UDC 81'25:004.89
DOI https://doi.org/10.32782/tps2663-4880/2025.41.1.44

AI TOOLS IN TRANSLATION PRACTICE

ІНСТРУМЕНТАРІЙ ШІ У ПРАКТИЦІ ПЕРЕКЛАДУ

Shkarban I.V.,
orcid.org/0000-0002-3450-910X
PhD in Philology,
Associate Professor at the English Language and Communication Department
Borys Grinchenko Kyiv Metropolitan University

This study examines the translation quality of AI tools (ChatGPT, Google Translate) compared to human translation (HT) across technical, news, travel, and literary texts using the Multidimensional Quality Metrics (MQM) framework. Forty-six first-year translation students have been involved in the evaluation of pre-translated English-Ukrainian and Ukrainian-English texts for accuracy, fluency, terminology, and style. The qualitative analysis revealed that AI systems, while fluent and grammatically accurate, struggled with stylistic accuracy, idiomatic expression, and metaphor translation. Google Translate often produced literal and mechanical renderings, whereas ChatGPT introduced lexical inconsistencies, particularly in culturally dense or poetic texts. Human translators displayed a consistent ability to preserve authorial voice and pragmatic nuances, especially in texts requiring emotional or aesthetic sensitivity. The study underscores the need for translator training programs to incorporate AI critically, equipping students with technical skills and the ability to evaluate, revise, and manage AI outputs. While AI tools offer valuable support for terminology management, basic comprehension, and first drafts, they cannot substitute the human translator's interpretive, creative, and ethical functions. Their limitations highlight the continued need for human oversight, especially in culturally rich or emotionally charged content. The study recommends integrating AI critically into translator training, enhancing post-editing, prompt engineering, and ethical awareness. Furthermore, the current evaluation model demonstrates the value of MQM as a robust framework for analysing translation quality across human and machine-produced texts. The use of convenience sampling and participants' novice status may affect the generalizability of results. Additionally, reliance on self-assessed translation proficiency introduces potential bias. Future research should employ standardised proficiency assessments and include more diverse and experienced participant cohorts. The need for advanced hybrid evaluation methods, combining automated and human assessment, remains pressing, particularly for culturally and stylistically complex texts.

Key words: Artificial intelligence (AI) tools, computer-assisted translation (CAT), human translation (HT), large language model (LLM) tools, machine translation (MT), Multidimensional Quality Metrics (MQM).

Статтю присвячено компаративному аналізу якості перекладу, здійсненого за допомогою інструментів штучного інтелекту (ChatGPT, Google Translate), у порівнянні з людським перекладом у технічних, новинних, туристичних та літературних текстах із використанням багатовимірних якісних метрик (MQM). Студентами першого курсу спеціальності «Переклад» здійснено оцінювання попередньо перекладених англо-українських та україно-англійських текстів за критеріями точності, грамотності, термінологічної відповідності та стилю. Якісний аналіз виявив, що, попри граматичну правильність і когерентність, системи ШІ демонструють труднощі з передачею стилістичних особливостей, ідіоматики та метафор. Переклади Google Translate переважно оцінені як буквальні й механістичні, переклади ChatGPT демонструють змістовні й лексичні неточності, що призводять до порушення когезії особливо у текстах із насиченим культурним чи поетичним контекстом. Переклади, здійснені людиною, послідовно зберігають авторський стиль і прагматичні нюанси, значною мірою в емоційно забарвлених і художньо виразних текстах. У статті наголошується на необхідності критичного впровадження інструментів ШІ у програми підготовки перекладачів, результатами навчання яких має бути не лише набуття технічних навичок цифрової компетентності, а й розвиток умінь оцінювати, редагувати та етично застосовувати продукти машинного перекладу. Обґрунтовано думку, що ШІ є ефективним інструментом на етапі первинного перекладу, обробки термінології та загального розуміння змісту тексту оригіналу, однак не здатен повноцінно замінити інтерпретативну, творчу й етичну функції людини-перекладача, оскільки виявлені недоліки вимагають збереження людського контролю над якістю перекладу, особливо стосовно культурно-специфічних та емоційно насичених текстів. Багатовимірні якісні метрики (MQM) є ефективним інструментом оцінки якості як машинного, так і людського перекладу. Водночас, застосований у цілях дослідження метод загальної вибірки та нерепрезентативна кількість учасників значною мірою обмежила узагальненість його результатів. Відсутність практичного досвіду та сформованого рівня перекладацької компетентності учасників дослідження потенційно сприяє похибкам в оцінці результатів. Майбутні дослідження повинні базуватися на стандартизованих тестах мовної та перекладацької компетентності та залучати ширшу і більш досвідчену аудиторію. Актуальним залишається питання розробки гібридних методів оцінювання, які поєднують автоматичну та експертну (людську) оцінку, особливо у випадках перекладу культурно й стилістично складних текстів.

Ключові слова: інструменти штучного інтелекту (ШІ), комп'ютеризований переклад (CAT), переклад людиною, інструменти великих мовних моделей (LLM), машинний переклад, багатовимірні якісні метрики (MQM).

Problem statement. Over the last century, a considerable amount of research has focused on Artificial intelligence (AI) technologies, such as machine translation (MT), parallel corpora, computer-assisted translation (CAT), and, more recently, large language model (LLM) tools, significantly enhancing the speed and accuracy of translation processes with the anticipation of fundamental role shift of human translation to post-editing, translation adapting, proof-reading, revising, and overall translation quality control [1, p. 1–2].

One of the revolutionary technologies that is massively utilized nowadays is ChatGPT (Generative Pre-Trained Transformer), which excessively depends on AI. It is a Natural Language Processing (NLP) model developed by OpenAI and trained on a vast dataset to generate human-like texts and provide coherent and contextually relevant responses. It has ushered in a new era of chatbot development across many key fields, such as education, healthcare, customer support, E-commerce, finance, human resources, entertainment, bias and transparency, and inaccurate content generation [2, p. 3]. ChatGPT possesses several notable strengths, including advanced natural language understanding, high scalability, multilingual capabilities, and cost-effectiveness. However, scholars have reported some drawbacks to ChatGPT, such as its reliability, accuracy, privacy, and self-confidence [2, p. 4].

Human translation (HT) constitutes a longstanding endeavour dedicated to fostering multicultural communication through transmitting cultural values and knowledge. Owing to its extensive historically established role in interlingual communication, HT is frequently regarded as the normative standard against which the performance and accuracy of machine translation (MT) systems are evaluated. As asserted by R. Al Rousan, R. Jaradat and M. Malkawi (2025), human translation is distinguishable from MT outputs due to its originality, interpretive depth, and creative dimension because the primary objective of the translator is to preserve and convey the semantic and pragmatic intent of the source text within the target text, thereby achieving a meaningful and culturally sensitive equivalence [2, p. 4–5]. Furthermore, effective translation requires advanced linguistic proficiency and a comprehensive understanding of the cultural frameworks embedded in both the source and target languages, enabling the production of a nuanced and contextually appropriate rendering [7, p. 171].

Literature Review. There have been conflicting views on the effectiveness of AI technology as a translation tool. A. Larroyed in the preliminary study of ChatGPT's performance and its impact on the current language regimes in Europe for patent translation (2023) claims that AI translation responses are logical and precise, resulting in outcomes that can be similar to the exceptional standards of human translators' proficiency as AI tool guarantees a faithful transmission of the source text's intended meaning. The author generalizes that ChatGPT can be an invaluable resource for resolving terminology-related issues because the vast amount of data available to ChatGPT allows it to recognize the various contexts in which the terminology is used and provide contextually appropriate translations. This enables ChatGPT to accurately identify and utilize pertinent terminology and provide a range of alternative translation options and explanations. Additionally, ChatGPT's ability to learn and adapt to new terminology in real-time means its performance steadily grows. Overall, her finding underscores the potential of ChatGPT as a valuable tool for specific translation tasks [5, p. 1017].

I. S. Bakhov, O. V. Stoliarenko, L. Y. Sidun and A. O. Sturba in the recent survey devoted to the comparison of traditional HT and AI-assisted simultaneous and consecutive translation (2025) argue that automatic speech recognition, speech synthesis and neural network translators definitely speed up the translation process, increase accuracy and significantly help to reduce the cognitive load on translators, but at the same time AI-assisted tools application requires new skills of adaptation and control over the quality of translation due to its limitations, in particular in contextual analysis, the transmission of cultural features and the recognition of ambiguous expressions. Aspects of ethics and responsibility in the use of automated translation systems are also considered [7, p. 158].

The rapid advancement of technology in the field of literary translation evokes both optimism and critical reflection. While interactive AI tools explicitly tailored for translation tasks may enhance user experience and engagement for literary translators, the application of MT to literary texts raises substantial concerns, particularly about quality issues and the preservation of authorial voice [1, p. 2]. Unlike legal or scientific translation, literary translation implies a comprehensive rendering of the source text's phonetic, lexical, stylistic, syntactic, and aesthetic dimensions. This is primarily due to the unique linguistic choices employed by the author, which are characteristic of their distinctive idiolect. A further area of complexity arises from the intertextual and polylogical nature of literary works. Readers across different cultural and historical contexts may interpret an author's style differently, highlighting the multiplicity of potential readings. Moreover, the act of literary translation itself is marked by significant variation, as individual translators bring their stylistic sensibilities to texts characterized by emotive and stylistically rich language. Given the inherently interpretive nature of literature, conveying the meaning of a literary text cannot be seen as a finite or objective goal as the translation of a metaphoric literary text nature extends beyond finding its semantic equivalence; it demands cultural and contextual adaptation within the target language in order to preserve its function and resonance [1, p. 3].

Within the educational context, particularly regarding the perspectives of both educators and learners, quantitative findings indicate that students generally responded positively to the use of ChatGPT. However, educators expressed reservations, noting that while ChatGPT holds promise for enhancing the efficiency of translation tasks, overdependence on such tools may impede students' creative capacities, especially when applied non critically [1, p. 4]. Language translation remains a fundamental aspect of English language education. It constitutes a core component of professional linguistic competence. However, its mastery requires an integration of theoretical understanding and practical application, namely the practical use of translation theory in authentic contexts aimed at refining translation proficiency [6, p. 1].

R. Al Rousan, R. Jaradat and M. Malkawi substantiate the idea that MT systems are vulnerable to criticism as the assessment of these systems helps in identifying the quality of their outputs and exploring the specific areas of MT that require enhancement in order to achieve more adequate performance [2, p. 5]. There are two procedures to evaluate the quality of MT outputs: automated evaluation and human evaluation. According to an extensive evaluation of automatic metrics for machine translation by T. Kocmi (2021), automated evaluation, such as BLEU and METEOeteoR, is concerned with finding the superior MT system by comparing the quality of a pair of MT systems, and technically, it is easier to apply than human evaluation. On the other hand, human evaluation is considered the best metric to evaluate MT output since it measures the quality of the MT system by comparing the MT outputs with HT [4, p. 216]. However, human evaluation is "costly and requires significant human labour, in addition to the difficulty of finding reliable bilingual annotators" [2, p. 5].

Emergent research has focused on improving translation outputs through prompt engineering. At the same time, other studies align with traditional MT research, comparing AI-generated translations with those by humans or legacy MT tools [1, p. 2].

Research Aims and Objectives. This study is driven by the lack of research about the quality of translation produced by AI, especially since it is still a new tool under trial. The research seeks to investigate potential gaps in the AI translation process and attempts to evaluate professionalism exhibited by the AI-based translation system compared to HT. In addition, as concerns have been raised over the possibility of AI technology replacing human translators, the article aims to examine the veracity of this claim. The current study contributes to advancing knowledge in the domain of MT, explicitly focusing on utilising AI technology in translation.

Presentation of the main material. HT is the conventional translation method that people have relied on throughout history. It refers to the conversion of the intended meaning of the source text (ST) from one language to another by a human translator able to create adequate and creative ST reflection in the target text (TT). Given that HT has a more extensive and well-established history than MT, it is represented as a standard by which the efficiency of any MT tool is evaluated [2, p. 2]. MT has been shown to produce a number of errors that require human post-editing, but the extent to which professional HT contains such errors shows a noted deficiency in prior studies [4, p. 215].

This study is a mixed-methods research that combines quantitative methods (pre-survey and questionnaire) and qualitative methods (reflection sheets) simultaneously to gather complementary data and examine the research question from multiple perspectives. Following the research design of A. A. Alsahil et al. [1] to ensure a more nuanced understanding of students' perspectives and practices of HT, MT and AI translation tools, the selected data were analyzed based on the latest version of the Multidimensional Quality Metrics (MQM) framework. The MQM error typology comprises eight dimensions: accuracy (faithfulness to source meaning), fluency (linguistic correctness independent of translation), terminology (domain-specific lexical choices), locale conventions (adherence to target audience norms), style (textual stylistic appropriateness), design (visual/textual formatting), verity (content suitability for the target audience), and other (errors not classified under the main dimensions). Pretranslated texts (technical, news report, travel, and song lyrics) were compiled in English-Ukrainian and Ukrainian-English pairs, interleaving AI, MT and HT segments. Forty-six first-year translation students (English – B2; Ukrainian – native) of Borys Grinchenko Kyiv Metropolitan University participated in a blind evaluation, identifying errors and post-editing the texts. Using the MQM framework, translations were assessed across four dimensions: accuracy, fluency, terminology, and style. Errors were categorised by severity as follows: minor (weight 1) – minimal impact on comprehension; major (weight 5) – affects usability but preserves meaning; and critical (weight 10) – distorts the intended meaning. These weights were used to calculate penalty scores for each translation sample, which in turn informed the overall quality score. The following equation was applied to compute the total penalty for each example: $P=\frac{minor\ issues\ \times 1+major\ issues\times 5+critical\ issues\times 10}{word\ count}$, where Table 1 presents error counts, penalties, word counts, and average quality scores for each category. Graphs accompanied the table to visualise the comparison between ChatGPT, Google Translate and HT outputs.

The comparative analysis of AI, MT and HT translation reveals no statistically significant differences between English-Ukrainian and Ukrainian-English pairs of Technical Text, News Report and Travel Brochure concerning accuracy, fluency, terminology, and style, as illustrated in Figure 1. However, it is noteworthy that HT song lyrics translation is highly valued over Google Translate (GT) and ChatGPT in the aforementioned dimensions. Students observed that GT frequently failed to preserve the original song lyrics' intended linguistic and stylistic features. A predominant concern was the lack of fluency, as the output appeared overly literal and synthetic. The system's emphasis on word-for-word accuracy and grammatical correctness came at the expense of conveying the source text's expressive means. As a result, figurative language such as metaphors, idioms, comparisons, and other literary devices was inadequately rendered, leading to translations that appeared diluted or semantically inconsistent.

GT, as an example of neural machine translation (NMT), is widely known for its instant, cost-effective access to multilingual content. Its reliance on data-driven models, powerful in processing vast amounts of linguistic data through deep learning algorithms, renders it susceptible to errors in idiomatic expressions, syntactic nuances, and culturally embedded meanings that require human interpretive competence [1, p. 10]. However, most GT-generated translations were contextually appropriate for everyday communicative situations such as travel, news reports, and preliminary content comprehension. Despite these strengths, GT also exhibits several limitations that constrain its applicability in the accurate translation of professional, technical and academic contexts.

AI technology exhibited frequent minor punctuation and typography issues. The major challenge of AI Travel Brochure translation involved translating culture-specific terms, while HT encountered difficulties accurately translating geographical terms using mostly the transliteration method. Although the AI and HT News Report translation scores are very close, the AI model produced a more fluent translation of the two, particularly regarding grammar and lexical consistency. However, AI tools generated

Table 1

Evaluation of accuracy fluency, terminology, and style

	Tech GPT	**Tech GT**	**Tech HT**	**News GPT**	**News GT**	**News HT**	**Travel GPT**	**Travel GT**	**Travel HT**	**Song GPT**	**Song GT**	**Song HT**
Minor errors	1	3	2	3	4	1	1	3	3	4	3	2
Major errors	1	4	2	0	2	0	3	3	1	4	6	0
Critical errors	0	1	2	0	0	0	0	1	0	1	4	0
Word count (TT)	500	450	556	345	322	445	567	312	603	144	134	136
Total Penalties	1.2	7,3	5.8	0.9	4.3	0.2	2.8	9	1.3	23.6	54.4	1.5
Quality Score	98.8	92.7	94.2	99.1	95.7	99.8	97.2	91	98.7	76,4	45.6	98.5

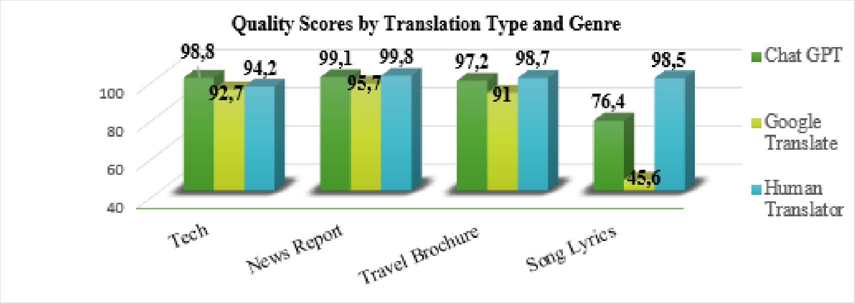

Fig. 1. Distribution of Quality Scores

superfluous words and phrases that impacted the accuracy and fluency of the song lyrics translation.

Conclusion. The findings underscore that NMT and LLMs AI tools demonstrate promising performance in translating technical, news, and travel texts, but remain considerably limited in translating literary texts, specifically song lyrics. The AI-generated translations were generally contextually appropriate for routine communicative purposes but exhibited notable deficiencies in conveying the source's text's cultural, idiomatic, and aesthetic dimensions. The human translations, by contrast, maintained a higher degree of interpretive depth and stylistic fidelity, particularly in cases demanding semantic nuance and cultural sensitivity. The error analysis based on the MQM framework revealed that AI systems often prioritise logical sequence, grammatical correctness and literal equivalence at the expense of expressive features, leading to diluted or inaccurate renderings in figurative and emotionally charged texts. Table 1 demonstrates that the average score for AI tool technical translation is 98.8%, while HT achieves 94.2%. Although both translation methods produced high-quality design scores, AI models showed higher scalability and fluency in technical translation tasks than HT. However, AI's vulnerability to punctuation and lexical inconsistencies, especially in culturally embedded and poetic content, highlights the indispensable role of human translators in quality assurance, adaptation, and meaning negotiation.

Prospects for further research. Despite the popularity of AI technology and its potential for MT, the study concludes that AI technology is not the most reliable method to generate accurate, fluent, and stylistically appropriate translations. The statistical findings indicate that AI and MT technology produced a substantially less accurate and fluent translation than the human translator, with the only exception of the technical text that can be hypothetically substantiated by the non-professional level of the novice translators in the domain-specific technical sphere. Empirical research involving larger and more diverse participant pools across educational and professional settings may offer broader insights into AI-assisted translation competencies and user perceptions. While the study offers valuable findings, certain limitations must be acknowledged. Firstly, the reliance on convenience sampling constrains the generalizability of the results, as the first-year novice translator students' participant pool may not adequately reflect the human translation proficiency. Secondly, the study's dependence on participants' self-assessed linguistic competence in translation introduces potential bias, as individuals may overestimate or underestimate their abilities. Future research should incorporate objective measures, such as standardised translation proficiency tests, to ensure more accurate assessment.

Given the growing integration of AI technologies in translation, several prospects for further research emerge. First, future studies should explore the long-term pedagogical implications of incorporating AI tools into translator training, particularly regarding their influence on students' critical thinking, creativity, and post-editing skills. Second, further inquiry into prompt engineering and its role in enhancing AI translation outputs is necessary. Customising prompts to improve semantic accuracy and contextual relevance may help mitigate some deficiencies observed in metaphorical or stylistically rich texts. Similarly, comparative studies involving domain-specific AI fine-tuning (e.g., for legal, literary, or medical translation) could provide more precise evaluations of model performance. Finally, given the limitations of current automatic evaluation metrics, further methodological advancements are needed to integrate hybrid evaluation approaches that combine computational precision with human judgment, particularly in assessing culturally and stylistically complex translations. In sum, while AI translation tools represent a significant advancement in translation technology, they should be regarded not as replacements for human translators but as complementary instruments requiring human oversight, adaptation, and ethical governance.

REFERENCES:

1. Abdelhalim S. M., Alsahil A. A., Alsuhaibani Z. A. Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts & Humanities*. 2025. 12(1). DOI: https://doi.org/10.1080/23311983.2025.2508031

2. Al Rousan R., Jaradat R., Malkawi M. ChatGPT translation vs. human translation: an examination of a literary text. *Cogent Social Sciences*. 2025. 11(1). DOI: https://doi.org/10.1080/23311886.2025.2472916

3. Calvo-Ferrer J. R. Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*. 2023. 32(6). P. 1115–1132. DOI: https://doi.org/10.1080/0907676X.2023.2268149

4. Fischer L., Läubli S. What's the difference between professional human and machine translation? A blind multi-language study on domain-specific MT [Paper presentation]. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2023. P. 215–224. DOI: https://doi.org/10.48550/arXiv.2006.04781

5. Larroyed A. Redefining Patent Translation: The Influence of ChatGPT and the Urgency to Align Patent Language Regimes in Europe with Progress in Translation Technology. *GRUR International*. 2023. V. 72, Issue 11. P. 1009–1017. DOI: https://doi.org/10.1093/grurint/ikad099

6. Wang Lan. The Impacts and Challenges of Artificial Intelligence Translation Tool on Translation Professionals. SHS Web of Conferences. 2023. 163. DOI: 10.1051/shsconf/202316302021.

7. Бахов І.С., Столяренко О.В., Сідун Л.Ю., Штурба А.О. Вплив інструментарію ШІ на особливості усного перекладу в англійській мові. *Закарпатські філологічні студії.* 2025. Вип. 39. Т. 1. С. 158–172. DOI: https://doi.org/10.32782/tps2663-4880/2025.39.1.28