English Studies at NBU, 2025
Volume 11, Issue 1, 17-42
https://doi.org/10.33919/esnbu.25.1.2

pISSN 2367-5705
eISSN 2367-8704
www.esnbu.org

TURKISH-TO-ENGLISH SHORT STORY TRANSLATION BY DEEPL: HUMAN EVALUATION BY TRAINEES AND TRANSLATION PROFESSIONALS VS. AUTOMATIC EVALUATION

Halise Gülmüş Sırkıntı

Marmara University, Istanbul, Türkiye

Abstract

This mixed-methods study aims to evaluate the quality of Turkish-to-English literary machine translation by DeepL, incorporating both human and automatic evaluation metrics while engaging translation trainees and professional translators. Raw MT output of two short stories, Mendil Altında and Kabak Çekirdekçi, evaluated by both groups via TAUS DQF tool and evaluators wrote reports on the detected errors. Additionally, BLEU was employed for automatic evaluation. The results indicate a consensus between trainees and professionals in assessing MT accuracy and fluency. Accuracy rates were 80.59% and 80.50% for Mendil Altında, and 73.08% and 82.35% for Kabak Çekirdekçi. Fluency rates were similarly close, 71.96% and 72.32% for Mendil Altında, and 66.81% and 62.09% for Kabak Çekirdekçi. Bleu scores, particularly 1-gram results, align with the human evaluators' results. Furthermore, reports show that trainees provided more detailed analysis, frequently using meta-language, suggesting that increased exposure to metrics enhances trainees' ability to identify fine-grained MT errors.

Keywords: literary translation, machine translation evaluation, human evaluation, automatic evaluation, BLEU

Article history:
Received: 20 February 2025;
Reviewed: 31 March 2025;
Accepted: 03 May 2025;
Published: 30 June 2025

Citation: Gülmüş Sırkıntı, H. (2025). Turkish-to-English short story translation by DeepL: Human evaluation by trainees and translation professionals vs. automatic evaluation. *English Studies at NBU, 11*(1), 17-42. https://doi.org/10.33919/esnbu.25.1.2

Halise Gülmüş Sırkıntı is an Assistant Professor in the Department of Translation and Interpreting at Marmara University, Türkiye. She holds a BA, MA, and PhD in Translation and Interpreting Studies. Her research focuses on literary translation, travel writing, and translation criticism.

E-mail: halisegulmus@gmail.com https://orcid.org/0000-0002-6585-5961

Translating Literary Texts via Machine Translation

The idea of using mechanical dictionaries for translation dates back to the seventeenth century, but concrete plans for machine translation (MT) emerged only in the twentieth century (Hutchins, 1995, p. 431). Between 1956 and 1966, MT saw high expectations, but the 1966 ALPAC report highlighted its limitations, temporarily curbing research interest (Hutchins, 1995, p. 434). Following its publication, research in English-speaking regions declined, as "MT became the victim of its own unrealistic expectations" (Quah, 2006, p. 61). However, research teams in other countries persisted in getting financing for MT projects (Poibeau, 2017). The EUROTRA (European Translation) project led to a revival of MT research in Europe from the 1970s to 1992, and advances in computational linguistics in the 1980s further facilitated progress in MT research (Quah, 2006, p. 62-63). From 1984 to 1992, MT underwent a phase of steady growth that was characterized by gradual advancement and improvement (Sin Wai, 2015, p. 5). In the early 1990s, technological advances in communication and computing technology reshaped the translation field, fostering the swift growth and widespread adoption of MT and computer-aided translation tools (Quah, 2006, p. 65). This transformative period also witnessed the shift from the dominance of rule-based machine translation (RBMT) in the 1950s to the 1980s, to the rise of statistical machine translation (SMT) in the 1990s. (Melby, 2020, p. 684; Yang and Min, p. 2015, p. 201). Nevertheless, MT wasn't completely revolutionized until the development of neural machine translation (NMT) in the twenty-first century. Unlike previous MT solutions, neural networks have the capability to generate words in the correct context, making the translation output more accurate and contextually appropriate (Taivalkoski-Shilov, 2019, p. 690). Typically, an NMT model has two parts: A decoder network generates the translation from a real-valued vector that an encoder network converts from the source text (Wang et al., 2022, p. 144). The promising results of NMT from the studies conducted (Bahdanau et al., 2015; Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Wu et al., 2016; Klubička et al., 2017; Shterionov et al., 2018) have sparked interest in post-editing as a human-machine collaboration, both among Translation Studies scholars and industry practitioners (O'Brien et al., 2014, p. vii; O'Hagan, 2020, p. 27). The increasing need for effective and cost-effective translation processes has also elevated the significance of MT and post-editing (Dillinger, 2014, p. ix).

Building on these progresses in machine translation, the focus shifted to exploring the feasibility and potential challenges of literary machine translation in the field. Literary texts, in contrast to technical texts, contain literary devices, cultural references, and aesthetic characteristics (Birkan Baydan, 2016) which pose significant problems for MT systems. As Maria Tymoczko (2014, p. 14-15) asserts "literary language is rich and complex" and literary works constitute the largest, most complex, and most representative collection of texts in terms of cross-cultural textual practices. Furthermore, according to Taivalkoski-Shilov (2019, p. 696) "the omnipresence and complexity of voice in literary text creates a great challenge for MT in literary translation". Translating literature necessitates a deep understanding of the context, emotions, and literary techniques, which poses challenges for MT systems. However, with the advances in MT, the field of literary MT has been the subject of some research in a variety of languages, including Chinese (Jiang & Niu, 2022), Dutch (Webster, 2020), Japanese (Gu, 2022), Korean (Mah, 2020), and Turkish (Şahin & Gürses, 2021; Ayık Akça, 2022; Aslan, 2024; Dallı et al., 2024; Gürses et al., 2024). While some studies focused primarily on the creative aspects of literary MT (Guerberof-Arenas & Toral, 2022), others have examined the ethical aspects of this activity (Taivalkoski-Shilov, 2019). Additionally, research on translation training and MT has also increased in recent years (Öner Bulut & Alimen, 2023; Trojszczak, 2022; Guerberof Arenas & Moorkens, 2019; Öner Bulut, 2019; Kenny & Doherty, 2014).

Machine Translation Evaluation

Evaluation techniques are not static components; rather, they develop similarly to the MT systems (Giménez & Màrquez, 2010, p. 77). In the same way that MT systems are constantly being developed and improved, evaluation methodologies also vary and advance with time. Throughout history, alongside the development of MT, there have been endeavors to assess its quality. In addition to introducing non-numerical programming on a computer for the first time, Georgetown University and IBM's initial MT demonstration in 1954 also marked the beginning of the first MT evaluation (Chunyu & Tak-ming, 2015, p. 214). Early studies like those by Miller and Beebe-Center (1956), which evaluated Russian-English systems based on human evaluations of elements like comprehensibility and fluency, are at the foundation of the history of MT evaluations. The European Commission thoroughly assessed Systran systems throughout the 1970s and the 1990s, and the 1990s saw specialized conferences addressing MT evaluation-related issues.

In global literature, assessment methods are categorized as either automated or human (manual) metrics (Chatzikoumi, 2019, p. 3). The human evaluation is concerned with how a human would rate or annotate the MT output. Although human evaluations initially were dominant, more recent efforts have concentrated on creating automatic or semi-automatic evaluation systems since they need less time and effort than human evaluations do (Hutchins, 2015, p. 130). The use of statistical analysis to evaluate MT systems automatically has been a significant result of the development of SMT models. The IBM group's BLEU (Papineni et al., 2002) was the first metric, and it was followed by the NIST (Hutchins, 2015, p. 130). By counting co-occurring n-grams in the MT output and reference sentences, BLEU quantifies the idea that greater similarity between machine and human translations signals higher quality (Chunyu & Tak-ming, 2015, p. 226). Assessing how closely the candidate translation adheres to the reference translations is the goal of BLEU, which "counts the number of matching n-grams (typically n ∈ {1, .., 4}) and computes a weighted average" (Shterionov et al., 2018, p. 222). With the use of these metrics, extensive analyses of numerous systems and language pair combinations can be carried out quickly and affordably (Chunyu & Tak-ming, 2015, p. 216). Unquestionably, automatic evaluation is also useful for tracking whether a given MT system has improved or not over time (Hutchins, 2015, p. 131). Although manual assessment takes much longer than automatic assessment, automatic assessment has frequently been criticized as having downsides (Webster et al., 2020, p. 2). It shouldn't be forgotten that automatic metrics only measure how closely a MT resembles a human-translated source text rather than evaluating the translation's quality; so, they are insufficient to fully check their reliability and consistency (Chunyu & Tak-ming, 2015, p. 232). As an example, BLEU, being the mostly preferred automatic MT evaluation metric (Shterionov et al., 2018, p. 222), has been criticized as having shortcomings in comparison to human evaluation (Callison-Burch et al., 2006; Smith et al., 2016). So, it can be said that despite improvements in automatic metrics, human evaluation continues to be essential for collecting nuanced details and contextual relevance. Within this regard, this study used both automatic and human MT evaluation metrics to have a comprehensive approach to translation evaluation.

Objectives and Methodology

This study employs a mixed-methods approach (Creswell, 2010) to comprehensively assess the quality of machine-translated Turkish stories into English through DeepL, integrating both human and automatic evaluation methods. Ethical approval for this study was obtained. For the automated evaluation, the study uses the BLEU metric, and human evaluation is conducted by both Gen Z translation trainees' and professional translators and the TAUS DQF served as the framework for human evaluation in this study. The DQF tools aim to standardize and enhance the evaluation process, promoting objectivity and transparency, making it the chosen framework for this study (Görög, 2014, p. 449). Projects were created for both of the MT outputs of literary narratives on TAUS DQF tools for the evaluators. The project type was chosen as "quality evaluation" and the evaluation types marked are adequacy and fluency. TAUS DQF tool provides numerical percentages that aid in the assessment of translation quality. In addition to utilizing TAUS DQF tool for manual evaluation, both trainees and professionals were requested to write reports on the detected errors and the key areas requiring more post-editing. A thematic analysis was conducted on the reports (Braun & Clarke, 2006). The linguistic similarity between the MT outputs and the human reference translations is measured using the BLEU metric. Trainees' and professionals' analyses using the DQF tool, together with their reports, were carefully examined and they were compared with the numeric BLEU results. Within this regard, this study aims to address the following research questions:

1. How effectively does the free version of DeepL handle the translation of Turkish short stories into English, and is it genuinely applicable?
2. How do Generation Z translation trainees and translation professionals evaluate the quality of machine translated literary texts using TAUS's evaluation metrics, and what are the similarities and differences in their evaluation results and reports?
3. Does the automatic evaluation of Turkish-English MT of literary texts align with the human evaluation of translation quality?

Participant Profiles

Gaining insights from both professional translators and Generation Z translation trainees, who have a natural inclination toward translation technologies, is valuable, especially as translation companies actively seek candidates with post-editing skills, prompting translator training institutions to integrate these competencies into their curriculum (Çetiner, 2021, p. 583). In line with this, seven undergraduate students were selected based on their successful completion of Information Technology for Translation I and II, taught by the author, demonstrating their foundational knowledge in MT and post-editing. While few had prior professional translation experience, they performed post-editing tasks on various texts during these courses. Over two semesters, they became familiar with TAUS's post-editing guidelines and error typology, applying them to their assignments. Additionally, three professional translators were selected based on their minimum of eight years of experience, all of whom graduated from a Translation Studies department. However, as NMT was not yet developed during their studies, they did not receive formal post-editing training. Before starting this project, they were briefed in a virtual meeting by the author on TAUS's guidelines and error annotation metrics. Prior to the experiment, both the students and the translators provided consent for participation. Participants received the raw translations and were given five days to complete their tasks at their own pace, with the flexibility to use external resources if needed. Ethical approval for this study was obtained from the Scientific Research and Publication Ethics Committee of Fatih Sultan Mehmet Vakıf University (Decision No. 314, dated July 6, 2023).

Materials Selected

To ensure a comprehensive evaluation process, short stories were intentionally selected so that they could be machine translated and post-edited completely, not partially. The short stories selected to be translated by free version of DeepL were *Mendil Altında* (Under the handkerchief) by Memduh Şevket Esendal and *Kabak Çekirdekçi* (Pumpkin seed seller) by Halide Edip Adıvar. These short stories were selected from *An Anthology of Turkish Short Stories*, a collection of Turkish short stories translated into English by Talat Sait Halman, who served as Turkey's Minister of Culture and translated several Turkish literary works into English. It is worth mentioning that to utilize the BLEU automatic metric, a reference translation was required, which is why these two stories were selected from Talat Sait Halman's collection, as his translations would be the reference ones for the automatic evaluation of MT.

Findings and Discussion

Trainees and professional translators' adequacy and fluency analyses on the DQF tool, as well as their reports, are comparatively examined in this section of the study. Subsequently, the automatic MT evaluation results, BLEU scores for each story translation, are presented and interpreted.

Adequacy and fluency analysis by trainees and professionals

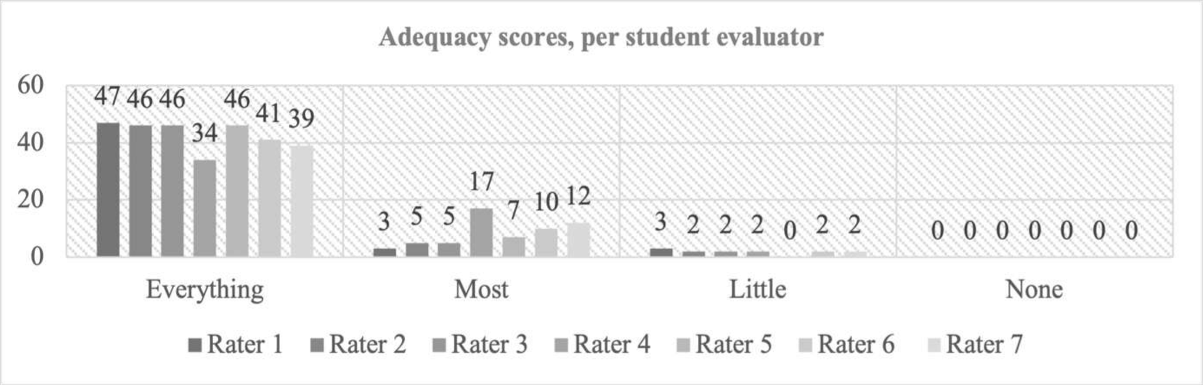
While adequacy is defined as "how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation", fluency is defined as to what extent the translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" (Görög, 2014, p. 161). The adequacy analysis categories, which encompass Everything, Most, Little, and None, evaluate the accuracy and suitability of translated content, while the fluency analysis, including Flawless, Good, Disfluent, and Incomprehensible, assess the naturalness and clarity of the translations.

Adequacy analysis of Mendil Altında by trainees

Seven trainees examined a total of 371 segments, with each trainee assessing 53 segments, to determine how closely the target translations reflected the meaning. The outcomes (Figure 1) showed that a total of 299 segments for all 7 students taken together had target translations that accurately captured every detail in the source text, highlighting that 80.59% of the segments were adequate. Additionally, 15.89% of the raw MT output conveyed the majority of the intended meaning from the source text, although not always perfectly. In a smaller subset of 3.51%, the target translations only conveyed a portion of the intended meaning, sometimes displaying notable omissions or alterations. Notably, 0 segments were discovered in which the target translations completely failed to convey any of the intended meaning, indicating an overall successful outcome.

Figure 1

Categorization of segments based on adequacy in the MT output of Mendil Altında by trainees



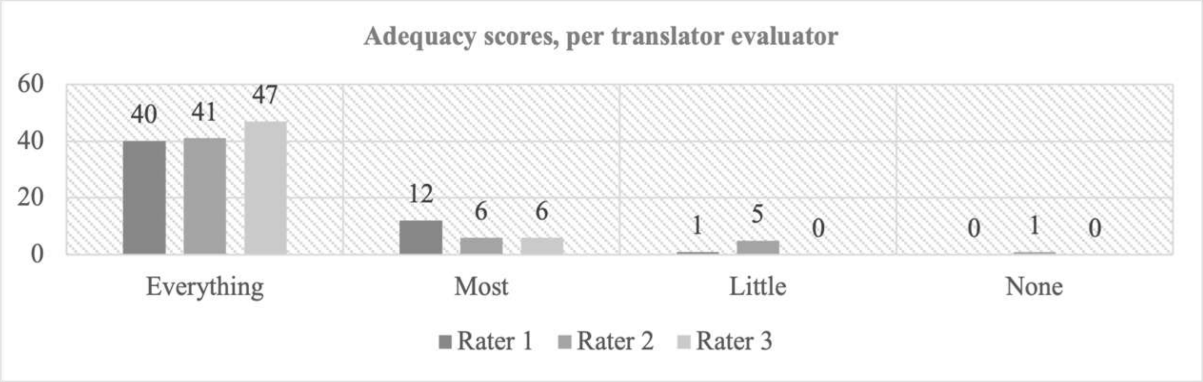
In the Everything category, a significant portion of the segments, ranging from 34 to 47, was assessed as fully preserving the meaning of the source text. Most evaluators were generally in agreement on this point. Evaluator ratings range from 3 to 17 segments in the Most category, indicating a high degree of interpretation variability. On the other hand, all these findings collectively demonstrate the strong adequacy of the MT output. The findings show that the categories of Little and None collectively represent a minority of the assessments. In the Little category, the counts vary slightly among the evaluators, with only 3, 2, 2, 2, 0, 2, and 2 segments falling into this category. Similarly, in the None category, no segments were rated as such by any of the seven evaluators.

Adequacy analysis of Mendil Altında by professionals

The findings of three translators' adequacy analysis of (Figure 2) demonstrate the machine translation's overall success in accurately conveying the source text's meaning. Out of the total 159 evaluated segments, a significant 128 segments (80.50%) received the Everything rating, suggesting that the translations were highly adequate. Furthermore, 15.09% of the segments were ranked in the Most category, indicating that although some segments were not flawless, they still represented a majority that was adequately translated with only minor issues. Furthermore, the Little and None categories, comprising 3.77% and 0.63% of the segments, respectively, collectively represent a minority, underscoring the overall effectiveness of the system in producing predominantly accurate translations.

Figure 2

Categorization of segments based on adequacy in the MT output of Mendil Altında by professionals



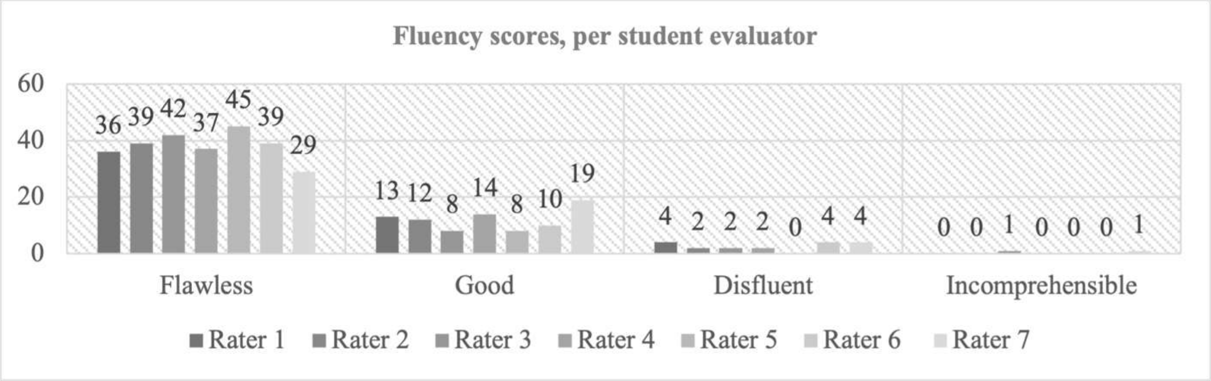
The number of segments in the Most category range from 6 to 12, meaning that although most translations were adequate, some small problems were noted. While the number of segments in the Little category range from 0 to 5, showing a variety of viewpoints on adequacy, the None category had very little representation, with only one segment, which suggest that segments completely lacking adequacy were quite rare.

Fluency analysis of Mendil Altında by trainees

371 segments in all were evaluated by seven raters (Figure 3) and 71.96% of the segments were classified as Flawless, indicating that a significant amount of the content had excellent fluency. With 22.61%, the Good category indicates that a considerable proportion of the passages maintained an acceptable level of fluency, despite not being flawless. A smaller subset of 4.85% fell into the Disfluent category. Reassuringly, only two segments were classified as Incomprehensible, indicating that the translations generally retained a high degree of fluency.

Figure 3

Categorization of segments based on fluency in the MT output of Mendil Altında by trainees



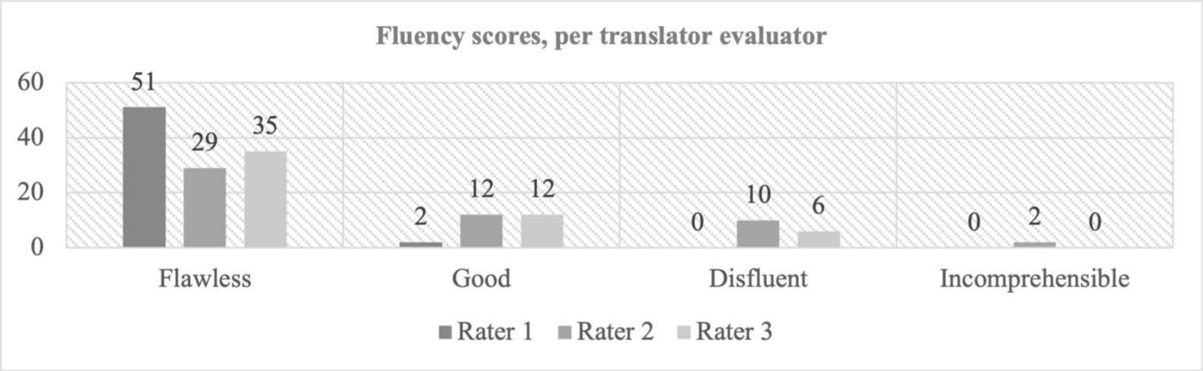
The number of segments classified under the Flawless category varied significantly, ranging from 29 to 45. This indicates that certain passages demonstrated particularly high fluency. This suggests that there were some very fluent passages in the text that satisfied the fluency requirements. The number of segments classified under the Good category ranged from 8 to 19, indicating that while the texts were generally fluent, some minor issues remained. On the other hand, the Disfluent and Incomprehensible categories had limited representation, with only a small percentage of segments classified as having fluency issues.

Fluency analysis of Mendil Altında by professionals

The analysis results (Figure 4) of the three professional translators showed that most of the segments, representing 72.32%, were rated as Flawless, indicating agreement on high fluency and natural language usage. An additional 16.35% of the segments were classified as Good, meaning that even though they weren't flawless, they still made up a sizable portion that were translated fluently with only a few minor problems. 10.06% of the segments fell into the Disfluent category, indicating a subset of translations that had obvious fluency problems. Just 1.26% of the segments fell into the Incomprehensible category, indicating an extremely small percentage of segments with particularly severe fluency problems.

Figure 4

Categorization of segments based on fluency in the MT output of Mendil Altında by professionals



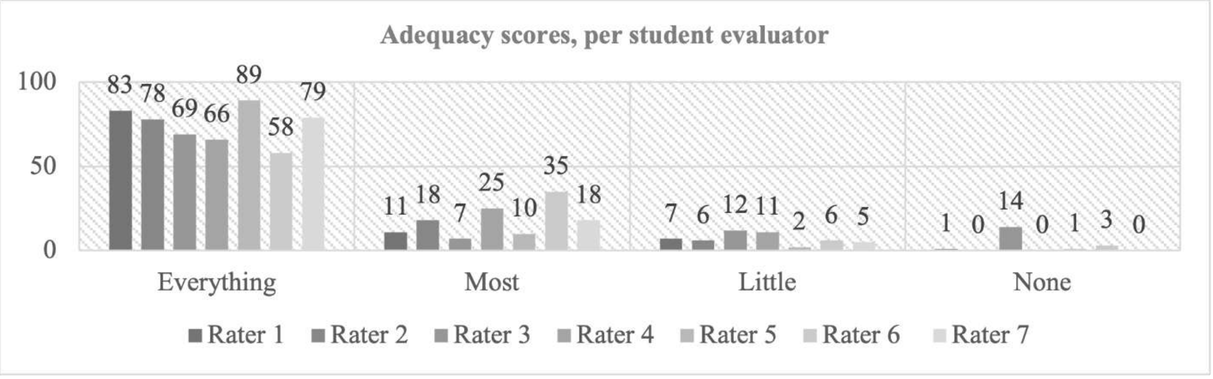
In the fluency analysis conducted by three different evaluators, a significant majority of the segments were classified as Flawless. Specifically, 51, 29, and 35 segments were identified as meeting the highest fluency standards. In the Good category, where segments were judged as reasonably fluent with minor issues, there were differing assessments, with 2, 12, and 12 segments. The Disfluent category received limited representation, with 10 segments rated as Disfluent by Rater 2 and 6 by Rater 3. In contrast, the Incomprehensible category had minimal representation, with only Rater 2 rating 2 segments.

Adequacy analysis of Kabak Çekirdekçi by trainees

Out of 714 segments, with each trainee assessing 102 segments (Figure 5), a substantial 73.08% were categorized as Everything, indicating their completeness in terms of adequacy, and highlighting that a sizeable amount of the machine-translated content was recognized for its high level of adequacy. Additionally, 17.36% of the segments were mostly accurate, indicating commendable adequacy. However, 6.86% of the segments were marked as having little adequacy and only 2.66% of the segments received a designation of no adequacy.

Figure 5

Categorization of segments based on adequacy in the MT output of Kabak Çekirdekçi by trainees



The evaluation results for the machine-translated story reveal varying degrees of adequacy as assessed by seven different raters. The number of segments classified in the Everything category ranges from 58 to 89, with Rater 5 identifying 89 segments in this category, the highest among all raters. This suggests that some translations were considered highly adequate. The Most category includes between 7 and 35 segments.

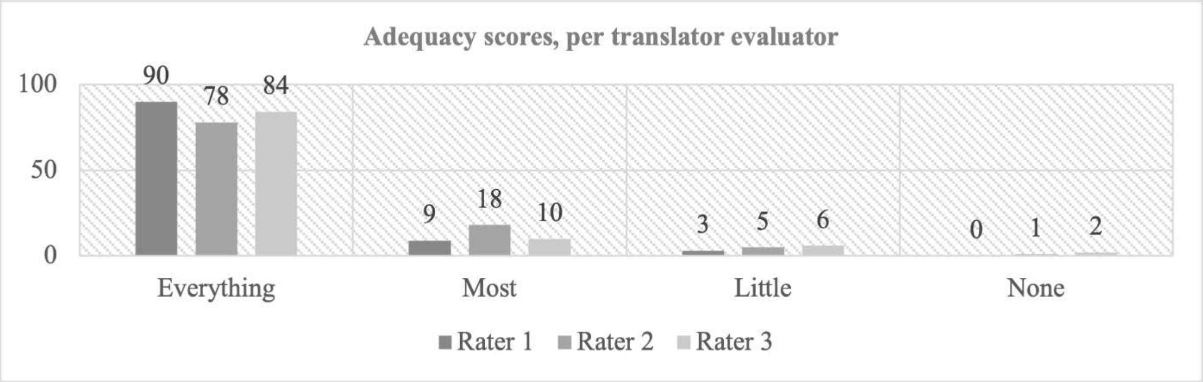
With the majority of the segments falling into the Everything category and receiving noticeably high ratings, it indicates that a sizable portion of the machine-translated content was considered to be highly adequate, with few to no errors in terms of adequacy, which is further supported by the lower rates in the 'little' and 'none' categories.

Adequacy analysis of Kabak Çekirdekçi by professionals

Three professional translators' examination of 306 segments in total (Figure 6) revealed that 82.35% of the machine translated segments translated the meaning of the source text, demonstrating the high adequacy of the MT output. Furthermore, 12.09% of the translations were classified as mostly adequate, implying that minimal post-editing is necessary to achieve a high degree of accuracy. Only 4.58% of the segments had limited adequacy, and 0.98% indicated no adequacy at all, necessitating extensive post-editing to increase accuracy, which are relatively small in comparison to the overall success of the MT system.

Figure 6

Categorization of segments based on adequacy in the MT output of Kabak Çekirdekçi by professionals



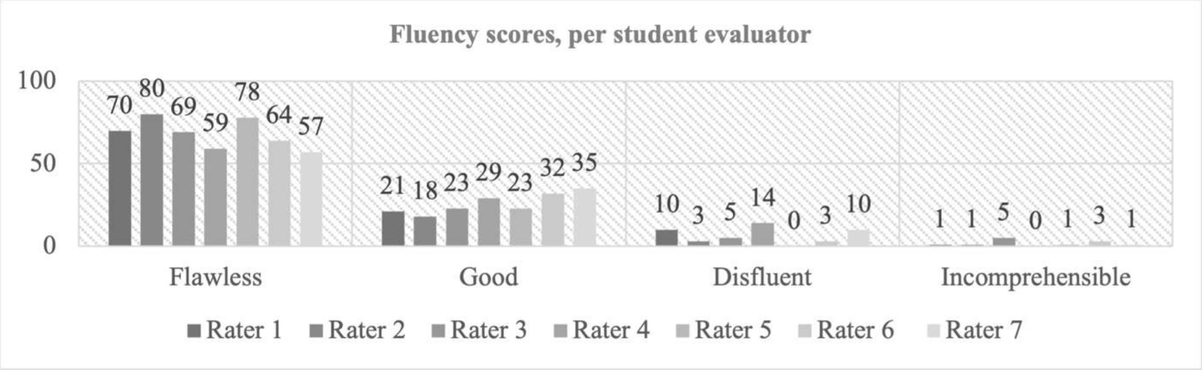
The three evaluators' adequacy scores indicate how well the evaluated MT translation performed. The translated segments were judged to be completely accurate in the Everything category, with 90, 78, and 84 segments classified under this category. This predominance in the Everything category demonstrate that a significant number of those segments were rendered accurately. Conversely, very few segments were scored in the Little category, which indicates only moderate accuracy, and even fewer in the None category.

Fluency analysis of Kabak Çekirdekçi by trainees

In the extensive evaluation of 714 segments (Figure 7) by seven different evaluators, notable patterns emerged in relation to fluency. Notably, 477 segments, representing 66.81% of the total, were designated as Flawless. This shows that a significant amount of the automatically translated content displayed a high level of fluency, characterized by seamless and coherent language flow. 25.35% of the segments, on the other hand, received a Good rating, indicating that many of them maintained a commendable level of fluency, despite a few minor hiccups. Additionally, 6.31% were labeled as Disfluent, indicating the existence of some segments with obvious fluency difficulties. Finally, only 1.54% of the segments were classified as Incomprehensible, showing that a small number of segments had serious problems with language comprehension.

Figure 7

Categorization of segments based on fluency in the MT output of Kabak Çekirdekçi by trainees



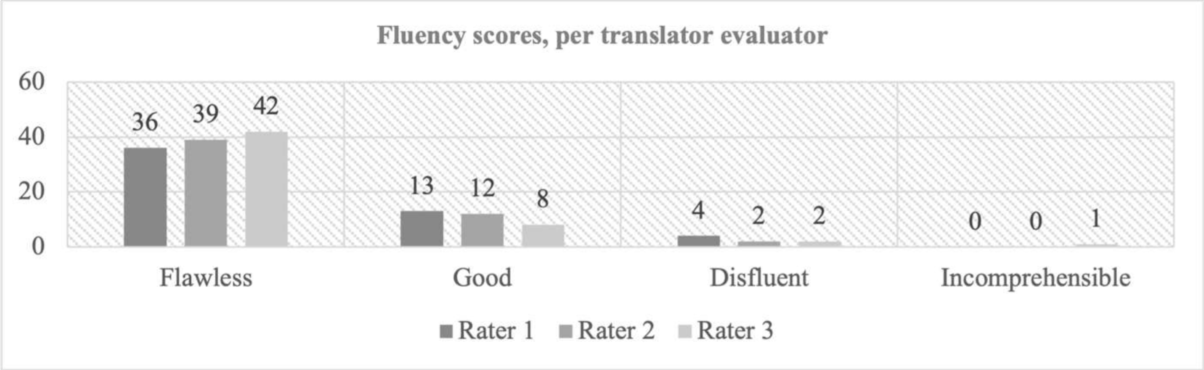
The number of segments classified under the Flawless category varied significantly, with Rater 2 identifying the highest count of 80 segments. This suggests that certain passages were considered particularly fluent. The average percentage for the Flawless category is 66.81% when all the provided ratings are considered, which highlights that a substantial portion of the text was deemed highly adequate by the raters.

Fluency analysis of Kabak Çekirdekçi by professionals

In the analysis of 306 segments focused on fluency by professional translators (Figure 8), 62.09% were marked as Flawless, demonstrating the effectiveness of MT in producing fluent language. 27.45% of the responses were rated as Good, suggesting fluency with only minor problems. Furthermore, 9.15% of the segments were classified as disfluent, and 1.31% were considered incomprehensible.

Figure 8

Categorization of segments based on fluency in the MT output of Kabak Çekirdekçi by professionals



The results of the fluency analysis, conducted by three different evaluators has shown that notably, the evaluators rated 36, 39, and 42 of the segments as Flawless, which represents the majority of the segments. The evaluators' agreement suggests that the translations achieved a natural language quality, demonstrating a high degree of fluency. The segments received lower counts in the Good category, with 13, 12, and 8, indicating a still-remarkable fluency with a few minor problems. On the contrary, a small percentage of the segments were classified as Disfluent or Incomprehensible, indicating that there may be problems with fluency that need to be addressed. The relatively lower percentages of disfluent and incomprehensible segments may indicate the effectiveness of the MT system in producing fluent translations, with only a small fraction exhibiting fluency issues.

Analysis of trainees' and professionals' reports on the MT output

After finishing the project utilizing TAUS DQF tool, the evaluators, both trainees and professional translators, were assigned to post-edit the machine-translated literary output. Their assignment not only included post-editing but also creating comprehensive reports documenting errors of the MT output and their individual post-editing process. A thematic analysis was conducted on the reports using the methodology described by

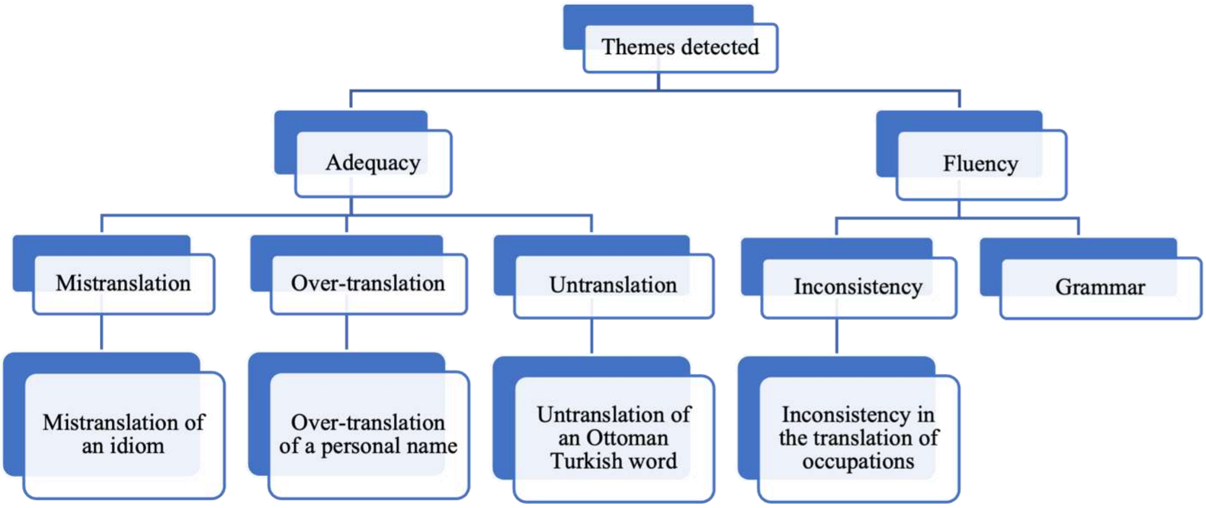
Braun and Clarke (2006). Initial codes were generated and then codes with similar content were combined. This led to the emergence of broad themes specific to each participant's post-editing process, especially the problems they detected. The next stage involved a detailed examination and improvement of these main themes along with the identification of subthemes. It is important to note that the themes detected from the reports highlight the problematic parts requiring post-editing and provide additional insights into the machine-translated text.

Analysis of trainees' reports on the MT output of Mendil Altında

Two major themes emerged from the analysis of the post-editing reports on Mendil Altında. It's important to note that the major themes, namely adequacy and fluency, correspond to the first two error categories in TAUS DQF error typology. However, the sub-themes of mistranslation, over-translation, undertranslation, inconsistency were independently formulated by the trainees through their analysis and post-editing of the MT output (Figure 9). Their analysis identified some specific examples, which have highlighted challenges within each subtheme. They mentioned that the MT rendered idioms literally. For instance, in the case of the idiom “para yetiştirmek” which conveys the meaning “to ensure having enough money to live on” MT translated it literally as “raising money”. Additionally, the subtheme of over-translation, particularly concerning personal names, unveiled examples where the translation exceeded the boundaries of necessity. As an illustration of overtranslation of a proper name, they mentioned that MT translated the Turkish name Meryem as Marry. Moreover, The Arabic word “mazbata” historically used, especially during the Ottoman Empire, to refer to “minutes,” remained untranslated by the MT, serving as an example of the untranslation of an Ottoman Turkish word. On the other hand, in terms of fluency issues inconsistency and grammatical problems were detected. As an example of inconsistency in the translation of occupations, the term “sicil memuru” which refers to the person responsible for ensuring the accuracy and proper maintenance of trade registry records, was translated by the MT as both “director” and “manager” interchangeably throughout the story.

Figure 9

Themes detected in the trainees' reports on the MT output of Mendil Altında

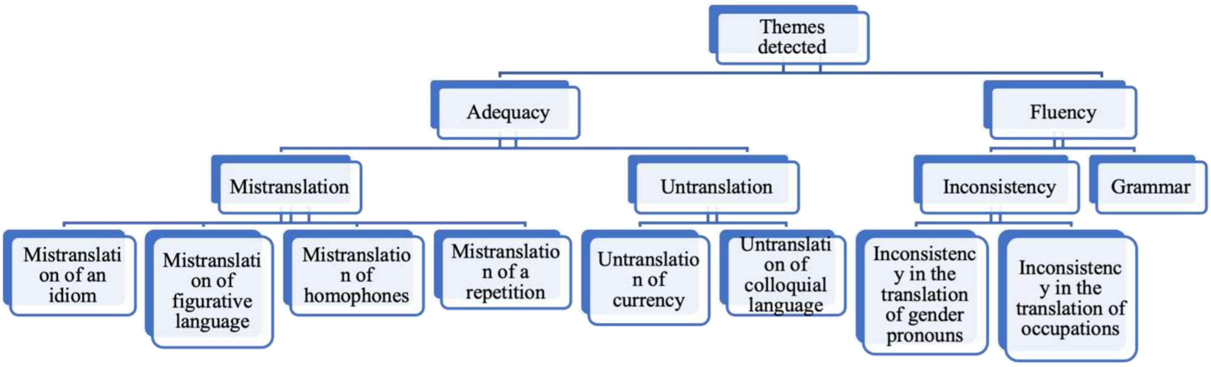

Themes detected
Adequacy
Fluency
Mistranslation
Over-translation
Untranslation
Inconsistency
Grammar
Mistranslation of an idiom
Over-translation of a personal name
Untranslation of an Ottoman Turkish word
Inconsistency in the translation of occupations

Analysis of trainees' reports on the MT output of Kabak Çekirdekçi

More themes were identified for the reports on the MT output of Kabak Çekirdekçi (Figure 10), highlighting that in this MT output there were more problematic parts necessitating post-editing. The theme mistranslation involves the trainees' examining occurrences of mistranslated metaphors, figurative language, homophones, slang, and repetitions in detail. As an example to mistranslation of a metaphor, the sentence of "gazeteye dayanamıyor", which aims to convey the pumpkin seed seller's strong attraction to reading newspapers, was given by the trainees. DeepL did not comprehend the metaphorical meaning of this sentence and translated it as "he can't stand newspapers". However, "can't stand" indicates a strong negative emotion or lack of patience with a particular thing or situation. It was reported that the machine's failure to grasp the metaphorical meaning resulted in a literal translation. Moreover, an instance of mistranslation in figurative language was detected in this sentence "Siyah gözleri eğlenip eğlenmediğimi anlak için yüzüme batıp çıkıyordu". In this sentence, the aunt explains that her nephew tried to understand whether she was having fun or not by examining her face. "yüzüne batıp çıkmak" was used figuratively to indicate that her nephew is closely observing her face. DeepL translated this sentence literally as "His black eyes darted in and out of my face to see if I was having fun" and couldn't render the figurative meaning in the source text. For the mistranslation of homophones, the trainees pointed the sentence of "altı olmayan kocaman düğmesiz iki potin", which emphasizes the poverty of the salesman by depicting the disintegrated shoe soles. However, In Turkish, “altı” is a homophone, encompassing both the meaning of the “number six” and indicating “being underneath” of something and in this context the narrator is talking about the underneath of the slaesman’s shoes. Although “altı” is used in the sense of “underneath,” in the source text, the MT rendered it as a number; “two huge, unbuttoned boots, maybe not six”, misinterpreting its intended meaning. Moreover, mistranslation of a repetition was also detected by the trainees, and an illustrative case was found in the phrase “siyah siyah gölgeleriyle”, which uses repetition for stress, a stylistic technique called epizeuxis, to vividly depict darkness of the shadows. However, DeepL translated this phrase as “black black shadows”, which is neither natural nor idiomatic in English. Two subthemes of untranslation were deduced from trainees’ reports: untranslation of currency and untranslation of colloquial language. They mentioned that “kuruş” referring to the Turkish currency worth one per cent of the lira, the official currency of Turkey, was not translated by the MT. As an instance of colloquial language left untranslated “kaabaak tazze tazze” can be given. In the story, the pumpkin seed seller calls out in the streets saying “kaabaak tazze tazze” to attract the attention of passers-by and advertise his pumpkin seeds on the streets. However, this phrase was not translated by the MT. Within the theme of fluency two significant sub-themes emerged, centered on inconsistency and grammar. Specifically, the subtheme of inconsistency explored the problems arising from gender pronouns and translating occupations. The characters in the story, Kabak Çekirdekçi, are an aunt and her niece. However, as the machine translation's handling of pronouns is inconsistent, it accidentally used “he” instead of “she”, which led to the MT mistranslating the girl as a “nephew”. Throughout the text, MT occasionally rendered the Arabic word “müsteşar” as “manager,” while at other times, it left the term untranslated.

Figure 10

Themes detected in the trainees’ reports on the MT output of Kabak Çekirdekçi

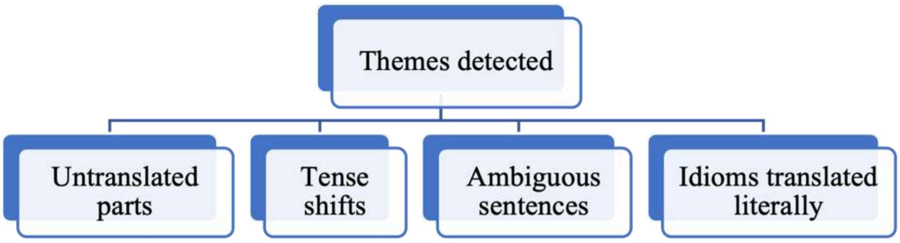

Themes detected
Adequacy
Fluency
Mistranslation
Untranslation
Inconsistency
Grammar
Mistranslation of an idiom
Mistranslation of figurative language
Mistranslation of homophones
Mistranslation of a repetition
Untranslation of currency
Untranslation of colloquial language
Inconsistency in the translation of gender pronouns
Inconsistency in the translation of occupations

Analysis of professionals' reports on the MT output of Mendil Altında

Different recurring themes emerged from the professional translators' post-editing reports (Figure 11). An apparent theme in the reports is the presence of untranslated parts. It was detected by the evaluators that some words such as "mazbata" and "bey" were not translated into English by the MT. Inconsistent use of tenses was mentioned as another critical concern, highlighting temporal inconsistencies in the translated material. However, tense differences were important for the plot of the story as it alternates between dream and real life, which is demonstrated by tense changes. Ambiguous sentences were also reported as problematic as they may lead to a lack of clarity or misinterpretation. Moreover, the literal translation of idiomatic expressions was recognized as a common issue, which may lead to the omission of cultural nuances and connotations and may even result in meaningless sentences. An example of this situation was explained as follow: "For instance, the Turkish idiom "ayaklarına kapanmak" which means "begging", was translated as "falls at the feet of the manager", losing its idiomatic meaning.".

Figure 11

Themes detected in the professionals' reports on the MT output of Mendil Altında

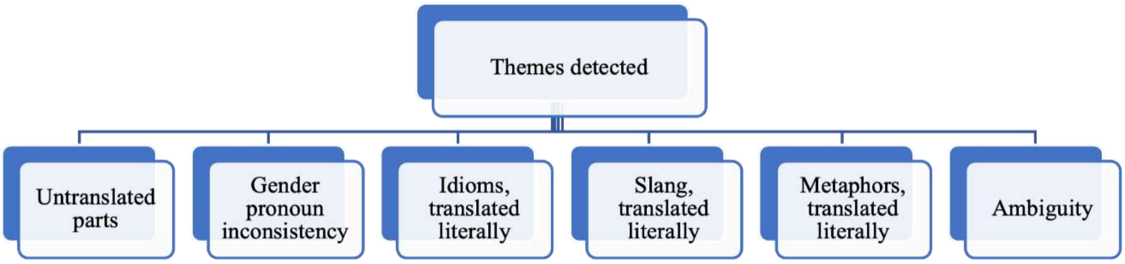


Analysis of professionals' reports on the MT output of Kabak Çekirdekçi

More themes detected in the reports (Figure 12) of Kabak Çekirdekçi, indicating that the MT output presented more challenges for post-editors than the story Mendil Altında. Untranslated parts posed a persistent issue in the MT output of Kabak Çekirdekçi as well. "Müsteşar", "yokuş" and "kuruş" were given as examples of untranslated words by the evaluators. Gender pronoun inconsistency emerged as another concern detected in the MT output. Moreover, the literal translation of idioms, slang, and metaphors was also reported to create significant problems, often resulting in the loss of cultural items and idiomatic expressions. Moreover, it was reported that "inconsistent translations of terms such as "kuruş" and "cent" introduced ambiguity in currency units, necessitating post-editing for a consistent approach to ensure clarity".

Figure 12

Themes detected in the professionals' reports on the MT output of Kabak Çekirdekçi


Themes detected
Untranslated parts
Gender pronoun inconsistency
Idioms, translated literally
Slang, translated literally
Metaphors, translated literally
Ambiguity

Automatic MT Evaluation by BLEU

While adequacy is defined as "how much of the meaning expressed in the gold-standard translation or the source… BLEU score was assessed through the Tilde platform's interactive BLEU score evaluator. As a reference point, translations with BLEU scores between 30 and 40 are considered "understandable to good", while scores between 40 and 50 indicate "high quality" translation. BLEU may face challenges when dealing with agglutinative languages such as Turkish, as even a small suffix added to a word, as opposed to the reference word, may result in a penalty within the BLEU score (Doğru, 2022). Thus, the analysis of the 1-gram segment of the BLEU outcome may provide a more comprehensive evaluation, especially when working with agglutinative languages such as Turkish (Ekinci, 2022). The cumulative BLEU score for the literary text, Mendil Altında, is 27.24 (Table 1). However, the cumulative score of 1-gram, which amounts to 69.47%, indicates a considerable level of precision when it comes to matching individual words between the machine-generated version and the reference translation.

Table 1

Results of automatic BLEU evaluation for Mendil Altında

Type	1-gram	2-gram	3-gram	4-gram
Individual	75.58	40.38	20.96	12.07
Cumulative	69.47	50.78	36.76	27.24

On the other hand, the cumulative BLEU score of 4-gram for Kabak Çekirdekçi is 37.56 (Table 2), which is under the category of "understandable to good", which falls into the "comprehensible to good" category. Moreover, the 1-gram cumulative match score for Kabak Çekirdekçi is 76.82, indicating a substantial match between individual words in the machine-generated output and the reference translation.

Table 2

Results of automatic BLEU evaluation for Kabak Çekirdekçi

Type	1-gram	2-gram	3-gram	4-gram
Individual	79.77	48.36	29.99	19.99
Cumulative	76.82	59.82	46.93	37.56

Discussion and Conclusion

In line with Chatzikoumi's (2020) suggestion that new studies should focus on specific linguistic phenomena within particular language pairs and challenging domains, this research employs a mixed-methods approach to comprehensively evaluate the quality of Turkish-to-English machine-translated short stories, integrating both human and automatic evaluation metrics. The evaluation of Mendil Altında using the TAUS DQF tool reveals a strong alignment in adequacy assessments between trainees and professionals, with accuracy rates of 80.59% and 80.50%, respectively. Similarly, fluency ratings were closely aligned, with trainees assessing 71.96% of segments as entirely fluent and professionals rating 72.32% as such. For Kabak Çekirdekçi, while trainees recorded a lower accuracy rate (73.08%) compared to professionals (82.35%), fluency ratings were more comparable, at 66.81% and 62.09%, respectively. This suggests that trainees may adopt a more selective approach when assessing accuracy. However, both groups demonstrated similar perspectives on fluency.

Thematic and document analysis of the reports indicated that trainees utilized meta-language, as reflected in the main themes identified - adequacy and fluency - which correspond to the main error categories of the TAUS's error typology. The second-level sub-themes identified in their reports also corresponded to the TAUS's error typology; however, the trainees themselves constituted the third-level sub-categories in terms of the issues identified in the MT output of the relevant story. In contrast, the professionals, despite receiving briefings on TAUS's error typology and using TAUS DQF tool for error annotation in the first part of this study, mostly refrained from using meta-language, including TAUS's error typology terminology, in their reports. Both trainees and professionals agreed that the MT of the narrative Kabak Çekirdekçi contained a higher number of errors, as the narrative was enriched with various literary devices such as colloquial language, repetitions, homophones, and metaphorical expressions.

Nevertheless, both groups were satisfied with the overall quality of the MT output for both texts.

Furthermore, the analysis indicates consistency between human and automatic MT evaluation, particularly in the 1-gram BLEU results with Under the Handkerchief scoring 69.47 and Kabak Çekirdekçi scoring 76.82. However, it is important to note that although the accuracy and fluency rates were rated higher for Mendil Altında by both trainees and professionals, there is a slight reverse difference in the BLEU scores, with Kabak Çekirdekçi performing better.

In conclusion, all the findings, both from trainees and professionals and BLEU, may be interpreted as indicating significant success of DeepL in the realm of literary MT between Turkish and English, especially in terms of short story translation. As illustrated by García (2014, pp. 430-436), students who are exposed to post-editing and MT evaluation metrics during their studies will be more skilled at time management, self-evaluation, and peer revision. Expanding on this, the current study suggests a subtle link between the introduction of these metrics and the cultivation of a broader skill set. The detailed analyses provided in the trainees' reports on the raw MT output emphasize their in-depth examination of the specific types of errors and their utilization of TAUS's error typology as a framework in their reports, despite not being specifically requested to do so, displayed their attention to detail and their application of a systematic approach in evaluating raw MT output. Furthermore, consistent with the analysis of both trainee and professionals' reports the study suggests that increased exposure to evaluation metrics could enhance the trainees' ability to identify subtle issues within the machine-translated output and to adopt a more systematic approach to post-editing. While BLEU has been criticized for assessing similarity to reference translations rather than accurately measuring the quality of translations (Castilho et al., 2018; Way, 2018), further research is required on specific language pairs, in conjunction with human evaluation, to provide insight into the effectiveness of automatic metrics. Such studies would offer valuable information regarding the relative efficacy of assessments conducted by humans and automated systems.

References

Aslan, E. (2024). Yapay zekâ destekli çeviri araçlarının edebi çevirideki yeterlilikleri üzerine karşılaştırmalı bir inceleme [A Comparative Study on the Adequacy of Artificial Intelligence-Assisted Translation Tools in Literary Translation]. *Istanbul University Journal of Translation Studies, 20*, 32-45. https://doi.org/10.26650/iujts.2024.1426435

Ayık Akça, T. (2022). Edebi metinlerde ve uzmanlık alan metinlerinde makine çevirisinin olanakları/olanaksızlığı: Çevirmenin değişen görev tanımlarına yeniden bakmak [The im/possibility of machine translation in literary and specialized texts: Rethinking translators' changing job descriptions]. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi,* (30), 1321-1343. https://doi.org/10.29000/rumelide.1188804

Adıvar, H. E. (1973). *Pumpkin seed seller*. In A. Alparslan (Ed.), *An anthology of Turkish short stories* (T. S. Halman, Trans.). RCD Cultural Institution.

Adıvar, H. E. (2001). *Gündelik adamlar: Kabak çekirdekçi* [Pumpkin Seed Seller]. In *Dağa çıkan kurt* [The Wolf on the Mountain] (pp. 33-38). Özgür Yayınları.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 6th International Conference on Learning Representations*. San Diego, USA. https://doi.org/10.48550/arXiv.1409.0473

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 257-267). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1025

Birkan Baydan, E. (2016). *Edebiyat çevirisinde sahneler ve aktörler* [The Scenes and Actors of Literary TranslationT]. Diye.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101. https://doi.org/10.1191/1478088706qp063oa

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249–256). Association for Computational Linguistics.

Castilho, S., O'Brien, S., Gaspari, F., Moorkens, J., & Way, A. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari, & A. Way (Eds.), *Translation quality assessment* (pp. 9-38). Springer. https://doi.org/10.1007/978-3-319-91241-7_2

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Chunyu, K., & Wong Tak-Ming, B. (2015). Evaluation in machine translation and computer-aided translation. In S.-W. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 213–237). Routledge.

Creswell, J. W. (2010). Mapping the developing landscape of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social and behavioral research* (2nd ed., pp. 45-68). Sage. https://doi.org/10.4135/9781506335193.n2

Çetiner, C. (2021). Sustainability of translation as a profession: Changing roles of translators in light of the developments in machine translation systems. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi, 9*, 575-586. https://doi.org/10.29000/rumelide.985014

Dallı, H., Dursun, O., Balal, Z., Hodjikj, E., Gürses, S., Güngör, T., & Şahin, M. (2024). Giving a translator's touch to the machine: Reproducing translator style in literary machine translation. *Palimpsestes, 38*, 15–56. https://doi.org/10.4000/12sp6

Dillinger, M. (2014). Introduction. In S. O'Brien, L. Winther Balling, M. Carl, et al. (Eds.), *Post-editing of machine translation: Processes and applications* (pp. ix–xv). Cambridge Scholars Publishing.

Doğru, G. (2022). Translation quality regarding low-resource, custom machine translations: A fine-grained comparative study on Turkish-to-English statistical and neural machine translation systems. *İstanbul Üniversitesi Çeviribilim Dergisi, 17*, 95–115. https://doi.org/10.26650/iujts.2022.1182687

Ekinci, S. (2022). *The effect of error annotation on post-editing effort and post-edited product: An experimental study on machine-translated subtitles of educational content* [unpublished Master's thesis]. 29 Mayıs University, Istanbul.

Esendal, M. Ş. (1973). *Under the handkerchief.* In A. Alparslan (Ed.), *An anthology of Turkish short stories* (T. S. Halman, Trans.). RCD Cultural Institution.

Esendal, M. Ş. (2023). *Mendil altında*. In *Mendil altında* (pp. 117–120). İletişim Yayıncılık.

García, I. (2014). Training quality evaluators. *Revista Tradumàtica, 12*, 430-436. https://doi.org/10.5565/rev/tradumatica.64

Giménez, J., & Màrquez, L. (2010). Asiya: An open toolkit for automatic machine translation (meta)evaluation. *The Prague Bulletin of Mathematical Linguistics, 94*, 77-86. https://doi.org/10.2478/v10108-010-0022-6

Gu, L. (2022). Translation of Japanese literature language and natural language environment understanding based on artificial neural network. *Journal of Environmental and Public Health, 22*, 1-12. https://doi.org/10.1155/2022/2015763

Guerberof Arenas, A., & Toral, A. (2022). CREAMT: Creativity and narrative engagement of literary texts translated by translators and NMT. In H. Moniz, L. Macken, A. Rufener, et al. (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 357–358). European Association for Machine Translation.

Guerberof-Arenas, A., & Moorkens, J. (2019). Machine translation and post-editing training as part of a master's programme. *JoSTrans: The Journal of Specialised Translation, 31*, 217-238. https://jostrans.soap2.ch/issue31/art_guerberof.php

Gürses, S., Şahin, M., Hodjikj, E., Güngör, T., Dallı, H., & Dursun, O. (2024). Çeviribilim çalışmalarında çevirmenin üslubu ve makinenin üslubu [The style of the translator and the style of the machine in translation studies]. *Çeviribilim ve Uygulamaları Dergisi, 36*, 100-124. https://doi.org/10.37599/ceviri.1468718

Hutchins, J. (1995). Machine translation: A brief history. In E. F. K. Koerner & R. E. Asher (Eds.), *Concise history of the language sciences: From the Sumerians to the cognitivists* (pp. 431–445). Pergamon Press. https://doi.org/10.1016/B978-0-08-042580-1.50066-0

Hutchins, J. (2015). Machine translation: History of research and applications. In S.-W. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 120–137). Routledge.

Jiang, Y., & Niu, J. (2022). How are neural machine-translated Chinese-to-English short stories constructed and cohered? An exploratory study based on theme-rheme structure. *Lingua, 273*, 103318. https://doi.org/10.1016/j.lingua.2022.103318

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. In M. Cettolo, J. Niehues, S. Stüker, et al. (Eds.), *Proceedings of the 9th International Workshop on Spoken Language Translation*. International Workshop on Spoken Language Translation.

Kenny, D., & Doherty, S. (2014). Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer, 8*(2), 276–294. https://doi.org/10.1080/1750399X.2014.936112

Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics, 108*, 121-132. https://doi.org/10.1515/pralin-2017-0014

Mah, S.-H. (2020). Defining language-dependent post-editing guidelines for specific content: The case of the English-Korean pair to improve literature machine translation styles. *Babel, 66*(4–5), 811–828. https://doi.org/10.1075/babel.00174.mah

Melby, A. K. (2020). Future of machine translation: Musings on Weaver's memo. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 419–436). Routledge. https://doi.org/10.4324/9781315311258-25

Miller, G. A., & Beebe-Center, J. G. (1956). Some psychological methods for evaluating the quality of translations. *Mechanical Translation, 3*(3), 73–80.

O'Brien, S., Winther Balling, L., & Carl, M., et al. (2014). Foreword. In S. O'Brien, L. Winther Balling, M. Carl, et al. (Eds.), *Post-editing of machine translation: Processes and applications*. Cambridge Scholars Publishing.

O'Hagan, M. (2020). Translation and technology: Disruptive entanglement of human and machine. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 26–59). Routledge.

Öner Bulut, S. (2019). Integrating machine translation into translator training: Towards 'human translator competence'? *TransLogos Translation Studies Journal, 2*(2), 1–26. https://doi.org/10.29228/transLogos.11

Öner Bulut, S., & Alimen, N. (2023). Translator education as a collaborative quest for insights into the re-positioning of the human translator (educator) in the age of machine translation: The results of a learning experiment. *The Interpreter and Translator Trainer, 17*(3), 375–392. https://doi.org/10.1080/1750399X.2023.2237837

Papineni, K., Roukos, S., & Ward, T., et al. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Philadelphia. https://doi.org/10.3115/1073083.1073135

Poibeau, T. (2017). *Machine translation*. MIT Press Essential Knowledge series. https://doi.org/10.7551/mitpress/11043.001.0001

Quah, C. K. (2006). *Translation and technology*. Palgrave Macmillan. https://doi.org/10.1057/9780230287105

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation, 32*(3), 217-235. https://doi.org/10.1007/s10590-018-9220-z

Sin-Wai, C. (2015). The development of translation technology. In S.-W. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 3–32). Routledge.

Smith, A., Hardmeier, C., & Tiedemann, J. (2016). Climbing Mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2017)* (pp. 269–281). European Association for Machine Translation.

Şahin, M., & Gürses, S. (2021). English-Turkish literary translation through human-machine interaction. *Tradumàtica: Tecnologies de la Traducció, 19*, 171–203. https://doi.org/10.5565/rev/tradumatica.284

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

Trojszczak, M. (2022). Translator training meets machine translation - Selected challenges. In *Language use, education, and professional contexts* (pp. 179-192). Springer International Publishing. https://doi.org/10.1007/978-3-030-96095-7_11

Tymoczko, M. (2014). Why literary translation is a good model for translation theory and practice. In J. Boase-Beier, A. Fawcett, & P. Wilson (Eds.), *Literary translation: Redrawing the boundaries* (pp. 11–31). Palgrave Macmillan. https://doi.org/10.1057/9781137310057_2

Wang, H., Wu, H., He, Z., et al. (2022). Progress in machine translation. *Engineering, 18*, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

Way, A. (2018). Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari, et al. (Eds.), *Translation quality assessment* (pp. 159–178). Springer. https://doi.org/10.1007/978-3-319-91241-7_8

Webster, R., Fonteyne, M., Tezcan, A., et al. (2020). Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics, 7*(32). https://doi.org/10.3390/informatics7030032

Yang, L., & Min, Z. (2015). Statistical machine translation. In S.-W. Chan (Ed.), *The Routledge encyclopedia of translation technology* (pp. 201–213). Routledge.

Reviewers:
1. Anonymous
2. Anonymous

Handling Editor:
Boris Naimushin, PhD
New Bulgarian University