attributs sont possibles : TA correcte bien corrigée (lorsque le post-éditeur introduit une variante ou une modification acceptable, mais pas strictement nécessaire) et TA correcte mal corrigée (lorsque le post-éditeur introduit une nouvelle erreur). Quand la TA est incorrecte, trois attributs sont possibles : TA erronée bien corrigée (le post-éditeur corrige l'erreur), TA erronée mal corrigée (le post-éditeur corrige l'erreur, mais pas correctement) et TA erronée non corrigée (le post-éditeur ne modifie pas l'erreur). Par conséquent, le schéma d'annotation utilisé dans le cadre de ce projet comprend à la fois une typologie de types d'erreurs et une typologie de types de modifications.

3.4.2.3. Scores de gravité

Dans la pratique d'évaluation de traductions, l'application de scores de gravité est largement utilisée. MQM présente d'ailleurs un exemple de *scorecard*[17], où l'on observe quatre niveaux de gravité : neutre (score 0), mineur (score 1), majeur (score 5) et critique (score 25). Nous avons par conséquent opté pour l'application d'un score de gravité, mais nous l'avons personnalisé pour qu'il corresponde à nos besoins :

- score de gravité 0 (neutre) : une meilleure traduction pourrait être proposée, mais la traduction proposée n'est pas réellement une erreur ;

[17] https://themqm.org/error-types-2/1_scorecards/.

- score de gravité 1 (mineur) : l'erreur a un (très) léger impact sur le texte cible, mais elle ne nuit pas à la lisibilité ou à la compréhension du contenu
- score de gravité 2 (majeur) : l'erreur a un gros impact sur le texte cible, c'est-à-dire qu'elle affecte la compréhension, la lisibilité ou la pertinence de celui-ci (par exemple, perte de sens ou de glissement de sens) ;
- score de gravité 3 (critique) : soit l'erreur rend le contenu totalement faux, soit l'erreur rend le contenu inexploitable, c'est-à-dire qu'une reformulation totale est nécessaire.

3.5. Manuel d'annotation

Un manuel d'annotation (voir Annexe 1) a également été rédigé dans le cadre de ce projet. Celui-ci vise à servir de guide pour l'annotation d'erreurs dans le cadre de traductions humaines, automatiques ou de post-éditions. Dans ce guide, est exposée la typologie d'erreurs, les principes fondamentaux à respecter lors de l'annotation et les différents attributs qui peuvent être ajoutés aux annotations d'erreurs. Enfin, le schéma d'annotation est illustré à l'aide de nombreux exemples provenant autant que possible du corpus de traductions automatiques et de post-éditions utilisé dans le cadre de ce travail.

La finalité de ce manuel d'annotation est de garantir un meilleur accord entre les différents annotateurs et de l'utiliser pour de futures tâches d'annotation.

3.6. Exploitation des données

Une fois les annotations réalisées, ces données ont été exportées et analysées. Pour chaque texte (source), une feuille Excel a été créée avec à la fois les données de la TA, de la post-édition de la « communauté » et de la post-édition des « traducteurs » pour la réalisation de *scorecards*. Une *scorecard* (cf. Figure 24) sert à évaluer la qualité globale d'une traduction automatique ou d'une post-édition.

Un score de qualité est calculé sur un total de 100 points. Pour calculer ce score, il a d'abord fallu calculer un « total de gravité » (cf. colonne rouge, Figure 24). Ce total est d'abord calculé par type d'erreur : les erreurs de niveau 0 ont un coefficient de 0, les erreurs de niveau 1 ont un coefficient de 1, les erreurs de niveau 2 ont un coefficient de 5, et les erreurs de niveau 3

ont un coefficient de 25[18]. Par exemple, dans la Figure 24, pour la catégorie « Omission », il n'y a qu'une seule erreur, qui est de niveau 2. Par conséquent, pour ce type d'erreur, le « total de gravité » est de 1 x 5, soit 5. La somme des totaux de gravité de chaque catégorie est calculée pour donner un score de gravité total (dans la Figure 24, le score de gravité total est de 23). Ce score de gravité est ensuite recalculé pour 100 mots en prenant comme référence le nombre de mots du texte source[19], ce qui donne la « pénalité totale par 100 mots » :

$$\frac{total\ de\ gravité}{nombre\ de\ mots\ du\ texte\ source}\ x\ 100$$

Enfin, pour calculer le score de gravité, il faut faire le calcul suivant :

$$100 - pénalité\ par\ 100\ mots - \%\ de\ l'attribut\ TACorrMalCor$$

Il semblait également important de prendre en compte, dans l'évaluation des *post-éditions*, le pourcentage de l'attribut de TA correcte mal corrigée, étant donné que celui-ci rend compte des erreurs non présentes dans la TA, mais introduites par le post-éditeur dans la PE.

En plus des statistiques sur les attributs, la *scorecard* comprend aussi des statistiques relatives à la distribution des niveaux de gravité (0, 1, 2 et 3).

[18] Ces coefficients semblent plus pertinents, car si on utilisait les coefficients correspondant aux scores (0, 1, 2 et 3), la plupart des traductions (notamment automatiques) obtenaient un score qui ne reflétait pas leur qualité. Dès lors, l'utilisation des coefficients 0, 1, 5 et 25 semblait plus judicieuse. C'est d'ailleurs ce que propose MQM.

[19] Le nombre de mots du texte source semble être la référence la plus juste, puisqu'elle permet d'évaluer de manière équitable la traduction automatique et ses deux post-éditions. Par ailleurs, ici aussi, si on utilisait comme référence le nombre de mots de la TA ou des post-éditions, certains textes obtenaient un score de qualité trop élevé au vu des erreurs annotées.

PEComm	Total erreurs	0	1	2	3	Total gravité
Transfert-contenu	0	0	0	0	0	0
Omission_TR-OM	1	0	0	1	0	5
Rajout_TR-AD	0	0	0	0	0	0
Distorsion_TR-DI	0	0	0	0	0	0
Indecision_TR-IN	0	0	0	0	0	0
Intrusion-langue-source	0	0	0	0	0	0
Non-traduit-traduisible_TR-SI-UT	0	0	0	0	0	0
Trop-litterale_TR-SI-TL	1	0	1	0	0	1
Unites-mesure-dates-nombres_TR-SI-UN	0	0	0	0	0	0
Intrusion-langue-cible	0	0	0	0	0	0
Traduction-unites-intraduisibles_TR-TI-TD	0	0	0	0	0	0
Trop-libre_TR-TI-TF	0	0	0	0	0	0
Langue	1	0	1	0	0	1
Syntaxe_LA-SY	0	0	0	0	0	0
Determination_LA-SY-DET	1	0	1	0	0	1
Mauvaise-preposition_LA-SY-PR	0	0	0	0	0	0
GNC_LA-SY-GNC	0	0	0	0	0	0
Flexion-accord	0	0	0	0	0	0
Temps-aspect_LA-IA-TA	0	0	0	0	0	0
Genre_LA-IA-GE	0	0	0	0	0	0
Nombre_LA-IA-NU	0	0	0	0	0	0
Typographie	0	0	0	0	0	0
Orthographie_LA-HY-SP	0	0	0	0	0	0
Accent-diacritiques_LA-HY-AC	0	0	0	0	0	0
Mauvaise-casse_LA-HY-CA	0	0	0	0	0	0
Ponctuation_LA-HY-PU	2	0	2	0	0	2
Registre	0	0	0	0	0	0
Incompatible-texte-source_LA-RE-IS	0	0	0	0	0	0
Inadapte-au-type-texte-cible_LA-RE-IT	0	0	0	0	0	0
Style	0	0	0	0	0	0
Formulation-maladroite_LA-ST-AW	2	0	0	2	0	10
Tautologie_LA-ST-TA	0	0	0	0	0	0
Style-titre_LA-ST-TS	0	0	0	0	0	0
Reference-pas-claire_LA-UR	0	0	0	0	0	0
Conventions-textuelles	0	0	0	0	0	0
Coherence_LA-TC-CE	0	0	0	0	0	0
Cohesion_LA-TC-CN	0	0	0	0	0	0
Terminologie-lexique	0	0	0	0	0	0
Choix-incorrect-Termino_LA-TL-INS	1	0	1	0	0	1
Choix-incorrect-Langue-Generale_LA-TL-ING	1	0	1	0	0	1
Mauvais-acronyme-abreviation_LA-TL-MAA	0	0	0	0	0	0
Faux-amis_LA-TL-FC	1	0	1	0	0	1
Terme-traduit-par-non-terme_LA-TL-NT	0	0	0	0	0	0
Collocation-incorrecte-Specialise_LA-TL-ICS	0	0	0	0	0	0
Collocation-incorrecte-Langue-Generale_LA-TL-ICG	0	0	0	0	0	0
Choix-incompatible-avec-texte-cible_LA-TL-IT	0	0	0	0	0	0
Incoherence-terminologique	0	0	0	0	0	0
Differents-termes-traduction_LA-TL-TI-DT	0	0	0	0	0	0
Differentes-abbreviations-traduction_LA-TL-TI-DA	0	0	0	0	0	0
Outils	0	0	0	0	0	0
Hallucination_OU-TAH	0	0	0	0	0	0
Conformite-corpus_OU-CC	0	0	0	0	0	0
Duplication_OU-DU	0	0	0	0	0	0
Choix-incompatible-glossaire_OU-GC	0	0	0	0	0	0
	11					23
Erreurs score 0	0	0%				
Erreurs score 1	8	73%				
Erreurs score 2	3	27%				
Erreurs score 3	0	0%				
TAEBienCorr	16	59%				
TACorrBienCorr	1	4%				
TAEMalCorr	1	4%				
TAENonCorr	7	26%				
TACorrMalCorr	2	7%				
Nombre de mots	188					
Pénalité par 100 mots	12,23					
Score de qualité	**80,77**					

Figure 24 : exemple de scorecard

4. Statistiques et analyse des résultats

Les données sur les traductions automatiques et leurs post-éditions sont analysées de deux façons. Dans un premier temps, dans le cadre du projet MATOS[20], l'équipe de l'Inria[21] a utilisé des métriques automatiques pour évaluer les différentes productions. Dans un second temps, grâce aux *scorecards* expliquées au point précédent, différentes statistiques de l'évaluation humaine seront présentées. Comparer les métriques d'évaluation automatique aux statistiques de l'évaluation humaine permet d'observer si ces deux méthodes d'évaluation quelque peu déconnectées aboutissent aux mêmes résultats, mais aussi de voir quels sont les apports et les avantages de chacune de ces méthodes.

4.1. Résultats des métriques automatiques

L'équipe de l'Inria a analysé automatiquement l'ensemble des traductions automatiques et des post-éditions, c'est-à-dire qu'elle n'a pas uniquement observé les textes qui ont deux post-éditions — comme cela a été fait dans le cadre de ce projet —, mais aussi les textes n'ayant qu'une seule post-édition. Dès lors, il est possible que cette différence engendre des résultats divergents entre les métriques automatiques et l'évaluation humaine. Toutefois, il est également possible que l'échantillon analysé de façon humaine soit représentatif de l'ensemble des textes, et que les résultats coïncident.

Avec ces métriques automatiques, l'Inria entend répondre à quatre questions :

1. Quel est l'effort nécessaire aux deux différents groupes pour post-éditer des traductions automatiques dans le domaine du TAL ?
2. Est-il possible de mesurer la qualité de ces post-éditions ?
3. Les différents systèmes de TA (DeepL, Systran et eTranslation) présentent-ils des résultats différents en termes de qualité ?
4. Quels sont les types d'erreurs qui posent des problèmes dans ce processus ?

Pour répondre à la question relative aux efforts nécessaires, ils ont utilisé la métrique HTER (cf. Figure 25), qui calcule la distance entre la post-édition et la traduction automatique

[20] Machine Translation for Open Science

[21] L'Institut national de recherche en sciences et technologies du numérique

en termes de modifications. Cette mesure est calculée pour chaque résumé séparément, puis une moyenne est faite pour chaque groupe de post-éditeurs.

Score HTER	
Moyenne totale	10.7
Moyenne Communauté	18.2
Moyenne Traducteurs	8.0

Figure 25 : Scores HTER

Ce tableau montre que le score HTER total moyen est de 10.7, un score relativement bas qui indiquerait que les traductions automatiques sont généralement de bonne qualité, puisqu'elles nécessitent peu d'efforts de post-édition. Par ailleurs, 13 documents sur 337 ont été entièrement laissés tels quels, sans aucune modification. Toutefois, on observe une différence de plus de 10 points entre le HTER de la communauté et celui des traducteurs. Ce résultat indique que les efforts de post-édition sont plus importants pour les membres de la communauté du TAL. Cette différence pourrait refléter la manière dont chaque groupe perçoit la tâche de post-édition. En effet, comme les membres de la communauté des traducteurs (étudiants et enseignants en traduction, traducteurs, etc.) sont censés être familiers avec la post-édition, ils ont tendance à suivre les consignes établies de la post-édition, c'est-à-dire d'apporter uniquement des modifications nécessaires. Les membres de la communauté du TAL, en revanche, sont plus susceptibles de reformuler des segments entiers, ce qui peut être superflu.

Pour mesurer la qualité des post-éditions et des traductions automatiques sans traductions de référence, l'Inria a utilisé Comet (cf. état de l'art). Voici les résultats obtenus :

	TA	PE	Amélioration
Communauté	77.8	78.6	+ 0,8
Traducteurs	76.3	77.0	+ 0,7

Figure 26 : score de qualité (métrique COMET)

Ce tableau montre qu'il y a très peu d'amélioration entre les traductions automatiques et les post-éditions, et ce, pour les deux groupes. Il montre aussi que les post-éditions des membres de la communauté du TAL sont légèrement meilleures que celles des traducteurs. Toutefois, les TA choisies par la communauté sont également meilleures que celles choisies par les traducteurs.

Pour comparer les différents systèmes de traduction automatique (DeepL, Systran et eTranslation), l'Inria a utilisé deux métriques : d'une part, le HTER ; d'autre part, le score BLEU (cf. état de l'art) pour chaque système de TA et pour chaque groupe (cf. Figure 27).

Groupe	DeepL	Systran	eTranslation
Communauté	81.7 / 13.8	73.1 / 19.7	68.3 / 24.4
Traducteurs	90.4 / 7.1	87.3 / 8.5	85.9 / 10.6

Figure 27 : comparaison des systèmes de TA (BLEU / HTER)

Ce tableau montre que le système de TA jugé comme étant le meilleur est DeepL, suivi de Systran, et enfin de eTranslation. En effet, plus le score BLEU baisse (plus la TA est jugée mauvaise), plus le HTER augmente (plus il y a des efforts de post-édition), ce qui est tout à fait logique.

Enfin, pour l'analyse des types d'erreurs, l'Inria a exploité les feedbacks donnés par les post-éditeurs lors de chaque tâche de post-édition (cf. Méthodologie). Ils ont observé que les erreurs de grammaire, de style et de ponctuation étaient principalement associées à des niveaux de gravité faibles, de même que les erreurs de fidélité au contenu et de cohérence. Toutefois, les erreurs de terminologie sont associées à des niveaux de gravité généralement plus élevés.

Pour conclure, les métriques automatiques montrent que les sorties brutes de la TA sont déjà assez bonnes, ce qui est démontré par des scores HTER relativement faibles. Une autre observation importante est que les experts du domaine ont tendance à reformuler des plus grandes portions de la TA, alors que les traducteurs sont plus susceptibles de rester proche de la sortie brute de la TA et de reformuler le nécessaire. Enfin, l'exploitation des feedbacks des post-éditeurs montre que les erreurs les plus fréquentes et les plus graves sont les erreurs de terminologie.

4.2. Résultats de l'évaluation humaine

Les *scorecards* réalisées à la suite des annotations ont permis de réaliser des statistiques sur différents points d'analyse : la qualité globale des traductions automatiques et des post-éditions, les types d'erreurs, les scores de gravité et les attributs.

4.2.1. Statistiques sur la qualité

Les métriques automatiques détaillées au point précédent montrent qu'il y a peu d'amélioration entre la TA et les post-éditions (0,8 point pour la communauté et 0,7 point pour les traducteurs). Par ailleurs, ces métriques semblent indiquer que DeepL est le système de TA qui génère les meilleures traductions, suivi par Systran, puis par eTranslation. Toutefois, toujours selon ces métriques, il n'y a pas de différence significative entre ces trois systèmes.

Les statistiques de l'évaluation humaine suivent la même tendance, mais avec des écarts plus importants entre certains systèmes de TA.

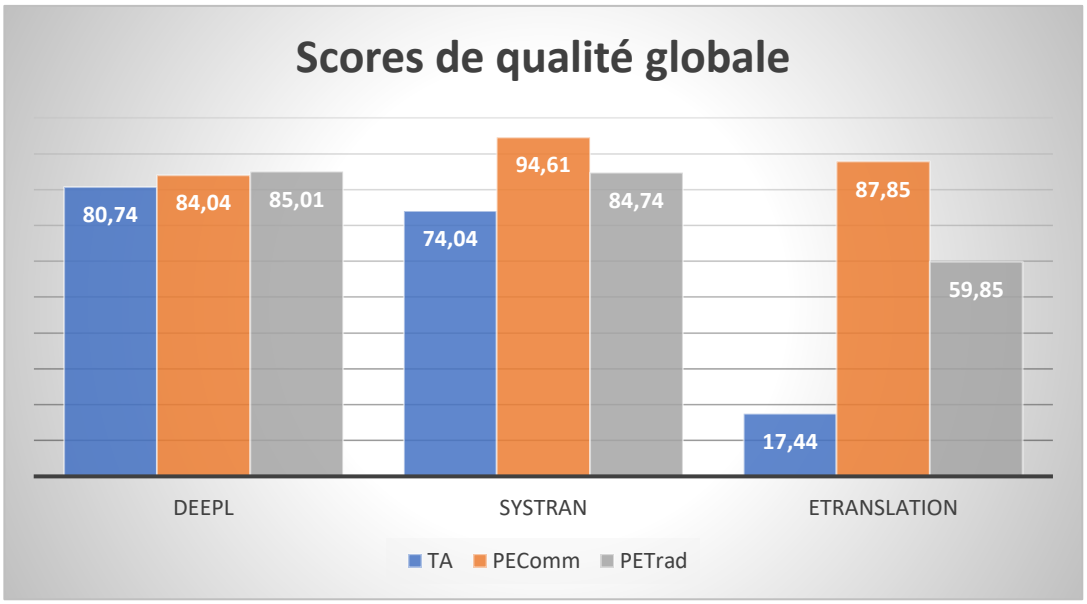

Figure 28 : scores de qualité globale

La Figure 28 montre effectivement que le meilleur système de TA est DeepL (score moyen de 80,74), suivi de Systran (74,04), et enfin de eTranslation (17,44). Toutefois, ce graphique ne corrobore pas complètement les observations formulées avec les métriques automatiques. En effet, on observe entre Systran et eTranslation un écart significatif (de presque 60 points). Ces tendances ne sont pas reflétées dans la qualité des post-éditions. Systran est le système dont les post-éditions sont les meilleures (89,68 en moyenne). Bien que DeepL propose les meilleures TA, ses post-éditions sont légèrement moins qualitatives (84,53 %) que celles de Systran. Les post-éditions de eTranslation, en revanche, ont une qualité qui varie fortement en fonction du post-éditeur. En effet, chez les membres de la communauté du TAL, le système de TA et la qualité de la TA ne semblent pas influencer la qualité de la post-édition. Néanmoins, chez les

traducteurs, une corrélation entre le système de TA et la qualité des post-éditions semble apparaître.

Il est également intéressant d'observer la différence entre la métrique automatique BLEU et l'évaluation humaine des traductions automatiques texte par texte (voir tableau ci-dessous).

Identifiant HAL	BLEU	Évaluation humaine	Différence
1011059	81,95	30,32	51,63
1425724	71,1	80	8,9
1522314	58,51	-153,33	211,84
1588171	83,91	76,52	7,39
1618388	100	85,86	14,14
1674140	86,17	85,6	0,57
1742378	36,38	-72,41	108,79
1847314	79,05	96,51	17,46
1990502	71,5	81,08	9,58
2343408	74	90,37	16,37
3014110	58,78	74,24	15,46
3014123	83,59	69,41	14,18
3029253	81,22	29,06	52,16
3029255	77,25	94,59	17,34
3424174	85,57	95,15	9,58
3686763	89,25	98,37	9,12
3714951	71,35	83,57	12,22
3720096	86,92	85,94	0,98
3812319	65,92	46,46	19,46
3834732	56,96	58,77	1,81
3901369	76,56	49,5	27,06
3933089	56,32	72,11	15,79
3977982	63,58	93,33	29,75
	73,73	58,74	25,83

Ce tableau montre que la différence moyenne entre le score BLEU et le résultat de l'évaluation humaine est de 25,83 points. Par ailleurs, l'évaluation automatique tend à surévaluer la qualité des traductions automatiques (moyenne de 73,73) en comparaison avec l'annotation (score moyen de 58,74). Pour plus de 65 % des textes (surlignés en orange), on observe que la différence entre le score BLEU et l'évaluation humaine est de plus de 10 points. Dès lors, ce tableau permet d'affirmer que l'évaluation humaine et l'évaluation automatique des TA sont totalement déconnectées et n'aboutissent pas à des résultats convergents.

La Figure 29 (ci-dessous) représente la répartition des scores de qualité par système de traduction automatique. Pour DeepL (ligne bleue), on remarque principalement une courbe qui croît, avec une exception pour les scores entre 0 et 50. Pour Systran et eTranslation, les courbes sont plus irrégulières. On observe que eTranslation est le seul système de TA qui génère des traductions obtenant un score négatif. C'est aussi le seul système pour lequel aucune des traductions n'a un score compris entre 90 et 100. Ce graphique montre effectivement que DeepL produit des traductions de meilleure qualité, et qu'eTranslation est loin derrière.

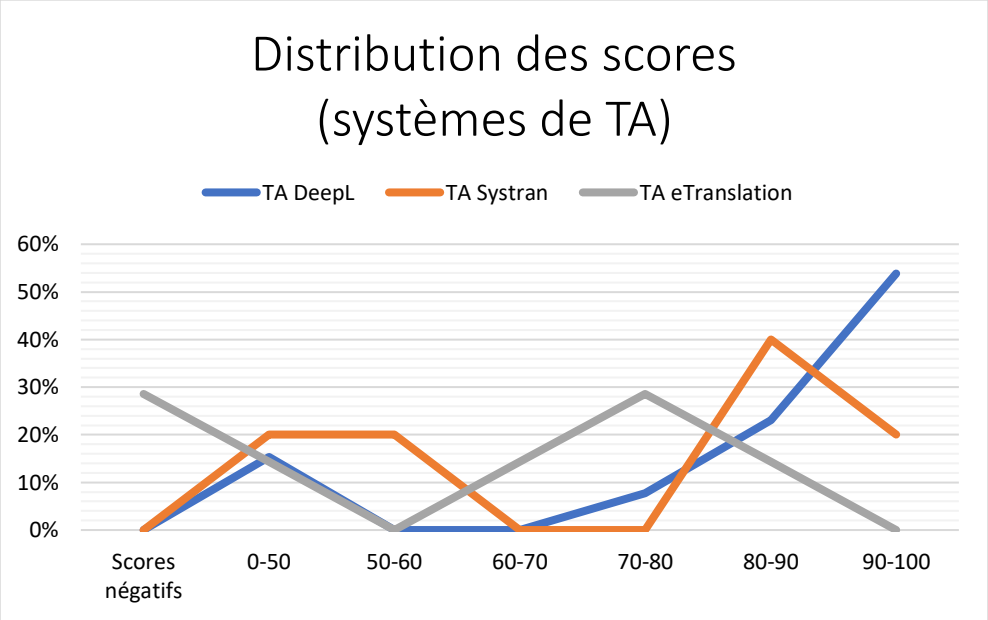

Figure 29 : distribution des scores par système de TA

En moyenne, les membres de la communauté du TAL génèrent des post-éditions jugées meilleures que celles des membres de la communauté des linguistes (cf. Figure 30).

TA	61,67
PEComm	87,22 (+ 25,55)
PETrad	77,91 (+ 16,24)

Figure 30 : scores moyens (TA, PEComm, PETrad)

Les résultats de la Figure 30 confirment en partie ce qui a été observé avec les métriques automatiques. Toutefois, avec les statistiques de l'évaluation humaine, on observe des améliorations significatives entre la TA et les PE, avec une amélioration plus importante pour les membres de la communauté du TAL.

Le graphique de la Figure 31 (ci-dessous) permet de visualiser la distribution des scores de qualité par groupe de post-éditeurs (communauté vs traducteurs). Différents phénomènes peuvent y être observés. Tout d'abord, on remarque que la communauté de spécialistes ne produit aucune post-édition jugée insuffisante (scores négatifs ou entre 0 et 50), à l'inverse des traducteurs. La courbe des post-éditions de la communauté est totalement croissante, ce qui est positif, puisque cela signifie qu'ils produisent plutôt des PE de (très) bonne qualité. La courbe des PE des traducteurs, quant à elle, est plus irrégulière et oscillante, bien qu'elle tende tout de même à croître. Enfin, on remarque que ce sont les membres de la communauté du TAL qui proposent le plus de PE de très bonne qualité (entre 90 et 100).

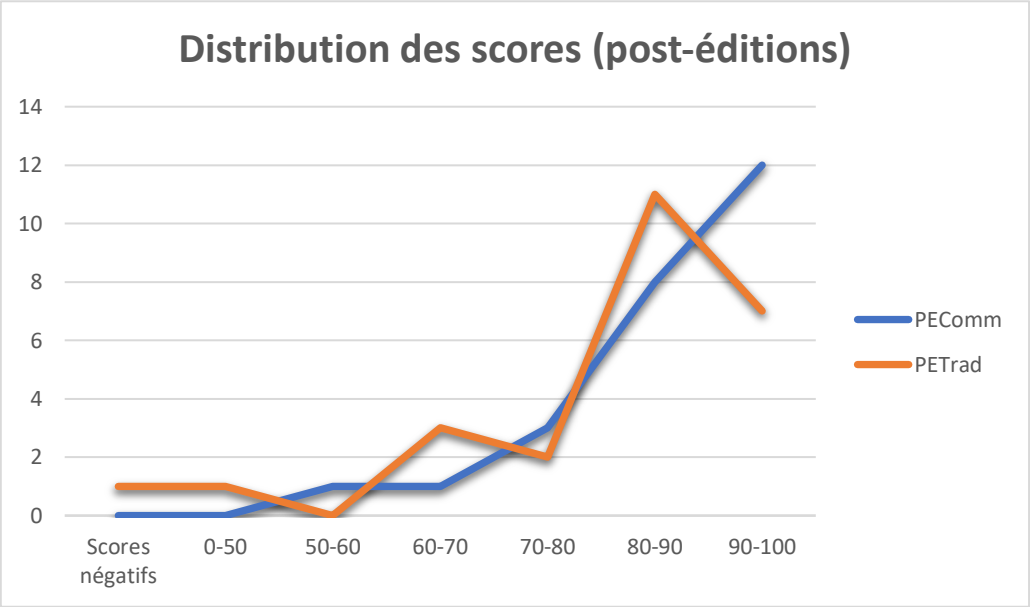

Figure 31 : distribution des scores par groupe de post-éditeurs

4.2.1.1. eTranslation

Il est intéressant d'observer les améliorations entre la TA et les PE au cas par cas, pour chaque système. La Figure 32 ci-dessous permet de visualiser ces relations pour eTranslation.

TA eTranslation	-153,33	85,60	-72,4	74,24	69,41	46,46	72,11
PEComm	92,38	89,65	88,51	90,91	79,24	90,43	83,88
PETrad	-28,71	88,42	24,14	92,42	76,71	84,34	81,63

Figure 32 : eTranslation - TA et PE

Dans ce tableau, chaque colonne représente un texte. Dans chaque colonne, on peut voir le score de la TA (ici eTranslation), ainsi que le score de chaque post-édition. Les cases rouges représentent la moins bonne version, les cases jaunes représentent les versions moyennes, et les cases vertes représentent les meilleures versions.

Pour le système eTranslation, on remarque que toutes les traductions automatiques sont jugées plus mauvaises que les post-éditions. Par ailleurs, on remarque qu'en moyenne, les TA de eTranslation sont beaucoup plus améliorées par la communauté (cf. Figure 28, score moyen de 87,85) que par les traducteurs (59,85). D'ailleurs, ces derniers produisent une PE avec un score négatif et une PE avec un score inférieur à 50. Pour une seule des TA, la PE des traducteurs est meilleure que celle de la communauté.

Cependant, la communauté ne produit presque que des PE avec un score supérieur à 80, malgré la mauvaise qualité des TA de eTranslation, qui ont en moyenne un score de 17,44.

4.2.1.2. Systran

Pour Systran, les résultats sont plus nuancés.

TA Systran	49,5	80,00	98,37	83,57	58,77
PEComm	93,00	84,71	99,19	96,14	100
PETrad	89,00	68,86	99,19	80,75	85,86

Figure 33 : Systran - TA et PE

Il y a, en effet, deux textes pour lesquels la traduction la moins bonne n'est pas celle de Systran, mais celle des traducteurs. La PE de la communauté est toujours la meilleure des trois traductions (à l'exception d'une égalité entre une PEComm et une PETrad). Par rapport à la TA de Systran, la qualité de la PE augmente en moyenne de 10 % environ chez les traducteurs et de 20 % environ chez les membres de la communauté du TAL (cf. Figure 28). Il s'agit du système pour lesquels les post-éditions sont les meilleures, bien que DeepL propose de meilleures traductions automatiques.

4.2.1.3. DeepL

Pour les TA de DeepL, en moyenne, ce sont les traducteurs qui réalisent les meilleures post-éditions (cf. Figure 28).

TADeepL	30,32	76,52	85,86	96,512	97,08	93,75	81,08	90,37	29,06	94,59	95,15	85,94	93,33
PEComm	80,77	68,52	85,86	78,67	80,66	97,22	88,51	91,52	90,60	59,99	100	100	70,15
PETrad	77,57	90,43	85,86	85,84	68,32	94,44	84,86	63,59	94,87	98,20	85,06	81,48	94,55

Figure 34 : DeepL - TA et PE

La particularité de DeepL est que c'est le seul des trois systèmes pour lequel une ou plusieurs des TA sont meilleures que les deux post-éditions ou égales aux deux PE (voir colonnes 3, 4 et 5). Il est également intéressant d'observer que les deux TA pour lesquelles les deux PE sont meilleures sont celles qui ont les meilleurs scores de qualité (96,51 et 97,08). On pourrait dès lors supposer que lorsque les TA sont de très bonne qualité, il y a plus de chances que les post-éditeurs fassent davantage de modifications superflues et introduisent par conséquent des erreurs (attribut « TA correcte mal corrigée »). Pour rappel, le pourcentage d'utilisation de cet attribut est pris en compte dans le calcul du score de qualité. Pour vérifier cette hypothèse, il peut être intéressant d'observer les statistiques relatives aux attributs.

4.2.2. Statistiques relatives aux attributs

Le tableau ci-dessous donne des statistiques intéressantes en termes d'erreurs corrigées, non/mal corrigées, de variantes proposées et d'erreurs introduites dans la PE.

	PEComm			PETrad	
TAEBienCorr	255	52,58 %		194	44,80 %
TACorrBienCorr	99	20,41 %		49	11,32 %
TAEMalCorr	30	6,19 %		33	7,62 %
TAENonCorr	68	14,02 %		131	30,25 %
TACorrMalCorr	33	6,80 %		26	6,00 %
TOTAL d'erreurs	**131**			**190**	

Figure 35 : distribution des attributs

Grâce à ce tableau, on remarque que ce sont les membres de la communauté qui corrigent (correctement) le plus d'erreurs. De plus, les membres de la communauté sont également ceux qui proposent le plus de variantes (TACorrBienCorr). Cela confirme une des observations formulées par l'Inria avec les métriques automatiques, qui affirmait que la communauté avait tendance à reformuler des fragments plus importants de la TA, alors que les traducteurs sont plus susceptibles de « respecter » les consignes établies de la post-édition et d'effectuer des modifications nécessaires. En revanche, par rapport à la communauté, les

traducteurs ont plus tendance à ne pas identifier des erreurs dans la TA, ce qui explique le pourcentage élevé de l'attribut « TAENonCorr » dans la PETrad. Ensuite, on observe également que ce sont les spécialistes du TAL qui introduisent le plus d'erreurs dans la post-édition (TACorrMalCorr), bien que l'écart soit faible entre la communauté et les traducteurs. Enfin, au total, les membres de la communauté ont commis 131 erreurs, alors que les traducteurs en ont commis 190.

L'hypothèse formulée au point précédent selon laquelle plus la TA est de bonne qualité, plus il y a de chance que l'attribut TACorrMalCorr (introduction d'une erreur non présente dans la TA) soit présent dans la PE, peut être vérifiée avec des statistiques sur cet attribut. Par exemple, la Figure 36 ci-dessous permet de visualiser la distribution de cet attribut par système de TA.

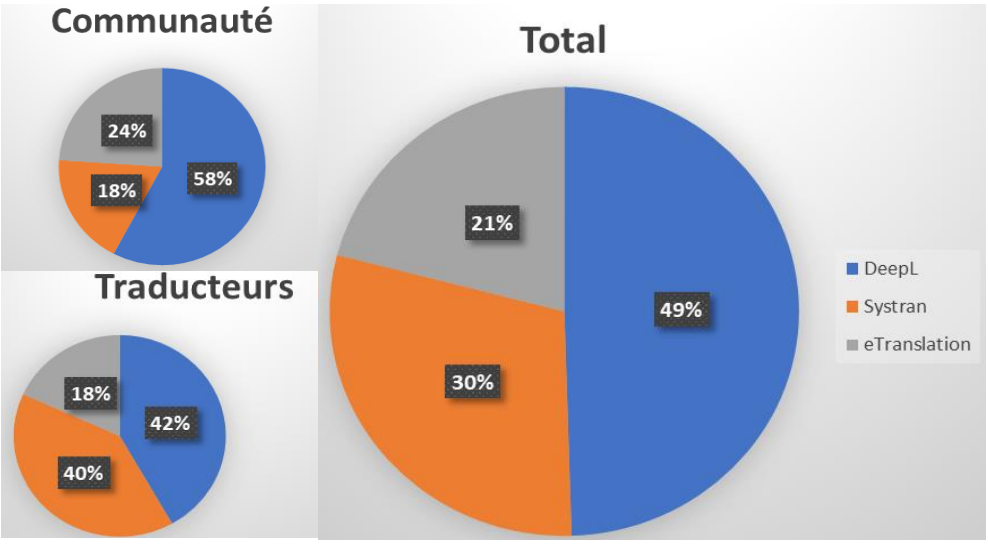

Figure 36 : distribution de l'attribut TACorrMalCorr par système de TA

Les trois graphiques indiquent effectivement que c'est avec la TA de DeepL (bleu) que l'attribut TACorrMalCorr est le plus représenté. Ce phénomène est particulièrement notoire pour les membres de la communauté (58 % de TACorrMalCorr pour les TA de DeepL). Par ailleurs, en moyenne, les textes traduits par eTranslation sont moins susceptibles d'entraîner l'introduction d'erreurs. Ces graphiques semblent donc corroborer l'hypothèse.

Pour une dernière vérification, on peut prendre les 10 textes dans lesquels il y a la plus grande représentation de l'attribut TACorrMalCorr, et observer si ceux-ci ont un score de qualité élevé (cf. Figure 37).

		TACorrMalCorr		
Système de TA	Score de qualité	PEComm	PETrad	Total
DeepL	90,37	1	2	3
Systran	83,57	0	3	3
DeepL	30,32	2	2	4
Systran	49,5	2	2	4
eTranslation	69,41	2	2	4
DeepL	94,59	4	0	4
DeepL	96,51	3	2	5
eTranslation	46,46	3	2	5
DeepL	97,08	3	3	6
DeepL	93,33	7	0	7

Figure 37 : 10 textes avec le plus d'attributs TACorrMalCorr

Le tableau ci-dessus ne permet pas de confirmer totalement l'hypothèse. En revanche, dans les 10 textes avec le plus d'attributs TACorrMalCorr, on remarque que 6 textes (surlignés en jaune) ont un score de qualité élevé (entre 80 et 100). Les deux textes qui ont le plus d'attributs TACorrMalCorr sont des traductions de DeepL ayant un score de qualité supérieur à 90. Dès lors, on peut affirmer qu'il y a une tendance à l'introduction d'erreurs dans les post-éditions de TA de bonne qualité, mais on ne peut pas affirmer que ce phénomène est propre aux TA de bonne qualité.

4.2.3. Statistiques relatives aux niveaux de gravité

Les scores de gravité attribués aux erreurs peuvent également donner des statistiques intéressantes sur les systèmes de traduction et sur les groupes de post-éditeurs.

a) Systèmes de TA

Le graphique en Figure 38 ci-dessous permet de visualiser la distribution des niveaux de gravité à travers les différents systèmes de traduction automatique.

Ce graphique peut expliquer, en partie, pourquoi eTranslation est le système qui obtient les moins bons scores de qualité. En effet, on observe que eTranslation commet plutôt des erreurs graves (niveau 2, coefficient de 5) et critiques (niveau 3, coefficient de 25). Ces poids d'erreurs contribuent de manière significative aux scores de qualité. En revanche, DeepL et Systran ont des profils similaires en termes de distribution des scores de gravité : pour chaque niveau de gravité, il y a environ 1 % d'écart entre les deux systèmes.

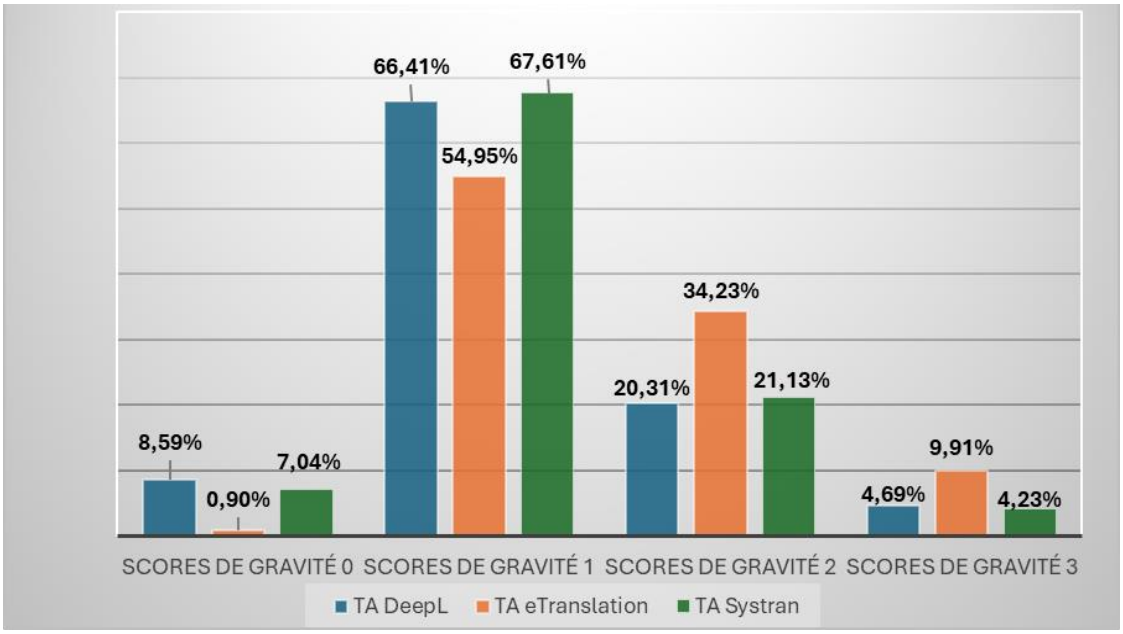

Figure 38 : distribution des scores de gravité (TA)

b) Groupes de post-éditeurs

Le graphique ci-dessous montre la répartition des niveaux d'erreurs entre les membres de la communauté du TAL (PEComm) et les membres de la communauté linguistique (PETrad).

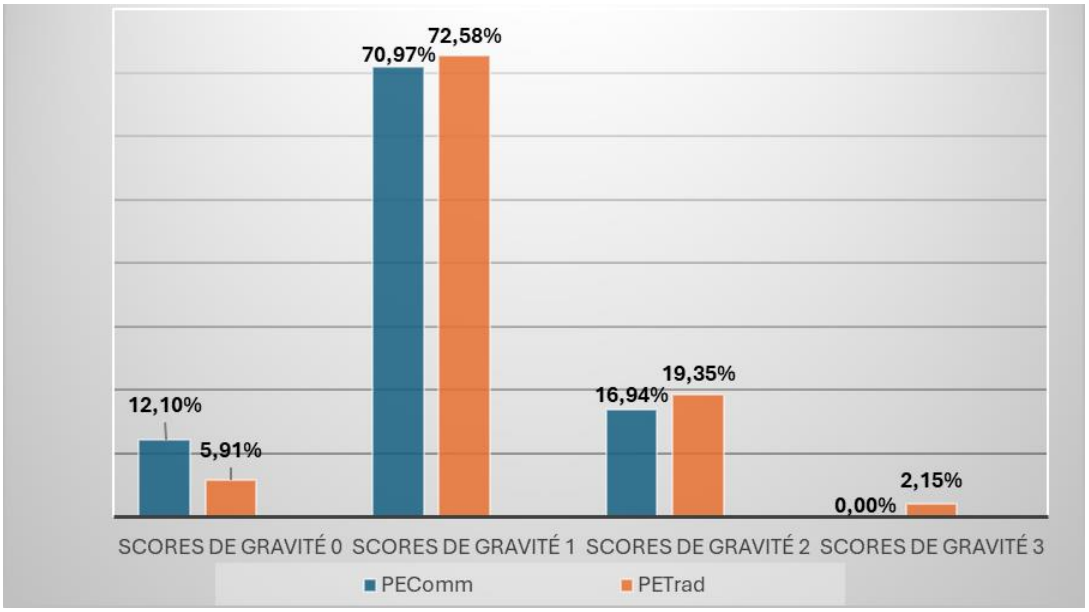

Figure 39 : distribution des scores de gravité (PE)

Premièrement, on remarque qu'en moyenne, il y a moins d'erreurs graves (niveaux 2 et 3) dans les post-éditions que dans les traductions automatiques (tous systèmes confondus). En moyenne, les post-éditeurs commettent plus d'erreurs de niveau 1 et de niveau 0, ce qui

explique également les scores de PE plus élevés que ceux des TA. En ce qui concerne les différences entre les spécialistes du TAL et les traducteurs, on remarque que la communauté fait davantage d'erreurs de niveau 0, et que les traducteurs font plus de « vraies » erreurs (niveaux 1, 2 et 3). Cette distribution des niveaux de gravité entre les deux groupes permet aussi d'expliquer les scores de qualité globale, qui sont meilleurs pour les membres de la communauté que pour les traducteurs.

Les erreurs de niveaux 3 effectuées par les traducteurs sont toutes les 4 concentrées dans la même post-édition. Il s'agit de la post-édition d'une TA générée par eTranslation, la plus mauvaise de toutes. La TA avait un score de -153,33 ; la PEComm avait un score de 92,38 ; et la PETrad restait très mauvaise (-28,71), étant donné que beaucoup d'erreurs n'étaient pas corrigées et qu'il reste 4 erreurs de niveau 3. Les erreurs de niveau 3 retrouvées dans cette PE sont des erreurs de groupes nominaux complexes, de distorsion et de choix incorrects terminologiques présentes dans la TA et particulièrement faciles à identifier. Par exemple, l'erreur terminologique était celle-ci :

Source	We model two important interfaces of constituency parsing with auxiliary tasks supervised at the word level […]
TA	Nous modélisons deux aspects importants de l'**analyse des circonscriptions** avec des tâches auxiliaires […]
PE	Nous modélisons deux aspects importants de l'**analyse des circonscriptions** avec des tâches auxiliaires […]

Le terme *circonscription* n'a aucun rapport avec le domaine du TAL, et une *analyse en constituants* est un terme courant dans ce domaine. Dès lors, on peut s'interroger sur le niveau d'attention du post-éditeur de ce texte, d'autant plus que toutes les erreurs de niveau 3 apparaissent uniquement dans cette PE. Dès lors, cette statistique sur la distribution des gravités d'erreurs entre la communauté et les traducteurs est à prendre avec précaution, étant donné que les erreurs de niveau 3 sont toutes concentrées dans un texte, et que le problème pourrait venir d'un seul post-éditeur, et non de l'ensemble.

4.2.4. *Statistiques sur les types d'erreurs*

4.2.4.1. *Erreurs fréquentes*

Enfin, les annotations humaines permettent aussi la création de statistiques sur les types d'erreurs effectuées et de comparer ainsi les types d'erreurs des différents systèmes de TA et des différents groupes de post-éditeurs. Toutes les erreurs sont définies dans le manuel d'annotation en Annexe 1.

Tout d'abord, les métriques automatiques semblant indiquer une prévalence des erreurs de terminologie, il est intéressant d'observer le ratio des erreurs terminologiques par rapport aux autres types d'erreurs.

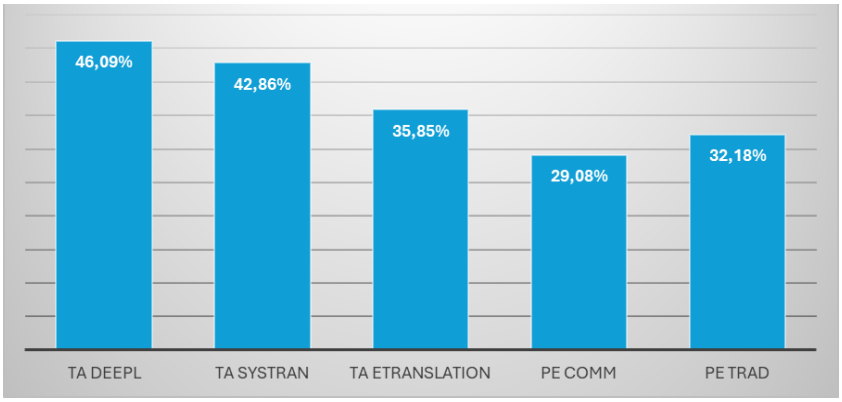

Figure 40 : prévalence des erreurs de terminologie

La Figure 40 montre effectivement que les erreurs terminologiques occupent une place significative dans l'ensemble de toutes les erreurs. Ce constat vaut surtout pour DeepL et Systran, dont plus de 40 % des erreurs commises sont des erreurs de terminologie. Bien qu'eTranslation propose les traductions jugées les plus mauvaises, c'est le système qui commet le moins d'erreurs de terminologie. Pour ce qui est des différents groupes de post-éditeurs, on constate que ce sont les traducteurs qui font le plus d'erreurs de terminologie, bien que l'écart entre les deux groupes soit faible (3,10 %).

Il est également intéressant d'observer les profils d'erreurs des trois systèmes de traduction automatique et de voir si ceux-ci commettent les mêmes types d'erreurs.

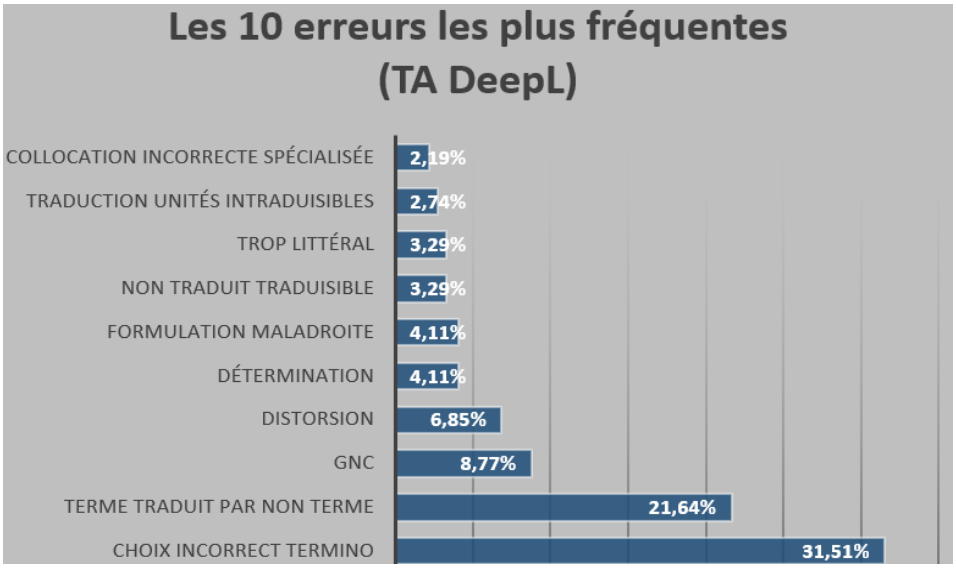

Figure 41 : DeepL - les 10 erreurs les plus fréquentes (%)

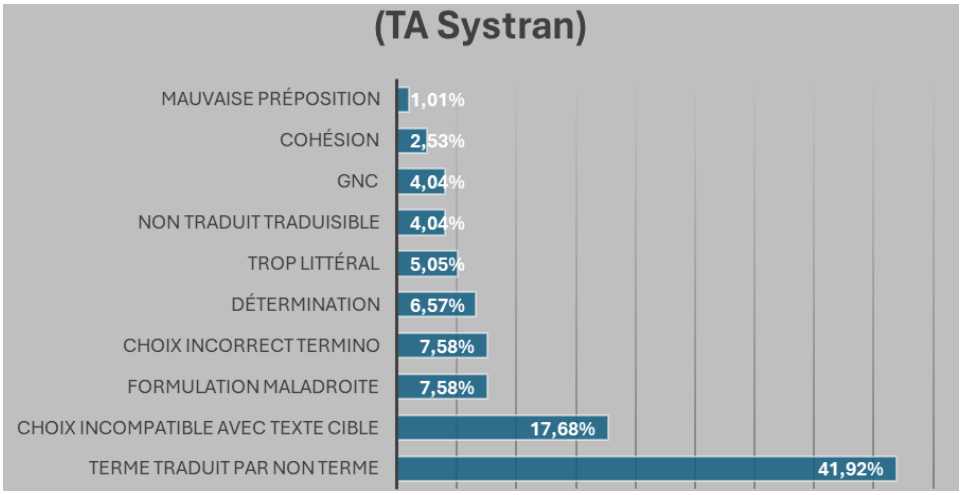

Figure 42 : Systran - les 10 erreurs les plus fréquentes (%)

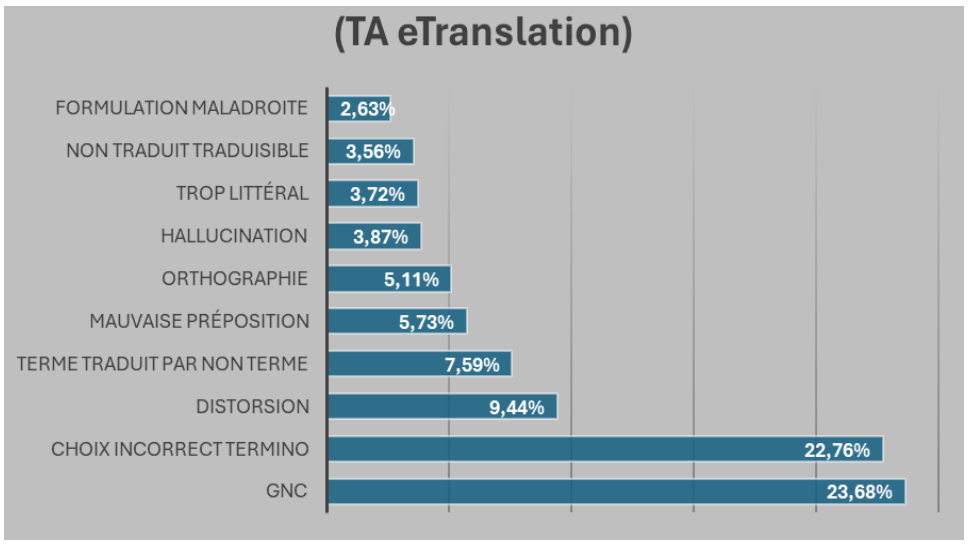

Figure 43 : eTranslation - les 10 erreurs les plus fréquentes (%)

Pour faciliter la comparaison des distributions des 10 erreurs les plus fréquentes par système de TA, la Figure 44 ci-dessous fournit un aperçu récapitulatif.

	DeepL	Systran	eTranslation
Traduction unités intraduisibles	•		
Collocation incorrecte spécialisée	•		
Trop littéral	•	•	•
Non traduit traduisible	•	•	•
Formulation maladroite	•	•	•
Détermination	•	•	
Distorsion	•		•
Groupes nominaux complexes	•	•	•
Terme traduit par non-terme	•	•	•
Choix incorrect terminologie	•	•	•
Mauvaise préposition		•	•
Cohésion		•	
Choix incompatible avec texte cible		•	
Hallucination			•
Orthographie			•

Figure 44 : synthèse - erreurs par système

Dans ce tableau, on remarque que 6 types d'erreurs figurent parmi les 10 erreurs les plus courantes pour chaque système de traduction automatique : les traductions trop littérales, les erreurs de « non traduit traduisible », les formulations maladroites, les groupes nominaux complexes, les termes traduits par des non-termes et les choix terminologiques incorrects.

En ce qui concerne les spécificités de DeepL, on observe que ce système produit plus d'erreurs de « traduction d'unités intraduisibles » et de collocation incorrecte en langue de spécialité que les deux autres systèmes. On remarque également que DeepL et eTranslation sont les deux seuls systèmes qui produisent beaucoup de distorsions, ce qui semble indiquer que Systran produit des traductions plus fidèles au contenu du texte source. Un phénomène notable pour eTranslation est qu'il est le seul système qui est sujet aux hallucinations. L'hallucination étant une erreur (par définition grave) où la traduction automatique est totalement déconnectée de l'énoncé de départ, ce phénomène démontre encore la qualité (beaucoup) plus faible de eTranslation par rapport aux autres systèmes, du moins dans le domaine du TAL. Les erreurs d'orthographie font également partie des spécificités de eTranslation. Enfin, Systran présente

deux spécificités en termes d'erreurs : les erreurs de cohésion et les erreurs terminologiques de choix incompatible avec le texte cible.

Un aperçu des distributions des grandes catégories d'erreurs (transfert de contenu, langue et outils) pour chaque système de TA peut également fournir des informations intéressantes.

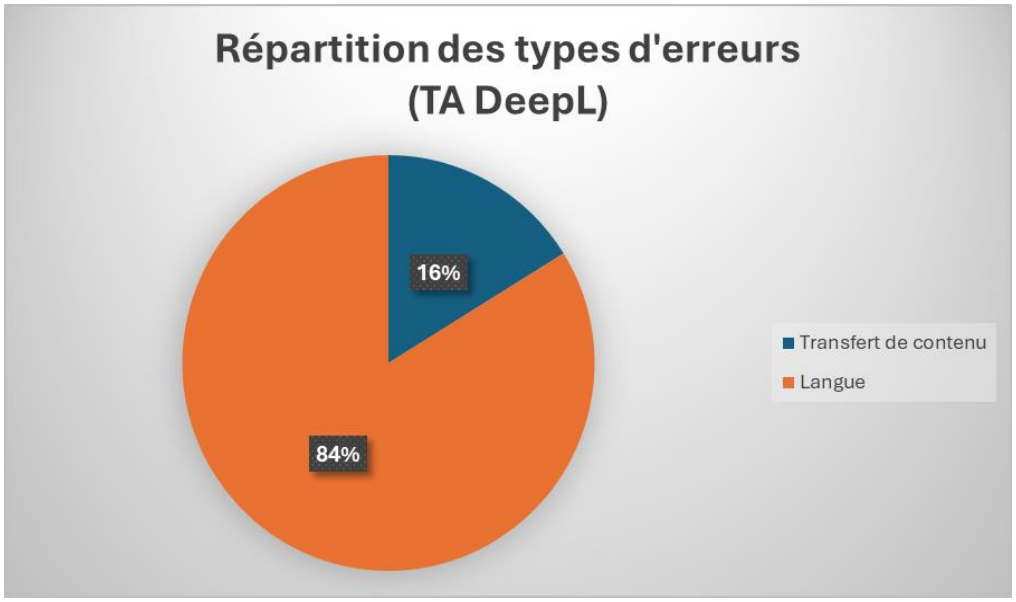

Figure 46 : distribution des catégories d'erreurs (DeepL)

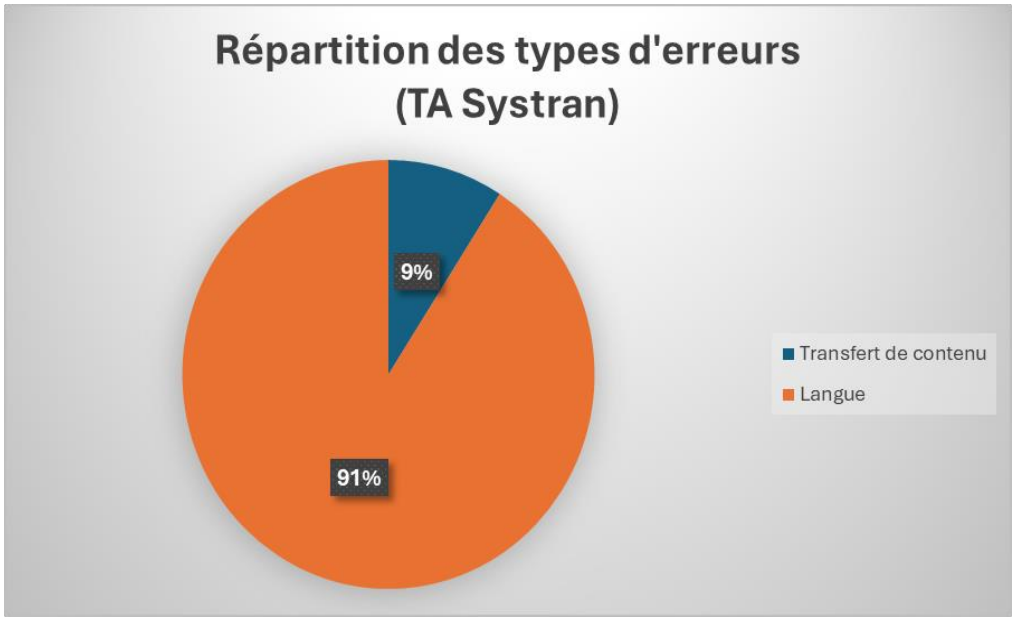

Figure 45 : distribution des catégories d'erreurs (Systran)

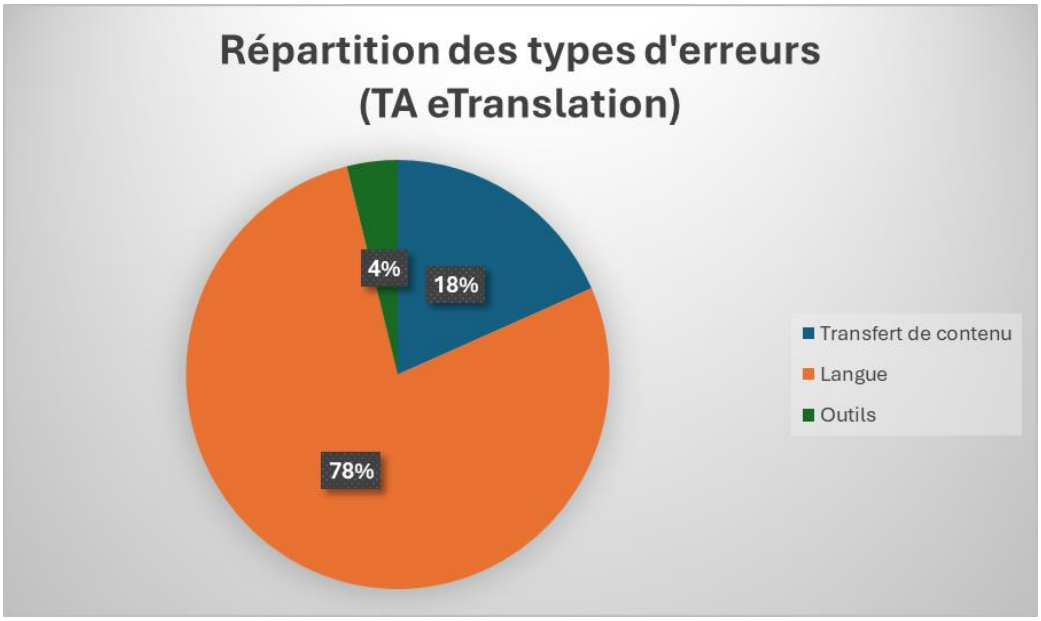

Figure 47 : distribution des catégories d'erreurs (eTranslation)

Les trois figures ci-dessus révèlent effectivement de subtiles nuances dans le profil des erreurs des différents systèmes de TA. DeepL et eTranslation ont des profils assez similaires en ce qui concerne la distribution entre les erreurs de transfert de contenu et de langue (environ 1/5 sont des erreurs de transfert de contenu et 4/5 des erreurs de langue).

En revanche, on remarque que eTranslation est le seul système qui génère des erreurs liées aux outils (uniquement des hallucinations de la TA).

Par ailleurs, on observe que Systran suit une tendance légèrement différente, puisqu'il génère moins d'erreurs de transfert de contenu que les deux autres systèmes, ce qui signifie *a priori* que Systran produit des TA dont le sens est plus proche du texte source que celles de DeepL et de eTranslation, comme observé au point précédent.

Il convient désormais d'observer si les deux groupes de post-éditeurs suivent des tendances similaires, ou si leurs profils divergent.

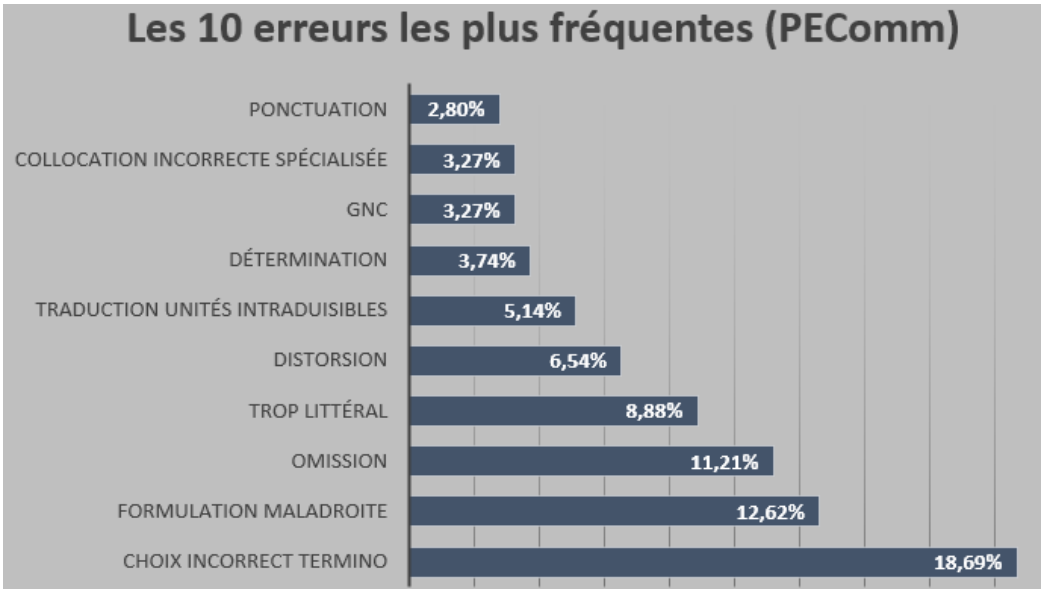

Figure 48 : PEComm - les 10 erreurs les plus fréquentes (%)

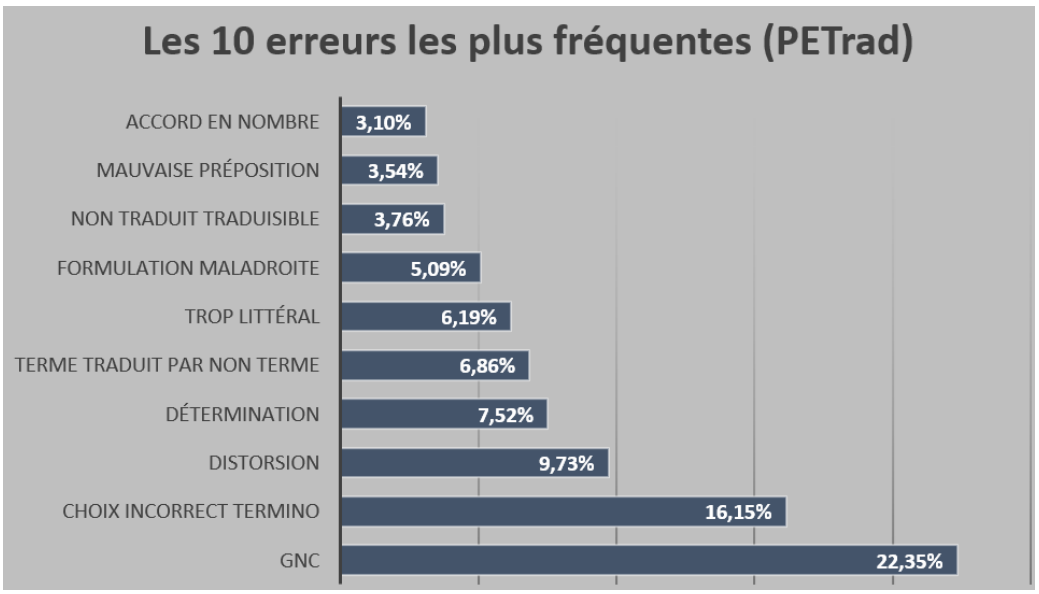

Figure 49 : PETrad - les 10 erreurs les plus fréquentes (%)

Pour rappel, la Figure 40 indiquait que les traducteurs commettaient plus d'erreurs de terminologie que les membres de la communauté du TAL. Or, on peut observer dans les deux graphiques ci-dessus que pour les spécialistes du TAL, l'erreur la plus fréquemment commise est une erreur de terminologie, ce qui n'est pas le cas pour les traducteurs. Pour y voir plus clair, voici un tableau récapitulatif :

	PEComm	PETrad
Choix incorrect terminologique	•	•
Formulation maladroite	•	•
Omission	•	
Trop littéral	•	•
Distorsion	•	•
Traduction unités intraduisibles	•	
Détermination	•	•
Groupes nominaux complexes	•	•
Collocation incorrecte spécialisée	•	
Ponctuation	•	
Terme traduit par non-terme		•
Non traduit traduisible		•
Mauvaise préposition		•
Accord en nombre		•

Figure 50 : synthèse - erreurs par groupe de post-éditeurs

Dans ce tableau, il y a uniquement 3 types d'erreurs qu'on ne retrouve pas dans les 10 erreurs le plus fréquentes pour les systèmes de TA : les omissions, les erreurs de ponctuation et d'accord en nombre. Dans ce tableau comparatif, on remarque qu'il y a 6 types d'erreurs partagés par les deux groupes de post-éditeurs, qui sont principalement des erreurs de langue et de terminologie. Toutefois, on observe aussi que les deux groupes commettent beaucoup d'erreurs de distorsion et de traductions trop littérales. Pour ce qui est des spécificités des post-éditions de la communauté, ce graphique révèle que les spécialistes du TAL ont plus tendance à faire des omissions (erreur de transfert de contenu) et, à commettre des erreurs de traduction d'unités intraduisibles (erreurs de transfert de contenu), de collocations spécialisées incorrectes (terminologie – langue) et de ponctuation (langue). En revanche, les traducteurs ont plus tendance à commettre des erreurs de prépositions (langue), d'accord en nombre (langue), de non-traduction d'unités traduisibles (transfert de contenu) et de termes traduits par des non-termes (terminologie – langue). Avec ce tableau, il est donc pertinent de supposer que les membres de la communauté du TAL commettent plus d'erreurs de transfert de contenu que les traducteurs ; à l'inverse, les traducteurs semblent commettre plus d'erreurs de langue.

Les graphiques suivants devraient permettre de corroborer ou d'infirmer cette hypothèse.

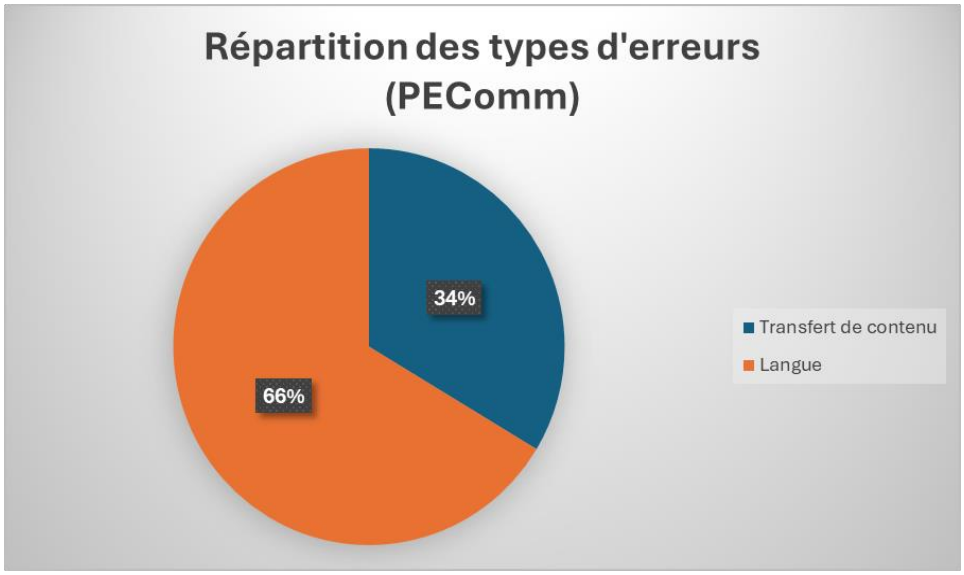

Figure 51 : distribution des catégories d'erreurs (PEComm)

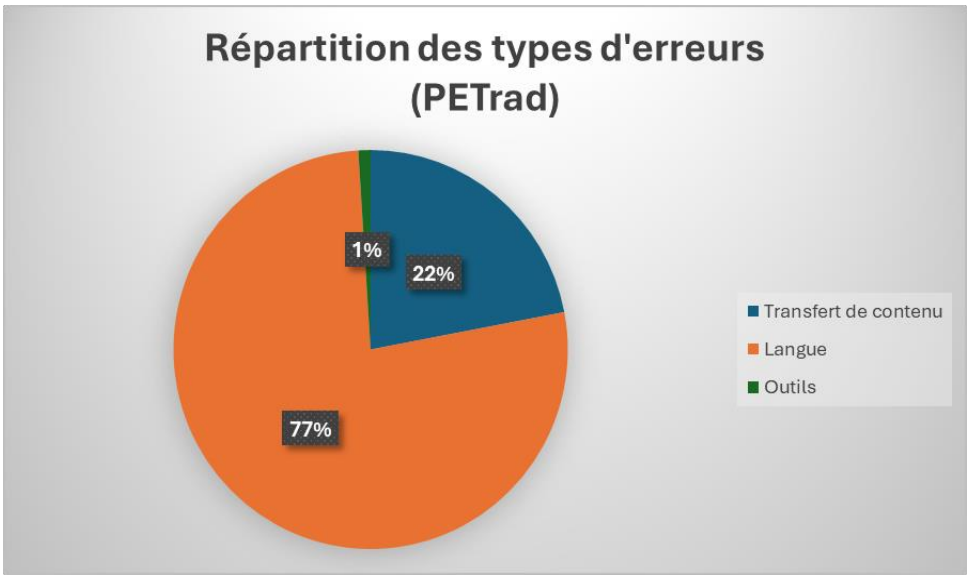

Figure 52 : distribution des catégories d'erreurs (PETrad)

Effectivement, ces deux graphiques corroborent l'hypothèse selon laquelle les traducteurs ont plutôt tendance à commettre des erreurs de langue (77 %, contre 66 % pour les membres de la communauté). Dès lors, la communauté du TAL produirait des post-éditions plus correctes sur les plans linguistique et terminologique, mais plus approximatives en termes de transfert de contenu. Par ailleurs, les traducteurs sont les seuls à produire des erreurs liées aux outils ; il s'agit d'une erreur de duplication (cf. Manuel d'annotation en Annexe 1).

4.2.4.2. Exemples des erreurs fréquentes

Nous illustrons ici chaque type d'erreurs fréquent à l'aide d'exemples provenant directement du corpus de TA et de PE annotées.

a) Traduction d'unités intraduisibles

Il y a une traduction d'unités intraduisibles « lorsqu'un élément est traduit dans la langue cible, alors qu'il convient de ne pas le traduire » (cf. Manuel d'annotation en Annexe 1).

Source	*Machine Translation, it's a question of style, innit? The case of English tag questions*
TA DeepL	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des **questions de tag** en anglais*

Dans cet exemple, DeepL a traduit le terme anglais *tag questions* par *questions de tag* en français, alors que ce terme n'a pas d'équivalent en français.

b) Collocation incorrecte spécialisée

Une *collocation* est une combinaison de « deux unités lexicales ou plus dont l'une au moins est un terme et dont la totalité des parties ne désigne pas un et un seul concept » (Brisson, 2019). Les erreurs de collocations incorrectes spécialisées concernent dès lors les collocations comprenant un terme du domaine de spécialité.

Source	*We train BPE-based attentive Neural Machine Translation systems with and without factored outputs using the open source nmtpy framework.*
TA eTranslation	*Nous **formons des systèmes de traduction automatique neuronales** attentifs basés sur BPE avec et sans sorties factorisées en utilisant le framework nmtpy open source.*

La collocation *former des systèmes de TA* est incorrecte. Dans le domaine du TAL, on parle d'*entraîner* les systèmes de TA.

c) Trop littéral

Les traductions trop littérales sont les traductions dont l'énoncé est trop proche de celui du texte source, impactant ainsi la lisibilité et le caractère naturel de la traduction.

Source	*Our approach uses automatically generated pairs of source sentences, where each pair tests one morphological contrast.*
TA DeepL	*Notre approche utilise des paires de phrases sources générées automatiquement, où **chaque paire teste un contraste morphologique**.*

Dans la TA, la formulation « chaque paire teste un contraste morphologique » est un anthropomorphisme. En français, on tend à éviter ces tournures. Cette erreur vient ici de la formulation du texte source. On pourrait se débarrasser de l'anthropomorphique en ajoutant le verbe *permettre*.

d) Non-traduit traduisible

Si un mot ou toute autre expression n'est pas traduite alors qu'une traduction est possible, il s'agit d'une erreur de non-traduit traduisible.

Source	*We also experiment with pre-trained word embeddings and Bertbased neural networks.*
TA DeepL	*Nous expérimentons également avec des **word embeddings** pré-entraînés et des réseaux neuronaux basés sur Bert.*

Le terme *word embeddings* dispose d'un équivalent en français : les *plongements lexicaux*.

e) Formulation maladroite

« Une formulation maladroite est une erreur de qui se caractérise par des choix de mots ou une structure de phrase peu idiomatiques, ce qui donne un aspect artificiel ou peu naturel dans la langue cible. Cette erreur peut affecter la lisibilité du texte traduit, le rendant souvent difficile à comprendre. » (Cf. Manuel d'annotation en Annexe 1)

Source	*The main approaches are presented from a largely historical perspective and in an intuitive manner, allowing the reader to understand the main principles without knowing the mathematical details.*
TA eTranslation	*Les approches principales sont présentées d'un point de vue largement historique et d'une manière intuitive, permettant au lecteur de comprendre les **principes principaux** sans connaître les détails mathématiques.*

Dans cet exemple, il n'y a aucune erreur à proprement parler, mais l'enchaînement des mots *principes* et *principaux*, qui ont la même racine, est dérangeant et peu naturel.

f) Détermination

Les erreurs de détermination concernent la non-utilisation ou la mauvaise utilisation des déterminants.

Source	*LIUM Machine Translation Systems for WMT17 News Translation Task*
TA eTranslation	*Systèmes de traduction automatique **Lium** pour WMT17 Nouvelles Tâche de Traduction*

Le LIUM est un laboratoire. Or, dans l'exemple ci-dessus, sans l'article défini *le* et la préposition *de*, on pourrait croire que Lium (mauvaise casse) est le nom du système de traduction automatique. Il s'agit donc d'une erreur de préposition *et* de détermination.

g) Distorsion

Les erreurs de distorsion sont des erreurs de déformation du sens du message source. En principe, les erreurs de distorsion sont causées par d'autres types d'erreurs.

Source	*We introduce a novel chart-based algorithm for span-based parsing of discontinuous constituency trees of block degree two, including ill-nested structures.*
TA DeepL	*Nous présentons un nouvel algorithme basé sur les diagrammes pour l'analyse syntaxique basée sur l'étendue des **arbres de circonscription** discontinus de degré de bloc deux, y compris les structures mal imbriquées.*

Dans cet exemple, il y a une erreur terminologique (*constituency trees = arbres en constituants*). Par ailleurs, le mot *circonscription* n'est plus en lien avec le domaine et ne veut ici rien dire. Dès lors, le sens du texte source est altéré.

h) Groupes nominaux complexes

Les erreurs de GNC, ou groupes nominaux complexes, sont liées à une « mauvaise identification de la tête du groupe nominal » ou à « une mauvaise factorisation des différents éléments du groupe nominal complexe » (cf. Manuel d'annotation en Annexe 1).

Source	*Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie*
TA DeepL	*Correction et validation de la dépendance syntaxique dans **le Rhapsodie de la banque d'arbres française parlée***

Le GNC de la traduction automatique de DeepL est totalement faux ; tout le groupe nominal a été mal analysé. On ne comprend dès lors plus que *Rhapsodie* est le nom du corpus, et qu'il s'agit d'un corpus arboré du français parlé.

i) Terme traduit par un non-terme

Cette erreur se produit lorsqu'un terme du texte source n'est pas traduit par un terme dans le texte cible.

Source	*Herein, we evaluate YASET on part-of-speech tagging and named entity recognition in a variety of text genres including articles from the biomedical literature in English and clinical narratives in French.*
TA DeepL	*Ici, nous évaluons YASET sur l'**étiquetage de la partie du discours** et la reconnaissance des entités nommées dans une variété de genres de textes, y compris des articles de la littérature biomédicale en anglais et des récits cliniques en français.*

Dans le domaine du TAL, on parle plutôt de l'étiquetage morpho-syntaxique.

j) Choix incorrect terminologique

À l'inverse, on considère qu'une erreur est un choix terminologique incorrect « lorsqu'un terme dans le texte source est traduit par un terme incorrect dans la traduction » (cf. Manuel d'annotation en Annexe 1).

Source	*To further characterize performance, we report distributions over 30 runs and different sizes of training datasets.*
TA DeepL	*Pour mieux caractériser les performances, nous rapportons les distributions sur 30 **exécutions** et différentes tailles d'ensembles de données d'entraînement.*

Dans le domaine du TAL, on parle plutôt d'*itérations*, et non d'*exécutions*.

k) Mauvaise préposition

Cette catégorie d'erreur concerne la mauvaise utilisation des prépositions.

Source	*The micro-syntactic annotation process, presented in this paper, includes a semi-automatic preparation of the transcription, the application of a syntactic dependency parser, […]*
TA DeepL	*Le processus d'annotation micro-syntactique, présenté dans cet article, comprend une préparation semi-automatique de la transcription, l'application d'un analyseur **de** dépendance syntaxique, […]*

Dans ce terme composé, on utilise plutôt la préposition *en.*

l) Cohésion

La cohésion textuelle peut se mesurer par le biais de marques linguistiques précises, telles que la coordination, les connecteurs logiques, les anaphores, etc. (voir, par exemple, Benali 2012 ; Halliday & Hasan 1976).

Source	*To further characterize performance, we report distributions over 30 runs and different sizes of training datasets.*
TA DeepL	*Pour mieux caractériser les **performances**, nous rapportons les distributions sur 30 exécutions et différentes tailles d'ensembles de données d'entraînement.*
PE	*Pour mieux caractériser les **performances de l'outil**, nous rapportons les distributions sur 30 itérations et différentes tailles de corpus d'entraînement.*

Dans la TA, on ne comprend pas directement de quelles performances il est question. Par ailleurs, dans la post-édition, le post-éditeur a décidé d'expliciter qu'il s'agissait des performances *de l'outil* (qui avait été évoqué plus tôt dans le texte).

m) Choix incompatible avec le texte cible

« On considère qu'un choix terminologique est incompatible avec le texte cible lorsqu'un terme du texte source est traduit par un terme théoriquement correct dans la langue cible, mais que le terme choisi n'est pas le terme approprié au vu de différents facteurs. » (Cf. Manuel d'annotation en Annexe 1)

Source	*We wrote the corresponding AZee discourse expressions for the entire video content, i.e. expressions capturing the forms produced by the signers and their associated meaning by combining known production rules, a basic building block for these expressions.*
TA Systran	*Nous avons écrit les expressions de discours AZee correspondantes pour l'ensemble du contenu vidéo, c'est-à-dire des expressions capturant les formes produites par les* ***signataires*** *et leur signification associée en combinant des règles de production connues, un bloc de base pour ces expressions.*

Le terme anglais *signer* peut avoir plusieurs traductions possibles. Toutefois, dans le domaine de la langue des signes, on parle de *signeurs*, qui sont ceux qui utilisent la langue des signes. L'équivalent *signataire* n'est dès lors pas en adéquation avec le texte et le domaine.

n) Hallucination

Une hallucination est une erreur — par définition grave — produite par un système de TA. Elle peut être définie comme une suite de « fragments de phrase complètement illogiques ajoutés ou remplacés dans la traduction ; [il peut s'agit de] termes inventés en raison d'une mauvaise segmentation ou factorisation des unités lexicales » (Hansen & Esperança-Rodier, 2022, traduction).

Source	*This paper describes LIUM submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs.*
TA eTranslation	*Cet article décrit les soumissions de LIUM à WMT17 News Translation Task pour l'anglais, l'allemand, l'anglais, l'anglais, le tchèque et l'anglais→****langue latine****.*

Ici, on peut supposer que les flèches présentes dans le texte source ont perturbé le système de TA (eTranslation), puisque *langue latine* est un élément totalement inventé dans la TA.

o) Orthographie

Une erreur d'orthographie concerne le non-respect des normes orthographiques établies.

Source	*We model two important in- terfaces of constituency parsing with aux- iliary tasks supervised at the word level: (i) part-of-speech (POS) and morpholog- ical tagging, (ii) functional label predic- tion.*
TA eTranslation	*Nous modélisons deux aspects importants de l'analyse des circonscriptions avec des tâches* ***aux iliaires*** *supervisées au niveau des mots: (I) la partie de la parole (POS) et l'étiquetage morphologique, (ii) la prédic-tion fonctionnelle de l'étiquette.*

Ici, le découpage du mot *auxiliary* dans le texte source a vraisemblablement perturbé le système de TA, qui a mal orthographié l'équivalent *auxiliaires*.

p) Omissions

On considère qu'il y a une omission quand il manque, dans la traduction, une idée que l'on retrouve dans le texte source.

Source	*We model two important interfaces of constituency parsing with auxiliary tasks.*
PEComm	*Nous modélisons deux aspects importants de l'analyse syntaxique avec des tâches auxiliaires.*

Dans cette exemple, le post-éditeur a omis l'idée des *constituants*. En effet, *constituency parsing* se traduit par *analyse syntaxique en constituants*.

q) Ponctuation

Cette catégorie d'erreur regroupe l'ensemble des « erreurs liées à la ponctuation (virgule omise ou superflue, point final manquant, espace (insécable) manquante avec un signe de ponctuation double, etc.) » (cf. Manuel d'annotation, Annexe 1).

Source	*Electronic versions of literary works abound on the Internet and the rapid dissemination of electronic readers will make electronic books more and more common.*
PEComm	*Les versions électroniques d'œuvres littéraires abondent sur Internet* ***et*** *la diffusion rapide des liseuses électroniques rendra les livres électroniques de plus en plus courants.*

Dans l'usage, il est préconisé d'utiliser une virgule avant *et* lorsque le sujet de la proposition change[22], ce qui est le cas ici. Sans cette virgule, il peut y avoir un risque d'équivoque, qui peut affecter la lisibilité et la clarté.

r) Accord en nombre

Cette catégorie sert à annoter les erreurs d'accord en nombre.

Source	*Parallel corpora can be leveraged to implement cross-lingual information retrieval or machine translation tools.*
PEComm	*Les corpus parallèles peuvent être utilisés pour mettre en œuvre des outils de recherche d'information ou de traduction automatique* ***multilingue****.*

Dans cet exemple, l'adjectif *multilingue* est censé s'accorder avec le substantif *outils*, ce qui n'est pas le cas dans cette post-édition, où il semble s'accorder avec le terme composé *traduction automatique*. Il s'agit également d'une erreur de groupe nominal complexe. Il n'est évidemment pas exclu qu'on observe plusieurs types d'erreurs pour un seul segment.

Plus d'informations sur toutes les catégories d'erreurs et davantage d'exemples de ces erreurs se trouvent dans le manuel d'annotation en Annexe 1.

4.2.4.3. Erreurs absentes

Il y a toutefois 7 types d'erreurs qui ne sont retrouvés ni dans les traductions automatiques, ni dans les post-éditions annotées dans le cadre de ce projet d'évaluation humaine. Il s'agit des erreurs suivantes :

- Traduction trop libre[23] (transfert de contenu),
- Accents diacritiques[24] (langue),

[22] https://www.btb.termiumplus.gc.ca/redac-chap?lang=fra&lettr=chapsect6&info0=6.1.

[23] « Une traduction trop libre est un traduction dont le sens diffère trop de celui transféré par le texte source, engendrant dès lors souvent une distorsion. » (Cf. Manuel d'annotation en Annexe 1)

[24] « Erreurs causées par la non-utilisation ou la mauvaise utilisation des accents, comme la confusion entre un accent aigu et un accent grave, par exemple. » (*loc. cit.*)

- Registre incompatible avec le texte source[25] (langue),
- Style du titre[26] (langue),
- Référence pas claire[27] (langue),
- Conformité au corpus[28]* (outils), et
- Choix incompatible avec le glossaire[29]* (outils).

L'absence des deux derniers types d'erreurs (*) peut s'expliquer par le fait qu'aucun corpus et aucun glossaire n'a été mis à la disposition des post-éditeurs dans le cadre de ce projet. Dès lors, il est impossible de retrouver ces erreurs dans les annotations.

[25] « Une erreur d'incompatibilité avec le texte source apparaît lorsque le registre utilisé dans la traduction ne correspond pas à celui employé dans le texte de départ (par exemple lorsqu'une expression vulgaire dans le texte source est « lissée » dans la traduction). » (*loc. cit.*)

[26] « Cette catégorie sert à annoter les erreurs de style dans les titres. En effet, les normes appliquées aux titres varient en fonction des langues. Par exemple, en anglais, on privilégie l'utilisation de la majuscule, ce qui n'est pas le cas en français. » (*loc. cit.*)

[27] « Une référence n'est pas claire quand l'élément auquel elle fait référence n'est pas immédiatement identifiable dans le texte. Cela peut se produire lorsqu'un pronom, un nom, un déterminant ou un autre élément est utilisé de manière ambiguë, ce qui rend difficile pour le lecteur de comprendre à quoi ou à qui il fait référence. » (*loc. cit.*)

[28] « Cette erreur regroupe […] les erreurs liées au non-respect du corpus. Par exemple, si un traducteur traduit un terme par un mauvais terme, il peut s'agir d'une erreur de choix incorrect terminologique, mais aussi d'une erreur de conformité au corpus. » (*loc. cit.*)

[29] Il y a une erreur de choix incompatible avec le glossaire lorsque le traducteur ne respecte pas les entrées terminologiques du glossaire ou de la base terminologique.

5. Interprétation et discussion des résultats

Dans un premier temps, le corpus des traductions automatiques et des post-éditions a permis à l'Inria d'appliquer des méthodes d'évaluation automatique à l'aide de métriques en vue de comparer les performances des systèmes de TA et des deux groupes de post-éditeurs. Le corpus annoté avec les types d'erreurs, les attributs et les scores de gravité a permis d'établir de nombreuses statistiques sur la qualité globale des TA et des PE, la distribution des scores de gravité et des attributs, ainsi que sur les différents types d'erreurs. Au vu de ces deux méthodes d'évaluation, il convient de répondre à deux questions :

a) L'évaluation humaine détaillée est-elle en accord avec les résultats des métriques automatiques ?
b) Ces évaluations permettent-elles d'établir des différences significatives entre les systèmes de traductions automatiques (DeepL, Systran et eTranslation) et d'affirmer que les deux groupes de post-éditeurs (spécialistes du TAL et traducteurs) ont des profils différents ?

D'un point de vue de la qualité des TA et des PE, les résultats des métriques automatiques ne sont pas réellement en adéquation avec ceux de l'évaluation humaine. Premièrement, l'évaluation automatique et l'évaluation humaine montrent toutes les deux que DeepL est le système de TA qui génère les meilleures traductions automatiques, suivi de Systran, puis de eTranslation. Les scores HTER et BLEU calculés dans le cadre de l'évaluation automatique indiquent que la différence entre les trois systèmes est peu significative. En revanche, l'évaluation humaine révèle une différence de qualité particulièrement importante entre Systran et eTranslation (près de 60 %), eTranslation générant des TA particulièrement mauvaises (score moyen d'environ 17).

Ensuite, en ce qui concerne les post-éditions, les deux types d'évaluation ne sont pas en accord. D'après les métriques automatiques, les post-éditions de la TA de DeepL étaient les meilleures, suivies de celles de Systran et enfin de celles de eTranslation. Or, selon les statistiques de l'évaluation humaine, ce sont les post-éditions de Systran qui sont de meilleure qualité, suivies par celles de DeepL, puis celles de eTranslation.

Les métriques automatiques indiquent, par ailleurs, une amélioration très légère entre les TA et les PE, et ce pour les deux groupes de post-éditeurs (moins de 1 % en moyenne). Or, l'évaluation humaine révèle des améliorations beaucoup plus significatives entre les TA et le

PE (+ 26 % en moyenne pour les membres de la communauté du TAL et + 16 % en moyenne pour les traducteurs). Les deux types d'évaluation s'accordent toutefois sur un aspect : les PE des membres de la communauté du TAL sont généralement de meilleure qualité que celles des traducteurs.

La métrique automatique HTER, qui sert à calculer la distance entre les TA et les PE, a révélé que les spécialistes du TAL effectuaient davantage de modifications que les traducteurs. Ce résultat est confirmé par l'évaluation humaine. En effet, on observe que l'attribut TACorrBienCorr est plus présent dans les post-éditions de la communauté du TAL que dans celles des traducteurs, ce qui signifie qu'ils introduisent davantage de variantes de traduction et qu'ils apportent — *a priori* — plus de modifications superflues que les traducteurs.

Pour ce qui est des erreurs commises et des niveaux de gravité de ces erreurs, les métriques automatiques donnent certaines pistes, mais cela ne se base que sur le feed-back des post-éditeurs, et non sur les TA et les PE elles-mêmes. Ce qu'on apprend avec les feed-backs des post-éditeurs, c'est que les erreurs qui sont, selon eux, les plus fréquentes et les plus graves dans les TA sont les erreurs de terminologie ; les erreurs de langue sont quant à elles moins fréquentes et associées à des niveaux de gravité plus faibles. Pour ce point, l'évaluation humaine permet des résultats plus précis et concrets. Cette dernière indique elle aussi que les erreurs de terminologie occupent une place de premier plan, tant dans les TA que dans les PE, avec une prépondérance plus marquée dans les TA, surtout pour DeepL et Systran. Pour ce qui est des deux groupes de post-éditeurs, on observe que les traducteurs commettent plus d'erreurs de terminologie que les spécialistes du TAL, bien que la différence soit minime (environ 3 %). En général, on constate que DeepL et eTranslation ont un profil assez similaire en termes de distribution des types d'erreurs : environ 1/4 des erreurs sont des erreurs de transfert de contenu, et 3/4 des erreurs sont des erreurs de langue (y compris de terminologie). En revanche, Systran commet davantage d'erreurs de langue et de terminologie (environ 90 %) que d'erreurs de transfert de contenu (environ 10 %). Par conséquent, Systran générerait des traductions plus correctes en termes de sens et de contenu que DeepL et eTranslation, qui eux fourniraient des TA plus correctes sur le plan linguistique et terminologique. On constate toutefois que eTranslation est le seul système sujet aux hallucinations (erreur liée aux outils), un phénomène par définition grave, ce qui explique sa plus mauvaise qualité globale. Enfin, les deux groupes de post-éditeurs suivent des tendances similaires en termes de distribution des types d'erreurs, mais avec quelques différences. On remarque en effet que les membres de la communauté du

TAL font plus d'erreurs de contenu que les traducteurs (respectivement 34 % contre 22 %), et moins d'erreurs de langue que les traducteurs (respectivement 66 % contre 77 %). Par ailleurs, les traducteurs sont les seuls à commettre des erreurs liées aux outils (1 %). Dès lors, on peut *a priori* affirmer que les membres de la communauté du TAL produisent des post-éditions plus correctes que les traducteurs sur le plan linguistique et terminologique. En revanche, les traducteurs produisent des post-éditions qui rendent mieux le sens et le contenu du texte source que les spécialistes du TAL.

En ce qui concerne la distribution des niveaux de gravité entre les systèmes de TA et entre les deux groupes de post-éditeurs, l'évaluation humaine permet également de faire des statistiques intéressantes. Pour les systèmes de TA, on remarque que DeepL et Systran présentent des profils similaires : ils commettent en grande majorité des erreurs de niveau 1, puis de niveau 2, et, en moindre mesure, des erreurs de niveau 3 et de niveau 0. En revanche, eTranslation commet plus d'erreurs graves (de niveau 3) que les deux autres systèmes, et moins d'erreurs de niveau 0, ce qui explique une fois de plus ses TA de plus mauvaise qualité. Quant aux deux groupes de post-éditeurs, ils suivent la même tendance. Toutefois, à l'inverse des traducteurs, les spécialistes du TAL ne commettent aucune erreur de niveau critique (3) ; en outre, les spécialistes du TAL commettent plus d'erreurs de niveau 0 que les traducteurs, ce qui peut expliquer leurs scores de qualité généralement plus élevés.

Enfin, les statistiques établies avec les attributs permettent de formuler une dernière observation. La présence de l'attribut « TACorrMalCorr », qui rend compte de l'introduction d'une erreur dans la PE non présente dans la TA, est légèrement plus importante dans les PE des spécialistes du TAL que dans celles des traducteurs. L'observation des *scorecards* présentant un grand nombre d'attributs TACorrMalCorr montre que ceux-ci semblent apparaître de manière prépondérante dans les textes qui ont une TA de bonne, voire de très bonne qualité. Ce phénomène aurait dès lors une implication : lorsque peu de modifications obligatoires sont à apporter, les post-éditeurs sont plus susceptibles d'apporter des modifications erronées et d'introduire des erreurs dans la post-édition.

Au vu des points précédents, on peut affirmer que les deux évaluations rendent compte, de manière générale, de tendances similaires, mais l'évaluation humaine, plus chronophage, permet des nuances plus fines et des résultats plus précis. Les métriques automatiques, et surtout l'évaluation humaine, permettent de comparer les différents systèmes de traduction et d'établir les profils des deux groupes de post-éditeurs.

6. Conclusion

L'objectif de ce travail était d'évaluer la qualité des traductions automatiques et post-éditions de résumés d'articles scientifiques dans le domaine du traitement automatique des langues. Les traductions automatiques ont été générées par trois systèmes de TA commerciaux différents, à savoir DeepL, Systran et eTranslation. Deux publics différents ont effectué les post-éditions : d'un côté, des spécialistes du traitement automatique des langues ; de l'autre, des membres de la communauté linguistique (enseignants et étudiants en traduction, traducteurs).

L'annotation des TA et des PE a permis de répondre à différentes questions : les systèmes de TA génèrent-ils des traductions de qualité comparable ? Les deux publics (spécialistes vs linguistes) ont-ils des profils de post-éditeurs différents ? Les métriques d'évaluation automatique et humaine présentent-elles des résultats différents ?

En ce qui concerne les performances des systèmes de TA, les deux types d'évaluation révèlent la même tendance, bien que l'évaluation humaine permette des observations plus fines. DeepL est le système qui génère les meilleures TA, suivi de près par Systran. En revanche, eTranslation est largement moins performant que les deux autres systèmes et produit souvent des traductions très mauvaises (scores négatifs). Quant à leur profil, l'évaluation humaine a permis d'observer que DeepL et eTranslation avaient des profils similaires en ce qui concerne la distribution entre les erreurs de transfert de contenu (environ 20 %) et les erreurs de langue (environ 80 %). Par ailleurs, eTranslation est le seul système sujet aux hallucinations. Systran génère quant à lui moins d'erreurs de sens que les autres systèmes (environ 10 %) et produit dès lors des TA plus correctes en termes de contenu.

Pour les données concernant les groupes de post-éditeurs, les deux types d'évaluation ne sont pas totalement en adéquation. L'évaluation humaine a permis de montrer que ce sont les post-éditions de Systran qui étaient les meilleures, suivies de celles de DeepL, puis celles de eTranslation. Les résultats montrent aussi que les post-éditions des spécialistes du TAL sont, en moyenne, meilleures que celles des traducteurs (10 % d'écart). En ce qui concerne leurs profils, l'annotation a mis en avant que les spécialistes du TAL commettent plus d'erreurs de transfert de contenu que les traducteurs, alors que les traducteurs commettent davantage d'erreurs de langue (y compris de terminologie). Ces résultats ne sont pas en accord avec les hypothèses formulées dans l'introduction. Dès lors, *a priori*, les membres de la communauté du TAL produisent des PE plus correctes sur le plan linguistique et terminologique, alors que

les membres de la communauté linguistique produisent des PE plus correctes d'un point de vue contenu et sens. Ces résultats corroborent un point déjà évoqué dans la littérature scientifique : les évaluations humaines permettent des observations plus précises et plus fines que les métriques automatiques.

L'évaluation humaine de cette étude consistait en une analyse qualitative de TA et de PE. Grâce à une typologie d'erreurs améliorée aux fins de ce travail, à un corpus de résumés d'articles scientifiques, de TA et de PE compilé dans le cadre du projet ANR MATOS et à un manuel d'annotation créé pour ce travail, une vingtaine de textes ont pu être annotés, ce qui a donné des statistiques instructives concernant les profils des différents systèmes de TA et des post-éditeurs.

Néanmoins, pour ce travail, seuls les résumés bénéficiant d'une post-édition par chaque groupe de post-éditeurs ont été annotés, laissant ainsi de côté plusieurs centaines de textes du corpus. Dès lors, les résultats obtenus avec cette analyse pourraient être corroborés ou nuancés en annotant une plus grande quantité de textes du corpus. Par ailleurs, il pourrait être pertinent de tester l'efficacité de la typologie d'erreurs et du manuel d'annotation développés dans le cadre de ce travail en mettant en place un protocole d'annotation avec plusieurs annotateurs. Celui-ci permettrait de calculer un score inter-annotateur, permettant ainsi d'évaluer l'efficacité de la typologie et du manuel. Enfin, il serait également intéressant de mener une étude similaire sur d'autres domaines de spécialité, pour voir si cela donnerait des résultats semblables.

L'évaluation humaine reste chronophage. Les métriques automatiques peuvent constituer une solution à ce problème. Toutefois, ce travail a permis de pointer du doigt les faiblesses des métriques automatiques par rapport aux méthodes d'évaluation humaine. Par exemple, les métriques automatiques constituent une méthode d'évaluation quantitative et non qualitative, puisqu'elles permettent uniquement de dire (en théorie) quelles sont les meilleures traductions, mais elles ne révèlent pas les forces et les faiblesses (a) des systèmes de TA et (b) des différentes traductions. Une solution judicieuse pourrait être la mise en place d'un cadre d'évaluation hybride et qualitatif en exploitant l'IA générative, combinée aux évaluateurs humains. L'exploitation de l'IA générative et des grands modèles de langage (LLM), en particulier de ChatGPT, commence à devenir un axe de recherche majeur en traitement automatique des langues, notamment pour la traduction (voir, par exemple, He, 2024 ; Jiao et al., 2023 ; L. Wang et al., 2023). Dans ce pan de recherche, les chercheurs semblent s'accorder sur un point : l'avenir de la traduction automatique sera lié aux LLM et à l'IA générative. En

toute logique, si l'IA générative — du moins les modèles plus avancés — enregistre des performances à l'état de l'art, ces modèles devraient pouvoir être exploités pour évaluer les traductions. Cet axe de recherche est, à l'heure actuelle, très peu développé. *A priori*, pour l'instant, seules deux expériences d'évaluation par l'IA générative ont été menées (cf. Kocmi & Federmann, 2023 ; Qingyu Lu et al., 2023). Ces deux études ont donné des résultats prometteurs quant aux performances des GPT (transformeurs génératifs pré-entraînés) dans l'évaluation de traductions grâce à l'élaboration de *prompts* appropriés, c'est-à-dire d'entrées textuelles formulées par l'utilisateur servant à guider le modèle d'IA dans sa production et sa sortie (Amatriain, 2024). Toutefois, aucune des études n'utilise une typologie d'erreurs précises pour l'évaluation par l'IA. Il serait dès lors intéressant d'exploiter la typologie d'erreurs ainsi que le manuel d'annotation développés dans le cadre de ce travail pour apprendre aux modèles d'IA générative à évaluer des traductions et voir dans quelle mesure les sorties produites par l'IA peuvent être exploitées par un humain dans le cadre d'une évaluation humaine assistée par l'IA. Nous pouvons donc avancer que cette méthode d'évaluation hybride rendrait l'évaluation manuelle moins chronophage.

Table des figures

Bibliographie

Corpus de textes sources

Ahmia, O., Béchet, N., & Marteau, P.-F. (2018). Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japon.

Bawden, R. (2017). Machine Translation, it's a question of style, innit ? The case of English tag questions. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 2497-2502.

Bawden, R., Bottala, M.-A., Gerdes, K., & Kahane, S. (2014). Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 1-6.

Bertin-Lemée, É., Braffort, A., Challant, C., Danet, C., Dauriac, B., & et al. (2022). Rosetta-LSF: an Aligned Corpus of French Sign Language and French for Text-to-Sign Translation. *13th Conference on Language Resources and Evaluation (LREC 2022)*. 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

Burlot, F., Garcia-Martinez, M., Barrault, L., Bougares, F., & Yvon, F. (2017). Word Representations in Factored Neural Machine Translation. *Conference on Machine Translation, Association for Computational Linguistics*, 43-55.

Burlot, F., & Yvon, F. (2017). Evaluating the morphological competence of Machine Translation Systems. *2nd Conference on Machine Translation (WMT17)*, 43-55.

Challant, C., & Filhol, M. (2022). A First Corpus of AZee Discourse Expressions. *Language Resources and Evaluation Conference,*. Language Resources and Evaluation Conference, Marseille, France.

Coavoux, M., & Crabbé, B. (2017). Multilingual Lexicalized Constituency Parsing with Word-Level Auxiliary Tasks. *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, 331-336.

Corro, C. (2020). Span-based discontinuous constituency parsing : A family of exact chart-based algorithms with time complexities from O(n^6) down to O(n^3). *Empirical Methods in Natural Language Processing*. Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic (online).

Corro, C., & Titov, I. (2019). Learning Latent Trees with Stochastic Perturbations and Differentiable Dynamic Programming. *57th annual meeting of Association for Computational Linguistics*. 57th annual meeting of Association for Computational Linguistics, Florence, Italy.

Duquenne, P.-A., Gong, H., Sagot, B., & Schwenk, H. (2022). T-Modules : Translation Modules for Zero-Shot Cross-Modal Machine Translation. *EMNLP 2022 - 2022 Conference on Empirical Methods in Natural Language Processing,*. EMNLP 2022 - 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates.

Fakih, A., Ghassemiazghandi, M., Hafeed Fakih, A., & K. M. Singh, M. (2024). Evaluation of Instagram's Neural Machine Translation for Literary Texts : An MQM-Based Analysis. *GEMA Online® Journal of Language Studies*, *24*(1), 213-233. https://doi.org/10.17576/gema-2024-2401-13

Fourrier, C., Bawden, R., & Sagot, B. (2021). Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task? *ACL-IJCNLP 2021 - Findings of the Association for Computational Linguistics*. ACL-IJCNLP 2021 - Findings of the Association for Computational Linguistics, Bangkok, Thailand.

Futeral, M., Schmid, C., Laptev, I., Sagot, B., & Bawden, R. (2023). Tackling Ambiguity with Images : Improved Multimodal Machine Translation and Contrastive Evaluation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5394-5413. https://doi.org/10.18653/v1/2023.acl-long.295

Garcia-Martinez, M., Caglayan, O., Aransa, W., Bardet, A., Bougares, F., & et al. (2017). LIUM Machine Translation Systems for WMT17 News Translation Task. *Second Conference on Machine Translation*, 288-295.

Ho, A. K. N., & Yvon, F. (2020). *Neural Baselines for Word Alignment* (arXiv:2009.13116). arXiv. http://arxiv.org/abs/2009.13116

Lerner, P., Ferret, O., & Guinaudeau, C. (2023). *Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2301.04366

Névéol, A., Jimeno Yepes, A., Neves, L., & Verspoor, K. (2018). Parallel Corpora for the Biomedical Domain. *International Conference on Language Resources and Evaluation*. International Conference on Language Resources and Evaluation, Miyazaki, Japon.

Pham, M.-Q., Crego, J.-M., & Yvon, F. (2022). Multi-Domain Adaptation in Neural Machine Translation with Dynamic Sampling Strategies. *Conference of the European Association for Machine Translation, European Association for Machine Translation*. Conference of the European Association for Machine Translation, European Association for Machine Translation, Ghent, Belgium.

Pham, M.-Q., Crego, J.-M., Yvon, F., & Senellart, J. (2020). A Study of Residual Adapters for Multi-Domain Neural Machine Translation. *Conference on Machine Translation*. Conference on Machine Translation, United States (online).

Pham, M.-Q., Xu, J., Crego, J.-M., Senellart, J., & Yvon, F. (2020). Priming Neural Machine Translation. *Conference on Machine Translation*. Conference on Machine Translation, United States (online).

Poibeau, T. (2017). Machine Translation. *MIT Press*.

Sadoun, D., Mkhitaryan, S., Nouvel, D., & Valette, M. (2016). ReadME generation from an OWL ontology describing NLP tools. *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, 46-49. https://doi.org/10.18653/v1/W16-3509

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., & Van Der Wal, O. (2022). You reap what you sow : On the Challenges of Bias Evaluation Under Multilingual Settings. *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models*, 26-41. https://doi.org/10.18653/v1/2022.bigscience-1.3

Wisniewski, G., Zhu, L., Ballier, N., & Yvon, F. (2021). Screening Gender Transfer in Neural Machine Translation. *Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for computational linguistics*. Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for computational linguistics, Punta Cana, Dominican Republic.

Yu, Q., Max, A., & Yvon, F. (2012). Aligning Bilingual Literary Works : A Pilot Study. *NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, ACL*, 36-44.

Références bibliographiques

Amatriain, X. (2024). *Prompt Design and Engineering : Introduction and Advanced Methods*. https://doi.org/10.48550/ARXIV.2401.14423

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65-72.

Bawden, R., Peng, Z., Bénard, M., Villemonte de la Clergerie, E., Esamotunu, R., Huguin, M., Kübler, N., Mestivier, A., Michelot, M., Romary, L., Zhu, L., Yvon, F. (2024). Translate your Own: a Post-Edition Experiment in the NLP domain. *The 25th Annual Conference of The European Association for Machine Translation (EAMT 2024)*. 24 - 27 June 2024. Sheffield, UK

Benali, A. (2012). Les problèmes de la catégorisation textuelle : Entre fondements théoriques et fondements structurels. *Synergies Algérie*, *17*, 35-49.

Bénard, M. (2020). *Évaluation de la qualité des systèmes neuronaux en matière de traduction des groupes nominaux complexes à prémodification nominale* [Mémoire].

Bénard, M., Bordet, G., & Kübler, N. (2022). Réflexions sur la traduction automatique et l'apprentissage en langues de spécialité. *ASp*, *82*, 81-98. https://doi.org/10.4000/asp.8113

Bénard, M., Kübler, N., Mestivier, A., Zhu, L., Bawden, R., de la Clergerie, E., Romary, L., Huguin, M., Nominé, J.-F., Peng, Z., & Yvon, F. (2023, juin). *MaTOS: traduction automatique pour la science ouverte*. Actes de CORIA-TALN 2023, Paris, France.

Brisson, F. (2019). *Les compétences terminologiques du traducteur : Pistes de réflexion pour un enseignement de la terminologie à l'usage de futurs traducteurs.* Université Savoie Mont Blanc.

Burchardt, A. (2013). Multidimensional quality metrics : A flexible system for assessing translation quality. *Proceedings of Translating and the Computer 35*.

Castagnoli, S., Ciobanu, D., Kerstin, K., Volanschi, A., & Kübler, N. (2011). Designing a Learner Translator Corpus for Training Purposes. *Corpora, Language, Teaching and Resources: from Theory to Practice*.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment : From Principles to Practice. In *Translation Quality Assessment. From Principles to Practice* (Springer, 1-1).

Comelles, E., Arranz, V., & Castellón, I. (2016). Guiding automatic MT evaluation by means of linguistic features. *Digital Scholarship in the Humanities*, fqw042. https://doi.org/10.1093/llc/fqw042

Comparin, L. (2016). *Quality in machine translation and human post-editing : Error annotation and specifications* [Mémoire].

Comparin, L., & Mendes, S. (2017). *Using error annotation to evaluate machine translation and human post-editing in a business environment*.

Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A., & Awad, M. (2010). Evaluation of Machine Translation Errors in English and Iraqi Arabic. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

Čulo, O., Gutermuth, S., Hansen-Schirra, S., & Nitzke, J. (2014). The Influence of Post-Editing on Translation Strategies. *Post-editing of Machine Translation: Processes and Applications*, 200-218.

Daems, J., De Clercq, O., & Macken, L. (2018). Translationese and Post-editese : How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies, 16*. https://doi.org/10.52034/lanstts.v16i0.434

Deneufbourg, G. (2019). Post-édition de traduction automatique : Se méfier des apparences. *ATA Journal*.

Depraetere, I. (2010). *What counts as useful advice in a university post-editing training context ? Report on a case study.* Communication présentée au 14th Annual EAMT Conference.

Dony, C., Kuchma, I., Neugebauer, T., Nominé, J.-F., Ševkušić, M., & Shearer, K. (2023). *Is there a case for accepting machine translated scholarly content in repositories?*

Farrús, M., Costa-jussà, M. R., Mariño, J. B., & R. Fonollosa, J. A. (2010). Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Federico, M., Negri, M., Bentivogli, L., & Turchi, M. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1643-1653. https://doi.org/10.3115/v1/D14-1172

Federmann, C. (2012). Appraise : An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics (PBML)*.

Fiorini, S., Garnier-Rizet, M., Hernandez Morin, K., Barbin, F., Humphreys, F., Josselin-Leray, A., Kübler, N., Loock, R., Martikainen, H., Nominé, J.-F., Plag, C., Rossi, C., & Yvon, F. (2020). *Rapport du groupe de travail Traductions et science ouverte*. Ministère de l'enseignement supérieur et de la recherche. https://doi.org/10.52949/20

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382. https://doi.org/10.1037/h0031619

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

Gledhill, C., Froeliger, N., Zimina-Poirot, M. (2023). "Intégrer des plateformes de traduction automatique neuronale dans l'enseignement de la traduction spécialisée". *Traduction*

humaine et traitement automatique des langues - Vers un nouveau consensus ?, Edizioni Ca Foscari.

Gledhill, C., & Zimina-Poirot, M. (2019). *The Impact of Machine Translation on a Masters Course in Web Translation: From Disrupted Practice to a Qualitative Translation/Revision Workflow*. Translating and the Computer 41, AsLing, The International Association for Advancement in Language Technology, Nov 2019, London, United Kingdom. pp.60-73.

Gordin, M. D. (2015). *Scientific Babel : How science was done before and after global English*. The University of Chicago Press.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English* (Longman).

Hansen, D., & Esperança-Rodier, E. (2022). *Human-Adapted MT for Literary Texts : Reality or Fantasy?* 178-190.

He, S. (2024). *Prompting ChatGPT for Translation : A Comparative Analysis of Translation Brief and Persona Prompts*. https://doi.org/10.48550/ARXIV.2403.00127

Hidalgo-Ternero, C. M. (2021). Google Translate vs. DeepL : Analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monografías de Traducción e Interpretación*, 154-177. https://doi.org/10.6035/MonTI.2020.ne6.5

House, J. (2015). *Translation quality assessment : Past and present*. Routledge.

Hutchins, W. J. (1995). Machine Translation : A Brief History. In *Concise History of the Language Sciences* (p. 431-445). Elsevier. https://doi.org/10.1016/B978-0-08-042580-1.50066-0

Isabelle, P., Cherry, C., & Foster, G. (2017). A Challenge Set Approach to Evaluating Machine Translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2486-2496. https://doi.org/10.18653/v1/D17-1263

Jia, Y., Carl, M., & Wang, X. (2019). How Does the Post-Editing of Neural Machine Translation Compare with From-Scratch Translation ? A Product and Process Study. *Journal of Specialised Translation*, *31*, 60-86.

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine*. https://doi.org/10.48550/ARXIV.2301.08745

Kamadjeu, R. (2019). English : The lingua franca of scientific research. *The Lancet Global Health*, *7*(9), e1174. https://doi.org/10.1016/S2214-109X(19)30258-X

Kirchhoff, K., Capurro, D., & Turner, A. (2012). Evaluating User Preferences in Machine Translation Using Conjoint Analysis. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*.

Koby, G. S., Fields, P., Hague, D., Lommel, A., & Melby, A. (2014). Defining Translation Quality. *Revista Tradumàtica: tecnologies de la traducció*, *12*, 413-420.

Kocmi, T., & Federmann, C. (2023). *Large Language Models Are State-of-the-Art Evaluators of Translation Quality*. https://doi.org/10.48550/ARXIV.2302.14520

Kübler, N. (2008). *MeLLANGE Final Report*. CLILLAC-ARP (EA_3967) - Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus.

Kübler, N., Martikainen, H., Mestivier, A., & Pecman, M. (2021). *Using corpora for post-editing neural MT in highly specialised domains : The case of complex noun phrases*.

UCCTS 2021 | Using Corpora in Contrastive and Translation Studies (6th edition), Sep 2021, Bertinoro, Italy.

Kübler, N., Mestivier, A., Pecman, M., & Zimina, M. (2016). *Exploitation quantitative de corpus de traductions annotés selon la typologie d'erreurs pour améliorer les méthodes d'enseignement de la traduction spécialisée.* JADT2016 Journées internationales d'Analyse statistique des Données Textuelles, Université de Nice Sophia Antipolis, CNRS, Jun 2016, Nice, France. pp.731-741.

Kübler, N., Pecman, M., & Mestivier, A. (2017). *Quand le corpus met K.O. la norme terminologique : le défi permanent des GN complexes anglais en traduction spécialisée*. 38ème colloque international du GERAS, Mar 2017, Lyon, France.

Kübler, N., Mestivier, A., & Pecman, M. (2022). Test corpora, status of the translation error, feedback on post-editing analysis and typologies of translation errors from learner's corpora. *TRALOGY III*.

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive Demand and Cognitive Effort in Post-Editing. *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, 73-84.

Lambert, P., Senellart, J., Romary, L., Schwenk, H., Zipser, F., Lopez, P., & Blain, F. (2012). *Collaborative Machine Translation Service for Scientific texts*. 11-15.

Lefer, M.-A., Piette, J., & Bodart, R. (2022). *Manuel MTPEAS: Machine Translation Post-Editing Annotation System*. *Version 1.0*.

Loffler-Laurian, A.-M. (1996). *La traduction automatique*. Presses universitaires du Septentrion. https://doi.org/10.4000/books.septentrion.74824

Maney, T., Sibert, L., Perzanowski, D., Gupta, K., & Schmidt-Nielsen, A. (2012). Toward determining the comprehensibility of machine translations. *Proceedings of the 1st PITR*, 1-7.

Martikainen, H., & Kübler, N. (2016). Ergonomie cognitive de la post-édition de traduction automatique : Enjeux pour la qualité des traductions. *ILCEA, 27*. https://doi.org/10.4000/ilcea.3863

Martikainen, H., & Mestivier, A. (2020, janvier). *Les outils de traduction nouvelle génération : Quel effet sur la qualité des textes traduits?* Journée d'études Traduction & Qualité 2020 : Biotraduction et traduction automatique, Université de Lille, France.

Minder, J. (2023). *Traduction automatique et marqueurs d'oralité : Analyse à partir d'un corpus parallèle allemand-français* [Mémoire].

Moorkens, J., Castilho, S., Gaspari, F., & Doherty, S. (Éds.). (2018). *Translation quality assessment : From principles to practice*. Springer.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU : A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Plenter, J. I. (2023). Advantages and pitfalls of machine translation for party research : The translation of party manifestos of European parties using DeepL. *Frontiers in Political Science, 5*, 1268320. https://doi.org/10.3389/fpos.2023.1268320

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Éds.), *Translation*

Quality Assessment (Vol. 1, p. 129-158). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_7

Popović, M., & Ney, H. (2007). Word Error Rates : Decomposition over POS classes and Applications for Error Analysis. *Proceedings of the Second Workshop on Statistical Machine Translation*, 48-55.

Qingyu Lu, Baopu Qiu, Ding, L., Liping Xie, & Dacheng Tao. (2023). *Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models : A Case Study on ChatGPT*. https://doi.org/10.13140/RG.2.2.17706.08647

Ragni, V., & Nunes Vieira, L. (2022). What has changed with neural machine translation? A critical review of human factors. *Perspectives*, *30*(1), 137-158. https://doi.org/10.1080/0907676X.2021.1889005

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). *COMET : A Neural Framework for MT Evaluation*. https://doi.org/10.48550/ARXIV.2009.09025

Robert, A.-M. (2010). La post-édition : L'avenir incontournable du traducteur ? *Traduire*, *222*, 137-144. https://doi.org/10.4000/traduire.460

Schmidhofer, A., & Mair, N. (2018). La traducción automática en la formación de traductores. *CLINA: Revista Interdisciplinaria de Traducción, Interpretación y Comunicación Intercultural*, *4*(2), 163. https://doi.org/10.14201/clina201842163180

Schumacher, P. (2019). Avantages et limites de la post-édition. *Revue française de la traduction*, *241*, 108-123.

Schumacher, P., & Sutera, A. (2022). Analyse comparative de post-édition et de traduction humaine en contexte académique. In C. Expósito Castro, M. del M. Ogea Pozo, & F. Rodríguez Rodríguez, *Theory and practice of translation as a vehicle for knowledge transfer* (p. 173-208). Editorial Universidad de Sevilla.

Screen, B. (2019). What effect does post-editing have on the translation product from an end-user's perspective? *Journal of Specialised Translation*, *31*, 133-157.

Shiwen, Y., & Xiaojing, B. (2015). Rule-based Machine Translation. In C. Sin-wai, *The Routledge Encyclopedia of Translation Technology* (Routledge, p. 186-200). Taylor & Francis Group.

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

Stymne, S., & Ahrenberg, L. (2012). *On the practice of error analysis for machine translation evaluation*. 1785-1790.

Swales, J. M. (1997). English as *Tyrannosaurus rex*. *World Englishes*, *16*(3), 373-382. https://doi.org/10.1111/1467-971X.00071

Taus. (2010). *Consignes relatives à la post-édition des traductions automatiques (Traduction de courtoisie Lexcelera)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Vela, M., Schumann, A.-K., & Wurm, A. (2014). Human Translation Evaluation and its Coverage by Automatic Scores. *Proceedings of MTE Workshop at LREC 2014.*

Vilar, D., Xu, J., D'Haro, L., & Ney, H. (2006). Error Analysis of Machine Translation Output. *International Conference on Language Resources and Evaluation (LREC)*, 697-702.

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in Machine Translation. *Engineering, 18*, 143-153. https://doi.org/10.1016/j.eng.2021.03.023

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). *Document-Level Machine Translation with Large Language Models*. https://doi.org/10.48550/ARXIV.2304.02210

Weiss, S., & Ahrenberg, L. (2012). Error profiling for evaluation of machine-translated text : A polish-english case study. *Proceedings of the Eighth LREC*, 1764-1770.

Wisniewski, G., Zhu, L., Ballier, N., & Yvon, F. (2021). Biais de genre dans un système de traduction automatique neuronale : Une étude préliminaire. *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*.

Wong Tak-ming, B., & Webster, J. J. (2015). Example-based Machine Translation. In C. Sin-wai, *The Routledge Encyclopedia of Translation Technology* (Routledge, p. 137-151). Taylor & Francis Group.

Yuan, W., Liu, P., & Neubig, G. (2021). BARTScore : Evaluating Generated Text as Text Generation. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.

Yulianto, A., & Supriatnaningsih, R. (2021). Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: The Asian Journal of English Language & Pedagogy, 9*(2), 109-127.

Annexe 1 – Manuel d'annotation

Manuel pour l'annotation de traductions humaines, automatiques et de post-éditions

réalisé par

MINDER Joachim

sous la direction de

KÜBLER Natalie

MESTIVIER Alexandra

en collaboration avec

BÉNARD Maud

COLIN Michele

GOUGEAUD Maëva

ZHU Lichao

Table des matières

Introduction

Ce manuel sert de guide pour l'annotation d'erreurs dans le cadre de traductions humaines, automatiques ou hybrides, c'est-à-dire des post-éditions.

Dans un premier temps, sera présenté un schéma d'annotation basé sur deux typologies d'erreurs existantes, à savoir la typologie MeLLANGE[1], qui a été créée pour l'annotation de traductions d'étudiants de master en traduction, ainsi que la typologie *Multidimensional Quality Metrics* (MQM)[2], élaborée pour l'évaluation de la qualité des traductions dans un contexte professionnel. Étant donné que l'équipe du CLILLAC-ARP avait déjà travaillé auparavant avec la typologie MeLLANGE, ce schéma d'annotation a été utilisé comme point de départ et a été par la suite enrichi de certains éléments provenant de la typologie MQM.

Par ailleurs, quelques principes fondamentaux à respecter lors de l'annotation des différents types de traduction seront présentés.

Ensuite, seront expliqués les différents attributs pouvant être ajoutés à l'annotation des types d'erreurs, les relations entre les différentes erreurs et les scores de gravité.

Enfin, la typologie d'erreurs sera expliquée en détail à l'aide d'exemples provenant de notre corpus de post-éditions dans les domaines du traitement automatique des langues (TAL) et des sciences de la terre, de l'environnement et de la planète (STEP).

[1] Castagnoli et al., 2011.

[2] https://themqm.org/the-mqm-full-typology/.

A. Présentation de la typologie d'erreurs

```
Transfert-contenu
        Omission_TR-OM
        Rajout_TR-AD
        Distorsion_TR-DI
        Indecision_TR-IN
        Type-annotateur_TR-UD
        Intrusion-langue-source
                Non-traduit-traduisible_TR-SI-UT
                Trop-litterale_TR-SI-TL
                Unites-mesure-dates-nombres_TR-SI-UN
                Type-annotateur_TR-SI-UD
        Intrusion-langue-cible
                Traduction-unites-intraduisibles_TR-TI-TD
                Trop-libre_TR-TI-TF
                Type-annotateur_TR-TI-UD
Langue
        Syntaxe_LA-SY
                Determination_LA-SY-DET
                Mauvaise-preposition_LA-SY-PR
                GNC_LA-SY-GNC
                Type-annotateur_LA-SY-UD
        Flexion-accord
                Temps-aspect_LA-IA-TA
                Genre_LA-IA-GE
                Nombre_LA-IA-NU
                Type-annotateur_LA-IA-UD
        Typographie
                Orthographie_LA-HY-SP
                Accent-diacritiques_LA-HY-AC
                Mauvaise-casse_LA-HY-CA
                Ponctuation_LA-HY-PU
                Type-annotateur_LA-HY-UD
        Registre
                Incompatible-texte-source_LA-RE-IS
                Inadapte-au-type-texte-cible_LA-RE-IT
                Type-annotateur_LA-RE-UD
        Style
                Formulation-maladroite_LA-ST-AW
                Tautologie_LA-ST-TA
                Style-titre_LA-ST-TS
                Type-annotateur_LA-ST-UD
        Reference-pas-claire_LA-UR
                Type-annotateur_LA-UD
        Conventions-textuelles
                Coherence_LA-TC-CE
                Cohesion_LA-TC-CN
                Type-annotateur_LA-TC-UD
        Terminologie-lexique
                Choix-incorrect-Termino_LA-TL-INS
                Choix-incorrect-Langue-Generale_LA-TL-ING
                Mauvais-acronyme-abreviation_LA-TL-MAA
                Faux-amis_LA-TL-FC
                Terme-traduit-par-non-terme_LA-TL-NT
                Collocation-incorrecte-Specialise_LA-TL-ICS
                Collocation-incorrecte-Langue-Generale_LA-TL-ICG
                Choix-incompatible-avec-texte-cible_LA-TL-IT
                Incoherence-terminologique
                        Differents-termes-traduction_LA-TL-TI-DT
                        Differentes-abbreviations-traduction_LA-TL-TI-DA
                Type-annotateur_TL-UD
Outils
        Hallucination_OU-TAH
        Conformite-corpus_OU-CC
        Duplication_OU-DU
        Choix-incompatible-glossaire_OU-GC
        Type-annotateur_OU-UD
```

i. Explications sur la typologie

Cette typologie d'erreurs est divisée en trois grandes catégories :

- transfert de contenu,
- langue,
- outils.

La catégorie *transfert de contenu* est liée aux erreurs de traduction, c'est-à-dire les erreurs altérant le sens du message source ou rendant son transfert et sa compréhension complexes. Cette catégorie englobe les omissions d'une partie du message, les ajouts dans la traduction, les distorsions, les indécisions, les erreurs provenant d'intrusions du texte source ainsi que les erreurs provoquées par une distance entre le texte source et la traduction.

Dans la deuxième catégorie, il ne s'agit plus d'erreurs de traduction, mais d'erreurs au niveau de la langue. On y retrouve des erreurs de syntaxe, de flexion et d'accord, de typographie, de registre, de style, des erreurs liées aux références, des erreurs de conventions textuelles ainsi que des erreurs de terminologie spécialisée et de lexique général.

Enfin, la dernière catégorie regroupe des erreurs liées aux outils ou à la maîtrise de ces outils. On y retrouve des hallucinations de la traduction automatique, des manques de conformité au corpus ou au glossaire fourni, le cas échéant, ainsi que des erreurs de duplication.

Toutes les sous-catégories seront expliquées plus en détail et illustrées à l'aide d'exemples plus loin.

Il est par ailleurs possible d'annoter un segment avec l'étiquette « Question ». Cela permet de faire remonter un point qu'on ne parvient pas à résoudre à d'autres personnes, par exemple à des spécialistes du domaine.

Différents scores de gravité peuvent être attribués à une erreur. Voici les différents scores de gravités disponibles :

- Score 0 : l'évaluateur considère qu'une meilleure traduction est possible, mais que la traduction proposée ne peut être pénalisée comme une erreur. L'« erreur » n'impacte en aucun cas la compréhension, la lisibilité ou le sens du message.
- Score 1 : l'erreur repérée a un impact très limité sur le texte cible, et celle-ci ne nuit pas à la lisibilité, à la compréhension ni à la pertinence du contenu.
- Score 2 : l'évaluateur considère que l'erreur a un impact majeur sur la traduction. Celle-ci affecte la compréhension, la lisibilité ou la pertinence du message.

- Score 3 : l'erreur de niveau 3 pose un obstacle à l'utilisation du texte. Par exemple, il y a une perte de sens ou une distorsion grave. Si une erreur majeure apparaît à un endroit important du texte (par exemple dans le titre), cela est un facteur pour la considérer comme une erreur de niveau 3.

Pour l'attribution du score de gravité, il convient naturellement de prendre en compte le genre textuel et la finalité du texte. Par exemple, une erreur de formulation maladroite n'aura pas le même poids dans un article scientifique que dans un texte littéraire.

Enfin, il convient d'ajouter aux annotations d'erreurs les attributs qualifiant les relations entre les différentes annotations. Ceux-ci ne sont à utiliser que lorsqu'on annote des post-éditions. Voici les différents attributs :

TA_Correct	Value:TACorBienCorr|TACorMalCorr|TACorNonCor
TA_Erronee	Value:TAEBienCor|TAEMalCor|TAENonCor

Leur utilisation est détaillée au point suivant.

ii. Les principes généraux

Le principe 1 est applicable aux différents types d'annotations (traductions humaines, automatiques ou post-éditions).

Principe 1 : choisir le niveau de granularité le plus précis

Lorsqu'on annote une traduction ou une post-édition, il ne faut pas choisir les grandes catégories comme « transfert de contenu », « syntaxe » ou encore « terminologie ». Il faut opter pour les sous-catégories les plus précises. Par exemple, si l'on observe une phrase au style peu naturel, il ne faut pas choisir la catégorie d'erreur « style ». Il est préférable de choisir une sous-catégorie pertinente, comme « formulation maladroite ». Dans certains rares cas, il est possible qu'une erreur ne puisse pas être annotée avec un niveau avancé de granularité. Uniquement dans ces cas, il est envisageable de sélectionner une catégorie générale.

Principe 2 : superposer les couches d'annotation

Il est tout à fait envisageable d'annoter un segment avec plusieurs étiquettes différentes. Par exemple, les distorsions sont souvent causées par d'autres erreurs sous-jacentes (exemple sur la Figure 1).

Sur la ligne 32 de la Figure 1, on constate que l'utilisation de parenthèses plutôt que l'utilisation de tirets cadratin (erreur de typographie) engendre une distorsion. Dans ce cas, il est nécessaire d'annoter les deux erreurs (ou plus).

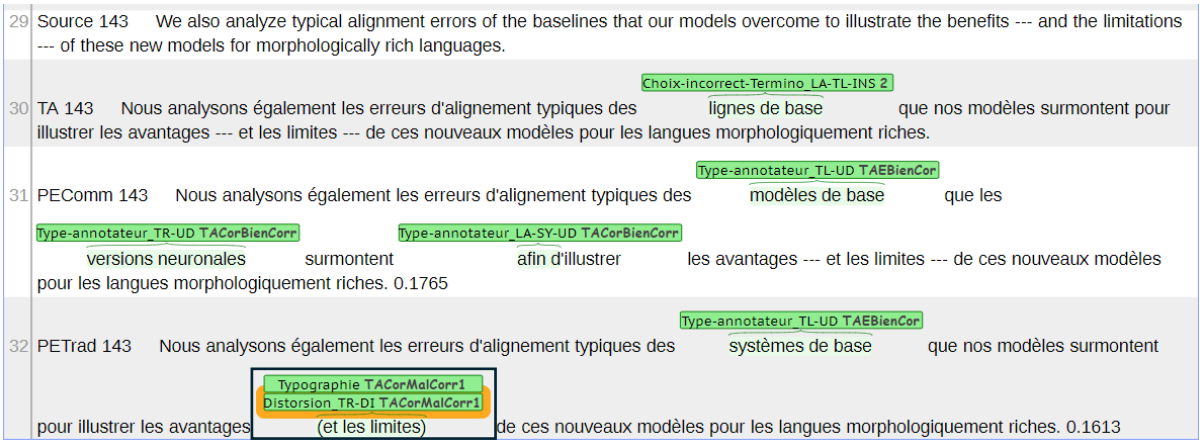

Figure 1 : superposition de couches d'annotation

Sur la ligne 32 de la Figure 1, l'utilisation de parenthèses plutôt que les tirets cadratin (erreur de typographie) engendre une distorsion. Par conséquent, il est nécessaire d'annoter les deux erreurs et de leur attribuer le même poids.

Les principes exposés ci-dessous s'appliquent uniquement lors de l'annotation de post-éditions. En effet, lorsqu'on annote des données avec un seul texte cible (c'est-à-dire une traduction humaine ou une traduction automatique), il convient simplement d'annoter l'erreur et d'ajouter le score de gravité. En revanche, pour l'annotation de post-éditions accompagnées de leurs traductions automatiques, il est important de suivre les principes suivants.

Principe 3 : utilisation de l'étiquette « Type annotateur »

Au point B, on remarque que dans chaque catégorie d'erreur, on retrouve l'étiquette *type-annotateur*. Cette étiquette est à utiliser lorsqu'on annote un segment qui ne contient pas d'erreur (et, dès lors, aucun score de gravité) dans les cas suivants :

- La traduction automatique comporte une erreur qui a été corrigée dans la post-édition. Dans ce cas, le segment de la post-édition correct correspondant au segment erroné dans la traduction automatique sera annoté avec l'étiquette *type-annotateur* de la catégorie d'erreur correspondante. Voici un exemple :

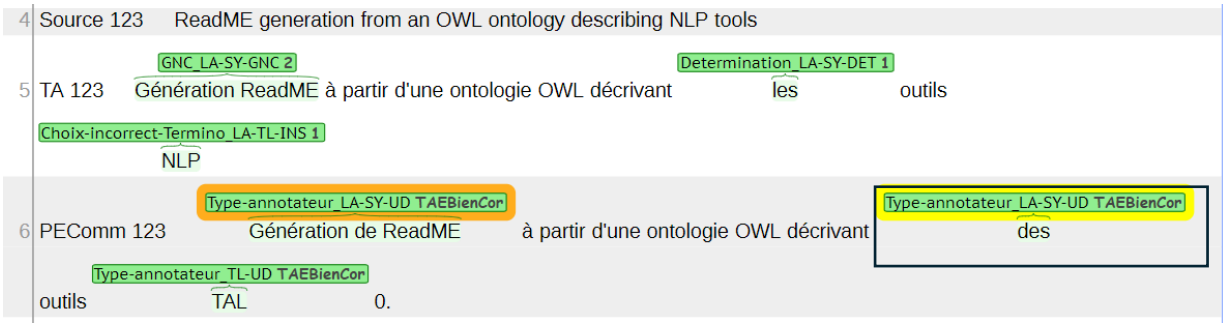

Figure 2 : type-annotateur — cas d'utilisation 1

Sur la Figure 2, le déterminant « les » (ligne 5, traduction automatique) est erroné. Il comporte donc l'étiquette d'erreur « Détermination » (LA-SY-DET). Toutefois, cette erreur a été corrigée dans la post-édition (ligne 6). Dès lors, comme il ne s'agit plus d'une erreur, mais d'une bonne solution, on utilise l'étiquette *type-annotateur* de la même catégorie (LA-SY-UD).

- La traduction automatique ne comporte aucune erreur, mais on observe une variante (correcte également) dans la post-édition. Dans ce cas, on n'annote pas le segment dans la traduction automatique, mais uniquement la variante dans la post-édition.

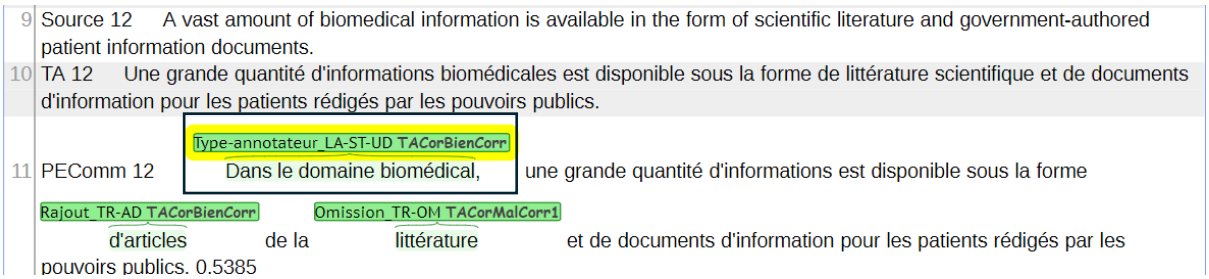

Figure 3 : type-annotateur — cas d'utilisation 2

Sur la Figure 3, on observe que l'adjectif « biomédicales » dans la traduction automatique (ligne 10), qui est correct, est devenu un groupe prépositionnel dans la post-édition (ligne 11). Il s'agit d'une amélioration, mais la traduction automatique ne comporte pas d'erreur. Dans ce cas, il convient d'annoter la variante proposée dans la post-édition et d'utiliser l'étiquette *type-annotateur* de la catégorie pertinente (ici, style, LA-ST-UD).

Principe 4 : utilisation des attributs

Comme expliqué brièvement au point précédent, il existe au total 5 attributs, divisés en 2 catégories différentes. Ces attributs ne peuvent être utilisés que dans la post-édition et sont ajoutés aux erreurs, aux variantes ou aux corrections annotées.

TA_Correct	Value:TACorBienCorr|TACorMalCorr
TA_Erronee	Value:TAEBienCor|TAEMalCor|TAENonCor

La première catégorie comporte les attributs à utiliser lorsque la traduction automatique (TA) est correcte :

- TA correcte bien corrigée (TACorBienCorr) : on utilise cet attribut *dans la post-édition uniquement* lorsqu'une variante est introduite dans la post-édition, mais qu'il n'y a aucune erreur dans la traduction automatique (voir par exemple Figure 3 ci-dessus) ;

- TA correcte mal corrigée (TACorMalCorr) : cet attribut est à utiliser lorsqu'il n'y a aucune erreur dans la TA, mais qu'une erreur est introduite à cet endroit dans la post-édition ;

La seconde catégorie comprend les attributs qu'il convient d'utiliser lorsque la traduction automatique comporte une erreur :

- TA erronée bien corrigée : il convient d'utiliser cet attribut *dans la post-édition* lorsqu'il y a une erreur dans la TA, mais que celle-ci a été corrigée dans la post-édition. Voici un exemple d'utilisation de cet attribut :

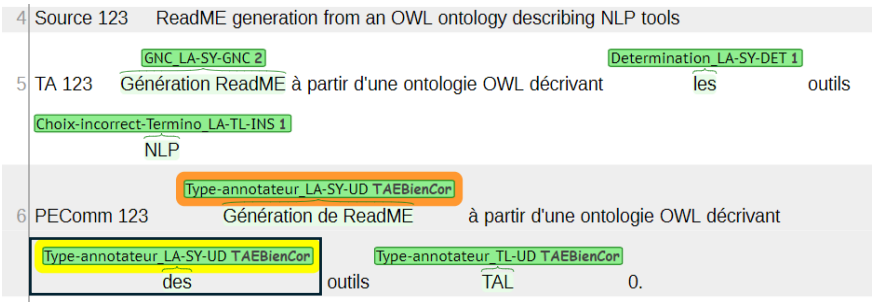

Figure 4 : attribut « TA erronée bien corrigée »

Sur la Figure 4, on observe que l'erreur dans la TA (détermination, ligne 5) a été corrigée dans la post-édition (type-annotateur, ligne 6). Par conséquent, lorsqu'on ajoute l'annotation « type-annotateur », il convient d'ajouter l'attribut TAEBienCor (TA erronée bien corrigée).

- TA erronée mal corrigée : il convient d'utiliser cet attribut *dans la post-édition* lorsqu'il y a une erreur dans la TA, mais que celle-ci a été mal corrigée dans la post-édition. Dans ce cas, il reste une erreur dans la post-édition, mais ce n'est plus la même que dans la TA. Voici un exemple d'utilisation de cet attribut :

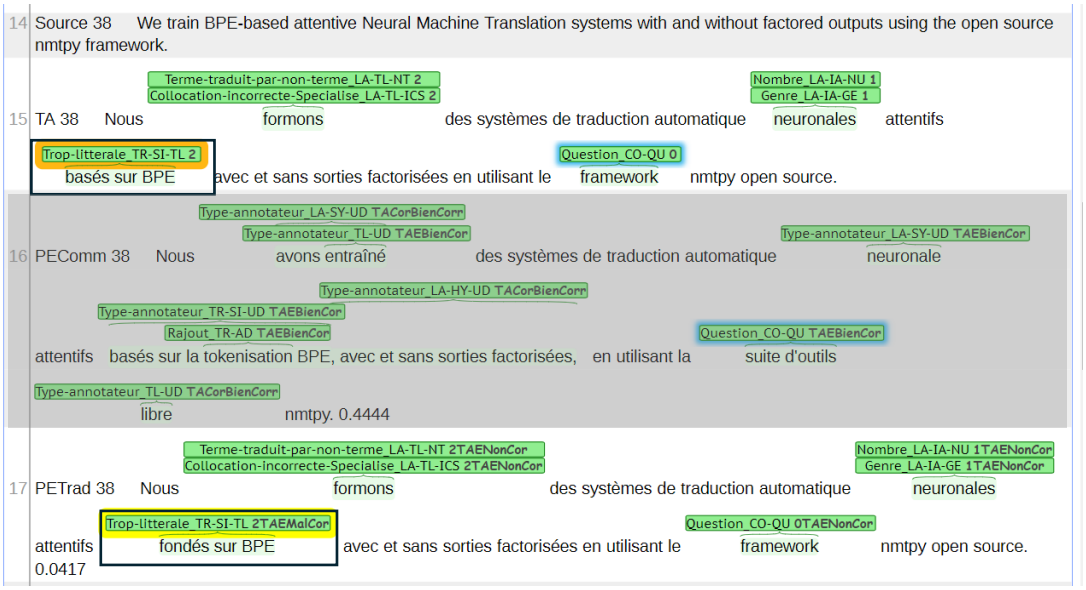

Figure 5 : attribut "TA erronée mal corrigée"

Quand on compare l'erreur « trop littérale » à la ligne 15 et l'erreur « trop littérale » à la ligne 17, on remarque qu'il y a une légère modification. Toutefois, la post-édition reste erronée. Dans ce cas, on utilise l'attribut « TAEMalCor » (TA erronée mal corrigée), et on conserve le type d'erreur (ici, trop littérale), et non l'attribut *type-annotateur*, puisqu'il y a toujours une erreur.

- TA erronée non corrigée : il convient d'utiliser cet attribut *dans la post-édition* lorsqu'il y a une erreur dans la TA, mais que celle-ci n'a pas été corrigée du tout dans la post-édition. Dans ce cas, on utilise la même étiquette d'erreur avec le même score de gravité que dans la traduction automatique.

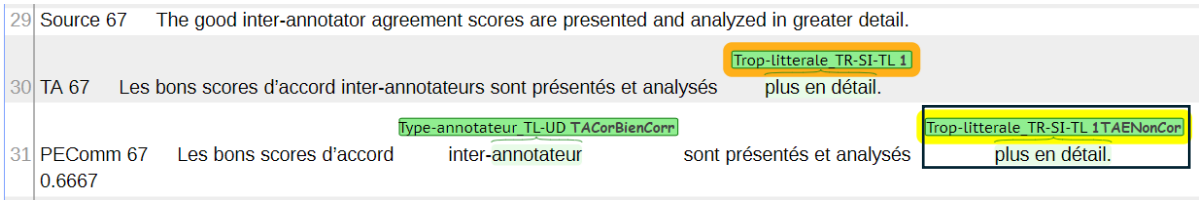

Figure 6 : attribut « TA erronée non corrigée »

Sur la Figure 6, on remarque que l'erreur « trop littérale » dans la TA (ligne 30) n'a pas du tout été corrigée dans la post-édition (ligne 31). Dès lors, on conserva la même étiquette d'erreur avec le même score de gravité que dans la traduction automatique (ici, trop littérale, score de gravité 1).

Principe 5 : quand et comment utiliser les scores de gravité ?

Par définition, les scores de gravité ne peuvent être utilisés *que* lorsqu'il y a une erreur. Par conséquent, lorsqu'il y a une étiquette *type-annotateur*, on ne peut pas utiliser de score de gravité.

- TA erronée bien corrigée : imaginons une TA erronée (score de gravité 2) avec une post-édition bien corrigée *type-annotateur*. Dans un tel cas, il n'y aura pas de score de gravité dans la post-édition, puisqu'il n'y a plus d'erreur.
- TA erronée non corrigée : si l'erreur présente dans la TA n'a pas été corrigée du tout dans la post-édition, alors le même score d'erreur sera conservé dans la post-édition.
- TA erronée mal corrigée : si l'erreur présente dans la TA a été mal corrigée (c'est-à-dire qu'il reste une erreur dans la post-édition, mais que ce n'est pas la même que dans la TA), alors le score de gravité peut être différent.
- TA correcte bien corrigée : si on introduit une variante dans la post-édition (*type-annotateur*), il n'y aura pas de score de gravité, car il n'y a pas d'erreur.

Principe 6 : erreur causée par le texte source

Il se peut qu'une erreur survenant dans la traduction automatique soit causée par le texte source (voir Figure 7 ci-dessous).

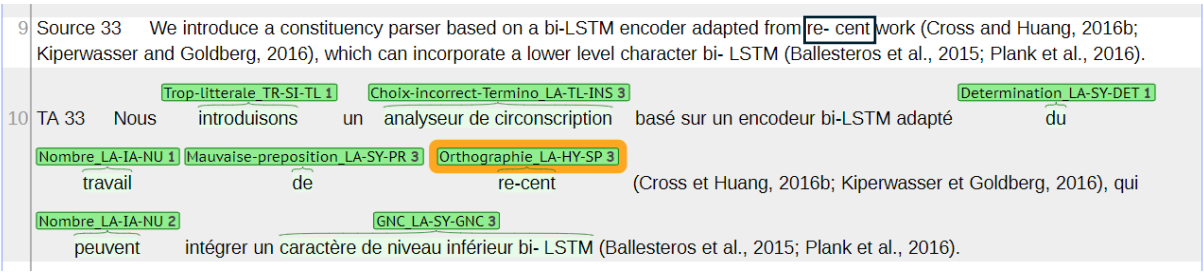

Figure 7 : erreur causée par le texte source

Ici, on remarque effectivement que l'erreur présente dans la TA (ligne 10) provient du mauvais découpage du texte source (ligne 9). Dans un tel cas, il convient de ne pas tenir compte du fait que l'erreur provient du texte source et d'évaluer l'erreur comme dans les autres cas. En outre, il ne faut **jamais annoter le texte source**.

Principe 7 : proposer une solution lorsque l'erreur n'est pas corrigée

Lorsqu'une erreur présente dans la traduction automatique n'est pas corrigée dans la post-édition (TA erronée non corrigée ou TA erronée mal corrigée), il convient de proposer une solution dans la section « Notes » lorsqu'on annote l'erreur (voir Figure 8 ci-dessous).

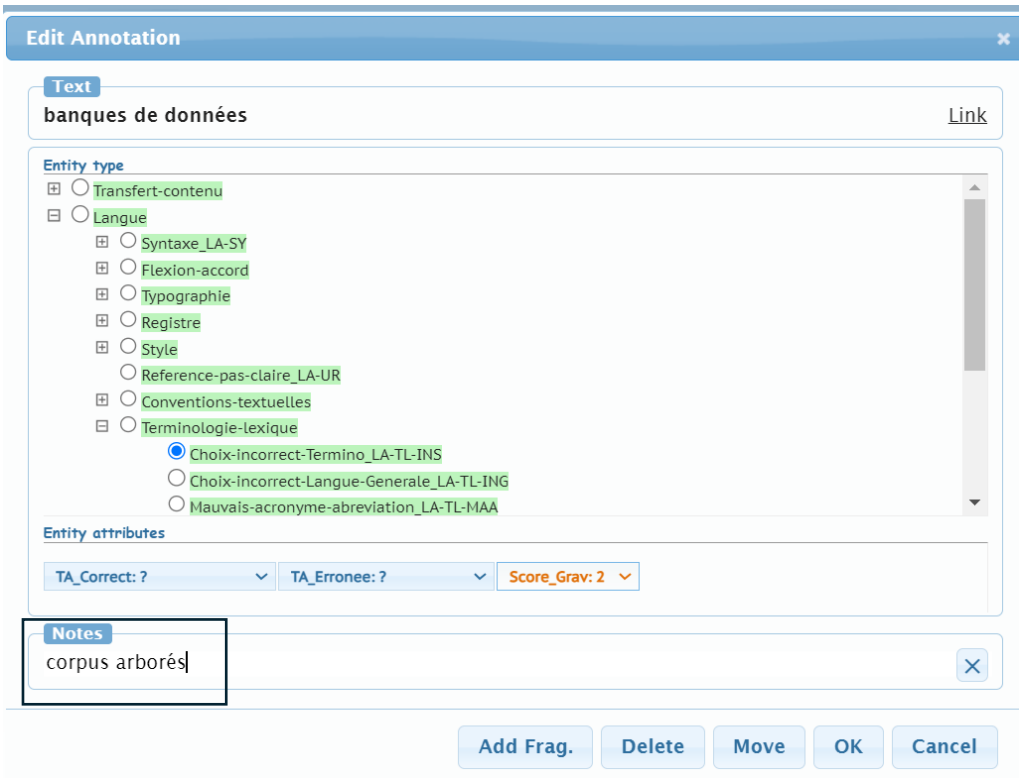

Figure 8 : proposer une solution

La fenêtre qui s'ouvre lors de la sélection d'un segment pour l'annoter permet d'ajouter un commentaire dans la section « Notes ». Cette case permet donc à l'annotateur d'ajouter la solution si l'erreur n'a pas été corrigée.

B. Les différentes erreurs en détail

1. Transfert-contenu

1.1. Omission_TR-OM

Une omission se produit lorsqu'il manque, dans la traduction, une idée qui est présente dans le texte source. Il ne faut pas confondre omission et implicitation. Une omission a lieu sans réelle raison valable, alors qu'une implicitation est un moyen d'éviter une surtraduction[3]. Cependant, nous ne faisons pas cette distinction dans notre typologie d'erreur. Ainsi, une implicitation n'est pas une erreur, il convient d'utiliser la catégorie « Omission » pour les implicitations, et d'utiliser un score de gravité de niveau 0. Pour les omissions, le score de gravité dépend de l'effet de celles-ci.

Source	***Despite of the services offered by the user-friendliness of the web site […]***
TA	*Malgré les services offerts par la convivialité du site web […]*
PE	*En dépit de la convivialité du site web […]*
ERREUR	Il manque la notion de « services ». En revanche, celle-ci est superflue et n'ajoute rien à la traduction. On peut dès lors considérer qu'il s'agit d'une implicitation.
ATTRIBUT	TA correcte bien corrigée → type-annotateur
SCORE	Aucun (puisque pas d'erreur)

Figure 9 : exemple d'omission (0)

Source	***We model two important interfaces of constituency parsing with auxiliary tasks.***
TA	*Nous modélisons deux aspects importants de l'analyse des circonscriptions avec des tâches auxiliaires.*
PE	*Nous modélisons deux aspects importants de l'analyse syntaxique avec des tâches auxiliaires.*
ERREUR	Omission de la notion de « constituency ». Ce n'est pas n'importe quelle analyse syntaxique, c'est une analyse syntaxique en constituants.
ATTRIBUT	TA erronée (*circonscriptions* est faux) mal corrigée
SCORE	1 : une nuance est perdue, mais l'erreur n'empêche pas la compréhension/lisibilité.

Figure 10 : exemple d'omission (1)

[3] Delisle 2003, p. 51.

1.2. Rajout_TR-AD

À l'instar de la différence entre *omission* et *implicitation*, on peut souligner une différence de nuance entre le rajout et l'explicitation. L'ajout est considéré comme une erreur, alors que l'explicitation peut s'expliquer par le fait que le traducteur ou le post-éditeur souhaite éviter la sous-traduction, auquel cas il convient d'utiliser l'étiquette *type-annotateur*.

Source	***One of the problem is being able to deal with multilingual generation of texts.***
TA	*L'un des problèmes est de pouvoir traiter la génération multilingue de textes.*
PE	*L'un des problèmes principaux est de pouvoir traiter la génération multilingue de textes.*
ERREUR	Le post-éditeur a ajouté la notion de « principaux », qui introduit une légère distorsion. Il convient dès lors d'ajouter également une étiquette *distorsion* avec le même score de gravité.
ATTRIBUT	TA correcte mal corrigée
SCORE	1 : l'impact est très limité sur le texte cible, mais légère distorsion.

Figure 11 : exemple de rajout (1)

Source	***This paper describes LIUM submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs.***
TA	*Cet article décrit les soumissions de LIUM à WMT17 News Translation Task pour l'anglais, l'allemand, l'anglais, l'anglais, le tchèque et l'anglais→langue latine.*
PE	*Cet article décrit les contributions du LIUM à la tâche de traduction d'articles de presse de* ***la conférence*** *WMT17 pour les paires de langues anglais↔allemand, anglais↔turc, anglais→tchèque et anglais→letton.*
ERREUR	Pas une erreur, mais un ajout justifié (il s'agit bien d'une conférence, c'est plus clair dans la post-édition).
ATTRIBUT	TA correcte bien corrigée
SCORE	Aucun, puisqu'il n'y a pas d'erreur (*type-annotateur*).

Figure 12 : exemple de rajout (0)

1.3. Distorsion_TR-DI

La distorsion est une déformation du sens du message source. Elle est, en principe, causée par une autre erreur. Dès lors, lorsqu'il y a une distorsion, il y a très souvent une autre étiquette qui l'accompagne, sauf si on ne parvient pas à détecter l'origine de la distorsion.

Source	***We also analyze typical alignment errors of the baselines that our models overcome to illustrate the benefits --- and the limitations --- of these new models for morphologically rich languages.***
TA	*Nous analysons également les erreurs d'alignement typiques des lignes de base que nos modèles surmontent pour illustrer les avantages --- et les limites --- de ces nouveaux modèles pour les langues morphologiquement riches.*
PE	*Nous analysons également les erreurs d'alignement typiques des systèmes de base que nos modèles surmontent pour illustrer les avantages (et les limites) de ces nouveaux modèles pour les langues morphologiquement riches.*
ERREUR	Erreur de typographie (parenthèses au lieu des tirets cadratin) causant une distorsion : l'incise sert à mettre en évidence, à l'inverse des parenthèses.
ATTRIBUT	TA correcte mal corrigée
SCORE	1 : légère nuance, mais la lisibilité n'est pas affectée et le sens est presque le même.

Figure 13 : exemple de distorsion (1)

Source	***A recent, lightweight approach, instead augments a baseline model with supplementary (small) adapter layers, keeping the rest of the model unchanged.***
TA	*Une approche récente et légère augmente plutôt un modèle de base avec des couches d'adaptateur supplémentaires (petites), gardant le reste du modèle inchangé.*
PE	*Une approche récente et moins coûteuse augmente plutôt un modèle de base avec des (petites) couches d'adaptateurs supplémentaires, gardant le reste du modèle inchangé.*
ERREUR	Erreur de terminologie (choix incorrect langue générale) causant une distorsion
ATTRIBUT	TA erronée (pas claire) mal corrigée
SCORE	2 : le sens n'est pas le même.

Figure 14 : exemple de distorsion (2)

1.4. Indecision_TR-IN

On considère qu'il y a une indécision lorsque le traducteur ou le post-éditeur propose plusieurs traductions possibles ou, dans le cas de post-éditions, il reste des traces de post-édition (par exemple, le post-éditeur a mal modifié la TA et il en reste des traces qui perturbent le texte cible). En voici un exemple :

Source	***Span-based discontinuous constituency parsing: a family of exact chart-based algorithms with time complexities from O(n^6) down to O(n^3)***
TA	*Analyse de constituants discontinus basée sur l'étendue : une famille d'algorithmes exacts basés sur des diagrammes avec des complexités temporelles de O(n^6) à O(n^3)*
PE	*Analyse de constituants discontinus basée sur les empans : une famille d'algorithmes tabulaires exacts basés avec des complexités temporelles allant de O(n^6) à O(n^3) 0.1026*
ERREUR	Ici, le post-éditeur a transformé « basés sur des diagrammes » par l'adjectif « tabulaires », mais il a omis de supprimer l'adjectif verbal « basés ». Par conséquent, la phrase n'a pas une syntaxe correcte.
ATTRIBUT	TA erronée mal corrigée
SCORE	2 : la lisibilité est affectée, et l'erreur se produit dans le titre (facteur aggravant).

Figure 15 : exemple d'indécision (2)

Voici un exemple de ce qu'on entend, en principe, par « indécision ».

Source	***After the intraoceanic subduction is initiated, the subduction of the left (Indian) plate dominates the system, which helps the plates remain attached to each other, and rapidly close the ocean basin.***
TRAD	*Après l'initiation de la subduction intra-océanique, la subduction de/au niveau de la plaque de gauche, soit la plaque indienne, domine le système. Cette situation aidera les plaques à rester collées l'une à l'autre et à fermer rapidement le bassin océanique.*
ERREUR	Ici, le traducteur a laissé deux propositions différentes.
ATTRIBUT	Aucun, car il n'y a pas de traduction automatique (traduction humaine).
SCORE	1 : la lisibilité est légèrement affectée, mais le sens reste correct.

Figure 16 : exemple d'indécision (1)

1.5. Type-annotateur_TR-UD

Il convient d'utiliser cet attribut dans la post-édition lorsqu'une des quatre erreurs ci-dessus (omission, rajout, distorsion ou indécision) dans la TA est bien corrigée dans la post-édition.

Source	***This paper describes LIUM submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs.***
TA	*Cet article décrit les soumissions de LIUM à WMT17 News Translation Task pour l'anglais, l'allemand, l'anglais, l'anglais, le tchèque et l'anglais→langue latine.*
PE	*Cet article décrit les contributions du LIUM à la tâche de traduction d'articles de presse de la conférence WMT17 pour les paires de langues anglais↔allemand, anglais↔turc, anglais→tchèque et anglais→letton.*
ERREUR	L'erreur de distorsion sur tout le segment surligné dans la TA a été corrigée dans la PE.
ATTRIBUT	TA erronée bien corrigée
SCORE	Aucun, puisque l'erreur est corrigée (type-annotateur).

Figure 17 : exemple type-annotateur

1.6. Intrusion-langue-source

Il convient, si possible, d'éviter d'annoter une erreur avec cette étiquette, puisqu'il s'agit d'une catégorie comprenant des sous-catégories (ci-dessous). Par conséquent, il faut privilégier l'utilisation des sous-catégories présentées ci-dessous.

1.6.1. Non-traduit-traduisible_TR-SI-UT

Cette catégorie couvre les erreurs causées par des mots, des syntagmes ou des segments n'étant pas traduits, alors qu'une traduction est possible dans la langue cible.

Source	***We also experiment with pre-trained word embeddings and Bertbased neural networks.***
TA	*Nous expérimentons également avec des word embeddings pré-entraînés et des réseaux neuronaux basés sur Bert.*
PE	*Nous expérimentons également avec des plongements lexicaux pré-entraînés et des réseaux neuronaux fondés sur Bert.*
ERREUR	Dans la TA, on retrouve le terme anglais *word embeddings*. Bien que l'on retrouve ce terme en anglais dans le corpus, la traduction française (que l'on retrouve d'ailleurs correctement dans la PE) est bien plus fréquente. Dès lors, l'erreur « non-traduit-traduisible » est à annoter dans la TA.
ATTRIBUT	TA erronée bien corrigée (type-annotateur) → type-annotateur
SCORE	Aucun, puisque l'erreur est corrigée (type-annotateur).

Figure 18 : exemple "non traduit traduisible" (1)

Source	***Multilingual Lexicalized Constituency Parsing with Word-Level Auxiliary Tasks***
TA	*Lexicalized Constituency Parsing multilingue avec des tâches auxiliaires de niveau Word*
ERREUR	Les deux parties surlignées ne sont pas traduites. Or, une traduction serait une meilleure solution.
ATTRIBUT	Aucun, puisque c'est la TA.
SCORE	2 pour chaque erreur, puisque cela gêne la lisibilité et l'erreur se produit dans le titre.

Figure 19 : exemple de "non traduit traduisible" (2)

1.6.2. Trop-litterale_TR-SI-TL

Une traduction est considérée trop littérale lorsqu'elle n'est pas idiomatique dans la langue cible et que ce manque de naturel est dû à une influence de la langue source.

Source	***Our approach uses automatically generated pairs of source sentences, where each pair tests one morphological contrast.***
TA	*Notre approche utilise des paires de phrases sources générées automatiquement, où chaque paire teste un contraste morphologique.*
PE	*Notre approche utilise des paires de phrases sources générées automatiquement, où chaque paire teste un contraste morphologique.*
ERREUR	L'anthropomorphisme est critiqué en français, il convient, par exemple, de dire « chaque paire permet de tester », puisque ce n'est pas la paire qui teste.
ATTRIBUT	TA erronée non corrigée
SCORE	1, cela reste compréhensible, mais manque d'idiomaticité.

Figure 20 : exemple de traduction trop littérale (1)

Source	***We train BPE-based attentive Neural Machine Translation systems with and without factored outputs using the open source nmtpy framework.***
TA	*Nous formons des systèmes de traduction automatique neuronales attentifs basés sur BPE avec et sans sorties factorisées en utilisant le framework nmtpy open source.*
PE	*Nous formons des systèmes de traduction automatique neuronales attentifs fondés sur BPE avec et sans sorties factorisées en utilisant le framework nmtpy open source.*
ERREUR	On ne comprend pas ce qu'est le « BPE » (un type de tokenisation).
ATTRIBUT	TA erronée mal corrigée
SCORE	2, les traductions manquent de clarté.

Figure 21 : exemple de traduction trop littérale (2)

1.6.3. *Unites-mesure-dates-nombres_TR-SI-UN*

Cette catégorie regroupe les erreurs liées au format, au transfert ou à la mauvaise retranscription des unités de mesure, des dates, des chiffres ou des nombres.

Source	***This contribution presents the discovery of ~3,700-Myr-old structures (Fig. 1) interpreted as stromatolites in an ISB outcrop of dolomitic rocks, newly exposed by melting of a perennial snow patch.***
Traduction	*Cet article présente la découverte de structures vieilles de ~3,700 millions d'années (Fig.1) identifiées comme des stromatolithes dans un affleurement de roches dolomitiques de la CSI, récemment exposées grâce à la fonte d'une couche de neige éternelle.*
ERREUR	Le nombre est calqué sur le format anglais. Il s'agit de 3 700, et non de 3,700 (format incorrect pour les milliers).
ATTRIBUT	C'est une traduction humaine, donc aucun attribut n'est nécessaire.
SCORE	1 : le format n'est pas correct, mais l'erreur ne peut pas être considérée comme grave.

Figure 22 : exemple d'erreur liée aux unités de mesure, aux dates et aux nombres (1)

1.6.4. *Type-annotateur_TR-SI-UD*

Il convient d'utiliser cet attribut dans la post-édition lorsqu'une des trois erreurs d'intrusion de la langue source ci-dessus (non traduit traduisible, trop littérale, unités de mesure/dates/nombres) dans la TA est bien corrigée dans la post-édition.

Source	***Multilingual Lexicalized Constituency Parsing with Word-Level Auxiliary Tasks***
TA	*Lexicalized Constituency Parsing multilingue avec des tâches auxiliaires de niveau Word*
PE	*Tâches auxiliaires au niveau des mots pour l'analyse syntaxique en constituants lexicalisés multilingue*
ERREUR	Deux erreurs de non-traduction dans la traduction automatique, qui ont été corrigées dans la post-édition.
ATTRIBUT	TA erronée bien corrigée
SCORE	Aucun

Figure 23 : exemple type-annotateur

1.7. *Intrusion-langue-cible*

Il convient d'éviter autant que possible d'annoter une erreur avec cette étiquette, puisqu'il s'agit d'une catégorie comprenant des sous-catégories (ci-dessous). Par conséquent, il faut privilégier l'utilisation des sous-catégories exposées ci-dessous.

1.7.1. Traduction-unites-intraduisibles_TR-TI-TD

Cette erreur s'utilise lorsqu'un élément est traduit dans la langue cible, alors qu'il convient de ne pas le traduire.

Source	***Machine Translation, it's a question of style, innit? The case of English tag questions***
TA	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des questions de tag en anglais*
PE	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des questions à étiquette en anglais*
ERREUR	Le terme anglais *tag questions* ne se traduit pas en français.
ATTRIBUT	TA erronée mal corrigée
SCORE	2 (pour la TA et la PE), car cela n'existe pas en français.

Figure 24 : exemple "traduction unités intraduisibles" (2)

Source	***YASET provides state-of-the-art performance on the CoNLL 2003 NER dataset […] and NCBI disease corpus (F1=0.81).***
TA	*YASET fournit des performances de pointe sur l'ensemble de données NER CoNLL 2003 […] et le corpus de maladies NCBI (F1=0,81).*
ERREUR	Ce corpus n'est pas traduit en français (corpus anglais).
ATTRIBUT	Aucun, puisque c'est la TA.
SCORE	2, car on ne retrouve pas ce corpus avec ce titre-là.

Figure 25 : 2e exemple "traduction unités intraduisibles" (2)

1.7.2. Trop-libre_TR-TI-TF

Une traduction trop libre est un traduction dont le sens diffère trop de celui transféré par le texte source, engendrant dès lors souvent une distorsion. En voici un exemple.

Source	***The rapid plate motion of India toward Eurasia remains a major tectonic puzzle.***
TA	*Le mouvement rapide des plaques de l'Inde vers l'Eurasie reste un casse-tête tectonique majeur.*
PE	*Le rapprochement rapide de la plaque indienne vers la plaque eurasiatique reste un phénomène tectonique inexpliqué.*
ERREUR	Le texte source n'évoque pas le fait qu'il s'agit d'un phénomène inexpliqué. Il s'agit d'une surtraduction.
ATTRIBUT	TA erronée mal corrigée
SCORE	2 : le sens diffère.

Figure 26 : exemple de traduction trop libre (2)

1.7.3. Type-annotateur_TR-TI-UD

Cette sous-catégorie de type-annotateur est à utiliser dans la post-édition lorsqu'une erreur de traduction d'unité intraduisible ou de traduction trop libre dans la TA et été corrigée dans la post-édition.

Source	***Machine Translation, it's a question of style, innit? The case of English tag questions***
TA	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des questions de tag en anglais*
PE	*La traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des « tag questions » en anglais*
ERREUR	Le terme anglais *tag questions* ne se traduit pas en français. L'erreur a été corrigée dans la post-édition.
ATTRIBUT	TA erronée bien corrigée
SCORE	Aucun

Figure 27 : exemple type-annotateur

2. Langue

2.1. Syntaxe_LA-SY

Il convient d'éviter autant que possible d'annoter une erreur avec cette étiquette générique, puisqu'il s'agit d'une catégorie comprenant des sous-catégories (ci-dessous). Par conséquent, il faut privilégier l'utilisation des sous-catégories exposées ci-dessous. Si une erreur de syntaxe ne peut pas être considérée comme une erreur de détermination, de préposition ou de groupe nominal complexe, alors il est possible d'utiliser l'étiquette générique « Syntaxe ».

2.1.1. Determination_LA-SY-DET

Cette catégorie regroupe les erreurs liées à l'utilisation, à la mauvaise utilisation ou à la non-utilisation de déterminants.

Source	***Machine Translation, it's a question of style, innit? The case of English tag questions***
TA	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des questions de tag en anglais*
PE	*Traduction automatique, c'est une question de style, n'est-ce pas ? Le cas des questions à étiquette en anglais*
ERREUR	Comme il s'agit d'une phrase complète, cette proposition serait plus naturelle avec le déterminant *la.*
ATTRIBUT	TA erronée non corrigée
SCORE	1 : l'erreur affecte légèrement la lisibilité et l'idiomaticité.

Figure 28 : exemple d'erreur de détermination (1)

Source	***LIUM Machine Translation Systems for WMT17 News Translation Task***
TA	*Systèmes de traduction automatique Lium pour WMT17 Nouvelles Tâche de Traduction*
PE	*Systèmes de traduction automatique LIUM pour WMT17 News Translation Task*
ERREUR	Le LIUM est un laboratoire. Ici, on peut avoir l'impression qu'il s'agit du nom des systèmes de traduction.
ATTRIBUT	TA erronée mal corrigée
SCORE	2 : l'erreur modifie le sens de la phrase.

Figure 29 : exemple d'erreur de détermination (2)

2.1.2. *Mauvaise-preposition_LA-SY-PR*

Cette catégorie s'applique aux erreurs de préposition.

Source	***The micro-syntactic annotation process, presented in this paper, includes a semi-automatic preparation of the transcription, the application of a syntactic dependency parser, transcoding of the parsing results to the Rhapsodie annotation scheme, manual correction by multiple annotators followed by a validation process, and finally the application of coherence rules that check common errors.***
TA	*Le processus d'annotation micro-syntactique, présenté dans cet article, comprend une préparation semi-automatique de la transcription, l'application d'un analyseur de dépendance syntaxique, le transcodage des résultats d'analyse au schéma d'annotation Rhapsodie, la correction manuelle par plusieurs annotateurs suivie d'un processus de validation, et enfin l'application de règles de cohérence qui vérifient les erreurs courantes.*
PE	*Le processus d'annotation micro-syntaxique, présenté dans cet article, comprend une préparation semi-automatique de la transcription, l'application d'un analyseur en dépendance syntaxique, le transcodage des résultats de l'analyse syntaxique au schéma d'annotation Rhapsodie, la correction manuelle par plusieurs annotateurs suivie d'un processus de validation, et enfin l'application de règles de cohérence qui vérifient les erreurs courantes.*
ERREUR	Erreur de préposition dans la traduction automatique. Cela peut aussi être une erreur de choix terminologique incorrect, si l'on observe le problème d'un point de vue terminologique. Les deux types d'erreurs peuvent donc être annotés.
ATTRIBUT	TA erronée bien corrigée
SCORE	1 : le terme est mal traduit à cause de cette préposition, mais il reste totalement compréhensible.

Figure 30 : exemple de mauvaise préposition (1)

Source	***Evaluation of a Sequence Tagging Tool for Biomedical Texts***
TA	*Évaluation d'un outil de marquage de séquences pour les textes biomédicaux*
PE	*Évaluation d'un outil de étiquetage de séquences pour les textes biomédicaux*
ERREUR	Le terme *marquage* (erronée) est bien corrigé dans la post-édition, mais le post-éditeur ajoute une erreur d'élision, considérée comme une erreur de préposition.
ATTRIBUT	TA erronée mal corrigée
SCORE	2 : l'erreur affecte uniquement la lisibilité, mais pas le sens. En revanche, comme elle apparaît à un endroit stratégique (dans le titre), l'erreur est considérée comme majeure.

Figure 31 : exemple de mauvaise préposition (2)

2.1.3. GNC_LA-SY-GNC

Cette catégorie concerne les erreurs liées au traitement des groupes nominaux complexes. Il peut s'agir, entre autres, d'une mauvaise identification de la tête du groupe nominal ou encore d'une mauvaise factorisation des différents éléments du groupe nominal complexe.

Source	***We present (i) the automatic annotation of English TQs in a parallel corpus of subtitles and (ii) an approach using a series of classifiers to predict TQ forms, which we use to post-edit state-of-the-art MT outputs.***
TA	*Nous présentons (i) l'annotation automatique des QT anglais dans un corpus parallèle de sous-titres et (ii) une approche utilisant une série de classificateurs pour prédire les formes de QT, que nous utilisons pour post-éditer les résultats de traduction automatique les plus récents.*
PE	*Nous présentons (i) l'annotation automatique des QT anglais dans un corpus parallèle de sous-titres et (ii) une approche utilisant une série de classificateurs pour prédire les formes de QT, que nous utilisons pour post-éditer les résultats de traduction automatique les plus récents.*
ERREUR	*State-of-the-art* est censé être relié à *MT*, et non à *outputs*. Or, dans la TA et dans la PE, l'adjectif est relié *résultats.* On devrait parler des *résultats des systèmes de TA à l'état de l'art.*
ATTRIBUT	TA erronée non corrigée
SCORE	2 : l'erreur modifie le sens de la phrase.

Figure 32 : erreur de GNC (2)

Source	***Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie***
TA	*Correction et validation de la dépendance syntaxique dans le Rhapsodie de la banque d'arbres française parlée*
PE	*Correction et validation de la dépendance syntaxique dans Rhapsodie, le corpus arboré du français parlé*
ERREUR	Dans la traduction automatique, le groupe nominal complexe est totalement faux. Les éléments ne sont pas correctement factorisés. De ce fait, on ne comprend pas que *Rhapsodie* est le nom du corpus, et que celui-ci est un corpus arboré du français parlé. Cette erreur est bien corrigée dans la post-édition.
ATTRIBUT	TA erronée bien corrigée
SCORE	Dans la TA : 3, car le texte n'a plus de sens, et l'erreur apparaît dans le titre. Dans la PE : aucun, puisque l'erreur est corrigée.

Figure 33 : erreur de GNC (3)

2.1.4. Type-annotateur_LA-SY-UD

Cette étiquette doit être utilisée dans la post-édition lorsqu'une erreur de syntaxe (détermination, préposition ou GNC) présente dans la traduction automatique est corrigée dans la post-édition.

2.2. Flexion-accord

Tant que cela est possible, il convient d'éviter d'annoter une erreur avec cette étiquette, puisque celle-ci comprend des sous-catégories (ci-dessous). Par conséquent, il faut privilégier l'utilisation des sous-catégories exposées ci-dessous. Si une erreur de flexion ou d'accord ne peut pas être considérée comme une erreur de temps/aspect, de genre ou de nombre, alors il est possible d'utiliser l'étiquette générique « Flexion-accord ».

2.2.1. Temps-aspect_LA-IA-TA

Les erreurs de temps et d'aspect concernent les erreurs de conjugaison, à savoir le choix d'un mauvais temps/aspect grammatical ou simplement une erreur de conjugaison.

Source	***Since the advent of computers, research has focused on the design of digital machine translation tools—computer programs capable of automatically translating a text from a source language to a target language.***
TA	*Depuis l'avènement des ordinateurs, la recherche s'est concentrée sur la conception d'outils numériques de traduction automatique — des programmes informatiques capables de traduire automatiquement un texte d'une langue source vers une langue cible.*
PE	*Depuis l'avènement des ordinateurs, la recherche s'est concentrée sur la conception d'outils numériques de traduction automatique — des programmes informatiques capables de traduire automatiquement un texte d'une langue source vers une langue cible.*
ERREUR	Le *since* anglais demande une conjugaison au *present perfect* anglais. En revanche, cela se traduit souvent par un présent en français. Par ailleurs, le présent serait plus idiomatique avec *depuis*.
ATTRIBUT	TA erronée non corrigée
SCORE	1 : l'erreur n'impacte ni le sens, ni la compréhension, mais rend le texte moins idiomatique.

Figure 34 : erreur de temps/aspect (1)

2.2.2. *Genre_LA-IA-GE*

Avec cette catégorie, il convient d'annoter les erreurs d'accord en genre, telles que celle ci-dessous.

Source	***We train BPE-based attentive Neural Machine Translation systems with and without factored outputs using the open source nmtpy framework.***
TA	*Nous formons des systèmes de traduction automatique neuronales attentifs basés sur BPE avec et sans sorties factorisées en utilisant le framework nmtpy open source*
PE	*Nous avons entraîné des systèmes de traduction automatique neuronale attentifs basés sur la tokenisation BPE, avec et sans sorties factorisées, en utilisant la suite d'outils libre nmtpy.*
ERREUR	Dans la traduction automatique, l'adjectif *neuronal* – qui est relié à *traduction automatique* – ne s'accorde ni avec le substantif *systèmes*, ni avec le terme composé *traduction automatique*.
ATTRIBUT	TA erronée bien corrigée
SCORE	1 : il s'agit d'une erreur, mais elle n'impacte ni le sens, ni la compréhension.

Figure 35 : erreur d'accord en genre (1)

2.2.3. *Nombre_LA-IA-NU*

Il convient d'utiliser cette catégorie pour annoter les erreurs d'accord en nombre.

Source	***Parallel corpora can be leveraged to implement cross-lingual information retrieval or machine translation tools.***
TA	*Les corpus parallèles peuvent être utilisés pour mettre en œuvre des outils de recherche d'informations ou de traduction automatique multilingues.*
PE	*Les corpus parallèles peuvent être utilisés pour mettre en œuvre des outils de recherche d'information ou de traduction automatique multilingue.*
ERREUR	L'erreur d'accord en nombre est introduite dans la post-édition. L'adjectif doit s'accorder avec *outils*.
ATTRIBUT	TA correcte mal corrigée
SCORE	1 : il s'agit d'une erreur, mais elle n'impacte ni le sens, ni la lisibilité.

Figure 36 : erreur d'accord en nombre (1)

2.2.4. *Type-annotateur_LA-IA-UD*

Il convient d'utiliser cette étiquette *type-annotateur* dans la PE lorsqu'une erreur de flexion/accord (temps/aspect, accord en genre ou accord en nombre) présente dans la TA est bien corrigée dans la post-édition.

2.3. Typographie

Il convient d'éviter, si cela est possible, d'annoter une erreur avec cette étiquette, puisque celle-ci comprend des sous-catégories (ci-dessous). Dès lors, il est préférable d'utiliser une des sous-catégories exposées ci-dessous. Si une erreur de typographie ne peut pas être considérée comme une erreur d'orthographe, d'accents diacritiques, de casse ou de ponctuation, alors il est possible d'utiliser l'étiquette générique « Typographie ».

2.3.1. Orthographie_LA-HY-SP

Une erreur d'« orthographie » est une faute dans la façon dont les mots sont écrits, par rapport aux règles d'orthographe établies.

Source	***On the SPMRL dataset, our parser obtains above state-of-the-art results on constituency parsing without requiring either predicted POS or morphological tags, and outputs labelled dependency trees.***
TA	*Sur l'ensemble de données SPMRL, notre analyseur obtient ci-dessus des résultats de pointe sur l'analyse des circonscriptions sans nécessiter une prévision de POS ou d'étiquettes morphologiques, et des sorties marquées d'arbres de dépendance.*
PE	*Sur l'ensemble de données SPMRL, notre analyseur obtient des résultats supérieurs à l'état de l'art en analyse syntaxique en constituents sans nécessiter de parties du discours prédites ni d'étiquettes morphologiques prédites, et permet de construire des arbres syntaxiques en dépendances étiquetées.*
ERREUR	Il y a une erreur dans l'orthographe de ce mot, qui doit s'écrire *constituants*.
ATTRIBUT	TA erronée mal corrigée
SCORE	1 : il s'agit d'une erreur, mais elle n'impacte ni le sens, ni la compréhension.

Figure 37 : erreur d'orthographe (1)

2.3.2. Accent-diacritiques_LA-HY-AC

Il s'agit des erreurs causées par la non-utilisation ou la mauvaise utilisation des accents, comme la confusion entre un accent aigu et un accent grave, par exemple.

Source	***The transient migrates to the northwest where it slowly decays beneath the locked zone.***
Traduction	*La transmission migre vers le nord-ouest ou il s'affaiblit lentement sous la zone bloquée.*
ERREUR	Il manque l'accent sur le <u>.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	2 : la phrase peut avoir un autre sens, la lisibilité est perturbée.

Figure 38 : erreur d'accent diacritique (2)

2.3.3. *Mauvaise-casse_LA-HY-CA*

Une erreur de casse se produit lorsqu'une lettre est utilisée avec une majuscule ou une minuscule incorrecte dans un mot.

Source	***LIUM Machine Translation Systems for WMT17 News Translation Task***
TA	*Systèmes de traduction automatique Lium pour WMT17 Nouvelles Tâche de Traduction*
ERREUR	Le nom du laboratoire LIUM s'écrit en lettres majuscules.
ATTRIBUT	Aucun, puisque l'erreur se produit dans la traduction automatique.
SCORE	2 : le sens est légèrement impacté. En effet, on pourrait croire qu'il s'agit du nom du système, et non du nom du laboratoire LIUM.

Figure 39 : erreur de casse (2)

2.3.4. *Ponctuation_LA-HY-PU*

Dans cette catégorie, il convient de regrouper les erreurs liées à la ponctuation (virgule omise ou superflue, point final manquant, espace (insécable) manquante avec un signe de ponctuation double, etc.).

Source	***Electronic versions of literary works abound on the Internet and the rapid dissemination of electronic readers will make electronic books more and more common.***
TA	*Les versions électroniques d'œuvres littéraires abondent sur l'internet et la diffusion rapide des lecteurs électroniques rendra les livres électroniques de plus en plus courants.*
PE	*Les versions électroniques d'œuvres littéraires abondent sur Internet et la diffusion rapide des liseuses électroniques rendra les livres électroniques de plus en plus courants.*
ERREUR	L'usage préconise l'utilisation de la virgule avant la conjonction de coordination *et* lorsque le sujet change[4]. En effet, sans la virgule, il y a un risque d'équivoque.
ATTRIBUT	TA erronée non corrigée
SCORE	1 : l'oubli de la virgule n'impacte ni le sens, mais légèrement la lisibilité.

Figure 40 : erreur de ponctuation (1)

2.3.5. *Type-annotateur_LA-HY-UD*

Il convient d'utiliser cette étiquette *type-annotateur* dans la PE lorsqu'une erreur de typographie (orthographie, accents diacritiques, mauvaise casse ou ponctuation) présente dans la TA est bien corrigée dans la post-édition.

[4] https://www.btb.termiumplus.gc.ca/redac-chap?lang=fra&lettr=chapsect6&info0=6.1.

2.4. Registre

Il convient d'éviter tant que possible d'annoter une erreur avec cette étiquette, puisque celle-ci comprend des sous-catégories (erreurs d'inadaptation au texte source ou au texte cible).

2.4.1. Incompatible-texte-source_LA-RE-IS

Une erreur d'incompatibilité avec le texte source apparaît lorsque le registre utilisé dans la traduction ne correspond pas à celui employé dans le texte de départ (par exemple lorsqu'une expression vulgaire dans le texte source est « lissée » dans la traduction).

Source	***The transition from blueschist or amphibolite to eclogite is expected to notably increase the viscosity of oceanic crust (3); however, here, we are considering sub-eclogite facies conditions.***
Traduction	*On s'attend à ce que le passage de la blueschiste ou de l'amphibolite à l'éclogite augmente considérablement la viscosité de la croûte océanique (3) ; cependant, nous considérons ici des conditions de faciès sub-éclogite*
ERREUR	Le registre n'est pas tout à fait le même que dans le texte source. En corpus, on remarque que la tournure passive *is expected to* est rarement traduite par une formulation en « on », mais plutôt avec le verbe *devoir* au conditionnel.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	1 : ce n'est pas tout à fait naturel, mais pas grave.

Figure 41 : erreur d'incompatibilité avec le texte source (1)

2.4.2. Inadapte-au-type-texte-cible_LA-RE-IT

Il y a une erreur d'inadaptation au texte cible lorsque le registre utilisé dans la traduction n'est pas conforme au registre attendu pour le type de texte en question (par exemple si le ton employé est trop informel dans un article scientifique).

Source	***In particular, we show that we can build variants of our parser with smaller search spaces and time complexities ranging from O(n^6) down to O(n^3)***
TA	*En particulier, nous montrons que nous pouvons construire des variantes de notre analyseur syntaxique avec des espaces de recherche plus petits et des complexités temporelles allant de O(n^6) à O(n^3).*
PE	*En particulier, nous montrons que nous pouvons construire des variantes de notre analyseur syntaxique avec des espaces de recherche restreints et des complexités temporelles allant de O(n^6) à O(n^3).*
ERREUR	Le groupe adjectival *plus petits*, qui est une formulation assez creuse, n'est pas tout à fait adapté au genre textuel scientifique.
ATTRIBUT	TA erronée bien corrigée
SCORE	0 : l'erreur n'affecte ni le sens, ni la compréhension, ni la lisibilité.

Figure 42 : erreur d'inadaptation au texte cible (0)

2.4.3. Type-annotateur_LA-RE-UD

Il convient d'utiliser cette étiquette lorsqu'une des deux erreurs ci-dessus a été corrigée correctement dans la post-édition (cf. Figure 42).

2.5. Style

Il existe différents types d'erreurs de style (voir 2.5.1., 2.5.2. et 2.5.3. ci-dessous). Il convient de privilégier autant que possible une de ces trois sous-catégories.

2.5.1. Formulation-maladroite_LA-ST-AW

Une formulation maladroite est une erreur de qui se caractérise par des choix de mots ou une structure de phrase peu idiomatiques, ce qui donne un aspect artificiel ou peu naturel dans la langue cible. Cette erreur peut affecter la lisibilité du texte traduit, le rendant souvent difficile à comprendre[5].

Source	***The main approaches are presented from a largely historical perspective and in an intuitive manner, allowing the reader to understand the main principles without knowing the mathematical details.***
TA	*Les approches principales sont présentées d'un point de vue largement historique et d'une manière intuitive, permettant au lecteur de comprendre les principes principaux sans connaître les détails mathématiques.*
ERREUR	L'enchaînement entre l'adjectif et le substantif – qui ont la même racine – n'est pas naturel et est dérangeant.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	1 : ce n'est pas tout à fait naturel, mais pas grave.

Figure 43 : exemple de formulation maladroite (1)

Source	***The impact of back-translation quantity and quality is also analyzed for English→Turkish where our post-deadline submission surpassed the best entry by +1.6 BLEU.***
TA	*L'impact de la rétro-traduction quantitative et de la qualité est également analysé pour English→Turkish où notre soumission post-date a dépassé la meilleure entrée de + 1,6 BLEU*
ERREUR	La phrase n'est pas claire en raison de sa formulation.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	2 : ce n'est pas tout à fait naturel, et la compréhension est affectée.

Figure 44 : exemple de formulation maladroite (2)

[5] De nombreuses erreurs peuvent être considérées comme des formulations maladroites. Dès lors, il convient — si cela est possible — d'utiliser des étiquettes plus précises pour catégoriser l'erreur.

2.5.2. *Tautologie_LA-ST-TA*

Une tautologie, aussi appelée pléonasme, est un « [p]rocédé rhétorique ou négligence de style consistant à répéter une idée déjà exprimée, soit en termes identiques (ex. *au jour d'aujourd'hui*), soit en termes équivalents (*monter en haut*) »[6].

Source	***This corpus offers a lot of future prospects, for instance concerning synthesis with virtual signers, machine translation or formal grammars for Sign Language.***
TA	*Ce corpus offre de nombreuses perspectives d'avenir, par exemple en matière de synthèse avec des signataires virtuels, de traduction automatique ou de grammaires formelles pour la langue des signes.*
PE	*Ce corpus offre de nombreuses perspectives pour le futur, par exemple en matière de synthèse avec des signeurs virtuels, de traduction automatique ou de grammaires formelles pour la langue des signes.*
ERREUR	Dans la TA, le groupe nominal *perspectives d'avenir* est un pléonasme, étant donné que *perspectives* englobe déjà l'idée d'avenir. Dans la post-édition, l'erreur est la même, sauf qu'elle est plus grave, étant donné que *perspectives pour le futur* est une collocation rare, contrairement à *perspectives d'avenir*.
ATTRIBUT	TA erronée mal corrigée
SCORE	TA : 0, puisque l'on retrouve cette expression assez souvent PE : 1, puisque la collocation est rare et moins naturelle

Figure 45 : exemple de tautologie (0 et 1)

2.5.3. *Style-titre_LA-ST-TS*

Cette catégorie sert à annoter les erreurs de style dans les titres. En effet, les normes appliquées aux titres varient en fonction des langues. Par exemple, en anglais, on privilégie l'utilisation de la majuscule, ce qui n'est pas le cas en français.

2.5.4. *Type-annotateur_LA-ST-UD*

Il convient d'utiliser cette étiquette *type-annotateur* dans la PE lorsqu'une erreur de style (formulation maladroite, tautologie ou style du titre) présente dans la TA est bien corrigée dans la post-édition.

[6] https://www.cnrtl.fr/definition/tautologie.

2.6. Reference-pas-claire_LA-UR

Une référence n'est pas claire quand l'élément auquel elle fait référence n'est pas immédiatement identifiable dans le texte. Cela peut se produire lorsqu'un pronom, un nom, un déterminant ou un autre élément est utilisé de manière ambiguë, ce qui rend difficile pour le lecteur de comprendre à quoi ou à qui il fait référence.

Source	***The unorthodox language phenomena observed as well as the rich-in-terminology scientific domains addressed in the educational video lectures, the language-independent nature of the approach, and the tackled three-class classification problem constitute innovative challenges of the work described herein.***
TA	*Les phénomènes langagiers non orthodoxes observés, ainsi que les domaines scientifiques riches en terminologie abordés dans les conférences vidéo éducatives, la nature de l'approche indépendante de la langue et le problème de classification à trois classes abordé constituent des défis innovants de l'oeuvre décrite ici.*
PE	*Les phénomènes langagiers singuliers observés, ainsi que les domaines scientifiques riches en terminologie abordés dans les conférences éducatives filmées, la nature de notre approche, indépendante de la langue, et le problème de classification en trois catégories abordé constituent les défis innovants de la présente étude.*
ERREUR	On ne comprend pas directement qu'il s'agit de l'approche adoptée par ceux qui ont mené l'étude. La post-édition est plus claire à cet égard.
ATTRIBUT	TA erronée bien corrigée
SCORE	TA : 1, pas clair, mais ne nuit pas vraiment à la lisibilité PE : aucun, puisqu'il n'y a plus d'erreur.

Figure 46 : exemple de référence pas claire (1)

2.6.1. Type-annotateur_LA-UD

Il convient d'utiliser cette étiquette *type-annotateur* dans la PE lorsqu'une erreur de référence présente dans la TA est bien corrigée dans la post-édition.

2.7. Conventions-textuelles

On distingue deux types d'erreurs liées aux conventions textuelles, à savoir les erreurs de cohérence et les erreurs de cohésion. Il convient de privilégier autant que possible une de ces deux sous-catégories. Il est important de distinguer ces deux notions. La cohésion se définit comme « l'ensemble des opérations qui permettent d'assurer le suivi d'une phrase à l'autre », alors que la cohérence « considère le texte d'un point de vue plus global » (Benali, 2012).

2.7.1. Coherence_LA-TC-CE

La cohérence d'un texte ne se mesure pas par des marques linguistiques précises, à l'inverse de la cohésion. La cohérence concerne l'enchaînement logique entre les différentes idées ou propositions dans le texte (Benali, 2012).

Source	***It then takes up the history of machine translation in more detail, describing its pre-digital beginnings, rule-based approaches, the 1966 ALPAC (Automatic Language Processing Advisory Committee) report and its consequences, the advent of parallel corpora, the example-based paradigm, the statistical paradigm, the segment-based approach, the introduction of more linguistic knowledge into the systems, and the latest approaches based on deep learning.***
TA	*Il reprend ensuite plus en détail l'histoire de la traduction automatique, décrivant ses débuts prénumériques, ses approches fondées sur des règles, le rapport de 1966 ALPAC (Automatic Language Processing Advisory Committee) et ses conséquences, l'avènement de corpus parallèles, le paradigme basé sur l'exemple, le paradigme statistique, l'approche par segment, l'introduction de plus de connaissances linguistiques dans les systèmes, et les dernières approches basées sur l'apprentissage profond.*
PE	*Il reprend ensuite plus en détail l'histoire de la traduction automatique, décrivant ses débuts pré-numériques, ses approches fondées sur des règles, le rapport ALPAC de 1966 (Automatic Language Processing Advisory Committee) et ses conséquences, l'avènement de corpus parallèles, le paradigme basé sur l'exemple, le paradigme statistique, l'approche par segment, l'introduction de plus de connaissances linguistiques dans les systèmes, et les dernières approches basées sur l'apprentissage profond.*
ERREUR	Dans la post-édition, la cohérence textuelle n'est pas respectée. En effet, il serait plus logique de lire la forme développée de l'acronyme ALPAC juste après celui-ci.
ATTRIBUT	TA correcte mal corrigée
SCORE	1, pas réellement une erreur, mais l'agencement perturbe légèrement la lecture.

Figure 47 : exemple d'erreur de cohérence (1)

2.7.2. Cohesion_LA-TC-CN

La cohésion, à l'inverse de la cohérence, se mesure par le biais de marques linguistiques précises (Benali, 2012). Les différentes opérations qui permettent la cohésion textuelle sont, entre autres, la coordination, les connecteurs logiques, les anaphores, etc. (voir par exemple Charolles, 2011 ; Halliday & Hasan, 1976). Voici un exemple d'erreur affectant la cohésion textuelle.

Source	***To further characterize performance, we report distributions over 30 runs and different sizes of training datasets.***
TA	*Pour mieux caractériser les performances, nous rapportons les distributions sur 30 exécutions et différentes tailles d'ensembles de données d'entraînement.*
PE	*Pour mieux caractériser les performances de l'outil, nous rapportons les distributions sur 30 itérations et différentes tailles de corpus d'entraînement.*
ERREUR	Dans la TA, il est plus difficile d'identifier de quelles performances il est question. C'est pourquoi le post-éditeur a choisi de recourir à une explicitation.
ATTRIBUT	TA erronée bien corrigée
SCORE	TA : 1, ce n'est pas faux, mais pas très clair. PE : aucun, puisqu'il n'y a plus d'erreur.

Figure 48 : exemple d'erreur de cohésion (1)

2.7.3. Type-annotateur_LA-TC-UD

Cette étiquette sert à annoter la post-édition lorsqu'une erreur de cohérence ou de cohésion présente dans la TA a été corrigée.

2.8. Terminologie-lexique

Il est nécessaire de privilégier une annotation plus granulaire et d'utiliser une des étiquettes ci-dessous qui regroupent des types d'erreurs terminologiques/lexicales précis.

2.8.1. Choix-incorrect-Termino_LA-TL-INS

On considère qu'une erreur est un choix terminologique incorrect lorsqu'un terme dans le texte source est traduit par un terme incorrect dans la traduction.

Source	***To further characterize performance, we report distributions over 30 runs and different sizes of training datasets.***
Traduction	*Pour mieux caractériser les performances, nous rapportons les distributions sur 30 exécutions et différentes tailles d'ensembles de données d'entraînement.*
ERREUR	Dans le domaine du traitement automatique des langues, on parle d'*itération*, et non d'exécution.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	1 : on peut comprendre, mais le terme n'est pas exact.

Figure 49 : exemple de choix incorrect terminologique (1)

Source	***Herein, we evaluate YASET on part-of-speech tagging and named entity recognition in a variety of text genres including articles from the biomedical literature in English and clinical narratives in French.***
TA	*Ici, nous évaluons YASET sur l'étiquetage de la partie du discours et la reconnaissance des entités nommées dans une variété de genres de textes, y compris des articles de la littérature biomédicale en anglais et des récits cliniques en français.*
PE	*Dans cet article, nous évaluons YASET sur l'étiquetage morphosyntaxique et la reconnaissance d'entités nommées dans une variété de corpus, dont des articles de la littérature biomédicale en anglais et des documents cliniques en français.*
ERREUR	La traduction automatique est correcte (*text genre = genre de texte*), bien que le terme *genre textuel* soit plus usité. En revanche, *corpus* n'est pas un synonyme.
ATTRIBUT	TA correcte mal corrigée
SCORE	PE : 2, il y a une distorsion causée par une erreur terminologique.

Figure 50 : exemple de choix incorrect terminologique (2)

2.8.2. Choix-incorrect-Langue-Generale_LA-TL-ING

On considère qu'une erreur est un choix incorrect de la langue générale lorsqu'un terme/élément lexical de la langue générale (et non spécialisée) dans le texte source est traduit par un terme/élément lexical incorrect dans la traduction.

Source	***The syntactic annotation contains two levels: a macro-syntactic level, containing a segmentation into illocutionary units (including discourse markers, parentheses ...) and a micro-syntactic level including dependency relations and various paradigmatic structures, called pile constructions, the latter being particularly frequent and diverse in spoken language.***
Traduction	*L'annotation syntaxique contient deux niveaux: un niveau macro-syntactique, contenant une segmentation en unités illocutionnaires (y compris des marqueurs de discours, des parenthèses...) et un niveau micro-syntactique comprenant des relations de dépendance et diverses structures paradigmatiques, appelées constructions de pile, ces dernières étant particulièrement fréquentes et diversifiées dans le langage parlé*
ERREUR	Dans cette phrase, l'adjectif anglais *diverse* ne signifie pas exactement *diversifié*, mais *divers.*
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition
SCORE	1, l'erreur n'est pas grave, mais il y a une légère nuance qui change.

Figure 51 : exemple de choix incorrect dans la langue générale (1)

2.8.3. Mauvais-acronyme-abreviation_LA-TL-MAA

Cette catégorie sert à annoter les erreurs liées aux acronymes et aux abréviations, par exemple lorsque l'acronyme ou l'abréviation ne correspond pas à la forme développée. Voici un exemple de ce type d'erreur.

Source	***In this paper, we address the problem of generating English tag questions (TQs) (e.g. it is, isn't it?) in Machine Translation (MT).***
TA	*Dans cet article, nous abordons le problème de la génération de questions à étiquette (TQ) en anglais (par exemple, it is, isn't it ?) dans le cadre de la traduction automatique (TA).*
PE	*Dans cet article, nous abordons le problème de la génération de questions à étiquette (QT) en anglais (par exemple, it is, isn't it ?) dans le cadre de la traduction automatique (TA).*
ERREUR	L'acronyme anglais TQ correspond au terme *tag questions*. En revanche, le post-éditeur a décidé d'inverser les deux lettres de l'acronyme, ce qui ne correspond pas à sa traduction française.
ATTRIBUT	TA correcte mal corrigée
SCORE	1 : pas grave, mais pas logique.

Figure 52 : exemple de mauvais acronyme/abréviation (1)

⚠ Une confusion peut parfois se produire entre cette catégorie et la catégorie « traduisible non traduit » (1.6.1.). Prenons l'exemple ci-dessous :

*« Dans cet article, nous proposons un cadre pour imiter le processus d'amorçage dans un contexte de traduction automatique neuronale (**NMT**). »*

Ici, le sigle anglais NMT reste en anglais, et celui-ci est correct, mais il est moins utilisé que son équivalent français TAN. Dans ce cas, cette erreur ne doit pas être considérée comme une erreur de mauvais acronyme/abréviation, mais comme une erreur de « traduisible non traduit », puisqu'une traduction française est largement utilisée.

2.8.4. Faux-amis_LA-TL-FC

Un faux-ami est une erreur de traduction où des mots similaires dans deux langues peuvent sembler équivalents, mais ont des significations différentes, par exemple lorsque *actually* en anglais est traduit en français par *actuellement*, plutôt que par *en fait* ou *en réalité*.

Source	***The syntactic annotation contains two levels: a macro-syntactic level, containing a segmentation into illocutionary units (including discourse markers, parentheses ...) and a micro-syntactic level including dependency relations and various paradigmatic structures, called pile constructions, the latter being particularly frequent and diverse in spoken language.***
TA	*L'annotation syntaxique contient deux niveaux: un niveau macro-syntactique, contenant une segmentation en unités illocutionnaires (y compris des marqueurs de discours, des parenthèses...) et un niveau micro-syntactique comprenant des relations de dépendance et diverses structures paradigmatiques, appelées constructions de pile, ces dernières étant particulièrement fréquentes et diversifiées dans le langage parlé.*
PE	*L'annotation syntaxique contient deux niveaux : un niveau macro-syntaxique, contenant une segmentation en unités illocutoires (y compris des marqueurs de discours, des parenthèses...) et un niveau micro-syntaxique comprenant des relations de dépendance et diverses structures paradigmatiques, appelées constructions de pile, ces dernières étant particulièrement fréquentes et diverses dans le langage parlé.*
ERREUR	Le terme anglais *language* peut effectivement avoir plusieurs traductions, dont *langage*. Toutefois, dans cette expression, on utilise le substantif *langue*, et non *langage*.
ATTRIBUT	TA erronée non corrigée
SCORE	1 : on comprend, mais ce n'est pas correct.

Figure 53 : exemple de faux-amis (1)

2.8.5. *Terme-traduit-par-non-terme_LA-TL-NT*

Cette catégorie sert à annoter les erreurs dans lesquelles un terme dans le texte source est traduit par un non-terme. Il convient de distinguer cette catégorie de l'erreur de choix terminologique incorrect, qui sert à identifier les erreurs où une terme du texte source est traduit par un terme incorrect.

Source	***We introduce a novel chart-based algorithm for span-based parsing of discontinuous constituency trees of block degree two, including ill-nested structures.***
TA	*Nous présentons un nouvel algorithme basé sur les diagrammes pour l'analyse syntaxique basée sur l'étendue des arbres de circonscription discontinus de degré de bloc deux, y compris les structures mal imbriquées.*
PE	*Nous présentons un nouvel algorithme tabulaire pour l'analyse syntaxique fondées sur les empans des arbres en constituants discontinus de degré de bloc deux, y compris les structures mal imbriquées.*
ERREUR	Dans le domaine du traitement automatique des langues, un *arbre en constituants* est un terme. Dans la TA, *circonscription* n'est pas un terme du domaine.
ATTRIBUT	TA erronée bien corrigée
SCORE	TA : 3, il n'y a plus de lien avec le domaine, c'est incompréhensible. PE : aucun, puisque l'erreur est corrigée.

Figure 54 : exemple de terme traduit par un non-terme (3)

Source	***Herein, we evaluate YASET on part-of-speech tagging and named entity recognition in a variety of text genres including articles from the biomedical literature in English and clinical narratives in French.***
Traduction	*Ici, nous évaluons YASET sur l'étiquetage de la partie du discours et la reconnaissance des entités nommées dans une variété de genres de textes, y compris des articles de la littérature biomédicale en anglais et des récits cliniques en français.*
ERREUR	Dans le domaine du traitement automatique, on parle plutôt d'étiquetage morpho-syntaxique, qui est un terme spécialisé.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	1 : cela reste compréhensible, mais ce n'est pas le terme spécialisé exact.

Figure 55 : exemple de terme traduit par un non-terme (1)

2.8.6. *Collocation-incorrecte-Specialise_LA-TL-ICS*

On entend par « collocation » une combinaison de « deux unités lexicales ou plus dont l'une au moins est un terme et dont la totalité des parties ne désigne pas un et un seul concept » (Brisson,

2019). Cette catégorie regroupe les erreurs de collocations spécialisées, c'est-à-dire des erreurs où la collocation comprend un terme du domaine en question. En voici deux exemples.

Source	***We evaluate our approach on German and English treebanks (Negra, Tiger, and DPTB) and report state-of-the-art results in the fully supervised setting.***
TA	*Nous évaluons notre approche sur des banques de données allemandes et anglaises (Negra, Tiger et DPTB) et rapportons des résultats de pointe dans un cadre entièrement supervisé.*
PE	*Nous évaluons notre approche sur des jeux de données en allemand et en anglais (Negra, Tiger et DPTB) et rapportons des résultats à l'état de l'art dans un cadre entièrement supervisé.*
ERREUR	*State-of-the-art*, dans le domaine du traitement automatique des langues, est souvent traduit par *à l'état de l'art*. Dès lors, la collocation de la TA n'est pas tout à fait correcte, bien que l'idée soit la même.
ATTRIBUT	TA erronée bien corrigée
SCORE	TA : 0, on retrouve tout de même cette collocation en corpus, mais en moindre mesure. PE : aucun, puisque l'erreur est corrigée.

Figure 56 : exemple de collocation incorrecte spécialisée (0)

Source	***We train BPE-based attentive Neural Machine Translation systems with and without factored outputs using the open source nmtpy framework.***
TA	*Nous formons des systèmes de traduction automatique neuronales attentifs basés sur BPE avec et sans sorties factorisées en utilisant le framework nmtpy open source.*
PE	*Nous avons entraîné des systèmes de traduction automatique neuronale attentifs basés sur la tokenisation BPE, avec et sans sorties factorisées, en utilisant la suite d'outils libre nmtpy.*
ERREUR	Le terme *système de traduction automatique neuronale* n'est pas accompagné du bon collocat. En effet, dans le domaine du traitement automatique des langues, on dit bien *entraîner des systèmes de TA*, et non *former des systèmes de TA*.
ATTRIBUT	TA erronée bien corrigée
SCORE	TA : 2, cette collocation est une collocation courante du domaine, et est donc totalement incorrecte, bien qu'on puisse toujours la comprendre. PE : aucun, puisque l'erreur est corrigée.

Figure 57 : exemple de collocation incorrecte spécialisée (2)

2.8.7. Collocation-incorrecte-Langue-Generale_LA-TL-ICG

Cette catégorie sert également à annoter les erreurs de collocation, mais ici, elle concerne les collocations de la langue générale, c'est-à-dire celles qui ne contiennent pas de terme du domaine de spécialité.

Source	***This paper proposes a new type of evaluation focused specifically on the morphological competence of a system with respect to various grammatical phenomena.***
TA	*Cet article propose un nouveau type d'évaluation axé spécifiquement sur la compétence morphologique d'un système par rapport à divers phénomènes grammaticaux.*
PE	*Cet article propose un nouveau type d'évaluation axé spécifiquement sur la compétence morphologique d'un système par rapport à divers phénomènes grammaticaux.*
ERREUR	Dans la langue scientifique française, on tend à éviter les anthropomorphismes, ce que l'on retrouve ici.
ATTRIBUT	TA erronée non corrigée
SCORE	1 : ce n'est pas naturel, mais l'erreur ne nuit pas à la compréhension ou au sens de la phrase.

Figure 58 : exemple de collocation incorrecte de la langue générale (1)

2.8.8. Choix-incompatible-avec-texte-cible_LA-TL-IT

On considère qu'un choix terminologique est incompatible avec le texte cible lorsqu'un terme du texte source est traduit par un terme théoriquement correct dans la langue cible, mais que le terme choisi n'est pas le terme approprié au vu de différents facteurs (registre, genre textuel, etc.).

Par exemple, dans un article scientifique du domaine des sciences de la terre, le terme *moonquake* peut être traduit par *tremblement de lune* dans la langue générale, mais sa traduction correcte en langue de spécialité est *séisme lunaire*. Par conséquent, si le traducteur décide de traduire ce terme par *tremblement de lune* dans un article scientifique, il s'agit d'un choix incompatible avec le texte cible.

2.8.9. Incoherence-terminologique

Dans cette typologie, on distingue deux types d'incohérences terminologiques (voir ci-dessous). Il convient de privilégier une de ces deux sous-catégories autant que possible.

2.8.9.1. Differents-termes-traduction_LA-TL-TI-DT

Le premier cas d'incohérence terminologie est lorsqu'on observe différentes traductions pour un seul terme dans la langue source. Voici un exemple.

Source	***We propose a method to inject priming cues into the NMT networkand compare our framework to other mechanisms that perform micro-adaptation during inference.*** ***[…] Overall, experiments conducted in a multi-domain setting confirm that adding priming cues in the NMT decoder can go a long way towards improving the translation accuracy. Besides, we show the suitability of our framework to gather valuable information for an NMT network from monolingual resources.***
Traduction	*Nous proposons une méthode pour injecter des signaux d'amorçage dans le réseau et nous comparons notre framework à d'autres mécanismes qui effectuent une micro-adaptation pendant l'inférence.* *[…]Dans l'ensemble, les expériences conduites dans un contexte multi-domaines confirment que l'ajout d'indices d'amorçage dans le décodeur du système peut contribuer grandement à améliorer la précision de la traduction.*
ERREUR	Ici, le terme *cues* a été traduit par *signaux*, puis par *indices*. Il s'agit donc d'une incohérence. Le terme correct est *signaux*.
ATTRIBUT	Aucun, puisque ce n'est pas une post-édition.
SCORE	1 : cela reste compréhensible, mais peut être perturbant à la lecture.

Figure 59 : exemple de différentes traduction pour un seul terme (1)

2.8.9.2. Differentes-abbreviations-traduction_LA-TL-TI-DA

Une autre source d'incohérence terminologique est lorsqu'on observe différentes abréviations, différents acronymes ou sigles pour un même terme dans la traduction.

Par exemple, il y a dans une traduction le terme *traduction automatique neuronale*. Le traducteur choisit tantôt le sigle *NMT*, tantôt le sigle *TAN*. Dans ce cas, il s'agit d'une erreur de différentes abréviations dans la traduction.

➔ Si ces deux abréviations/acronymes/sigles sont utilisés dans la langue cible, l'annotateur choisit un score de gravité 1, car l'erreur n'est pas grave, mais elle peut perturber le lecteur ;

➔ Si une de ces deux abréviations n'est pas utilisée dans la langue cible, l'erreur peut être considérée comme une erreur plus grave (score 2).

2.8.10. Type-annotateur_TL-UD

Cette étiquette sert à annoter la post-édition lorsqu'une des erreurs de terminologie répertoriées ci-dessus est présente dans la TA, mais que celle-ci a été corrigée dans la post-édition.

3. Outils

Les erreurs liées aux outils ou à leur maîtrise sont plus rares, mais pas exclues. Il convient d'utiliser, si cela est possible, une des quatre sous-catégories présentées ci-dessous.

3.1. Hallucination_OU-TAH

Une hallucination peut être définie comme une suite de « fragments de phrase complètement illogiques ajoutés ou remplacés dans la traduction ; [il peut s'agit de] termes inventés en raison d'une mauvaise segmentation ou factorisation des unités lexicales » (Hansen & Esperança-Rodier, 2022, traduction). Il s'agit dès lors d'une sortie de la TA qui est totalement déconnectée du texte source. Par conséquent, les hallucinations sont très souvent des erreurs considérées comme des erreurs graves. En voici un exemple :

Source	***This paper describes LIUM submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs.***
TA	*Cet article décrit les soumissions de LIUM à WMT17 News Translation Task pour l'anglais, l'allemand, l'anglais, l'anglais, le tchèque et l'anglais→langue latine.*
PE	*Cet article décrit les contributions du LIUM à la tâche de traduction d'articles de presse de la conférence WMT17 pour les paires de langues anglais↔allemand, anglais↔turc, anglais→tchèque et anglais→letton.*
ERREUR	Ici, on peut supposer que le formatage du texte source (flèches) a perturbé le système de TA. Dès lors, en plus de la mauvaise reproduction de ce formatage, on remarque qu'il a traduit *Latvian* par *langue latine,* ce qui est absolument faux.
ATTRIBUT	TA erronée bien corrigée
SCORE	3 : la traduction est totalement déconnectée du texte source, et le sens est complètement faux.

Figure 60 : exemple d'hallucination (3)

3.2. Conformite-corpus_OU-CC

Cette catégorie n'est à utiliser que si un corpus a été mis à disposition du traducteur est que celui-ci avait pour consigne d'utiliser ce corpus. Cette erreur regroupe dès lors les erreurs liées au non-respect du corpus.

Par exemple, si un traducteur traduit un terme par un mauvais terme, il peut s'agir d'une erreur de choix incorrect terminologique, mais aussi d'une erreur de conformité au corpus. Les deux étiquettes doivent dès lors être utilisées.

3.3. Duplication_OU-DU

Une erreur de duplication se produit lorsque le traducteur ou le post-éditeur ne relit pas correctement sa production et qu'il y laisse plusieurs fois le même mot.

Par exemple, si on remarque que le traducteur a écrit « j'y suis allé le *le* matin », il s'agit d'une erreur de duplication. Cette erreur n'est pas grave (1), puisqu'elle n'affecte pas le sens ni la compréhension, mais elle peut gêner légèrement la lecture.

3.4. Choix-incompatible-glossaire_OU-GC

Comme la catégorie 3.2., celle-ci ne s'utilise que lorsqu'un glossaire ou une base terminologique a été fournie au traducteur, et que celui-ci ne respecte pas les entrées de ce glossaire ou de cette base. Ici aussi, l'erreur peut comporter plusieurs étiquettes.

Par exemple, s'il s'agit d'une erreur de terminologique (ce qui est probable, puisqu'un glossaire n'est censé recenser que des termes), l'erreur comportera une étiquette d'erreur terminologique et une étiquette de choix incompatible avec le glossaire.

3.5. Type-annotateur_OU-UD

Cette étiquette est utilisée dans la post-édition lorsqu'une erreur liée aux outils présente dans la TA est corrigée correctement dans la post-édition.

Liste des figures

Sources

Benali, A. (2012). « Les problèmes de la catégorisation textuelle : entre fondements théoriques et fondements structurels ». *Synergies Algérie*, 2012, 17, pp. 35-49.

Brisson, F. (2019). *Les compétences terminologiques du traducteur : pistes de réflexion pour un enseignement de la terminologie à l'usage de futurs traducteurs*. Université Savoie Mont Blanc.

Charolles, M. (2011). *Cohérence et cohésion du discours*. Holker, K. ; Marello, C. *Dimensionen der Analyse Texten und Diskursivent - Dimensioni dell'analisi di testi e discoursi*, Lit Verlag, pp. 153-173.

Delisle, J. (2003). *La traduction raisonnée. Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. Ottawa, Presses de l'Université d'Ottawa.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London, Longman

Hansen, D. & Esperança-Rodier, E. (2022). "Human-Adapted MT for Literary Texts: Reality or Fantasy?" *NeTTT 2022*, Jul 2022, Rhodes, Greece. pp. 178-190.

Tautologie (définition du Centre National de Ressources Textuelles et Lexicales, CNRTL). Lien : https://www.cnrtl.fr/definition/tautologie

Tautologie (définition Termium). Lien : https://www.btb.termiumplus.gc.ca/redac-chap?lang=fra&lettr=chapsect6&info0=6.1.