Empowering future language professionals: Findings from a classroom experiment on MT Quality Evaluation in collaboration with DGT

Romane Bodart
UCLouvain
romane.bodart@uclouvain.be

Christine Pasquier
UCLouvain
christine.pasquier@uclouvain.be

Marie-Aude Lefer
UCLouvain
marie-aude.lefer@uclouvain.be

Abstract

Recent advances in translator education encompass technology training, including machine translation (MT) (e.g. the updated 2022 version of the European Master's in Translation competence framework). In this paper, we present the findings of a hands-on teaching unit focused on human evaluation of MT quality through error annotation. The teaching unit was part of a larger international project led by the European Commission's Directorate-General for Translation (DGT). DGT's objectives for this project were to disseminate DGT's methodology for translation quality evaluation, gather empirical data on the quality of eTranslation outputs, and collect feedback for further enhancing eTranslation (DGT 2024). Our objectives as lecturers were to train translation students in machine translation evaluation using a real-life error taxonomy and, more broadly, to equip them as language professionals who will interact with MT output in their future careers. Specifically, we aimed to make them aware of the current limitations of neural MT systems through extensive hands-on experience. Thirty-two students took part in the project and worked in pairs to assess the quality of the machine-translated texts they were assigned. In this article, we analyze the results obtained and we take stock of the teaching unit.

1 Introduction: machine translation quality evaluation in translator education

Recent developments in translator education have increasingly incorporated technology training, including machine translation (MT). As MT continues to evolve, driven by advances in artificial intelligence and large language models, and becomes increasingly prominent in the language services industry, it is crucial for translation students to acquire the skills needed to effectively post-edit machine-generated translations. Building on O'Brien's work (2002), several authors, such as Doherty and Kenny (2014), Koponen (2015), Mellinger (2017), and Guerberof Arenas and Moorkens (2019), have designed course frameworks that incorporate MT and post-editing (PE) into translation curricula.

O'Brien (2002) proposed a structured approach that integrates both theoretical and practical MT components into translator training. In her view, students must not only understand how MT works but also be able to identify and address its shortcomings. This foundational work set the stage for subsequent researchers, such as Doherty and Kenny (2014), Koponen (2015), Mellinger (2017), and Guerberof Arenas and Moorkens (2019), who developed similar course frameworks. These frameworks feature a clear division between theoretical instruction and practical exercises. The theoretical portion often covers a wide array of topics, including the history and development of MT and PE, controlled languages, pre-editing for MT, common MT errors, MT quality evaluation, PE types, post-editing quality, PE effort, and PE skills.

On the practical side, researchers consistently emphasize the importance of hands-on experience with MT and PE, both during classroom hours and through independent work. Practical activities range from comparing different MT outputs to performing pre-editing tasks and evaluating MT quality. O'Brien (2002) and Doherty and Kenny (2014) place particular

emphasis on students developing programming skills to understand the mechanics of MT. For example, Doherty and Kenny (2014) require students to compute automatic evaluation metrics, an exercise that helps them gain a deeper understanding of MT performance. Meanwhile, Guerberof Arenas and Moorkens (2019) and Koponen (2015) focus more on the linguistic aspects of post-editing, such as identifying errors in the MT output and correcting them accurately.

Moorkens (2018) conducted a classroom experiment comparing statistical and neural MT systems, using three quality assurance metrics: adequacy, post-editing productivity (measured by temporal effort), and error annotation. This experiment underscored the complexity of evaluating MT quality. It is especially relevant as new MT types, including systems based on large language models (LLMs), emerge. As the technology progresses, it has become clear that regardless of the type of MT system – whether statistical, neural, or LLM-based – students need to develop a strong capacity for MT quality evaluation.

The importance of teaching MT quality evaluation is further highlighted in a 2021 study by Ginovart Cid and Colominas Ventura. This study surveyed 53 translation educators from European Master's in Translation (EMT) institutions to explore current practices related to PE training. Of the respondents, 45 indicated that they teach human MT evaluation as part of their courses, while 27 reported that they also cover automatic evaluation metrics. This consensus reflects a broad recognition among trainers that students must be able to assess MT quality and detect MT errors, irrespective of the specific MT technology being used.

Automatic quality metrics, such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), chrF (Popovic, 2015), and COMET (Rei et al., 2020), are frequently introduced to students as tools for evaluating MT output. These metrics are often praised for their objectivity and efficiency, as they provide a quick, automated assessment of translation quality. However, they are also criticized for being less detailed and less accurate than manual evaluation. While automatic metrics can provide a useful starting point, they do not replace the need for human evaluation, which requires students to engage more deeply with the translation output and to identify specific error types, which in turn improves their post-editing performance.

Recognizing the importance of error detection, researchers have developed a variety of error classification schemes tailored specifically to MT. These include the schemes proposed by Llitjós et al. (2005), Farrús et al. (2010), Bojar (2011), Kirchoff et al. (2012), Comelles et al. (2012), Federico et al. (2014), and Castilho et al. (2017). Although these taxonomies were primarily designed for MT, they can also be applied to human translation. They typically classify errors into categories such as grammar, terminology, and content. One widely used taxonomy is the Multidimensional Quality Metrics (MQM) framework (2014), which is popular among language service providers, even though it was not specifically created for MT error annotation. These error classification schemes are valuable tools for translation trainers aiming to address human MT evaluation, as they offer structured frameworks for assessing quality.

Several applied studies have examined the ability of translation students to detect and correct errors in MT output, revealing varying levels of accuracy and highlighting key challenges in post-editing across different language pairs and text types. For instance, Koponen and Salmi (2017) conducted a study to assess the accuracy and necessity of post-editing corrections made by five Finnish-speaking translation students. The participants, four master's students and one bachelor's student, were tasked with post-editing English-to-Finnish machine translations. The researchers found that students failed to detect and correct 3% of the MT errors, a relatively low rate compared with other studies. However, among the errors that were identified, 9% were incorrectly edited. In these cases, students either failed to correct the errors

(34% of the time) or introduced new errors (66% of the time). This highlights the dual challenge students face: detecting errors in MT output and making the appropriate corrections.

In contrast, Pavlović and Antunović (2021) conducted a study with 44 students working with the English-Croatian language pair and reported a much higher error detection failure rate. Students missed 30% of the errors in the MT output, and 12% of the corrections they made failed to improve the translation. This rate of missed errors is comparable to that found by Kübler et al. (2022), who analyzed how students handled complex noun phrases in specialized texts during post-editing. Their study showed that students failed to detect errors 31% of the time, often due to overconfidence in the MT output. Additionally, students failed to correct the MT appropriately in 49% of cases.

Bodart, Piette, and Lefer (2024) reported an even higher error detection failure rate in their study, which involved analyzing 30 post-edited texts produced by master's students working with the French-English language pair in the financial and legal domains. The researchers found that students missed 47.5% of MT errors and failed to correct errors appropriately in 4.7% of cases. These findings show that error detection remains a persistent challenge for students, particularly when working with complex or specialized texts.

Overall, this overview highlights two key trends in empirical research. First, students consistently struggle with detecting errors in MT output. This difficulty is compounded by the fact that errors in MT can be subtle and may not be immediately obvious. Second, once students successfully identify an error, they are generally able to correct it effectively. The exception to this trend is seen in Kübler et al.'s study, where students detected more errors but were unable to correct them appropriately.

Given these empirical findings, it is clear that MT error detection and classification is a critical component of technology training in translator education. Translation lecturers must ensure that students are equipped with the tools and skills necessary to identify MT errors and to correct them accurately. By focusing on error detection in both theoretical and practical training, translation programs can help students overcome one of the most significant challenges they face in post-editing machine translations.

2 Background: a collaborative project with the European Commission's DGT

In this paper, we present the results of a practical teaching unit designed to engage students in the human evaluation of MT quality through error annotation. This teaching unit, conducted during the spring semester of 2023, formed part of a broader international initiative coordinated by the Directorate-General for Translation (DGT) of the European Commission. The project involved collaboration between DGT and six European Master's in Translation (EMT) universities, including ours, with a focus on assessing the quality of MT outputs produced by eTranslation's General Text engine, specifically in connection with news articles published on the website of the Joint Research Centre (JRC).

The objectives set by DGT for this project included dissemination of its established methodology for translation quality evaluation, collection of empirical data on the quality of eTranslation outputs, and gathering feedback to inform future improvements to the system (DGT 2024). As lecturers, our primary goal was to train translation students in the evaluation of machine translation by applying a real-world error taxonomy. Additionally, we aimed to equip them with the skills needed to critically assess and work with MT output, an increasingly vital aspect of their future roles as language professionals. Through this hands-on experience, we sought to deepen students' understanding of the current limitations of neural machine translation systems, thus preparing them for the challenges they may face as translators in a technology-driven environment.

3 Data and methodology

3.1 MT error taxonomy used

At the start of the project, DGT provided materials related to its internal error taxonomy, including relevant documentation and access to two e-learning modules. The DGT's taxonomy, largely based on the MQM framework, categorizes errors into six types: accuracy, terminology, linguistic norms, job-specific style, general style, and design (DGT 2020) (see Figure 1).

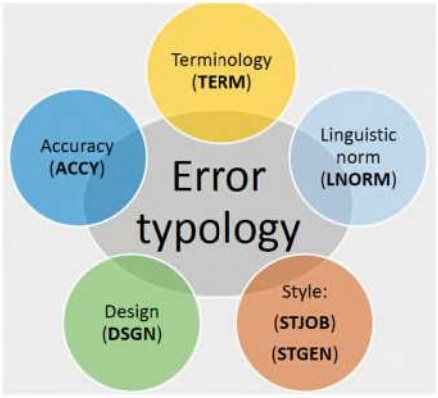

Figure 1: DGT's error typology (DGT 2020)

Accuracy errors (ACCY) relate to the content or meaning between the source and target text, such as mistranslations (e.g. distortion, mismatched names of places or numbers, ambiguity), additions, omissions, or untranslated content. Terminological errors (TERM) arise when the translation does not adhere to accepted terminology within a specific domain or does not comply with a term base or reference document provided by the contractor. DGT defines a term as "a lexical unit comprising one or more words that corresponds to a concept in a particular subject field or application area. Terms are used for expert communication and in that sense are different from purely linguistic and/or stylistic expressions" (DGT 2020:4). Linguistic norm errors (LNORM) concern the linguistic "well-formedness" of the text, which can be evaluated independently of whether the text is a translation. These errors involve formal language aspects, such as grammar, punctuation, and spelling, which are governed by linguistic norms. Style errors refer to grammatically and linguistically correct formulations that are nevertheless inappropriate because they deviate from organizational style guides or job-specific instructions (STJOB), or they exhibit an inappropriate general style, such as tone, register, and stylistic appropriateness (STGEN). Design errors (DSGN) pertain to the presentation of the translated product, including issues such as text or paragraph formatting, layout, the proper integration of graphical elements, and mark-up. This category excludes typographical and stylistic errors. Design errors can be identified either within a document (e.g. a second-level heading formatted as a first-level heading) or in relation to the source text (e.g. inconsistently formatted headings between the source and target). DGT developed a decision tree to assist users in selecting the appropriate error category. It is graphically represented in Figure 2. Examples of each category are provided in Table 1.

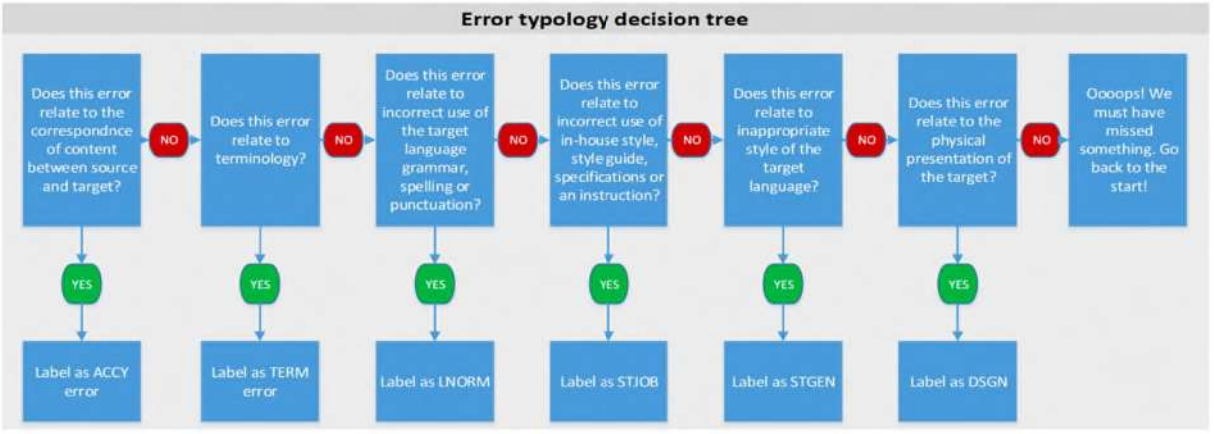

Figure 2: DGT's error typology decision tree (DGT 2020)

Error category	Example	Correct version
ACCY	*ST: The first Zero Pollution Outlook by the Joint Research Centre analyses whether the EU is on track to reach its zero pollution targets* ***with current and newly proposed EU policies****.* *MT: Les premières perspectives de pollution zéro par le Centre commun de recherche analysent si l'UE est sur la bonne voie pour atteindre ses objectifs de zéro pollution* ***avec les politiques actuelles et proposées par l'UE****.*	[…] *l'UE est sur la bonne voie pour atteindre ses objectifs de zéro pollution* ***avec les politiques actuelles et les nouvelles politiques proposées par l'UE****.*
TERM	*ST: In the new CAP (starting in 2023), the share of* ***UAA*** *that will receive CAP support for organic farming is higher.* *MT: Dans la nouvelle PAC (à partir de 2023), la part des* ***UAA*** *qui bénéficieront d'un soutien de la PAC pour l'agriculture biologique est plus élevée.*	*UAA = utilised agricultural area* ***SAU*** *= superficie agricole utile*
LNORM	*ST: […] 27 national digital contact tracing apps (21 Member States, 2 EAA countries, Switzerland and United Kingdom (3 apps)) were examined in this study […]* *MT: […] 27 applications nationales de traçage numérique des contacts (21 États membres, 2 pays de l'EEE, la Suisse et le Royaume-Uni (3 applications) ont été examinées dans cette étude […]*	*[…] 27 applications nationales de traçage numérique des contacts (21 États membres, 2 pays de l'EEE, la Suisse et le Royaume-Uni (3 applications)) ont été examinées dans cette étude […]*
STJOB	*ST: More than* ***100*** *organisations provided feedback on the interim report and the feedback received was generally positive.* *MT: Plus de* ***100*** *organisations ont fourni des commentaires sur le rapport provisoire et les commentaires reçus ont été généralement positifs.*	*Plus de* ***cent*** *organisations …* Cf. Code de rédaction interinstitutionnel (2022), section 10.4

STGEN	*ST: Organic farms have lower yields* ***on average*** *[...]* *MT: Les exploitations biologiques ont des rendements* ***en moyenne*** *inférieurs [...]*	***En moyenne,*** *les exploitations biologiques ont des rendements inférieurs [...]*
DSGN	*ST: All EU Member States stand united in the determination that any form of* ***racism,*** ***antisemitism*** *and* ***hatred*** *have no place in Europe.* *MT: Tous les États membres de l'UE sont unis dans leur détermination à ce que toute forme de racisme, d'antisémitisme et de haine n'ait pas sa place en Europe.*	Bold face is missing in the MT

Table 1: Examples of DGT's error categories

3.2 Data used: workbooks prepared by DGT

At the start of the project, it was decided to provide students with full texts for analysis, allowing MT quality to be evaluated at text level rather than segment level, as is often the case in MT evaluation campaigns. The JRC news texts selected for assessment by DGT were systematically organized into eight separate Excel workbooks, each dedicated to a specific topic: biodiversity, creating digital society, organic farming, security, trade agreements, public health, digital education, and sustainable finance. Within each workbook, news items were provided on separate spreadsheets, totaling approximately 200 segments. In total, across all eight workbooks, there were 56 full texts comprising 1,514 segments.

For ease of comparison and error identification, the English source texts were aligned side by side with their corresponding French machine translations. Alongside these, additional columns were provided for annotating any errors using the DGT's error typology, as well as for providing comments or clarifications as needed. This layout facilitated detailed tracking and assessment of MT quality.

Each workbook also included a set of comprehensive instructions, a translation brief, and links to the original news items available online, enabling students to consult the broader context of each article as necessary. Additionally, to encourage accountability, the first spreadsheet in each workbook was devoted to a logbook where students were required to indicate the amount of time they had spent on the project. This time log also captured whether the students had worked individually or collaborated. This structure aimed to ensure a well-documented approach to the MT quality evaluation task.

3.3 Outline of the teaching unit

The teaching unit was embedded within a first-year master's course focused on revision and post-editing. Thirty-two students participated in Spring 2023, working in pairs. At the outset of the project, all students signed an informed consent form. Each Excel workbook was assigned to two pairs of students who worked independently on the same material. The structure of the teaching unit followed a multi-step process, represented in Figure 3: (1) a short e-learning module on DGT's error taxonomy, (2) an in-class introductory session led by the three lecturers, which included the project presentation, a detailed overview of DGT's taxonomy, and an introductory MT quality evaluation exercise, (3) a more comprehensive e-learning module, (4) pair work on MT outputs, (5) an in-class Q&A session to address challenges in identifying and annotating MT errors, and (6) pair work to finalize the assignment.

Figure 3: Workflow of the teaching unit

In the first phase (1), students were introduced to DGT's error taxonomy through a brief e-learning module, which provided an overview of its core principles and concepts. This module, designed to be completed in 15 minutes, included short explanatory videos and quizzes to test comprehension. Students were required to complete this module before attending the in-class session (2), during which the three lecturers presented the project and provided further insights into DGT's taxonomy, complementing the e-learning material. They also organized an introductory evaluation exercise that mirrored the students' final task: identifying MT errors, categorizing them using DGT's taxonomy, and performing post-editing (this last PE step was not part of the collaborative project with DGT, but was added by the three lecturers). Following this, students were assigned a second e-learning module (3), estimated to take around one hour, which delved deeper into DGT's taxonomy. Each error category was covered in a dedicated video, with examples provided for clarity, followed by exercises to check understanding. Once students felt confident in their grasp of the material, they were expected to collaborate in pairs on their assigned workbooks (4). A Q&A session (5) was held in class a few weeks later, providing an opportunity to address any difficulties encountered during the initial stages of pair work. Students shared the challenges they faced in identifying and annotating errors, allowing for a collective discussion and clarification of concepts. Finally, students were given several weeks to complete their assignments outside of class (6), ensuring they had sufficient time to apply the knowledge gained and finalize their work before submission.

4 Results and discussion

The full dataset comprised 56 texts, consisting of 1,514 segments and a total of 30,800 words. Based on the students' evaluations, the proportion of machine-translated segments without any errors, as assessed by the two pairs of students annotating the same workbook, was 20.5% (n=311/1,514; see Table 2). For the remaining 1,203 segments, students identified 2,653 errors in total, i.e. an average of 2.2 errors per segment.

Total number of MT segments	1,514 (100%)
Error-free MT segments	311 (20.5%)
Erroneous MT segments	1,203 (79.5%)

Table 2: Error-free vs erroneous MT segments in the workbooks (full dataset)

As shown in Figure 4, when considering all annotated MT errors, regardless of whether they were identified by one or both pairs of students, stylistic errors emerged as the most frequent, accounting for roughly one-third of all MT errors. These were followed by errors related to linguistic norms and accuracy, with terminology errors ranking just behind them. Design-related errors remained marginal, representing only a small fraction of the total errors.

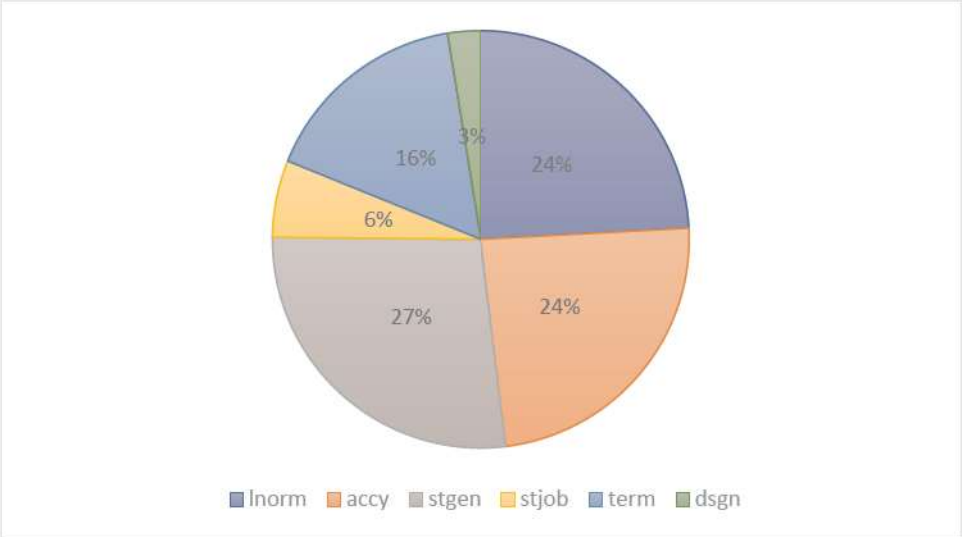

Figure 4: Categories of MT errors (full dataset; n=2,653 MT errors)

However, it is crucial to highlight that of these 2,653 errors, only about one-fourth (i.e. 648 errors) were consistently identified by both pairs of students analyzing the same workbook. These 648 errors were spread across 501 MT segments, representing one-third of the MT data analyzed (see Table 3). Unlike the higher figure presented in Table 2 (79.5% of erroneous MT segments), here, we can assert with a high level of confidence that these 501 segments are indeed erroneous, given students' agreement.

Total number of MT segments	1,514 (100%)
Error-free MT segments	1,013 (66.9%)
Erroneous MT segments	501 (33.1%)

Table 3: Error-free vs erroneous MT segments in the workbooks (cases of agreement across student pairs)

As can be seen in Figure 5, among the 648 errors consistently identified by both pairs of students evaluating the same MTs, approximately one-third were related to accuracy (30%), meaning that source-text content was not transferred appropriately. This was followed by errors relating to linguistic norms (23%) and terminological errors (23%). Errors related to general style accounted for 14%, while task-specific stylistic errors made up 8%, and design-related errors were the least frequent at 2%. These findings, combined with the trends represented in Figure 4 above, suggest that while there was some consistency in identifying critical errors such as accuracy, linguistic norms, and terminology, style-related categories were less consistently annotated.

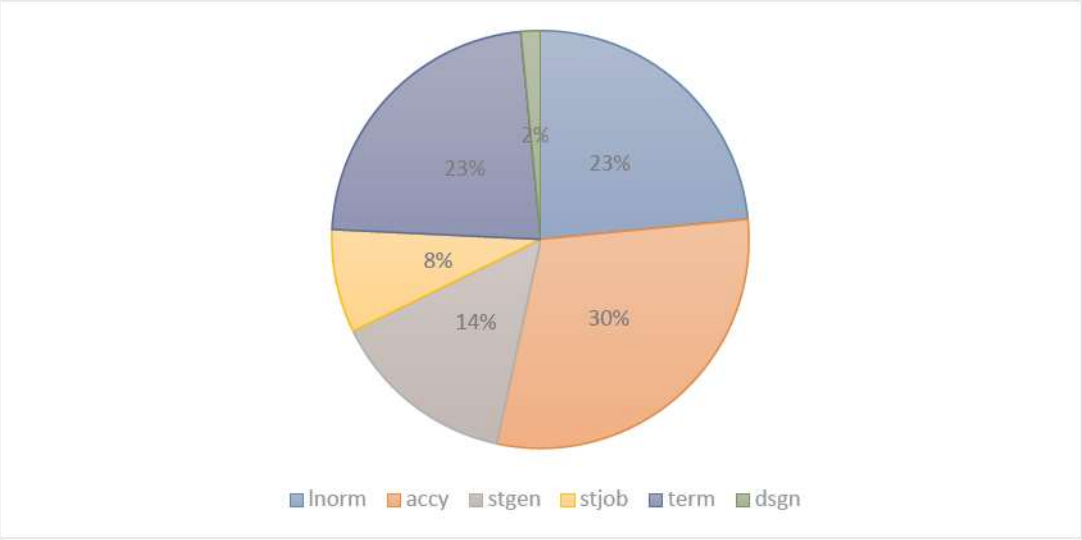

Figure 5: Categories of the errors identified by both pairs of students evaluating the same workbook (n=648 MT errors)

Interestingly, as shown in Table 4, there was substantial variation in the results obtained for the different workbooks provided to students. Depending on the workbook (and, hence, the topic at hand), different error types – linguistic norms, accuracy, terminology, or general style – occupied the top position. This variability suggests that eTranslation generated different error types depending on the sets of source texts used as input. For instance, terminological errors rank first in the economic texts related to trade agreements and sustainable finance, while accuracy is the most problematic area for the texts that deal with organic farming and security.

	Biodiversity	Creating digital society	Organic farming	Security	Trade agreements	Public health	Digital education	Sustainable finance
LNORM	10	27	7	9	16	11	66	6
ACCY	8	25	32	17	15	34	39	24
STGEN	11	9	21	6	5	11	20	10
STJOB	0	0	6	0	0	3	13	30
TERM	4	15	15	6	21	36	14	36
DSGN	0	3	0	4	0	0	3	0

Table 4: Breakdown per workbook of the errors identified by both student pairs (n=648 MT errors)

Our dataset includes 2,005 MT errors identified by only one of the two student pairs. These cases fall into two main categories: (1) MT errors that one pair overlooked, and (2) words, expressions, or constructions incorrectly identified as erroneous by one pair, despite being appropriate in French. Among these 2,005 errors, those related to general style were the most

frequent, comprising nearly one-third of the discrepancies (31%). Errors associated with linguistic norms ranked second, accounting for 25%, while accuracy-related errors followed closely at 22%. Together, these three categories represent the most common types of error observed by only one student pair (see Figure 6). In addition to these, terminological errors made up 14% of the remaining discrepancies, while job-specific stylistic errors accounted for 5%, and design-related errors were the least frequent, constituting 3%. These findings reveal the varying degrees of student attention to different aspects of translation quality, particularly those involving general style, linguistic norms, and accuracy. The variability in detecting stylistic errors probably reflects the inherently subjective nature of style, making it more prone to variation in assessment. By contrast, the high number of segments identified as erroneous by only one of the two student pairs in the other error areas confirms the critical need for thorough training in MT quality evaluation.

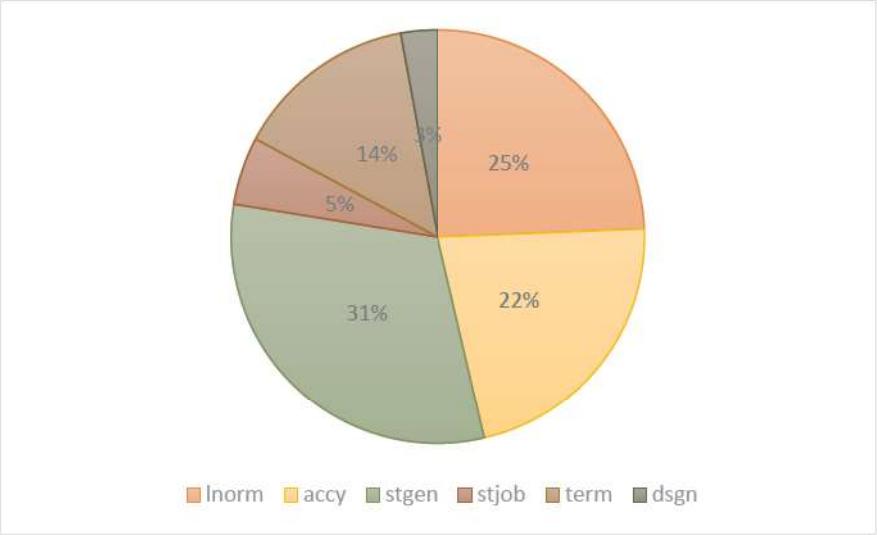

Figure 6: Categories of the errors identified by only one of the two student pairs evaluating the same workbook (n=2,005 MT errors)

Overall, our results show that human MT quality evaluation remains a challenging task for first-year master's students. They further indicate that it is necessary to integrate large-scale hands-on training into translation curricula in order to help students develop the critical evaluation skills required to interact effectively with MT outputs.

5 Student experience: retrospective questionnaire

Following the project, students were invited to complete a questionnaire, which we developed collaboratively with DGT, to provide feedback on their experience. Sixteen of the 32 UCLouvain students who participated in the project responded to the questionnaire. All respondents found the project to be engaging and valuable. They reported that it helped them learn to distinguish between different error types. Additionally, they noted the significance of evaluating whether certain edits are truly necessary during the post-editing process, among other insights.

However, one commonly reported drawback was the project's time demands. Many students felt that it was too time-consuming, with most reporting that they spent over 20 hours

on completing the project (average time spent: 24 and a half hours; see Table 5). One group, in particular, reported spending nearly 46 hours on the evaluation task.

Biodiversity	Pair 1	19h25min
	Pair 2	16h58min
Creating digital society	Pair 1	21h55min
	Pair 2	14h45min
Organic farming	Pair 1	26h15min
	Pair 2	25h45min
Security	Pair 1	15h05min
	Pair 2	13h40min
Trade agreements	Pair 1	34h17min
	Pair 2	19h23min
Public health	Pair 1	22h25min
	Pair 2	38h54min
Digital education	Pair 1	19h40min
	Pair 2	26h54min
Sustainable finance	Pair 1	31h25min
	Pair 2	45h54min
Average		**24h32min**

Table 5: Time spent on the evaluation task (self-reported, on the basis of a project logbook)

In the questionnaire, students were also asked whether before participating in the project they had feared that MT engines might replace human translators, and if their views had changed after completing the project. Almost half of respondents (7 out of 16) indicated that they had not been concerned about MT replacing human translators, and their opinion remained unchanged after the project. However, nine participants replied that they were initially concerned that MT engines could eventually replace human translators. Three of these changed their minds after completing the project, while six maintained their initial concern (see Table 6). The students who see a promising future for human translators provided several reasons for their optimism. They highlighted the poor quality of the MT outputs, noting that MT engines are not yet sufficiently advanced to compete with human translators, particularly because they cannot fully grasp the nuances of language in the way a human can. Additionally, students pointed out that MT engines still produce a considerable number of errors, including gender bias, incoherence, and repetitions, further emphasizing the ongoing need for human oversight and expertise in the translation process.

Number of students	Opinion before and after the project
7	No → No
3	Yes → No
6	Yes → Yes

Table 6: Students' opinions as to whether MT engines will replace human translators

The primary reason cited by the six students who believe that MT engines will eventually replace human translators is the expectation that MT technology will continue to improve over time. However, upon closer examination of their comments, it appears that two of these students

held a more nuanced view. One student clarified that the translator's role is unlikely to disappear entirely but will evolve to meet market needs and technological advances, resulting in more post-editing assignments rather than traditional translation work. Similarly, another student noted that post-editing is particularly appealing to clients seeking to reduce costs, suggesting that while the profession may shift toward more post-editing, it is unlikely to vanish altogether. This indicates that both students foresee an adaptation of the profession rather than its complete replacement by MT.

6 Concluding remarks

In this article, we have presented the results of a practical teaching unit developed in collaboration with DGT, aimed at engaging 32 first-year master's students in the human evaluation of MT quality at text level through error annotation. Our study shows that students often encounter challenges in accurately detecting and categorizing errors within machine-translated texts. This observation provides empirical support for the widely held belief that students must be continuously trained not only in post-editing but also in MT error detection in order to be fully prepared for careers in the language services industry. The complexity of these tasks highlights the need for ongoing, hands-on practice to develop the necessary skills for effective MT evaluation and post-editing work. However, it is important to stress that if we had conducted this study with second-year master's students, different trends might have emerged, potentially reflecting their more advanced skills and deeper familiarity with MT error detection and post-editing.

For future research, it would be valuable to examine the quality of students' post-edited texts to ensure alignment with the MT errors they identified. This could help pinpoint students' weaknesses in post-editing, enabling the development of targeted exercises or dedicated sessions within post-editing courses to address these challenges. Such practical initiatives would make meaningful contributions to translator education by refining students' skills in post-editing. Additionally, tracking students' progress over the course of their master's program would provide valuable insights into their evolving skills, allowing educators to adapt training methods to better support their development in MT error detection and post-editing. Finally, it would be important to replicate the experiment with the updated version of the eTranslation General Text engine so as to engage students with MT outputs of LLM-based technologies.

References

Banerjee, Satanjeev, and Alon Lavie. 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, MI, pages 65-72.

Bodart, Romane, Justine Piette, and Marie-Aude Lefer. 2024. The Machine Translation Post-Editing Annotation System (MTPEAS), A standardized and user-friendly taxonomy for student post-editing quality assessment. *Translation Spaces*, Online First. https://doi.org/10.1075/ts.24002.bod

Bojar, Ondrej. 2011. Analyzing error types in English-Czech machine translation. *The Prague Bulletin of Mathematical Linguistic*, (95): 63–76.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone, and Maria Gialama. 2017. A comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI: Research Track*, Nagoya Japan, pages 116–131.

Comelles, Elisabet, Jordi Atserias, Victoria Arranz, and Irene Castellón. 2012. VERTa: Linguistic features in MT evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association (ELRA), pages 3944–3950.

DGT (Directorate B – Translation, ABCD Quality Managers). 2024. Study to evaluate the quality of eTranslation output from multilingual platforms. Report summary.

DGT. 31 March 2020. DGT Guidelines for Evaluation of Outsourced Translations (TRAD19). Key aspects of evaluation under TRAD19 – Quick Reference.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co- occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California, USA, pages 138-145.

Doherty, Stephen, and Dorothy Kenny. 2014. The design and evaluation of a Statistical Machine Translation syllabus for translation students. *The Interpreter and Translator Trainer*, vol. 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. Cham, Springer, pages 95–106.

EMT Expert Group. 2022. Competence Framework. https://commission.europa.eu/system/files/2022-11/emt_competence_fwk_2022_en.pdf [last accessed October 6, 2024].

Farrús, Mireia, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France, European Association for Machine Translation, pages 167–173.

Federico, Marcello, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effect Models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics, pages 1643–1653.

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. London/New York, Routledge, pages 226–246.

Guerberof Arenas, Ana, and Joss Moorkens. 2019. Machine translation and post-editing training as part of a master's programme. *The Journal of specialized Translation*, vol. 31: 217-238.

Kirchhoff, Katrin, Daniel Capurro, and Anne Turner. 2012. Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, Trento, Italy, European Association for Machine Translation, pages 119–126.

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, pages 2–15.

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.

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. London: Bloomsbury, pages 237–266.

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

Moorkens, Joss. 2018. What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4): 375-387.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *6th EAMT Workshop Teaching Machine Translation*, pages 99-106.

Office des publications de l'Union européenne, 2022. *Code de rédaction interinstitutionnel*. Office des publications de l'Union européenne : https://data.europa.eu/doi/10.2830/445722

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pages 311–318.

Pavlovic, Natasa, and Goranka Antunovic. 2021. Towards acquiring post-editing abilities through research-informed practical tasks. *Strani jzici: Strani jezici: časopis za primijenjenu lingvistiku*, vol. 50(2): 185-205.

Popovic, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal. Association for Computational Linguistics, pages 392–395.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pages 2685–2702.

Snover Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA, Association for Machine Translation in the Americas, pages 223–231.