# Representing emotive discourse in Ukrainian-English literary translation: A multi-method performance evaluation of Large Language Models, Neural Machine Translation and Computer-Assisted Translation tools

**Olena Karpina** [a, *], **Serhii Zasiekin** [a, b]

[a] *Lesya Ukrainka Volyn National University, Ukraine*
[b] *University College London, UK*

*Received June 18, 2025; Revised September 3, 2025; Accepted November 17, 2025*

**Abstract.** The study examines the capacity of modern translation technologies to render Ukrainian literary texts into English. Lesya Ukrainka's Letter to Serhii Merzhynskyi was chosen as the original text for translation analysis. It is a piece of emotive discourse marked by vivid imagery, nuanced stylistic features, expressive syntactic patterns and archaic vocabulary. Six translation services were tested. They included general-purpose Neural Machine Translation (NMT) services, Computer-Assisted Translation (CAT) tools and Large Language Models (LLMs). Their output was evaluated using a three-step methodological framework. First, automatic evaluation was conducted using a Bilingual Evaluation Understudy (BLEU) metric to provide initial quantitative comparability across the systems' output. Second, a qualitative analysis was undertaken through the concept of literariness, focusing on literature-specific features, aesthetic and stylistic peculiarities that distinguish literary texts from non-literary ones. In the final stage, human evaluation was employed, with five human annotators – native speakers with advanced linguistic proficiency, professional translators and scholars – ranking sentences to assess machine translation performance. The results of human evaluation and qualitative analysis revealed that the top-performing translation technologies were LLMs ChatGPT-5 and DeepSeek, which not only met a baseline level of translation adequacy but also consistently surpassed human translation in contextual and emotional sensitivity and overall naturalness and fluency. By contrast, automatic evaluation using the BLEU metric assigned the highest score to Google Translate output, highlighting the metric's limitations for literary text. Despite the notable efficiency of modern translation technologies, certain errors persist to varying degrees across all tested tools. These errors are connected with rendering imagery, handling syntactic constructions with long-range dependencies, translating pronouns, handling register mismatches, disrupting tone and other similar issues.

***Keywords:*** quality evaluation, neural machine translation, literary translation, BLEU, Computer-Assisted Translation tools, Large Language Model, emotive discourse.

* Corresponding author. Olena Karpina, 0000-0001-9520-074X, karpina@vnu.edu.ua

*East European Journal of Psycholinguistics, 12(2), 178–203,* https://doi.org/10.29038/kar

**Карпіна Олена, Засєкін Сергій. Відтворення емотивного дискурсу в українсько-англійському літературному перекладі: комплексне оцінювання продуктивності Великих Мовних Моделей, систем нейронного машинного та автоматизованого перекладу.**

**Анотація.** У статті автори аналізують здатність сучасних технологій перекладу відтворювати українські літературні тексти англійською мовою. Лист Лесі Українки до Сергія Мержинського слугує текстом оригіналу для подальшого перекладацького аналізу. Текст є фрагментом емотивного дискурсу, який характеризує емоційна образність, різноманітні стилістичні особливості, експресивні синтаксичні конструкції та наявність архаїчної лексики. Застосовано шість перекладацьких технологій, – від універсальних сервісів нейронного машинного перекладу (NMT) до інструментів автоматизованого перекладу (CAT) та великих мовних моделей (LLMs). Триетапна методологічна рамка оцінювала їхню ефективність. Спершу проведено автоматизоване оцінювання за допомогою метрики BLEU для забезпечення кількісно обґрунтованої зіставності результатів роботи систем. На другому етапі здійснено якісний аналіз крізь призму концепту літературності, що його зосереджено на літературно-специфічних особливостях, естетичних і стилістичних відмінностях, які відмежовують літературні тексти від нелітературних. На заключному етапі було проведено експертне оцінювання, під час якого носії мови з високим рівнем мовної компетентності, професійні перекладачі й науковці здійснили експертне оцінювання якості машинного перекладу, проаналізувавши кожне речення окремо. Результати експертного оцінювання та якісного аналізу засвідчили, що найпродуктивнішими технологіями перекладу є великі мовні моделі ChatGPT-5 та DeepSeek, які не лише забезпечили стандартну адекватність перекладу, а й систематично перевершували переклад людини і за контекстуальною та емоційною чутливістю, і загальною природністю та плавністю викладу. Натомість за результатами автоматичного оцінювання, проведеного за використання метрики BLEU, найвищі показники було зафіксовано у Google Translate, що засвідчує обмежені можливості цієї метрики в роботі з літературними текстами. Незважаючи на очевидну ефективність сучасних технологій перекладу, певні помилки так чи інакше зберігаються у всіх протестованих інструментах. Ці помилки стосуються відтворення образів, синтаксичних конструкцій із віддаленим зв'язком, перекладом займенників, невідповідністю регістру, порушенням тону та іншими подібними недоліками.

***Ключові слова:*** оцінювання якості, машинний переклад, літературний переклад, BLEU, інструменти автоматизованого перекладу, великі мовні моделі (LLMs), емотивний дискурс.

«Торік бувало тут, над сим потоком,
Звивала я тобі вінки барвисті,
Тоді ж у мене і квітчасті вірші
Жартуючи, лилися з-під пера»

*Леся Українка*
*"На пам'ять 31 липня 1895 року"*

By this same stream, but one year past,
I braided wreaths of blooms for you,
And all my verses, bright and vast,
Spilled laughing from my pen anew.
(Translated by DeepSeek)

# Introduction

The advent of Artificial Intelligence (AI) continues to reshape industries, driving significant changes in translation practices. The rise of sophisticated neural networks in recent decades has significantly enhanced translation quality, accelerating progress in the field of machine translation (MT) (Castilho & Knowles, 2024). Such rapid development of the field is stipulated by the availability of a large amount of data, advancement in computing technology, and the evolution of sophisticated algorithms – deep learning models – deployed in critical sectors, including Neural Machine Translation (NMT) (Mienye et al., 2024; Noll et al., 2025). Professional translators need to reconsider their role in translation by introducing MT post-editing into the translation workflow on a regular basis. The applicability of MT to different types of texts is addressed in a broad array of scholarly works (Fonteyne et al., 2020; Guerberof-Arenas & Toral, 2022; Karpina, 2023; Lan & Zhao, 2021; Zasiekin & Kalishchuk, 2025). However, despite continuous progress in NMT, some generic MT systems still perform poorly in out-of-domain translation (Emad et al., 2024).

## A Review of Recent Research into AI- and Machine Translating of Literary Texts

The translation of literary texts is considered the greatest challenge for both humans and machines (Karpina, 2020; Rybicki, 2024; Toral & Way, 2018; Wu et al., 2024). Due to its stylistic and structural complexity, it is called "the last bastion of human translation" (Toral & Way, 2014, p. 174). If literal translation – the fact machines are often accused of – is acceptable for translating technical documentation, the expectations of the readers of literary text are much higher; they are not limited to mere rendering of the meaning of the text, as readers expect to preserve the reading experience of the original. Furthermore, low-quality translations may negatively affect the author's original work (Zhang et al., 2024). Some scholars argue that even pre-translation of literary data with MT systems limits the creativity of human potential, making their output unfit for publication (Guerberof-Arenas & Toral, 2022).

Continuous progress in the field of MT, particularly with the advent of large language models (LLMs), offers promising prospects for its application. Since most MT advancements have occurred predominantly at the sentence level, recent efforts are focused on enhancing translation quality by incorporating broader contextual information (Wu et al., 2024). In this regard, the translation of literary texts has become a next-level challenge for MT

(Zhang et al., 2024). Current research demonstrates mixed results regarding the effectiveness of MT in the literary domain. While some studies claim that MT can match human translation quality (Wu et al., 2024; Toral & Way, 2018), others argue that even the most advanced LLMs produce more literal, less diverse translations than humans (Zhang et al., 2024; Karpinska et al., 2023).

Recent advances in MT go beyond refining individual MT systems, moving towards the creation of a sophisticated multi-agent framework that simulates various roles in a human translation company. For instance, Wu et al. (2024) introduced *TransAgents*, a large language model-based system that followed the traditional book translation workflow involving a diverse range of roles: from CEO to localisation specialists. Their findings show that such collaborative practices are more effective at capturing the nuances of literary texts than single-system MT models. Insights from the two-step evaluation strategy confirm the superiority of TransAgents' final output, which, in some evaluation steps, was rated higher than human translation. As for the quality of translation at the initial translation stage, it was generally high; however, it contained specific errors related to cultural adaptation and terminology management, which were eliminated during the localisation and proofreading stages.

Another essential issue is the methodology for evaluating literary MT. This issue has received increased attention as traditional evaluation metrics and schemes such as BLEU (Bilingual Evaluation Understudy) metric or MQM (Multidimensional Quality Metrics) for human evaluation widely used for non-literary translation may not always guarantee objective results: automatic metrics demonstrate limited results, functioning on the sentence level and falling short of taking the context into account (Toral & Way, 2018). As most of them rely on a reference translation, adequate translations can be penalised if they differ significantly from the reference. In human evaluation practices, on the other hand, inadequate evaluation procedures may lead to rushed judgments (Freitag et al., 2021), as the quality of human evaluation is highly dependent on the expertise of evaluators and the complexity of evaluation schemes, most of which remain untested for literary translation (Zhang et al., 2024; Zhang et al., 2025a).

Many evaluation frameworks – both automatic and human – focus on accuracy and comprehensibility as the main evaluation criteria, overlooking artistic expression. Eventually, it may lead to a significant decline in cultural authenticity and overall translation quality (Zhang et al., 2025b). In recent years, the impact of LLMs, which have demonstrated remarkable capabilities in various linguistic tasks, extended to the translation evaluation domain. Researchers began incorporating them alongside traditional evaluation frameworks to address their limitations. To address the shortcomings of existing evaluation procedures, Zhang and a team of researchers (2025b) developed LITRANSPROQA, an LLM-based framework specifically designed for literary translation evaluation. This system focuses on critical issues in literary assessment, such as stylistic devices, cultural understanding, tone, and authorial voice. Another prominent example of LLM-based evaluation solutions is GEMBA-MQM – a GPT-based evaluation metric designed to estimate translation quality by identifying and marking error spans (Kocmi & Federmann, 2023).

LLMs such as ChatGPT have been tested not only in translation evaluation tasks but also by incorporating them into post-editing of machine-translated literary texts (Macken et al., 2024). These findings show that ChatGPT improved lexical diversity over MT translated texts. However, it still solves fewer errors than human post-editors and introduces more problems by making unnecessary edits despite the preliminary instructions.

## Errors Made by MT in Literary Translation

A comprehensive analysis of errors is a crucial step in each natural language processing task, as it helps outline the prospects of future research. Error taxonomies proposed by scholars usually fall into linguistic categories or types and are influenced by the idiosyncrasies of the language examined. For instance, the research carried out by Angela Costa et al. (2017) considers error types relevant to European languages, categorising them into Orthography, Lexis, Grammar, Semantic, and Discourse types that specifically indicate the language level at which the error occurs. The researcher singles out contraction as a separate subtype within the grammar category of errors, as contraction issues are typical for Romance and also Germanic languages.

As stated above, regardless of the type of MT system applied, errors still persist in the literary domain as literary translation requires not only accuracy and fluency, but also the preservation of specific textual qualities of the original, which in some research are conceptualised under the notion of‘literariness’ *(*Toral et al., 2024). In computational literary studies, literariness is understood as a set of linguistic, aesthetic, and formal properties that distinguish literary texts from non-literary ones. These encompass a wide array of syntactic constructions and a rich semantic structure, characterised by topic variability, unexpected semantic shifts, lexical diversity, vivid imagery, etc. A high degree of creativeness, reflected in the originality of the text and non-routine patterns, as well as its aesthetic value, is often described as text “beauty” or poeticity (Jacobs & Kinder, 2022; van Cranenburgh & Bod, 2017; Toral et al., 2024).

Various qualitative and quantitative studies consistently show that, apart from general errors such as mistranslations, omissions, and fluency issues, literature-specific errors persist despite advances in domain-adapted and discourse-oriented NMT. Notably, they include flattening of the author’s style, diminished emotional engagement by omitting crucial narrative details, semantic discrepancies with the original, and literalisation of imagery. A specific example may illustrate the author's omission of stylistic repetition for emphasis. Another example of stylistic errors concerns changes in register, which were characterised by human evaluators as “excessively pompous” (Toral et al., 2024) or as more formal and less conversational than in the original. Karpinska and Iyyer (2023) also mention cultural nuances that can lead to the assignment of an inappropriate pronoun in a culture-bound context.

These findings demonstrate that despite significant advancements in NMT and LLM-powered translation, critical errors persist. Current translation services face challenges while rendering literary discourse, highlighting the need for further research into literary translation, particularly within the Ukrainian-English language pair.

The aim of this study is to evaluate and compare the quality of translation output produced by AI-powered translation tools when applied to emotionally charged literary discourse.

# Methods

The analysis compares the linguistic, stylistic, and cultural adequacy of translations of Lesya Ukrainka’s *Letter* to Serhii Merzhynskyi, “Your letters always smell of withered roses,” using six translation tools. These tools span three categories: DeepL and Google Translate are general-purpose neural machine translation (NMT) services; Matecat and Smartcat are computer-assisted translation (CAT) platforms that integrate machine translation (MT); and DeepSeek and ChatGPT-5 are large language models (LLMs) with translation capabilities. This distinction enables direct comparison across different types of translation technology.

The evaluation followed a structured methodological framework to identify differences in the capture of lexical, grammatical, stylistic, and cultural nuances across the outputs of six tested tools.

## Tools

### Translation Technologies Tested in the Experiment

Google Translate and DeepL are NMT services. They were originally built on neural sequence-to-sequence architectures and later transitioned to

Transformer-based models. These models capture long-range linguistic dependencies, resulting in more coherent, natural-sounding translations (Wu et al., 2016). DeepL and Google Translate are consistently ranked among the top free translation solutions in professional industry surveys and scholarly research. A survey conducted by the Association of English Companies found DeepL to be the most widely used MT provider, serving over 100,000 businesses and government customers across various sectors (PR Newswire, 2024). With 82% of businesses and language service providers using DeepL, it surpassed Google Translate (46%), Microsoft Translator (32%), and Amazon AWS (17%). Recent studies show DeepL generally outperforms Google Translate not only among European languages (Yulianto & Supriatnaningsih, 2021) but also in translations into Asian languages, delivering high accuracy and readability (Venita & Hasnah, 2024).

ChatGPT-5 and DeepSeek are not dedicated translators, but these services have emerged in the translation industry due to their profound contextual sensitivity and creative potential. Both systems belong to generative pre-trained transformer, a type of large language model based on deep learning architecture (OpenAI, 2025; Xu et al., 2024). While their primary capability is to process and generate text, this function can be directed towards task-specific applications.

Matecat and Smartcat are alternative cloud-based solutions for professional translators. These tools integrate human translation and MT in a single professional environment. They combine multiple tools: translation memories, terminology management, post-editing and quality assurance features, and MT engines. All are designed to streamline the translation workflow (Federico, 2014). Professionals widely prefer these platforms, but further clarification is needed on the amount of human effort required to refine their output and whether it is acceptable without human intervention. Matecat employs a Modern MT light translation engine. This is a simplified version of Modern MT real-time adaptation to users' corrections and translation memories (Matecat, n.d.). Smartcat integrates several industry-leading translation solutions. These include Google Translate, Microsoft Translator, Yandex Translate, Baidu, DeepL, Amazon MT, and modern MT. The intelligent routing mechanism automatically detects the text and selects the best MT engine for each specific language pair (Smartcat, n.d.). In our experiment, we selected the "Use for free with feedback" setting. This allows the Smartcat community to use the document's content to improve their translation performance.

The implementation of these tools, spanning different paradigms of translation technology – from conventional NMT services to LLMs and CAT tools – allows us to approach the problem of rendering Ukrainian emotive discourse from multiple perspectives, facilitating the search for the optimal translation solution for the English-Ukrainian language pair with reference to literary text.

## Material

Lesya Ukrainka's correspondence has long been the subject of scholarly interest, attracting the attention of literary theorists, historians, and textual critics, who examined her letters through various critical lenses. As evidenced in Lesya Ukrainka's creative biography, the relationship with Serhii Merzhynskyi represents an emotional and creative turning point in her life. Serhii Merzhynskyi was a 27-year-old Belarusian revolutionary, a committed social democrat and intellectual, who highly appreciated Lesya Ukrainka's creative genius. Their bond, often romanticised in historical and biographical writing, yet marked by a deeper emotional connection, high spiritual values, shared views on life and literature, an idealistic attitude toward the liberation of people, as well as deep humanity, sincerity and honesty (Koliada, 2021). Their acquaintance soon developed into a close friendship, which, for Lesya, deepened into a profound, though unrequited, love. The condition of Serhii Merzhynskyi, who was gravely ill with tuberculosis and whom she had come to take care of, was so desperate that it cost her tremendous efforts to suppress an overwhelming emotional response. This emotional experience transformed into powerful creative energy giving rise to numerous lyrical texts among which is a "Your letters always smell of withered roses" Letter (hereinafter *The Letter*), which serves as the case study of our research (Ukrainka, 2021, V. 12). Presumably, this letter was written in November, 1900, four months before the death of Serhii Merzhynskyi, influenced by the relationship between the young people, who met in 1897 while both were undergoing treatment in Yalta.

The text of *The Letter* consists of 29 sentences, each of which is translated using six translation services. It should be noted that the free versions of DeepL and Google Translate used in our experiment impose a limit on the number of characters processed at once. For the purpose of this study, we chunked the text into two segments, which narrowed the contextual environment and could, in turn, affect overall translation quality.

The language of *The Letter* is emotionally charged; it abounds in stylistic devices, cultural references, and vivid imagery. Some sentences comprise complex syntactic constructions, are longer, and are complicated by long-distance dependencies, which pose specific challenges for MT tools and increase the likelihood of mistranslations, semantic distortions, and reduced fluency.

## Methodological Framework for MT Evaluation

The evaluation framework was carried out using a structured, multi-step approach comprising three successive steps: automatic evaluation, qualitative analysis, and human evaluation.

For automatic evaluation, we employed the BLEU metric (Papineni et al., 2002). BLEU is an algorithm designed to evaluate the quality of MT output, matching it against a human reference translation. The scores are calculated for each sentence and averaged across the corpus. It compares the similarity of individual words and collocations comprising 2, 3, and 4 words, called n-grams. The value spans from 0 to 1. The more matches it detects, the higher the value, and the more similar the MT output is to the human gold-standard translation. For the human translation, we selected the version published by Stephen Komarnyckyj (2022), which includes a date indicating that no MT assistance was used. Although BLEU is not particularly effective for certain text types (Thai et al., 2022), we use it as a preparatory step to establish a baseline for further analysis, which will be complemented by human evaluation in subsequent stages of the research.

In the following step, we undertake a qualitative analysis through the lens of the concept of 'literariness', commonly discussed in translation studies. Literariness, as defined in linguistic studies, is a set of linguistic and formal features, aesthetic and stylistic qualities that provide distinctiveness to literary texts (Jacobs & Kinder, 2022; Toral et al., 2024). Empirical research often highlights that apart from general errors in accuracy and fluency, MT challenges include flattening, or 'levelling out', of the author's artistic style, the reduction of emotional impact on the reader, semantical losses, literalisation, or 'normalisation', of idiomatic language (Fonteyne et al., 2020; Guerberof-Arenas & Toral, 2022; Toral et al., 2024; Zasiekin, 2019; Zhang et al., 2025a). During the qualitative analysis stage, we manually examined the output of 29 sentences across six translation services, focusing not merely on surface-level translation errors but also on the other distinctive properties of the evaluation sample that shape it as a literary artefact. By focusing on the rendering of imagery, the preservation of emotional resonance, register consistency, and semantic faithfulness, we evaluated the general capacity of MT technologies in the literary domain that goes beyond basic accuracy. Furthermore, by examining the output of various translation technologies, we compared the potential for literary translation of each individual tool.

To ensure a greater objectivity, we commissioned the next level of evaluation to human annotators. In this stage, the evaluation was conducted by professional translators, linguists, and native English speakers with advanced knowledge of English. Each annotator received an evaluation table, where each line corresponded to the original sentence, paired with its translations, each translation displayed in a separate column. The columns labelled 1, 2, 3, etc., presented the translations of a sentence produced by the translation technologies, without revealing their names. Such a layout allowed the annotators to observe and compare the translation of one sentence in parallel, which made the ranking more comfortable. Additionally, one human translation was included in the translation set as a reference point to reduce bias during evaluation; the annotators were informed of this.

The annotators were asked to rank the translation of each sentence individually, from best to worst, and to support their evaluation rank with brief examples. To decide which translation was superior compared to the others, the annotators were instructed to primarily check if the intended meaning of the original sentence was conveyed and whether the translation sounded natural. If two or more sentences preserved the meaning to a similar extent, the annotators were asked to consider the number of errors (lexical, syntactic, and stylistic). The ranking protocol was adapted from Toral and Way (2018). Given seven versions of translation, the annotators were instructed to rank them, following these guidelines:

- Assign the higher rank (Rank 1) to a translation if its quality is higher than that of another, which should receive a lower rank (Rank 2).
- If two or more translations are of equal quality, assign them the same highest rank. For example, Translations 1 (T1) and 2 (T2) are equally good, and both outperform Translation 3 (T3). In this case, T1 and T2 are marked with Rank 1, and T3 with Rank 2. Avoid leaving unnecessary gaps between rankings.
- If it is impossible to rank a particular set of translations, the item can be skipped.

The final ranking of each translation was calculated by summing the results of the sentence-level rankings to ensure consistent evaluation across the entire translation. Such a sentence-level procedure is conceptually similar to the Best-Worst Scaling (BWS) approach adopted by Kiritchenko and Mohammad (2017), as it relies on relative judgements of annotators rather than absolute scores. But instead of selecting the single best and worst translations, the annotators were to assign an explicit rank to each sentence.

By combining automatic BLEU-based evaluation with qualitative analysis and human evaluation in our methodological framework, we expected to achieve a balanced and objective assessment of MT performance in the literary domain, integrating measurable comparability, interpretative analysis, and professional expertise.

# Results and Discussion

## Automatic Evaluation of MT Performance Using the BLEU Metric

To automatically evaluate the output of the six selected translation tools, we employed the BLEU metric, which measures n-gram overlap between the human reference translation and the computerised translation output. The evaluation was carried out using the Interactive BLEU score evaluator available on the Tilde website (Tilde, n.d.). It enables evaluating and comparing two translation outputs, aligned with the human translation.

Table 1 presents the results of the automatic evaluation for six translation tools, organised and evaluated by type: NMT-powered generic online translators (DeepL vs Google Translate), LLM-powered translation tools (GPT-5 vs DeepSeek), and CAT tools with integrated MT (Matecat vs Smartcat).

Table 1

*BLEU Scores with N-Gram Precision Across All Translation Tools*

| Translation Technology | BLEU Score | 1 gram % | 2 gram % | 3 gram % | 4 gram % |
|---|---|---|---|---|---|
| Google Translate | 26.09 | 54.09 | 41.16 | 32.31 | 26.09 |
| DeepL | 1.40 | 7.50 | 3.69 | 2.25 | 1.40 |
| Matecat | 1.66 | 7.52 | 1.99 | 1.01 | .51 |
| Smartcat | 1.04 | 7.10 | 1.47 | .66 | .17 |
| GPT-5 | 16.86 | 34.22 | 17.97 | 13.33 | 9.85 |
| DeepSeek | 1.40 | 7.50 | 3.69 | 2.25 | 1.40 |

The results of the automatic evaluation demonstrate considerable disparity in their BLEU scores across all translation tools. The highest scores (26.09) were achieved by the NMT-powered tool Google Translate. Its output showed a rather moderate translation quality, suggesting that the system captured the meaning but exhibited considerable mismatches with the reference translation. According to the n-gram precision, more than half of the individual words (54 %) overlap with the human reference translation. However, its score decreased sharply for 2–4-word strings, indicating that the system is quite accurate at rendering individual words but struggles to produce coherent multiword sequences.

Conversely, DeepL, another popular generic translation tool, scored only 1.40, showing a slightly higher degree of lexical overlap at the unigram level (7.5), while the match drops significantly when it comes to higher-order n-grams. Such a large performance gap indicates that Google Translate achieved greater lexical and structural proximity to the reference translation than DeepL, which is usually expected to be well-suited to translation tasks.

These results should be interpreted with caution, as they may suggest literal renderings that achieve surface-level similarity but do not necessarily preserve the stylistic and emotional nuances of the original text. DeepL's low scores can be explained by its tendency for lexical variability and paraphrasing, which helps maintain idiomaticity in translation. By contrast, Google Translate's higher scores may reflect a tendency toward surface-level matches, which, in subsequent evaluation steps, could lead to stylistic inadequacy and emotional insensitivity.

Surprisingly, the LLM-powered systems, ChatGPT-5 and DeepSeek, which tend to produce translations that are semantically faithful and stylistically natural, scored much lower than expected. ChatGPT-5 demonstrated low to medium performance, with a cumulative score of 16.86, significantly underperforming Google Translate in terms of lexical and syntactic similarity to the reference. Although it seems relatively modest by conventional standards, ChatGPT-5 substantially outperformed DeepSeek, particularly in cumulative n-gram matches (34.22 % for unigrams and 16.86 % for 4-grams). This suggests that ChatGPT-5 demonstrates greater sensitivity to lexical and syntactic accuracy, but its output remains less consistent than the baseline MT engine in terms of literal reference matching. By contrast, the DeepSeek system produced poor output, as indicated by BLEU (1.40), with minimal n-gram overlap (7.50 % at the unigram level and >4 % for bigrams). Such results suggest severe divergence from the reference translation, producing an output that is lexically and structurally different.

Matecat and Smartcat, leading CAT tools valued for their efficiency in translation, rely on integrated MT engines whose translation quality is considerably lower than that of NMT systems such as Google Translate or DeepL. Both Smartcat and Matecat demonstrated very low BLEU scores – 1.04 % and 1.66 %, respectively. The Smartcat CAT-tool demonstrated the lowest performance at all n-gram levels, with minimal unigram match (approx. 7 %) and less than 1 % for quadrigrams. Matecat, another CAT tool, produced a slightly better translation quality, with the overall BLEU score of 1.66. While Matecat performed slightly better, neither system can generate translations of human quality.

Such low BLEU scores across the most widely used and highly regarded translation tools do not necessarily call for reconsideration of their applicability in literary translation; rather, they raise a broader methodological concern. It should be emphasised that, despite its popularity, BLEU has certain limitations, as it tends to reject lexical and syntactic mismatches not found in a reference translation. It is particularly problematic in literary translation, where semantic nuances, imagery, creativity, and emotional undertones cannot be evaluated solely by surface-level overlap. As BLEU rewards literal renderings and penalises creative paraphrasing, the output with significant lexical and syntactic divergences will be poorly scored despite being qualitatively superior from a literary standpoint. Conversely, literal renderings with high n-gram overlap may receive inflated scores, even being stylistically and semantically inadequate. This explains why LLM-powered systems failed in BLEU evaluation, with a tendency toward greater variability, while traditional MT systems, such as Google Translate, are better optimised for standardised evaluation procedures.

Thus, while the BLEU metric provides a suggestive baseline for observing tendencies across translation tools, it does not capture the full range of translation quality in the literary domain. Its significant limitations highlight the need for human-centred qualitative analysis.

## Interpretative Evaluation Through the Lens of Literariness: A Qualitative Analysis

The central image of Lesya Ukrainka's *Letter* – the metaphor "зів'ялі троянди / зів'ялий квіт" ("withered roses / withered flower") – is vital for the emotional palette of the text, as it represents her dying beloved. Of the seven translation versions, including a human gold-standard translation, three translated the collocation as "withered flower" or "withered rose", whereas a human translator and DeepSeek chose the adjective "faded". The other two translation tools opted for the collocations with the adjective "wilted". Frequency data from Google Books indicate that the collocations "withered flower" and "faded flower" are more common in English, particularly in American English, where the collocation "withered flower" was encountered 3044 times. The colocation "wilted flower" is less common, with the number of occurrences 1177 in American English and only 119 in British English.

However, a thorough analysis of dictionary definitions of the term "withered" occasionally carries certain semantic associations of old age. For instance, in Collins COBUILD Advanced Learner's Dictionary, "if you describe a person or a part of their body as **withered**, you mean that they are thin and their skin **looks old**" (Collins English Dictionary, n.d.). Similarly, in Oxford Learner's Dictionaries, this term is applied to people's appearance, carrying the meaning of looking old: "looking old because they are thin and weak and have very dry skin" (Oxford Learner's Dictionaries, n.d.). Such an interpretation distorts the portrayal of the young and handsome, though terminally ill man. In this regard, the term "wilted" or "faded" used by DeepL, Matecat, and Human Translator seems to better render the intended meaning.

In literary translation, it is crucial to maintain consistency in rendering central imagery to achieve a cumulative effect throughout the text. However, several tools showed inconsistent behavior when replacing the key metaphor with synonymous collocations. In this respect, Google Translate and ChatGPT-5 consistently preserved the central image, while DeepL, Matecat, Smartcat, and DeepSeek alternated between synonymous expressions, blurring its symbolic impact.

Serious mistranslations arose from ambiguity caused by polysemous words. For example, the word "листи" was incorrectly translated as "leaves" (Google Translate, Smartcat) rather than "letters", resulting in serious semantic distortion. Another example of semantic misinterpretation is observed in rendering the extract "я піду до тебе з найщільніших обіймів". DeepL translated it as "I will **come** to you **with** the tightest hugs", failing to render the intended meaning "I would **go** to you **from** the tightest embraces."

Beyond inconsistency and semantic misinterpretation, another recurrent issue was literalism. For example, the plural noun in the collocation "легкі, тонкі пахощі" was literally translated into English as "fragrances". Although technically correct, it does not preserve the aesthetic elegance of the Ukrainian word "пахощі", which conveys a richer poetic quality than "запах" or "аромат". The more naturally sounding variant "fragrance" was employed by LLMs and neural-powered translation tools.

Literal translation also affected syntax, with certain tools mirroring the structure of the Ukrainian sentences – an error often made by Google Translate or DeepL. It occasionally led to awkward phrasing and semantic shifts. Consider the translation of the following original sentence performed by Google Translate:

> **Original:** І ніщо так не вражає тепер мого серця, як сії пахощі, тонко, легко, але невідмінно, невідборонно нагадують вони мені про те, що моє серце віщує і чому я вірити не хочу, не можу.
>
> (1) **Google Translate:** And nothing strikes my heart so much now as this **fragrance**, subtly, lightly, but invariably, irresistibly **they** remind me of what my heart foretells and what I do not want to believe, **cannot**.

In this sentence, the final word “cannot” appears detached and fragmented. The use of the pronoun “them” refers to the noun “fragrance” as a result of too literal translation from Ukrainian – the singular word “fragrance” corresponds to the Ukrainian word “пахощі”, which requires certain changes in the further context, replacing this lexical unit with the pronoun “it” instead of “them”.

Difficulties also emerged from syntactically and lexically complex Ukrainian sentences, which contain multiple clauses and extended structures that computerised tools often struggle to render accurately and fluently. This issue is clearly seen in the following comparison of ChatGPT-5 and Matecat translation versions:

> (2) **GPT-5:** And nothing now strikes my heart so deeply as this fragrance—fine, subtle, yet inevitably, irresistibly reminding me of what my heart foretells, and what I do not wish, cannot bring myself, to believe.
>
> (3) **Matecat**: And nothing now strikes my heart so much as sowing incense, subtly, easily, but indelibly, indefatigably reminding me of what my heart foretells and why I do not want to believe, I can not.

While LLM-powered translation managed to preserve the poetic nature of the text, opting for a balanced use of vocabulary and syntax, the CAT tool proceeds to lexical inaccuracies and syntactic incoherence, which breaks the emotional flow and diminishes the expressive power of the text (e.g. adverbs “easily”, “indefatigably” seem contextually irrelevant, the collocations “sowing incense”, “indelibly, indefatigably reminding me” appear overly complicated, which reduces emotional intensity, while disrupting the fluency of the passage).

Likewise, the Matecat tool showcased semantic inaccuracy along with stylistic and syntactic clumsiness in the following example:

> (4)This is nothing that you never hugged me, **this is nothing that was not between us and the memory of kisses**, oh, I will go to you from the tightest hugs, from the sweetest kisses!

Further difficulties involved tautology and semantic reduction in translating two distinct Ukrainian terms into a single translation equivalent. The Ukrainian verbs “томить” and “мучить” were both occasionally translated as “torment” by most translation tools. LLMs, however, along with the Google Translate NMT tool, managed to preserve semantic variation and avoided tautology, as illustrated in the following example:

> (5) **GPT-5:** Everything that **wears me down**, everything that **torments** me—I know you will lift away with your thin, trembling hand <…>

In contrast to the weaknesses mentioned above, computerised translation tools produce stylistically powerful translations, particularly in LLM-powered output. For example, the phrase: “О, я знала ще інше життя” literally translated by NMT and CAT tools as “Oh, I knew another life” was rendered as “Oh, I once knew another life, too” and “Oh, I knew yet another life” by ChatGPT-5 and DeepSeek, respectively. Both versions are more emotionally nuanced and more closely aligned with the original.

Despite greater stylistic capacity and higher emotional sensitivity, LLM-powered tools are prone to improvisation, a trait inherent to their generative nature. Such translations, although stylistically appealing, can be unacceptable in the translation sphere, where the meaning and the form of the original are prioritised. In this respect, ChatGPT-5 achieved greater balance, maintaining semantic coherence and stylistic closeness. DeepSeek occasionally demonstrated semantic and formal divergencies. For example, the Ukrainian word “простір”, translated by all the translation tools as “space”, acquired additional connotations of emptiness in DeepSeek’s version as “void”.

It should be noted that the stylistic distinctiveness of Lesya Ukrainka’s writing is deeply rooted in several factors. Firstly, her language bears traces of its time of creation, reflected in numerous archaic constructions no longer found in contemporary language. Secondly, the author’s style embodies the peculiarities of the broader Romantic movement, which manifest in high expressiveness, vivid imagery, and symbolism (Lihus & Grinchenko, 2021). These features are intensified by the biographical foundation of her writing, enhancing the emotional force and making the text deeply personal and touching.

The elevated register of Lesya Ukrainka’s writing is considerably affected by lexical choices in the translation output. The use of such emotionally charged words as “perish” (vs. “die”), “land (vs. “country”), and “amidst” (vs. “in the middle”) enhanced the stylistic colouring of translation. Similarly, parallelism and inversion were used as compensatory techniques to generate a more stylistically resonant translation. In the following translation, LLM employed inversion, combined with parallelism, to focus on the object of the action – the person for whom the action was carried out, closely mirroring the Ukrainian version: “Я ж для тебе почала нову мрію життя, я для тебе вмерла і воскресла”. ChatGPT-5 renders it as “For you I have begun a new dream of life, for you I have died and risen again.” Likewise, translation compensation employed by LLM-powered translation tools is observed in the use of an archaic negative construction, “it matters not”, to compensate for the old-fashioned Ukrainian phrase “Се нічого”. The emphatic construction in the

Ukrainian text is rendered with the help of an intensifier placed at the beginning of the sentence.

(6)**Original**: “Ціною нових молодощів **і то** я не хочу життя”

(7) **DeepL**: Even at the price of new youth I do not want life.

This technique was employed by most translation tools. CAT-based tools, on the other hand, showed semantic inaccuracies, rendering the meaning incomplete or even distorted. Matecat, for example, misinterpreted the meaning of the passage, personalising the abstract concept “молодощі” used by the author as a poetic synonym for “youth” – the state of being young, and translating it as “new young people”, which makes it completely out of context. Additionally, the literal translation of the collocation “і то” as “and then” ruined the emphasis of the sentence.

(8)**Matecat**: At the cost of **new young people**, **and then** I don’t want life.

Register mismatches also influenced the quality of the translation output. The elevated, prayer-like tone of *The Letter,*characteristic of Romanticism in literature and shaped by its biographical context, makes the use of contractions inappropriate, breaking the solemnity of the original. Among the translation tools tested, only ChatGPT-5 consistently observed full forms throughout the translation. Some translation tools occasionally use overly colloquial language. For example, the solemnity of the phrase “Мій друже, мій друже, невже я одинока згину?” was broken by Google Translate in its flattened colloquial version “My friend, my friend, will I really die alone?” ChatGPT-5 and DeepSeek, on the other hand, offered more stylistically attuned translations: “My friend, my friend, shall I perish all alone?” and “My friend, my friend – must I perish alone?” respectively. Smartcat, however, failed to convey the intended meaning, reducing it to the incomplete phrase “My friend, my friend, am I alone?”, omitting crucial information about the dynamics of loneliness.

Errors connected to sense reduction are not occasional in the output, produced by CAT-based tools. Matecat, for instance, overlooked a meaningful part of the sentence “Я так боюся жити!” translating it as “I am so afraid”, which should have specified the fear of living.

Turning to the issues of accuracy, several instances of severe grammatical errors have been observed: For example, Smartcat used a singular article “a” to a plural noun – “a child’s sobs”. Google Translate introduced distorted temporal relations, arising from sentence-level translation that focuses on markers in isolation rather than the broader context.

> (9)**Google Translate:** And then both happiness and grief broke as suddenly as a child's sobbing, and I saw you. I **have seen** you before, but not so transparently <...>

In this excerpt, the use of the present perfect tense ("I have seen you") is wrong, as the intended meaning refers to the past event occurring before another past action, which makes the past perfect an appropriate choice ("I had seen").

Further issues that compromise translation accuracy across the tools tested include unnecessary capitalisation (Smartcat), omissions (DeepSeek), and incorrect punctuation (Matecat). As for the pronoun use, only LLMs maintained neutrality by employing gender-neutral pronouns "them" and "it" with reference to a child. However, the use of the pronoun "her" can be contextually justified, reflecting the author's self-identification with the image of a weeping child, as specified in the example: "...like a crying child goes into the arms of the one who pities her".

These observations modify the picture provided by BLEU metric evaluation, offering a more detailed perspective on the performance of the tested tools, while the next step of the evaluation procedure – human evaluation by professional translators and native speakers – enhances the credibility and objectivity of the study.

In terms of translation universals (Chesterman, 2011) or regularities (Zasiekin, 2019), the studied LLMs tended to 'normalisation' and 'complication' in the target versions. In contrast, the NMT tools demonstrated an overall tendency to 'implicitation' and 'levelling out'. The CAT tools, on the other hand, showed primarily 'simplification', producing less sophisticated versions of *The Letter*, abundant in punctuation and stylistic errors.

## Human Evaluation: Expert Ranking of Translation Output

The human evaluation procedure ranked the translation outputs produced by different translation tools. Given the diversity of the annotators' backgrounds, ranging from native speakers with advanced proficiency in English to professional linguists and translators, it was expected that the native speakers would primarily focus on general sentence readability, while the linguists would engage more deeply with semantic and stylistic subtleties. By adopting such an evaluation strategy, we expected to integrate a dual perspective, combining the expertise of bilingual professionals with the intuitive judgments of native speakers, who are particularly attuned to fluency and naturalness rather than accuracy. Although the evaluation is marked by noticeable differences, the variations were shaped by subjective preferences rather than by linguistic experience or native speakers' intuition, which seem complementary to the former.

Based on five human annotators' rankings, the best-performing translation tools were language model-based DeepSeek and ChatGPT-5, both of which consistently outperformed human translation across most evaluation metrics. This tendency of LLMs to be selected as the best-performing translators is particularly noticeable across multi-clause, emotionally intense sentences, which prompted detailed comments from the annotators. It highlights that, in addition to linguistic accuracy, LLMs are more sensitive to emotional and stylistic nuances.

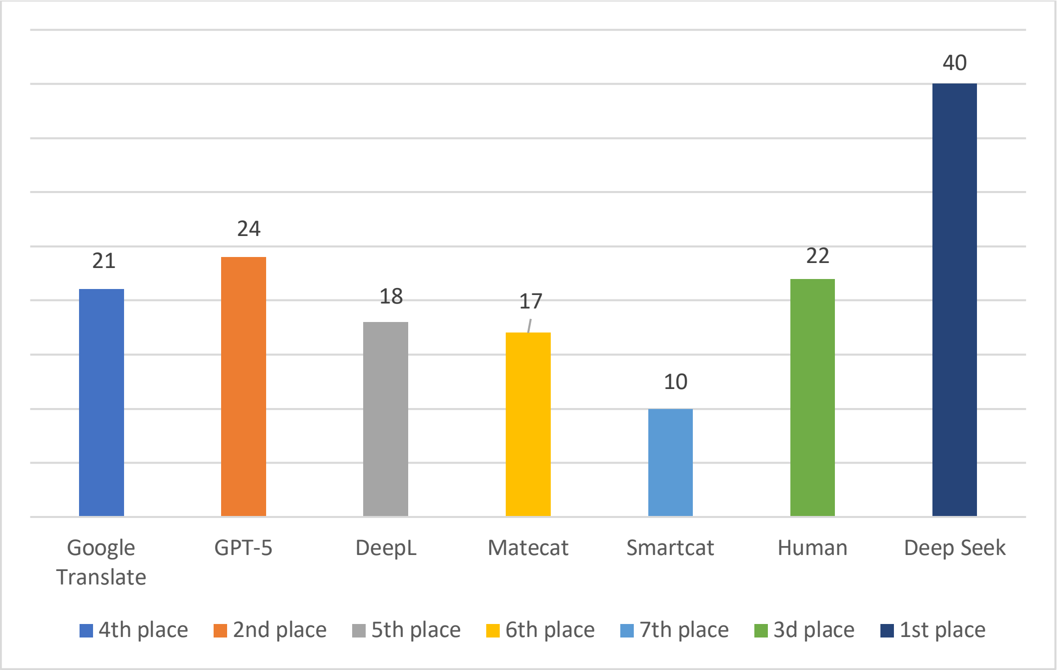
These tools are followed by neural-powered Google Translate and DeepL. As for Google Translate, its outcome is mostly driven by unexpected preferences of some annotators to choose simplicity bordering on literalness over emotional intensity. Its output was occasionally tied to that of other, more sophisticated tools in short, formulaic sentences, which were treated interchangeably. Google Translate and DeepL often received a relatively similar number of preferences across evaluation forms. To resolve this, we normalised the final score by the number of instances in which the system was ranked last. Our assumption was that a tool that received a few top rankings, even if it also received some of the lowest rankings, would have performed better than a tool that was never selected as the best but avoided the lowest rankings. The latter case does not necessarily demonstrate excellence, but rather suggests reliability and consistency.

Matecat and Smartcat manifested the lowest results, being unanimously ranked the worst by both native speakers and professional translators. These calculations are based not only on the scarcity of instances in which the systems were rated best, but also on the high number of cases in which they received the worst ranking. This tendency was traced not only across long and complex sentences but also in shorter ones, where other tools generally handled translation better.

A generalised overview of the systems' translation performance is illustrated in Fig. 1 below.

Most of the annotators' observations concerned stylistic and semantic subtleties, referring either to the level of expressiveness or contextual appropriateness rather than serious grammatical inaccuracies or lexical mismatches. The annotators rarely marked the machine output as utterly unacceptable; some of them explicitly highlighted that "none of them are bad, which points out that the overall output is broadly intelligible for native readers. Although occasional errors occurred, the overall evaluation suggests that MT has finally achieved a sufficient level of adequacy and fluency to excel in the literary domain and compete with human translation skills. The feedback provided through occasional comments in the annotation form revealed certain tendencies, concerning lexical choices, grammar inaccuracies affecting pragmatics, modality errors, morphological distortions, punctuation issues, and stylistic and tonal mismatches.

Figure 1
*Rank Distribution Across the Translation Tools*



Some annotators, native speakers without translation expertise, showed a tendency toward surface-level fluency. They mostly favoured ChatGPT-5 or DeepSeek output intuitively, justifying their decision “as a matter of choice”. Such an evaluation could not assess fidelity to the source, limiting the judgment solely to the immediate readability of the content in English. It also explains tied rankings or skipping rankings, occasionally indicating that “there was hardly any difference between most of them” (sentences).

A frequent issue flagged by the annotators concerns the selection of lexical units in a given context. Some words, although technically acceptable at the denotative level, slightly modify the meaning, introducing extraneous connotations different from those intended in *The Letter* or failing to align with the tone of the original. Such contrast pairs as “wilted” vs “withered”, “pity” vs “console”, “vanished” vs “bygone”, “void” vs “space”, “joy” vs “happiness”, “exile” vs “foreign land” and the like illustrate the instances of translation, where vocabulary choice shifted the nuances of the meaning. At the same time, variations between the words “darling” and “love” disrupted the register of the source text.

In other cases, difficulties were identified with the attribution of words to morphological categories. For example, in the collocation “my poor withered blooming”, “blooming” was treated as a noun despite its verbal nature. Similarly, “beloved”, according to some annotators, needed to be expanded to “beloved friend” to ensure accuracy and natural sounding. In some sentences, the choice of a morphological category or a vocabulary unit was influenced by the original phrasing, resulting in literal renderings that were either grammatically incorrect, awkward, or semantically distorted. The examples highlighted by the annotators included “this is nothing” (це нічого), “densest hugs” (найщільніші обійми), “see transparently” (бачити прозоро), among others.

Other grammar-related errors, which downgraded the translation output, concerned modality shifts, e.g., “will/shall I die alone” vs “must I perish alone”, where “shall” and “will” flatten the perception of surprise or rhetorical doubt. Another observation focused on the verb choice in the sentence “Тільки ти вмієш рятувати мене від самої себе”, where the occasional preference of the verb “can” (вмієш) to the verb “will” obscures the intended meaning of ability, replacing it with the implication of certainty of the future action.

It is noteworthy that, in several sentences, native-speaking annotators requested additional explanations, as certain lines remained obscure to them despite being well-structured and accurately rendered in vocabulary choice. Even after being provided with the explanations, they continued to express doubt regarding the intended contextual meaning, admitting that the overall situation “doesn’t make any sense”. In another instance, an annotator sought clarification with a line mentioning blue roses, where the choice of the colour was intentional, alluding to Lesya Ukrainka’s drama “Blue Rose”, highly valued by the recipient of *The Letter*. Such cases indicate that certain translation decisions are closely tied to extralinguistic knowledge, highlighting interpretative challenges faced by individuals unfamiliar with the Ukrainian cultural mindset and relevant context.

Overall, the annotators were actively engaged in the evaluation process and identified some positive instances that convey the emotional cadence and stylistic qualities of *The Letter*. One of the examples praised by the annotators refers to the phrase “and nothing strikes my heart so deeply”, where “strikes deeply” was commended for its natural sounding and emotional impact, illustrating the capacity of some translation engines to achieve stylistically convincing results. Some annotators favoured the translation output that, contrary to the common choice of the term “space” in that context, opted for “void”. Such language choice shifted the meaning towards the concept of “emptiness”, which the annotators considered emotionally appealing given the context. Interestingly, one annotator further contextualised the choice of the term “void” through the intertextual parallel to Ozzy Osbourne’s creative legacy, referencing his song “Into the Void” – a speculative, yet ungrounded suggestion.

Annotators paid much attention to lexical precision but rarely noted syntactic inaccuracy. Long syntactic constructions, initially seen as challenging, did not attract significant criticism in evaluators’ feedback. This suggests modern translation tools performed better than expected. The main limitations lie in lexical and semantic disruptions.

## Conclusions

The evaluation of Lesya Ukrainka’s *Letter* using a multi-method approach – automatic BLEU metric, qualitative analysis, and human evaluation – provides a detailed overview of the performance of contemporary translation technologies. The fact that LLMs, such as ChatGPT-5 and DeepSeek, achieved the highest evaluation scores in human assessment stages suggests that computerised translation services have come closer than ever to the long-standing dream of the early MT pioneers – Fully Automated High Quality Translation. However, certain errors, such as semantic shifts, syntactic incoherence, over-creativity, and occasional tone inconsistency, persist even in the output of the highest-ranked translation systems, underscoring the need for human expert-driven analysis as an essential part of the translation workflow and demonstrating that the human factor remains indispensable in translation practices.

The study also confirms the BLEU metric's limited applicability for evaluating literary translation, leading to a misleading impression of translation quality. The findings underscore the need to complement automatic evaluation with expert human judgment or to develop more nuanced evaluation metrics that capture the subtleties of emotive literary discourse.

## Acknowledgements

The authors express sincere gratitude to Tony Palmer, Russell Mackenzie, Nataliia Dobzhanska-Knight, and Nataliia Voloshynovych for their selfless assistance in conducting the human annotation work.

## Disclosure Statement

The authors reported no potential conflicts of interest.

# References

Alghamdi, E. A., Zakraoui, J., & Abanmy, F. A. (2024). Domain adaptation for Arabic machine translation: Financial texts as a case study. *Applied Sciences, 14*(16), 7088. https://doi.org/10.3390/app14167088

Castilho, S., & Knowles, R. (2025). A survey of context in neural machine translation and its evaluation. *Natural Language Processing, 31*(4), 986–1016. Cambridge University Press. https://doi.org/10.1017/nlp.2024.7

Costa, A., Ling, W., Luís, T., Correia, R., & Coheur, L. (2017). A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation, 31*(3–4), 229–244. https://doi.org/10.1007/s10590-017-9205-4

Federico, M., Cattelan, A., & Trombetti, M. (2014). *The MateCat tool.* In *Proceedings of the COLING 2014: System Demonstrations* (pp. 129–132). Association for Computational Linguistics. https://www.aclanthology.org/C14-2028.pdf

Fonteyne, M., Tezcan, A., & Macken, L. (2020). Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)* (pp. 3790–3798). European Language Resources Association. https://aclanthology.org/2020.lrec-1.468/

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9,* 1460–1474. https://doi.org/10.1162/tacl_a_00437

Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces, 11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue.

Jacobs, A. M., & Kinder, A. (2022). Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large corpus of English literature. *arXiv Preprint arXiv:2201.04356.* https://arxiv.org/abs/2201.04356

Karpina, O. (2020). Komparatyvnyi analiz literaturnoho i mashynnoho perekladiv (na materiali frahmentiv romanu S. Plath “The Bell Jar”) [Comparative analysis of literary and machine translations (a case study of the excerpts *The Bell Jar* by S. Plath)]. *Current Issues of Foreign Philology, 3,* 94–101. https://doi.org/10.32782/2410-0927-2020-12-16

Karpina, O. (2023). Evaluating the quality of machine translation output with HTER in domain-specific textual environment. *Linguistic Studies, 46,* 85–99. https://doi.org/10.31558/1815-3070.2023.46.8

Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv Preprint arXiv:2304.03245.* https://arxiv.org/abs/2304.03245

Kiritchenko, S., & Mohammad, S. M. (2017). Best–worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 465–470). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2074

Kocmi, T., & Federmann, C. (2023). *GEMBA-MQM:* Detecting translation quality error spans with GPT-4. *In Proceedings of the Eighth Conference on Machine Translation* (pp. 768–775). Association for Computational Linguistics. https://arxiv.org/abs/2310.13988

Koliada, I. (2021, February 25–26). "Ukrainska Biatrice" i "Donka Prometeia": Zhertovne kokhannia u zhytti Hanny Barvinok i Lesi Ukrainky ["Ukrainian Beatrice" and "Daughter of Prometheus": Sacrificial love in the life of Hanna Barvinok and Lesya Ukrainka]. In *Ideolohynia natsionalnoi arystokratii (na poshanu 150-richchia vid dnia narodzhennia Lesi Ukrainky)* [*The Ideologist of the National Aristocracy (in honor of the 150th anniversary of Lesya Ukrainka's birth*)], Proceedings of the International Conference (pp. 326–335). Lviv Danylo Halytskyi National Medical University.

Lan, M., & Zhao, L. (2021). Contrasting and analyzing machine and human translation: A case study on *Red Sorghum*. In *Proceedings of the 2021 6th International Conference on Modern Management and Education Technology (MMET 2021)* (pp. 684–690). Atlantis Press. https://doi.org/10.2991/assehr.k.211011.123

Lihus, O., & Grinchenko, B. (2021). Ukrainian Romanticism in the context of European culture of the 19th – early 20th centuries: Interdisciplinary historiographical analysis. *Mundo Eslavo, 20,* 147–157. https://revistaseug.ugr.es/index.php/meslav/article/view/21627/22597/79552

Macken, L. (2024). *Evaluating ChatGPT's Ability to Automatically Post-Edit Literary Texts.* In B. Vanroy, M.-A. Lefer, L. Macken, & P. Ruffo (Eds.), *Proceedings of the 1st Workshop on Creative-text Translation and Technology* (pp. 65–81). European Association for Machine Translation. https://aclanthology.org/2024.ctt-1.7/

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information, 15*(9), 517. https://doi.org/10.3390/info15090517

Noll, R., Berger, A., Kieu, D., *et al.* (2025). Assessing GPT and DeepL for terminology translation in the medical domain: A comparative study on the human phenotype ontology. *BMC Medical Informatics and Decision Making, 25,* 237. https://doi.org/10.1186/s12911-025-03075-8

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

Rybicki, J. (2025). Can machine translation of literary texts fool stylometry? *Digital Scholarship in the Humanities, 40*(1), 268–276. https://doi.org/10.1093/llc/fqaf010

Sun, S., Liu, K., & Moratto, R. (2025). Navigating the paradigm shift-translation studies in the age of AI. In S. Sun, K. Liu, & R. Moratto (Eds.), *Translation Studies in the Age of Artificial Intelligence (pp. 1*–17). Routledge.

Thai, K., Karpinska, M., Krishna, K., Ray, W., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). Exploring document-level literary machine translation with parallel paragraphs from world literature. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 9882–9902). Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.672

Toral, A., & Way, A. (2014). Is machine translation ready for literature? In *Translating and the Computer 36* (pp. 174–176). Dublin City University. https://aclanthology.org/2014.tc-1.23.pdf

Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text? *arXiv Preprint arXiv:1801.04962.* https://doi.org/10.48550/arXiv.1801.04962

Toral, A., van Cranenburgh, A. V., & Nutters, T. (2024). Literary-adapted machine translation in a well-resourced language pair: Explorations with More Data and Wider Contexts. In A. Rothwell, A. Way, & R. Youdale (Eds.), *Computer-assisted literary translation* (pp. 27–52). Routledge. https://doi.org/10.4324/9781003357391-3

Van Cranenburgh, A., & Bod, R. (2017). *A Data-Oriented Model of Literary Language.* In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Long Papers)* (pp. 1228–1238). Association for Computational Linguistics. https://aclanthology.org/E17-1115

Vennita, R. & Hasnah, Y. (2024). A probe into the comparison of human translation and deep translation in translating English text into Indonesian. *Journal of English Development*, 4(2), 303-330. https://doi.org/10.25217/jed.v3i01.4503

Wu, M., Xu, J., Yuan, Y., Haffari, G., Wang, L., Luo, W., & Zhang, K. (2024). (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv Preprint arXiv:2405.11804*. https://arxiv.org/abs/2405.11804

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv Preprint arXiv:1609.08144*. https://doi.org/10.48550/arXiv.1609.08144

Xu, Y., Zhang, B., et al. (2024). DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv Preprint arXiv:2401.02954*. https://arxiv.org/abs/2401.02954

Yulianto, A., & Supriatnaningsih, R. (2021). Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy, 9*(2), 109–127. https://doi.org/10.37134/ajelp.vol9.2.9.2021

Zasiekin, S. (2019). Investigating cognitive and psycholinguistic features of translation universals. *Psycholinguistics*, *26*(2), 114–134. https://doi.org/10.31470/2309-1797-2019-26-2-114-134

Zasiekin, S. & Kalishchuk , D. (2025). Can machines communicate psychotrauma? Affective and cognitive shifts in AI-translated Russia-Ukraine war narratives. *Psycholinguistics*, *38*(2), 58-76. https://doi.org/10.31470/2309-1797-2025-38-2-58-76

Zhang, R., Zhao, W., & Eger, S. (2025a). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. arXiv Preprint arXiv:2410.18697. https://doi.org/10.48550/arXiv.2410.18697

Zhang, R., Zhao, W., Macken, L., & Eger, S. (2025b). TransProQA: An LLM-based literary translation evaluation metric with professional question answering. *arXiv Preprint arXiv:2505.05423*. https://doi.org/10.48550/arXiv.2505.05423

## Sources

Collins Dictionary. (n.d.). *Withered*. In *Collins English Dictionary*. https://www.collinsdictionary.com/dictionary/english/withered

DeepL. (2024, October 9). DeepL is 2024's most-used machine translation provider worldwide among language service companies [Press release]. PR Newswire. https://www.prnewswire.com/

DeepSeek. (2025, June 27). https://chat.deepseek.com/

Komarnyckyj, S. (2022). Lesya Ukrainka's "Your Letters Always Smell of Withered Roses" (1947): A new translation. *Volupté: Interdisciplinary Journal of Decadence Studies*, 5(1), 92–94. https://doi.org/10.25602/GOLD.v.v5i1.1624.g1738

Matecat. (n.d.). *Free online CAT tool with integrated MT*. https://www.matecat.com

OpenAI. (2025, August 7). Introducing GPT-5. https://openai.com/index/introducing-gpt-5

Oxford Learner's Dictionaries. (n.d.). *Withered*. In *Oxford Learner's Dictionaries*. https://www.oxfordlearnersdictionaries.com/definition/english/withered?q=withered

Smartcat. (n.d.). *AI-powered CAT tool. Free online translator*. https://www.smartcat.com/cat-tool/

Interactive BLEU score evaluator (n. d.). *Tilde*. https://translate.tilde.ai/bleu#/

Українка Л. (2021) Повне академічне зібрання творів: у 14 томах. Том 12. Листи. 1897–1901 / ред. О. Полюхович; упоряд. В. Прокіп (Савчук); комент. В. Прокіп (Савчук), В. Агеєва. Луцьк: Волинський національний університет імені Лесі Українки,. 608 с., С. 321–322.

Ukrainka, L. (2021). Povne akademichne zibrannia tvoriv: u 14 tomakh. Tom 12. Lysty, 1897–1901 [Complete academic collection of works: in 14 volumes. Vol. 12. Letters, 1897–1901] / O. Poliukhovych (Ed.), compiled by V. Prokip (Savchuk) with comments from V. Prokip (Savchuk), V. Aheieva. (pp. 321–322). Lesya Ukrainka Volyn National University.