Translation Quality

Joss Moorkens and Sheila Castilho, School of Applied Language and Intercultural Studies/ADAPT Centre, Dublin City University, Dublin, Ireland

Key Points

- Many definitions of translation quality have been proposed by academics and industry, but the variability of translation domains and processes means that none have been universally accepted.
- The scope of academic definitions of quality has broadened over time, tending towards heuristics rather than measurable quantitative metrics.
- For human evaluation of human, machine, and hybrid translation, the modular error annotation metric MQM appears to be most widely used.
- Translation quality evaluation in the translation industry can focus on the translation process and/or the product, tending towards a variable fit-for-purpose quality requirement.
- Automatic evaluation of machine translation can be divided into untrained and trained metrics, the latter being the state of the art, and furthermore into reference-based and reference-free metrics.
- AI-driven translation systems challenge traditional evaluation methodologies, necessitating new context-aware approaches that account for ethical considerations, bias, and real-world applicability.

Abstract

Definitions of translation quality in translation studies have expanded in scope over time, while the translation industry has tended towards quantitative measures, created to formally assess many dimensions of quality for a given purpose. With the advent of highly technologized industry workflows aided by machine translation and large language models, the need to establish systematic and effective translation quality assessment has become vital. This chapter introduces translation quality and assessment methods used in human, automated, and hybrid translation workflows. It investigates the theoretical frameworks, industry practices, and technological advancements in translation quality assessment while highlighting critical issues and potential solutions in the field.

Introduction

What makes a good translation? This seemingly simple question is difficult to answer. Translation is a complex cognitive, linguistic, social, cultural, and technological process. Translation processes, and thus translation quality assessment (TQA), reflect this complexity. Definitions of translation quality attempt to capture these dimensions and their interactions to devise a means of formally assessing translation quality for a given purpose. With the advent of highly technologized industry workflows aided by machine translation (MT) and large language models (LLMs), the need to establish systematic and effective TQA methodologies has become vital. This is because automated systems can produce high volumes of translations with variable and sometimes unpredictable quality, ranging from overtly poor translations to fluent but inaccurate ones. TQA is key to identifying inconsistencies and ensuring reliability across outputs. Quality needs may also vary according to norms and reader/user expectation, fitness for purpose, and levels or types of risk of mistranslation, from the risk of miscommunication to the risk of injury or death (see Canfora & Ottmann, 2020). This chapter introduces translation quality and TQA methods used in human, automated, and hybrid translation workflows. It investigates the theoretical frameworks, industry practices, and technological advancements in TQA while highlighting critical issues and potential solutions in the field.

Defining Translation Quality

Translation quality has long been a topic of debate—as Drugan (2013) writes, "theorists and professionals overwhelmingly agree there is no single objective way to measure quality"—and definitions often vary significantly between academic research and industry practices. Scholars who evaluate student work may emphasize theoretical and pedagogical dimensions, whereas industry professionals are more likely to focus on practical and standardized error typologies to ensure client satisfaction, and there is some evidence of the use of these in translation classrooms, particularly as part of experiential or project-based learning. The fundamental dichotomy between accuracy (source-oriented) and fluency (target-oriented) continues to influence TQA frameworks. The increasing integration of automation using MT and LLMs has further diversified these methodologies, necessitating a re-evaluation of existing practices. In the following section, we look at how the scope of academic definitions of translation quality have expanded over time, before we return to industry practices.

Academic Perspectives on Translation Quality

From the linguistic approaches of Nida (1964) and Holmes (1972) through functionalist models like Skopos Theory (Reiß & Vermeer, 1984), to the polysystems of Toury (2012), the purview of translation quality in translation studies publications has tended to expand over time to take into account the cultural and social environment and related translation norms. Equivalence-based models initially focused on achieving "optimal equivalence" (Lauscher, 2000, p. 151), formal correspondence between source and target texts. Nida (1964) introduced the variability of formal and dynamic equivalence, highlighting the importance of maintaining the source text's content and style while ensuring comprehensibility for the target audience. However, critics argue that strict adherence to equivalence may overlook cultural nuances and pragmatic considerations. In particular, Gentzler (2001) is distrustful of the acceptability of making changes to a text in order to achieve dynamic equivalence (i.e. to provoke an equivalent reader response).

According to Munday et al. (2022, p. 23), over time there has been a move away from prescriptive definitions of translation quality in translation studies towards descriptive evaluation methods that vary depending on "the goals of the research and the researchers". Functionalist approaches shift the focus from linguistic accuracy to communicative effectiveness. According to Reiß and Vermeer (1984), the translation process should align with the text's intended purpose (skopos) and the target audience's needs. This perspective encourages flexibility and creativity, particularly in domains like marketing and advertising, where cultural adaptation is crucial. However, the subjectivity inherent in functionalist models presents challenges for standardized TQA practices. House (2015, p. 11) writes that it is not clear "how one can determine whether a given translation fulfils its skopos".

This is true of many academic proposals for TQA, which might be best considered as heuristics rather than replicable instruments for measurement, such as Abdallah's (2014) proposal for three dimensions of quality, looking at the product, process, and social quality. The notions of product and process quality are widely used in industry (as we shall see in the following section), using an appropriate measure for the product, and assuming that, by following certain prescribed process steps, product quality is likely to be higher. Abdallah's proposal is that social quality—who does what and under what sort of working conditions—is indivisible from the other two dimensions, affecting the final translation quality.

The proposal by Moorkens et al. (2024) built on this idea of social quality, proposing a 'triple bottom line' for automatic translation quality whereby the effect on people and the planet should be just as important as performance (i.e. the quality of the translation). 'People' include all stakeholders in the production of the MT or LLM system, and 'planet' includes environmental impact such as ICT production and disposal, and CO_2 emissions and water use in datacenters.

Industry Practices in TQA

Industry practices tend to prioritize efficiency, consistency, and cost-effectiveness. As we wrote in Moorkens et al. (2018, p. 3), the ever-increasing amount of text to translate across many language pairs and directions has "led to a new level of pragmatism in large translation service providers, whereby a sharpened focus on a targeted end-user […] has added a new meticulously calibrated variability to translation quality requirements". That is, quality expectations are now often articulated through broad or loosely defined categories depending on the client's needs.

Fields et al. (2014) describe five approaches commonly used to define translation quality. The first, often used in academia, is the transcendent approach, which involves deciding that a "product or service possesses excellence based on its subjective relationship to some standard" (Fields et al., 2014, p. 406). A product-based approach involves measurement based on certain attributes or fitness for a purpose; a user-based approach prioritizes satisfying the needs of the user; a production-based approach (as we have seen) focuses on fulfilling certain production criteria; and a value-based approach involves a balance of costs and benefits in calibrating translation quality.

The translation and localization industry commonly employs product-based approaches, involving practical TQA typologies such as multidimensional quality metrics (MQM) (Lommel, 2018), which is also widely used in translation research. These typologies typically categorize errors based on severity and assign penalty points to quantify quality. Earlier typologies employed a one-size-fits-all approach that often failed to accommodate the diverse needs of different projects, but MQM represents a shift towards customizable and modular TQA methods that can integrate human and automated evaluation, with a choice of error categories and severity levels to choose from. MQM evaluation can be quite subjective, but clear instructions and examples to differentiate error types can help to maximize inter-annotator agreement. MQM has also been incorporated into production-based approaches that assume that quality will be improved by following defined steps, such as ISO 5060 and ASTM WK54884 standards.[1]

These error typologies are used retrospectively once the translation has been produced, but language service providers may also use translation management system-integrated or standalone tools like ApSIC Xbench or QA Distiller to perform real-time quality checks or to semi-manually annotate errors. Many of these tools identify issues related to terminology consistency, formatting, punctuation, and numerical accuracy. However, automated tools often produce 'false positives' (flagging errors that turn out not to be errors) and there are concerns regarding their ability to detect context-dependent errors, such as cultural mismatches or pragmatic inconsistencies.

In the broader context of process quality, ISO11669 is a foundational industry standard for translation quality. It categorizes translation specifications into five areas: source content, target requirements, production tasks, environmental factors, and relational dynamics. The standard promotes collaboration between clients and providers, ensuring clarity regarding quality expectations. However, critics argue that its generic nature may not address the nuanced requirements of specialized projects and that the standard is no longer fit for purpose in the era of LLMs.

Human TQA Methods

Human evaluators assess translation quality using measures like adequacy, fluency, readability, comprehensibility, and acceptability, or by annotation using MQM or other error typologies. Adequacy and fluency are the most commonly employed criteria in both research and industry contexts. Adequacy (sometimes known as accuracy) evaluates semantic fidelity to the source text, while fluency focuses on linguistic correctness in the target language. Adequacy assessment requires bilingual proficiency to compare source and target texts, while fluency evaluation demands target-language expertise to identify grammatical, lexical, and stylistic deviations. In practice, these criteria are often used together, as it is difficult to assess adequacy and fluency in isolation—fluency issues can obscure adequacy problems, while adequacy errors can affect overall fluency.

Human evaluation may be augmented by manual or automated measures of acceptability. Readability tests, such as the Flesch-Kincaid readability scale, quantify text complexity based on sentence length, word frequency or type-token ratio. While readability and comprehensibility concern the ease of text processing by users, acceptability measures the alignment with user expectations. Acceptability assessments, in the broader context of process quality, are often used in usability studies and gauge user satisfaction with translation outputs. These assessments often utilize Likert scales, cloze tests, and usability testing to gather both subjective and objective data. Similarly, comprehensibility evaluations measure a text's ease of understanding from the reader's perspective. Techniques like cloze tests and recall tasks provide insights into cognitive processing and information retention. These methods collectively contribute to a holistic understanding of translation quality.

Automated TQA Methods

Automated metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), TER (Snover et al., 2006), chrF (Popović, 2015), BERT (Devlin et al., 2019) and COMET (Rei et al., 2020), offer quick, cost-effective and consistent means of evaluating MT (and LLM translation) output to a limited extent. The original intention of many of these metrics was to give a quick indication of quality and to signal whether the system output had changed after retraining. Automated metrics can be divided into untrained and trained metrics, and furthermore, into reference-based and reference-free metrics.

Reference-based metrics compare an MT segment (usually a sentence) with a human reference translation, which is presumed to be of high quality. The comparison primarily relies on n-gram[2] matching between the MT output and reference translation, with recent models incorporating linguistic features to improve accuracy. These are also known as "lexical similarity-based metrics", as they focus on exact surface-level word matches. Thus, these metrics are unable to capture semantic and pragmatic nuances.

For a long time, BLEU (Bilingual Evaluation Understudy) was the most widely used lexical similarity-based metric in MT evaluation. However, BLEU's inability to account for synonymy and sentence-level coherence prompted the development of more sophisticated alternatives, along with criticism of overuse (e.g. Kocmi et al., 2021). TER (Translation Edit Rate) measures the minimum number of edits required to transform MT output into an acceptable target text and is thus sometimes used to approximate post-editing effort.

Trained metrics such as COMET seem to correlate better with human judgment than lexical similarity-based metrics, as long as sufficient previous judgments are available for training. These metrics compute source and target word embeddings to propose

[1] See Strandvik (2025) on the development of translation standards and specifications.

[2] An n-gram is a string of text made up of a number of words. "The 'n' represents the number of words, so that a 2- gram ('bigram') is a 2- word string of text, a 3- gram ('trigram') a 3- word string of text and so on" (Moorkens et al., 2025, p. xix).

scores, but still may have trouble with negation, wrong proper nouns, and prioritization of critical errors (Rei et al., 2023). Despite their practical applications, automated metrics tend to struggle with discourse-level phenomena like anaphora resolution and logical consistency, although LLM-based evaluation seems to improve this (Kocmi & Federmann, 2023).

Beyond reference-based approaches, reference-free metrics, such as COMETKiwi (Rei et al., 2022), assess MT output directly against the source text without requiring a reference translation. These metrics typically leverage multilingual language models and contextual embeddings to determine both the lexical similarity of the candidate translation to the source and its natural language fluency. By bypassing the need for a reference, they aim to provide a more flexible and scalable evaluation of translation quality.

Hybrid Approaches in HT-MT Workflows

The integration of human translation (HT) and MT in modern workflows has blurred the lines between these traditionally distinct processes. Post-editing of MT output exemplifies this convergence, requiring evaluators to address linguistic errors while maintaining the communicative intent of the text. Research into post-editing effort, including cognitive, temporal, and technical dimensions, has provided insights into optimizing TQA in these hybrid workflows (e.g. Nunes Vieira, 2020; Rico, 2022).

Post-editing involves correcting machine-generated text to meet predefined quality standards. Light post-editing focuses on basic error correction, a process required less often as MT quality improves, while full post-editing aims for near-human translation quality. Studies such as Krings (2001) and the meta-analysis by Koponen (2016) highlight the cognitive demands of post-editing, emphasizing the role of domain knowledge, linguistic competence, and tool proficiency.

Challenges and Future Directions

Although both automatic and human evaluation metrics represent the state of the art in TQA, the increasing complexity and improved output of NMT and LLM systems have highlighted the need for more comprehensive and rigorous evaluation approaches. As efforts to integrate context into these systems have grown, there is a pressing need to develop context-aware evaluation methodologies that assess translation quality in a broader, more meaningful way (Castilho & Caseli, 2023). This is because it has become evident that context-aware evaluation traditional sentence-level evaluation is insufficient, as it fails to capture the full extent of these systems' advancements. However, research in this area remains in its early stages, and few studies have systematically addressed how to implement such evaluations effectively (Castilho & Knowles, 2024). Furthermore, LLM-driven translation introduces additional challenges, including ethical considerations, bias detection, and the alignment of evaluation methods with real-world applications, making the development of robust context-aware assessment frameworks even more critical.

Therefore, several persistent challenges in TQA can be listed:

- **Lack of Standardization:** The absence of universally accepted quality metrics hinders effective communication between academic and industry stakeholders, although MQM is now widely accepted as a human evaluation standard.
- **Inconsistency in Human Evaluations:** Subjective biases and varying expertise levels contribute to inconsistent assessments.
- **Contextual Evaluation Challenges**: Incorporating context into evaluation—whether human or automatic—remains difficult. As context can be multi-layered, including co-text (surrounding text) and non-text (real-world knowledge), human evaluators may struggle with consistency in applying context, while automated metrics often fail to capture its nuances, leading to misaligned assessments.
- **Human-Machine Discrepancies:** Disparities between human judgments and automated scores complicate the evaluation process.
- **Social and Ethical Considerations:** The growing reliance on contingent labor and the ethical implications of translation practices necessitate further investigation.

Addressing these challenges requires collaborative efforts to establish common standards, develop advanced evaluation tools, and promote educational initiatives in translation studies. Future research should focus on refining TQA methodologies to account for contextual variability, user-specific requirements, and technological advancements.

Conclusion

Translation quality has no universally agreed definition and TQA remains a dynamic and complex field that must continuously adapt to technological advancements and evolving industry needs. The convergence of human and machine translation, along with the development of more sophisticated evaluation metrics, offers avenues for future research. Establishing standardized, context-sensitive TQA frameworks will enhance the reliability and applicability of TQA across diverse settings.

However, as NMT and LLM translation systems become more sophisticated, the need for rigorous, context-aware evaluation becomes even more pressing. Traditional sentence-level assessments fail to capture real-world complexities, overlooking issues such as bias, ethical risks, and the impact of translation choices on people and society. Addressing these challenges requires scalable methodologies that integrate human expertise, contextual understanding, and automated evaluation in a meaningful way.

The interplay between theoretical insights and practical applications is essential for addressing TQA challenges. By fostering collaboration between researchers, educators, industry professionals, and policymakers, the field can move toward a more cohesive and effective understanding of translation quality. Ultimately, the goal is to support high-quality translations that facilitate cross-cultural communication, knowledge exchange, and global engagement, while upholding ethical standards and preserving the human touch in translation.

References

Abdallah, K. (2014). Social quality. Key to collective problem solving in translation production networks. In *Proceedings of the 6th Riga symposium on pragmatic aspects of translation* (pp. 5—18).

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65—72). Association for Computational Linguistics. https://aclanthology.org/W05-0909/.

Canfora, C., & Ottmann, A. (2020). Risks in neural machine translation. *Translation Spaces, 9*(1), 58—77. https://doi.org/10.1075/ts.00021.can

Castilho, S., & Caseli, H. de M. (2023). Tradução Automática: Abordagens e Avaliação. In H. M. Caseli, & M. G. V. Nunes (Eds.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* (1st ed.). BPLN (chapter 18) https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-mt/cap-mt.html.

Castilho, S., & Knowles, R. (2024). A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 1—31. https://doi.org/10.1017/nlp.2024.7. Online first.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171—4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Drugan, J. (2013). *Quality in professional translation: Assessment and improvement*. Bloomsbury. https://www.bloomsbury.com/uk/quality-in-professional-translation-9781441149541/.

Fields, P., Hague, D. R., Koby, G. S., Lommel, A., & Melby, A. (2014). What is quality? A management discipline and the translation industry get acquainted. *Tradumàtica Tecnologies de La Traducció, 12*, 404—412. https://doi.org/10.5565/rev/tradumatica.75

Gentzler, D. (2001). *Contemporary translation theories*. Clevedon: Multilingual Matters.

Holmes, J. S. (1972). The name and nature of translation studies. *Translation Section of the Third International Congress of Applied Linguistics*, 66—79. https://books.google.is/books?id=ad8pHQAACAAJ.

House, J. (2015). *Translation quality assessment: Past and present*. Routledge. https://doi.org/10.4324/9781315752839

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *arXiv*. https://doi.org/10.48550/ARXIV.2302.14520

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the sixth conference on machine translation* (pp. 478—494). https://aclanthology.org/2021.wmt-1.57.

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *Journal of Specialised Translation, 25*, 131—148.

Krings, H. P., & Koby, G. S. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent: Kent State University Press.

Lauscher, S. (2000). Translation quality assessment: Where can theory and practice meet? *The Translator, 6*(2), 149—168. https://doi.org/10.1080/13556509.2000.10799063

Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In J. Moorkens, S. Castilho, S. Doherty, & F. Gaspari (Eds.), *Translation quality assessment. From principles to practice* (pp. 109—127). Springer. https://doi.org/10.1007/978-3-319-91241-7_6

Moorkens, J., Castilho, S., Gaspari, F., & Doherty, S. (2018). *Translation quality assessment from principles to practice*. Springer. https://doi.org/10.1007/978-3-319-91241-7

Moorkens, J., Castilho, S., Gaspari, F., Toral, A., & Popović, M. (2024). Proposal for a triple bottom line for translation automation and sustainability: An editorial position paper. *Journal of Specialised Translation, 41*, 2—25. https://doi.org/10.26034/cm.jostrans.2024.4706

Moorkens, J., Way, A., & Lankford, S. (2025). *Automating translation*. London: Routldege.

Munday, J., Pinto, S. R., & Blakesley, J. (2022). *Introducing translation studies: Theories and applications* (5th ed.). London: Routledge.

Nida, E. (1964). *Toward a science of translating: With special reference to principles and procedures involved in bible translating* (2nd ed.). BRILL. https://doi.org/10.1163/9789004495746

Nunes Vieira, L. (2020). Post editing of machine translation. In M. O'Hagan (Ed.), *The routledge handbook of translation and technology*. London: Routledge.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311—318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, et al. (Eds.), *Proceedings of the tenth workshop on statistical machine translation* (pp. 392—395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Rei, R., Guerreiro, N. M., Treviso, M., Coheur, L., Lavie, A., & Martins, A. F. T. (2023). The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.48550/ARXIV.2305.11806

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 2685—2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., & Martins, A. F. T. (2022). CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, et al. (Eds.), *Proceedings of the seventh conference on machine translation (WMT)* (pp. 634—645). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.60/.

Reiß, K., Vermeer, H. J., with Dudenhofer, M., & Nord, C. (1984). *Towards a general theory of translational action: Skopos theory explained*. Routledge. https://doi.org/10.4324/9781315759715

Rico, C. (2022). Mind the gap: The nature of machine translation post-editing. *Babel, 68*(5), 697—722. https://doi.org/10.1075/babel.00288.ric

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (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* (pp. 223—231). https://aclanthology.org/2006.amta-papers.25.

Strandvik, I. (2025). Translation quality and the role of specifications — How standards can help the translation sector today. Across Languages and Cultures, 26(S), 5—24. https://doi.org/10.1556/084.2025.01057

Toury, G. (2012). *Descriptive translation studies—And beyond*. Amsterdam: John Benjamins.