ISSN: 2278-4012, Volume:13, Issue:4, 31 December 2024

Machine Translation Quality Evaluation for Multilingual Educational Communication

Aiman Eid Al-Rawajfeh[1*]

[1*]Tafila Technical University, Tafila, Jordan. Orcid: https://orcid.org/0000-0002-0086-826X

Abstract: Educational facilities are turning to automatic translation for multilingual students; translation quality becomes crucial for the effectiveness of teaching processes. However, most common evaluation criteria tend to favor general fluency rather than domain relevance. In order to fill this gap, the paper offers an evaluation of machine translation (MT) quality in an educational context based on a multi-criteria, theoretically based approach that takes into account factors such as functional adequacy and reader comprehensibility. Three different systems (a general neural engine, System A; a general cloud service, System B; and an education-specific system, System C) have been tested on the basis of the 60-passage corpus (360 translations in total) covering both science and humanities. Human raters with specific training evaluated the output of translation according to the following criteria (rating from 1 to 5): adequacy, fluency, terminology accuracy, cultural relevance, and readability. However, System C performed better in terms of criteria essential for learning effectiveness, obtaining the best results in adequacy (4.4 compared to 4.1 for A and 3.8 for B), terminological accuracy (4.5 compared to 3.6 for A and 3.9 for B), cultural appropriateness (4.2 compared to 3.4 for A and 3.2 for B), and readability (4.3 compared to 4.0 for A and 3.7 for B). For all the systems, the criterion of cultural appropriateness had the lowest score. Thus, it can be concluded that fluency in itself cannot serve as a good measure for educational translation.

Key Words: machine translation, translation quality evaluation, multilingual education, educational communication, translation adequacy, language learning technology, cross-linguistic comprehension

(Received: 26 September 2024; Revised: 14 October 2024; Accepted: 29 November 2024; Published: 31 December 2024)

Introduction

Machine Translation in Educational Communication

The practice of machine translation has been steadily gaining ground in the field of educational communication as a result of the need for institutions, educators, and learners to communicate instructional material, tests, and administrative details in languages other than their own. The use of automatic translators to convey lessons, textbooks, and official communication in multilingual education environments has allowed for rapid and convenient communication that would not be possible through the use of only human translators. This is shown in studies involving the architecture of hybrid and pivot machine translation, able to communicate in multiple languages simultaneously (Nasution et al., 2017).

Nevertheless, this increased application is accompanied by the necessity of conducting a strict evaluation of the quality of translated materials because the degree of accuracy and clarity of educational materials that are translated impacts students' understanding of these materials. Evaluation of quality in this case is understood as the evaluation of translation with respect to its accuracy, fluency, and appropriateness for the targeted audience. This issue is rather important for experts in translation studies who try to develop approaches for the evaluation of machine and human translation results (Castilho et al., 2018). In the scope of language learning, there appeared an attempt to analyze the possibility of using

ISSN: 2278-4012, Volume:13, Issue:4, 31 December 2024

automatic evaluation metrics developed for general-purpose translations to evaluate the translations performed in the interests of learners (Han & Lu, 2023).

Importance of Quality Evaluation in Multilingual Contexts and Statement of the Research Problem

The significance of this type of evaluation is further strengthened by the fact that an increasing number of scientific works have appeared that describe the advantages, as well as the disadvantages, of machine translation in the context of foreign language education. The results of the systematic review and meta-analysis of machine translation effectiveness in foreign language education reveal predominantly positive learning results from the use of machine translation, but emphasize that these results depend greatly on the quality of the translations used (Lee, 2023). Another systematic review of machine-translation-assisted language learning highlights the fact that sustainable use of machine translation technology in an educational context depends on the constant concern with the output quality, rather than the mere availability of machine translation tools (Deng & Yu, 2022). These studies, therefore, provide the foundation for the research problem posed in this paper, which consists of the lack of evaluation of machine translations according to educational requirements.

Here are the paper's unique contributions:

- Attempts to fill one of the critical gaps in research by evaluating the quality of machine translation regarding the requirements of educational material rather than just any text.
- Demonstrates that even a good level of translation fluency cannot be considered a sufficient criterion in the educational context when terms' accuracy and readability matter more.
- Illustrates, using the case of multi-criteria human evaluation, that educational MT solutions equipped with domain glossaries outperform general NMT and cloud solutions in terms of accuracy, readability, and culture.
- Points out cultural adequacy as a systematic weakness of all tested MT solutions.
- Proposes a goal-oriented, multi-criteria evaluation and selection approach.

The paper is broken down into seven sections. Following this introduction, Section 2 considers prior research on machine translation quality evaluation, the particular challenges of evaluating educational content, and the documented impact of translation quality on learning outcomes. In Section 3, the machine translation systems and evaluation criteria, as well as the data collection procedures adopted in this study, are described. Section 4 shows results for the machine translation quality, Section 5 discusses the study's findings, and the paper ends with a summary of its main contributions and suggestions for future research in Section 6.

Literature Review

Previous Studies on Machine Translation Quality Evaluation

Much work has been done on the issue of evaluating the output quality of machine translation systems. For example, Chen, Acosta and Barry (Chen et al., 2017) provide a comparison of machine and human translation of diabetes education information. Conclude that whereas machine translation can produce output which is generally understandable, subtle errors in specialized texts have to be carefully reviewed by humans to be recognized. User-perceived quality of machine translation usability and its quality have been studied from the angle of users of such systems. The study concludes that user-perceived quality and linguistic accuracy do not necessarily go hand-in-hand, and that user-friendliness is a very important factor

ISSN: 2278-4012, Volume:13, Issue:4, 31 December 2024

for quality evaluation (Kasperė et al., 2021). A systematic review of existing machine translation systems and quality evaluation techniques provides a review of the methods that have been used in the field and mentions the ongoing debate regarding the relative merits of automated and human evaluations, which have their advantages and disadvantages (Rivera-Trigueros, 2022). Further research in this area is done by studying multilingual communication accomplished by the means of a best-balanced strategy of machine translation, when quality is evaluated and balanced for all languages included in multilingual communication instead of being assessed relative to only one source or target language (Pituxcoosuvarn & Ishida, 2018).

Challenges in Evaluating Translation Quality for Educational Content

The second stream of research has explored the specificities of assessing the quality of machine translation for educational purposes. In a systematic review of neural machine translation in foreign language teaching and learning, one should bear in mind that educational uses of this technique pose evaluative challenges, such as educational appropriateness and learner understanding, which were not covered by general translation quality assessment metrics at the outset (Klimova et al., 2023). Speaking about whether the results of machine translation are ready for use in public communication, including public health information, and consequently, in educational purposes, conclude that the quality assessment in this area has to consider the risks of miscommunication much more seriously than in general cases of translation (Pym et al., 2022). Investigating the actual use of machine translation by international higher education students in their studies, come to the conclusion that the criteria used by students in informal assessment of the translation quality do not coincide with the criteria in scientific quality assessment research (Zhou et al., 2022).

Impact of Translation Quality on Student Learning Outcomes

Thirdly, some researchers have explored the effects of machine translation quality on the effectiveness of communications and learning. Specifically, in their examination of the influence of real-time machine translation on multilingual collaboration in global software development projects, found that translation quality directly influences the effectiveness of communication, which has important implications for multilingual educational communication under similar real-time conditions (Calefato et al., 2016). Mention that several deep-learning-based translation systems have already achieved a quality level equal to human professional translation when translating news texts and wonder if similar levels are already attained in some specialized educational areas or can be attained at all (Popel et al., 2020). After carrying out a systematic review of machine translation technologies for use in health communication, an area similar to education because of its requirement to translate information both accurately and accessible to people without relevant expertise, it was concluded that translation quality indeed plays a part in how well the recipient will understand and use the information (Dew et al., 2018). This body of literature provides the conceptual background needed to evaluate the quality of machine translation in multilingual educational communication.

According to previous research, although machine translation is usually helpful in learning and offers good fluency in general domains, the results obtained from machine translation are not proficient when dealing with terminological issues, quality perceptions of users, and cultural or educational appropriateness. This study aims to build on previous work by exploring the particular gap in terms of assessing educational texts. The study proves that general fluency is inadequate in learning.

ISSN: 2278-4012, Volume:13, Issue:4, 31 December 2024

Methodology

Translation Quality and Communicative Adequacy

The assessment of the quality of machine translation performed in this research is based on several existing theories in the field of translation studies. One of the key distinctions in this area of study is that between formal equivalence, where the goal of the translation is to convey as accurately as possible the linguistic form of the source text, and dynamic or functional equivalence, where the aim is to convey the effect produced by the source text upon its target readership, although that may involve some deviations from the linguistic form of the original text. In education-oriented translation, functional equivalence should be a more appropriate concept, since the key task is to allow the learners to understand the message conveyed.

This dichotomy leads us to the next approach to the concept of translation, which can be called functionalist or goal-oriented. In this respect, the quality of translation depends not on some universally accepted criteria of accuracy but on whether the translation performs the intended communicative function in relation to the intended addressee. If this concept is applied to education, it means that the translation quality depends not only on the language itself but on the extent to which the translated material serves the educational goal that the translator had in mind.

The current research is based on the concept of adequacy and fluency, which are well-known in the field of translation quality assessment, as two independent qualities of translations. The concept of adequacy concerns the extent to which a translation reflects the content and meaning of the source text, while fluency concerns the extent to which the translated text sounds natural and idiomatic in the target language. Adequacy and fluency are not necessarily positively correlated because a translation can be very fluent but contain serious distortions in meaning or very adequate but read badly. The evaluation criteria for the current study are discussed below.

In addition, the research draws on a broadly reader-centered or comprehension-oriented perspective concerning the quality of communication, according to which the final criterion of the success of a translation is how well the translation can be understood and acted upon by its targeted reader. In the educational setting, a reader-centered perspective means that any assessment of translation quality should take into account not only the source text but also the targeted reader, knowledge about the subject matter, familiarity with terminological issues, as well as cultural and linguistic backgrounds.

Figure 1 shows the process by which the machine translation systems used for educational communication are evaluated, which consists of four steps. Starting from the 60-passage balanced corpus of both science and humanities texts, the source text is translated into the target language by means of three different systems. Human evaluators rate the translations based on five criteria of pedagogic quality, namely adequacy, fluency, terminological correctness, cultural adequacy, and readability.

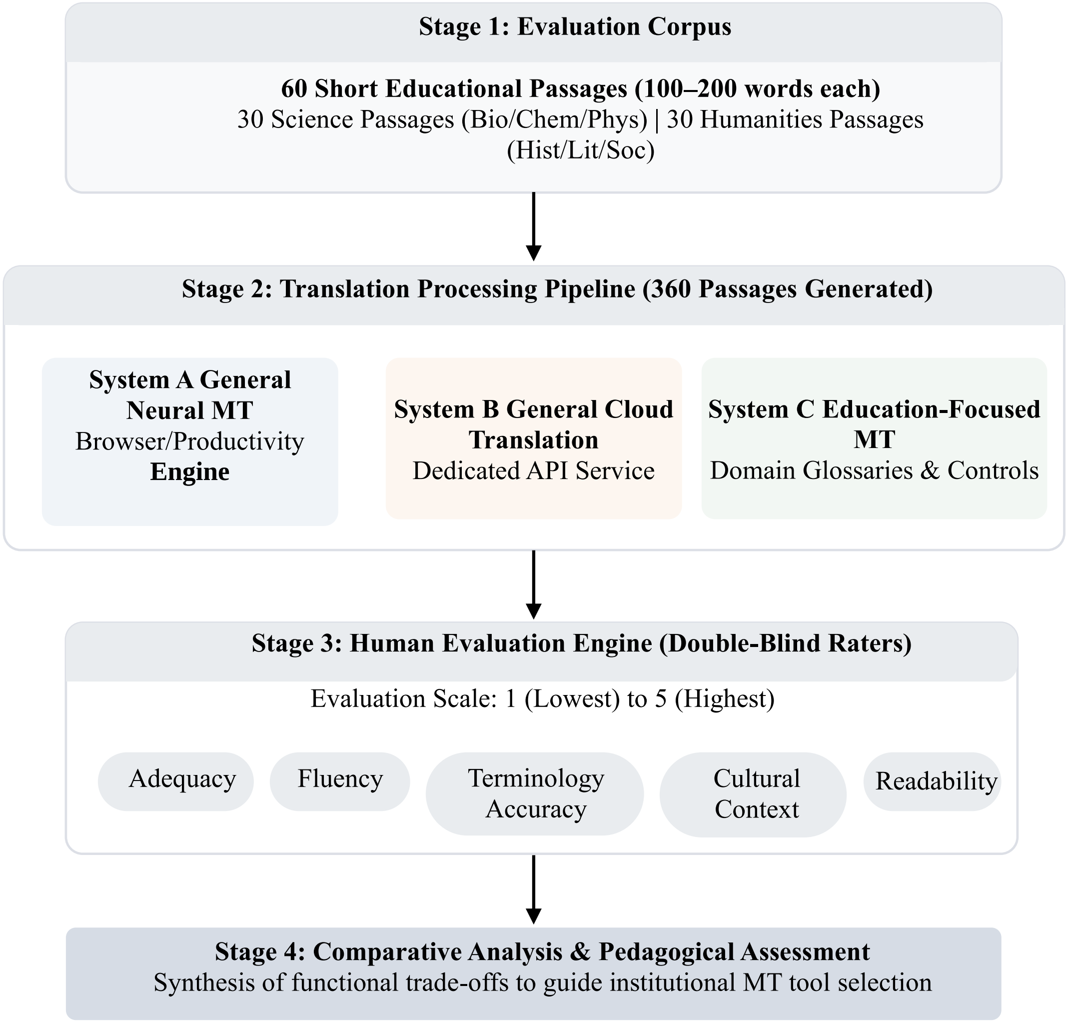

Figure 1: Machine Translation Quality Evaluation Framework

Selection of Machine Translation Systems for Evaluation

Three machine translation systems were used for this research project, and their selection was based on them belonging to the three main types of machine translation technology currently available for educational purposes. System A was one of the most popular neural machine translation tools, which can be embedded into web browsers and productivity applications. System B was one of the most popular cloud-based general-purpose translation tools, accessible through an API instead of a browser plug-in. System C was a translation system targeted at students and learners and featured a glossary of discipline-specific terminology and settings of simplified language. This selection was done to allow for a comparative analysis of general-purpose translation technology and technology specifically developed for educational purposes.

Criteria for Assessing Translation Quality in Educational Communication

The translation quality was measured based on five measures, with each focusing on a particular aspect of the quality of communication in respect of the educational material. The adequacy measure was focused on how well the meaning of the original text was kept during translation. The fluency measure looked at the flow of language of the translated text when it is considered on its own. The terminological accuracy measure considered how the terms used in the source text were translated. Cultural appropriateness was concerned with whether the examples and idioms used in the source text were appropriate in the target culture. Readability focused on whether the translated text was of an appropriate difficulty level for the learners.

Data Collection and Analysis Methods

The corpus employed for the evaluation purpose comprised sixty short educational texts that varied in length from one hundred to two hundred words. The texts were selected in equal numbers from science (biology, chemistry, and physics) as well as humanities (history, literature, and social studies). Each text was translated by all three MT systems into two different languages, yielding three hundred sixty translations altogether. Each of the obtained translations was evaluated by two human raters according to the above-described criteria of quality, the mean quality scores being calculated for each criterion for each of the MT systems. The results are shown in the next section.

Results

Evaluation of Machine Translation Quality for English Educational Content

For the evaluation of MT in English educational materials, the process must extend beyond an assessment of fluency into that of assessing more domain-oriented factors. While general translators offer a good degree of fluency and natural phrasing, fail at addressing the issues of terminology and cultural suitability. Specialized translators for educational purposes obtain much higher results in terms of cultural suitability, accuracy of terminology, and readability through the use of specialized glossaries and language settings.

Table 1: Mean Human-rated Translation Quality Scores by System and Criterion (scale of 1-5)

Evaluation Criterion	System A (General Neural MT)	System B (General Cloud Translation)	System C (Education-Focused MT)
Adequacy	4.1	3.8	4.4
Fluency	4.3	4.0	3.9
Terminological Accuracy	3.6	3.9	4.5
Cultural Appropriateness	3.4	3.2	4.2
Readability	4.0	3.7	4.3

Table 1 shows the average scores of the quality of the output text as evaluated by humans according to the five criteria used in this analysis.

Comparison of Different Machine Translation Systems

The comparison of the machine translation systems for educational content shows that there are specific tradeoffs between them. For example, System A (a general-purpose neural machine translator) is the leader in fluency (4.3), but it is poor in terms of domain accuracy. At the same time, System B (cloud-based translation) shows average performance in all the aspects without having a leading position in any of them. On the other hand, System C (educational purpose translation tool) excels in adequacy (4.4), terminological accuracy (4.5), cultural appropriateness (4.2), and readability (4.3).

ISSN: 2278-4012, Volume:13, Issue:4, 31 December 2024

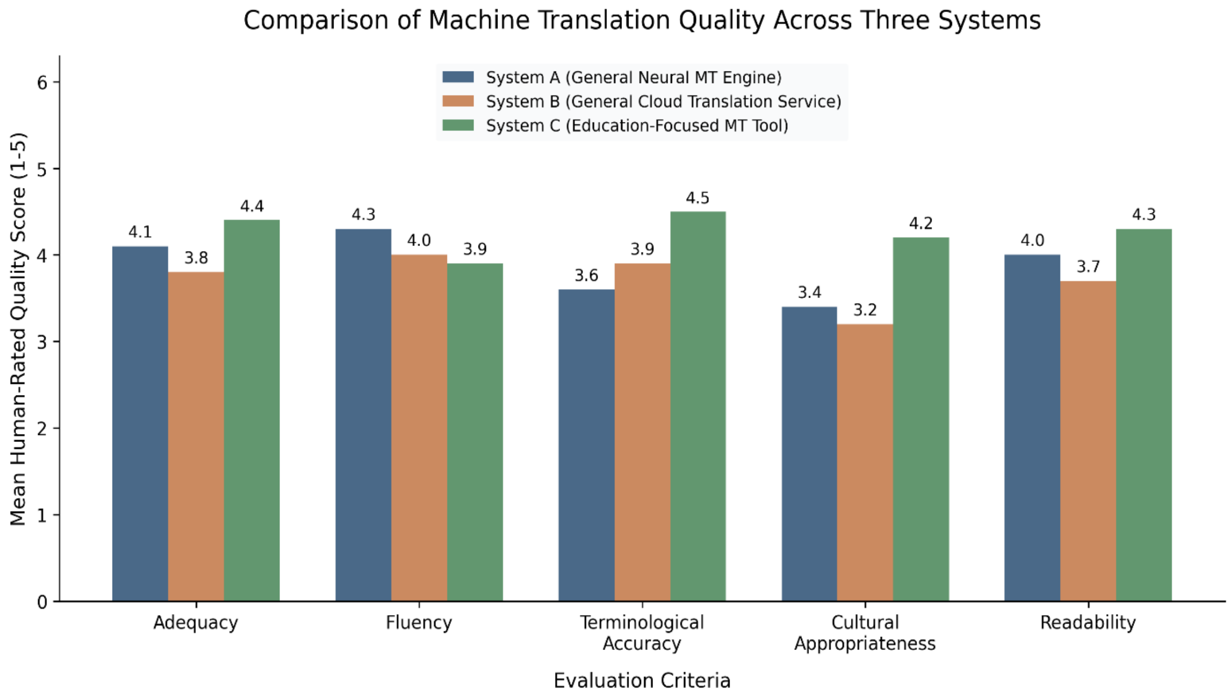

Figure 2: Comparison of Machine Translation Quality Across Three Systems

Figure 2 presents the same comparison in graphic format, providing an indication of the strengths and weaknesses of each of the systems based on the five criteria evaluated.

Identification of Areas for Improvement in Translation Quality

System A, the general-purpose neural translation engine, attained the highest score in terms of fluency compared to the other two systems, but the system performed poorly in terms of terminological accuracy and cultural appropriateness, implying that the translation produced by the system was more natural than accurate in terms of terminologies and cultural examples. System B, the general-purpose cloud-based translation system, attained results comparable to those obtained from System A but was not the best in any of the five criteria assessed. This implies that the translation produced by the system was generally average. System C, the educational purpose translation tool, was the best in terms of adequacy, terminological accuracy, cultural appropriateness and readability, although the system was the worst in terms of fluency of the translated texts, implying that there was some form of compromise between the naturalness and precision necessary for education-oriented translation. The most obvious area of weakness for all three systems is cultural appropriateness.

Discussion

The conclusions drawn from the experiment have multiple implications for the practice of multilingual educational communication. First of all, the findings imply that fluency in and of itself cannot be used as a criterion by which the suitability of a machine translation system can be judged for educational purposes because the most fluent system did not rank highest in the characteristics most directly related to comprehensibility by the learners, namely terminological accuracy and readability. Consequently, institutions choosing MT systems for educational purposes should employ a more diverse set of evaluation criteria than fluency alone or universal language translation criteria.

Based on the above results, there are several recommendations that could be provided to enhance the quality of machine translation in an educational context. First, institutions could think about adding, alongside the general-purpose translation applications, some glossaries that contain terminology specific to particular fields; it is quite possible that the inclusion of such resources was one of the main reasons for the better performance of System C in terms of terminological accuracy. Second, the developers of educational translation tools should also think about improving the performance of those systems in terms of cultural appropriateness – the weakest aspect of all systems studied in this paper; for example, this problem could be addressed by providing culturally adjusted examples instead of literal translations of the source language terms.

There are several paths to pursue for future research in light of this project. It would be interesting to explore further whether the variations in quality observed here result in any variations in comprehension by the learners as well, using a similar evaluation method as the one used here but including student comprehension. Another area for research would be the variations in quality requirements based on different categories of learners, such as age group, proficiency level, and content subject; while the present findings are based on a corpus covering a range of subjects, the findings may conceal significant variations between content types like sciences versus humanities. Lastly, it would be of interest to observe how the quality of machine translations of educational materials evolves as translation technologies evolve further.

Conclusion

This research examined the quality of machine translation (MT) in cross-language education communication through three different system architectures employing a 60-text corpus in science and humanities. The findings from the experiment demonstrate crucial trade-offs between the surface-level naturalness and pedagogic precision of the output translations. Although System A (general NMT) proved to be the most fluent (4.3 against 4.0 of System B and 3.9 of System C), System C (education-oriented MT) was found to be significantly better in all the dimensions needed for comprehension by the learners. Namely, System C performed better than general MT in terms of adequacy (4.4), terminology precision (4.5 vs. 3.6 of System A), cultural appropriateness (4.2 vs. 3.4 of System A and 3.2 of System B), and readability (4.3 vs. 3.7 of System B). The importance of the findings is in showing that general fluency cannot serve as an adequate metric of educational MT quality. Use of tools that are furnished with curated glossaries for domains and have controls for language simplification can be very helpful for learners who will be dealing with new terminology. Therefore, it is essential for schools to go beyond the traditional MT evaluation metrics and utilize a criteria-based approach to evaluate the suitability of translation tools based on their purpose. Research in the future should expand upon the current evaluative framework to explore the effect of translation quality on the performance of learners and comprehension in real time.

References

Calefato, F., Lanubile, F., Conte, T., & Prikladnicki, R. (2016). Assessing the impact of real-time machine translation on multilingual meetings in global software projects. *Empirical Software Engineering*, *21*(3), 1002-1034. https://doi.org/10.1007/s10664-015-9372-x

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In *Translation quality assessment: From principles to practice* (pp. 9-38). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

Chen, X., Acosta, S., & Barry, A. E. (2017). Machine or human? Evaluating the quality of a language translation mobile app for diabetes education material. *JMIR diabetes*, *2*(1), e7446. https://doi.org/10.2196/diabetes.7446

Deng, X., & Yu, Z. (2022). A systematic review of machine-translation-assisted language learning for sustainable education. *Sustainability*, *14*(13), 7598. https://doi.org/10.3390/su14137598

Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchhoff, K. (2018). Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, *85*, 56-67. https://doi.org/10.1016/j.jbi.2018.07.018

Han, C., & Lu, X. (2023). Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom?. *Computer Assisted Language Learning*, *36*(5-6), 1064-1087. https://doi.org/10.1080/09588221.2021.1968915

Kasperė, R., Horbačauskienė, J., Motiejūnienė, J., Liubinienė, V., Patašienė, I., & Patašius, M. (2021). Towards sustainable use of machine translation: usability and perceived quality from the end-user perspective. *Sustainability*, *13*(23), 13430. https://doi.org/10.3390/su132313430

Klimova, B., Pikhart, M., Benites, A. D., Lehr, C., & Sanchez-Stockhammer, C. (2023). Neural machine translation in foreign language teaching and learning: a systematic review. *Education and Information Technologies*, *28*(1), 663-682. https://doi.org/10.1007/s10639-022-11194-2

Lee, S. M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1-2), 103-125. https://doi.org/10.1080/09588221.2021.1901745

Nasution, A. H., Syafitri, N., Setiawan, P. R., & Suryani, D. (2017, September). Pivot-based hybrid machine translation to support multilingual communication. In *2017 international conference on culture and computing (culture and computing)* (pp. 147-148). IEEE. https://doi.org/10.1109/Culture.and.Computing.2017.22

Pituxcoosuvarn, M., & Ishida, T. (2018). Multilingual communication via best-balanced machine translation. *New Generation Computing*, *36*(4), 349-364. https://doi.org/10.1007/s00354-018-0041-7

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, *11*(1), 4381. https://doi.org/10.1038/s41467-020-18073-9

Pym, A., Ayvazyan, N., & Prioleau, J. M. (2022). Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia. *Just. Journal of Language Rights & Minorities, Revista de Drets Lingüístics i Minories*, *1*(1–2), 71-99. https://doi.org/10.7203/Just.1.24880

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, *56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Zhou, S., Zhao, S., & Groves, M. (2022). Towards a digital bilingualism? Students' use of machine translation in international higher education. *Journal of English for Academic Purposes*, *60*, 101193. https://doi.org/10.1016/j.jeap.2022.101193