11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.19997

Bridging Theory and Practice: The Role of Language Quality Assessment in Translation Technology Training

Katrin Herget
Department of Languages and Cultures /CLLC, University of Aveiro, Portugal.

How to cite: Herget, K. (2025). Bridging Theory and Practice: The Role of Language Quality Assessment in Translation Technology Training. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.19997

Abstract

In today's globalized economy, delivering high-quality translations is essential for effective cross-cultural communication and professional localization. Machine Translation Evaluation (MTE), supported by Language Quality Assessment (LQA) frameworks within Translation Management Systems, provides systematic methods for evaluating the accuracy, fluency, and contextual relevance of machine-generated translations. This study explores the integration of LQA workflow into Translation Technology education at the University of Aveiro, where 14 Master's students in Specialized Translation were tasked with evaluating machine-translated texts. Through hands-on activities, students identified, categorized, and assessed translation errors, gaining critical competencies in quality assessment and error analysis. The findings highlight the importance of integrating LQA training into academic curricula to prepare future translators for the demands of a rapidly evolving industry.

Keywords: *Translation Technology; Language Quality Assessment; Machine Translation Evaluation; error analysis.*

1. Introduction

Translation quality management has long been a cornerstone of both professional practice and academic research. Yet, as Vela-Valido (2021, p. 95) notes, the field is characterized by terminological inconsistency, with overlapping and sometimes ambiguous terms such as "translation quality evaluation," "translation quality control," "translation quality assurance," and "translation quality assessment." These variations reflect differences in approach, context, and purpose—whether academic or professional, product-focused or process-focused—often resulting in conceptual discrepancies.

This study bypasses the ongoing terminological debate to focus on a practical classroom project conducted with Master's students specializing in Translation Technology. At the University of Aveiro, 14 Master's students participated in a project involving Linguistic Quality Assessment

(LQA) using the Phrase TMS tool. The project aimed to evaluate machine translation (MT) output by categorizing predefined errors, providing students with hands-on experience in linguistic review and quality evaluation.

Training students in LQA is essential for preparing them to meet the modern translation industry's demands. Through error annotation, students develop critical analytical skills by systematically identifying, categorizing, and addressing translation issues such as mistranslations, terminological inconsistencies, omissions, and contextual mismatches, while ensuring domain-specific relevance.

Integrated into computer-assisted translation (CAT) tools like Phrase, MemoQ, and Smartling, LQA provides a structured workflow for assessing machine-translated content. This not only helps students evaluate translations systematically but also equips them with critical thinking and industry-relevant skills in quality assessment and error analysis.

This study explores how integrating LQA into translation technology education enhances learning. By incorporating LQA workflows into the curriculum, students gain valuable hands-on experience with industry-standard tools, preparing them to meet the evolving demands of the profession.

2. Machine Translation Evaluation

The advent of machine translation (MT) systems has transformed the translation industry while introducing the critical need for robust quality evaluation. The primary goal of Machine Translation Evaluation (MTE) is to ensure that the MT output meets the required standards of accuracy, fluency, and contextual appropriateness for professional use.

MTE methods can be divided into manual and automatic evaluation. According to Moorkens et al. (2025, p. 84), "Manual evaluation can provide a detailed view of MT quality, depending on the skill of the evaluators, but is likely to be slow and expensive." Its subjective nature also makes consistency a challenge. To address this, Inter-Annotator Agreement (IAA) measures the reliability of error annotation across evaluators, ensuring a standardized approach (Artstein, 2017). However, as Lommel (2018, p. 120) observes, "When evaluating a translation, it is typically not enough to know how many errors are present. Evaluators also need to know (a) how severe they are and (b) how important the error type is for the task at hand. Severity and importance are distinct concepts in MQM".

For automatic evaluation, MT output is typically compared to a human-generated reference translation. Metrics such as BLEU, METEOR, and COMET are widely used for this purpose. Among these, COMET (Rei et al., 2020) stands out for leveraging pre-trained neural models to evaluate semantic similarity and fluency, aligning more closely with human judgments.

This study focuses on manual annotation of MT output, within the context of a classroom project. Master's students participated in hands-on quality evaluation, annotating errors in MT output using the structured framework provided by Phrase's LQA tools. This process included identifying, categorizing, and assessing errors, offering students practical exposure to the complexities of translation evaluation.

3. The Role of Quality Assessment in Translation Technology

LQA offers a structured workflow for evaluating both machine-translated and human-translated texts. In this way, it plays a crucial role in ensuring translation quality, regardless of whether the output is produced by machines or humans. Translation Management Systems like Phrase, Memoq, and Smartling incorporate LQA features to streamline error detection and quality control. As described by Phrase (s.d.), "LQA provides visibility on translation quality based on pre-configured criteria. LQA can be added as a workflow step to help linguists review translations (i.e., human translations, machine translations, or machine translations with edits) according to predefined error categories applied in the project." The LQA process typically involves three stages: a) Defining and customizing error categories; b) conducting translation reviews to identify and evaluate errors; and c) automatically calculating LQA scores to measure overall translation quality (Phrase, s.d.). This systematic approach allows for a more objective, data-driven assessment of translation quality, enabling both translators and reviewers to identify areas for improvement efficiently.

By incorporating LQA into translation workflows, professionals can significantly enhance the reliability and accuracy of machine-generated translations. In educational settings, LQA plays a critical role in preparing future translators for industry demands.

4. Study Design and Methodology

The classroom project began with an introductory session on LQA, emphasizing its importance in translation technology and workflows. Students were already familiar with the translation processes in Phrase and were subsequently introduced to its built-in LQA features.

Central to the project was the adoption of the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), which was designed to "be applicable to any sort of translated text (human or machine translated) and to any type of text" (Lommel et al., 2014, p. 456), not aiming, however, to be a "one-size fits-all model for evaluating translation quality". The MQM framework is versatile, offering predefined categories and subcategories that can be customized based on project-specific needs.

Each student was assigned a machine-translated text for evaluation. The selected text was a news article reporting on severe snowfall in South Korea's capital, which led to significant

disruptions in transportation. Originally published by *The Guardian* on November 28, 2024, the article was machine-translated and analyzed using Phrase, where a LQA was conducted based on the MQM framework. A news text was chosen for its use of general language, making it accessible to students while also offering a range of linguistic features that are valuable for translation quality assessment. The students' task involved systematically identifying and categorizing errors within the translation according to the MQM framework in Phrase. While the MQM framework includes predefined categories, its flexibility allowed students to adapt and refine these categories and subcategories to address the specific challenges posed by their assigned text (Figure 1).

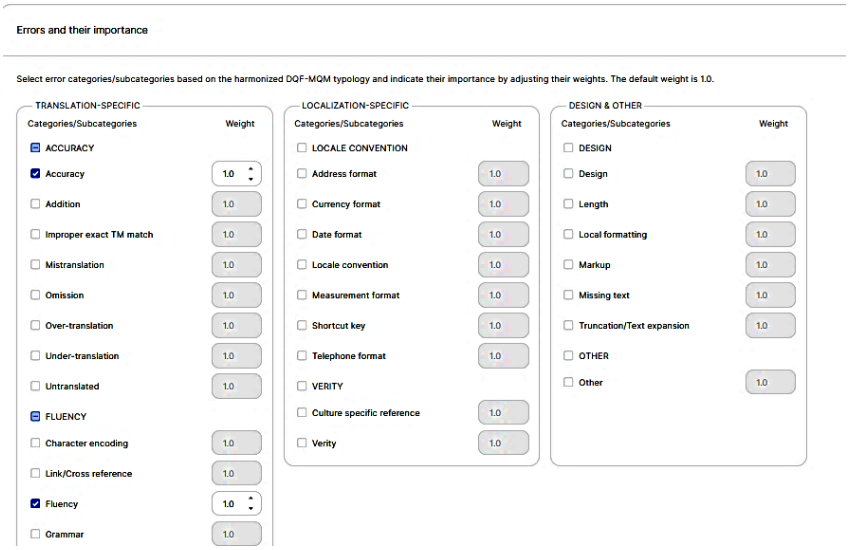

Figure 1: Predefined MQM Categories in LQA

This hands-on approach exposed students to error types such as accuracy, fluency, terminology, and style, while encouraging critical thinking about the most appropriate classifications for various translation issues. During the evaluation process, students were instructed to critically assess the quality of machine-generated translations. They classified errors, assigned severity ratings, and proposed corrections, thereby applying theoretical knowledge in a practical setting.

The study design prioritized project-based learning, aiming to bridge the gap between classroom instruction and real-world industry practices.

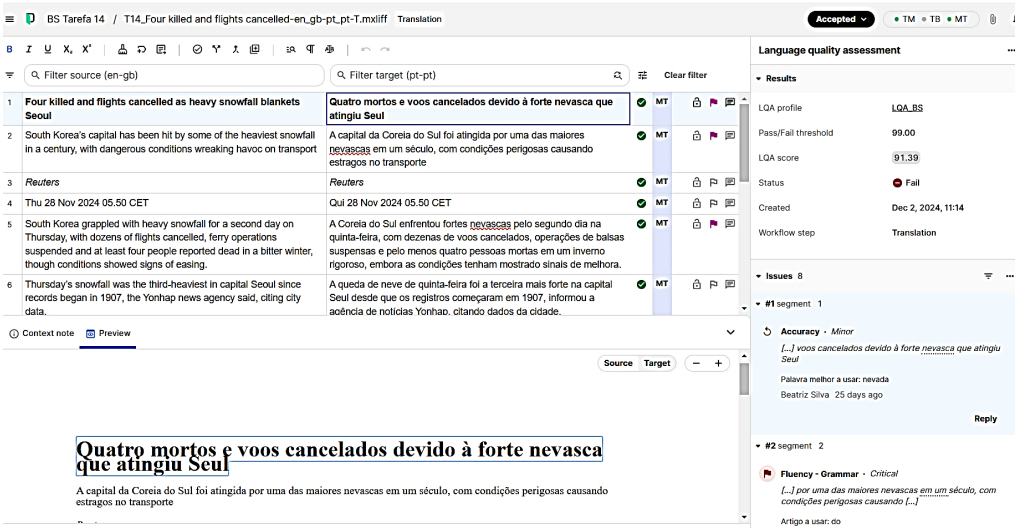

Figure 2. Example of LQA annotation by an MA student using the Phrase editor.

Following the completion of the LQA task, the annotations were reviewed for consistency in error categorization and severity assignment. Discrepancies were addressed through a structured feedback session where students explained their reasoning. These annotations were then analyzed for patterns in categorization and severity assessment, with particular attention given to variations across students.

5. Results and Discussion

The analysis of student performance in LQA revealed both strengths and challenges faced by the participants. One of the most notable discoveries was the variability in how students selected and applied subcategories within the MQM framework. While the MQM framework provides predefined categories for error classification, its flexibility allows for adaptation to the specific characteristics of the text being evaluated. This adaptability, while beneficial for customizing the framework to different contexts, also resulted in inconsistencies in how students interpreted and applied category definitions.

In general, students demonstrated competence in identifying key translation errors. However, there was significant divergence in how they rated the severity of these errors. For instance, an error classified as "Minor" by one student was sometimes rated as "Major" by another, suggesting that the severity of errors was often a subjective judgment.

Furthermore, the choice of error categories was not always consistent. Some students focused primarily on linguistic and grammatical issues, while others gave more weight to stylistic or cultural adaptation concerns. This divergence emphasizes the need for aligning category

selection with the intended purpose and context of the translation. Table 1 provides a detailed summary of the final LQA scores and Pass/Fail statuses for each evaluated text, highlighting these inconsistencies.

The Pass/Fail results in this study were determined using a scoring model based on the MQM framework, with a pre-established threshold of 99.0%. According to this model, a text must achieve a final score of at least 99% to pass. The scoring formula incorporates:

$$\text{Score} = 1 - \frac{\text{Penalty Total}}{\text{Word Count}} \quad (1)$$

Here, the Penalty Total is calculated based on the number of errors, their severity, and the associated weights assigned to error categories. For example, critical errors may carry higher penalties compared to minor errors. These penalties are summed to determine the total impact on the score.

- Pass: Achieved when the final score is equal to or greater than the threshold (≥ 99.0%).
- Fail: Occurs when the score falls below the threshold (< 99.0%).

The only text in the dataset that achieved a "Pass" status was Tarefa_14_DC. This result reflects its minimal error count and relatively high accuracy, resulting in a score that met the 99% target. All other texts failed due to either a high number of errors or a greater penalty-to-word ratio, which caused their scores to drop below the required threshold.

Table 1: Summary of LQA Scores and Statuses

Report #	Project Name	Final Score	Status	Total Errors	Accuracy Errors	Fluency Errors	Terminology Errors	Style Errors
1	Rvo_Tarefa14	96.63	FAIL	9	3	0	0	2
2	tarefa 14	92.13	FAIL	21	0	0	15	0
3	T14 - Four killed and flights cancelled	97.38	FAIL	7	1	1	0	2
4	Tarefa 14 - GL - Korean Snow News	71.91	FAIL	75	1	1	35	0
5	BS Tarefa 14	91.39	FAIL	23	5	2	0	1
6	tarefa 14 liane	95.88	FAIL	11	10	0	1	0
7	Tarefa 14 Sara	66.67	FAIL	89	5	0	78	1
8	Tarefa 14 AMRP	90.64	FAIL	25	15	3	0	0
9	Tarefa 14 MN	98.13	FAIL	5	5	0	0	0
10	Tarefa_14_DC	99.63	PASS	1	0	1	0	0
11	T14_MG	98.88	FAIL	3	2	1	0	0
12	Tarefa 14 LX	94.78	FAIL	18	8	2	6	2
13	Tarefa 14 J	94.12	FAIL	12	6	1	3	2
14	Tar 14 QT	97.76	FAIL	8	3	0	0	1

While there was general agreement on the types of errors present in the machine-translated texts, the application of severity ratings and subcategory assignments varied considerably. These findings emphasize the need for more structured guidance and practical exercises to help students apply error categorization frameworks and improve overall consistency.

6. Conclusions

This study highlights the critical role of LQA in translation technology education, underscoring its importance in preparing students to meet the expectations of the professional translation industry. Incorporating LQA training into academic curricula not only provides students with the skills to evaluate translation quality but also familiarizes them with industry-standard tools and methods for assessing both machine-generated and human translations. The analysis of students' LQA performance indicates a general trend in the types of errors identified, with most students consistently recognizing issues such as accuracy, fluency, and style. However, there was noticeable variation in the total number of errors and the assignment of error categories, particularly regarding the severity levels. This variability suggests that while students were aligned on some key aspects of translation quality, further training and calibration are needed to ensure a more consistent application of error categorization across the group. To address these challenges, the development of standardized reference materials, such as annotated examples or calibration exercises, could be instrumental. These resources would provide students with a clear benchmark for applying error categories, helping them align their evaluations with established best practices.

In professional translation workflows, LQA is becoming increasingly indispensable, ensuring that all outputs meet client expectations and uphold rigorous quality standards. Future research could explore and compare the effectiveness of LQA workflows provided by different Translation Management Systems, focusing on their impact on the accuracy, consistency, and efficiency of translation quality assessments.

References

Artstein, R. (2017). Inter-annotator Agreement. In: Ide, N., Pustejovsky, J. (eds) *Handbook of Linguistic Annotation.* Springer, Dordrecht. https://doi.org/10.1007/978-94-024-0881-2_11

MemoQ (s.d.). https://docs.memoq.com/current/en/Concepts/concepts-linguistic-quality-assurance.html Accessed: 13.01.2025.

Lommel, A. (2018) 'Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies', in J. Moorkens et al. (eds) *Translation Quality Assessment.* Cham: Springer International Publishing (Machine Translation: Technologies and Applications), pp. 109–127. Available at: https://doi.org/10.1007/978-3-319-91241-7_6

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): a Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, Núm. 12, 455-463.

Moorkens, J., Way, A., & Lankford, S. (2025). *Automating Translation.* Routledge.

Phrase (s.d.). https://support.phrase.com/hc/en-us/articles/5709599557020-Language-Quality-Assessment-TMS. Accessed: 13.01.2025.

Rei, R., Stewart, C., Farinha, A.C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics.

Vela-Valido (2021). Translation Quality Management in the AI Age: New Technologies to Perform Translation Quality Management Operations. *Revista Tradumàtica. Tecnologies de la Traducció*, 19, 93-111