Preprint version
Article to appear in: ***Translation Spaces: A multidisciplinary, multimedia, and multilingual journal of translation***
https://doi.org/10.1075/ts.24002.bod

The *Machine Translation Post-Editing Annotation System* (MTPEAS): A standardized and user-friendly taxonomy for student post-editing quality assessment

Romane Bodart | Justine Piette | Marie-Aude Lefer
UCLouvain

Abstract

Machine translation post-editing quality evaluation has received relatively little attention in translation pedagogy to date. It is a time-consuming process that involves the comparison of three texts (source text, machine translation and student post-edited text) and the systematic identification and correction of students' edits (or absence thereof) of machine translation (MT) output. There are as yet no widely available, standardized, user-friendly annotation systems for use in translator education. In this article, we address this gap by describing the *Machine Translation Post-Editing Annotation System* (MTPEAS). MTPEAS includes a taxonomy of seven categories that are presented in easy-to-understand terms: Value-adding edits, Successful edits, Unnecessary edits, Incomplete edits, Error-introducing edits, Unsuccessful edits, and Missing edits. We then assess the robustness of the MTPEAS taxonomy in a pilot study of 30 students' post-edited texts and offer some preliminary findings on students' MT error identification and correction skills.

Keywords

machine translation post-editing, post-editing quality evaluation, translator education, post-editing training, annotation system, inter-rater agreement

1. *Introduction: post-editing quality evaluation in translator education*

In recent years, given rapid advances in machine translation (MT), MT post-editing (PE) training has become one of the key technological components of translator education (Plaza Lara 2019; Ginovart Cid and Colominas Ventura 2021). The updated version of the European Master's in Translation (EMT) Competence Framework, for example, acknowledges that

"MT literacy and awareness of MT's possibilities and limitations is an integral part of professional translation competence" (EMT Expert Group 2022, 7) and includes competences related to PE.

The first fully fledged pedagogical proposals to integrate PE training into translator education date back to the early 2000s-2010s (e.g. O'Brien 2002; Doherty and Kenny 2014; Kenny and Doherty 2014; Koponen 2015). The last decade has witnessed a wealth of proposals and initiatives directed at MT and PE training in translation programmes (see e.g. the contributions in Massey, Piotrowska and Marczak 2023 and in Froeliger, Larsonneur and Sofo 2023). As pointed out by Kenny (2020, 509) and Konttinen, Salmi and Koponen (2021, 189-190), two main approaches can be distinguished: (1) stand-alone translation technology modules targeting MT and PE, fully or in part (e.g. Guerberof and Moorkens 2019; Nitzke, Tardel and Hansen-Schirra 2019) and (2) PE practical tasks in language-pair-specific translation modules (e.g. Kübler, Mestivier and Pecman 2022). This shows that a balance needs to be struck in translator education between cross-cutting, general introductions to PE and language-pair- or domain-specific practical activities.

To integrate the two approaches and thereby achieve cross-curriculum implementation of PE training, as advocated among others by Mellinger (2017, 280), several hurdles need to be overcome by translation programme directors and trainers. One is the continuous professional development of translation trainers not yet familiar with MT and PE (see Rico and Gonzalez Pastor 2022). Another is the quality evaluation of students' productions, with a view to providing effective feedback and thereby boosting students' PE skills. While the quality evaluation of students' translations has received much attention in applied translation studies to date (e.g. Huertas Barros and Vine 2018),[1] especially in the field of learner translation corpus research (Granger and Lefer 2023, 16-18), the evaluation of student PE quality remains relatively under-researched in translation pedagogy. For example, while various error taxonomies have been developed to assess student translations (e.g. the MeLLANGE typology by Castagnoli et al. 2011), there are as yet no widely available, standardized, user-friendly PE annotation systems for use in translator education.

Despite this relative lack of applied research in translation pedagogy, the topic of human evaluation of PE quality has been explored empirically in numerous studies, with scholars relying on different typologies to assess post-editors' edits. An oft-mentioned taxonomy is de

[1] Note, however, that according to Doherty et al. (2018), the topic of translation quality assessment (TQA) itself is relatively neglected in translator education.

Almeida (2013)'s, where four categories are distinguished to characterize edits: 'Essential changes', 'Preferential changes', 'Essential changes not implemented', and 'Introduced errors'. In this typology, "a change is considered as essential when, if the change is not implemented, the sentence (or part of it) is either: a) Grammatically incorrect (i.e. it obviously breaches a grammatical rule specified in accepted grammar books), or b) Grammatically correct, but not accurate in comparison to the source text (i.e. it does not contain all the information that is present in the source text, or it contains extra information that is not present in the source text). Conversely, a change is considered preferential if the sentence from the raw MT output would still be grammatically correct, intelligible and accurate in relation to the source text, even if the change in question was not implemented" (ibid, 100). Essential changes not implemented result in sentences or segments that are grammatically incorrect or inaccurate with respect to the source text, because the raw MT output has not been edited (ibid). Finally, introduced errors are errors introduced by post-editors themselves, i.e. errors that were not originally present in the raw MT output (ibid, 101).

Another influential work in the field is Koponen and Salmi (2017), who approach PE quality along two dimensions, namely correctness and necessity, where correctness refers to edits that are "accurate in terms of both ST [source text] meaning and proper grammar and spelling of the TL [target language] form" and necessity means that "the change was needed to make the TL sentence comprehensible, grammatically correct or accurate in meaning" (ibid, 142)[2] (see also Koponen, Salmi and Nikulin, 2019). In their view, edits made to MT output can be classified as either correct or incorrect and as either necessary or unnecessary. This view gives rise to four categories of edit: correct and necessary, correct and unnecessary, incorrect and necessary, and incorrect and unnecessary. Koponen and Salmi (2017)'s typology has been further refined by Stasimioti and Sosoni (2021), who suggest adding five additional options, represented in bold in Figure 1 and defined in Table 1. These options are combined alongside the correctness and necessity dimensions to give the following six categories: Correct and necessary edit, Correct and necessary no edit, Incorrect and necessary edit, Incorrect and unnecessary edit, Edit missing and required, and Redundant and unnecessary edit.

INSERT FIG 1 ABOUT HERE

[2] It is important to recognize that the distinction between necessary and unnecessary (preferential) edits largely hinges on the type of post-editing being undertaken, such as light vs full post-editing, as outlined, for example, in post-editing guidelines (see Hu and Cadwell, 2016).

Correctness: **correct no edit**, correct edit, incorrect edit, **edit missing**, **redundant edit**
Necessity: **necessary no edit**, necessary edit, unnecessary edit, **edit required**

Figure 1: Categories added by Stasimioti and Sosoni (2021, 92) to Koponen and Salmi (2017)'s dimensions of correctness and necessity

---

INSERT TABLE 1 ABOUT HERE

Table 1: Definitions of the categories added by Stasimioti and Sosoni (2021, 92) to Koponen and Salmi's (2017) dimensions of correctness and necessity

Additional correctness categories	
Correct no edit	"the post-editor was right to leave the MT output unedited given that there was no error in the MT output"
Edit missing	"cases when an error in the MT was not corrected by the post-editor"
Redundant edit	"a preferential change made by the post-editor"
Additional necessity categories	
Necessary no edit	"there was no error in the MT output and therefore no edit was required"
Edit required	"there was an error in the MT which was not corrected by the post-editor"

Focusing on studies that examine student post-edited texts specifically, we can also mention Pavlović and Antunović (2021), who report on an English-to-Croatian post-editing experiment involving 44 translation students untrained in PE. Among necessary edits, the authors distinguish the following five categories: edits that provide an acceptable solution, edits that are themselves erroneous, edits that are partial, edits that introduce a new error, and missed necessary edits (i.e. essential changes not implemented, to use de Almeida 2013's terminology) (Pavlović and Antunović 2021, 192). As for unnecessary edits, Pavlović and Antunović use four categories: edits that introduce a solution that is more appropriate than the MT (a category that is not used in the studies discussed so far), edits that introduce a less appropriate solution than the MT, edits that can be considered neutral (i.e. that are neither better nor worse than the MT), and edits that result in a new error (ibid, 192-193). An interesting approach is also offered by Kübler et al. (2022, 251) in their study of errors related to complex noun phrases in student post-editing. They distinguish the following categories:

'Overconfidence in MT', when the incorrect MT output is left unchanged, 'Underconfidence in MT', when students modify correct MT output, and 'Failure to correct MT', when students have identified an error in the MT but have failed to correct it appropriately. Yamada (2019), by contrast, takes a more restrictive view and operationalizes post-editing quality as the proportion of MT errors detected and correctly revised by student post-editors, focusing only on major errors. Yet another option taken in empirical studies devoted to student PE is to rely on typologies which are commonly used in revision research, i.e. necessary changes, underrevisions (incorrect necessary changes), hyperrevisions (preferential changes) and overrevisions (introduced errors) (see e.g. Robert, Schrijver and Ureel 2023).

Finally, it is important to add that some PE quality studies do not rely on a systematic comparison of MT and PE. Instead, they assess post-edited texts like any other type of translation (whether human or machine), i.e. irrespective of edits made to raw MT output, relying for instance on manually computed scores for adequacy (accuracy) and fluency (acceptability) (see e.g. Castilho et al. 2018 on the use of these two constructs in translation quality assessment). This is typically done when post-edited texts are compared with human translations (e.g. Daems et al. 2017; Schumacher 2023). We are not in favour of adopting such an approach in translator education, given that it does not take into consideration the MT output students are asked to post-edit. PE training, at least in its first stages, is precisely about teaching students how to handle MT output to meet the quality requirements included in the PE brief. The former approach is better suited to training scenarios where MT output is fully integrated into the translation workflow alongside translation memories and terminological databases, for instance in a computer-aided translation environment. In such cases, it arguably becomes more complex and less relevant to go back systematically to the MT output to assess students' productions. The same holds true for interactive and adaptive MT PE scenarios.

As can be seen from this overview, which includes only a selection of representative examples, many different typologies are available with which to assess PE quality manually (a discussion of automatic methods is beyond the scope of this article). In our view, the time is ripe for the design and adoption of a PE quality evaluation system for translator education that would meet the following two criteria, which are largely inspired by feedback research in the field of education (Lipnevich and Panadero, 2021) and our own pedagogical practice:

(1) **standardization**, with a view to allowing comparability across courses, lecturers, educational settings and/or experiments. Standardization is closely linked with thorough documentation of the annotation system, taking the form of a detailed manual that provides definitions and examples for each category within the typology to be used;

(2) **user-friendliness**, for translation lecturers and students alike, which requires that there should be a well-described methodology to follow and that the labels used in the typology should be transparent and easy to use.

We believe that this type of endeavour can be beneficial for both training and (applied) research. For example, it would contribute to a faster take-up of language- and domain-specific PE training in practical translation courses, because trainers are encouraged to use a ready-made and well-described framework, and improve the comparability of the empirical results obtained from the collection and analysis of student post-edited texts. The aim of this article is to fill precisely this gap by (1) presenting the annotation system that we have developed to evaluate PE quality in translator education and (2) assessing the robustness of the annotation taxonomy on the basis of a pilot study. In the pilot study, we examine inter-rater agreement on quality as a proxy for the reliability of the system we have developed, assessing its usability with authentic data and the effectiveness of its taxonomy in capturing the quality features of student post-edited texts.

The remainder of the article is structured as follows. In Section 2, we describe the Machine Translation Post-Editing Annotation System and the corresponding taxonomy we use to annotate and assess students' post-edited texts. Section 3 is devoted to the data and methodology used for the pilot study, while Section 4 presents and discusses the results of the study, from the perspective of both inter-rater agreement and students' post-editing skills, as reflected in their productions. Section 5 concludes the article by summarizing our insights into PE quality evaluation in translator education and sketching some avenues for future work in the field.

2. *The Machine Translation Post-Editing Annotation System (MTPEAS)*

The Machine Translation Post-Editing Annotation System (MTPEAS) and its taxonomy were developed within the framework of the Post-Edit Me! Project (Lefer et al., 2023). They were designed as pedagogical tools aimed at translation lecturers and students in the context of PE training in translator education at Master's level. In this section, we describe the MTPEAS annotation system and its taxonomy.

2.1. *The MTPEAS annotation system*

The annotation system comprises two steps, both of which are executed at the text level: (1) the lecturer's detection of errors in the MT that students will be asked to post-edit and (2) evaluation of the final post-edited texts produced by students. During the MT error detection

stage, lecturers detect and tag the translation errors found in the MT output, before assigning the corresponding PE task to their students. We recommend the use of the Translation-oriented Annotation System for this step (Granger and Lefer, 2021; see Section 3.2). This preliminary annotation of the errors in the MT doubly facilitates the evaluation process. First, it allows lecturers to assess the level of difficulty of the PE task at hand and to adapt their PE guidelines if necessary. Second, it speeds up the correction process by allowing edits (or absence thereof) to be more easily and consistently identified in students' post-edited texts. As error detection in the MT is an integral part of the PE skill set that students are expected to acquire, this preliminary annotation is shared with the students only when they have completed the PE task and submitted their work (e.g. when they are given access to their lecturer's feedback). During the PE evaluation stage, lecturers can check that students have made the expected and necessary edits in their post-edited texts. Thanks to the MTPEAS taxonomy, lecturers can highlight the presence of students' edits in the PE, or the lack thereof, and determine whether the changes made to the MT are accurate or not.

2.2. *The MTPEAS taxonomy*

The MTPEAS taxonomy has been designed for pedagogical purposes and easy use by both translation lecturers and students across language pairs and domains. It is crucial to clarify from the beginning that in MTPEAS, we adopt the concept of a logical edit, in contrast to a mechanical edit, meaning that an edit represents a modification that makes sense linguistically (see Blain et al., 2011). In this approach, edits can involve changes to multiple words, thus incorporating a range of edit operations (insertions, deletions, substitutions, and shifts). Logical edits are therefore evaluated at the word level and beyond, including multiword terms and expressions, phrases, and clauses. Furthermore, this approach involves considering edits that arise as a result of propagation from another edit together with the originating edit (for example, a change in the gender of an adjective resulting from the substitution of a noun). The automatic extraction of logical edits can be quite challenging, a factor that does not pose a problem in our context, as edits are identified by lecturers.

Starting from existing taxonomies (see Section 1), we analysed a dozen authentic post-edited texts produced by our students in order to establish the taxonomy, which is as follows: (1) Value-adding edit (NE-PLUS); (2) Successful edit (E-OKAY); (3) Unnecessary edit (NE-SUPF); (4) Incomplete edit (E-INCO); (5) Error-introducing edit (NE-INTR); (6) Unsuccessful edit (E-FAIL); (7) Missing edit (E-MISS). The MTPEAS tags start with the

prefixes ‘E’ (Error) or ‘NE’ (No Error), which refer respectively to whether or not the corresponding segments[3] in the MT were tagged as erroneous by the lecturer during the MT annotation stage. The seven MTPEAS categories are defined in Figure 2, which categorizes edits according to their impact on the overall quality of the final products, distinguishing them as having a positive, neutral, or negative effect. These are represented by the colours green, yellow, and orange/red, respectively. The MTPEAS labels that are used to tag erroneous segments in the final post-edited texts can be combined with the Translation-oriented Annotation System (TAS) taxonomy (Granger and Lefer 2021) to provide students with valuable information about the nature of the remaining or introduced errors: Grammar and Syntax, Lexis and Terminology, Content, Mechanics, Discourse and Pragmatics, Register and Style, Culture, and Brief (see definitions in Section 3.2). The rationale for using TAS over widely adopted translation quality evaluation systems in the language services industry, such as DQF-MQM, lies in TAS’s design orientation towards pedagogical purposes. Specifically, it aims to aid students in enhancing their translation and linguistic skills. It is essential to highlight, however, that MTPEAS is compatible with any translation error taxonomy, allowing for flexible integration.

INSERT FIGURE 2 ABOUT HERE

---

[3] In this article, the term *segment* refers to punctuation marks, single words, and larger sequences (such as multiwords, word strings, phrases, or clauses) deemed worthy of quality-related annotation, whether in machine translations or in student post-edited texts. This usage does not align with the concept of segment as employed in the context of sentence-by-sentence segmentation in computer-aided translation tools.

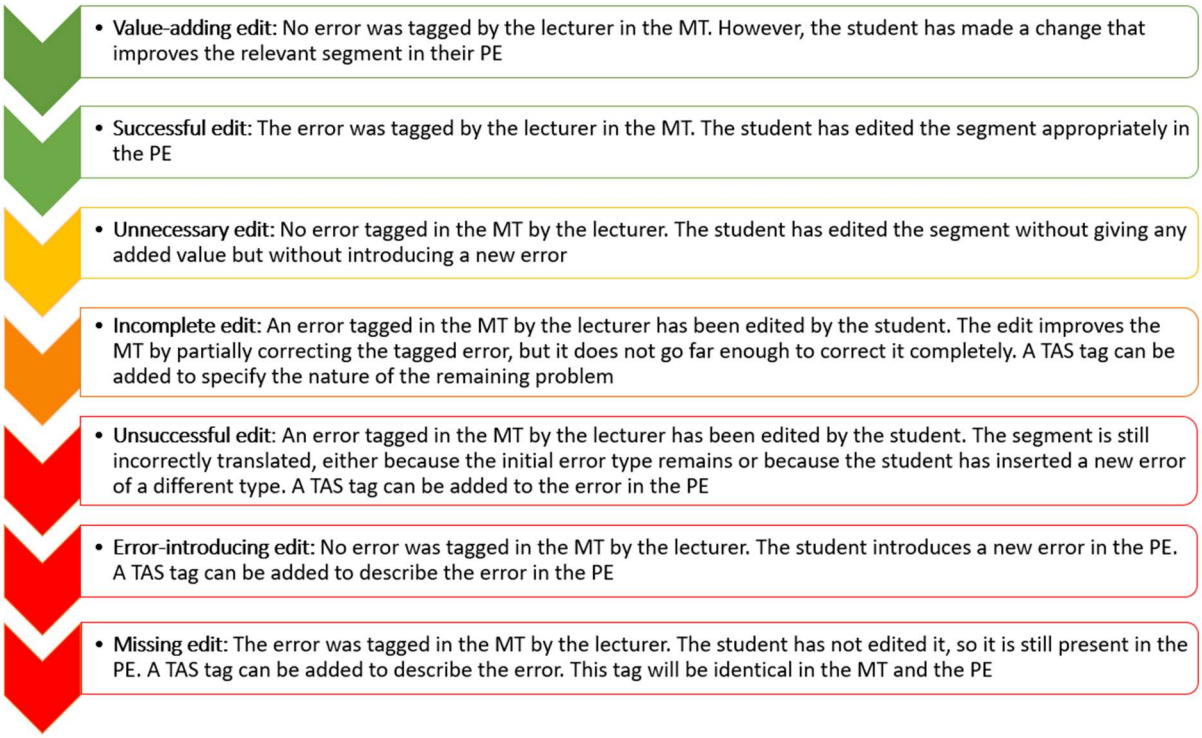

Figure 2: Definitions of MTPEAS categories – Colours correspond to the impact of edits on quality, with green indicating a positive impact, yellow neutral, and orange and red a negative impact

2.3. *The MTPEAS decision tree: a user-friendly, graphical representation of the annotation system*

The MTPEAS decision tree in Figure 3 represents graphically the PE correction process. Intended as a resource to help MTPEAS users familiarize themselves with the annotation system, it guides lecturers to the relevant error category by structuring the questions that typically arise when assessing the quality of a post-edited text.

INSERT FIGURE 3 ABOUT HERE

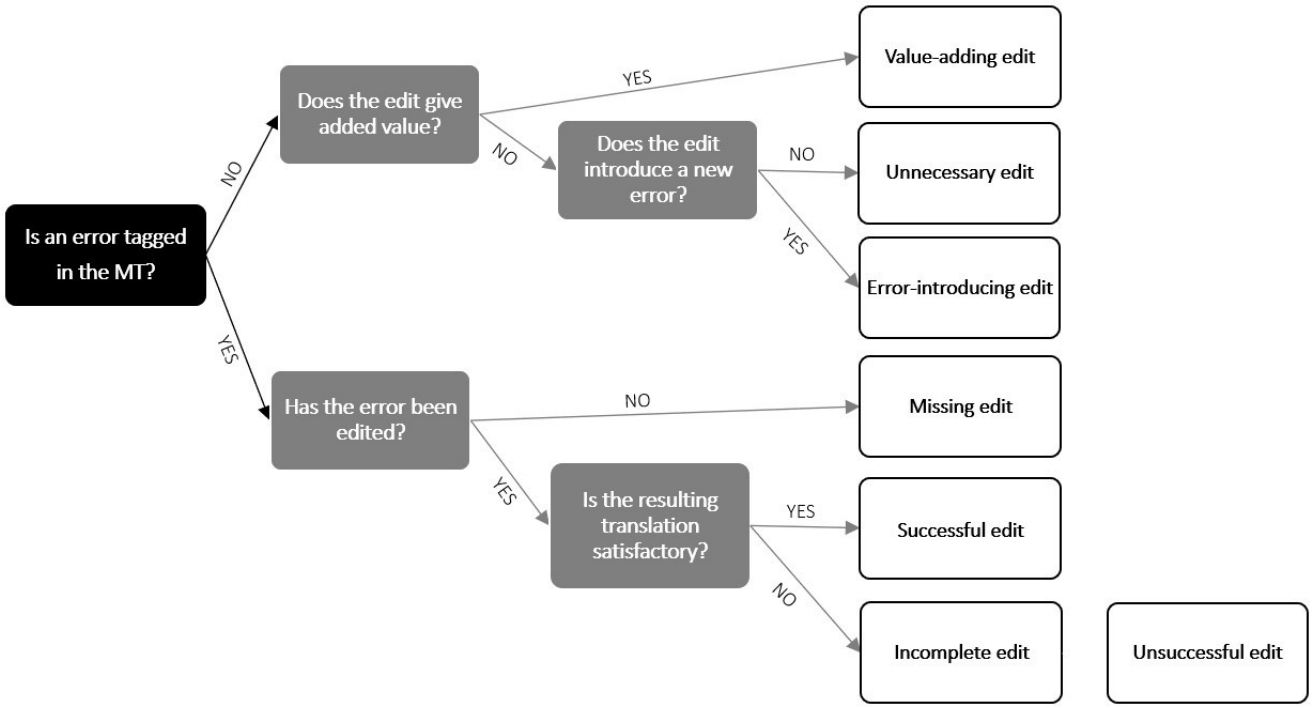

Figure 3: The MTPEAS decision tree

The starting point is the final PE: whenever the student has made a change to the MT, the lecturer checks to see if this edit corresponds to an error tagged in the MT ('Is an error tagged in the MT?'). If no error is tagged in the MT, the edit might still be justified ('Does the edit give added value?'). If it gives added value, this is a case of a **value-adding edit**. If it gives no added value, it may or may not introduce a new error ('Does the edit introduce a new error?'). If no new error is introduced, this is a case of an **unnecessary edit**. If a new error is introduced, it is a case of an **error-introducing edit**. Where there is a tagged error in the MT, the lecturer must ascertain whether it has been edited by the student ('Has the error been edited?'). If the student has left the MT error unchanged, this is a case of a **missing edit**. If the student has detected the error and changed it, the lecturer assesses the appropriateness of the edit ('Is the resulting translation satisfactory?'). An affirmative answer means there has been a **successful edit**. If the resulting translation is not satisfactory, an error remains in the MT despite the edit. The error is either due to an edit by the student that is justified but does not go far enough – an **incomplete edit** – or to a new error introduced by the student (whether or not similar in nature to the error in the MT) – an **unsuccessful edit**. In certain instances, categories may be combined to describe edits, as long as these categories are not mutually exclusive. For instance, if a word in the MT is replaced to correct a lexical error and the resulting translation is satisfactory, the edit will be labelled as Successful. Additionally, if that

specific word is also relocated within the sentence, the annotator revisits the decision tree. As a result, the edit receives an additional tag, for example, Value-adding if it contributes significant value to the final text by improving information structuring. The integration of MTPEAS with TAS enables a clearer understanding of why an edit is deemed both Successful (pertaining to Lexis and Terminology) and Value-adding (related to Discourse and Pragmatics). Edits that combine both correct and incorrect elements are marked as Incomplete, as they are still flawed to some degree. In essence, within MTPEAS, an edit cannot be simultaneously tagged as Successful and Unsuccessful.

3. *Pilot Study: Data and methodology*

In this section, we describe the data and methodology used for the pilot study, i.e. the post-editing data, the annotation methodology and the statistical test used.

3.1. *Data*

The pilot study is based on 30 English-to-French texts post-edited by second-year Master's students in translation at UCLouvain, all of them previously trained in MT and PE through a dedicated compulsory course in the first year of their Master's programme. The PE texts were produced in authentic classroom settings within the framework of two three-day intensive courses in legal and financial translation. For both PE tasks, students were given a detailed post-editing brief and asked to produce a full PE (publishable quality). The source texts were respectively entitled *What makes a good legal contract translation?* (674 tokens) and *Responsible investment strategies* (683 tokens). Students were all provided with the same machine translation of these texts, produced with the free version of DeepL in October 2022. The French machine translations contained 795 and 872 tokens respectively. Students were all familiar with the specialized topics, terminology and text types at hand, given that the intensive courses within which the data were collected were specifically devoted to contract translation and sustainable finance. The tasks were individual, and not marked (they were part of students' formative assessment). Students were allowed to use resources (e.g. glossaries, terminological databases, DIY specialized corpora uploaded on SketchEngine, Internet, CAT tools, MT) and were given three hours to complete the PE task, allowing them to work at their own pace. Given that at the time PE tasks were only beginning to be introduced in our translation classes, students were not

encouraged to use any particular PE software tool and chose to work in Microsoft Word. Table 2 provides an overview of the PE data used for the pilot study.

INSERT TABLE 2 ABOUT HERE

Table 2: PE data used for the pilot study

Domain	Number of PE	Total number of PE tokens
Financial translation	20	17466
Legal translation	10	7963

3.2. *Methodology*

The two machine translations used in the pilot study were error-annotated by two raters working independently (they were both trained translators and translator trainers). The raters relied on the first level of the error taxonomy of the multi-level *Translation-oriented Annotation System* (TAS) developed within the framework of the *Multilingual Student Translation* project (Granger and Lefer 2020, 2021).[4] TAS is a pedagogical translation-error taxonomy that consists of eight main categories: Content, Lexis and terminology, Grammar and syntax, Mechanics, Discourse and pragmatics, Register and style, Culture, and Translation brief. The categories are defined in Table 3.

INSERT TABLE 3 ABOUT HERE

Table 3: TAS error taxonomy used to tag MT errors (Granger and Lefer 2021)

Content (CT)	CT groups together all errors arising from a failure to reflect accurately the content of the source text (mistranslations, additions, omissions).
Lexis and terminology (LT)	LT subsumes problems with single words, multiwords and terms.

[4] The TAS annotation manual is available at https://cdn.uclouvain.be/groups/cms-editors-cecl/cecl-papers/TAS-2.0_annotation_manual_2021-10-26.pdf

Grammar and syntax (GR)	GR groups together sentence grammar errors and word grammar (i.e. lexico-grammatical) errors. It also includes inflectional morphology errors.
Mechanics (ME)	ME groups together all errors related to spelling, punctuation and typography.
Discourse and pragmatics (DP)	DP refers to issues related to cohesion and information structuring, such as the use of connectors and thematic progression.
Register and style (REG)	REG groups together errors resulting from the use of words or structures which do not fit with the communicative function(s) of the translation, and clumsy, awkward chunks of text that do not contain any errors as such but that one would want to reformulate. It also encompasses cases of redundant or unnecessary words that would be better left out.
Culture (CL)	CL refers to errors related to culture-specific references. It covers instances where explanations of realia or cultural concepts are missing. CL also subsumes cases where realia or proper names are translated when they should have been left untranslated, and vice versa.
Translation brief (TB)	TB refers to all cases of failure to comply with the specifications of the translation brief.

The two raters identified erroneous segments in the two machine translations and used TAS tags to specify the nature of the errors. Once they had finished, they shared their annotations and discussed them in order to agree on a set of MT errors that second-year Master's students should be able to detect and correct, given the curriculum's requirements, expected PE skills and the PE brief. In other words, they did not rely on an absolute notion of MT error but rather considered the pedagogical context in which students were asked to post-edit. In total, 49 segments were annotated as erroneous in the financial MT and 50 in the legal MT. The breakdown of MT errors per text is given in Table 4. Errors related to Register and Style are most prevalent, followed by Lexis and Terminology, and Discourse and Pragmatics. There were no MT errors related to Culture or Translation Brief.

INSERT TABLE 4 ABOUT HERE

Table 4: Breakdown of errors in the two machine translations used in the pilot study

Error category	Total errors	Total errors in legal MT	Total errors in financial MT	Examples
REG	42	23	19	ST: The simplest answer to this question **is that it's not recommended**.

				MT: *La réponse la plus simple à cette question* ***est que ce n'est pas recommandé****.* [clumsy formulation, translated literally from English] Correct translation: *Pour répondre simplement à cette question :* ***c'est déconseillé****.* ['it's unadvised']
LT	16	8	8	ST: A translation that obscures the meaning of key sentences or substitutes legal terms for less precise **alternatives** […]. MT: *Une traduction qui obscurcit le sens de phrases clés ou qui remplace des termes juridiques par des* ***alternatives*** *moins précises […].* Correct translation: *formulations* ['formulations']
DP	14	9	5	ST: The integration process focuses on the potential impact of ESG issues on company financials **(positive and negative)**, which in turn may affect the investment decision. MT: *Le processus d'intégration se concentre sur l'impact potentiel des questions ESG sur les finances de l'entreprise* ***(positif et négatif)****, qui peut à son tour affecter la décision d'investissement.* [wrong information structuring in French, the two adjectives modify the noun *impact* in the source text] Correct translation: *Le processus d'intégration se concentre sur l'impact potentiel* ***(positif et négatif)*** *des questions ESG sur les finances de l'entreprise […].*
ME	11	5	6	ST: (…) machine translation tools often fail to recognize the context and syntax of a sentence, which […] can render a text absolutely illegible – and thus, make your contract unsignable. MT: *(…) les outils de traduction automatique ne reconnaissent souvent pas le contexte et la syntaxe d'une phrase, ce qui, ajouté aux difficultés du langage juridique courant, peut rendre un texte absolument illisible* **-** *et donc, rendre votre contrat non signable.* [clumsy punctuation in French] Correct translation: […] *peut rendre un texte absolument illisible****,*** *et donc, rendre votre contrat non signable.* [use of a comma]
CT	8	1	7	ST: Impact investments can be made **in both emerging and developed markets** (…). MT: *Les investissements d'impact peuvent être réalisés* ***sur les marchés émergents et développés*** *(…).* [incorrect meaning, which does not reflect the nuance conveyed by *both* in the source text] Correct translation: […] *tant sur les marchés émergents que sur les marchés développés* ['both on the emerging markets and on the developed markets']
GR	8	4	4	ST: Thematic funds focus on specific or multiple issues related **to ESG**. MT: *Les fonds thématiques se concentrent sur des questions spécifiques ou multiples liées* ***à l'ESG****.* [wrong determiner, ESG is plural when it refers to ESG criteria, as in this context] Correct translation: […] *aux ESG.* [plural use]

In a second stage, the two raters annotated students' PE with the help of the MTPEAS taxonomy (see Section 2.2). Both had been actively involved in the development of the taxonomy and were therefore familiar with it. Again, they worked independently and each annotated the 30 post-edited texts in an Excel spreadsheet. The spreadsheets prepared by the two annotators were then compared and merged to allow for

computation of inter-rater agreement. The merged dataset contained a total of 2179 distinct annotations.

Two types of inter-rater agreement scores were computed: (1) percent agreement on PE segment identification, with a view to determining whether the raters had identified the same PE segments as being worthy of annotation, and (2) percent agreement on MTPEAS categories, to check whether, for segments that had been annotated by both raters, they had selected for the same MTPEAS category to annotate them (for this, we also relied on Kappa scores; see Kunilovskaya 2015 for a similar methodology). The inter-rater agreement scores serve as a critical measure for evaluating the robustness of the MTPEAS annotation system and its underlying taxonomy. High inter-rater agreement would indicate that the system is intuitive and the taxonomy is comprehensive, ensuring that different raters can consistently identify edits and classify them into the appropriate MTPEAS categories.

After this second step, the raters discussed cases of disagreement in order to resolve them (as regards both segment identification and MTPEAS annotation). A third rater was called in to resolve the most problematic cases. This was done in group discussions at which the three raters were present. After discussion, it was decided that segments that had been identified as Unnecessary edits (NE-SUPF) by one rater and as Value-adding edits by the other (NE-PLUS) would neither be further discussed nor resolved. The reason for this was that the distinction between the two categories, which both pertain to preferential changes, involves the highest degree of subjectivity in post-editing quality assessment. Consequently, it became apparent that disambiguating these cases was practically unfeasible. They were therefore coded as NE-SUPF/PLUS. When all other cases had been resolved, we obtained a final dataset of 2125 MTPEAS annotations.

4. *Pilot study: results and discussion*

In this section, we discuss inter-rater agreement on segment identification and MTPEAS annotation before sharing some preliminary findings.

4.1. *Inter-rater agreement on segment identification*

First, we examined inter-rater agreement on PE segment identification with a view to determining whether the two raters agreed on the segments to be annotated with the

MTPEAS taxonomy in students' post-edited texts. As shown in Table 5, the raters agreed on 82% of the segments to be annotated (n=1778/2179), with practically no difference between the two subject areas under scrutiny. This figure, which is much higher than percent agreements usually reported for translation evaluation (see e.g. Kunilovskaya 2015), shows that segment identification in PE quality evaluation is a fairly straightforward task. For the remaining 18% (n=401/2179), one of the raters considered them worthy of annotation, while the other rater left the segments unannotated.

INSERT TABLE 5 ABOUT HERE

Table 5: Inter-rater agreement on PE segment identification (percent agreement)

	All	Legal	Financial
Agreement	1778 [82%]	641 [81%]	1137 [82%]
Disagreement	401 [18%]	149 [19%]	252 [18%]
Total	2179 [100%]	790 [100%]	1389 [100%]

A close examination of the data shows that the cases of disagreement were mainly due to lapses of concentration, for instance in the case of repeated errors within the same post-edited text. It is important to recall here that the raters worked in an Excel spreadsheet and that the edits in the PE texts had not been automatically identified in any way prior to the annotation in Excel. Raters had to identify the segments to be annotated completely manually (as is often the case in translator education). The top three categories that were overlooked by one of the raters were the following: Missing edits (n=126), Unnecessary edits (n=76) and Value-adding edits (n=57) (see examples in Table 6).

INSERT TABLE 6 ABOUT HERE

Table 6: Examples of disagreement on PE segment identification: cases where only one of the raters identified a segment as worthy of annotation

	ST	MT	PE

Missing edits	An approach that excludes specific investments or classes of investment from the investible universe such as companies, sectors, or countries.	Une approche qui exclut des investissements ou des catégories d'investissements spécifiques de l'univers investissable, **tels que** des entreprises, des secteurs ou des pays. [DP error tagged in the MT: unclear reference]	Une approche qui exclut des investissements ou des catégories d'investissements spécifiques de l'univers investissable, **tels que** des entreprises, des secteurs ou des pays. [MT error still present in the PE]
Unnecessary edits	Aside from the aforementioned obstacles, depending on the languages and countries of origin of the parties involved, legal contract translation may pose the following challenges: (…).	Outre les obstacles susmentionnés, en fonction des langues et des pays d'origine des parties concernées, la traduction de contrats juridiques peut poser les défis suivants : (…).	Outre les obstacles susmentionnés, **et** en fonction des langues et pays d'origine des parties concernées, la traduction de contrats juridiques peut poser les défis suivants : (…). [addition of *et* 'and']
Value-adding edits	[…] can give one of the signing parties an **erred** understanding of the agreement.	[…] peut donner à l'une des parties signataires une compréhension **erronée** de l'accord.	[…] peut engendrer une **mauvaise** compréhension de l'accord par l'une des parties signataires. [this is a more idiomatic collocation]

4.2. *Inter-rater agreement on MTPEAS annotation*

We then focused on the segments that had been identified by both raters as being worthy of annotation (n=1778) to check whether the raters agreed on the MTPEAS category to be used to tag these segments. For this part of the pilot study, we relied on percent agreement and Cohen's Kappa (Landis & Koch 1977) to compute inter-rater agreement scores. The results indicate that raters agreed on MTPEAS categorization in 76% of the cases (n= 1359/1778; kappa=0.66; substantial agreement), with notable differences between the two texts under scrutiny: 69% in the legal text (kappa=0.60; moderate agreement) and 80% in the financial text (kappa=0.69; substantial agreement) (see Table 7).

INSERT TABLE 7 ABOUT HERE

Table 7: Inter-rater agreement on MTPEAS annotation (percent agreement and Kappa scores)

	All	Legal	Financial
Agreement	1359 [76%]	445 [69%]	914 [80%]
Disagreement	419 [24%]	196 [31%]	223 [20%]
Total	1778 [100%]	641 [100%]	1137 [100%]
Kappa score	0,66 *substantial agreement*	0,60 *moderate agreement*	0,69 *substantial agreement*

The agreement scores, which are quite high overall, suggest that the MTPEAS taxonomy is a reliable tool for assessing the quality of students' post-edited texts. The consistency observed among raters indicates that the system is robust, confirming that the categories are well-defined and the guidelines sufficiently clear to handle the complexity of the data being annotated. However, it is important to acknowledge that the following sets of MTPEAS categories led to cases of disagreement: (1) Value-adding edits vs Unnecessary edits (n=95), (2) Error-introducing edits vs Unnecessary edits (n=74) and (3) Unsuccessful edits vs Successful edits (n=73) (see Table 8). Examples are provided in Table 9.

INSERT TABLE 8 ABOUT HERE

Table 8: Sets of MTPEAS categories with the highest rates of disagreement

	All	Financial	Legal
Value-adding edits vs Unnecessary edits	95	53	42
Error-introducing edits vs Unnecessary edits	74	33	41
Unsuccessful edits vs Successful edits	73	42	31

INSERT TABLE 9 ABOUT HERE

Table 9: Examples of sets of MTPEAS categories with the highest rates of disagreement

	ST	MT	PE
Value-adding edits vs Unnecessary edits	But legal translations can't be automated from beginning to end, **because** machine translation tools often fail to recognize […]	Mais les traductions juridiques ne peuvent pas être automatisées du début à la fin, **car** les outils de traduction automatique ne reconnaissent souvent pas […] ['because', correct in the MT]	Mais les traductions juridiques ne peuvent pas être automatisées du début à la fin. **En effet, souvent,** les outils de traduction automatique […] [MT sentence split into two parts + fronting of the adverb *souvent*]
Error-introducing edits vs Unnecessary edits	Social issues vary from community-related aspects, such as the improvement of health and education, to workplace-related issues, […].	Les questions sociales vont des aspects liés à la communauté, tels que l'amélioration de la santé et de l'éducation, aux questions liées au lieu de travail […].	Les questions sociales concernent des aspects liés à la communauté, tels que l'amélioration de la santé et de l'éducation, **mais aussi** des questions liées au lieu de travail, […]. [addition of *mais aussi* 'but also', which puts emphasis on the element listed]
Unsuccessful edits vs Successful edits	This type covers explicit consideration of ESG factors **alongside** financial factors in the mainstream analysis of investments.	Ce type couvre la prise en compte explicite des facteurs ESG **à côté des** facteurs financiers dans l'analyse générale des investissements. ['next to the']	Cette stratégie couvre la prise en compte explicite des facteurs ESG **au même titre que** celle des facteurs financiers dans l'analyse générale des investissements. ['in the same way as']

Unsurprisingly, raters mainly disagreed on Value-adding vs Unnecessary edits, as choosing between these two categories is quite subjective. As regards the other two causes of disagreement (Error-introducing vs Unnecessary and Unsuccessful vs Successful edits), these were related to lexis and terminology, content, and register and

style (n=118; see examples in Table 10). Very few instances of grammatical phenomena and mechanics feature among these cases of disagreement, given that they lend themselves to fairly objective decisions relating to linguistic rules in the target language (here, French) (n=21).

INSERT TABLE 10 ABOUT HERE

Table 10: Examples of Content, Lexis and terminology, and Register and style phenomena giving rise to disagreement on MTPEAS category

	Content	Lexis and terminology	Register and style
Error-introducing vs Unnecessary edits	ST: Formatting **is also an important part** of legal contract translation. MT: Le formatage **est également une partie importante** de la traduction de contrats juridiques. ['is also an important part', correct in the MT] PE: La mise en forme **fait partie intégrante** de la traduction de contrats juridiques. ['forms an integral part']	ST: Formatting is also an important **part** of legal contract translation. MT: Le formatage est également une **partie** importante de la traduction de contrats juridiques. ['part', correct in the MT] PE: La mise en forme est également une **étape** importante de la traduction de contrats juridiques. ['step']	ST: can produce **costly** inaccuracies. MT: peut entraîner des inexactitudes **coûteuses.** ['costly', correct in the MT] PE: peut entraîner des inexactitudes **qui vont vous coûter cher.** ['that will cost you a lot of money', informal]
Unsuccessful vs Successful edits	ST: Stable, **well-functioning** and well governed social, environmental and economic systems MT: Systèmes sociaux, environnementaux et économiques stables, **fonctionnant bien** et bien gouvernés ['functioning well', tagged as a CT error in the MT] PE: Systèmes sociaux, environnementaux et économiques stables, **opérationnels** et efficaces ['operational']	ST: Legal translators ought to be up-to-date with the latest **developments** of the legal systems at hand. MT: Les traducteurs juridiques doivent être au courant des derniers **développements** des systèmes juridiques concernés. ['developments', tagged as an LT error in the MT] PE: Les traducteurs juridiques doivent être au courant des dernières **avancées** des systèmes juridiques concernés. ['breakthroughs']	ST: **Examples include** greenhouse gas emissions […]. MT: **Les exemples incluent** les émissions de gaz à effet de serre […]. ['The examples include', tagged as a REG error in the MT] PE: **Il s'agit par exemple** des émissions de gaz à effet de serre […]. ['It is the case for example']

4.3. *Preliminary findings*

As mentioned in Section 3.2, cases of disagreement were resolved in order to obtain a final dataset. This final dataset, which contains 2125 annotated segments, allows us to formulate some tentative initial observations about our participants' post-editing skills. A general overview of our results is given in Table 11.

INSERT TABLE 11 ABOUT HERE

Table 11: General overview of the results (final dataset, n = 2125)

	Legal	Financial	All	All (%)
E-MISS	301	709	1010	47.5
E-OKAY	120	171	291	13.7
NE-SUPF	93	134	227	10.7
NE-INTR	92	104	196	9.2
NE-PLUS	56	93	149	7.0
E-FAIL	46	53	99	4.7
NE-PLUS/SUPF	42	53	95	4.5
E-INCO	38	20	58	2.7
TOTAL	788	1337	2125	100

First, it can be seen that Missing edits (E-MISS) make up almost half of the dataset, while Successful edits (E-OKAY) represent only 13.7% of the annotations. Figure 3 shows Successful edits in relation to the total number of errors in the MT. On average, students corrected about one fifth of the MT errors successfully, which is a very low figure compared with Yamada (2019), for instance, who reports 68%. This substantial discrepancy can be attributed to Yamada's analysis focusing solely on major MT errors. It is important to note that our students' profiles differ quite dramatically: the proportion of appropriately corrected MT errors varies from 4% to 38% across students. These preliminary results suggest that MT error identification cannot be taken for granted in translator education, not even with final-year Master's students

who have been formally trained in post-editing in a stand-alone module (see Pavlović and Antunović 2021, 193 for a similar observation regarding translation students untrained in PE). This clearly illustrates the need for cross-curriculum PE practice in translator training.

INSERT FIGURE 3 ABOUT HERE

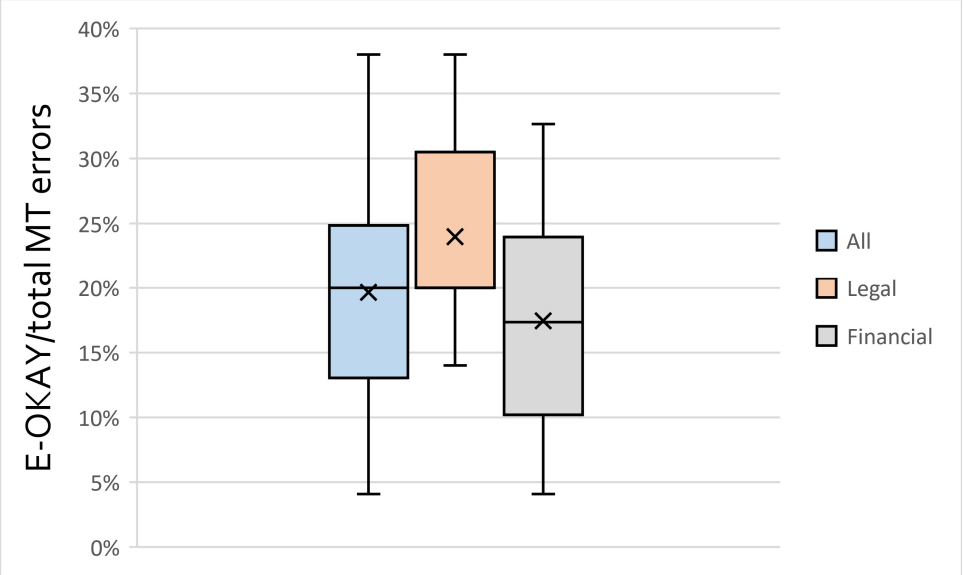

Figure 3: Successful edits (E-OKAY) out of the total of MT errors

We also examined the number of erroneous segments that were still present in the final, post-edited products, namely the aggregate sum of Error-introducing (NE-INTR), Unsuccessful (E-FAIL) and Missing (E-MISS) edits. On average, there are more than 50 remaining translation errors per 1000 tokens. These were either introduced by students where the MT was error-free or were already present in the MT and either left unchanged or wrongly edited. Once again, there is individual variation, ranging from 43 to 71 erroneous segments per post-edited text (see Figure 4).

INSERT FIGURE 4 ABOUT HERE

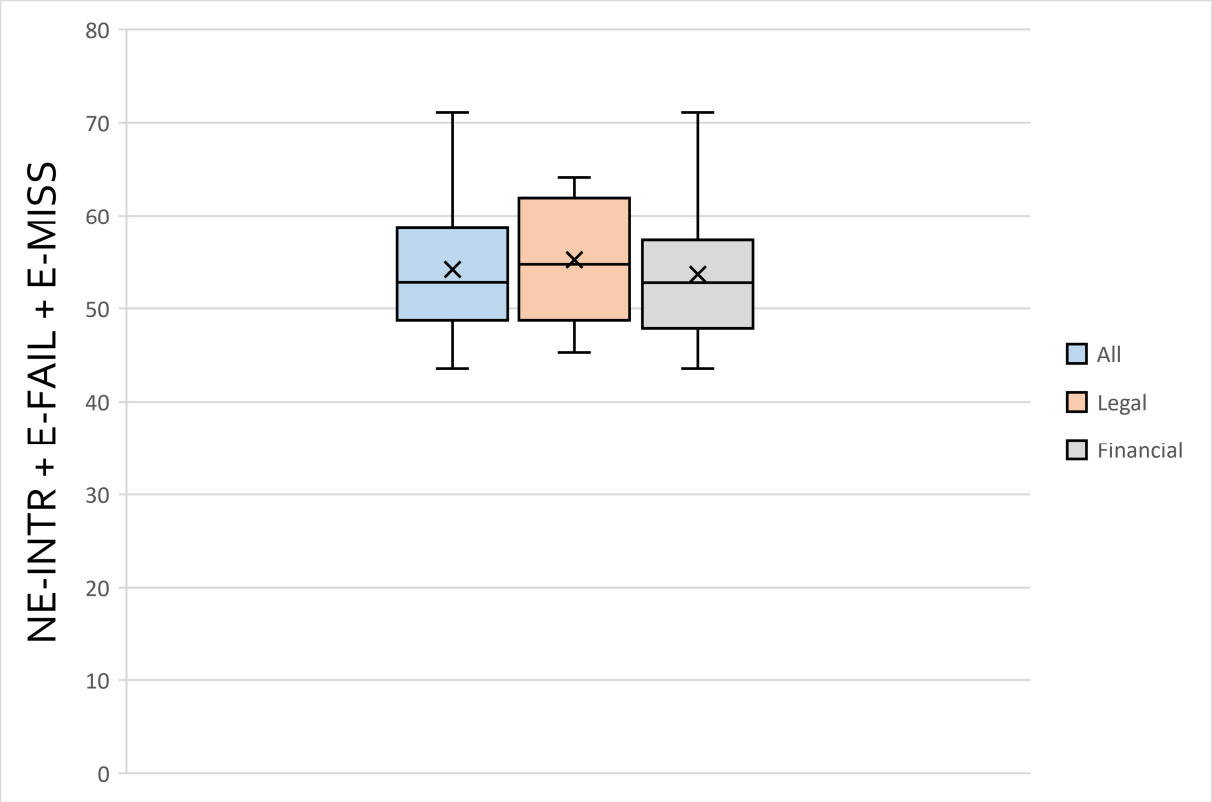

Figure 4: Erroneous segments in the final post-edited texts, per 1000 tokens

Figure 5 provides a breakdown of erroneous segments (E-MISS, NE-INTR and E-FAIL). What we see here is that most errors are in fact MT errors that went undetected (E-MISS): there are about 40 MT errors left per 1000 tokens, with little variation across the two tasks. The second source of erroneous segments is introduced errors, i.e. cases where the MT output was correct but wrongly edited. These are clear cases of students worsening the quality of the final product. They represent ca 10 erroneous segments per 1000 tokens, with strong variation across students (from 1 to 24 introduced errors). Cases of Unsuccessful edits (E-FAIL) come last, and are quite rare. These are cases where students have detected the MT error but introduced a wrong solution into the final product (6 E-FAIL per 1000 tokens, on average). This means that when students detect MT errors, they usually edit them appropriately.

These results lead us to conclude that, taken overall, the quality of the final post-edited products is inferior to raw MT quality. Should these preliminary trends be confirmed by larger-scale studies and replicated for other language pairs, domains and cohorts, this would call into question the viability of the 'junior post-editor' profile currently advertised in the language services industry (see e.g. Gene & Guerrero 2022) and

indicate that PE tasks, very much like revision of human translations, can be handled much more successfully by experienced language professionals.

INSERT FIGURE 5 ABOUT HERE

Figure 5: Breakdown of erroneous segments in the final post-edited texts, per 1000 tokens

A final note is in order here regarding preferential changes. In the MTPEAS taxonomy, we decided to distinguish between Unnecessary edits (NE-SUPF) and Value-adding edits (NE-PLUS) with a view to empowering translation students. However, we were well aware when integrating the distinction into MTPEAS that the categorization would be highly subjective: what one rater sees as adding value may be perceived as superfluous by another. As mentioned in Section 3.2, when resolving cases of disagreement between the two raters it was decided that cases where a segment was tagged as an Unnecessary edit by one rater and as a Value-adding edit by the other would not be resolved. Instead, we classified these instances as NE-SUPF/PLUS. They are relatively rare (4 edits per 1000 tokens).
Unnecessary and Value-adding edits taken together represent 42.2% of all edits in the two PE tasks (legal and financial) (see Table 12). This corresponds closely to figures reported by Koponen and Salmi (2017) (38% of unnecessary edits) and Pavlović and

Antunović (2021) (40% of unnecessary edits) in experiments involving translation students (for the English-Finnish and English-Croatian language pairs, respectively).

INSERT TABLE 12 ABOUT HERE

Table 12: Value-adding and Unnecessary edits vs other edits (excluding Missing edits)

	Number of edits	%
NE-SUPF, NE-PLUS & NE-PLUS/SUPF	471	42.2
Other edits	644	57.8
Total	1115	100

Figure 6 shows only NE-PLUS and NE-SUPF cases on which raters agreed. What emerges from our dataset is that there is marked individual variation when it comes to Unnecessary edits tagged as such by the two raters (where we can assume with some certainty that the edits were indeed superfluous): in the financial task, for instance, these vary from 1 to 25. An example of an unnecessary edit is the substitution of *tests sur les animaux* ('test on the animals') by *essais sur les animaux* ('trials on the animals') to render *animal testing*. Even though unequivocal Value-adding edits are quite rare (around 4 per 1000 tokens), we believe that they are worth tagging as such in order to identify cases where students have managed to improve the MT output significantly even though it was not erroneous in the first place. For instance, to render *the quirks of legal prose*, a student edited *les bizarreries de la prose juridique* ('oddities') to *la singularité de la prose juridique* ('singularity'), which is more appropriate stylistically. Future research will show whether this is correlated with a relative lack of productivity (e.g. in terms of temporal effort).

INSERT FIGURE 6 ABOUT HERE

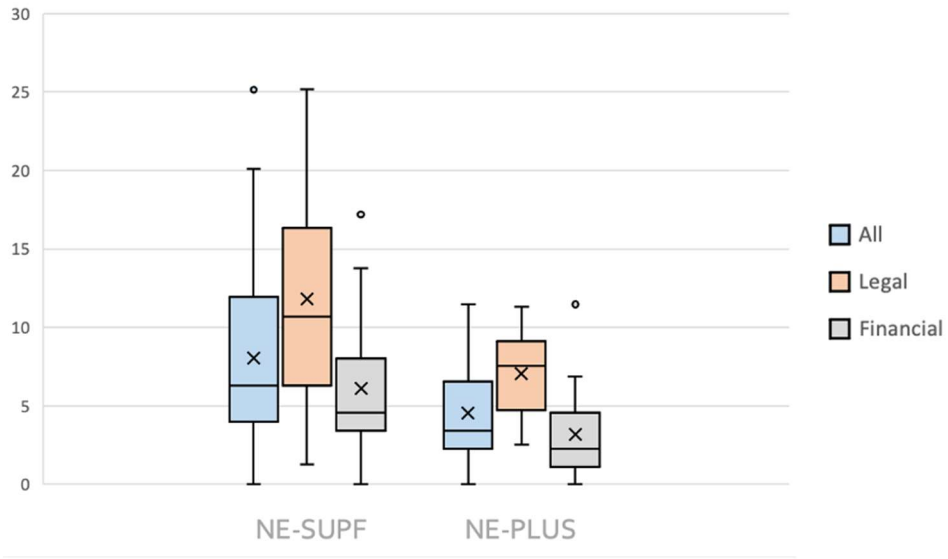

Figure 6: Preferential changes – Unequivocal Value-adding edits and Unnecessary edits, per 1000 tokens

5. *Concluding remarks*

In this article, we have presented the MTPEAS taxonomy and applied it to a first sample of 30 authentic post-edited texts produced by final-year Master's students in translation working on the English-to-French language pair. Our pilot study shows that the taxonomy is quite robust, as it produces quite satisfactory inter-rater reliability scores, and is fine-grained enough to reveal interesting trends in texts post-edited by Master's students. Furthermore, despite its small scale, the study reveals that students overlook a significant portion of MT errors and introduce errors when post-editing. This finding challenges the common perception within the language services industry that post-editing is a straightforward task suitable for junior linguists.

It is our hope that the MTPEAS system will be used by other PE trainers and will foster new insights into PE training in translator education, be they related to PE competence models, PE cross-curriculum implementation or PE materials design. A considerable number of issues remain open at this stage and need to be explored in the future. These include evaluation of the effectiveness of MTPEAS-based feedback. It is important to examine whether students understand the feedback provided by their lecturers in the form of MTPEAS annotations and whether it helps them to overcome their particular problems and come up with appropriate translation solutions. This could be done through think-aloud exercises that would be recorded

and analysed, along the same lines as Strobl (2022). Another key pedagogical question, from a longitudinal perspective, is whether students' PE skills improve over time if lecturers offer MTPEAS-based structured feedback on a regular basis, across language pairs and domains, which allows specific student profiles to emerge on the basis of annotation statistics that are cumulatively generated for each student. It would also be helpful to design teaching materials on the basis of the trends that emerge from the annotation statistics (such as databases of MT errors that typically go unnoticed by students or typical cases of unnecessary edits). Finally, on a more practical level, there is a pressing need to automatize the PE quality assessment process and generation of statistics, including edit distance metrics and automatic grade calculations. To achieve this goal, a web-based application named postedit.me has been developed (Lefer et al., under review). It relies on the MTPEAS taxonomy introduced in this article and is freely accessible to universities for teaching and internal research purposes.

Funding

This article was supported by the UCLouvain - Fonds de Développement Pédagogique.

References

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. "Qualitative analysis of post-editing for high quality machine translation." In *Proceedings of Machine Translation Summit XIII: Papers*, 164-171.

Castagnoli, Sara, Dragos Ciobanu, Natalie Kübler, Kerstin Kunz, and Alexandra Volanschi. 2011. "Designing a learner translator corpus for training purposes." In *Corpora, language, teaching, and resources: From theory to practice*, edited by Natalie Kübler, 221-248. Peter Lang.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. "Approaches to Human and Machine Translation Quality Assessment." In *Translation Quality Assessment*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 9-38. Cham: Springer.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, and Lieve Macken. 2017. "Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators." *Meta* 62(2):245-270. Accessed January 19, 2024. DOI 10.7202/1041023ar.

De Almeida, Giselle. 2013. *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages.* PhD Thesis, Dublin City University.

Doherty, Stephen, and Dorothy Kenny. 2014. “The design and evaluation of a Statistical Machine Translation syllabus for translation students.” *The Interpreter and Translator Trainer* 8 (2): 295-315.

Doherty, Stephen, Joss Moorkens, Federico Gaspari, and Sheila Castilho. 2018. “On education and training in translation quality assessment.” In *Translation Quality Assessment*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 95-106. Cham: Springer.

EMT Expert Group. 2022. “Competence Framework”. Accessed January 19, 2024. https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf

Froeliger, Nicolas, Claire Larsonneur, and Giuseppe Sofo (eds). 2023. *Traduction humaine et traitement automatique des langues: Vers un nouveau consensus?/Human Translation and Natural Language Processing: Forging a New Consensus?* Studi e ricerche 35. Edizioni Ca’ Foscari - Venice University Press: Venice.

Gene, Viveta, and Lucía Guerrero. 2022. “A Common Machine Translation Post-Editing Training Handbook for Academia, Clients, LSPs and Post-Editors.” GALA MTPE Training Special Interest Group.

Ginovart Cid, Clara, and Carme Colominas Ventura. 2021. “The MT post-editing skill set: Course descriptions and educators’ thoughts.” In *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian S. Mossop, Isabelle Robert, and Giovanna Schocchera, 226-246. London/New York: Routledge.

Granger, Sylviane, and Marie-Aude Lefer. 2023. “Learner translation corpora: Bridging the gap between learner corpus research and corpus-based translation studies.” *International Journal of Learner Corpus Research* 9 (1): 1-28.

Granger, Sylviane, and Marie-Aude Lefer. 2021. *“Translation-oriented Annotation System manual (Version 2.0).” CECL Papers 3*. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain. Accessed January 19, 2024. https://uclouvain.be/en/research-institutes/ilc/cecl/cecl-papers.html

Granger, Sylviane, and Marie-Aude Lefer. 2020. “The Multilingual Student Translation corpus: A resource for translation teaching and research.” *Language Resources and Evaluation* 54: 1183-1199.

Guerberof, Ana, and Joss Moorkens. 2019. “Machine translation and post-editing training as part of a master’s programme.” *The Journal of Specialised Translation* 31: 217-238.

Hu, Ke, and Patrick Cadwell. 2016. “A Comparative Study of Post-editing Guidelines.” *Baltic J. Modern Computing* 4 (2), 346-353.

Huertas Barros, Elsa, and Juliet Vine (eds). 2018. *New Perspectives in Assessment in Translator Training*. London: Routledge.

Kenny, Dorothy. 2020. “Technology and translator training." In *The Routledge Handbook of Translation and Technology*, edited by Minako O’Hagan, 498-515, London: Routledge.

Kenny, Dorothy and Doherty Stephen. 2014. “Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators.” *The Interpreter and Translator Trainer* 8 (2): 276-294.

Konttinen, Kalle, Leena Salmi, and Maarit Koponen. 2021. “Revision and post-editing competences in translation education.” In *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian S. Mossop, Isabelle Robert, and Giovanna Schocchera, 187-202. London/New York: Routledge.

Koponen, Maarit. 2015. “How to teach machine translation post-editing? Experiences from a post-editing course.” In *Proceedings of 4th Workshop on post-editing technology and practice, Miami, USA, October 30 to November 3, 2015, 2-15*. Accessed January 19, 2024. https://mt-archive.net/15/MTS-2015-W1-Koponen.pdf

Koponen, Maarit and Leena Salmi. 2017. “Post-editing quality: Analysing the correctness and necessity of post-editor corrections.” *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 16: 137-148.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. 2019. “A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output.” *Machine Translation* 33 (1): 61-90.

Kübler, Natalie, Alexandra Mestivier, and Mojca Pecman. 2022. “Using comparable corpora for translating and post-editing complex noun phrases in specialized texts. Insights from English-to-French specialized translation.” In *Extending the Scope of Corpus-Based Translation Studies*, edited by Sylviane Granger and Marie-Aude Lefer, 237-266. London: Bloomsbury.

Kunilovskaya, Maria. 2015. “How far do we agree on the quality of translation?” *English Studies at NBU* 1 (1): 18-31.

Landis, J. Richard, and Koch Gary G. 1977. “The measurement of observer agreement for categorical data.” *Biometrics* 33: 159-174.

Lefer, Marie-Aude, Romane Bodart, Adam Obrusník, and Justine Piette. 2023. The Post-Edit Me! project. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 493-494.

Lefer, Marie-Aude, Romane Bodart, Justine Piette, and Adam Obrusník. 2024, under review. “MTPE quality evaluation in translator education: the postedit.me app.” EAMT2024, Sheffield.

Lipnevich, Anastasiya A., and Ernesto Panadero. (2021). “A review of feedback models and theories: Descriptions, definitions, and conclusions.” *Frontiers in Education* 6: 720195.

Massey, Gary, Maria Piotrowska, and Mariusz Marczak. 2023. “(Re-) profiling T&I education: meeting evolution with innovation”. *The Interpreter and Translator Trainer* 17 (3): 325-331. Accessed January 19, 2024. DOI 10.1080/1750399X.2023.2237321

Mellinger, Christopher D. 2017. “Translators and machine translation: knowledge and skills gaps in translator pedagogy.” *The Interpreter and Translator Trainer* 11 (4): 280-293.

Nitzke, Jean, Anke Tardel, and Silvia Hansen-Schirra. 2019. “Training the modern translator – the acquisition of digital competencies through blended learning”. *The Interpreter and Translator Trainer* 13 (3): 292-306.

O’Brien, Sharon. 2002. “Teaching post-editing: a proposal for course content.” In *Proceedings of the 6th EAMT Workshop: Teaching machine translation, Manchester, England, November 14-15, 2002*, 99-106. Accessed January 19, 2024. https://aclanthology.org/2002.eamt-1.11.pdf

Pavlović, Nataša, and Goranka Antunović. 2021. “Towards acquiring post-editing abilities through research-informed practical tasks.” *Strani jezici: časopis za primijenjenu lingvistiku* 50 (2): 185-205.

Plaza Lara, C. 2019. “SWOT Analysis of the Inclusion of Machine Translation and Post - Editing in the Master's Degrees Offered in the EMT Network.” *The Journal of Specialised Translation* 31: 260-280.

Rico, Celia, and Diana Gonzalez Pastor. 2022. “The Role of Machine Translation in Translation Education: A Thematic Analysis of Translator Educators’ Beliefs.” *Translation & Interpreting* 14 (1): 177-197.

Robert, Isabelle S., Iris Schrijver, and Jim JJ Ureel. 2023. “Comparing L2 translation, translation revision, and post-editing competences in translation trainees: An exploratory study into Dutch–French translation.” *Babel* 69 (1): 99-128.

Schumacher, Perrine. 2023. *La post-édition de traduction automatique en contexte d’apprentissage.* PhD Thesis, University of Liège and University of Geneva.

Stasimioti, Maria, and Vilelmini Sosoni. 2021. “Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair.” In *Translation, interpreting, cognition: The way out of the box*, edited by Tra&Co Group, 79–104. Berlin: Language Science Press. DOI: 10.5281/zenodo.4545037

Strobl, Carola. 2022. “Revision behaviour of translation students in response to electronic feedback. Insights from a pilot study using Hypal4MUST.” Paper presented at *EST conference 2022*, Oslo, Norway, 22-25 June 2022.

Yamada, Masaru. 2019. “The impact of Google neural machine translation on post-editing by student translators.” *The Journal of Specialised Translation* 31 (1): 87-106.

Address for correspondence

Marie-Aude Lefer
UCLouvain
Place Cardinal Mercier 14/L3.06.03
1348 Louvain-la-Neuve

Belgium
marie-aude.lefer@uclouvain.be
0000-0002-6172-6150

Co-author information

Romane Bodart
UCLouvain
romane.bodart@uclouvain.be

Justine Piette
UCLouvain
justine.piette@uclouvain.be