Étude des Protocoles d' Évaluation Humaine pour la Traduction de Documents

Maud Bénard, Natalie Kübler, Alexandra Mestivier, Joachim Minder, Lichao Zhu

;

Étude des Protocoles d'Évaluation Humaine pour la Traduction de Documents

Maud Bénard, Natalie Kübler, Alexandra Mestivier,
Joachim Minder et Lichao Zhu

CLILLAC-ARP, Université Paris Cité

Septembre 2024

MaTOS — Livrable D4-1.1

Machine Translation for Open Science - ANR-22-CE23-0033

Étude des Protocoles d'Évaluation Humaine pour la Traduction de Documents

septembre 2024

Contributeurs :
Maud Bénard, CLILLAC-ARP, Université Paris Cité
Natalie Kübler, CLILLAC-ARP, Université Paris Cité
Alexandra Mestivier, CLILLAC-ARP, Université Paris Cité
Joachim Minder, CLILLAC-ARP, Université Paris Cité
Lichao Zhu, CLILLAC-ARP, Université Paris Cité

Résumé

Ce rapport fait le point sur les différents protocoles qui permettent d'évaluer la qualité de la traduction humaine, de la traduction automatique et/ou de la post-édition. Après un bref résumé sur les métriques automatiques développées en TALN, nous nous concentrons sur les protocoles d'évaluation mis en oeuvre par des humains. Les approches psychologistes sont distinguées des approches textuelles ou discursives. Nous abordons de manière plus approfondie la description des approches textuelles, à savoir, principalement des typologies d'erreurs, dans les contextes théoriques, professionnels et pédagogiques pour évaluer la qualité de la traduction humaine et automatique et de la post-édition. Finalement, nous développons la nouvelle typologie adaptée à ces trois types de production qui est mise en œuvre dans le projet MaTOS. Le manuel de cette typologie est présenté en annexe.

Mots-clé : évaluation humaine, traduction, traduction automatique, post-édition, typologie d'erreurs, annotation d'erreurs

Abstract

This report provides an overview of the various protocols for evaluating the quality of human translation, machine translation and/or post-editing. After a brief summary about the automatic metrics developed in NLP, we focus on the evaluation protocols

implemented by humans. Psychological approaches are distinguished from textual or discursive approaches. We discuss in more detail the description of textual approaches, i.e. mainly error typologies, in theoretical, professional and pedagogical contexts for evaluating the quality of human and machine translation and post-editing. Finally, we develop the new typology adapted to these three types of production, which is being implemented in the MaTOS project. The manual for this typology is presented in the appendix.

Keywords: human evaluation, translation, machine translation, post-editing, error typology, error annotation

Sommaire

3.2.7. Conventions-textuelles
3.2.7.1. Coherence_LA-TC-CE
3.2.7.2. Cohesion_LA-TC-CN
3.2.7.3. Type-annotateur_LA-TC-UD
3.2.8. Terminologie-lexique
3.2.8.1. Choix-incorrect-Termino_LA-TL-INS
3.2.8.2. Choix-incorrect-Langue-Generale_LA-TL-ING
3.2.8.3. Mauvais-acronyme-abreviation_LA-TL-MAA
3.2.8.4. Faux-amis_LA-TL-FC
3.2.8.5. Terme-traduit-par-non-terme_LA-TL-NT
3.2.8.6. Collocation-incorrecte-Specialise_LA-TL-ICS
3.2.8.7. Collocation-incorrecte-Langue-Generale_LA-TL-ICG
3.2.8.8. Choix-incompatible-avec-texte-cible_LA-TL-IT
3.2.8.9 Incohérence-terminologique
3.2.8.9.1. Différents-termes-traduction_LA-TL-TI-DT
3.2.8.9.2. Differentes-abbreviations-traduction_LA-TL-TI-DA
3.2.8.10. Type-annotateur_TL-UD

- 3.3. Outils

3.3.1. Hallucination_OU-TAH
3.3.2. Conformite-corpus_OU-CC
3.3.3. Duplication_OU-DU
3.3.4. Choix-incompatible-glossaire_OU-GC
3.3.5. Type-annotateur_OU-UD

1. L'évaluation des traductions humaines, traductions automatiques et post-éditions

1.1. L'évaluation des traductions, une question ancienne

La question de la qualité d'une traduction se confond avec la question de son évaluation. Il s'agit de l'une des préoccupations majeures du domaine de la traduction, de la recherche à l'industrie ; pourtant aucune méthode d'évaluation ne fait consensus entre les acteurs, que ce soit entre les formateurs, les professionnels ou les chercheurs (Secară 2005 ; House 2015), dont les approches diffèrent parfois fortement (Castilho , Doherty et al. 2018, p. 11).

House souligne que l'évaluation dans le domaine de la traduction est profondément liée aux théories traductologiques (House 2011, p. 222) : différentes visions de ce qu'est la traduction conduisent à des perceptions différentes de sa qualité et donc de la manière de l'évaluer. Les critères varient en fonction des théories et des approches traductologiques. Jusque dans la première moitié du XXe siècle, l'analyse de la traduction était avant tout subjective, sans critères explicites ni méthodologie solide (Secară 2005 ; House 2011). Une approche plus scientifique a ensuite émergé dans les années 1960 en lien avec l'émergence de la traduction comme sujet d'étude et d'enseignement (Casagrande 1954). Elle s'inscrivait dans la recherche d'une approche plus méthodique et rigoureuse de l'évaluation.

L'essor de la traduction, comme discipline d'enseignement et pratique professionnelle, s'est accompagné de l'élaboration d'échelles de correction et de notation, qui ont conduit à l'établissement de typologies pour l'analyse des erreurs. Cette pratique reste aujourd'hui largement répandue, car elle fournit un cadre permettant une classification rigoureuse, systématique et économe en ressources (Secară 2005, p. 39).

1.2. Des approches théoriques, professionnelles et pédagogiques qui diffèrent

La recherche en traductologie distingue les approches psychologistes (*response-based approaches*), comprenant les approches comportementales et fonctionnalistes, des approches textuelles ou discursives (*text- and discourse-oriented approaches*), dont font partie les approches linguistiques et descriptives (House 2015, p. 10-14, House 2011).

Les approches psychologistes se concentrent sur la fonction communicative du texte source (House 2015, p. 11). C'est le cas par exemple de la théorie du *skopos* définie par Vermeer (1996), des travaux de Nida (1964) sur l'équivalence dynamique entre la source et la cible ou de l'approche cognitive de Carroll (1966) et de Gutt (2014) pour évaluer la traduction automatique. Elles reposent essentiellement sur la réponse du lecteur au texte

cible en cherchant à déterminer si celle-ci 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, et les critères comme « l'intelligibilité » ou « le caractère informatif » du texte sont souvent imprécis.

Les approches textuelles ou discursives reposent sur une analyse des textes sources et cibles pour identifier les régularités. S'inscrivent par exemple dans cette tendance les travaux de Vinay et Darbelnet (1968),de Mounin (1976) et de Chuquet & Paillard (1989) sur les liens entre traduction et linguistique, les travaux de Toury (1995) sur la description des opérations de traduction ou ceux de Venuti (1995) sur le contexte socioculturel à l'origine des traductions. Dans l'approche linguistique, les textes sources et cibles sont comparés pour mettre en évidence les régularités de transfert entre les deux langues du point de vue syntaxique, sémantique, stylistique ou pragmatique (House 2015, p. 13). En recherche appliquée (comme en traduction automatique) et dans le monde professionnel, les frontières sont plus floues : si les approches linguistiques prédominent pour analyser les erreurs produites, elles font appel, ne serait-ce qu'implicitement, à la finalité des traductions pour évaluer la gravité des erreurs commises. Les typologies d'analyses des erreurs, massivement utilisées dans l'industrie de la traduction, combinent ainsi une prise en compte des erreurs langagières, mais également les attentes du client (guide de rédaction, terminologie propre…) et les contraintes du processus de production (temps passé, délai de livraison…) ; chacun de ces éléments est pondéré en fonction de sa gravité (O'Brien 2012).

Les traductologues se concentrent plutôt sur une approche pédagogique et théorique destinée à la compréhension du processus de traduction ou à la formation des professionnels. Au contraire, l'évaluation en contexte professionnel (Gile 2005) s'inscrit dans une visée pragmatique qui dépasse les seules questions linguistiques (prise en compte du cahier des charges, rapidité de la prestation…) et s'inscrit dans l'évaluation d'un processus (échanges client/fournisseur…), tandis que la recherche en traduction automatique s'inscrit plutôt dans un processus de développement d'une méthode de traitement. Pour les chercheurs de ce dernier domaine, il s'agit de montrer une amélioration ou une dégradation notable de la traduction produite afin de comparer des systèmes, ou de juger de la pertinence d'une adaptation logicielle apportée à un système (Stymne 2018, p. 2-4).

Compte tenu de ces objectifs différents, les typologies d'erreurs utilisées diffèrent, plus ou moins fortement, entre la formation, le monde professionnel ou la recherche en traduction automatique (Secară 2005 ; Popović 2018). Les techniques d'évaluation sont également différentes. Pour cette raison, certaines approches spécifiques aux différentes théories traductologiques ou aux professionnels de l'industrie (ISO 17100, ISO 18587, le LISA QA Model…) ne seront pas développées en détail. Nous n'aborderons que les aspects pertinents pour une évaluation dans le cadre de la recherche académique. On peut en effet noter que l'utilisation croissante des outils d'aide à la traduction et l'essor de la post-

éditionpost-édition brouillent peu à peu la limite entre l'évaluation des systèmes de traductions automatiques et celle des traducteurs humains (Castilho , Doherty et al. 2018, p. 27). Dans ce contexte, l'harmonisation des évaluations fait l'objet d'une recherche active. Par exemple, le projet européen QTLaunchPad[1] visait à créer une métrique d'évaluation applicable à la fois aux traductions humaines et aux traductions automatiques.

1.3. L'évaluation de la traduction automatique

1.3.1. Des approches multiples et complémentaires

Trois grandes classes de techniques sont classiquement utilisées en matière d'évaluation des traductions automatiques en recherche expérimentale. La première consiste à demander à un panel d'évaluateurs monolingues ou bilingues (traducteurs ou experts du domaine) de juger des traductions produites. La deuxième consiste à juger la qualité de façon indirecte en évaluant la performance sur une tâche *downstream*, par exemple *gap filling* (Ageeva et al. 2015). La troisième se base sur des métriques automatiques. L'automatisation de l'évaluation des traductions constitue un domaine de recherche à part entière depuis les années 1990. Elle s'inscrivait à l'époque dans une volonté d'obtenir une technique d'évaluation répétable, rigoureuse et à faible coût dans un contexte où l'industrie de la traduction automatique prenait son essor (Castilho 2019 ; Poibeau 2017, p. 133).

En effet, malgré de nombreuses tentatives d'encadrement, l'évaluation humaine reste profondément subjective, coûteuse, chronophage et non répétable (Castilho 2019 ; Stymne 2018 ; Way 2018). Il est notamment difficile d'arriver à un accord entre les annotateurs sur les erreurs identifiées. Pour essayer de pallier ce problème, plusieurs méthodes ont donc été développées pour essayer de déterminer la fiabilité de scores et l'accord effectif entre les évaluateurs ou annotateurs. L'interprétation des scores inter annotateur ainsi produits fait cependant débat. De plus, les compétences des évaluateurs jouent un rôle essentiel dans l'évaluation humaine, ce qui complexifie leur recrutement : si l'évaluation de la grammaire ou de la fluidité est relativement simple, l'évaluation de l'adéquation et de la fidélité au texte source implique des connaissances bilingues approfondies (Castilho 2019). Pour finir, même lorsque des consignes précises sont données, le résultat de cette évaluation manuelle est difficilement reproductible en raison des grandes variations entre les notes données par les évaluateurs. L'évaluation humaine présente cependant l'avantage d'être plus efficace pour évaluer les phénomènes linguistiques complexes.

Les évaluations humaines et automatiques regroupent chacune des techniques différentes d'évaluation. Le recours à l'une ou l'autre des techniques dépend de ce qui est évalué et des ressources disponibles.

1 https://cordis.europa.eu/project/id/296347

Castilho (2019) a ainsi synthétisé en un arbre de décision les conditions menant à utiliser une technique plutôt qu'une autre (voir la figure 1) : la distinction principale repose sur la volonté de recourir ou non à des humains lors de l'évaluation de la qualité. Les métriques automatiques (*automatic evaluation metrics*) utilisées dans le domaine de la recherche se distinguent du contrôle qualité tel que pratiqué dans le milieu professionnel (*Quality Assessment ou Quality Assurance*, noté QA) principalement par l'utilisation d'une traduction de référence. La portée des évaluations humaines est plus large, permettant notamment d'intégrer l'utilisabilité effective des textes traduits par les utilisateurs finaux (*usability*) ou d'obtenir une vision fine de la performance du système par une catégorisation des erreurs produites (*error typology*).

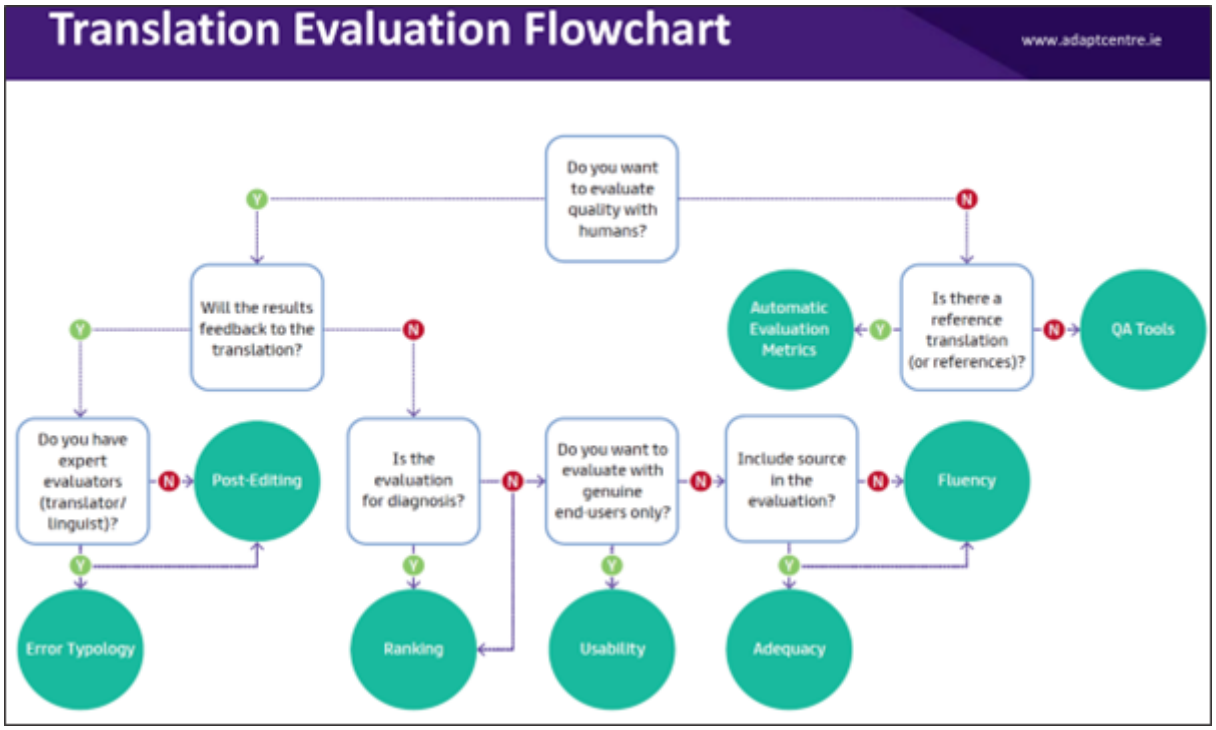


Figure 1 – Graphe de répartition entre les différentes techniques d'évaluation de la traduction automatique (Castilho 2019, p. 11)

Compte tenu des besoins d'évaluation du domaine, les campagnes d'évaluation de la TA se sont multipliées depuis le début des années 2000. La plus importante est organisée tous les ans depuis 2006 par le *Workshop on Machine Translation* (WMT) et depuis 2016 par la *Conference on Machine Translation*. Celle-ci comporte également un workshop sur l'évaluation des nouvelles métriques (WMT Metrics Task). Celle-ci porte sur plusieurs paires de langues dans lesquelles l'anglais est toujours impliqué (français-anglais, espagnol-anglais, allemand-anglais, tchèque-anglais et russe-anglais). Ces dernières années, le chinois a été bien représenté, mais selon les pistes, il y a aussi eu des langues plus rares (khmer, tamil, pashto, inuktitut, etc.).

Le présent rapport évoque brièvement les métriques automatiques, pour plus de détail voir le livrable 4.3 "Survey of existing language-based MT evaluation metrics and of document-level metrics" du projet MATOS (https://anr-matos.github.io/).

1.3.2.Les évaluations humaines

Les évaluations humaines représentent les types d'évaluation les plus anciens. Ces évaluations sont plus précises que celles qui sont basées sur des métriques automatiques, car elles permettent d'identifier précisément les forces et les faiblesses des systèmes sur des phénomènes linguistiques précis. Elles s'inscrivent alors dans des approches linguistiques et typologiques à travers l'identification, le dénombrement et la catégorisation des erreurs. Certaines évaluations humaines permettent également de prendre en compte le ressenti des utilisateurs finaux sur l'utilisabilité effective des traductions produites. On retrouve ainsi des approches plus fonctionnalistes centrées sur la fonction ou la visée du texte source ou cible.

L'évaluation la plus fréquente repose sur l'analyse de la fidélité, *accuracy* en anglais (le respect du sens du texte source), et de la fluidité, *fluency* (le respect des règles de la langue cible et de son caractère idiomatique), à l'échelle du segment ou de la phrase. Des évaluations peuvent également être menées sur la base de l'utilisabilité et de l'acceptation de la traduction par l'utilisateur final. La lisibilité, *readability* (la facilité de lecture liée à la rédaction même), et la compréhension du texte, *intelligibility* (la facilité de lecture liée au lecteur visé), peuvent également être évaluées. Il est par ailleurs possible de classer plusieurs phrases du texte cible sur la base de critères ou de concepts prédéfinis. Cette dernière approche est particulièrement utilisée dans les campagnes internationales en raison de sa rapidité et de son efficacité (Castilho, Doherty et al. 2018).

Il est généralement demandé aux évaluateurs de noter les traductions sur une échelle de 1 à 4 ou de 1 à 5, dans laquelle 1 indique une traduction de très mauvaise qualité. À l'origine, des échelles de valeur plus grande étaient utilisées (notation de 1 à 10 par exemple, voire de 1 à 100 (Graham et al., 2016: 3126), mais il s'est avéré que les évaluateurs éprouvent des difficultés à noter avec une grande granulométrie. De plus, une échelle plus restreinte facilite l'atteinte d'un accord entre les évaluateurs (Way 2018, p. 164). Il est également possible de demander aux évaluateurs de classer les traductions d'un même segment (*Relative Ranking*) afin de comparer des moteurs (Popović 2018, p. 130).

Jusqu'à l'apparition de la traduction automatique neuronale, une note élevée en fluidité était corrélée à une fidélité au texte source. Les progrès récents liés à la traduction neuronale bouleversent cette analyse : il est désormais possible pour un système de produire une phrase parfaitement formée, mais qui ne correspond pas à la traduction attendue du texte source. Way souligne que cette situation incite à revoir l'évaluation de

la traduction neuronale, car la fluidité observée peut conduire les post-éditeurspost-éditeur à accepter une traduction pourtant erronée :

> *« [Neural] MT output can be deceptively fluent; sometimes perfect target language sentences are output, and less thorough translators and proofreaders may be seduced into accepting such translations, despite the fact that such translations may not be an actual translation of the source sentence at hand at all! » (Way 2018, p. 164)*

De plus, les scores produits ne permettent pas d'évaluer l'effort de post-édition à fournir pour atteindre une traduction correcte. Ils n'indiquent pas non plus quelles sont les réussites ou les erreurs effectives d'un moteur de traduction.

On retrouve ici des problématiques similaires à celles des métriques automatiques. Les techniques d'analyse et de classification des erreurs semblent à même de répondre à ces difficultés. Elles permettent de relever précisément les points faibles et les points forts d'un moteur, les difficultés spécifiques de post-édition ou les subtilités de performance entre les moteurs. C'est pourquoi une évaluation basée sur des tâches précises (Way 2018, p. 164 et p. 170) ou une analyse des erreurs (Popović 2018, p. 131) semblent plus appropriées pour une évaluation fine de la traduction automatique neuronale.

Il faut noter que, dans le domaine de la recherche en TA, le recours à des évaluateurs professionnels semble l'exception, probablement en raison de contraintes financières. L'usage semble être de faire appel à des étudiants ou à des évaluateurs amateurs. De manière générale, la compétence linguistique des évaluateurs ou leur expertise sur le type de texte évalué ou la tâche elle-même ne sont pas souvent détaillées dans les études, alors même qu'aucune formation n'est généralement dispensée. Cette approche soulève de nombreuses interrogations, notamment sur la possibilité de biais et le manque de fiabilité des analyses (Doherty 2017, p. 11) et la différence de jugements entre experts et profanes ou non-experts (Poibeau 2022).

1.3.3.Des métriques automatiques qui doivent évoluer

Les métriques automatiques sont historiquement des algorithmes calculant la similarité entre une traduction à évaluer, dite « hypothèse de traduction », et une ou plusieurs traductions de référence. Il existe aussi des métriques sans références (qui se basent uniquement sur une comparaison entre la traduction automatique et le texte source), par exemple Quality Estimation (QE) metrics sous forme de CometKiwi (Rei et al. 2022). Cette section présente brièvement le premier type de métrique ; pour une typologie plus complète des métriques, voir le livrable 4.3 "survey of existing language-based MT evaluation metrics and of document-level metrics" du projet Matos.

Pour les métriques impliquant la confrontation de la TA avec une ou plusieurs traductions de référence, la similarité observée est traduite par un score de qualité compris généralement entre 0 (aucune correspondance) et 1 (similarité parfaite). Plus la

traduction fournie par la machine se rapproche d'une traduction humaine, plus elle est jugée de bonne qualité. Aucune métrique ne répondant à elle seule aux différents besoins d'évaluation, les propositions de métriques se sont donc multipliées depuis vingt ans. La plus populaire reste *Bilingual Evaluation Understudy* (BLEU), développée en 2002 par Papineni, Roukos, Ward et Zhu. Le score BLEU repose sur le dénombrement des segments identiques entre le texte évalué et la traduction de référence. Par ailleurs, les métriques automatiques sont considérées comme plus objectives et répétables que les évaluations humaines, car elles limitent l'intervention humaine à la production des traductions de référence et les résultats obtenus sont identiques, quel que soit l'opérateur. De plus, elles ne nécessitent pas d'évaluateurs bilingues ou spécialisés, ce qui en réduit les coûts. Néanmoins, cette objectivité n'est que relative : le nombre de traductions de référence utilisées étant limité, la comparaison ne tient pas compte de l'ensemble des possibilités effectives de traduction (Doherty 2017, p. 5 ; Castilho, Doherty et al. 2018, p. 26).

Toutefois, l'essor des systèmes neuronaux vient remettre en cause la prédominance des métriques automatiques comme BLEU ou METEOR (Banerjee et Lavie, 2005) : le score défini ne permet plus d'identifier précisément les difficultés des systèmes. Il est aujourd'hui reconnu que les métriques automatiques classiques ne sont plus adaptées à l'évaluation des systèmes neuronaux et que de nouvelles techniques d'évaluation doivent être développées (Burlot et Yvon 2018). Burlot et Yvon distinguent ainsi quatre grandes familles récentes en matière de diagnostic (*ibid.*) :

- Deux familles basées sur l'analyse directe de la traduction produite par le système de TA :
  - Une catégorisation des erreurs, manuelle ou automatique, dans les sorties des systèmes ;
  - Une analyse fondée sur des jeux de test (*test suites* en anglais) visant à analyser la capacité de traduction d'un système sur un problème linguistique particulier ;
- Deux familles basées sur une analyse indirecte :
  - Une évaluation basée sur les scores donnés par le système et non la traduction qu'il produit : la qualité du système est alors liée à sa capacité à attribuer un meilleur score à une phrase correcte par rapport à une phrase erronée ;
  - Une comparaison des plongements lexicaux, ou enchâssement de mots (*word embedding* en anglais), appris par le système de traduction au regard de leur capacité à prédire les traductions (par exemple, COMET, BertScore, etc.). L'analyse ne porte plus sur les textes produits, mais sur la vectorisation des mots (la représentation mathématique des mots) effectuée par le système lors de son apprentissage. Cette technique repose sur l'idée que les mots apparaissant dans des contextes similaires

possèdent des vecteurs (c'est-à-dire une représentation mathématique) qui sont relativement proches.

Parmi ces quatre familles, l'analyse directe des erreurs avec ses typologies et ses méthodes d'analyse apparaît comme la technique la plus appropriée pour évaluer finement les performances des systèmes neuronaux (Burlot et Yvon 2017). De nombreux chercheurs tentent de développer de nouvelles métriques automatiques menant une analyse linguistique plus approfondie. C'est le cas des travaux de Burlot et Yvon sur l'évaluation des compétences morphologiques des systèmes de TA (Burlot et Yvon 2017 ; Burlot et Yvon 2018). Certains chercheurs (Giménez et Màrquez 2007 ; Wong et Kit 2012 ; Guzmán et al. 2015) tentent également de développer des métriques automatiques assurant une évaluation à l'échelle du texte entier pour en vérifier la cohérence. Ces dernières années, quelques études ont été menées pour identifier les difficultés linguistiques propres aux systèmes neuronaux. Leur ambition est de se concentrer sur des phénomènes linguistiques spécifiques (lexicaux ou grammaticaux), afin de constater la façon dont ils sont gérés par les différents systèmes, puis de les comparer à des traductions humaines ou à des textes directement rédigés dans la langue source (Loock 2018). Cette approche contribue à l'amélioration ou à la sélection d'un système en fonction de la tâche envisagée (Macketanz et al. 2017). Par exemple, dans le cas d'un système dédié à la post-édition, il est possible de se concentrer sur les points particulièrement difficiles à post-éditer. Dans le cas d'un système dédié à fournir des informations à des utilisateurs finaux, l'accent pourra au contraire être mis sur la lisibilité du produit final. Une telle priorisation n'est pas possible avec les métriques automatiques existantes aujourd'hui, bien que l'on constate une tendance récente qui consiste à évaluer la TA avec des Grands Modèles de Langue (LLMs) en demandant au système d'identifier et d'expliquer les erreurs - par exemple GEMBA et GEMBA-MQM (Kocmi et Federmann 2023). Malgré les recherches en cours, les évaluations humaines restent encore les plus adaptées pour une analyse fine des traductions automatiques.

1.4. Conclusion

Nous avons exposé en quoi les évaluations humaines classiques basées sur la seule fluidité et les métriques automatiques classiques ne sont plus appropriées à l'évaluation de la qualité des traductions produites par les systèmes neuronaux. Les évaluations basées sur des phénomènes linguistiques précis ou des tâches spécifiques avec des jeux de tests apparaissent comme une réponse plus appropriée (Way 2018 ; Burlot et Yvon 2018). Ces approches impliquent de mener une analyse des erreurs constatées. Généralement, l'analyse des erreurs vise soit à obtenir un profil et une distribution des types d'erreurs pour une traduction donnée, soit à comparer des systèmes en étudiant la distribution des erreurs dans les traductions produites. En traduction automatique, des analyses plus spécifiques peuvent également être menées comme le lien entre certains types d'erreurs et les préférences de post-édition ou l'incidence des différents types d'erreurs sur l'effort de post-édition (Popović 2018, p. 131).

Historiquement, ces analyses étaient réalisées manuellement. Compte tenu de la non-répétabilité des évaluations humaines et de leur caractère chronophage, la recherche de systèmes d'analyse automatiques est particulièrement dynamique depuis les années 2000. Les métriques développées (WER, PER, Addicter…) reposent sur une comparaison entre une ou plusieurs traductions de référence et la traduction à évaluer. Elles sont capables d'identifier un grand nombre de types d'erreurs mises en évidence par l'évaluation humaine : les erreurs de flexions, d'ordre des mots, d'omissions, d'ajouts ou les non-traductions d'un terme. Les systèmes automatiques d'analyse des erreurs présentent cependant le même défaut que les métriques automatiques : le nombre de traductions de référence utilisées étant limité, la comparaison ne tient pas compte de l'ensemble des possibilités effectives de traduction (Popović 2018, p. 142-43).

Il faut cependant noter que l'évaluation manuelle permet de distinguer un plus grand nombre de types d'erreurs que les systèmes automatiques disponibles aujourd'hui. De plus, les systèmes automatiques sont plus enclins à confondre certains types d'erreurs que les évaluateurs humains (*loc. cit.*). Toutefois, ces systèmes peuvent être utiles dans le cas d'un prétraitement des erreurs qui sera ensuite amendé par un évaluateur humain. Cette approche semi-automatique présente par exemple un intérêt dans le cadre de la post-édition (Popović 2018, p. 154).

2. Analyse des typologies d'erreurs existantes

2.1. Les typologies d'erreurs dans la recherche en traduction automatique

En se basant sur le schéma de classification de Llitjós et al. publiés en 2005, David Vilar et al. ont proposé, en 2006, une typologie d'erreurs de traduction à trois niveaux pour l'analyse de traductions statistiques du chinois vers l'anglais, de l'espagnol vers l'anglais et de l'anglais vers l'espagnol (Vilar et al. 2006). Cette typologie d'erreurs (voir la figure 2 ci-dessous) identifie cinq catégories principales d'erreurs de traduction (Vilar et al. 2006 in Esperança-Rodier et Becker 2018) :

— « Mots manquants », pour les mots qui n'ont pas été traduits : elle permet de vérifier si le sens complet de la phrase a été conservé ou non ;
— «Ordre des mots», pour un mauvais ordre des mots dans la séquence traduite : elle indique si l'erreur de traduction entraîne une réorganisation des mots eux-mêmes ou un réordonnancement des segments dans la phrase. Elle permet de localiser à quel niveau, lexical ou sémantique, le système a échoué ;
— « Mots incorrects », pour une erreur de traduction : elle vise à distinguer la raison de l'erreur de traduction ;
— « Mots inconnus », pour les mots restés dans la langue source : elle permet de distinguer si le lemme des mots était connu par le système ou non ;
— « Ponctuation », lorsque les règles de ponctuation de la langue cible n'ont pas été respectées.

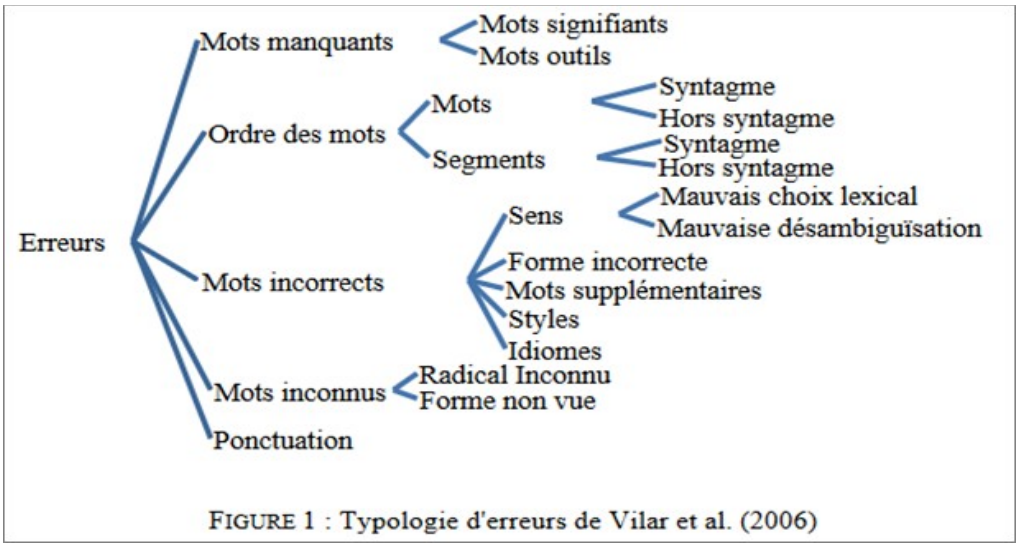


Figure 2 – Traduction de la typologie d'erreurs de Vilar et al. (2006) par Esperança-Rodier et Becker 2018

Cette typologie d'erreurs regroupe cependant sous un même type d'erreur des occurrences lexicales, morphologiques et syntaxiques très différentes. Conscients de cette limite, Emmanuelle Esperança-Rodier et Nicolas Becker (2018) envisageaient de proposer une nouvelle typologie. En 2022, Esperança-Rodier publie avec Damien Hansen une nouvelle typologie d'erreurs pour évaluer manuellement les traductions automatiques d'œuvres de fiction (voir Figure 3). Cette nouvelle typologie se démarque profondément de celle de Vilar et al. (2006). D'un côté, elle prend en compte des erreurs bien connues des systèmes de traduction automatique neuronale telles que "stuttering" (la répétition d'un même mot sans raison) ou "hallucination" (la génération de contenus qui n'ont pas de sens et qui sont infidèles au texte source (Ji et al. 2023). De l'autre, elle se détache de la classification en arborescence propre à la majorité des typologies d'erreurs en ne proposant aucun regroupement des erreurs identifiées. Outre les travaux d'Esperança-Rodier et Nicolas Becker et Hansen & Esperança-Rodier, plusieurs autres typologies d'erreurs ont été développées[2] depuis 2006.

Farrús Caberceran et al. (2010) développent une typologie d'erreurs simple, comprenant un seul niveau et en cinq classes, pour l'évaluation bidirectionnelle de traduction espagnol-catalan. Ils observent que ce sont les erreurs lexicales et sémantiques qui influencent le plus la perception des évaluateurs sur la qualité des traductions. En 2014, Federico, Negri, Bentivogli et Turchi développent une typologie d'erreurs similaires pour l'analyse des traductions de l'anglais vers l'arabe, le chinois ou le russe. Ils ont pour objectif de vérifier l'impact du type d'erreurs sur la perception générale de la qualité d'un texte. Ils constatent ainsi que la corrélation la plus importante se fait avec les erreurs lexicales ou les mots manquants dans la traduction. Ils montrent également que la perception de qualité ne dépend pas forcément de la fréquence d'une erreur donnée. Des typologies similaires ont été utilisées par différents chercheurs, dont Castilho, les années suivantes pour comparer les résultats de traduction basée sur les fragments et ceux de traductions neuronales (Popović 2018).

En 2012, Kirchhoff, Capurro et Turner proposent un schéma d'annotation (voir la table 1) dans le cadre d'une étude qui, sur la base de traductions automatiques de l'anglais vers l'espagnol, vise à comparer les annotations d'erreurs à une estimation de la qualité de chaque TA établie par conjointement par plusieurs utilisateurs par *crowdsourcing*. Leur étude montre par ailleurs l'importance donnée aux erreurs sémantiques et d'ordre des mots, tandis que la fréquence de certaines erreurs comme les erreurs morphologiques ne semblent pas influencer la perception de qualité.

La même année, Stymne et Ahrenberg sont les premiers à prendre en compte le taux d'accord inter-annotateur dans la classification des erreurs (Stymne et Ahrenberg 2012). Ils développent également une typologie d'erreurs détaillée à deux niveaux. Cette typologie a

2 Maja Popovic en a résumé un grand nombre dans un article détaillé publié en 2018 (Popović 2018) dont nous partageons les analyses. Ces principales typologies sont présentées dans les pages suivantes.

été testée sur des traductions de l'anglais vers le suédois. L'analyse a été réalisée par des Suédois natifs en deux fois, une première fois sans exemples de classification (seule la typologie était fournie) et une seconde fois avec des exemples et des consignes. L'étude a montré que dans le premier cas, le taux d'accord inter-annotateur 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. Cette étude a mis en valeur l'intérêt de fournir des consignes et des exemples en accompagnement de la typologie.

Ces différents travaux ont permis de mettre en évidence les limites à l'utilisation des typologies d'erreurs et le rôle que jouent certaines catégories dans l'évaluation de la qualité générale d'une traduction. Ainsi, une catégorie classique comme « Ordre des mots » (« *Word order* ») pourrait correspondre à des phénomènes très différents pour l'analyse d'un syntagme nominal complexe comme une mauvaise attribution d'un modifieur ou une mauvaise identification de la tête (Kübler et al. 2022). Or, cette catégorie est présente de manière directe ou indirecte (avec des catégories comme « *Misordering* » ou « *Reordering errors* ») au niveau le plus fin des typologies présentées précédemment. De même, les erreurs portant sur les erreurs lexicales sont abordées dans ces typologies soit de manière très générale (« *Lexical errors* », « *lexical choice* »), soit par l'ajout ou l'omission de mots (« *extra words* », « *missing word* », « *Omission* », « *Addition* », « *Untranslated* », etc.), et sous son aspect sémantique (« *semantics* », « *confusion of senses* ») uniquement dans deux typologies (Costa et al. 2015 ; Stymne et Ahrenberg 2012). Or, en langues de spécialité, une erreur portant sur un terme n'aura probablement pas le même impact qu'une erreur portant sur un mot relevant de la langue générale. Cette distinction permettrait également de vérifier l'efficacité de la spécialisation d'un système dans le domaine considéré. À l'exception de la typologie de Costa et al. (*ibid*.) et de celle de Hansen & Esperança-Rodier (2022), aucune autre typologie ne semble prendre en compte les questions de style et de variation qui sont pourtant importantes en langue s de spécialité.

Omission (OMI) 21.6% of all errors
Word or idea that is present in the source but not in the target (even if also omitted in the reference).

Addition (ADD) 0.5% of all errors
Word or idea that is present in the target but not in the source.

Overtranslation (OTR) 0.6% of all errors
Translation that is correct but that leads to a redundancy or adds a nuance in the target that was not present in the source.

Undertranslation (UTR) 3.0% of all errors
Translation that is correct but that leads to the loss of a nuance in the target.

Stuttering (STU) 0.2% of all errors
Words repeated for no apparent reason by the MT system.

Hallucination (HAL) 1.0% of all errors
Completely illogical fragments of sentence added or replaced in the target; invented terms due to subword segmentation.

Non-translation (NTR) 0.0% of all errors
Source term left untranslated in the target.

Mistranslation (MTR) 16.0% of all errors
Terms or syntagmas wrongly translated or translated with the wrong sense.

Opposite meaning (OME) 0.9% of all errors
Translation leading to a meaning that is contradictory to the source.

Nonsense (NON) 2.0% of all errors
Translation that does not make any sense.

Shift in meaning (SME) 3.4% of all errors
Translation leading to a different meaning than the one expressed in the source.

Referential cohesion (REF) 2.5% of all errors
Break in the logical relation of co-referring items (e.g. anaphora): pronoun resolution, lack of antecedent, lexical choice...

Relational cohesion (REL) 0.2% of all errors
Break in the logical articulation and flow of a text's sentences or clauses.

Logic (LOG) 4.8% of all errors
Sentence that is grammatically correct but syntactically or semantically inaccurate based on the source sentence or story.

Gender (GEN) 1.5% of all errors
Any issue related to a grammatical or character gender (excluding pronoun resolution).

Number (NUM) 0.5% of all errors
Any issue related to agreement based on grammatical number.

Tense (TEN) 4.1% of all errors
Wrong tense; problem with the sequence of tenses.

Person (PER) 0.6% of all errors
Subject-verb agreement.

Function words (FUN) 4.3% of all errors
Mistranslation of a determiner, preposition, etc. (anything but content words).

Punctuation (PUN) 1.6% of all errors
Any punctuation issue, with the exception literary-specific typographic conventions (e.g. related to dialogues).

Style (STY) 17.4% of all errors
Literal translation, repetition, unnatural wording or collocation, translation that makes the text more difficult to understand...

Register (REG) 3.0% of all errors
Confusion between the French "*tu*" and "*vous*"; character speaking with an inappropriate tone.

Unfitting paraphrase (PAR) 0.6% of all errors
Term rendered by an equivalent syntagma but leading to a ponderous or poor translation in the given context.

Adaptation (ADA) 1.2% of all errors
Text fragment that requires a particular translation solution due to language differences, cultural context, formal constraints...

Coherence with previous volumes (COH) 2.3% of all errors
Translation of a common or series-specific term that is not in line with previous volumes.

Loss (LOS) 0.8% of all errors
Term that is specific to the series's universe and translated with a flat and neutral instead of an original or obligatory solution.

Dialogues (DIA) 3.5% of all errors
Problem tied to the typographic conventions of the series.

Case (CAS) 1.6% of all errors
Translation going against previous choices regarding the capitalization of series-specific terms (always capitalized in English).

Figure 3 - Typologie d'erreurs de traduction automatique littéraire développée par Hansen et Esperança-Rodier (2022)

Typologie de Farrús et al. (2010)

morphological errors
lexical errors
orthographic errors
syntactic errors
semantic errors

Typologie de Federico et al. (2014)

morphological errors
lexical choice
additions
omissions
casing and punctuation
reordering errors
too many errors

Typologie de Kirchhoff et al. (2012)

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	
capitalization	
untranslated	medical term
	proper name
	other
pragmatics	
diacritics	
other	

Typologie de Stymne et Ahrenberg (2012)

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)	

Typologie de Costa et al. (2015)

level 1	level 2	level 3
Orthography	Punctuation	
	Capitalization	
	Spelling	
Lexis	Omission	
	Addition	
	Untranslated	
Grammar	misselection	Word Class
		Verbs
		Agreement
		Contraction
	Misordering	
Semantic	Confusion of senses	
	Wrong choice	
	Collocational errors	
	Idioms	
Discourse	Style	
	Variety	
	Should not be translated	

Table 1 – Principales typologies d'erreurs dans le domaine de la recherche en traduction automatique telles que schématisées par Popović (2018)

D'autres travaux se sont penchés sur l'élaboration d'une typologie portant sur des difficultés linguistiques précises, sources potentielles d'erreurs lors de la traduction par les moteurs de TA (traduction des pronoms de l'anglais vers le français…). Ces phénomènes sont généralement spécifiques à une paire de langues, voire à une direction de traduction, et vont bien au-delà des classes d'erreurs classiques et présentées au chapitre précédent (Popović 2018).

Par exemple, dans le cadre d'une analyse des erreurs en post-édition, Blain et al. (2011) se sont intéressés aux erreurs concernant les syntagmes nominaux. Pour cela, ils ont développé une typologie à un niveau comprenant huit types d'erreurs (voir Table 2 ci-après). Cette typologie présente un intérêt certain : les erreurs liées aux concurrents semblent être prises en compte à travers les erreurs « *Noun stylistic change* » (les concurrents terminologiques dont les synonymes ou quasi-synonymes, ne sont pas interchangeables dans tous les contextes) et « *NP structure change* » (un terme ne peut être confondu avec un groupe et son expansion).

Noun-Phrase (NP) — related to lexical changes.

- Determiner choice — change in determiner
  SRC: *enable a drawing preview of the DWG overlay*
  TGT: *activer l'aperçu du dessin de la superposition DWG*
  PE : *activer un aperçu du dessin de la superposition DWG*
- Noun meaning choice — a noun is replaced by another noun **changing its meaning**
  SRC: *the border displays as stripes*
  TGT: *la bordure s'affiche sous forme de rayures.*
  PE : *la bordure s'affiche sous forme de bandes.*
- Noun stylistic change — a noun is replaced by a synonym (no meaning change)
  SRC: *[…]that placing[…]*
  TGT: *[…]que le placement[…]*
  PE : *[…]que le positionnement[…]*
- Noun number change
  SRC: *[…]their proper locations*
  TGT: *[…]leur_ emplacement_ approprié_*
  PE : *[…]leurs emplacements appropriés*
- Case change
- Adjective choice — change in adjective choice for better fit with modified noun
  SRC: *[…]regardless of the active project.*
  TGT: *[…]quel que soit le projet en cours.*
  PE : *[…]quel que soit le projet actif.*
- Multi-word change — multiword expression change (meaning change)
  SRC: *credit card*
  TGT: *carte bancaire*
  PE : *carte de crédit*
- NP structure change — structure change of NP but the sense is preserved
  SRC: *preview color*
  TGT: *couleur de l'aperçu*
  PE : *couleur d'aperçu*

Table 2 – Extrait de la typologie de Blain et al. (2011)

Cette typologie présente cependant des défauts pour les langues de spécialité : les erreurs « *Multi-word change* » et « *Noun meaning choice* » comprennent en réalité des erreurs

d'ordre terminologique qui devraient être fusionnées. Ainsi, une « carte de crédit » ne correspond pas à la même notion qu'une « carte bancaire » dans le domaine bancaire. En se basant sur les définitions de la base TERMIUMPlus[3], une carte de crédit est une « [c]arte émise par un établissement financier ou une société, qui permet à son titulaire de régler, sans versement immédiat, le paiement de biens et de services », tandis qu'une carte bancaire est une « [c]arte de crédit émise par une banque à charte, une compagnie de fiducie, une coopérative de crédit ou une caisse populaire » ; le second terme est donc un hyperonyme du premier. Nous sommes donc face à une erreur terminologique qui aurait pu concerner un terme composé d'un seul mot. L'erreur n'est donc pas liée au fait qu'il s'agisse d'un terme composé, mais au fait qu'il s'agisse d'un terme du domaine bancaire.

Popović donne ainsi l'exemple des difficultés de traduction résultant de la structure des verbes allemands qui peuvent se traduire par les erreurs classiques de « ordre des mots », « mots [verbes] manquants » ou de « mauvaises traductions ». Parmi ces approches, on peut citer celle de Comelles et al. (2012). VERTa (voir la Figure 4 ci-dessous) est une métrique automatique, motivée linguistiquement, qui combine des connaissances linguistiques à différents niveaux (lexical, morphologique, syntaxique et sémantique). Cette typologie étend celle plus basique proposée par Farrús et al. (*op.cit.* p. 130).

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 4 – Typologie de Comelles et al. (2012) extraite de Popović (2018)

Isabelle, Cherry et Forster (2017) ont également proposé une approche linguistique des problèmes de traductions automatiques par fragments et de traductions neuronales, de l'anglais vers le français. Il s'agissait d'identifier manuellement les faiblesses des systèmes neuronaux. Pour cela, un corpus a été spécifiquement créé, un jeu de test ou *challenge set*, avec 100 phrases contenant chacune un exemple de phénomène linguistique connu pour

3 Dernière consultation le 26 septembre 2023

poser des difficultés de traduction (divergences morphosyntaxiques, lexico-syntaxiques ou syntaxiques). Les erreurs ont donc été classées en fonction de ces trois grandes catégories.

Ces typologies mettent en évidence l'intérêt de construire une terminologie ad-hoc pour un phénomène linguistique afin d'évaluer la performance des systèmes de traduction automatique.

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 5 – Typologie de Isabelle, Cherry et Forster (Isabelle et al., 2017)

Nous constatons donc qu'aucune typologie d'erreurs ne semble faire l'unanimité dans le domaine de l'évaluation de la traduction automatique. Cela peut s'expliquer par la difficulté à prendre compte l'ensemble des erreurs liées à un phénomène linguistique précis ou une paire de langues données.

2.2.Les typologies d'erreurs, évaluation humaine et métriques chez les professionnels de la traduction

L'évaluation en contexte professionnel s'inscrit dans une visée pragmatique qui dépasse les seules questions linguistiques (prise en compte du cahier des charges, rapidité de la prestation…) et s'inscrit dans l'évaluation d'un processus (échanges client/fournisseur…). En raison de cette complexité, de nombreuses métriques, typologies d'erreurs de traduction et standards d'évaluation sont utilisés, parfois même au sein d'une même entreprise, en fonction des projets ou des clients. Au regard de l'importance que présente cette question pour les professionnels, une norme ISO[4] portant sur l'évaluation des traductions est en préparation depuis 2020 et à l'étape de son approbation depuis août 2023. Cependant, à la date de rédaction de ce document, cette norme n'est toujours pas publiée et n'est donc pas consultable.

La SAE J2450 est une métrique de qualité développée par la Society of Automotive Engineers en collaboration avec General Motors. L'objectif était de développer une métrique de qualité applicable à toutes les productions linguistiques de l'industrie automobile, et ce, quelle que soit la langue ou la méthode de traduction utilisée. Cette métrique est applicable depuis 2001 et sa dernière révision remonte à 2016[5]. Secară a détaillé et commenté la version initiale et la révision de 2005 (Secară 2005) : cette métrique repose sur 7 catégories d'erreurs nommées *Wrong Term*, *Syntactic Error*, *Omission*, *Word structure or Agreement Error*, *Misspelling*, *Punctuation Error* et *Miscellaneous Error*. Chacune de ces catégories se voit attribuer un poids déterminant son niveau de gravité. Secară souligne que cette métrique est fortement orientée vers une évaluation terminologique des traductions et qu'elle ne serait donc pas appropriée à un domaine où les erreurs de grammaire, d'orthographe ou de style seraient également à prendre en compte. Cette limitation semble toujours présente dans la version révisée de 2016. Ainsi, la page officielle de la métrique indique « *[…] the current version of the metric does not measure errors in style, making it unsuitable for evaluations of material in which style is important (e.g., owner's manuals or marketing literature)* ».

Trois métriques se distinguent dans le monde professionnel en raison de leur diffusion et de l'accent mis sur la détection des erreurs de traduction et leur sévérité : la typologie *Dynamic Quality Framework* (DQF), la *Multidimensional Quality Metrics* (MQM) et la *LISA Quality Assurance* (QA).

La métrique développée par la Localisation Industry Standards Association (LISA) reste très utilisée en entreprise, quand bien même elle n'a plus été mise à jour depuis 2011, car de nombreuses métriques internes ad-hoc ont été développées à partir de celle-ci (Castilho , Doherty et al. 2018). Cette métrique a été développée dans les années 1990 pour répondre aux besoins des industries de développement logiciel dans le but de maintenir la qualité et la précision du processus de révision. Elle repose sur une catégorisation des

4 . https://www.iso.org/standard/80701.html

5 . https://www.sae.org/standards/content/j2450_201608

erreurs en 20, 25 ou 123 catégories en fonction des besoins des évaluateurs et des exigences du client, auxquelles est associé un système de points. Le nombre maximum de points d'erreur autorisés est basé sur le nombre de mots traduits (généralement moins de 1% du nombre total de mots). Trois niveaux de gravité d'erreur sont pris en compte : critique, majeur et mineur. Plus l'erreur est grave, plus le nombre de points attribués à la gravité de l'erreur est élevé. La métrique d'origine ne semble cependant plus répondre aux exigences des professionnels et d'autres métriques ont été développées depuis les années 2010, dont la typologie *Dynamic Quality Framework / Multidimensional Quality Metrics*.

La Dynamic Quality Framework / Multidimensional Quality Metrics (DQF / MQM) résulte de la fusion de deux systèmes d'évaluation pré-existants : la Multidimensional Quality Metrics développée par le laboratoire DFKI sur la base de la LISA QA, dans le cadre du projet européen QT Launch Pad, et la Dynamic Quality Framework développée par la société TAUS. Son intégration dans l'outil Trados depuis 2016 fait que cette métrique est régulièrement utilisée en milieu professionnel. Elle a été développée dans le cadre du projet QT21, financé par l'Union européenne de 2015 à 2018. QT21 s'attaquait aux barrières linguistiques qui entravaient la circulation de l'information en Europe. Cette typologie a été conçue dans le souci de proposer un standard pour évaluer une traduction, quel que soit son mode de production, dans un contexte professionnel (Castilho, Doherty et al. 2018). Cette typologie a ensuite été intégrée au sein de la MQM, puis remplacée par la MQM-Core[6] fin 2021. Cette dernière présente des évolutions pertinentes en matière de traitement des erreurs terminologiques ou des erreurs de fidélité au texte source. Par exemple, la sous-catégorie *Do not translate* apparaît désormais clairement dans *Accuracy* et le fait que le mauvais terme puisse avoir été utilisé est pris en compte par la catégorie *Wrong term* dans *Terminology*. Depuis, la typologie MQM est régulièrement remise à jour. Elle a connu une dernière actualisation en 2024 et comprend désormais huit grandes catégories : *terminology*, *accuracy*, *linguistic conventions*, *style*, *locale conventions*, *audience appropriateness* et *design and markup* (voir Figure 6), qui se déclinent ensuite en sous-catégories.

6 `https://themqm.org`

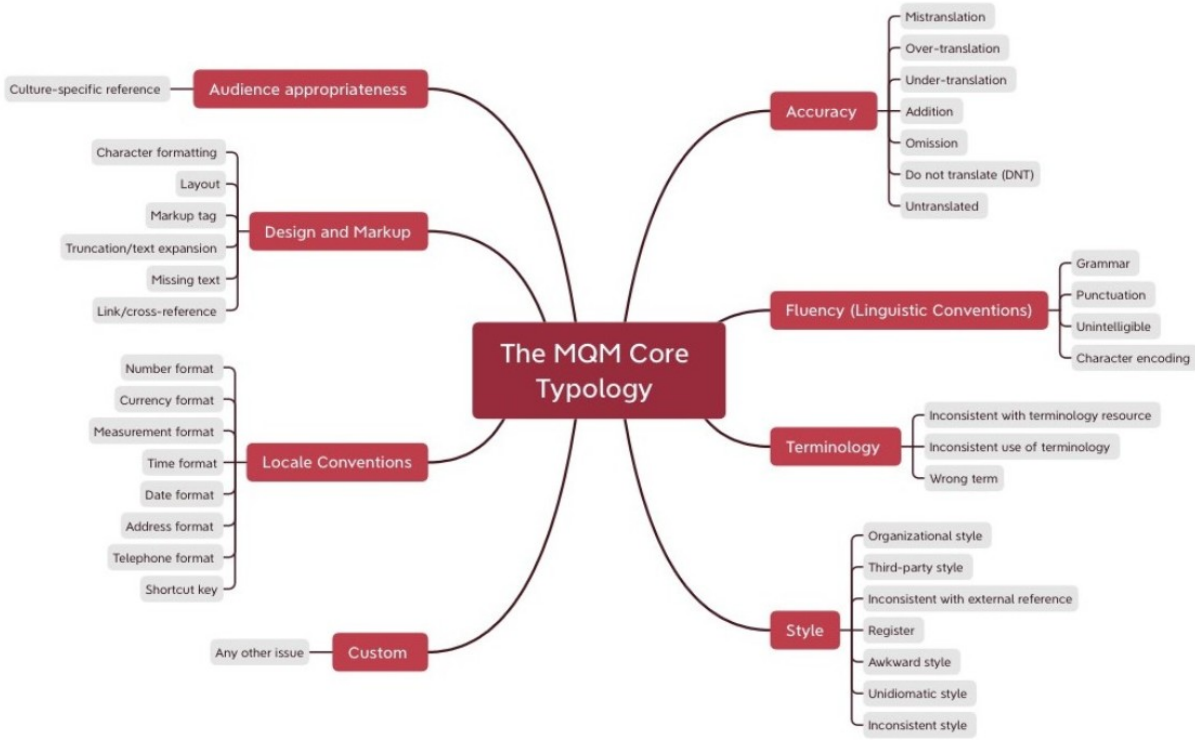


Figure 6 - Dernière version de la typologie MQM en date (Fakih et al. 2024)

L'*American Translators Association* (ATA) utilise également une typologie d'erreurs dans le cadre de son examen de certification afin d'évaluer rapidement si une traduction répond aux critères de qualité attendus (American Translators Association 2023c). Cette évaluation repose sur une grille de 25 types d'erreurs, prenant en compte la terminologie, le registre textuel, la ponctuation, etc. (voir Figure 7). Chacune de ces erreurs est ensuite pondérée en fonction de sa gravité (voir Figure 8). Une note de 18 ou plus est éliminatoire.

L'*Institute of Translation and Interpreting* (ITI), l'association professionnelle représentant les traducteurs, les interprètes et les entreprises de services linguistiques au Royaume-Uni, propose aussi une grille de notation en 8 points dans le cadre de son examen de certification (Institute of Translation and Interpreting 2021). L'évaluation repose sur 6 catégories d'erreurs faisant perdre des points aux traducteurs (erreur de transfert, de terminologie, de grammaire…) et 1 catégorie donnant au contraire des points en cas de proposition excellente. Une seule erreur jugée sérieuse est éliminatoire, ainsi qu'une note finale inférieure à 63.

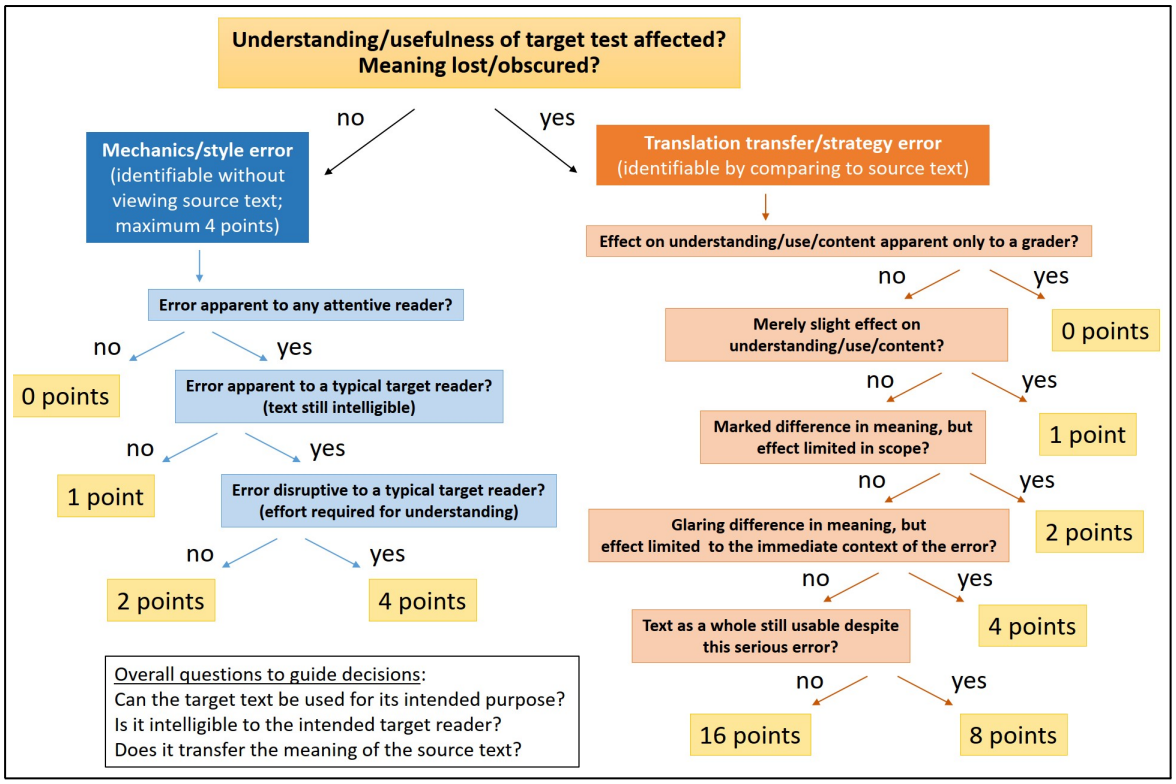


Figure 7 – Logigramme pour pondérer les erreurs observées dans les traductions lors de l'examen de certification de l'ATA (American Translators Association 2023a)

Parmi ses défauts, il faut noter que cette typologie regroupe dans une même catégorie les erreurs terminologiques, celles liées au genre textuel et celles liées aux collocations. De même, les erreurs de grammaire et de syntaxe ne sont pas distinguées.

Nous constatons donc que, dans le domaine de la traduction professionnelle (agences de traduction, indépendants…), si les approches linguistiques prédominent pour analyser les erreurs produites, elles font appel, ne serait-ce qu'implicitement, à la finalité des traductions pour évaluer la gravité des erreurs commises. Les typologies d'analyses des erreurs, comme la typologie Dynamic Quality Framework (DQF), la Multidimensional Quality Metrics (MQM) ou la LISA QA, combinent la prise en compte des erreurs langagières, les attentes du client (guide de rédaction, terminologie propre…), les nécessités de distinction entre différentes variantes linguistiques (par exemple, entre anglais britannique ou américain, ou entre français de France ou français canadien) et les contraintes du processus de production (temps passé, délai de livraison…) ; chacun de ces éléments est pondéré en fonction de sa gravité.

ATA CERTIFICATION PROGRAM
FRAMEWORK FOR STANDARDIZED ERROR MARKING
Version 2017

Exam No.: _____

Passage No.: _____

1	2	4	8	16	Code	Reason
Errors that concern the form of the exam						
Treat missing material within the passage as an omission.					UNF	Unfinished (If a passage is substantially unfinished, do not grade the exam.)
					ILL	Illegibility
					IND	Indecision, gave more than one option
Meaning transfer or strategic errors: Negative impact on clarity or usefulness of target text. Use one of the categories below whenever possible. If none are applicable, use OTH-MT						
					A	Addition
					AMB	Ambiguity
					COH	Cohesion
					F	Faithfulness (translation strays too far from ST meaning)
					FA	Faux ami (false friend)
					L	Literalness
					MU	Misunderstanding of source text (if identifiable)
					O	Omission
					T	Terminology, word choice
					TT	Text type (failure to follow Translation Instructions) This category also covers register and style.
					VT	Verb Tense (grammar correct, but conveys wrong meaning)
					OTH-MT	Other (describe; use a separate page if needed)
Mechanical errors: Negative impact on overall quality of target text. Points may vary by language. Maximum 4 points. Use one of the categories below whenever possible. If none are applicable, use OTH-ME						
					G	Grammar (use one of next two sub-categories if applicable)
					SYN	— Syntax (phrase/clause/sentence structure)
					WF/PS	—Word form /Part of speech
					P	Punctuation
					SP/CH	Spelling/Character (usually 1 point, maximum 2; if more than 2 points, another category must apply)
					D	— Diacritical marks/Accents
					C	— Capitalization
					U	Usage
					OTH-ME	Other (describe; use a separate page if needed)
	__x 2 = ___	__x 4 =___	___x 8 =___	_x 16 =___		**Column totals**
A grader may stop marking errors when score reaches 46 error points (mark such exams 46+)			A grader may award a quality point for each of up to three instances of exceptional translation.			Quality points are subtracted from the error point total to yield a final score. A passage with a score of 18 or more points receives a grade of Fail.
Total error points (add column totals):			**Quality points (maximum 3):**			**Final passage score (subtract quality points from error points):**

Figure 8 – Grille d'analyse des erreurs dans les traductions lors de l'examen de certification de l'ATA (American Translators Association 2023b)

Ces typologies professionnelles présentent les mêmes limites que les typologies utilisées dans le domaine de la TA : elles s'avèrent inadaptées à l'analyse des erreurs linguistiques portant sur un seul phénomène linguistique, en raison du fort accent mis sur les erreurs de localisation pour répondre aux besoins d'un client spécifique, ce qui rend difficile leur utilisation sur une grande diversité de genres textuels. Si nous considérons par exemple le phénomène linguistique que sont les syntagmes nominaux complexes en anglais, ces typologies s'avèrent trop générales pour prendre en compte les spécificités syntaxiques des erreurs portant sur les syntagmes qui posent difficulté aux traducteurs : erreurs portant sur les constituants ou l'identification de la tête ou encore, sur des variations terminologiques inappropriées au sein du syntagme (León Araúz & Cabezas-Garcia 2020, Kübler et al. 2022). Si ces typologies peuvent permettre de distinguer une bonne traduction d'une mauvaise sur la base de grandes catégories d'erreurs, elles ne sont cependant pas adaptées à une analyse fine des difficultés rencontrées par un système.

De plus, on peut s'interroger sur la pertinence de certaines des catégories proposées, et ce, quelle que soit la métrique considérée. Certaines catégories de la typologie DQF/MQF sont ainsi difficiles à circonscrire : les erreurs de terminologie ne semblent prendre compte que les problèmes de cohérence au sein du texte (*Inconsistent with termbase et Inconsistent use of terminology* ; comment distinguer entre un style *unidiomatic* et *awkward* ?, etc.). Concernant la typologie utilisée par l'ATA, certaines catégories d'erreurs posent également question. La distinction faite entre les catégories Faux Ami et Terminology nous paraît contestable. Dans la description de l'erreur Terminology, l'ATA explique : « If the erroneous word or term is based on the choice of a target-language cognate that has a different meaning, the subcategory Faux Ami (FA) may be used […] ». Une telle distinction impliquerait d'avoir une justification de ses choix par le traducteur afin de déterminer si l'erreur provient d'une confusion sémantique effective entre des mots anglais et français similaires graphiquement ou homophones (erreur du type *Faux Ami*) ou de la sélection d'une variation terminologique qui aurait pu être en usage dans le domaine de spécialité considéré par analogie avec l'anglais, même si celle-ci semble incorrecte au regard de la grammaire de langue générale (erreur du type *Terminology* en raison de la traduction d'un terme par un non-terme).

2.3.Les typologies d'erreurs à visée pédagogique

Les typologies d'erreurs à visée pédagogique sont orientées vers des applications en enseignement, soit pour la traduction, soit pour l'acquisition d'une L2. En traduction, elles permettent d'annoter des corpus de textes produits par des apprenants en traduction. L'évaluation dans un contexte pédagogique diffère des évaluations en milieu professionnel ou dans la recherche. La notion d'évaluation à visée pédagogique vise non seulement à déterminer si les étudiants ont acquis le niveau de compétences attendu, mais également à leur fournir un retour sur leurs difficultés et leurs possibilités d'amélioration. On produit ainsi des corpus d'apprenants de la traduction permettant à la fois l'évaluation des traductions et la conception de matériel pédagogique pour le cours de traduction/post-édition (Castagnoli et al. 2011) . Si le premier objectif est proche de celui d'une évaluation en contexte professionnel (une traduction est recevable ou non, et ce, malgré la présence de quelques erreurs), le second implique une évaluation beaucoup plus fine des erreurs rencontrées afin de pouvoir conseiller les étudiants et produire des exercices basés sur les difficultés des apprenants.

Bowker et Bennison (2003) étudient l'utilisation des corpus d'apprenants de la traduction à des fins d'enseignement et de recherche en traduction, tout comme Uzar et Walinski (2001) qui développent le corpus PELCRA. Ces deux études visaient à explorer les difficultés des apprenants en matière de traduction, à l'aide d'une typologie d'erreurs. En effet, les corpus d'apprenants de la traduction, annotés à l'aide d'une typologie d'erreurs, fournissent une base utile pour étudier la qualité de la traduction et développer du matériel pédagogique. Par exemple, Espunya (2013) exploite un corpus d'apprenants anglais-catalan, annoté à l'aide d'informations linguistiques et d'étiquettes d'erreurs de traduction, pour rechercher les différents types d'erreurs de traduction lexicales, telles que les faux amis et les choix lexicaux erronés ou imprécis. La typologie Multilingual e-Learning in Language Engineering (MeLLANGE)[7] a été développée dans le cadre du projet européen du même nom de 2004 à 2007 (Kübler 2008, Castagnoli et al. 2011). Le projet avait pour objectif de couvrir les différents types d'erreurs (syntaxe, terminologie, erreurs de registre ou de style) potentiellement produites par les apprenants en traduction dans six langues pour faciliter le commentaire de leurs traductions par leurs enseignants et l'harmonisation de ces commentaires entre intervenants. Par ailleurs, le corpus d'apprenants multi-directionnel et multilingue devait amener à créer du matériel pédagogique pour la formation (Kunz et al. 2010). Il faut noter que cette typologie a été développée dans une visée descriptive des erreurs. Par conséquent, la typologie n'inclut aucun élément permettant un jugement sur l'erreur observée : contrairement aux métriques professionnelles présentées précédemment dans ce rapport, il n'y a donc pas de notion de pondération des erreurs en fonction d'un niveau de gravité prédéterminée. La typologie produite résulte d'un consensus entre les chercheurs des 10 partenaires du projet, répartis dans sept pays européens différents et couvrant 6 langues différentes dont l'anglais et le

7 . http://mellange.eila.univ-paris-diderot.fr/

français. La typologie se compose de deux grandes catégories distinguant entre les erreurs liées au contenu, *Content transfer*, et celles liées à la langue, *Language* ; catégories elles-mêmes divisées en sous-catégories (voir la figure 9). Il est possible d'attribuer plusieurs catégories d'erreurs à un même segment de texte. Štěpánková (2014) utilise une version adaptée de la typologie d'erreurs MeLLANGE pour le corpus d'apprenant de la traduction tchèque-anglais. Elle souligne la nécessité d'adapter les typologies d'erreurs aux différentes paires de langues, car certaines erreurs sont liées à la langue. Par exemple, le tchèque n'a pas de déterminant, ce qui pose des problèmes de choix de déterminants pour les langues à déterminants telles que l'anglais. Il est donc nécessaire d'ajouter une catégorie 'déterminant' dans la typologie d'erreurs.

Récemment, Kübler, Mestivier et Pecman se sont intéressées aux erreurs de distorsion faites par les apprenants en traduction lors de la traduction des « groupes nominaux complexes ». Leur typologie d'erreurs repose sur la typologie MeLLANGE qui a été complétée de manière à prendre en compte les spécificités liées à ces groupes (Kübler, Mestivier et Pecman 2022). Pour cela, elles ont créé une sous-catégorie d'erreurs « *CNP* » à la catégorie « *Syntax* » de la typologie MeLLANGE. Cette catégorie se combinait le plus souvent avec d'autres catégories d'erreurs comme « *Distortion* » ou « *Awkward* ». L'ensemble des erreurs recensées dans les corpus d'apprenant ont été regroupées en seize catégories différentes (voir la Figure 9). Cette typologie résulte d'un inventaire sur plusieurs années portant sur les erreurs faites par les apprenants en traduction lors de la post-édition d'articles de recherche du domaine des sciences de la Terre et de la Planète. Ce travail d'inventaire se poursuit aujourd'hui grâce à un cours annuel du type « Travaux dirigés » en première année de master en traduction à l'université Paris Cité, dans lequel l'évaluation de la TA et de la post-édition a été introduite en 2020 (Mestivier & Martikainen, Kübler et al. 2021, 2022) . Cela permet d'envisager une comparaison sur le long terme entre les erreurs faites par les apprenants en traduction et celles produites par les systèmes de traduction automatique.

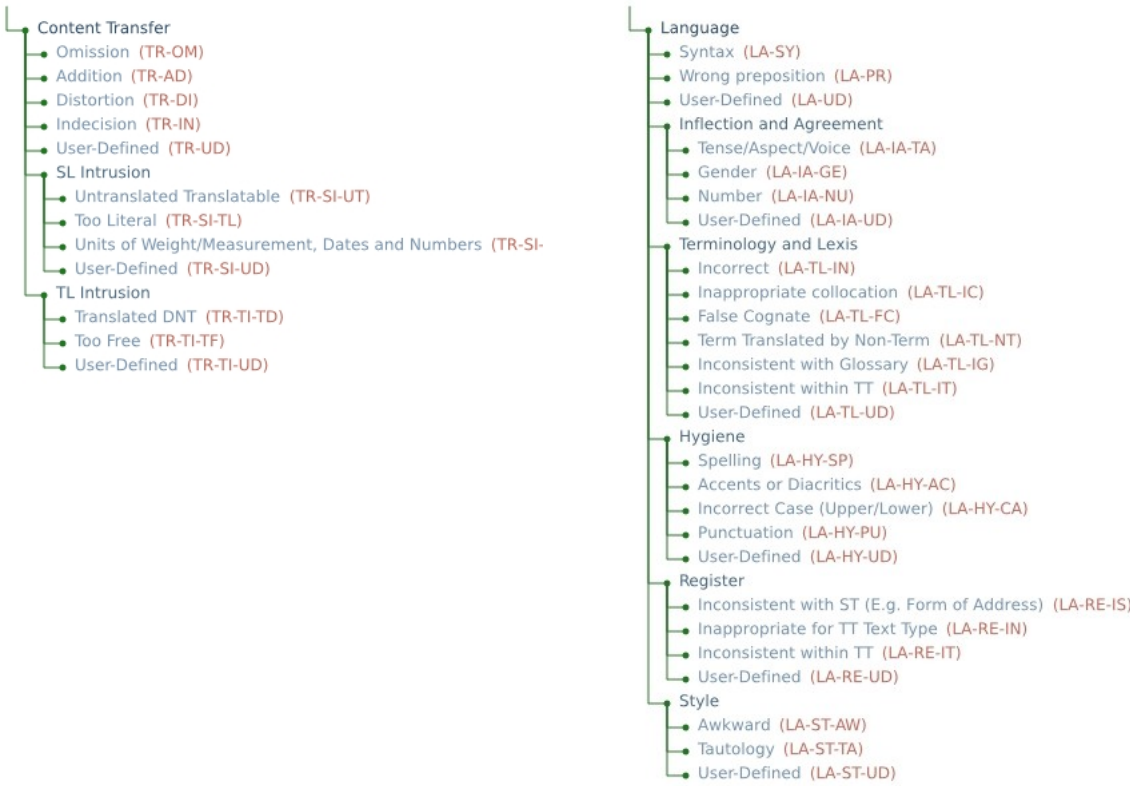


Figure 9 – Schéma d'annotation d'erreurs MeLLANGE (Castagnoli et al. 2011)

2.4.Des typologies spécifiques à la post-édition

En ce qui concerne l'évaluation de la post-édition et outre l'exploitation de typologies d'erreurs, l'introduction d'une classification des modifications effectuées entre la soirtie de traduction automatique et le texte post-édité.L'équipe du CLILLAC-ARP a par exemple ajouté à la typologie MeLLANGE, des annotations permettant les différents types de modifications entre la TA et le résultat de la post-édition (Kübler et al. 2021, 2022). Les annotations se répartissent en deux types de sorties de traduction automatique: TA correcte vs TA incorrecte et dans ces deux types, les modifications effectués lors de la PE sont classées comme correcte, incorrecte ou non corrigée, ce qui donne cinq possibilités : *TACorBienCorr*, recouvre les propositions de variantes lors de la PE ; *TACorMalCorr*, recouvre les cas où la TA est correcte et a été modifiée de manière erronée lors de la PE ; *TAEBienCor*, la TA est erronée et a été bien corrigée lors de la PE ; *TAEMalCor*, la TA est erronée et a été mal corrigée lors de la PE ; *TAENonCor*, la TA est erronée et n'a pas été corrigée lors de la PE.

Cette classification est celle qui est utilisée actuellement et s'applique conjointement avec une version modifiée de la typologie MeLLANGE (cf. annexe1). Cette annotation des erreurs de TA et PE associée aux annotations de modifications entre la TA et la PE est exploitées comme outil pédagogique, mais permet aussi de créer des corpus

d'apprenants de la post-édition richement annotés. Les étudiants poste-éditent les traductions automatiques de textes; ensuite, les enseignants annotent la TA et la PE à l'aide de ce double système d'annotation, ce qui permet ensuite aux étudiants qui y ont accès de réviser leurs post-éditions. Par ailleurs, les corpus d'apprenants de la PE ainsi compilés sont exploités pour créer des exercices de post-édition l'année suivante.
Lefer et al. (2022) proposent une approche un peu similaire, 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."* (Lefer et al. 2022, p. 5)

Le processus d'évaluation est déterminé par un arbre décisionnel (voir Figure 10 ci-dessous).

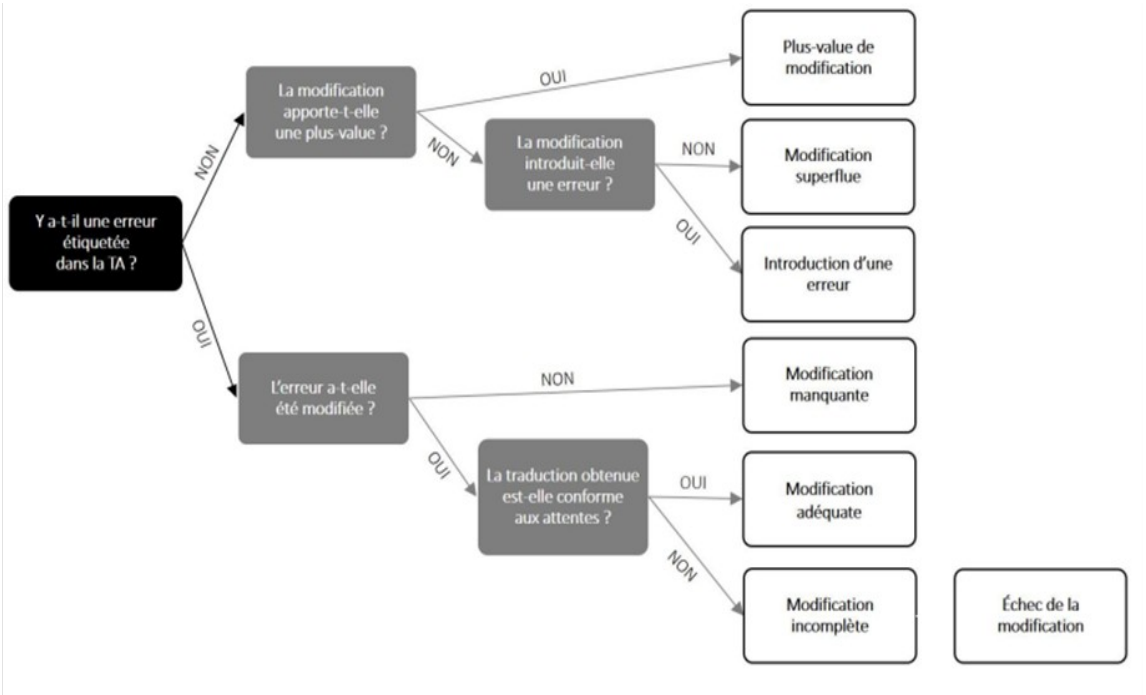


Figure 10 : arbre décisionnel élaboré par Lefer et al. (2022, p. 6)

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 (Lefer et al. 2022, 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.

3. Proposition d'une nouvelle typologie

○3.1. Déclinaison et fusion de typologies existantes

On constate que les différentes typologies présentent de nombreux éléments en commun. Cependant, elles ont toutes des finalités différentes. Dans le cadre du projet MaTOS, l'objectif consiste à élaborer une typologie permettant d'évaluer à la fois la traduction humaine, la traduction automatique et la post-édition. D'une part, l'équipe du CLILLAC-ARP avait l'habitude de travailler avec la typologie MeLLANGE (voir Figure 9), une typologie basée sur une approche linguistique et est destinée à l'évaluation de traductions dans un cadre d'enseignement. D'autre part, MQM est une typologie très largement utilisée pour l'évaluation de la qualité de traductions, à la fois humaines et automatiques (Freitag et al. 2021), voire hybrides (post-éditions) et propose des types d'erreurs pertinents non présents dans MeLLANGE. Toutefois, certaines des erreurs répertoriées dans la typologie MQM posent problème (par exemple, la différence entre *coherence* et *cohesion* reste difficile à distinguer). 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 avec certains types d'erreurs retrouvés dans MQM (voir Figure 11).

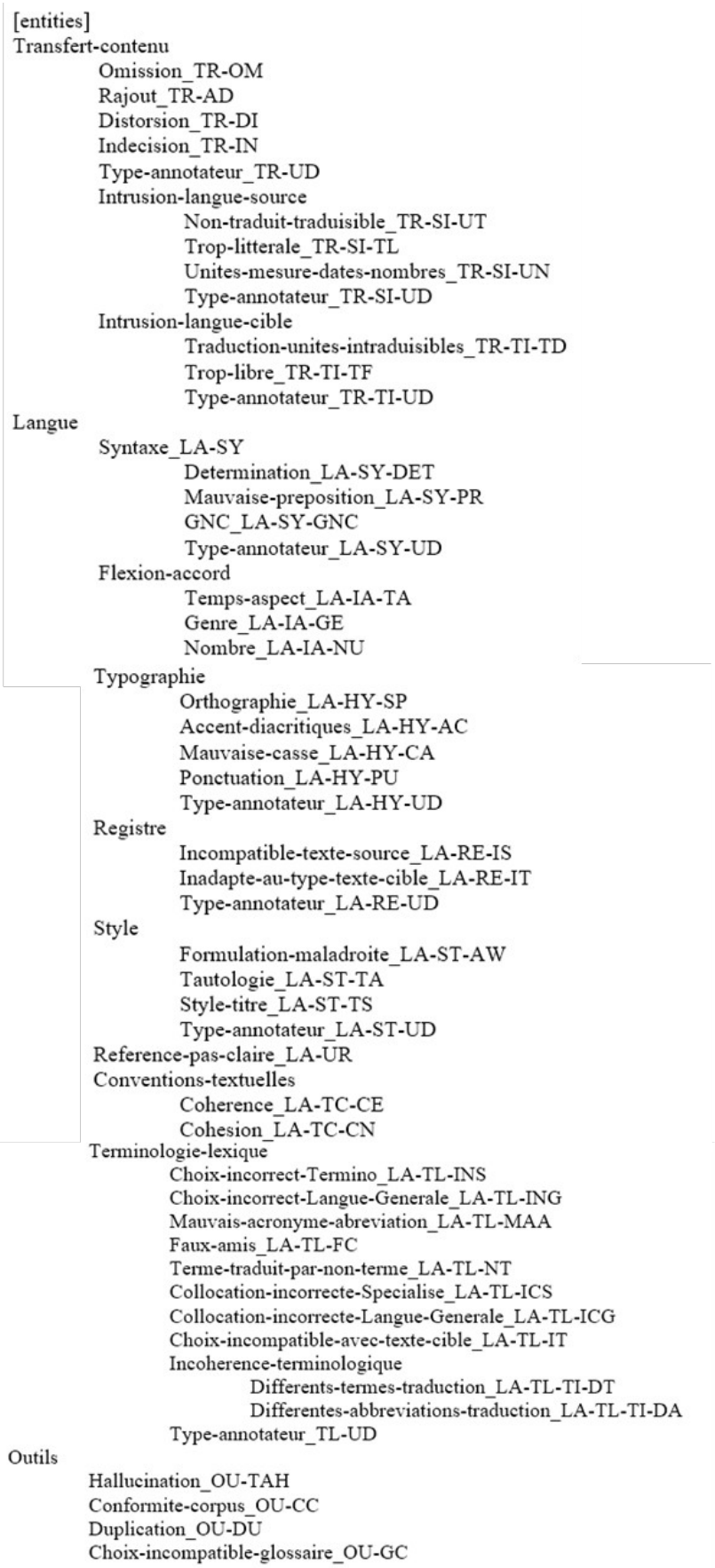


Figure 11 - Typologie d'erreurs MeLLANGE - V2

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 et 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 » comprend 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 dans le manuel d'annotation en annexe. Celui-ci vise à servir de guide pour l'annotation d'erreurs dans le cadre de traductions humaines, automatiques ou de post-éditions.

○ 3.2. Définition d'attributs

Les attributs ajoutés représentent tout d'abord les modifications prenant place entre la TA et la PE que nous avons présentées plus haut (Kübler et al. 2021) (Figure 12).

```
[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 12 - Attributs de la typologie d'erreurs

Ces attributs sont à ajouter lorsqu'on annote une post-édition, et non une traduction automatique ou humaine. C'est ce schéma d'annotation qui est utilisé dans le projet MaTOS.
.

○3.3. Définition de 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*[8], où l'on observe quatre niveaux de gravité : neutre (score 0), mineur (score 1), majeur (score 5) et critique (score 25). Par conséquent un score de gravité a été ajouté au schéma d'annotation (Minder 2024), mais il a été 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 ;
- 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.

Ces attributs et les scores de gravité sont également expliqués en détail dans le manuel d'annotation (annexe 1).

Cette typologie sera désormais appliquée dans le cadre du projet, notamment dans le but d'automatiser l'évaluation de la TA à l'aide de l'IA générative.

Références bibliographiques

Ageeva, Ekaterina, Tyers, Francis M., Forcada, Mikel L. & Pérez-Ortiz, Juan Antonio. (2015). "Evaluating machine translation for assimilation via a gap-filling task". In : Proceedings of the 18th Annual Conference of the European Association for Machine Translation, Antalya, Turkey. European Association for Machine Translation.

8 https://themqm.org/error-types-2/1_scorecards/.

American Translators Association, ATA. (2023). *Flowchart for Error Point Decisions*. url : www.atanet.org/certification/how-the-exam-is-graded/error-points/ (visité le 10/10/2023).

Banerjee, Satanjeev & Lavie, Alon. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In : Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan. Association for Computational Linguistics, p. 65–72.

Benali, Abdelkader. (2012). "Les problèmes de la catégorisation textuelle: entre fondements théoriques et fondements structurels". Synergies Algérie, 17, p. 35-49.

Blain, Frédéric, Senellart, Jean, Schwenk, Holger, Plitt, Mirko, & Roturier, Johann. (2011). "Qualitative Analysis of Post-Editing for High Quality Machine Translation". In : Proceedings of Machine Translation Summit XIII: Papers, Xiamen, China.url : https://aclanthology.org/2011.mtsummit-papers.17.pdf.

Bowker, Lynne & Bennison, Peter.. (2003). "Student translation archive: Design, development and application". Corpora in Translator Education. p. 103-117.

Burlot, Franck & Yvon, François. (2017). "Evaluating the morphological competence of Machine Translation Systems". In : Proceedings of the Second Conference on Machine Translation. Copenhagen, Denmark : Association for Computational Linguistics, p. 43-55.

Burlot, Franck & Yvon, François. (2018). "Évaluation morphologique pour la traduction automatique : adaptation au français (Morphological Evaluation for Machine Translation : Adaptation to French)". In : Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN. Rennes, France : ATALA, p. 61-74.

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

Carroll, John B. (1966). *An Experiment in Evaluating the Quality of Translation,* In : Mechanical Translation and Computational Linguistics, vol 9, p. 55–66.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler & Alexandra Volanschi. (2011). "Designing a Learner Translator Corpus for Training Purposes". In : Corpora, Language, Teaching, and Resources : From Theory to Practice, p. 221-248. url : https://u-paris.hal.science/hal-01135016.

Castilho, Sheila, Doherty, Stephen, Gaspari, Federico & Moorkens, Joss. (2018). "Approaches to Human and Machine Translation Quality Assessment" : From Principles to

Practice". In : Joss Moorkens et al. (Eds) Translation Quality Assessment : From Principles to Practice. Springer International Publishing, p. 9-38.

Castilho, Sheila. (2019). *Machine Translation Evaluation : know your essentials. Présentation. Varna, Bulgaria.* url : http://ranlp.org/archive/ranlp2019/NMT-ranlp2019-Sh.Castilho.pdf.

Charolles, Michel. (2011). "Cohérence et cohésion du discours". In : Holker & Marello. *Dimensionen der Analyse Texten und Diskursivent - Dimensioni dell'analisi di testi e discoursi*, Lit Verlag, p. 153-173.

Comelles, Elisabet et al. (2012). « VERTa : linguistic features in MT evaluation ». In : Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turkey. url : http://www.mt-archive.info/LREC-2012-Comelles.pdf.

Costa, Angela, Ling, Wang & Luís, Tiago & Correia, Rui & Coheur, Luisa. (2015). "A linguistically motivated taxonomy for Machine Translation error analysis. Machine Translation". 29. p. 127-161.

Chuquet, Hélène & Michel Paillard (1989). *Approche linguistique des problèmes de traduction anglais-français*. Paris : Ophrys Édition révisée. 451 p.

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

Doherty, Stephen (2017). "Humans Issues in (Machine) Translation Quality Assessment". In : *Human Issues in Translation Technology.* 1st Edition. The IATIS Yearbook. London : Routledge, p. 131-148. isbn : 978-1-138-12329-8.

Esperança-Rodier, Emmanuelle & Nicolas Becker (2018). "Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs". In : Actes de TALIA 2018. Nancy, France. url : https://pfia2018.loria.fr/wp-content/uploads/2018/06/Talia-Esperan%c3%a7a-Rodier_Becker.pdf.

Espunya, Anna. (2013). "Investigating lexical difficulties of learners in the error-annotated UPF learner translation corpus". In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)* (Vol. 1, p. 129). Presses universitaires de Louvain.

Fakih, Altaf, Ghassemiazghandi, Mozhgan, Fakih, Abdul-Hafeed, Mehar Singh & Manjet Kaur. (2024). "Evaluation of Instagram's Neural Machine Translation for Literary Texts : An

MQM-Based Analysis". *GEMA Online® Journal of Language Studies*, n°24, vol.1, p. 213-233.

Farrús, Mireia, Costa-jussa, Marta, Acebal, Jose, Fonollosa, José. (2010). "Linguistic-based evaluation criteria to identify statistical machine translation errors". In : 14th Annual Conference of the European Association for Machine Translation, p. 167-173.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, & Wolfgang Macherey. (2021). "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation". Transactions of the Association for Computational Linguistics, vol 9, p. 1460–1474.

Gile, Daniel. (2005). "La qualité dans la traduction professionnelle. Les fondements". In : La traduction. La comprendre, l'apprendre. Linguistique nouvelle. Paris, France : Presses Universitaires de France, p. 37-68.

Giménez, Jesús & Lluís Màrquez. (2007). "Linguistic Features for Automatic Evaluation of Heterogenous MT Systems". In *Proceedings of the Second Workshop on Statistical Machine Translation*, 256-64. StatMT'07. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1626355.1626393.

Graham, Yvette, Baldwin, Timothy, Dowling, Meghan, Eskevich, Maria, Lynn, Teresa & Tounsi, Lamia. (2016). "Is all that Glitters in Machine Translation Quality Estimation really Gold?", In : Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Gutt, Ernst-August. (2014). *Translation and relevance: Cognition and context*. Routledge.

Guzmán, Francisco, Abdelali, Ahmed, Temnikova, Irina, Sajjad, Hassan & Vogel, Stephan. (2015). "How do Humans Evaluate Machine Translation". In : Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisboa, Portugal : Association for Computational Linguistics, p. 457-466. url : http://www.statmt.org/wmt15/pdf/WMT59.pdf

Halliday, Michael & Hasan, Ruqaiya. (1976). *Cohesion in English*. London, Longman

Hansen, Damien & Esperança-Rodier, Emmanuelle. (2022). *Human-Adapted MT for Literary Texts: Reality or Fantasy?*. In *NeTTT 2022*, Jul 2022, Rhodes, Greece. p.178-190. https://hal.science/hal-04038025

House, Juliane. (2011). "Quality". In : Routledge of Encyclopedia of Translation Studies. Second Edition. Routledge, p. 222-225.

House, Juliane. (2015). *Translation Quality Assessment : Past and present*. Routledge.

Institute of Translation and Interpreting, ITI (2021). MITI assessment Applicant Handbook. Standard Assessment Guide for Translators. url : https : / /www.iti.org.uk/asset/FA7E6E31%2D415B%2D4731%2D9F34B3EEF2F9A712/

Isabelle, Pierre, Cherry, Colin & Foster, George. (2017). "A Challenge Set Approach to Evaluating Machine Translation". In : Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, p. 2486–2496, url : http://arxiv.org/abs/1704.07431.

Ji, Ziwei, Lee, Nayeon, Frieske, Rita, Yu, Tiezheng, Su, Dan, Xu, Yan, Ishii, Etsuko, Bang, Yejin, Chen, Delong, Dai, Wenliang, Madotto, Andrea & Fung, Pascale. (2023). "Survey of hallucination in natural language generation". *ACM Computing Surveys*, n°55, vol 12, p. 1-38.

Kocmi, Tom & Federmann, Christian. (2023). GEMBA-MQM: Detecting translation quality error spans with GPT-4. *arXiv preprint arXiv:2310.13988*.

Kübler, Natalie (2008). A comparable learner translator corpus: Creation and use. In *Proceedings of the Comparable Corpora Workshop of the LREC Conference*, p. 73-78.

Kübler, Natalie, Martikainen, Hanna, Mestivier, Alexandra & Pecman, Mojca. (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, Natalie, Mestivier, Alexandra & Pecman, Mojca. (2022a). "Using comparable corpora for translating and post-editing complex noun phrases in specialized texts." Extending the scope of corpus-based translation studies (2022), p. 237-266.

Kübler, Natalie, Pecman, Mojca & Mestivier, Alexandra (2022, April). "Test corpora, status of the translation error, feedback on post-editing analysis and typologies of translation errors from learner's corpora". In *TRALOGYIII*.

Kunz, Kerstin, Castagnoli, Sara & Kübler, Natalie. (2010). "Corpora in translator training." In : Gile Daniel., Pokorn Nike K. and Hansen Gyde (eds), *Why Translation Studies Matter. Benjamins Translation Library (BTL)*.

Lefer, Marie-Aude, Justine Piette, & Romane Bodart. (2022). *Manuel MTPEAS: Machine Translation Post-Editing Annotation System. Version 1.0*, https://oer.uclouvain.be/jspui/bitstream/20.500.12279/829/9/MTPEAS_manual_EN_final_CC.pdf

León Araúz, Pilar & Melania Cabezas García. (2020). "Term and translation variation of multi-word terms." In : Mogorrón Huerta, Pedro (ed.) Análisis multidisciplinar del fenómeno

de la variación fraseológica en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue 6, p. 210-247.

Loock, Rudy. (2018). “Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus”. In : Méta : journal des traducteurs / Meta : Translators’ Journal 63.3, p. 786-806.

Macketanz, Vivien, Avramidis, Eleftherios, Burchardt, Aljoscha, Helcl, Jindřich & Srivastava, Ankit. (2017). “Machine translation : Phrase-based, rule-based and neural approaches with linguistic evaluation”. In : Cybernetics and Information Technologies, n°17, vol 2, p. 28-43.

Mounin, Georges. (1976). *Les problèmes théoriques de la traduction.* Gallimard.

Nida, Eugene. (1964). *Toward a Science of Translation.* Leiden: Brill.

O’Brien, Sharon. (2012). “Translation as human–computer interaction”. In : Translation Spaces, n°1, vol 1, p. 101-122.

Popović, Maja. (2018). “Error Classification and Analysis for Machine Translation Quality Assessment”. In : Joss Moorkens et al. (Eds) Translation Quality Assessment : From Principles to Practice. Springer International Publishing, p. 129-158. url : https://doi.org/10.1007/978-3-319-91241-7%5C_7.

Poibeau, Thierry. (2017). *Machine Translation. Essential Knowledge.* The MIT Press.

Poibeau, Thierry. (2022). “On “Human Parity” and “Super Human Performance””. In : Machine Translation Evaluation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, p. 6018–6023, Marseille, France. European Language Resources Association.

Rei, Ricardo, Marcos Vinícius Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, T. Glushkova, Duarte M. Alves, Alon Lavie, Luísa Coheur & André F. T. Martins. (2022). “CometKiwi: IST-unbabel, submission for the quality estimation shared task”. *arXiv preprint arXiv:2209.06243*, https://aclanthology.org/2022.wmt-1.60

Secară, Alina. (2005). “Translation evaluation : A state of the art survey.” In : *Proceedings of the eCoLoRe/MeLLANGE Workshop.* Leeds (UK) : Centre for Translation Studies, University of Leeds, p. 39-44.

Specia, Lucia & Kashif, Shah. (2018). “Machine Translation Quality Estimation : Applications and Future Perspectives”. In : *Translation Quality Assessment. From Principles*

to Practice. 1sd. ed. Machine Translation : Technologies and Applications 1. Springer International Publishing, p. 159-178.

Stymne, Sara & Ahrenberg, Lars. (2012). "On the practice of error analysis for machine translation evaluation". In : Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey : European Language Resources Association (ELRA), p. 1785-1790. url : http://www.lrec-conf.org/proceedings/lrec2012/pdf/717_Paper.pdf.

Stymne, Sara. (2018). *Machine Translation Evaluation. Cours (Lecture).* Uppsala. url : https://cl.lingfil.uu.se/kurs/MT18/slides/f2-eval.pdf

Toury, Gideon. (1995). *Descriptive Translation Studies and Beyond.* Amsterdam: Benjamins.

Uzar, Rafal & Jacek Tadeusz Waliński. (2001). "Analysing the fluency of translators." *International journal of corpus linguistics*, n°6, vol 3, p. 155-166.

Venuti, Lawrence. (1995). *The Translator's Invisibility: A History of Translation.* London: Routledge.

Vermeer, Hans J. (1996). *A skopos theory of translation: (some arguments for and against)*. Heidelberg: TextconText Verlag.

Vilar, David, Xu, Jia, D'Haro, Luis Fernando, Ney, Hermann. (2006). "Error Analysis of Machine Translation Output". In : Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. European Language Resources Association (ELRA), p. 697-702.

Vinay, Jean-Paul & Darbelnet, Jean. (1958). *Stylistique comparée du français et de l'anglais: méthode de traduction.* Paris: Didier.

Way, Andy. (2018). "Quality Expectations of Machine Translation". In : Translation Quality Assessment. Joss Moorkens et al. (Eds) Machine Translation : From Principles to Practice. Springer International Publishing, p. 159-178.

Wong, Billy TM & Kit, Chunyu. (2012). "Extending machine translation evaluation metrics with lexical cohesion to document level." *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*.

Annexe 1 : Manuel d'annotation de la nouvelle typologie

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 , qui a été créée pour l'annotation de traductions d'étudiants de master en traduction, ainsi que la typologie Multidimensional Quality Metrics (MQM), élaborée pour l'évaluation de la qualité des traductions dans un contexte professionnel.

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 le domaine du traitement automatique des langues.

o 1. Présentation de la typologie d'erreurs

La typologie d'erreurs adaptée pour ce projet (cf. figure 11) 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
TA_Erronee Value:TAEBienCor|TAEMalCor|TAENonCor

Leur utilisation est détaillée au point suivant.

o 2. Les principes généraux

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

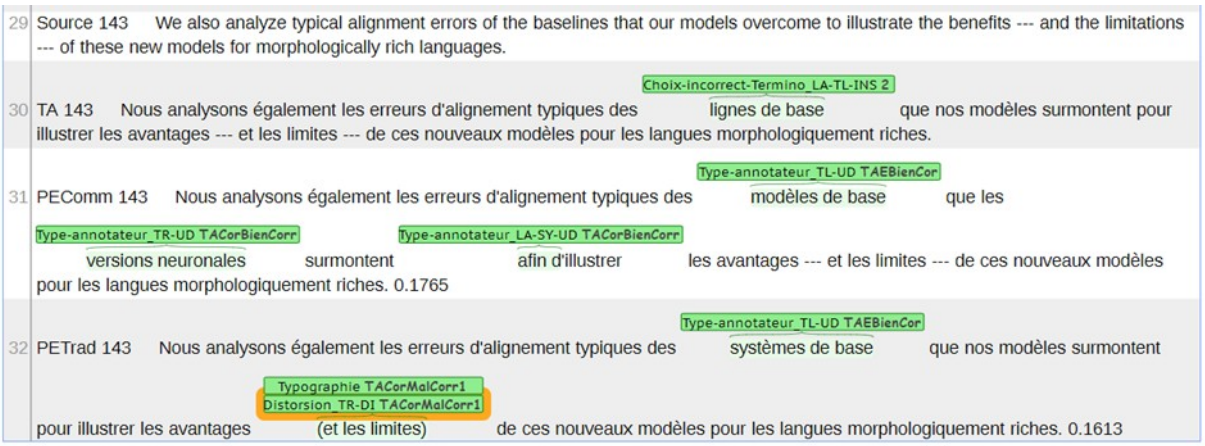
- **Principe 1 : choisir le niveau de granularité le plus précis**

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 ». Il arrive cependant que, dans certains rares cas, 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.



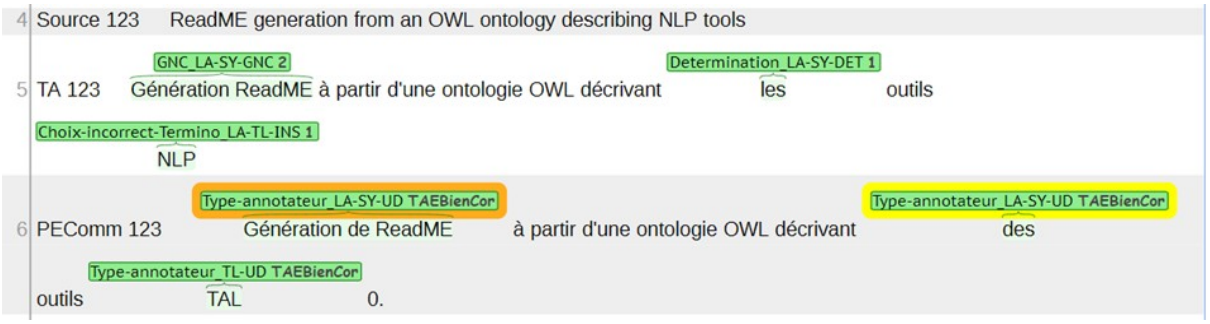
À la ligne 32, on constate que l'utilisation de parenthèses plutôt que l'utilisation de tirets cadratin (erreur de typographie) engendre une distorsion. Dans ce cas, il est nécessaire d'annoter les deux erreurs (ou plus).

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 »**

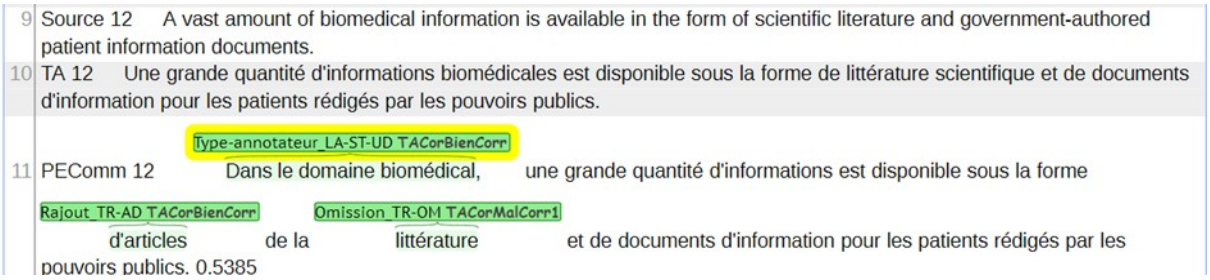
Dans chaque catégorie d'erreur, on retrouve l'étiquette *type-annotateur*. Cette étiquette est à utiliser lorsqu'on annote un segment qui ne contient pas d'erreur (et, dès lors, aucun score de gravité) dans les cas suivants :

- 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 :



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.



Ici, 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

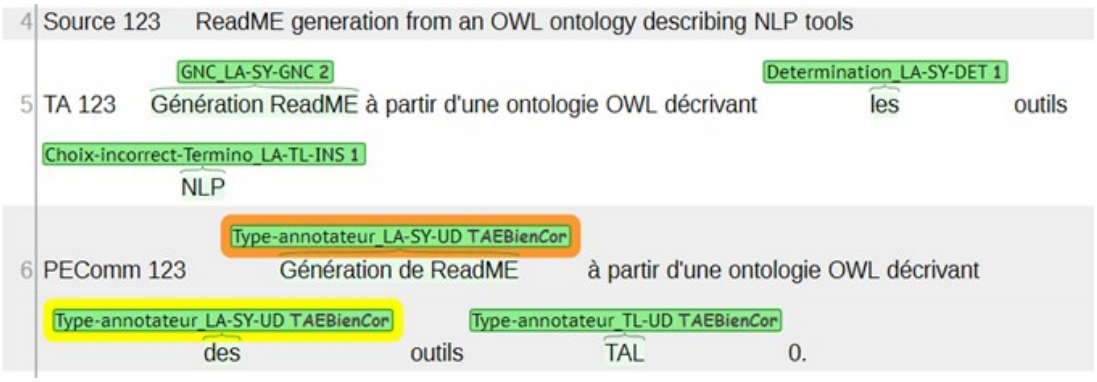
La première catégorie comporte les attributs à utiliser lorsque la traduction automatique (TA) est correcte :

- 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 ;

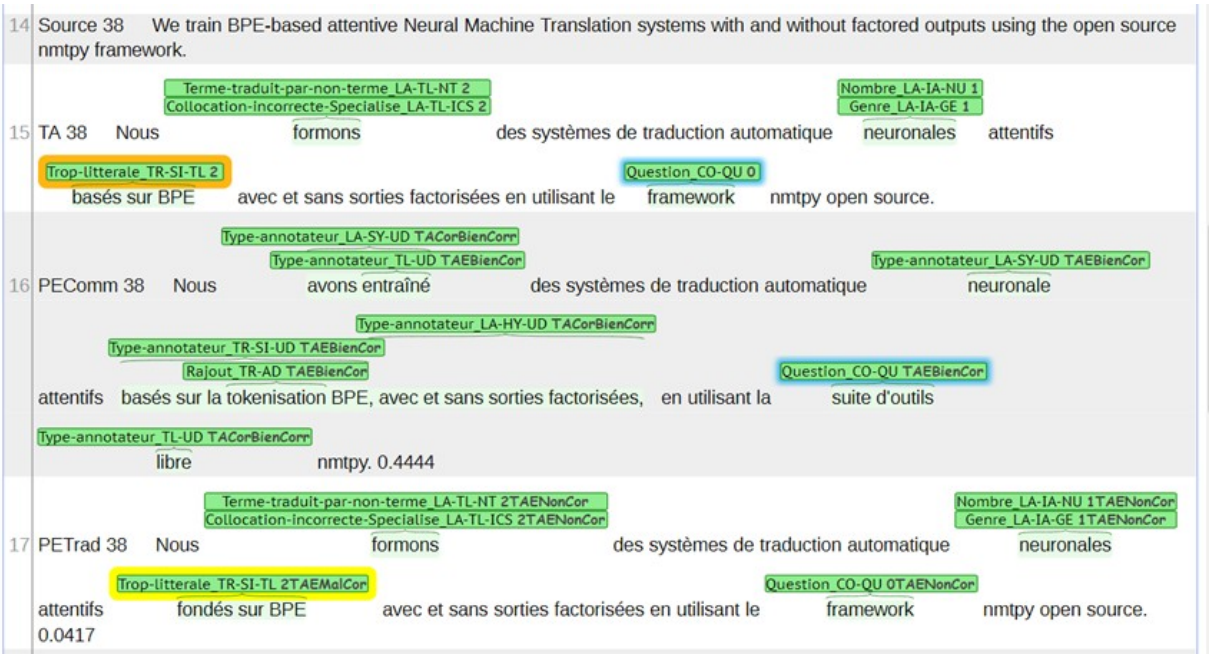
La seconde catégorie comprend les attributs qu'il convient d'utiliser lorsque la traduction automatique comporte une erreur :

- 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 :



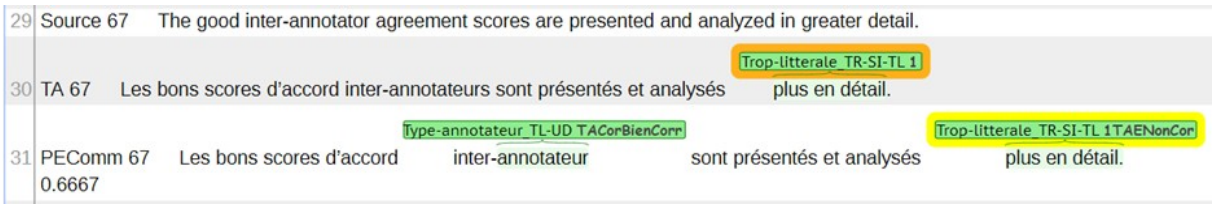
Ci-dessus, on observe que l'erreur dans la TA (détermination, ligne 5) a été corrigée dans la post-édition (type-annotateur, ligne 6). Par conséquent, lorsqu'on ajoute l'annotation « type-annotateur », il convient d'ajouter l'attribut TAEBienCor (TA erronée bien corrigée).

- 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 :



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.



Ici, 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 conserve 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.

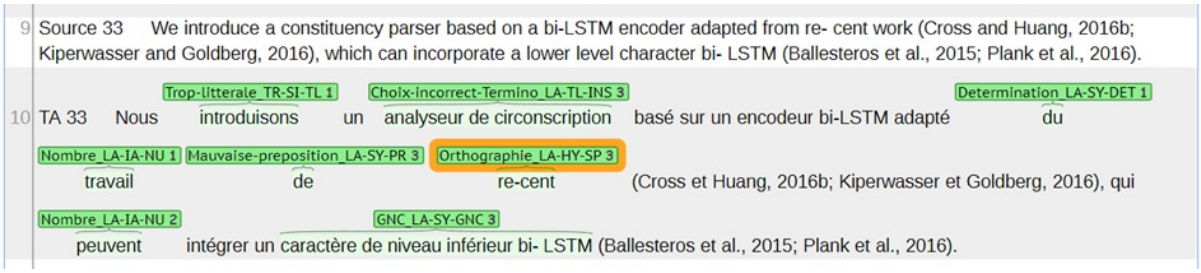
- TA erronée mal corrigée : si l'erreur présente dans la TA a été mal corrigée (c'est-à-dire qu'il reste une erreur dans la post-édition, mais que ce n'est pas la même que dans la TA), alors le score de gravité peut être différent.

- 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**

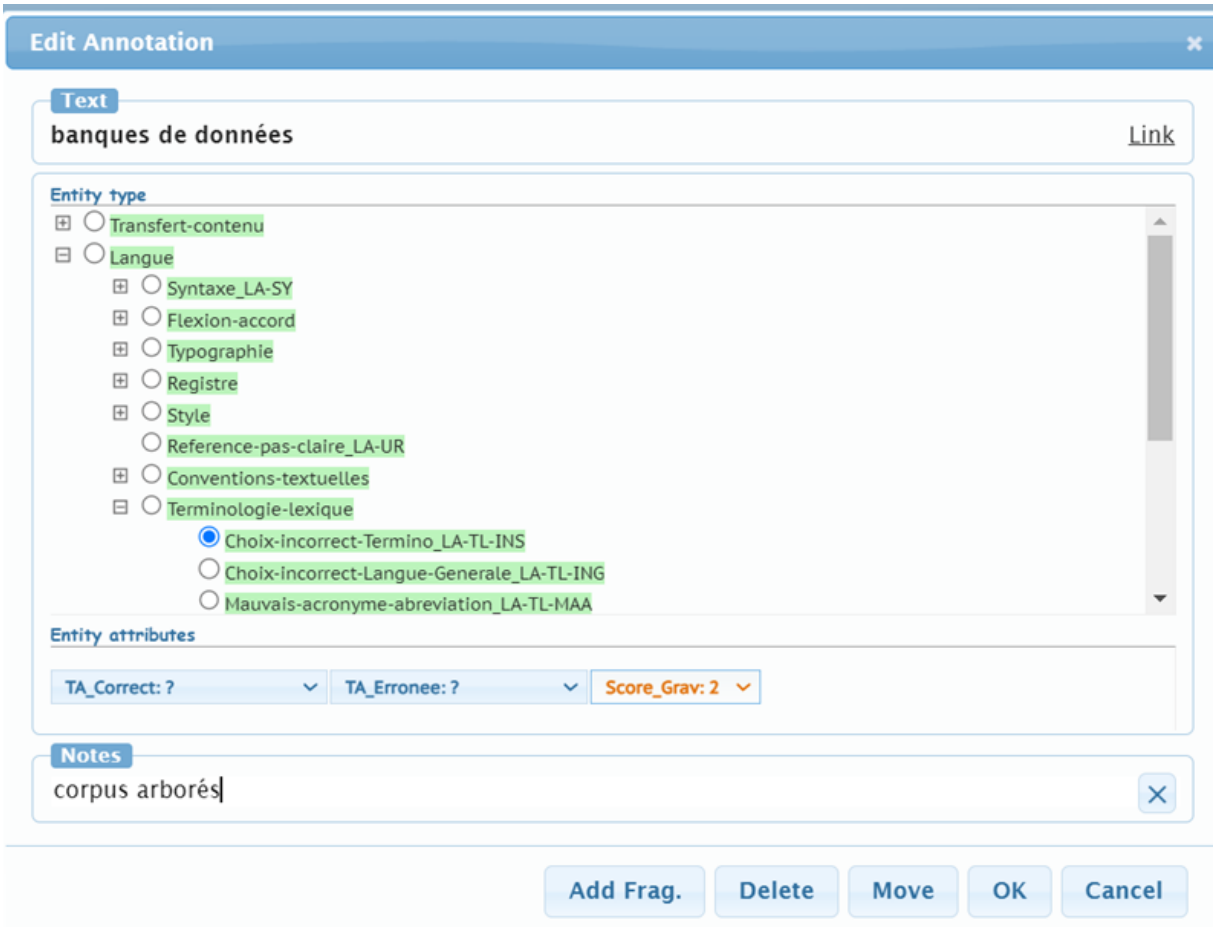
Il se peut qu'une erreur survenant dans la traduction automatique soit causée par le texte source.



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 ci-dessous).



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.

3. Les différentes erreurs en détail

3.1. Transfert-contenu

3.1.1. 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 (Delisle 2003, p. 51). 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)

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.

3.1.2. 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.

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*).

3.1.3. 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.

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.

3. 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).

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.

3.1.5. 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).

3.1.6. 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.

3.1.6.1. 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) à type-annotateur

SCORE	Aucun, puisque l'erreur est corrigée (type-annotateur).

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.

3.1.6.2. 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.

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.

3.1.6.3. 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.

3.1.6.4. 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

3.1.7. 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.

3.1.7.1. 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.

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.

3.1.7.2. Trop-libre_TR-TI-TF

Une traduction trop libre est une 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.

3.1.7.3. 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

- **3.2. Langue**

3.2.1. 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 ».

3.2.1.1. 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.

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.

3.2.1.2. 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.
ATTRIBUT	TA erronée bien corrigée
SCORE	1 : le terme est mal traduit à cause de cette préposition, mais il reste totalement compréhensible.

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.

3.2.1.3. 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 postéditer 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 postéditer 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.

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.

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.

3.2.2. 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 ».

3.2.2.1. 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.

3.2.2.2. 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.

3.2.2.3. 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.

3.2.2.4. 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.

3.2.3. 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 ».

3.2.3.1. 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.

3.2.3.2. 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.

3.2.3.3. 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.

4.3.2.3.4. 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. 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.

3.2.3.5. Type-annotateur_LA-HY-UD

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

3.2.4. 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).

3.2.4.1. 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.

3.2.4.2. 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.

3.2.4.3. 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.

3.2.5. 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.

3.2.5.1. 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. 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.

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.

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.

3.2.5.2. 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*) »[1].

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

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

3.2.5.3. 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.

3.2.5.4. 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.

3.2.6. 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'œuvre 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.

3.2.6.1. 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.

3.2.7. 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).

3.2.7.1. 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.

3.2.7.2. 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.

3.2.7.3. 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.

3.2.8. 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.

3.2.8.1. 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.

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.

3.2.8.2. 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.

3.2.8.3. 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.

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.

3.2.8.4. 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.

3.2.8.5. 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.

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.

3.2.8.6. 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.

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.

3.2.8.7. 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.

3.2.8.8. 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.

3.2.8.9 Incohérence-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.

3.2.8.9.1. Différents-termes-traduction_LA-TL-TI-DT

Le premier cas d'incohérence terminologique 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.

3.2.8.9.2. 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).

3.2.8.10. 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.

- **3.3. 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.

3.3.1. 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.

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.

3.3.3. 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.

3.3.4. 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 terminologie (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.

3.3.5. Type-annotateur_OU-UD

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