University of
Sistan and Baluchestan

Iranian Journal of Applied Language Studies

Print ISSN: 2008-5494 Online ISSN: 2322-3650

Homepage: https://ijals.usb.ac.ir

Uppsala University

Google Translate and Microsoft Bing Translator's Challenges in Rendering Camus's *The Stranger* from English to Persian

Niloufar Amiri[1] , Ali Beikian[2]

[1]Department of English Language, Chabahar Maritime University, Chabahar, Iran, Email: tedbeheshti@gmail.com

[2]*Corresponding author*, Assistant Professor, Department of English Language, Chabahar Maritime University, Chabahar, Iran, Email: a_beikian@yahoo.co.uk

Abstract

Machine translation (MT) of literary texts presents unique challenges due to their stylistic complexity and cultural nuances. This study evaluated the performance of Google Translate (GT) and Microsoft Bing Translator (MBT) in translating Camus's *The Stranger* from English to Persian. Data collection for this study involved automated evaluation using the Bilingual Evaluation Understudy (BLEU) metric and human evaluation conducted by three experts using the Localization Industry Standards Association (LISA) rubric. The results showed that GT significantly outperformed MBT across nearly all dimensions. GT achieved a BLEU score of 21.57 compared to MBT's 6.36, with superior n-gram precision at all levels. The human evaluation phase also revealed GT's fewer critical and major errors in almost all categories compared to MBT. However, both systems struggled to preserve the aesthetic and philosophical richness of *The Stranger*. These findings highlight the persistent limitations of MT in literary translation, particularly for linguistically distant pairs like English and Persian. While MT shows potential as a supplementary tool, it remains unsuitable as a replacement for human translators in capturing the depth and artistry of literary works.

Keywords: Google Translate, literary translation, machine translation challenges, Microsoft Bing Translator, Persian translation

Received: July 19, 2024 Revised: November 20, 2024 Accepted: November 29, 2024 Published: January 10, 2025

Article type: Research Article DOI: 10.22111/ijals.2025.50602.2487

Publisher: University of Sistan and Baluchestan

How to cite: Amiri, N., & Beikian, A., (2025). Google translate and Microsoft Bing translator's challenges in rendering Camus's the stranger from English to Persian. *Iranian Journal of Applied Language Studies*, *17*(2), 101-120. https://doi.org/10.22111/ijals.2025.50602.2487

1. Introduction

Literary translation presents unique and multifaceted challenges due to its dual requirement for linguistic precision and cultural sensitivity. These challenges arise from the complex nature of literary texts, which transcend conventional grammatical constraints and demand careful attention to stylistic complexity, heterotopic elements, and artistic nuances (Baker, 2018; Cui, 2021; Delabastita, 2010, 2019; Katan & Taibi, 2021). Unlike other textual genres, literary works are crafted to evoke emotional responses and generate entertainment through imaginative language, poetic devices, and layered meanings (Cui, 2021; Wittman, 2013). This intricate interplay of linguistic and cultural elements necessitates systematic approaches to translation, as target readers expect adaptations that bridge temporal and cultural gaps while preserving content comprehension (Landers, 2001).

The linguistic aspects of literary translation further complicate the process, as they involve interactions across lexical, syntactic, and macrostructural levels. These interactions become particularly pronounced when translating between linguistically distant language pairs, such as English and Persian, where structural differences, such as the SVO configuration of English and the SOV structure of Persian, require systematic adjustments to maintain semantic equivalence and textual integrity (Baker, 2018; Hatim & Mason, 2014; Munday et al., 2022; Nord, 2014). Such challenges are exacerbated by the incorporation of figurative language, cultural references, and stylistic devices, which demand both linguistic competence and cultural understanding (Katan & Taibi, 2021; Snell-Hornby, 2006). As noted by Newmark (1988) and Vinay and Darbelnet (2000), the preservation of literary elements in translation is often hindered by structural and cultural disparities between source and target languages.

Despite significant advancements in machine translation (MT) technology, its application to literary translation remains fraught with limitations. MT systems, while effective for technical and formal texts, struggle to address the nuanced demands of literary translation, particularly in handling cultural subtleties, contextual understanding, and stylistic features (Naveen & Trojovský, 2024; Taivalkoski-Shilov, 2019; Toral & Way, 2015). These challenges are especially pronounced when translating linguistically distant pairs like English and Persian, where structural differences further impede the preservation of literary elements (Baker, 2018; Jones & Irvine, 2013; Toral & Way, 2015; Vinay & Darbelnet, 2000). Given these considerations, a critical evaluation of MT systems' performance in literary translation is essential to identify their strengths, weaknesses, and potential applications.

Guided by these insights, this study investigates the linguistic challenges faced by Google Translate (GT) and Microsoft Bing Translator (MBT) in translating *The Stranger* (Camus, 1989) from English to Persian. The analysis focuses on their comparative performance in terms of translation quality and readability, with particular attention to their ability to address cultural elements, contextual nuances, and stylistic complexities inherent in literary texts. By evaluating these aspects, the study aims to provide a comprehensive understanding of the capabilities and limitations of MT systems in handling the unique demands of literary translation. To achieve this objective, the study is guided by the following research questions:

1. What challenges does GT encounter when translating *The Stranger* from English to Persian?
2. What challenges does MBT face when translating *The Stranger* from English to Persian?
3. What are the differences between GT and MBT in terms of their challenges when translating *The Stranger* from English to Persian?

The paper is organized as follows: First, we address the difficulties of literary translation and the constraints faced by machine translation systems. Next, we detail the research design and data analysis methods used in the study. This is followed by a comprehensive presentation of the results, in which we evaluate and compare the performance of GT and MBT. Subsequently, we engage in a thorough discussion of the findings, emphasizing their significance and implications. Finally, we conclude by reflecting on the broader outcomes of the study and proposing directions for future research.

2. Challenges of MT Systems in Translating Literature

Literary texts demand a nuanced approach, as their stylistic, cultural, and aesthetic richness often surpass the capabilities of current MT systems. Voigt and Jurafsky (2012) identified referential cohesion as a significant barrier for MT in literary contexts. Their study revealed GT's limitations in handling cohesive devices effectively, with human translations consistently outperforming machine outputs in achieving coherence and fluency. These findings underscore the substantial challenges faced by MT systems in preserving the stylistic depth of literary texts.

Jones and Irvine (2013) extended this critique by demonstrating that MT systems frequently fail to capture idiomatic expressions and cultural subtleties. Their comparative analysis of GT and Moses highlighted systemic struggles with fidelity and stylistic complexity, exacerbated when translating between linguistically distant pairs. Such deficiencies, rooted in both linguistic and cultural constraints, complicate efforts to achieve high-quality literary translations through MT.

Adding depth to this discourse, Toral and Way (2015) emphasized the importance of linguistic proximity in MT performance. Their findings revealed that translations within closely related language families, such as French and Italian, were significantly more accurate and fluent than those between distant pairs like French and English. Kuzman et al. (2019) expanded on this by examining neural machine translation (NMT) systems tailored for morphologically complex languages like Slovenian. They found that specialized systems demonstrated improved performance, yet general-purpose MT tools, including GT, struggled with longer sentences and contextual integrity, further highlighting the inadequacy of generic MT for literary translation.

Although the evolution from statistical machine translation (SMT) to NMT marks a transformative advancement in computational linguistics, persistent challenges continue to impede the attainment of consistently reliable and contextually nuanced translations. Matusov (2019) observed notable improvements in lexical richness and syntactic cohesion with customized NMT systems for English-to-Russian and German-to-English literary translations. However, unresolved issues such as pronoun ambiguity and semantic inconsistencies remained. Similarly, Wang (2021) documented the struggles of NMT systems in addressing rhetorical and creative elements in English-to-Chinese translations. These findings collectively reinforce the argument that, while NMT represents a technological leap, it cannot fully replicate the interpretative and creative nuances demanded by literary texts.

This limitation is further evidenced by Guerberof-Arenas and Toral (2022), who demonstrated that human translation consistently outperformed both MT and post-editing (PE) in terms of creativity when translating Kurt Vonnegut's short story. They argued that MT's literal solutions constrained the creativity of post-editors, resulting in translations deemed unfit for publication. Thai et al. (2022) echoed these concerns, noting that literary translators overwhelmingly preferred human outputs over MT-generated paragraphs in the Par3 dataset. Their analysis identified discourse-disrupting errors and stylistic inconsistencies in MT outputs, further emphasizing the challenges in modeling and evaluating literary MT.

The phenomenon of "post-editese," where post-edited (PE) outputs mirror machine-translated texts more closely than human translations, adds another layer of complexity. Castilho and Resende (2022), using works like *Alice's Adventures in Wonderland*, found post-editese traits in PE versions of *The Girl on the Train*, undermining fidelity to the source text. These findings align with those of Guerberof-Arenas and Toral (2022), illustrating the constraints imposed by post-editing on achieving high-quality literary translations.

Further complicating this landscape, comparative evaluations reveal mixed results for different MT systems. Zahroh et al. (2023) found that GT excelled in active-passive transformations, while MBT performed better in pronoun resolution. However, both systems exhibited challenges with verb tense management, reflecting broader limitations in stylistic accuracy. Jibreel (2023) highlighted similar shortcomings in translating proverbs, where MT systems often defaulted to literal interpretations, distorting the intended meaning. MBT and GT were found to provide relatively better semantic equivalents, but errors in idiomatic expression translation persisted.

Efforts to refine MT systems for literary applications have shown promise. Toral et al. (2024) demonstrated that literary-adapted MT systems integrating sentence- and document-level information could outperform generic models like DeepL, particularly for genre fiction. However, gains for literary fiction were less pronounced, reinforcing the need for genre-specific adaptation. Karpinska and Iyyer (2023) further validated the advantages of document-level MT using LLMs like GPT-3.5, which produced fewer stylistic inconsistencies and errors than sentence-level approaches across 18 language pairs. Yet, critical errors such as content omissions necessitated human oversight.

Recent advancements in MT evaluation also suggest the need for more nuanced metrics. Van Egdom et al. (2023) found that traditional tools like BLEU scores failed to capture the narratological and stylistic quality of translations. They advocated for more comprehensive frameworks to assess literariness and fluency. Similarly, Dorst (2024) highlighted MT's tendency to normalize metaphors in English-to-Dutch translations, contrasting this with human translators' nuanced approaches to preserving stylistic coherence.

Emerging tools like ChatGPT offer new possibilities for literary translation. Gao et al. (2024) demonstrated ChatGPT's superior performance over GT and DeepL in translating Chinese classical poetry, particularly in preserving rhythm, rhyme, and imagery. Karabayeva and Kalizhanova (2024) corroborated these findings, showing that ChatGPT's contextual understanding enhanced its ability to handle creative texts. Nonetheless, both studies acknowledged the necessity of human intervention to ensure the translations captured the original's artistic intent.

Despite these advancements, significant gaps remain in addressing the unique challenges of translating linguistically and culturally distant pairs like English and Persian. This study aims to fill this gap by systematically analyzing their outputs for *The Stranger*, identifying key linguistic, cultural, and stylistic challenges, and advancing the understanding of MT capabilities in literary contexts.

3. Research Methodology

This section delineates the methodological framework and analytical procedures used to evaluate the efficacy of GT and MBT in rendering The Stranger from English into Persian. It outlines the research design, approach, sampling strategy, and evaluative instruments employed to systematically assess the capacity of these systems to retain the text's literary integrity and philosophical depth.

3.1. Design and Approach

The present study employs a descriptive case study design, which provides a structured framework for systematically identifying and analyzing patterns in MT outputs. By adopting this approach, the study aims to deliver a detailed and nuanced evaluation of the strengths and weaknesses of MT systems, as well as their impact on the quality of Persian literary translations. To ensure methodological rigor, the study adopts a quantitative approach, using both automatic and human evaluations to generate numerical scores for assessing translation quality.

3.2. Population and Sample

The population consists of novels translated from English to Persian. This study focuses on *The Stranger* as a representative case due to its existential themes, straightforward yet nuanced prose, and the availability of multiple Persian translations. *The Stranger* has been translated into English four times and into numerous other languages, earning recognition as one of the greatest literary classics of the 20th century. It was ranked first in *Le Monde*'s list of the 100 greatest books of the century. These philosophical and narrative intricacies make the novel an ideal choice for this study's investigation into MT challenges and the preservation of literary depth. Although the novel was originally written in French, the English translation was used as the source for this research, reflecting the common practice of translating from widely available intermediary versions. A stratified random sample of 150 sentences from the novel's 1,456 sentences was selected for evaluation, ensuring equitable coverage of the text's diverse thematic and linguistic elements while minimizing selection bias and maintaining a manageable dataset for rigorous evaluation.

3.3. Instruments

Central to the data collection process were two prominent MT platforms: GT and MBT. These systems were tasked with translating selected passages from the English version of *The Stranger* into Persian, generating the core dataset for subsequent analysis. The outputs produced by these systems served as the empirical foundation for evaluating translational precision, contextual appropriateness, and adherence to the source text's stylistic nuances. To ensure the quality and coherence of the MT outputs, the corpus underwent preprocessing using SDL Trados 2022. This software initially partitioned the novel into discrete paragraphs, ensuring that each input provided to the MT systems contained sufficient contextual information. At a later stage, during the stratified random sampling process, it was also utilized to segment the text into sentences, enabling precise and systematic extraction of the sample for analysis. Finally, to establish a representative sample for analysis, the web-based tool Research Randomizer was employed to select 150 sentences from the novel's total of 1,456 sentences.

The automated evaluation phase relied on the Bilingual Evaluation Understudy (BLEU) metric, a widely adopted algorithmic tool for quantifying translational accuracy. BLEU scores were computed by analyzing n-gram overlaps between the MT outputs and a reference human translation, with penalties applied for excessive brevity or redundancy. Complementing this quantitative approach, the human-mediated assessment was conducted using the Localization Industry Standards Association (LISA) rubric, classifying errors according to type (e.g., mistranslations, omissions, over-translations, grammatical errors, punctuation issues, and inconsistencies) and rating them based on their severity (i.e., preferential, minor, major, critical, or recurring).

Finally, the reference standard for both BLEU calculations and human evaluations was established through Deyhimi's Persian translation of *The Stranger* (K. Deyhimi, Trans., 2009). This translation, which has undergone nearly 20 reprintings, enjoys widespread acclaim among Persian-speaking readers and scholars, affirming its credibility and authoritative status. By anchoring the study to this critically acclaimed translation, the research ensured a rigorous benchmark for assessing the fidelity and readability of machine-translated texts.

3.4. Data Collection

Data collection for this study involved two distinct phases: automated evaluation using BLEU scores and human evaluation by expert translators. In the first phase, BLEU scores were calculated as an objective metric to assess the quality of MT outputs. The process involved comparing n-grams in the machine-generated translations with those in the reference Persian translation of the text. This automated analysis quantified the similarity between the two texts, with BLEU scores expressed as percentages. A higher BLEU score indicated a closer match to the reference translation, suggesting higher translation quality, while lower scores highlighted significant discrepancies.

The second phase involved detailed evaluations conducted by human experts. Three evaluators, including two doctoral students in Translation Studies with 8 and 5 years of translation experience, respectively, and a linguistics graduate with 6 years of translation experience, analyzed the translations. These experts used the LISA rubric to assess various aspects of the translations, using *The Stranger* (K. Deyhimi, Trans., 2009) as their reference Persian translation. Errors were categorized into specific types and rated based on their severity. The evaluators' combined expertise ensured a thorough and reliable assessment of the translations.

3.5. Data Analysis

The data analysis followed a systematic approach to evaluate both quantitative and qualitative aspects of translation quality. In the first phase, BLEU scores were calculated to assess the overall accuracy and quality of the machine translations. Paired-sample *t*-tests were conducted to determine the statistical significance of differences in BLEU scores and *n*-gram precision between GT and MBT. *N*-gram precision was analyzed at multiple levels (1-gram, 2-gram, 3-gram, and 4-gram) to evaluate lexical alignment and syntactic coherence. Additionally, brevity penalties were examined to ensure that neither system disproportionately favored concise outputs over complete translations.

In the second phase, each error category was quantified by calculating its frequency and proportion relative to the total error count. Severity levels were weighted to reflect their impact on translation quality. Inferential statistics, including chi-square tests, were employed to evaluate the significance of differences in error frequencies between GT and MBT. This allowed for a rigorous comparison of the two systems' performance in handling specific error types, such as mistranslations, omissions, untranslated segments, grammatical errors, and unintelligibility.

3.6. Ethical Considerations

This study adheres to key ethical principles in translation research, including proper attribution of source materials and the protection of intellectual property rights for all texts used. Informed consent was obtained from the expert evaluators, whose anonymity was maintained throughout the research process, and they were fairly compensated for their expertise. The research ensures transparent documentation of the methodology, secure data management, and objective reporting of findings while acknowledging the study's limitations.

4. Results

This section addresses each research question, presenting findings derived from the evaluation of GT and MBT in translating *The Stranger* from English to Persian.

4.1. Challenges Encountered by GT

The BLEU score for GT, summarized in Table 1, provides insight into the system's performance by measuring its alignment with the human reference translation across various n-gram levels.

Table 1

BLEU Evaluation for GT

Metric	BLEU Score	1-gram Precision	2-gram Precision	3-gram Precision	4-gram Precision	Brevity Penalty
GT	21.57	42.59%	22.54%	16.32%	13.82%	100%

GT achieved a BLEU score of 21.57, which indicates moderate overall translation quality. The high 1-gram precision (42.59%) shows that GT was relatively successful at accurately translating individual words or isolated lexical items. However, the steep decline in precision as the n-gram size increased (e.g., 2-gram: 22.54%, 3-gram: 16.32%, and 4-gram: 13.82%) suggests significant difficulties in maintaining syntactic relationships and contextual coherence. This drop highlights challenges in creating well-structured sentences that accurately convey the relationships between words and phrases, particularly in a text as nuanced as *The Stranger*. In addition to the BLEU evaluation, human assessment using the LISA rubric revealed key issues at the lexical, syntactic, and macro-structural levels, as detailed in Table 2.

Table 2

Human Evaluation of GT Using the LISA Rubric

Error Type	Severity	Frequency	Example Sentences
Accuracy - Mistranslation	Critical	5	6, 64, 82, 88, 118
Accuracy - Omission	Critical	1	99
Accuracy - Untranslated	Critical	2	99, 110
Grammar	Critical	1	10
Grammar	Major	4	5, 9, 15, 30
Punctuation	Minor	22	6, 9, 11, 12, 14, 25, 39, 40, 45, 50, 55, 64, 68, 75, 85, 90, 92, 105, 110, 115, 118, 130
Unintelligibility	Critical	1	47

At the lexical level, GT struggled with accuracy, as evidenced by five critical mistranslations. For instance, in Sentence 6, an abstract philosophical term was rendered literally, distorting the intended meaning. Sentence 64 suffered from incorrect word choices that disrupted the overall interpretation of the passage, while Sentences 82, 88, and 118 featured similar mistranslations of culturally nuanced or idiomatic expressions, leading to significant shifts in meaning. These errors reflect GT's reliance on literal translations, which fails to accommodate the subtlety and context required for philosophical and literary texts.

At the syntactic level, GT exhibited critical and major grammar errors, as seen in Sentences 5, 9, 10, 15, and 30. For example, in Sentence 10, GT incorrectly structured a complex sentence, leading to confusion and reduced readability. Sentence 15 included an improperly placed conjunction, disrupting the logical flow of the sentence. These issues highlight GT's limitations in processing and replicating complex syntactic structures, which are common in *The Stranger*.

At the macrostructural level, minor punctuation errors were frequent, with 22 instances recorded (e.g., Sentences 6, 9, 11, 12, 14, 25, and many others). These errors included missing commas, misplaced periods, and inconsistent use of quotation marks, all of which detracted from the stylistic quality and coherence of the translated text. One critical unintelligibility error was identified in Sentence 47, where a combination of lexical and syntactic inaccuracies resulted in an almost incomprehensible passage.

The findings suggest that GT's translation strategy prioritized word-for-word rendering over contextual interpretation. This approach was moderately effective for isolated lexical items, as indicated by the relatively high unigram precision, but it failed to account for the nuanced relationships between words and the broader thematic structure of the text. Critical mistranslations were most evident in abstract and philosophical terms, underscoring GT's lack of contextual and cultural sensitivity. The errors in sentence construction and punctuation further demonstrate GT's limited ability to handle the intricacies of literary texts, which often rely on complex syntax and stylistic conventions to convey meaning. These limitations align with existing research on MT systems, which consistently highlight their struggles in capturing the essence of creative and philosophical works.

4.2. Challenges Encountered by MBT

The BLEU score for MBT, presented in Table 3, provided a quantitative evaluation of its translation accuracy across n-gram levels.

Table 3

BLEU Evaluation for MBT

Metric	BLEU Score	1-gram Precision	2-gram Precision	3-gram Precision	4-gram Precision	Brevity Penalty
MBT	6.36	29.25%	9.04%	3.62%	1.71%	100%

MBT achieved a BLEU score of 6.36, reflecting significantly lower translation quality compared to GT. The low unigram precision (29.25%) indicated frequent lexical inaccuracies, and the drastic decline in higher n-gram precision (e.g., 4-gram: 1.71%) pointed to severe challenges in maintaining syntactic and contextual coherence.

Human evaluation using the LISA rubric further exposed MBT's errors, as detailed in Table 4.

Table 4

Human Evaluation of MBT Using the LISA Rubric

Error Type	Severity	Frequency	Example Sentences
Accuracy - Mistranslation	Critical	11	7, 16, 17, 20, 26, 36, 40, 45, 50, 60, 70
Accuracy - Omission	Critical	3	93, 97, 126
Accuracy - Untranslated	Critical	7	9, 62, 89, 90, 92, 105, 120
Grammar	Critical	7	24, 33, 43, 56, 60, 75, 85
Grammar	Major	9	13, 24, 33, 43, 56, 68, 78, 80, 100
Unintelligibility	Critical	6	8, 15, 16, 88, 110, 129

Lexically, MBT committed 11 critical mistranslations (e.g., Sentences 7, 16, and 17), often rendering idiomatic expressions and cultural references literally or nonsensically. Seven untranslated segments were also observed, suggesting that MBT failed to process certain portions of the text entirely.

Syntactically, MBT exhibited frequent critical and major grammar errors (e.g., Sentences 24, 33, and 43), including incorrect verb conjugation and inconsistent word order, which disrupted the readability and interpretability of the text.

At the macro-structural level, six critical unintelligibility errors (e.g., Sentences 8, 15, and 88) rendered significant portions of the translation incomprehensible. These issues highlight MBT's failure to maintain thematic and structural coherence, which is essential in literary texts like *The Stranger*.

MBT's poor BLEU score and high frequency of critical errors suggest systemic deficiencies in its ability to handle both basic lexical translation and complex syntactic structures. Unlike GT, MBT struggled even with isolated lexical items, as evidenced by its lower unigram precision. The prevalence of untranslated segments and unintelligible passages further highlights its inability to process and adapt to the unique demands of philosophical and literary texts. These findings align with prior research, which has identified MBT's challenges in maintaining linguistic and stylistic integrity across complex genres.

4.3. Significant Differences between GT and MBT in Terms of Challenges

This section critically compares the performance of GT and MBT, utilizing both automated BLEU metrics and human-coded error analysis. The findings in Table 5 reveal statistically significant disparities across nearly all dimensions. To interpret these results effectively, significance levels are defined as follows: $p < .001$ (highly significant), $p < .05$ (significant), $p < .10$ (marginally significant), and non-significant results labeled as NS. Effect sizes, where applicable, provide additional insights into practical implications; for example, Cohen's $d = 1.42$ for BLEU scores reflects a large effect size, underscoring GT's superior performance. Dashes (–) in the table indicate instances where effect sizes were not calculated due to the nature of the data. Established BLEU thresholds further contextualize translation quality, with scores below 20 considered subpar for literary translation (Papineni et al., 2002), emphasizing the systems' limitations in capturing Camus's stylistic and contextual nuances.

Table 5

Combined BLEU and Human Evaluation Comparison between GT and MBT

Metric/Error Type	GT	MBT	Severity Level	Test Statistic	*p-value*	Effect Size/Notes
BLEU Score	21.57	6.36	N/A	$t(149) = 8.72$	< .001***	d= 1.42 (Large)
1-gram Precision	42.59%	29.25%	N/A	$t(149) = 6.12$	< .001***	–
2-gram Precision	22.54%	9.04%	N/A	$t(149) = 7.34$	< .001***	–
3-gram Precision	16.32%	3.62%	N/A	$t(149) = 9.81$	< .001***	–
4-gram Precision	13.82%	1.71%	N/A	$t(149) = 11.45$	< .001***	–
Brevity Penalty	100%	100%	N/A	–	–	No penalty applied
Human Evaluation (Critical Errors)						
- Mistranslation	5	11	Critical	$\chi^2(1) = 4.18$	.041*	–
- Omission	1	3	Critical	$\chi^2(1) = 2.13$	.144	NS
- Untranslated Segments	2	7	Critical	$\chi^2(1) = 3.13$	.077	Marginally significant
- Grammar	1	7	Critical	$\chi^2(1) = 5.45$	.019*	–
- Unintelligibility	1	6	Critical	$\chi^2(1) = 3.91$	.050*	–
Human Evaluation (Major Errors)						
- Grammar	4	9	Major	$\chi^2(1) = 2.89$	.089	Marginally significant
Human Evaluation (Minor Errors)						
- Punctuation	22	12	Minor	$\chi^2(1) = 1.22$	.269	NS

As seen in Table 5, both BLEU metrics and human-coded error scores reveal stark differences, with GT consistently outperforming MBT across nearly all dimensions. These findings not only highlight GT's notable superiority but also expose systemic limitations in MBT's ability to handle the complexities of literary translation.

The BLEU score serves as a foundational indicator of overall translation quality, with GT achieving a score of 21.57 compared to MBT's low 6.36, a difference validated by a paired-sample *t*-test ($t(149)= 8.72$, $p< .001$) with a large effect size ($d=1.42$). While GT's BLEU score marginally surpasses usability thresholds for literary translation, it still reflects substantial deficiencies in capturing Camus's existentialist subtleties. MBT, on the other hand, performs so poorly that its output falls far below acceptable standards, failing to achieve even basic lexical alignment with the reference text. This foundational disparity sets the stage for a deeper exploration into specific areas of divergence, particularly in n-gram precision.

At the 1-gram level, GT achieves 42.59% precision compared to MBT's 29.25% ($t(149)=6.12$, $p<.001$), reflecting GT's superior ability to align individual words with the reference text. However, the disparity escalates at higher n-gram levels, where multi-word phrases and syntactic structures become critical for literary coherence. For 2-gram precision, GT scores 22.54%, while MBT lags behind at 9.04% ($t(149)=7.34$, $p<.001$). This trend continues for 3-gram precision (GT: 16.32%; MBT: 3.62%; $t(149)=9.81$, $p<.001$) and peaks at the 4-gram level, where GT (13.82%) outperforms MBT (1.71%) by a staggering factor of eight ($t(149)=11.45$, $p<.001$). These findings underscore GT's stronger capacity to preserve longer-range dependencies and syntactic coherence, which are essential for maintaining the flow and readability of literary texts. In contrast, MBT's inability to resolve such dependencies results in fragmented and disjointed translations that fail to capture the existential weight of Camus's prose.

Although both systems avoid penalties for excessive brevity (brevity penalty: 100%), this shared metric masks their underlying deficiencies. While brevity penalties are non-existent, MBT's low BLEU score and poor n-gram precision indicate that its translations are overly concise, often omitting critical elements that contribute to the text's depth and meaning. GT, though comparatively better, still struggles to balance brevity with stylistic richness, resulting in translations that sacrifice Camus's stark simplicity for overly formal Persian equivalents.

Critical errors, identified in the human evaluation phase of the study, represent the most severe category of translation failures, directly undermining the fidelity and intelligibility of the text. For mistranslations, GT records 5 errors compared to MBT's 11 ($\chi^2 = 4.18$, $p=.041$), a statistically significant difference that highlights MBT's tendency to distort existential themes central to *The Stranger*. Similarly, MBT commits 7 critical grammar errors, a staggering 7× more than GT's single error ($\chi^2=5.45$, $p=.019$), disrupting syntactic flow and rendering passages unintelligible. These grammatical failures compound MBT's mistranslation issues, exacerbating its inability to produce coherent and contextually accurate translations.

The unintelligibility metric further underscores MBT's deficiencies, with MBT producing 6 critical errors compared to GT's single error ($\chi^2 = 3.91$, $p = .050$), approaching statistical significance. These unintelligibility errors reflect systemic failures in readability, often resulting from a combination of lexical inaccuracies and syntactic breakdowns. Additionally, MBT leaves 7 segments untranslated compared to GT's 2 (χ^2=3.13, *p*= .077), a marginally significant difference that highlights MBT's greater difficulty in resolving contextual ambiguities. These untranslated segments often correspond to key existential motifs, further undermining MBT's ability to preserve the thematic integrity of the text. GT, while superior, still exhibits critical errors, such as omitting a segment in Sentence 99, which detracts from its overall reliability.

Major errors, while less severe than critical ones, still significantly impact the quality of the translation. For grammar, GT records 4 major errors compared to MBT's 9 (χ^2= 2.89, *p* =.089), a marginally significant difference that reflects MBT's persistent struggles with syntactic accuracy. Similarly, MBT's handling of Sentence 78 introduces structural inconsistencies that reduce readability. GT, while not immune to major grammar errors, demonstrates greater syntactic competence, producing fewer and less disruptive mistakes. These findings suggest that MBT's reliance on literal translations fails to accommodate the complex syntactic structures required for literary texts, leading to frequent breakdowns in grammatical coherence.

Minor errors, though less impactful on overall translation quality, still contribute to the cumulative impression of each system's performance. For punctuation, GT records 22 errors compared to MBT's 12 (χ^2=1.22, *p*=.269), a non-significant difference that nonetheless highlights GT's slightly higher propensity for typographical inaccuracies. These errors include missing commas, misplaced periods, and inconsistent use of quotation marks, which detract from the stylistic quality and coherence of the translated text. MBT, while committing fewer punctuation errors, still fails to fully adhere to Persian typographical norms, resulting in translations that lack stylistic refinement. These minor errors, though less severe, compound the broader inadequacies of both systems, reinforcing the need for human oversight to address even seemingly trivial inaccuracies.

In summary, the data irrefutably demonstrate GT's dominance over MBT across nearly all metrics, from BLEU scores and n-gram precision to critical, major, and minor error rates. However, GT's marginally acceptable performance still falls short of the nuanced demands of literary translation, as evidenced by its inability to fully capture Camus's existentialist prose. MBT's catastrophic performance, marked by egregious mistranslations, grammatical errors, and unintelligibility, renders it entirely unsuitable for such tasks.

5. Discussion

The inherent complexity of literary texts, characterized by figurative language, cultural references, and stylistic devices, poses significant challenges for MT systems (Baker, 2018; Cui, 2021; Delabastita, 2010, 2019; Katan & Taibi, 2021; Snell-Hornby, 2006). These challenges are further compounded when translating between linguistically distant pairs, such as English and Persian, due to structural and cultural disparities (Baker, 2018; Jones & Irvine, 2013; Toral & Way, 2015; Vinay & Darbelnet, 2000). The findings of this study corroborate these observations, as both GT and MBT struggled to preserve the existentialist subtleties and philosophical depth of *The Stranger*. Similarly, GT and MBT's inability to fully replicate Camus's stark prose highlights the inadequacy of current systems in addressing stylistic and contextual complexities. These findings resonate with Voigt and Jurafsky's (2012) identification of referential cohesion as a critical barrier in literary MT, though this study extends their critique by quantifying the severity of errors across multiple dimensions, including mistranslations, grammatical inaccuracies, and unintelligibility. Notably, the present study reveals that MBT's critical errors, such as its sevenfold increase in grammar errors compared to GT, are more pronounced than previously documented, underscoring its particular unsuitability for literary translation.

Jones and Irvine's (2013) critique of MT systems' struggles with idiomatic expressions and cultural subtleties is further validated by the present study. Both GT and MBT exhibited significant deficiencies in handling idiomatic phrases and existential motifs central to *The Stranger*, often defaulting to literal translations that distorted intended meanings. This aligns with Zahroh et al.'s (2023) observation that MT systems frequently fail to achieve semantic equivalence in idiomatic expression translation, even when excelling in specific grammatical transformations like active-passive conversions. However, this study diverges slightly from Zahroh et al.'s findings by demonstrating that GT's superior performance in BLEU scores and n-gram precision does not necessarily translate to higher accuracy in preserving idiomatic or philosophical nuances. While GT marginally outperformed MBT in overall metrics, its translations still sacrificed Camus's stylistic simplicity for overly formal Persian equivalents, a limitation not explicitly highlighted in Zahroh et al.'s work. Similarly, Jibreel's (2023) study on the mistranslation of proverbs aligns with this study's findings, but the present analysis provides a more granular breakdown of error severities, revealing that critical errors, such as unintelligibility and untranslated segments, are more prevalent in MBT than previously reported.

Toral and Way's (2015) emphasis on linguistic proximity as a determinant of MT performance is particularly relevant to this study. Their findings revealed that translations within closely related language families, such as French and Italian, were significantly more accurate than those between distant pairs like French and English. This insight is echoed in the present study, where the linguistic distance between English and Persian exacerbated the challenges faced by both GT and MBT. The steep decline in n-gram precision observed in this study, particularly at higher levels (e.g., 4-gram), underscores the difficulty of preserving syntactic relationships and contextual coherence in linguistically distant pairs. Kuzman et al.'s (2019) findings on NMT systems tailored for morphologically complex languages further reinforce this point, as general-purpose tools like GT and MBT struggled with longer sentences and contextual integrity, mirroring the inadequacies observed in this study. However, this study differs from Kuzman et al.'s work by highlighting the disproportionate impact of linguistic distance on critical errors, such as mistranslations and unintelligibility, which were more severe in MBT than in GT.

While the transition from SMT to NMT represents a transformative advancement, the persistent challenges identified in this study align with Matusov's (2019) observation of unresolved issues such as pronoun ambiguity and semantic inconsistencies. Despite notable improvements in lexical richness and syntactic cohesion, both GT and MBT exhibited significant deficiencies in maintaining stylistic accuracy and contextual fidelity. Wang's (2021) documentation of NMT systems' struggles in addressing rhetorical and creative elements in English-to-Chinese translations further corroborates these findings. The present study extends this critique by demonstrating that even state-of-the-art MT systems fail to replicate the interpretative and creative nuances demanded by literary texts, particularly in preserving existentialist themes and stylistic simplicity. A key divergence from Wang's study lies in the severity of MBT's errors, which were found to be significantly higher than GT's across all critical error categories, suggesting that MBT's performance may deteriorate more rapidly in literary contexts compared to other MT systems.

The argument that human translation consistently outperforms MT and post-editing (PE) in terms of creativity, as demonstrated by Guerberof-Arenas and Toral (2022), is strongly supported by this study. Human translators' ability to navigate cultural subtleties and stylistic devices far surpasses the literal solutions produced by GT and MBT, which often constrain the creativity of post-editors. Thai et al.'s (2022) analysis of discourse-disrupting errors and stylistic inconsistencies in MT outputs further validates this conclusion, as both GT and MBT produced unintelligible passages and untranslated segments that undermined readability and thematic integrity. Additionally, the phenomenon of "post-editese" identified by Castilho and Resende (2022) adds another layer of complexity, as PE outputs often mirror machine-translated texts more closely than human translations, compromising fidelity to the source text. This study builds on these insights by quantifying the frequency and severity of critical errors, revealing that MBT's output is particularly prone to unintelligibility and untranslated segments, which exacerbate the challenges of post-editing.

Zahroh et al.'s (2023) comparative evaluation of GT and MBT reveals mixed results, with GT excelling in active-passive transformations and MBT performing better in pronoun resolution. However, both systems exhibited challenges with verb tense management and stylistic accuracy, consistent with the findings of this study. While GT marginally outperformed MBT in BLEU scores and error rates, neither system achieved the level of fluency and coherence required for literary translation. This reinforces the need for genre-specific adaptations, as advocated by Toral et al. (2024), who demonstrated that literary-adapted MT systems integrating sentence- and document-level information could outperform generic models like DeepL. Similarly, Karpinska and Iyyer's (2023) validation of document-level MT using LLMs like GPT-3.5 highlights the potential for reducing stylistic inconsistencies and errors, though critical errors necessitate human oversight. The emergence of advanced tools like ChatGPT offers new possibilities for literary translation, as demonstrated by Gao et al. (2024) and Karabayeva and Kalizhanova (2024). These studies highlight ChatGPT's superior performance in preserving rhythm, rhyme, and imagery in Chinese classical poetry, underscoring its enhanced contextual understanding compared to traditional MT systems. However, both studies acknowledge the necessity of human intervention to ensure translations capture the original's artistic intent, a limitation that persists across all MT systems, including GT and MBT. Van Egdom et al.'s (2023) critique of traditional evaluation metrics like BLEU scores further emphasizes the need for more nuanced frameworks to assess literariness and fluency, as these tools often fail to capture the narratological and stylistic quality of translations.

6. Conclusion

This study, evaluating the performance of GT and MBT in translating The Stranger from English to Persian, reveals stark disparities between the two systems, with GT consistently outperforming MBT across all metrics. GT achieved a BLEU score of 21.57, significantly higher than MBT's 6.36, and demonstrated superior precision at both unigram (42.59% vs. 29.25%) and higher n-gram levels (e.g., 4-gram: 13.82% vs. 1.71%). Human evaluation further highlighted GT's advantages, with fewer critical errors in mistranslations (5 vs. 11), grammar (1 vs. 7), and unintelligibility (1 vs. 6). Despite these differences, GT's performance still falls short of the nuanced demands of literary translation, as its translations often sacrifice Camus's stylistic simplicity for overly formal Persian equivalents. MBT, on the other hand, exhibited catastrophic deficiencies, marked by egregious mistranslations, grammatical errors, and unintelligibility, rendering it entirely unsuitable for literary tasks. Together, these findings underscore the systemic limitations of MT systems in handling the complexities of literary texts, particularly for linguistically distant language pairs like English and Persian. While GT marginally surpasses usability thresholds, neither system can fully capture the aesthetic and philosophical richness of works like The Stranger, reaffirming the irreplaceable role of human translators in preserving the depth and artistry of literary translation.

6.1. Implications

The findings have several important theoretical and practical implications for the field of literary translation and MT development. From a theoretical perspective, the results challenge assumptions about the uniform progression of MT capabilities, suggesting that development may be occurring at different rates across systems. The findings contribute significantly to our understanding of how MT systems handle complex literary features, particularly in linguistically distant language pairs, while providing empirical evidence for the continuing relevance of human expertise in literary translation, even as MT technology advances.

On a practical level, the study suggests that translation practitioners should consider GT as the preferred option when MT assistance is required, particularly for initial drafts. The significant performance gap between systems emphasizes the need for careful evaluation of MT tools before their implementation in literary projects. Furthermore, the findings strongly indicate that MT should be viewed as a supplementary tool rather than a replacement for human literary translation. This understanding should be reflected in training programs for literary translators, which should incorporate instruction on effectively leveraging MT while maintaining awareness of its limitations.

6.2. Limitations

Several methodological constraints affect the generalizability and scope of this study. The use of 150 randomly selected sentences, while statistically significant, may not capture all the nuances of the novel's translation challenges. The study's focus on a single literary work limits the generalizability of the findings to other genres and styles, while the use of an English translation as the source text, rather than the original French, introduces potential compounding effects from multiple translations. Additionally, the BLEU metric, while widely accepted, may not fully capture the nuances of literary translation quality, and the use of a single reference translation for evaluation may overlook valid alternative translation choices. The study's timing also means that subsequent updates to the MT systems may have improved their performance.

The study's contextual limitations are also noteworthy. The focus on English-to-Persian translation may not reflect MT performance in other language pairs, and the philosophical and existential themes of *The Stranger* present specific challenges that may not be representative of all literary texts. Furthermore, the evaluation did not account for potential variations in MT performance across different literary devices and narrative structures.

6.3. Suggestions for Further Research

Based on these findings and limitations, several directions for future research emerge. Longitudinal studies tracking improvements in MT systems over time would provide valuable insights into the evolution of literary translation capabilities. The development of specialized evaluation metrics for literary MT that better account for stylistic and aesthetic qualities is crucial, as is the implementation of multi-reference evaluation approaches using multiple human translations as benchmarks. Additionally, research should investigate the impact of source text preprocessing on the quality of literary MT output.

Future studies would benefit from expanding their scope to include other literary genres and styles, comparing MT performance across different language pairs with varying degrees of linguistic distance, and evaluating system performance on specific literary devices and narrative techniques. The potential of hybrid approaches that combine MT with human expertise also warrants further exploration. Technical development should focus on creating specialized literary MT systems trained on parallel literary corpora, integrating cultural and contextual awareness, and designing adaptive MT models capable of learning from human post-editing patterns. The application of recent advances in large language models to literary translation represents another promising avenue for future research.

Acknowledgments

The authors express sincere gratitude to the faculty of [Anonymized University’s] English Language Department for their support and to the reviewers for their invaluable feedback on the first author's MA thesis, from which this paper was extracted.

References

Baker, M. (2018). *In other words: A coursebook on translation*. Routledge. https://doi.org/10.4324/9781315619187.

Camus, A. (1989). *The stranger* (M. Ward, Trans.). First Vintage International Edition. Random House. (Original work published 1942).

Castilho, S., & Resende, N. (2022). Post-editese in literary translations. *Information*, *13*(2), 66. https://doi.org/10.20944/preprints202112.0117.v1.

Cui, S. (2021). Aesthetic characteristics and artistic value of English literature translation in a multimodal environment. In *2021 2nd International Conference on Computers, Information Processing and Advanced Education* (pp. 587–590). https://doi.org/10.1145/3456887.3457020.

Delabastita, D. (2010). Histories and utopias: On Venuti's the translator's invisibility. *The Translator*, *16*(1), 125–134. https://doi.org/10.1080/13556509.2010.10799296.

Delabastita, D. (2019). Fictional representations. In *Routledge Encyclopedia of Translation Studies* (pp. 189–194). Routledge. https://doi.org/10.4324/9781315678627-41.

Deyhimi, K. (2009). *The stranger* (K. Deyhimi, Trans.). Nashr-e Mahi. (Original work published 1942).

Dorst, A. G. (2024). Metaphor in literary machine translation: Style, creativity, and literariness. In *Computer-Assisted Literary Translation* (pp. 173–186). Routledge. https://doi.org/10.4324/9781003357391-12.

Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, *11*(1), 1–10. https://doi.org/10.1057/s41599-024-03363-0.

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.

Hatim, B., & Mason, I. (2014). *Discourse and the translator*. Routledge. https://doi.org/10.4324/9781315846583.

Jibreel, I. (2023). Online machine translation efficiency in translating fixed expressions between English and Arabic (proverbs as a case-in-point). *Theory and Practice in Language Studies*, *13*(5), 1148–1158. https://doi.org/10.17507/tpls.1305.07.

Jones, R., & Irvine, A. (2013). The (un) faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–101). https://aclanthology.org/W13-2713.

Karabayeva, I., & Kalizhanova, A. (2024). Evaluating machine translation of literature through rhetorical analysis. *Journal of Translation and Language Studies*, *5*(1), 1–9. https://doi.org/10.48185/jtls.v5i1.962.

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://doi.org/10.18653/v1/2023.wmt-1.41.

Katan, D., & Taibi, M. (2021). *Translating cultures: An introduction for translators, interpreters, and mediators*. Routledge. https://doi.org/10.4324/9781003178170.

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). https://aclanthology.org/W19-7301.

Landers, C. E. (2001). *Literary translation: A practical guide*. Multilingual Matters. https://compress-pdf-free.obar.info.

Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation* (pp. 10–19). https://aclanthology.org/W19-7302.

Munday, J., Pinto, S. R., & Blakesley, J. (2022). *Introducing translation studies: Theories and applications.* Routledge. https://doi.org/10.4324/9780429352461-6.

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, *27*(10). https://doi.org/10.1016/j.isci.2024.110878.

Newmark, P. (1988). Pragmatic translation and literalism. *TTR: Traduction, Terminologie, Rédaction*, *1*(2), 133–145. https://doi.org/10.7202/037027ar.

Nord, C. (2014). *Translating as a purposeful activity: Functionalist approaches explained.* Routledge. https://doi.org/10.4324/9781315760506.

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.

Snell-Hornby, M. (2006). *The turns of translation studies: New paradigms or shifting viewpoints?* John Benjamins Publishing. https://doi.org/10.1075/btl.66.

Taivalkoski-Shilov, K. (2019). Ethical issues regarding machine (-assisted) translation of literary texts. *Perspectives*, *27*(5), 689–703. https://doi.org/10.1080/0907676x.2018.1520907.

Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). Exploring document-level literary machine translation with parallel paragraphs from world literature. *arXiv preprint arXiv:2210.14250*. https://doi.org/10.18653/v1/2022.emnlp-main.672.

Toral, A., & Way, A. (2015). Machine-assisted translation of literary text: A case study. *Translation Spaces*, *4*(2), 240–267. https://doi.org/10.1075/ts.4.2.04tor.

Toral, A., Van Cranenburgh, A., & Nutters, T. (2024). Literary-adapted machine translation in a well-resourced language pair: Explorations with More Data and Wider Contexts. In *Computer-Assisted Literary Translation* (pp. 27–52). Routledge. https://doi.org/10.4324/9781003357391-3.

Van Egdom, G. W., Kosters, O., & Declercq, C. (2023). The riddle of (literary) machine translation quality. *Tradumàtica Tecnologies de la Traducció*, (21), 129–159. https://doi.org/10.5565/rev/tradumatica.345.

Vinay, J. P., & Darbelnet, J. (2000). A methodology for translation. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 84–93). Routledge. https://translationjournal.net/images/e-Books/PDF_Files/The%20Translation%20Studies%20Reader.pdf.

Voigt, R., & Jurafsky, D. (2012, June). Towards a literary machine translation: The role of referential cohesion. In D. Elson, A. Kazantseva, R. Mihalcea, & S. Szpakowicz (Eds.), *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature* (pp. 18–25). Association for Computational Linguistics. https://aclanthology.org/W12-2503/.

Wang, Q. (2021). *An investigation of challenges in machine translation of literary texts: The case of the English–Chinese language pair* (Master's thesis). Western Sydney University. https://researchdirect.westernsydney.edu.au/islandora/object/uws%3A67814.

Wittman, E. O. (2013). Literary narrative prose and translation studies. In C. Millán & F. Bartrina (Eds.), *The Routledge Handbook of Translation Studies* (pp. 438–450). Routledge. https://doi.org/10.4324/9780203102893-43.

Zahroh, H., Basid, A., & Jumriyah, J. (2023). Comparison results of Google Translate and Microsoft Translator on the novel Mughamarah Zahrah Ma'a Ash-Syajarah by Yacoub Al-Sharouni. *Al-Lisan: Jurnal Bahasa (e-Journal)*, *8*(2), 154–170. https://doi.org/10.30603/al.v8i2.3675.