Forum for Linguistic Studies 2024; 6(1): 1994.
Original Article

AI-translated poetry: Ivan Franko's poems in GPT-3.5-driven machine and human-produced translations

Viacheslav Karaban[1,*], Anna Karaban[2]

[1] *Department of the Theory and Practice of Translation from English, Taras Shevchenko National University of Kyiv, Kyiv 03191, Ukraine*

[2] *Department of English Philology and Intercultural Communication, Taras Shevchenko National University of Kyiv, Kyiv 03191, Ukraine*

*** Corresponding author:** Viacheslav Karaban, v.karaban@knu.ua

ARTICLE INFO

Received: 16 September 2023
Accepted: 30 October 2023
Available online: 31 January 2024

doi: 10.59400/fls.v6i1.1994

Copyright © 2024 Author(s).

Forum for Linguistic Studies is published by Academic Publishing Pte. Ltd. This article is licensed under the Creative Commons Attribution License (CC BY 4.0). http://creativecommons.org/licenses/by/4.0/

ABSTRACT: The article presents a detailed comparative analysis of translations of twelve great Ukrainian poet Ivan Franko's poems done by translator Percival Cundy and the GPT-3.5 AI language model. Using various manual and automatic analytical research methods and techniques, we analyzed the translations' merits, demerits, and eight essential qualitative and quantitative linguistic and poetic characteristics to verify a hypothesis that human and GPT-3.5-driven machine translations can be quite comparable in terms of their quality and poetic features. The results obtained sufficiently prove the hypothesis and suggest that developing AI translation potential for poetry translation can help build more capable, diversified, and nuanced large language models. The AI revolutionary breakthrough in translation makes it quite possible to acquaint satisfactorily the wider public with the poetic heritage of the world's nations, especially those using minor languages, whose poetry is evidently under-translated. A follow-up study is desirable to assess the progress made by GPT4.0 and its possible later versions in poetry translation, as compared with GPT-3.5.

KEYWORDS: AI translation; Ivan Franco's poems, human translation; comparative analysis; poetry translation; translation from Ukrainian into English

1. Introduction

The realm of machine translation is rapidly evolving, capturing the attention of translators, translation scholars, linguists, computer scientists, and literary enthusiasts. This is especially pertinent to literary translation, a captivating area of interest for scholars and aspiring professionals. Among literary forms, translating poetry stands as an intricate endeavor that navigates linguistic frontiers. Translating poetry requires a delicate balancing act between maintaining the structural integrity, semantic essence, and emotional resonance of the original text. This often results in a precarious trade-off, where the pursuit of accuracy or aesthetic appeal may be compromised, and occasionally, both may suffer. In this era of remarkable advancements in artificial intelligence striving to emulate and potentially surpass human translators, significant progress has been made in the efficacy, speed, and accuracy of machine translation. Key developments include Google Translate's proficient translations of Portuguese poetry into English Humblé (2019), and the notable shift in AI machine translation evaluations, transitioning from "satisfactory" to being recognized as a "clear success" as per Poibeau (2022). However, the translation of poetry remains a largely uncharted frontier and poses a formidable challenge for machine translation

algorithms, given its complex interplay of emotions and cultural nuances woven into each verse line.

AI-driven systems have evolved beyond industrial applications and now hold potential for education and culture, including creative tasks like poetry composition and translation for broader cultural exposure. As these systems improve in accuracy and nuance, a critical question arises: how do AI-generated translations compare to human-produced poetic translations in terms of quality, creativity, and human-like attributes? Translating poetry, especially for AI systems, involves navigating complex nuances. Beyond linguistic fidelity, poetry requires understanding imagery, unique ethno-specific lexicons, and distinct poetic structures in various syntactic arrangements, necessitating advanced AI training and algorithm enhancements.

Given the current lacuna in scholarly literature focused on AI-driven poetic translations within the English-Ukrainian linguistic context, our research endeavors to address this. By systematically analyzing this niche area, our study aims to advance the understanding of AI-driven poetic translation processes. This study aims to investigate how an AI generative language model navigates the intricate challenge of poetic translation in contrast to a human translator. In analyzing these translations, we try to explore the potential and limitations of such cutting-edge translation technologies, refine research methodologies, assess the obstacles faced in attempting to imbue artificial intelligence with the essence of poetry, reflect on ways to improve automated literary translation, specifically, poetry, and consider the implications of this technological progress for literary creation, cultural preservation, and the future of human artistic expression.

2. Related work

2.1. Machine translation of literary works

Research on literary machine translation has experienced a recent surge due to the advancement of AI-driven systems in tackling the complexities of literary translation (Boulenouar 2022). The following presents a succinct synthesis of the scholarly discourse encompassing general literary machine translation, machine translation of poetry, and the comparative analysis of human and machine translation. In the realm of general literary machine translation, Toral and Way (2018) conducted an evaluation of MT systems and found that native speakers considered 17% to 34% of NMT translations comparable to those produced by professional human translators. A perspective contrary to the impending threat of AI was expressed by Hadley (2020), who believed that AI systems would not pose significant challenges to human literary translators in the near future, advocating instead for tools that aid these translators. Conversely, professional translator Hans-Christian Oeser criticized machine translation for its failure to capture the aesthetic and stylistic nuances of literature (Kenny and Winters 2020). In contrast, Matusov's (2019) optimistic stance challenged the assumption of machine translation's unsuitability for literary works, suggesting that NMT could overcome the challenges of literary translation. Kuzman et al. (2019) found that Google Neural Machine Translation outperformed custom NMTs in rendering literary translations for low-resource language pairs, such as English-Slovene. Visby (2020), president of the European Council of Literary Translators' Association, acknowledged the quality of machine translation for genre literature while admitting the need for human editors to correct errors. Grace et al. (2018) predicted the surpassing of human translation, including non-professional translators, by high-level machine intelligence by the end of the 2020s. Zong (2018) proposed a hybrid approach of blending machine and human translation to enhance efficiency and effectiveness.

2.2. Machine translation of poetry

Moving to the realm of poetry translation, the early exploration of NLP creativity was centered around poetry generation, exemplified by the Bairon system generating poems mimicking the styles of selected writers based on user input (Badura et al. 2022). The research by Vincent (2019) delved into various forms of creativity in poetry writing. Genzel et al. (2010) investigated the intricate task of poetry translation using machine techniques, showcasing the feasibility of maintaining metrical constraints in poetry translation through statistical machine translation. Studzińska (2020) acknowledged the enhancement in automated poetry translation quality but left open the question of whether algorithms could achieve human-like translations. Advancing in this domain, Ghazvininejad et al. (2018) introduced an approach for automatic poetry translation that preserves target rhythm and rhyme patterns, utilizing neural translation techniques to improve translation quality while adhering to specified constraints. Chakrabarty et al. (2021) recognized the complexities of automatic poetry translation due to semantic, stylistic, and figurative language preservation challenges, noting the effectiveness of multilingual fine-tuning on poetic text. Ma and Wang (2020) introduced a linguistic framework for analyzing and contrasting poetry translations, albeit not aimed at contrasting human and machine translation.

2.3. Comparative analysis of machine and human translations

Comparative studies analyzing human and machine translations of literary works have gained traction among translation scholars. Seljan et al. (2020) underscored the effectiveness of machine poetry translation for low-resource languages. Dai et al. (2022) discussed the limitations of machine translation for poetry, citing its struggle to convey the beauty of ancient Chinese traditional culture while noting the occasional enrichment of language through figures of speech. Humblé (2019) found Google Translate's translations of E. Dickinson's poems into Portuguese to be surprisingly satisfactory. In 2023, Alowedi and Al-Ahdal (2023) conducted a linguistic analysis comparing Arabic poem translations, concluding that machine translation falls short in capturing nuances and cultural context, advocating for machine-assisted translation with post-editing as a cost-effective solution. To our knowledge, the realm of GPT-3.5's (OpenAI 2023) poetry translation remains so far unexplored in the framework of a comparative human vs. machine translation analysis.

3. Research

3.1. Research problem and hypothesis

While the translation of most textual forms has benefited from AI-driven enhancements, the question remains if poetry translation is one of such domains, given its constraints of structural form, semantic depth, and emotional tenor. With poetry's deep reliance on nuance, figurative language, meter, rhyme, and cultural context, there is a distinct need to enhance machine translation methods to handle poetry specifically. This article seeks to investigate the proficiency and nuances of AI-generated poetry translations against their human counterparts, striving to understand both the potential and constraints of AI in capturing poetic artistry. The problem under investigation is the current capabilities and anticipated outcomes of advancements in AI-driven translation, with a specific focus on ensuring the preservation of both form and meaning in poetic translations from low-resource languages such as Ukrainian. Our hypothesis posits that AI-enabled translation can yield poetic renditions that align closely with human-generated translations in terms of translation quality. This encompasses the preservation of intrinsic poetic features, structural fidelity, semantic accuracy, and emotive consistency.

3.2. Research materials and methodology

In this study, we compare twelve AI-translated poems from Ivan Franko's (2021) collection of poems "Faded Leaves" to those translated by an Anglo-American translator, Percival Cundy (Franko et al. 1948). Franko, a pivotal figure in Ukrainian poetry, offers a robust basis for examining AI's capabilities in translating poetic texts. Utilizing the selected poems, amounting to 253 lines, we juxtaposed AI's target texts against Cundy's translations, first published between 1929 and 1931. The source texts comprised 1080 words, which is fewer than both the Cundy translations at 1426 words and the AI translations at 1428 words, but, at the same time, the translators retained all the twelve source-text lines. Intriguingly, the word counts of the human and machine translations are almost identical, a result that was quite unexpected.

For a comparative translation analysis, we got GPT-3.5 LLM AI to generate translations of the selected poems from Ukrainian to English, with prompts to the AI model instructing it to preserve the original rhyme scheme, meter, and verse length wherever feasible. The resultant AI translations were then compared with Percival Cundy's translations according to 8 criteria to better capture both form and meaning as salient for poetic translation and then preprocessed human and AI translations to make them suitable for language processing, which involved removing any formatting and ensuring the text was in the appropriate UTF-8 character encoding. We then tokenized the texts, making them ready for language processing tools. The translations were assessed based on a combination of quantitative and qualitative equivalence measures, using the metrics put forward by Dastjerdi et al. (2011) and considerations in Boase-Beier (2013) to shape these criteria. They encompass lexical density and diversity, part-of-speech (POS) composition, adjective-to-verb ratio, rhyme scheme and meter, tropes and figures of speech, as well as emotion and tone.

In concordance with the findings of Kulchytskyi et al. (2018), we posit that quantitative metrics serve as reliable indicators for establishing correspondences in translation. Such metrics hold significant promise for contrasting human and machine translation methodologies. Several important poetry criteria were selected to cover both the formal linguistic features and the artistic aspects of the poems, allowing for a comprehensive assessment of the translations. By utilizing these diverse criteria and tools, we aimed to capture the complex interplay of linguistic and poetic elements in both AI-generated and human-produced translations. The aforementioned metrics were analyzed through a variety of tools:

The quantitative metrics were analyzed with the help of a variety of tools:

1) Similarity of the translations was assessed through establishing the BLEU score utilizing the Natural Language Toolkit (NLTK) library (Bird et al. 2009);
2) Lexical density and diversity, POS composition, and adjective-to-verb ratio were established through quantitative calculations, where lexical density was determined by dividing the number of lexical words by the total number of words in a poem and lexical diversity was calculated by dividing the number of unique lexical words by the total number of lexical words in the poem; the POS composition was derived by tokenizing and tagging words through Natural Language Toolkit (NLTK) and calculating the proportions of each category; adjective-to-verb quotient was computed by dividing the number of adjectives by the number of verbs;
3) The Zeuscansion tool (Agirrezabal et al. 2016) was used to determine the rhythmic pattern (meter) of a poem.

The qualitative metrics, such as tropes and figures of speech, translation inaccuracies and rhyme schemes were directly observed and identified by the authors of this article.

4. Results

First and foremost, both human-produced and AI-generated translations adhere to the formal criteria of what is typically recognized as a poem, namely, structured verse, rhythmic and rhyming patterns, cohesive thematic elements, and the intentional use of stylistic devices such as metaphors, similes, and imagery to convey deeper meanings and evoke emotional responses. To ascertain that the AI language model did not source chunks of existing translations of Franko's work from web-scraping, we subjected both human and AI-generated translations to the Bilingual Evaluation Understudy (BLEU) metric (Interactive BLEU score evaluator; n. d.). The derived BLEU scores yielded a mean value of 3.72 (**Figure 1**), suggesting a level of originality in the translations produced.

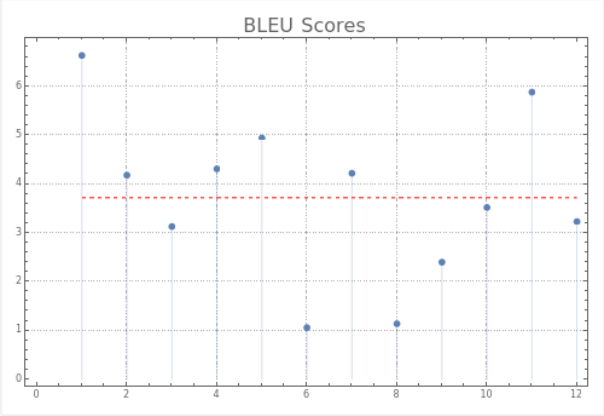

Figure 1. Mean of BLEU scores of the human (reference) to AI-generated (target) translations.

4.1. Poetic form

4.1.1. Meter

Franko's poetry exhibits no strict adherence to a singular meter, drawing inspiration from the stress patterns of Ukrainian folk songs. These songs traditionally utilize a diverse array of meters, defying easy categorization. We have relied on the analysis of Franko's poetry conducted by Bunchuk (2009) for the reference values of the Ukrainian original and the Zeuscansion tool's results for English translations (both human and AI-generated). Franko's work often gravitates towards various manifestations of trochee — a meter prevalent in Ukrainian poetry, as outlined in the "Literary Encyclopedia" (Kovaliv 2007). Particularly, the trochaic tetrameter is recognized as closely tied to the verse length of folk songs (Johansen 1922). and is a metric structure frequently employed by Franko.

The translations, both human-produced by Cundy and AI-generated by GPT-3.5, demonstrate distinctive metric preferences. Cundy primarily oscillates between trochaic and iambic trimeters, deftly interchanging between the trochee and iamb's varied feet, with singular instances of both molossus dimeter and antibacchic trimeter. In contrast, GPT-3.5 exhibits an inclination for the iambic pentameter—a meter deeply entrenched in English poetry and resonant with balladic forms, analogous to Franko's style of folk songs in the original. Still, the AI model displays a balanced usage of trochaic, spondic, and bacchic meters, punctuated by occasional molossus trimeter and antibacchic dimeter employments—paralleling Cundy though differing in specific translation alignments.

In the task of adhering to Franko's metrical style, both the human translator and the AI language model encountered challenges. Predominantly, this arises from the inherent variability in Franko's poetic structure as both the human and AI translations display greater metrical consistency, which

unsurprisingly results in a discrepancy when juxtaposed with Franko's original rhythm. Interestingly, the human and AI translations bear a closer metrical resemblance to each other than to Franko's original, suggesting a potential alignment with established English poetic conventions rather than following in the 'metric' footsteps of the original. Cundy's translations manage to achieve two instances of exact concordance in both stress pattern and metrical foot with Franko, whereas GPT-3.5 secures no full matches and only one partial alignment regarding stress pattern. In terms of matching the metrical foot of Franko's original, Cundy does so four times, outperforming GPT-3.5, which achieves this thrice. This narrow margin suggests that the human translator was able to align more closely with Franko's work in terms of both meter type and length.

4.1.2. Rhyming scheme

In selected Franko's works, the predominant use of the alternate rhyming scheme can be observed, consistent with the most popular schemes in Ukrainian poetry (Hromyak et al. 1997). This scheme, akin to the structure of the traditional English sonnet, is not only widespread but is also among the simplest, lending itself to potential ease of replication in translation.

Interestingly, the results are a mixed bag (**Figure 2**). The AI model, in its translations, closely adhered to Franko's alternate rhyming scheme, diverging only in two instances. In contrast, Cundy demonstrates a marked inclination towards the ABCB pattern, implementing it in two-thirds of translations regardless of the pattern in the original. This deviation is intriguing, given the accessibility and familiarity of the alternate rhyming scheme in both Ukrainian and English poetry. Another salient observation is the performance of both when confronted with irregular and more complex rhyming patterns. While both Cundy and GPT-3.5 struggled to accurately replicate an unsystematic rhyming scheme in "Poludne", they exhibited remarkable fidelity in rendering intricate patterns like ABABABCC ("Hoch ti ne budesh kvitkoyu…") and ABABABCB ("Poklin tobi, Buddo!"). Statistically, GPT-3.5 displayed greater alignment, accurately mirroring the rhyming patterns of eight out of the twelve original pieces (66.67%); conversely, Cundy's translations align in only 5 instances (41.67%), which suggests that the AI language model outperformed the human translator in this respect.

Figure 2. Rhyming pattern matches in human (Cundy) and AI-generated (GPT-3.5) translations.

4.1.3. Linguistic complexity

Drawing inspiration from Simonton's (1990) insights on the importance of intra-textual variance as a potent metric for gauging poetic allure beyond mere structural intricacy, our research sought to investigate how this direct positive correlation with poetic aesthetics is reflected in the quality of translations of poetry. We examined key metrics indicative of such complexity: lexical density, lexical diversity, and the adjective-verb ratio to discern patterns and distinctions in the linguistic architecture of both human and AI-generated translations.

Lexical density

In our analysis, the following observations emerged regarding lexical densities within the translations (**Figure 3**). Cundy's translations presented a lexical density spectrum from a low of 35% in

"Oj, zhalyu mij, zhalyu…" to a peak of 57% in both "Chervona kalino…" and "Poklin tobi, Buddo!". Similarly, the AI-driven translations by GPT-3.5 exhibited a range from 31% in "Oj, zhalyu mij, zhalyu…" to 56% in "Yak vil v yarmi…". Both the human and AI registered minimum lexical densities in "Oj, zhalyu mij, zhalyu…", implying a possible resonance in this specific text. Yet, no uniform pattern emerged across the entire dataset; in almost half of the instances, each showcased a heightened lexical density, evidencing the variability in translation approaches. Such fluctuating proximities are exemplified in "Tvoyi ochi yak te more…", "Ya ne zhaluyus na tebe…", and "Poludne", where lexical densities between Cundy and GPT-3.5 converge closely; conversely, disparities like that in "Chervona kalino…"—with a 14% divergence between Cundy's 57% and GPT-3.5's 43%—highlight the unpredictable nature of translational linguistic choices. Both Cundy and GPT-3.5 exhibit significant variability in lexical density across the poems, indicating that neither maintains a strong advantage over the other regarding lexical richness. However, Cundy does exhibit a slight edge, demonstrating higher lexical density in 58.33% of the instances. On average, Cundy's translations possess a lexical density of 49%, marginally surpassing GPT-3.5's mean of 48%. In our analysis, the following observations emerged regarding lexical densities within the translations (**Figure 3**). Cundy's translations presented a lexical density spectrum from a low of 35% in "Oj, zhalyu mij, zhalyu…" to a peak of 57% in both "Chervona kalino…" and "Poklin tobi, Buddo!". Similarly, the AI-driven translations by GPT-3.5 exhibited a range from 31% in "Oj, zhalyu mij, zhalyu…" to 56% in "Yak vil v yarmi…". Both the human and AI registered minimum lexical densities in "Oj, zhalyu mij, zhalyu…", implying a possible resonance in this specific text. Yet, no uniform pattern emerged across the entire dataset; in almost half of the instances, each showcased a heightened lexical density, evidencing the variability in translation approaches. Such fluctuating proximities are exemplified in "Tvoyi ochi yak te more…", "Ya ne zhaluyus na tebe…", and "Poludne", where lexical densities between Cundy and GPT-3.5 converge closely; conversely, disparities like that in "Chervona kalino…"—with a 14% divergence between Cundy's 57% and GPT-3.5's 43%—highlight the unpredictable nature of translational linguistic choices. Both Cundy and GPT-3.5 exhibit significant variability in lexical density across the poems, indicating that neither maintains a strong advantage over the other regarding lexical richness. However, Cundy does exhibit a slight edge, demonstrating higher lexical density in 58.33% of the instances. On average, Cundy's translations possess a lexical density of 49%, marginally surpassing GPT-3.5's mean of 48%.

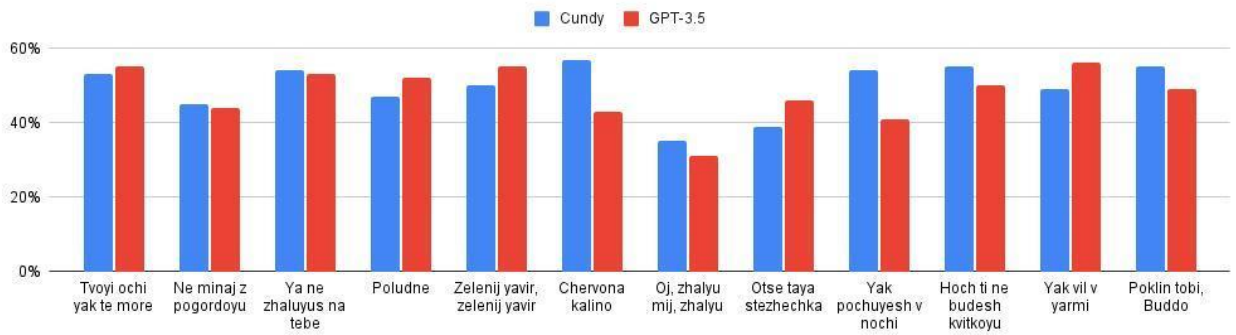

Figure 3. Lexical density in the human (Cundy) and AI-generated (GPT-3.5) translations.

Lexical diversity

We discerned salient patterns of lexical diversity within the translations, as illustrated in **Figure 4**. Both Cundy and GPT-3.5 present translations that stretch over a broad spectrum of lexical densities. These span from 49% for Cundy and 44% for GPT-3.5 at the lower end, scaling to peak values of 81%

and 85% respectively. A substantial majority of translations by both surpass a lexical density threshold of 60%, underscoring the rich vocabulary diversity they bring into their respective translations. While certain parallels exist—for instance, both achieve an identical lexical density of 69% in "Poludnie" and exhibit their lowest densities in "Chervona kalino…" (with Cundy at 49% and GPT-3.5 at 44%)—there is no dramatic superiority of one over the other. However still, GPT-3.5 displays a marginal advantage in terms of lexical density: it exhibits more pronounced increases in density compared to the human counterpart (with two instances where the difference exceeds 10%), is more lexically diverse in two-thirds of translations, and has a marginally elevated mean of 71%, juxtaposed against Cundy's average of 69%.

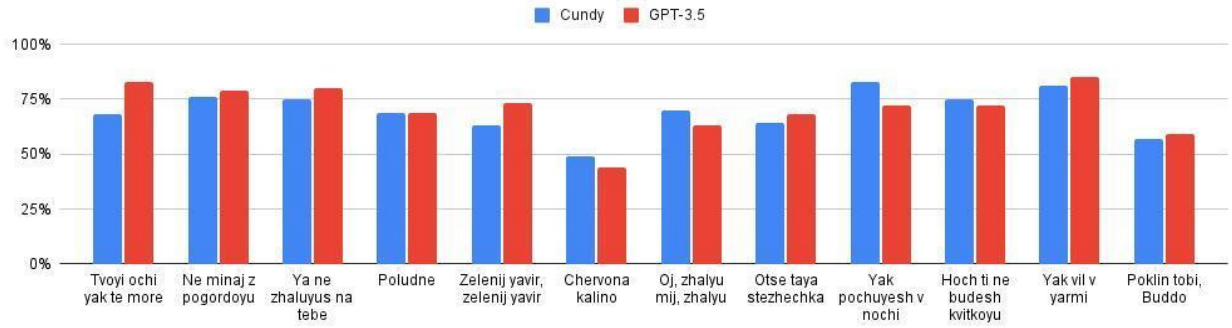

Figure 4. Lexical diversity in the human (Cundy) and AI-generated (GPT-3.5) translations.

Adjective-verb quotient

To derive the adjective-to-verb quotient, we conducted part-of-speech tagging on the tokenized texts from both the human translator and the AI language model. In a cursive examination of the general part-of-speech composition, we observed that the translations produced by both human and AI sources displayed a significant degree of alignment. However, GPT exhibited a slight preference for nouns and adjectives, suggesting a more descriptive style. In contrast, Cundy's translations contained more pronouns and were more verb-dense, inclusive of modal verbs, and coordinate conjunctions indicating a possibly more dynamic and personal translation style. The similarities in adverb, preposition, and determiner usage suggest that both translations adhere to similar structural norms in these areas.

Subsequently, we computed the adjective-to-verb quotient by dividing the aggregate adjective count by the total verb count as seen in **Table 1**. Although neither Cundy nor GPT-3.5 attains a Shakespearian caliber in this metric, they near that of English language poetry on average, with their respective quotients exhibiting remarkable comparability. Based on this particular metric, the AI language model's translations exhibit a marginally elevated adjective-verb quotient in comparison to human translations.

Table 1. Adjective-verb quotient (AVQ) in human (Cundy) and AI-generated (GPT-3.5) translations.

	Cundy	**GPT-3.5**
Adjective count	123	127
Verb count	252	245
Adjective-verb quotient	0.488	0.518

4.2. Poetic imagery

4.2.1. Creativity

Quantifying creativity has long been a topic of academic inquiry, with multiple metrics proposed to address this challenge. For the purpose of our study, we elected to employ the metric of forward flow, a

concept rooted in latent semantic analysis (LSA), which quantifies the average semantic distance between a given thought and its preceding thoughts. Elevated forward flow values signify an innovative, divergent thought process, while lower values indicate cyclical or repetitive thought patterns. This relationship between forward flow and creativity has been empirically validated in various research endeavors. Notably, Jacobs and Kinder (2022) conducted a seminal analysis of an extensive literary corpus, revealing that poets consistently outperformed other English writers in forward flow measurements, hence this metric should lend itself well to measuring the creativity of poetry translations into English. For our analysis, translations were systematically segmented into discrete lines, which were subsequently subjected to the forward flow analytical framework available on a specialized online platform (http://www.forwardflow.org/).

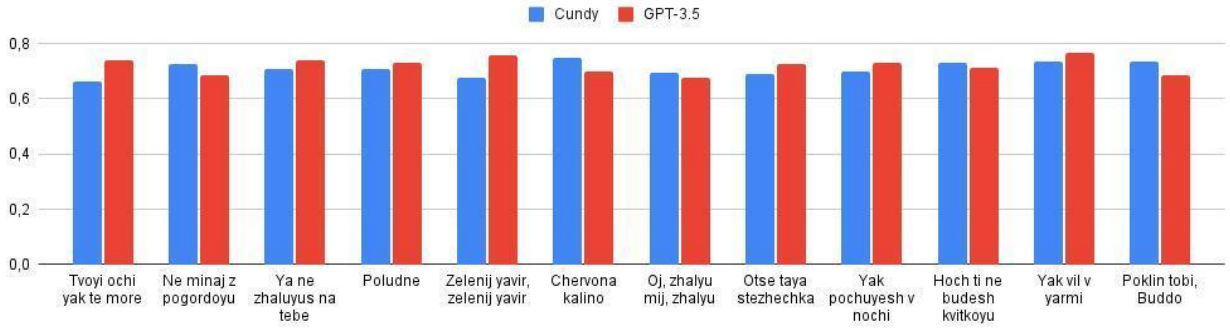

Figure 5. Average semantic distance (forward flow) in the human (Cundy) and AI-generated (GPT-3.5) translations.

The forward flow metrics derived from both Cundy's and GPT-3.5's translations (**Figure 5**) markedly differ from the values observed in original English poetry as presented in Jacobs (2018). In their study, poets consistently registered forward flow means between 0.8 and 0.9, while both human and AI-generated translation in our study fall short of 0.8. It is crucial, however, to contextualize these disparities by considering the inherent complexities of translation. Unlike original composition, which already presents its set of challenges, translation is governed by the imperative fidelity to the source text, which can often limit the scope for creative divergence.

Understanding these challenges and constraints, we have derived the following distinctions from the average semantic distances metric. GPT-3.5 registered an average forward flow of 0.721, marginally surpassing Cundy's mean of 0.710. This higher mean, coupled with the AI outperforming Cundy in 58.33% of instances, might hint at a slight edge in the creative propensity for GPT-3.5. Notably, GPT-3.5 also exhibited both the highest (0.766) and lowest (0.676) values among the two. Yet, as can be observed from the data, the performance spread between the two is relatively narrow and this relative parity underscores the AI model's performance when set against the benchmarks established by the human translator.

4.2.2. Stylistic devices

Upon direct observation, we identified various stylistic devices in both the source and translated texts. For instance, in Franko's work, we found epithets such as "сміх твій нинішній, Срібний та дзвінкий", translated by Cundy as "with silvery laugh" and by GPT-3.5 as "laughter, so silver and sweet". Similes were evident in phrases like "Мов пилинка в них тоне́", with Cundy's translation reading "Like a speck, sinks out of sight" and GPT-3.5's as "Like dust, they vanish out of sight". Metaphors in Franko's "Що з мойого сердечка щастя унесла; Гіркий не помалу" were rendered by Cundy as "Who took from out my bosom; Its joyous, happy song; I ache with bitter pain" and by GPT-3.5 as "From my own heart, so tenderly", among others.

As seen in **Table 2**, both versions consistently prioritize the use of epithets and metaphors, suggesting a shared focus on creating vivid imagery and layered meanings in the poems. Similes are less favored in both versions, exhibiting similar preferences for stylistic devices. A high correlation coefficient between the stylistic devices used by Cundy and GPT-3.5 indicates a synchronized tendency: when one employs a device frequently, so does the other. The standard deviations point towards a fairly consistent stylistic approach across the poems in translations by both human and AI.

Table 2. Stylistic choices mean in the human (Cundy) and AI-generated (GPT-3.5) translations.

	Cundy (mean)	**GPT-3.5 (mean)**
Epithet	8.42	10.8
Metaphor	7	8.91
Simile	1.58	2

Both Cundy and GPT-3.5 exhibit comparable tendencies in their translations, yet GPT-3.5 frequently employs epithets more than Cundy, as evidenced in **Figure 6**. This inclination corresponds with GPT-3.5's slightly elevated adjective-to-verb quotient. Despite these differences, the uniform use of epithets by both translators underscores their aptitude for preserving the original poems' thematic core. However, substantial standard deviations in epithet counts (6.35 for Cundy and 6.76 for GPT-3.5) highlight variances across individual poems, suggesting that epithet deployment can significantly fluctuate. Further, a positive correlation (0.89) between the two datasets reveals a congruence: a high epithet count in one often mirrors a high count in the other.

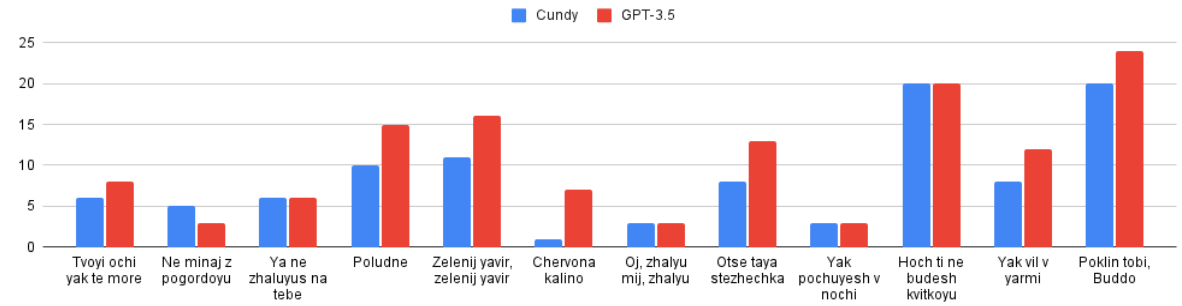

Figure 6. Epithet use in the human (Cundy) and AI-generated (GPT-3.5) translations.

GPT-3.5 generally employs more metaphors than Cundy, as evident in **Figure 7**, although this distinction is not as consistent. For metaphors, significant standard deviations (5.16 for Cundy and 4.89 for GPT-3.5) underline variability in their usage across both the human and AI, with Cundy displaying a marginally broader range.

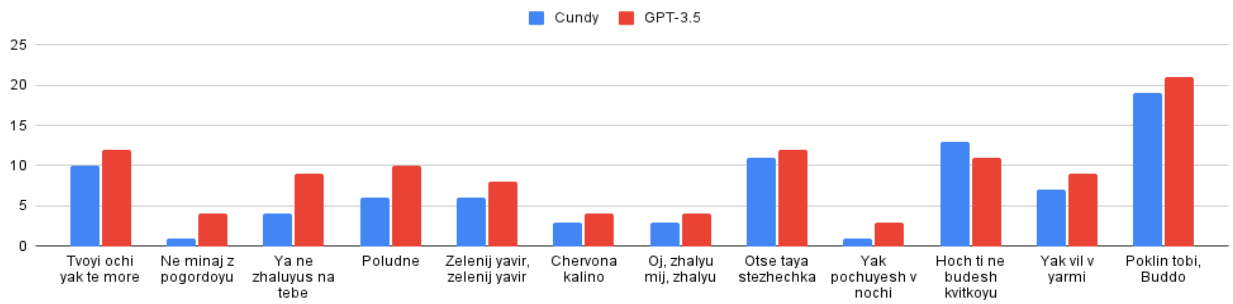

Figure 7. Metaphor use in the human (Cundy) and AI-generated (GPT-3.5) translations.

Meanwhile, the usage of similes remains largely similar across both datasets, as depicted in **Figure 8**, without a discernible trend favoring one over the other. The standard deviations for simile counts are narrower (1.12 for Cundy and 1.15 for GPT-3.5), indicating greater consistency in their application than metaphors. While strong correlations for both metaphors (0.94) and similes (0.95) indicate that, regardless of count differences, the two datasets exhibit analogous rank order concerning these stylistic devices across the texts.

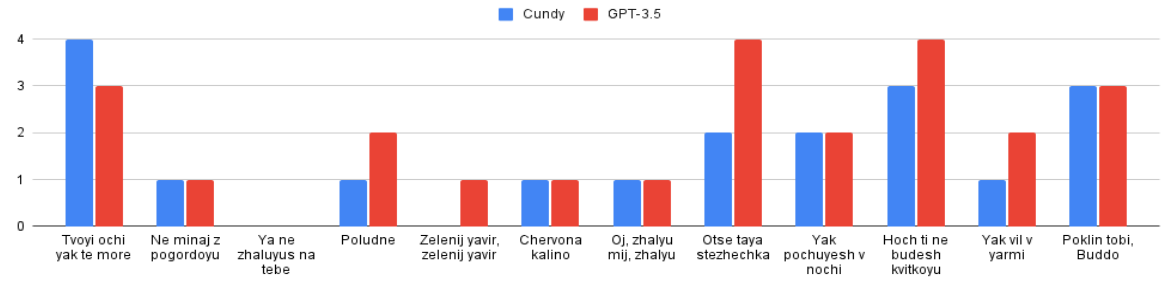

Figure 8. Simile use in the human (Cundy) and AI-generated (GPT-3.5) translations.

The close correlations in stylistic device usage across certain translations may indicate that some poems lend themselves to more literal translations as exemplified through "Oj, zhalyu mij, zhalyu….", while others, like "Otse taya stezhechka…" allow (or require) more creative liberties. Although not strictly an indicator of fidelity—given that a metaphor in the original Ukrainian might be rendered as a simile in English (or vice versa) while retaining analogous imagery—these translation metrics explicate the stylistic capacities of both human and AI translators in poetry, with the balance between fidelity to the source text and recreative expression remaining a central challenge in poetry translation.

While the goal was to compare the linguistic properties of human and machine translation, an interesting observation came up that both translations have a higher usage of epithets and metaphors compared to the original, while the original Franko text has a slightly higher usage of similes compared to the translations. Despite the close resemblance among the three stylistic devices—epithets, metaphors, and similes—their total numbers in the translations significantly diverge from the original, as depicted in **Table 3**, except for similes, which might suggest that the explicit character of similes makes them more amenable to translation. On the other hand, the translations, especially those of the AI model, might amplify imagery to compensate for the nuanced, transient quality of the original's imagery. This could indicate the AI's strategy to address the subtlety of imagery through increased volume.

Table 3. Tropes in the original (Franko), human translations (Cundy), and the AI translations (GPT-3.5).

	Franko	**Cundy**	**GPT-3.5**
Epithet	62	101	130
Metaphor	81	84	107
Simile	25	19	24

Such observations led us to cross-check with the original text to discern if the augmented use of tropes—evident in translations—mirrored a similar trend in stylistic figures. Contrary to expectations, the original Franko text exhibited substantially greater use of alliteration and assonance compared to both translations, as can be seen in **Table 4**. Notably, the two translations displayed comparable counts of these stylistic figures, suggesting a similar translation strategy concerning these devices. The significant reduction in the use of these devices in the translations could be due to the challenges of preserving such

stylistic elements, in particular. This could be attributed to the inherent characteristics of the English language, which might not accommodate assonance or alliteration as seamlessly as Ukrainian might. Alternatively, both human and AI translators might have deemed the verse's musicality less vital in English. It is worth noting that in Ukrainian poetry, such sound patterns are as prevalent and pivotal as tropes. While the impact of stylistic device dominance on translation fidelity remains an open question, our primary objective was to juxtapose the target texts of both human and AI translations, which proved to be quite comparable.

Table 4. Figures in the original (Franko), human translations (Cundy), and the AI translations (GPT-3.5).

	Franko	**Cundy**	**GPT-3.5**
Alliteration	145	27	28
Assonance	166	21	21

4.3. Accuracy of translation

The main problematic areas in terms of accuracy in the translations of Ukrainian poems using the language model are as follows.

4.3.1. Inaccuracies in AI-driven translations

Grammatical complexities

In "Otse taya stezhechka…", GPT misinterprets the sentence "Обливав з сльозами я Пил із її ніг…" due to verse structure (in particular, capitalization) an inverted predicate, resulting in the inaccurate translation "With tears, my cheeks were traced, As dust offered no relief" or in "Poklin tobi, Buddo" fails to observe a causal structure and translates "Безсмертне лиш тіло, Бо жаден атом Його не пропаде…" as "The body may perish, But atoms won't cease…".

Cultural contexts

The phrase "Щоб запалася!" in "Otse taya stezhechka…" misses the traditional Ukrainian curse and instead translates it as the misspelled "запалалась" - "glow". Not only does it show that a significant layer of Ukrainian cultural context is missing in its training data, but it also attests to GPT's tendency to treat unfamiliar items as misspellings and, instead of omitting, overtranslate. Much like in "Я не жалуюсь на тебе, доле…" we observe misidentification of dialect "отінив" (overshadow) potential as "одтіснив" (repel) which also is dialectal but connected to Russian "оттеснил" can be indicative either of cross-contamination of Ukrainian and Russian training data.

Dialectal forms

Words like "хоре" ("хворе" in standard Ukrainian) which remained untranslated in "Poludne" and the dual meaning of "сімя" (both "seed" and "family" in Ukrainian) in "Я не жалуюсь на тебе, доле…" where the wrong word was translated exemplify GPT's struggle with Ukrainian dialectal and non-standard words.

Botanical terms

Words like "левкоя" in "Hoch ti ne budesh kvitkoyu…" and "явір" in "Zelenij yavir, zelenij yavir…" were translated as "lily' and "fir' correspondingly, instead of their equivalents 'gillyflower' and 'plane tree', which might be indicative of either of the following: GPT could simplify terminology that doesn't fit into the poetic register or opts for shorter words for the sake of adhering to the form of the poem better.

4.3.2. Inaccuracies in human translations

Shifts in style

The translator frequently interprets dialectal Ukrainian in archaic English forms, leading to abundant uses of 'thou' and 'thee'. This imparts an unintended antiquated and elevated tone to the translations.

Shifts in meaning

There are instances where meanings undergo alterations as in "Otse taya stezhechka…", the phrase "Я вагувася" is translated as quite the opposite "I never gave it though" instead of the more accurate "hesitated". Or in "Poludne" where "І сверщики в травах тріщать" becomes "And grasshoppers flit through its blades" instead of the closer "chirp in the grass", extending the meaning from literal noise to the noise of them moving to simply an act of movement. The language of the human translator is overly ornate, thus overcomplicating simple lines of the original.

Omission and addition of detail

In "Зелений явір, зелений явір," the phrase "Моргають серед ночі" loses the mention of "night" in "Blink in the summer sky", while the clause "Till I doze" in "Poludne" is an addition not present in the original. Such changes affect the mood created through the verse thus co-authorship of the translator comes into the picture.

Mistakes by GPT are often technical and can be refined with model iterations and editing (the latter presumably already present in the human translations)—areas like enhanced cultural and dialectal data training might offer improvements. In contrast, the human translator's errors stem from personal interpretation and stylistic choices. This individual touch, whether aligned with the original author's voice or not, provides a distinctly human quality to the translation, infusing it with personal nuance and imagery. While GPT's translations might lack this inherent personality, the human translator's rendition, albeit sometimes divergent from the source, offers a consistent and personal voice.

5. Discussion

Machine translations, powered by evolving AI technologies have made substantial progress in tackling the multifaceted challenge of poetry translation, which involves structural, linguistic, poetic, semantic, and emotional elements. Central to our investigation was the capability of AI to compete with human expertise in translating poetry, particularly for resource-limited languages like Ukrainian. Our hypothesis posited that AI could potentially match human translations in preserving poetry's form, meaning, and emotional depth.

In many cases, the disparity between AI and human translations has become increasingly narrow. Our data confirms AI's ability to retain inherent poetic elements like verse structure, rhythm, and thematic unity, suggesting potential applications in accurately representing cultural poetry. GPT-3.5's translations, however, adhered more closely to specific rhyme schemes, highlighting advancements in AI's capacity for poetic translation. AI demonstrated a particular aptitude for retaining poetic forms and thematic consistency. However, both modalities encountered challenges in maintaining meter and rhyme. Which is consistent with a broader context in the translation of poetry, deeply rooted in Ukrainian folk traditions like that of Franko.

Linguistic complexities, assessed via metrics like lexical density and diversity, showed a balance between linguistic intricacy and artistic expression in both translation methods. AI tended towards a

descriptive style, whereas human translations were more verb-centric. Notably, the AI system showed a slight advantage in the creativity metric, while both exhibited proficiency in vital stylistic techniques.

Accuracy evaluations highlighted challenges for both modalities. AI faced difficulties with grammatical nuances and cultural contexts, while human translations occasionally deviated in meaning or style due to personal interpretation. Such deviations in human translations could arise from individual perspectives, whereas AI's limitations could be attributed to algorithmic restrictions and potential biases in training data.

Upon criteria-based analysis, GPT-3.5 equals or surpasses human translations in six of the eight major parameters. This parity in performance supports our hypothesis of AI's potential in poetry translation, emphasizing its ability to capture poetic form and depth.

A notable strength of AI systems is their consistent output, uninfluenced by external factors. However, such consistency occasionally yielded rigid translations. In terms of cultural sensitivity, human translations exhibited a profound understanding of cultural nuances and contexts, emphasizing the importance of embedding AI systems with comprehensive contextual knowledge.

In conclusion, GPT-3.5's translations of Ukrainian poetry from the late 19th century demonstrated a caliber comparable to human efforts. Future research could endeavor advancements in subsequent AI iterations to gain deeper insights into the evolving capabilities of AI in the realm of poetry translation for low-resource languages.

Author contributions

Conceptualization, VK and AK; methodology, VK and AK; validation and analysis, VK and AK; resources, VK and AK; writing—original draft preparation, review and editing, VK and AK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

Agirrezabal M, Astigarraga A, Arrieta B, et al. (2016). ZeuScansion: A tool for scansion of English poetry. Journal of Language Modelling. 2016, 4(1). doi: 10.15398/jlm.v4i1.102

Alowedi N, Al-Ahdal A (2023). Artificial Intelligence based Arabic-to-English machine versus human translation of poetry: An analytical study of outcomes. Journal of Namibian Studies : History Politics Culture. 2023, 33. doi: 10.59670/jns.v33i.800

Badura M, Lampert M, Dreżewski R (2022). System Supporting Poetry Generation Using Text Generation and Style Transfer Methods. Procedia Computer Science. 2022, 207: 3310-3319. doi: 10.1016/j.procs.2022.09.389

Bird S, Klein E, Loper E (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Boase-Beier J (2013). Poetry Translation. In: Baker M, Saldanha G (editors). *The Routledge handbook of translation studies*. Routledge. pp. 475-487

Boulenouar M (2022). A Comparative Study of Machine and Human Translation: The Case of English-Arabic Literary Translations. *Journal of Language and Linguistic Studies*, 18 (Special Issue 1): 176-191.

Bunchuk B (2009). The innovation of Ivan Franko's verse form as a sign of the modernist character of his poetry. In: *The Bible and Culture* 11: 112-117.

Chakrabarty T, Saakyan A, Muresan S (2021). Don't Go Far Off: An Empirical Study on Neural Poetry Translation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Published online 2021. doi: 10.18653/v1/2021.emnlp-main.577

Dai J, Shen H, et al. (2022). The Differences Between Machine Translation and Human Translation from the

Perspective of Literary Texts. *International Journal of Arts and Social Science*. 5(10): 112-133.
Dastjerdi HV, Khosravani Y, Shokrollahi M, et al. (2011). Translation Quality Assessment (TQA): A Semiotic Model for Poetry Translation. Lebende Sprachen. 2011, 56(2). doi: 10.1515/les.2011.021
Interactive BLEU score evaluator (n. d.) *Tilde MT.* Available from: https://www.letsmt.eu/Bleu.aspx
Franko I (2021). *Faded leaves*. Litopys.
Franko I, Cundy P, Manning CA (1948). *Ivan Franko, the Poet of Western Ukraine.* Philosophical Library.
Genzel D, Uszkoreit J, Och F (2010). "Poetic" statistical machine translation: rhyme and meter. *EMNLP*. Available from: https://research.google/pubs/pub36745/
Ghazvininejad M, Choi Y, Knight K (2018). Neural Poetry Translation. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Published online 2018. doi: 10.18653/v1/n18-2011
Grace K, Salvatier J, Dafoe A, et al. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research. 2018, 62: 729-754. doi: 10.1613/jair.1.11222
Hadley J (2020). Literary Machine Translation. *Counterpoint*. (4): 15-18. Available from: https://www.ceatl.eu/wp-content/uploads/2020/12/Counterpoint_2020_04_article_04.pdf
Humblé P (2019). Machine translation and poetry. The case of English and Portuguese. *Ilha do Desterro*, 72: 41-56. doi: 10.5007/2175-8026.2019v72n2p41
Hromyak RT, Kovaliv YI, Teremko VI (1997). *Dictionary of Literary Studies*. Akademia. p. 331.
Jacobs AM (2018). The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses. Frontiers in Digital Humanities. 2018, 5. doi: 10.3389/fdigh.2018.00005
Jacobs AM, Kinder A (2022). Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large Corpus of English Literature. arXiv:2201.04356.
Johansen M (1922). *Elementary laws of versification*. Vseukrlitkom.
Kenny D, Winters M (2020). Machine translation, ethics and the literary translator's voice. Translation Spaces. 2020, 9(1): 123-149. doi: 10.1075/ts.00024.ken
Kovaliv Y (2007). *Literary encyclopedia*. Kyiv: Academia Publishing House.
Kulchytskyi I, Tsiokh L, Malaniuk M (2018). Quantitative Equivalence Level in Poetry Translation. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). Published online September 2018. doi: 10.1109/stc-csit.2018.8526715
Kuzman T, Vintar Š, Arcan M (2019). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the qualities of literary machine translation* (pp. 1-9). Available from: https://aclanthology.org/W19-7301.pdf
Ma Y, Wang B. Description and Quality Assessment of Poetry Translation: Application of a Linguistic Model. Contrastive Pragmatics. 2020, 3(1): 89-111. doi: 10.1163/26660393-bja10015
Matusov E (2019). The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation* (pp. 10-19). Available From: https://aclanthology.org/W19-7302.pdf
OpenAI (2023). *ChatGPT* (Mar 14 version) [Large language model]. Available From: https://chat.openai.com/chat
Poibeau T (2022). On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation. In *Language Resource and Evaluation Conference.* June 2022, Marseille, France.
Seljan S, Dunder I, Pavlovski M (2020). Human Quality Evaluation of Machine-Translated Poetry. 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO). Published online September 28, 2020. doi: 10.23919/mipro48935.2020.9245436
Simonton DK (1990). Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. Computers and the Humanities. 1990, 24(4): 251-264. doi: 10.1007/bf00123412
Studzińska J (2020). Turing test for (automatic) translation of poetry (Polish). Porównania. 2020, 26: 299-313. doi: 10.14746/por.2020.1.17
Toral A, Way A (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? Translation Quality Assessment. Published online 2018: 263-287. doi: 10.1007/978-3-319-91241-7_12
Vincent R (2019). *Multilingual Poetry Generation* [Master's thesis]. NTNU (Norwegian University of Science and Technology).
Visby M (2020). The future relationship of literary translation and AI: Reflections from CEATL president. *Counterpoint*. 4: 28-31.
Zong Z (2018). Research on the Relations Between Machine Translation and Human Translation. Journal of Physics: Conference Series. 2018, 1087: 062046. doi: 10.1088/1742-6596/1087/6/062046