Human Translation versus Post-Editing of Machine-Translated Figurative Expressions: A Pilot Study

Ogareet Khoury[1]

[1] Al-Ahliyya Amman University, Jordan

Correspondence: Ogareet Khoury, Al-Ahliyya Amman University, Jordan.

Received: September 1, 2025 Accepted: January 14, 2026 Online Published: March 13, 2026

doi:10.5430/wjel.v16n4p169 URL: https://doi.org/10.5430/wjel.v16n4p169

Abstract

This research paper presents a pilot study of temporal effort and translation quality in human translation (HT) and machine translation post-editing (MTPE) by trainee translators. It integrates a process- and product-oriented approach comparing HT with MTPE with respect to efficiency (translation process) and effectiveness (translation product) in translating figurative expressions from Arabic into English. Ten senior translation students were selected to perform the two tasks in different sessions. In the HT task, participants were asked to translate Arabic figurative expressions into English without the assistance of machine translation (MT). In the MTPE task, they were asked to post-edit comparable English figurative expressions from the same source text that had been translated by Google Neural Machine Translation (GNMT). The Arabic novella, *Returning to Haifa,* by Ghassan Kanafani, served as the source text from which the figurative expressions were extracted. Its published English translation by Harlow and Riley (2000) was used as the reference translation. The study employs a mixed quantitative-qualitative approach. For observing the process, the keylogging software, *Translog II*, was used to track the time of the tasks. The human-translated and post-edited texts were evaluated by human raters using the Multi-Dimensional Quality Metrics (MQM). The study primarily investigates HT versus MTPE while including the BLEU score of the initial unedited machine translation output. Findings reveal that MTPE improved processing speed, while HT produced better quality. Thus, the study raises certain concerns about MTPE of figurative expressions, which should be considered when designing courses in post-editing of literary discourse.

Keywords: BLEU, figurative language, MQM, machine translation, post-editing, Translog-II

1. Introduction

1.1 Translation Process and Product Research

Empirical translation studies can be approached from two main perspectives: process-oriented and product-oriented. Process-oriented research tends to investigate the efficiency of the translation by using methods such as human observation, keystroke logging, and eye tracking (Carl, 2003). Product-oriented research, on the other hand, tends to describe, explain, and evaluate the linguistic, textual, and functional characteristics of the target texts in relation to the source text. Before the emergence of automated assessment tools in the late 1990s, translation product assessment relied on human judgement using methods such as linguistic error typology. The four basic evaluation criteria have been identified as language fluency, translation accuracy, readability, and acceptability of the target text (Newmark, 1988; White, 1995; Koehn & Monz, 2006; Lommel, Brümmer & Uszkoreit, 2014). Accuracy in translation refers to preserving the semantic content of the source text, as faithfully as possible, in the target text (Williams, 2004). Acceptability, on the other hand, indicates the extent to which the target text reads as a natural text and conforms to the norms and expectations of the target culture (Toury, 1995).

1.2 Tracking and Assessment Tools

Translog II keylogging software has been widely used in empirical process-oriented studies. It records the translation activity in detail, capturing keystrokes, typing pauses, and time spent on tasks, providing insights into temporal and cognitive effort. It includes two components: *Translog II* Supervisor, for configuring experiments and analyzing data, and *Translog II* User, for conducting translation tasks.

Automated translation assessment tools have emerged to allow the assessment of MT output. These include Bilingual Assessment Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Translation Edit Rate (TER). BLEU, one of the earliest and most influential tools, was introduced by Papineni et al. (2002). Since its introduction, it has been widely used in product assessment studies in different language pairs including Arabic and English (e.g. Habib et al., 2025). It evaluates a candidate translation by comparing it to a reference translation.

The Multidimensional Quality Metrics (MQM) is a translation evaluation framework that was developed by the EU-funded *QTLaunchPad* project, under the direction of the German Research Center for Artificial Intelligence (Lommel, Brümmer & Uszkoreit, 2014). It is maintained by the MQM Council and is widely used in empirical studies that involve human evaluation of translations. MQM is employed to assess a given translation by breaking it down into specific error typologies (e.g. language fluency, translation accuracy, naturalness and acceptability of the target text), assigning severity levels to them (minor, major, or critical errors). According to Lommel, Brümmer, and Uszkoreit (2014), minor errors are those errors which have a limited effect on accuracy and fluency without impeding understandability of the content whereas major errors seriously affect understandability. Critical errors, on the other hand, were identified as those errors that render the target text unfit for its purpose.

1.3 Translating Figurative Language

Figurative expressions are linguistic constructions in which speakers use non-literal language to convey abstract, emotional, and complex meaning. They encompass devices including metaphor, simile, personification, hyperbole, and idioms, conveying meanings that cannot be deduced from the individual words themselves (Abrams, 1999). Some empirical studies in corpus linguistics have revealed how these expressions are used in our everyday communication (e.g. Abu Manie et al., 2025).

The translation of figurative language is widely regarded as complex due to its reliance on non-literal meaning and cultural nuance (Venuti, 2008). It requires accurate meaning while creating a comparable effect in the target culture (Bassnett & Lefevere, 1998). Translators are required in such cases to act as active mediators to bridge any communication gaps (e.g. Almommani, 2024).

Several studies showed that despite the recent developments, neural machine translation has undergone, it remains limited when used in the translation of stylistically and culturally rich genres for its literal renderings (Matusov, 2019; Guerberof-Arenas & Toral, 2022; Dankers, Lucas & Titov, 2022). Due to the inherent deficiencies of MT, MTPE began to be widely adopted in the translation industry while simultaneously becoming a focus of empirical research, investigating its efficiency and effectiveness (e.g. Toral & Way, 2018; Gaspari, Naskar & Way, 2014). This paper is structured as follows: it begins with the theoretical framework, followed by a review of related literature, the statement of the problem and the methodology; it then presents the analysis and discussion of the findings. The study is then concluded with the conclusion, limitations and recommendations.

2. Theoretical Framework

The present pilot study adopts the Dynamic Quality Framework (DQF) developed by the Translation Automation User Society (TAUS) as its primary theoretical framework. It is a holistic approach to evaluating translation performance by integrating time and quality. It was developed to guide translation quality assessment in machine-translation contexts. DQF supports the combination of product and process studies as it seeks to investigate efficiency (time) and effectiveness (quality) which serves the purpose of this study. Within DQF, time efficiency can be tracked through keylogging. Quality, on the other hand, can be examined by employing models such as the Multi-Dimensional Quality Metrics (MQM) for human-generated translations and automated tools for machine-generated texts. The basic evaluation criteria were identified as accuracy, fluency, and acceptability for the target language and culture (e.g. Castilho & O'Brien, 2017; Fiederer & O'Brien, 2009). By framing the present research within DQF, the study captures the dynamic interplay between HT and MTPE in terms of temporal effort and the quality of the outcome in light of the initial GNMT output.

3. Review of Related Literature

3.1 Figurative Language Translation

Early theorists such as Nida (1964) and Dagut (1976) stated that figurative expressions are generally untranslatable due to their deep linguistic and cultural roots. Van den Broeck (1981) and Newmark (1988) pointed out that figurative language translation is certainly possible but inherently difficult. This assumption was later challenged by scholars such as Reiss (2000) who acknowledged that while culture-bound expressions may pose difficulties, they can still be rendered through linguistic and cultural equivalents.

Some translation strategies were suggested for figurative expressions. Literal translation was advocated when the linguistic and cultural norms of the target language permit it, suggesting that translators can resort to literal translation as a first option, adjusting and explicating the target text when cultural incompatibility occurs as much as the target language system allows (Youssef & Albarakati, 2023; Park, 2009). Similarly, Dagnev and Chervenkova (2020) revealed that literal translation of several figurative expressions from English into Bulgarian was widely acceptable.

On the other side of the spectrum, several scholars such as Venuti (1995) and Baker (2018) argue that literal translation fails to convey the cultural and emotional depth embedded in literary texts, advocating a context-based translation strategy. Even with recent developments of NMT, it is still characterized by a tendency toward literal translation, especially for figurative expressions that are naturally culture-bound (Naveen & Trojovsky, 2024). Thus, MTPE remains an essential step for literary MT output and it is still worth investigating whether MTPE can fine-tune this inherently literal output. Section 3.2 below examines the empirical research conducted on MTPE of literary figurative language.

3.2 Post-editing Figurative Language

Machine Translation (MT) has significantly enhanced speed and efficiency in various translation domains (Hartley, 2009; Plitt & Masselot, 2010; Martin & Serra, 2014). Despite its industrial success, literary texts remain a major challenge for machine translation, particularly when it comes to creative and figurative language (Guerberof-Arenas & Toral, 2022; Castilho & Resende, 2022; Afzal, 2018). In the translation industry, professional literary translators remain skeptical of relying completely on MT for its limitations in handling figurative language (Matusov, 2019). There has been an argument that with the development of Neural Machine Translation (NMT), translating figurative language is becoming more feasible provided that it is followed by MTPE (Dorst, 2023). Within the Arabic-English context, the empirical study conducted by Alsharif and Khasawneh (2017) found that MT systems relied on literal renderings struggling in capturing the cultural nuances of figurative language.

Jia and Lai (2022) conducted a study on MTPE of metaphor-rich English texts machine-translated into Chinese. Findings showed that

MTPE significantly improved translators' productivity and accuracy compared with translating from scratch. Editors predominantly relied on direct transfer; a strategy that may overlook deeper semantic subtleties. Similarly, Koglin (2015) found that MTPE of metaphors required less cognitive effort than manual translation, evidenced by reduced fixation durations and fewer keystroke pauses, though it produced less creative texts. Toral, Sprugnoli and Gervás (2018) empirically assessed MTPE of rule-based and phrase-based machine translation for novels comparing it to human translations. The study unveiled that productivity increased by 36% in MTPE compared with human translations produced from scratch, with final text quality closer to human translation than to raw unedited machine translation. Kuzman, Vintar, and Arčan (2019) evaluated NMT for literary texts from English to Slovene, using BLEU and METEOR, and found that MTPE achieved both efficiency and effectiveness. Guerberof-Arenas and Toral (2022) investigated creativity in an empirical comparative study of HT versus MTPE. They concluded that, in translating fiction from English into Catalan, HT scored significantly higher than MTPE on creativity, narrative engagement, and reader reception. A recent study by Castilho and Resende (2022) using the MultEval assessment tool found that MTPE of literary texts was more similar to raw MT outputs than to human translations produced from scratch, suggesting a compromise of creativity. Besacier and Schwartz (2015) investigated phrase-based MT for English–French literary translation, revealing that post-editing improved readability and acceptability but not creativity. Toral and Way (2018), who evaluated different NMT systems trained on English-to-Catalan literary data, concluded that human intervention remains crucial in preserving creativity and textual nuance of literary discourse.

Review of empirical comparative studies on MTPE versus HT generally reveals that unedited MT of literary figurative language is likely to fail in maintaining the emotionality and creativity of figurative language. Research on MTPE of literary figurative language proved empirically that while MTPE saves time and effort in the translation process, it also struggles with this culture-specific domain as it tends to flatten the depth and creative resonance of figurative expressions when compared to HT. The tendency of MT toward overly literal translation of figurative language is increased when the source and target languages are linguistically and culturally remote, as is the case with Arabic and English (Tawfiq, 2015).

To the best of the author's knowledge, the existing body of literature lacks empirical studies that directly compare MTPE with HT in the context of literary texts from Arabic into English. Therefore, the present study aims to explore how a highly emotive text such as *Returning to Haifa* by Kanafani can be handled in HT versus MTPE. By focusing on BA students, this study not only fills a gap in Arabic-English literary translation research but also provides pedagogical insights that are directly relevant to translator training.

3. Statement of the Problem

Despite the growing interest in investigating MTPE for literary figurative language, research in this area remains in its infancy, especially for pedagogical implications. Thus, further studies are needed to compare HT versus MTPE from two main aspects: time and quality. This should contribute to research in this area by providing insights into literary translation in general and post-editing MT-generated literary output in particular.

This study seeks to answer two research questions:

1. What is the temporal effort required for each task at both the individual and group levels?
2. Which method, human translation (HT) or machine translation post-editing (MTPE) produces higher quality translations according to MQM evaluation?

4. Methodology

To address the research questions, the study employed a mixed-methods approach combining quantitative and qualitative analyses. Time-tracking was performed with *Translog II*, while quality assessment was conducted by human raters using the MQM framework. Additionally, the initial MT output was assessed using the BLEU metric to provide a baseline assessment of the initial MT output before post-editing.

4.1 Sampling Procedure

The study employed a within-subjects design in which ten senior translation students were selected to perform HT and MTPE tasks. All participants were native speakers of Arabic with English as their first foreign language. They were considered to have comparable translation skills; given that they were in the final term of a four-year translator training program. They had completed all practical courses in general and specialized translation, and were enrolled in an on-the-job training course. Participants were recruited on a voluntary basis and provided written informed consent. They were provided with full instructions for the tasks and oriented to the nature of the tasks, the tracking and assessment tools, access to the source text and CSV sheet, as well as the purpose of the study.

The independent variable was the type of task (HT and MTPE), while the dependent variables were the time spent on each task and the translation quality.

To ensure internal validity, several factors were controlled: 1) The academic and linguistic background of the participants; 2) the experiment's keylogging tool and translation environment (identical lab conditions, software access, sources), 3) task uniformity (comparable figurative expressions and task instructions).

Machine translation websites were blocked; however, access to online dictionaries was allowed. Gender was not considered a variable in the analysis. The inclusion of both male and female students was only to mirror natural classroom settings, not intended to investigate gender differences.

The participants were monitored during the tasks by the researcher and a lab supervisor. Although the translation tasks were not time-bound, translation time was recorded and compared between tasks. To control the order effects of tasks and protect internal validity, five participants performed the HT task first followed by the MTPE task, while the other five performed the MTPE task first followed by the HT task.

4.2 Experiment Setup & Assessment Tools

Kanafani's Arabic novella *Returning to Haifa* was chosen as the source text for the study. Its published English translation by Harlow and Riley (2000) was used as a reference translation to assess the candidate translations. This English translation has been widely praised for preserving the emotional and stylistic power of the original Arabic (e.g., Bamia, 2002). Based on Gibbs' (1994) definition of figurative expressions, a total of 65 expressions were identified in the 76-page novella (see Appendix A).   Most of these were metaphors and idioms. The rest consisted of similes and instances of personification. From the 65 identified expressions, 40 were selected for use in the tasks. They were evenly divided into two sets (20 each), with comparable lengths and a balanced distribution of metaphors, similes, personifications, and idioms. For each task, the expressions were entered into the User Mode of *Translog II* to be processed by the participants. For the HT task, Arabic expressions were inserted, as they appear in the source text, to be translated into English. For the MTPE task, the participants were asked to post-edit the 20 GNMT-generated English expressions. Google neural machine translation was chosen because it is widely used in Arabic-English pedagogical contexts (Alhaisoni & Alhaysoni, 2017)

To ensure precise reference to the source text, a comma-separated value sheet was prepared (CSV). It was serially and systematically annotated with page, paragraph, and sentence numbers. The Excel-prepared CSV was made available to the participants for reference to the contextual meaning of their assigned expressions. It served as the master list for all subsequent procedures, including task preparation, rater evaluation, and data analysis.

Translog II is a keylogging software that tracks and records time, keyboard, and mouse behavior during translation. In this study, it was used to track time, calculating total seconds taken per target word (STW), as well as overall time spent on each task. Data for *Translog II* were prepared in plain UTF-8 text format (translog_input.txt) so that the software could load the translation material for the task.

BLEU is a quantitative metric that calculates the degree of similarity between a candidate translation and a reference translation by computing n-gram overlap. Scores range from 0 to 100 (or 0.0 to 1.0), with 100 indicating a perfect match. The BLEU score is calculated using the following formula: ∑_{n=1} ^N *Wn* log *Pn*, where *Pn* is the modified precision for n-grams, *Wn* is the weight for each n-gram, and *BP* is the brevity penalty applied if the candidate translation is shorter than the reference translation. Each line in the candidate file was aligned with its counterpart in the reference file. The study used TILDE Interactive Evaluator, a standardized web-based BLEU assessment tool, to compute corpus-level BLEU scores of the initial GNMT output, providing a baseline quality assessment before any post-editing. The text was prepared as plain UTF-8 text (bleu_reference.txt).

For human evaluation, two experienced translation teachers were asked to rate the ten HT texts and the ten MTPE texts within the Multi-Dimensional Quality Metrics (MQM) framework with which they are well acquainted. Within MQM, translation quality is assessed by decomposing the target texts into error typologies and standardized severity levels. As the participants of the present study were translation students rather than professionals, a simplified-reduced severity MQM was used (annotating only minor and major errors). For focused and targeted evaluation, several empirical studies employed the simplified MQM-based evaluation with binary error severity (e.g., Koehn and Monz, 2006; Castilho et al., 2017). Errors were categorized within four main categories that serve the purpose of the present study. These were: handling of figurative language, accuracy of translation, choice of terminology and fluency of language with two levels of error severity (minor and major).

For data analysis, the latest version of *IBM SPSS Statistics* (version *31.0.1.0*) was used to produce descriptive statistics and run nonparametric tests (e.g., Wilcoxon signed-rank tests) as well as Cohen's Kappa reliability tests. Results extracted from *Translog II,* BLEU, and the human raters are discussed in Section 5.

5. Analysis

The performance of HT versus MTPE revealed distinct patterns in terms of translation quality (effectiveness) and task duration (efficiency). In Section 5.1, the MQM evaluation by the two raters of the human-translated and post-edited texts is presented. In addition, BLEU assessment of the GNMT output is reported to provide an automated assessment prior to post-editing in the MTPE task. Section 5.2 presents and discusses the temporal effort of the HT and MTPE tasks as tracked and recorded by *Translog II*.

5.1 Translation Quality

Each target text (HT and MTPE) produced by each participant was evaluated by annotating major errors (-5 points) and minor errors (-1 point) across the four categories: figurative language, accuracy, terminology, and fluency. In the MQM framework, penalties for error severity levels are typically weighted such that major errors carry a greater penalty than minor errors. The MQM common scheme assigns -5 for major errors and -1 for minor errors as adopted in some empirical studies (e.g., Lu et al., 2025; Feng et al., 2025) with an overall rating out of 100.

Table 1 presents the ratings out of 100, as annotated by each rater in each task for the ten participants.

Table 1. HT and MTPE scores by two raters (out of 100)

Participants	HT (R1)	MTPE (R1)	HT (R2)	MTPE (R2)
P1	72	62	72	59
P2	65	59	59	53
P3	63	53	55	51
P4	64	54	61	50
P5	66	49	63	49
P6	65	57	60	51
P7	56	43	54	44
P8	69	60	71	56
P9	74	64	73	57
P10	78	65	77	63

To ensure inter-rater reliability of the two raters' assessments, quadratic weighted Cohen's Kappa was run in SPSS. The translation scores were categorized into three ordinal levels (High > 70, Medium: scores 60-69, and low < 60). Reporting the weighted Cohen's Kappa is appropriate here because the translation scores are ordinal. This means that the difference between categories (low to medium versus medium to high) is meaningful. Weighted Cohen's Kappa accounts for the degree of disagreement between the raters' assessments, giving partial credit when ratings are close rather than treating all disagreements as equal. According to the common interpretation of Cohen's Kappa quadratic weighted test, a substantial agreement is reflected if k = 0.61-0.80 while a score of > 0.80 signifies strong agreement. The Cohen's Kappa was k = 0.81 (strong agreement) for the HT assessments and k = 0.76 (substantial to strong agreement) for the MTPE assessments. These values indicate that the two raters were generally consistent in their assessments.

The scores in Table 1 reveal that the HT assessments by the two raters tend to be higher than the MTPE assessments for most participants. For instance, participants P1, P8, P9 and P10 fall within the high range category (> 70) whereas their MTPE scores mostly fall within the medium and low range categories (55-65). The trend shows that most participants' assessments were within medium-to-high scores in the HT assessments, but medium-to-low scores in the MTPE assessments

A Wilcoxon signed-rank test for the averaged HT and MTPE scores was run. It indicates that all ranks are positive (HT > MTPE). Given that all differences are positive; it is highly significant ($p < 0.01$) indicating that HT scores are consistently higher than MTPE scores with a meaningful effect size ($r > 0.5$). Table 2 presents the mean, median, and standard deviation for the HT assessment scores and their MTPE counterparts.

Table 2. Mean, Median, and SD of HT and MTPE scores

Measure	Mean	Median	SD (Sample)
HT (R1)	67.2	65.5	6.27
MTPE (R1)	56.6	58	6.95
HT (R2)	64.5	62	8.11
MTPE (R2)	53.3	52	5.52
HT Average = 65.85			
MTPE Average = 54.95			

According to MQM, raters annotated errors in the target texts as minor errors (-1 point) and major errors (-5 points) across the four categories: figurative language, accuracy, terminology, and fluency. The following table shows the total minor and major errors committed by the participants in the two tasks at the individual and aggregate levels.

Table 3. Total errors in HT by rater and participant

Raters	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
HT Rater 1	16	19	13	20	18	23	24	15	14	10	= 172
HT Rater 2	16	25	17	23	21	28	22	13	11	11	= 187
Total errors annotated by the two raters: 359 (81 major and 278 minor errors)											

Table 4. Total errors in MTPE by rater and participant

Raters	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
MTPE Rater 1	10	13	15	18	14	19	25	24	20	10 = 168
MTPE Rater 2	13	15	13	14	19	21	24	28	19	13 = 179
Total errors annotated by the two raters: 347 (137 major and 210 minor errors)										

The quadratic weighted Cohen's Kappa was k ≈ 0.784 for the HT total errors and ≈ 0.827 for the MTPE total errors by the two raters. This indicates substantial to strong agreement between the number of errors annotated. The distribution of errors shows a slightly higher number of errors in HT than in MTPE. Although the total number of errors was slightly higher in the HT output, the vast majority were minor in nature, in contrast to MTPE, where major errors were more frequent. According to the MQM, major errors are those errors that impair readability, acceptability, distort the intended meaning or lead to mistranslations and are thus highlighted in figure 1 below. Figure 1 illustrates the distribution of major errors across the four categories in the assessments of the two tasks.

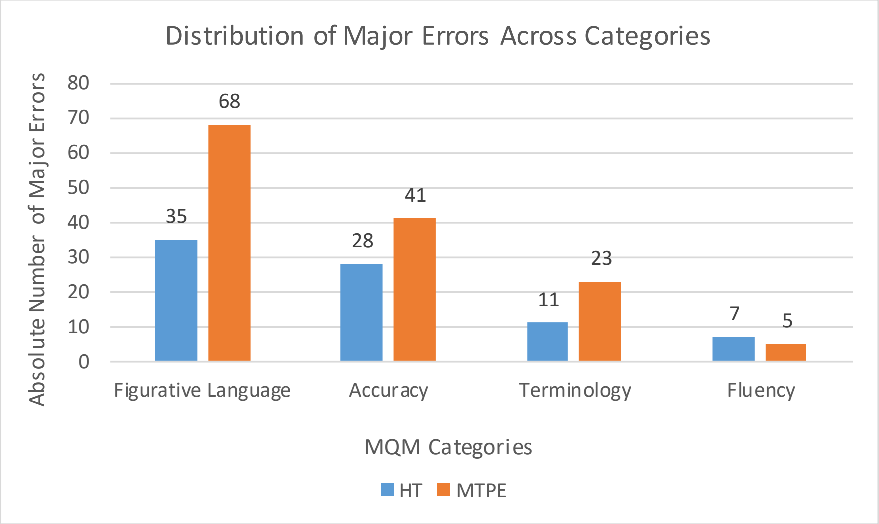


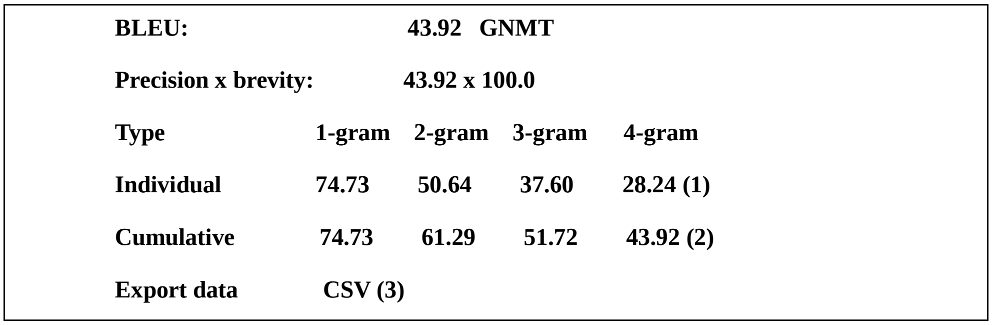
Figure 1. Major Errors in the HT and MTPE Tasks

Results in Figure 1 show that MTPE consistently reflects more major errors than HT across all MQM categories. This generally indicates that participants experienced greater difficulty in the MTPE task, particularly in dealing with nuanced culture-specific expressions and idioms (figurative language category) and subsequently the preservation of the intended meaning (accuracy category). These two categories reflect the highest discrepancies between the performance of participants in the HT and MTPE (a difference of 33 and 13 major errors in figurative language and accuracy, respectively). Terminology is not an exception, as the major errors in MTPE were also almost double those committed in the HT task. The category of fluency, on the other hand, shows a smaller gap between the major errors of the two tasks. This suggests that handling figurative expressions from Arabic into English is clearly problematic in MTPE based on empirical evidence that machine translation often produces literal renderings which normally distort the meaning in figurative language (Matusov, 2019; Guerberof-Arenas & Toral, 2022).

The higher number of errors in all categories in the MTPE task can be attributed to minimal post-editing effort and a tendency among post-editors to maintain most renderings of the MT which are literal by nature. MTPE appears more penalized not only because errors are more frequent but because they are more severe, affecting readability and meaning. However, the pattern of error distribution shows that figurative language accounts for the highest number of major errors, followed by accuracy, terminology, and fluency in descending order in both tasks. Overall results and their implications are discussed further in Sections 6 and 7. The following section presents BLEU scores for the GNMT output on which the participants worked in the MTPE task. Presenting BLEU scores of the unedited MT output provides insights on the quality of the machine translation at the surface level, based on n-gram similarity between the candidate and reference translations.

5.1.1 BLEU Metrics

BLEU was used to evaluate the quality of the machine-translated text by comparing it to the published English translation (Harlow & Riley, 2000) as the reference translation. The data illustrated in Figure 2 show BLEU scores in relation to the precision of n-gram overlaps between the candidate and reference translations.



BLEU:	**43.92 GNMT**			
Precision x brevity:	**43.92 x 100.0**			
Type	**1-gram**	**2-gram**	**3-gram**	**4-gram**
Individual	**74.73**	**50.64**	**37.60**	**28.24 (1)**
Cumulative	**74.73**	**61.29**	**51.72**	**43.92 (2)**
Export data	**CSV (3)**			

Figure 2. BLEU metrics scores per n-gram individually and cumulatively

The BLEU overall score for the GNMT output is 43.92. BLEU 1-gram measures unigram (single-word) matches, BLEU 2-gram measures bigram (two-word sequence) matches; BLEU 3-gram and 4-gram measure the matching of three-word and four-word sequences. This explains why the score decreases as BLEU calculates multi-word sequences. As the n-value increases, exact matches become rarer. Such data are particularly significant for the present study as metaphorical and idiomatic expressions are considered conventionalized phrases with restricted form and meaning (Cowie, 1988). This indicates that while the English reference translation reflects an idiomatic, functional translation, the GNMT renderings were literal resulting in either complete or partial mismatching. To contextualize the present BLEU scores, it is worth mentioning how they were interpreted in the literature. In most BLEU-based studies, scores ranging between 30-50 were interpreted as indicating translation quality from understandable to good (Toral, & S ánchez-Cartagena, 2017). According to Denkowski and Lavie (2010), scores above 30 generally reflect understandable translations, and above 50 reflect good to fluent translations. as illustrated below:

30 ______________35______________40_________________45______________50

Understandable *Moderate* *Adequate* *Good* *Excellent*

However, BLEU primarily evaluates the lexical and phrasal similarity between a candidate translation and a reference translation by quantifying n-gram(s) overlap. According to the interpretation of these scores, the score of the GNMT output is closer to *adequate* than it is to *good*. This result can be attributed to the inherent limitations of GNMT in capturing metaphors, idioms, and cultural references, where literal renderings fail to convey the intended meaning (Papineni et al., 2002). In light of this, and considering that the MTPE performance was lower than the HT performance (Table 1), it can be assumed that the participants may be biased toward minimal edits in the MTPE, leading to the preservation of errors inherent in the GNMT baseline. Several studies have revealed that MTPE is typically biased toward minimal changes, refraining from reformulation of the MT output (e.g., Koponen et al., 2012). This tendency implies that errors embedded in the initial MT output are likely to persist in the final edited version. In contrast, unaided human translation was seemingly able to re-contextualize the figurative meaning, aligning with target-language conventions which may explain the higher HT scores in the MQM evaluation (Table 1). The divergence in translation quality among the initial GNMT output, the post-edited versions, and the human-translated texts is discussed further in Section 7.

5.2 Temporal Effort

Two types of data were extracted from the Translog II log files: the total time spent on each task by each participant and the word count of each target text, where the latter was used to calculate the average number of seconds per target word. Tables 5 and 6, and Figure 3 below present the data, showing individual and aggregate results.

Table 5. Total time spent on HT in Minutes

Time Spent on HT Task in Minutes									
P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
40	38	42	41	39	44	43	40	41	42
Mean ≈ 41 Minutes									

Table 6. Total time spent on MTPE in Minutes

Time Spent on MTPE Task in Minutes									
P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
34	32	33	34	35	31	33	33	36	30
Mean ≈ 33 Minutes									

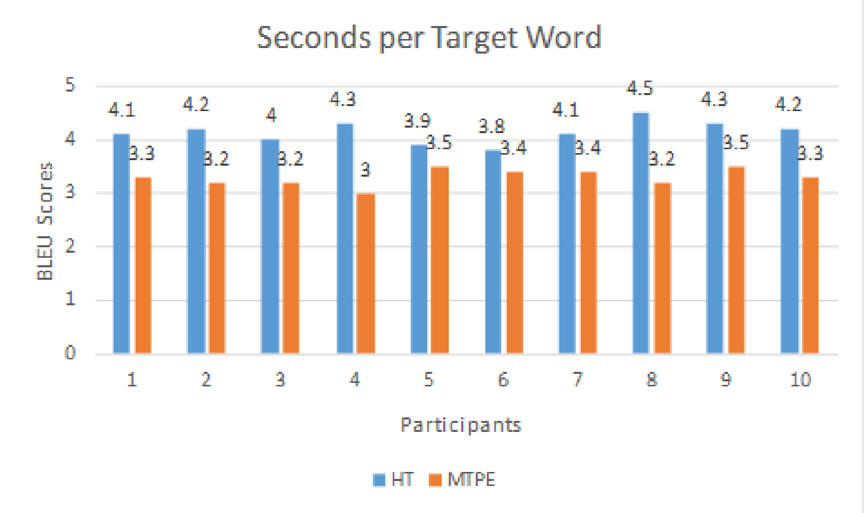


Figure 3. Time spent by each participant in seconds per target word

The time data show a clear difference in task duration between HT and MTPE, with implications for cognitive load and work efficiency. The Wilcoxon paired test did not show any negative ranks. The p-value is < 0.01 (very significant) with a strong effect size. Analysis reveals that participants spent significantly more time on the HT task (mean ≈ 41 minutes) than on the MTPE task (mean ≈ 33 minutes). This represents about a 20% reduction for the MTPE task, supporting the assumption that post-editing is more time-efficient because the pre-translated output reduces the need for full syntactic and semantic generation. Notably, time variability across participants was identical in both tasks (a range of 6 minutes), although absolute times were lower in MTPE. This consistency in time variability suggests that post-editing introduced a more uniform workflow, possibly because the machine-generated text standardized the difficulty level. However, this may also lead to a more mechanical process that risks overlooking subtleties in figurative meaning.

These results are consistent with the underlying logic of post-editing workflows: the machine provides a preliminary draft that reduces the cognitive and linguistic load on the translator. The findings reveal that student translators showed lower average seconds per target word when post-editing machine-translated text compared to performing human translation from scratch.

Although the present research does not examine the cognitive effort involved in translating versus post-editing, it is worth noting that several studies have linked temporal effort to cognitive effort. Evidence-based studies have concluded that longer processing times are indicators of higher cognitive effort (e.g., Koponen et al., 2012; Wang, Zhang & Chen, 2022).

The HT task required more average time, indicating deeper engagement with the source text and greater cognitive effort, resulting in an MQM average score of 65.85. The MTPE task, on the other hand, required less time (20%-time reduction), resulting in an MQM average score of 54.95. In other words, the 20% increase in time consumption was at the expense of quality rating with a difference of 11 points in favor of human translation. Section 6 provides a combined discussion of the results followed by the conclusion and pedagogical implications.

6. Discussion

The present pilot study examined how translation trainees perform human translation (HT) and machine translation post-editing (MTPE) when working with figurative expressions from Arabic into English. The study intended to focus on efficiency (temporal effort) and effectiveness (translation quality). The findings reveal a consistent and meaningful trade-off between speed and quality. MTPE resulted in an appreciable reduction in task duration (approximately 20%), whereas HT produced higher-quality translations, with an average MQM score advantage of 11 points.

6.1 Efficiency Gains vs. Depth of Processing

The reduced temporal effort observed in the MTPE task confirms that the availability of a pre-generated draft is likely to increase efficiency in the translation workflow. The shorter task duration suggests that MTPE shifts the translator’s role from text generation to revision and evaluation, reducing the need for continuous lexical search and text structuring. From a process-oriented perspective, this supports that post-editing is a fundamentally different cognitive activity from translating a text from scratch. In other words, post-editors tend to focus on surface-level adjustments to meaning, while translators engage more deeply in both semantic and syntactic construction which is inherently time-consuming. Nevertheless, the time savings observed in the MTPE task must be interpreted cautiously. While reduced task duration may indicate lower cognitive effort, the quality outcomes suggest that speed gains may come at the expense of deeper engagement with the source text. In the context of figurative language, which requires cultural awareness and creative reformulation, the availability of a machine-generated draft may constrain the translator’s problem-solving processes.

6.2 Quality Outcomes and the Nature of Errors

Although the total number of errors in HT was slightly higher than in MTPE, the error typology and severity profile reveal a more critical distinction. MTPE was associated with a higher number of major errors across all MQM categories, particularly in figurative language and accuracy. This finding is especially significant, as major errors directly affect meaning transfer (accuracy), readability, and communicative adequacy.

This pattern suggests that MTPE does not merely reflect more errors, but rather different types of errors that are more disruptive. The predominance of major errors in figurative language indicates that machine-generated literal renderings often survive the post-editing process, even when they distort the intended meaning. In contrast, human translation, despite involving more minor inaccuracies, tends to preserve the intended meaning.

6.3 Minimal-Edit Bias and Trainee Vulnerability

One plausible explanation for the observed quality gap lies in the well-documented phenomenon of minimal-edit bias in post-editing. The BLEU score of 43.92 suggests that the GNMT output was understandable at a surface level, which may have encouraged trainees to accept the machine’s solutions with limited scrutiny. As previous studies have shown, novice post-editors tend to refrain from extensive reformulation, especially when the output appears grammatically fluent.

For figurative expressions, this tendency is particularly problematic since these expressions are often conventionalized and culturally embedded, requiring translators to move beyond lexical equivalence. When a given MT output provides a literal but fluent rendering, trainees may fail to recognize semantic mismatches, leading to the retention of meaning-distorting solutions. This helps explain why MTPE texts were closer to the MT baseline and why major errors persisted despite post-editing.

6.4 Creativity, Agency, and Cognitive Engagement

The longer processing times observed in the HT task may reflect deeper cognitive engagement and greater translator agency. Human translation requires continuous decision-making and meaning reformulation, especially when dealing with figurative language. This aligns with views of translation as an interpretive and problem-solving activity rather than a purely technical operation (Toury, 1995).

By contrast, MTPE may reduce opportunities for creative problem-solving by anchoring the translator to the MT output. While this anchoring effect improves efficiency, it may also narrow the range of solutions, particularly for non-literal language. The findings thus support arguments that MTPE workflows, if applied uncritically, may limit creative reformulation and deeper semantic processing.

Taken together, the results suggest that MTPE is not a universally optimal solution, especially for texts rich in figurative and culturally specific language. While MTPE offers clear efficiency advantages, its effectiveness appears limited when meaning negotiation and creative adaptation are required. For professional practice, this implies that task suitability should be carefully evaluated before adopting MTPE workflows. From a pedagogical perspective, the findings highlight the need for targeted training in post-editing strategies which are highlighted further in the next section.

7. Conclusion and Pedagogical Implications

The present study concludes that in the HT task, the participants were likely more engaged in interpreting and reformulating the contextual meaning and culture-specific elements embedded in figurative expressions. This deep engagement in the task is imposed by the increased interpretive effort on the source text with the absence of any pre-existing template. Thus, the HT task required 20% more temporal effort than the MTPE task. The existence of a pre-translated output in post-editing encourages over-reliance on the first draft, reducing the sense of ownership over the text, thus creating an anchoring effect (Tversky & Kahneman, 1974).

Interestingly, the same participants showed different behaviors across the two tasks for similar figurative expressions. In the HT task, the participants appeared more engaged and tended toward functional translation strategies. In contrast, during the MTPE task, their decisions were influenced by the GNMT output, which they often accepted with minimal intervention.

The present study suggests that MTPE may limit critical thinking and creativity under the anchoring effect. Trainee translators seem to rely heavily on machine output structure without considering the idiomatic meaning or reevaluating the figurative content. Translation of figurative language is error-prone because it involves cultural and pragmatic knowledge. When such expressions are not translated functionally, accuracy of meaning is adversely affected as observed in the distribution of errors across categories. Terminology and fluency are, on the other hand, language- rather than culture-related categories in which MT systems are trained to produce grammatically correct texts.

The fact that the same participants performed better in the HT than they did in the MTPE shows that the practice of post-editing is not well-developed in translator training. This becomes more problematic when dealing with figurative language, in which machine translation has its own limitations. Although MTPE reduced the temporal effort by approximately 20%, this saving came at the expense of the final translation quality. The difference in time, while measurable, was not large, indicating that interpreting, and correcting a pre-translated text still requires significant cognitive engagement. Therefore, the modest time savings offered by MTPE do not appear to justify the compromise in handling figurative expressions. Based on the results, it can be assumed that the low performance in the MTPE is not entirely due to the participants' ignorance of handling figurative language, because they performed better in the from-scratch human translation for similar expressions. The low performance in the MTPE can be attributed to two factors. Firstly, students seem to have more trust in machine translation than they have in their own abilities. Secondly, they seem to have insufficient awareness of the limitation of machine translation in domain-specific texts between Arabic and English.

However, from another perspective, the major errors reflected in the HT task should not be neglected, especially that the highest MQM score given by the raters was 78% with a mean score of 65.85% where the majority of errors were reflected in the handling of figurative language and accuracy. Having a similar pattern of error distribution in the two tasks indicates that figurative language translation from Arabic into English is likely to pose a certain level of difficulty for trainees in human translation as well as in post-editing.

8. Study Limitations and Recommendations

This research has several limitations that should be acknowledged. The small sample size limits the generalizability of the results. A larger sample with extended assessment methods would allow for more robust statistical analysis and finer-grained comparisons. In addition, the task focused on 20 figurative expressions in one language pair. Future studies can possibly test whether these findings hold across different genres (e.g., legal, medical, technical texts) in different language pairs or translation directions.

Based on the findings and discussion above, the following can be recommended for translator training:

- Incorporate balanced training in MTPE for domain-specific texts, especially those involving figurative language.
- Emphasize critical evaluation of machine translation output, where students are trained not only to correct surface language-related errors but also to ensure adequacy of target cultural norms.
- Use of process-oriented tools such as *Translog II* in translator training to engage students in time management and enhance their awareness of the cognitive effort it requires.

Acknowledgments

The author confirms that this research was conducted independently and that no individuals or institutions contributed to the preparation of this manuscript.

Authors' contributions

Ogareet Yacoub Khoury is the sole and only author of the manuscript.

Funding

This work did not receive any funding from any individual or entity.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

Copyrights

References

Abrams, M. H. (1999). *A glossary of literary terms* (7th ed.). Heinle & Heinle.

Abu Manie, H., Al-Sayyed, S., Alkhawaja, L. & Rababa'h, B. (2025). Congratulation strategies of Crown Prince Hussein's Wedding: A socio-pragmatic study of Facebook comments. *Open Cultural Studies, 9*(1), 2025-0051. https://doi.org/10.1515/culture-2025-0051

Afzal, M. T. (2018). Challenges of machine translation for literary texts. *Journal of Translation Studies*, *12*(2), 45-60.

Alhaisoni, E., & Alhaysony, M. (2017). An investigation of Saudi EFL university students' attitudes towards the use of Google Translate. *International Journal of English Language Education*, *5*(1), 72-82. https://doi.org/10.5296/ijele.v5i1.10696

Almommani, O. (2024). Navigating the gray zone: When interpreters become mediators and communication facilitators. *Journal of Language Teaching and Research, 15*(4), 1372-1380. https://doi.org/10.17507/jltr.1504.35

Alsharif, B., & Khasawneh, R. (2025). Strategies of rendering metaphor from Arabic into English: A comparative study of ChatGPT and Matecat. *World Journal of English Language, 16*(1), 45-51. https://doi.org/10.5430/wjel.v16n1p45

Baker, M. (2018). *In other words: A coursebook on translation* (3rd ed.). Routledge. https://doi.org/10.4324/9781315619187

Bamia, A. (2002). Review of Karen Riley's translation of *Returning to Haifa*. *Journal of Third World Studies*, *19*(1), 123-130.

Bassnett, S., & Lefevere, A. (1998). *Constructing cultures: Essays on literary translation.* Multilingual Matters. https://doi.org/10.21832/9781800417892

Besacier, L., & Schwartz, L. (2015). Phrase-based machine translation and post-editing for English-French literary texts. *Machine Translation*, *29*(3-4), 267-284.

Carl, M. (2003). From translation process research to the study of user–translator interaction. In F. Alves (Ed.), Triangulating translation: Perspectives in process-oriented research (pp. 91-109). John Benjamins.

Castilho, S., & O'Brien, S. (2017). Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia, New Series – Themes in Translation Studies, 16,* 120-136. https://doi.org/10.52034/lanstts.v16i0.430

Castilho, S., & Resende, C. (2022). Post-editing of literary texts: Creativity versus productivity. *Translation Spaces*, *11*(1), 56-73.

Cowie, A. P. (1988). The treatment of idioms and metaphoric expressions in machine translation. In F. J. P. Dale, C. J. Mellish, & M. Zock (Eds.), *Current issues in computational linguistics*: In honour of Don Walker (pp. 187-202). Springer.

Dagnev, D., & Chervenkova, V. (2020). Literal translation of figurative language: English to Bulgarian. *Bulgarian Journal of Linguistics*, *34*(2), 99-114.

Dagut, M. (1976). Can "metaphor" be translated? *Babel, 22*(1), 21-33. https://doi.org/10.1075/babel.22.1.05dag

Dankers, V., Lucas, C. G., & Titov, I. (2022). Can Transformer be too compositional? Analysing idiom processing in neural machine translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3608-3626). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.252

Denkowski, M., & Lavie, A. (2010). Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010) (pp. ...). Association for Machine Translation in the Americas.

Dorst, A. (2023). Neural machine translation and figurative language: Post-editing prospects. *Journal of Artificial Intelligence and Translation*, *5*(1), 12-28.

Feng, Z., Su, J., Zheng, J., Ren, J., Zhang, Y., Wu, J., Wang, H, & Liu, Z. (2025). M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, *1*, 7084-7107, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.351

Fiederer, R., & O'Brien, S. (2009). Quality and machine translation: A realistic objective? *The Journal of Specialised Translation, 11,* 52-74. https://doi.org/10.26034/cm.jostrans.2009.639

Gaspari, F., Toral, A., Naskar, S. K., Groves, D., & Way, A. (2014). Perception versus reality: measuring machine translation post-editing productivity. In S. O'Brien, M. Simard & L. Specia (Eds.), *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas* (pp. 60–72). Association for Machine Translation in the Americas.

Gibbs, R. W., Jr. (1994). *The poetics of mind: Figurative thought, language, and understanding.* Cambridge University Press.

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

Habib, R., Alkhawaja, L., Khoury, O., & Al-Sayyed, S. (2025). Six NMT systems, one language pair: Which best translates Arabic-English? *World Journal of English Language, 16*(1), 1-16. https://doi.org/10.5430/wjel.v16n1p1

Harlow, B. & Riley, K. (2000). *Palestine's Children: Returning to Haifa & Other Stories*. Lynne Rienner Publishers: USA.

Hartley, A. (2009). Machine translation in practice: Overview and challenges. *Translation Journal*, *13*(3), 40-50.

House, J. (1977). *A model for translation quality assessment.* Gunter Narr Verlag. https://doi.org/10.7202/003140ar

Jia, L., & Lai, K. (2022). Post-editing metaphor-rich English texts into Chinese: Efficiency and accuracy. *Translation and Interpreting Studies*, *17*(2), 220-237.

Kanafani, G. (1969). *Returning to Haifa* [Arabic novella].

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In Proceedings of the Workshop on Statistical Machine Translation (pp. 102-121). Association for Computational Linguistics. https://doi.org/10.3115/1654650.1654666

Koglin, A. (2015). Cognitive effort in post-editing metaphors versus manual translation: An eye-tracking study. *Translation & Interpreting, 7*(1), 126-141. https://doi.org/ti.106201.2015.a06

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In S. O'Brien, M. Simard & L. Specia (Eds.), *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice* (pp. 11-20). Association for Machine Translation in the Americas.

Kuzman, G., Vintar, Š., & Arčan, I. (2019). Evaluating NMT for literary texts: English to Slovene. *Machine Translation*, *33*(1-2), 127-142.

Lommel, A., Brümmer, M., & Uszkoreit, H. (2014). MQM: A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: Tecnologies de la Traducció, 12,* 455-463. https://doi.org/10.5565/rev/tradumatica.77

Lu, Q., Ding, L., Zhang, K., Zhang, J., & Tao, D. (2025). MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, & S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics (pp. 5570-5587). Association for Computational Linguistics.

Martin, T., & Serra, J. (2014). The evolution of machine translation in the industry. *Translation Journal*, *18*(2), 55-68.

Mason, I. (1982). The untranslatability of idioms reconsidered. *The Translator*, 1(2), 89-98.

Matusov, E. (2019). Professional translators' perspectives on MT for literary texts. *Journal of Translation Technology*, *11*(1), 33-47.

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, *27*(10), 110878. https://doi.org/10.1016/j.isci.2024.110878

Newmark, P. (1988). *A textbook of translation*. Prentice Hall.

Nida, E. A. (1964). *Toward a science of translating*. E.J. Brill. https://doi.org/10.1163/9789004495746

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In 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

Park, J. S. (2009). Literal translation as a primary strategy in figurative language translation. *Translation Quarterly*, *65*, 25-42.

Pérez, C. (2024). Re-thinking machine translation post-editing guidelines. *The Journal of Specialized Translation, 41*(1), 26-47. https://doi.org/10.26034/cm.jostrans.2024.4696

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics, 93,* 7-16. https://doi.org/10.2478/v10108-010-0010-x

Reiss, K. (2000). *Type, kind and individuality of text: Decision making in translation*. St. Jerome.

Tawfiq, A. (2015). Translating emotive culture-specific figurative language: Context sensitivity matters. *Translation Journal*, *19*(4), 10-25.

Toral, A., & Way, A. (2018). Human post-editing of neural machine translation output for literary texts. *Journal of Machine Translation, 32*(3-4), 235-254.

Toral, A., Sprugnoli, R., & Gervás, P. (2018). Post-editing in literary translation: Efficiency and quality assessment. *Translation and Interpreting Studies*, *13*(2), 190-211.

Toury, G. (1995). *Descriptive translation studies and beyond.* John Benjamins. https://doi.org/10.1075/btl.4

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Van den Broeck, R. (1981). The limits of translatability exemplified by metaphor translation. *Poetics Today*, *2*(4), 73-87. https://doi.org/10.2307/1772487

Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

Wang, S., Zhang, L., & Chen, Y. (2022). Advances in neural machine translation. *Computational Linguistics*, *48*(1), 1-20.

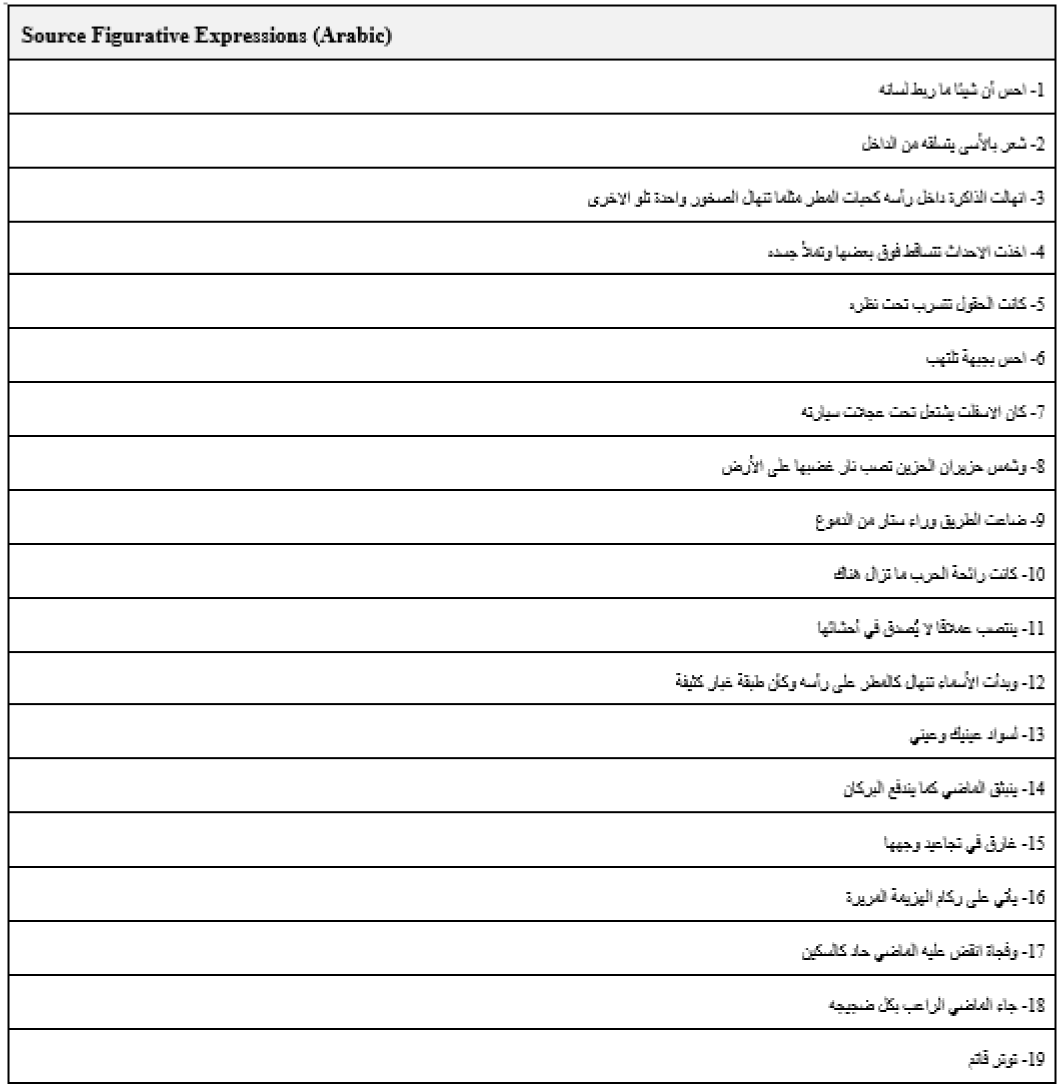
White, J. (1995). Approaches to black-box machine translation evaluation. In Proceedings of MT Summit V (pp. 386–393).

Williams, M. (2004). *Translation quality assessment: An argumentation-centred approach.* University of Ottawa Press.

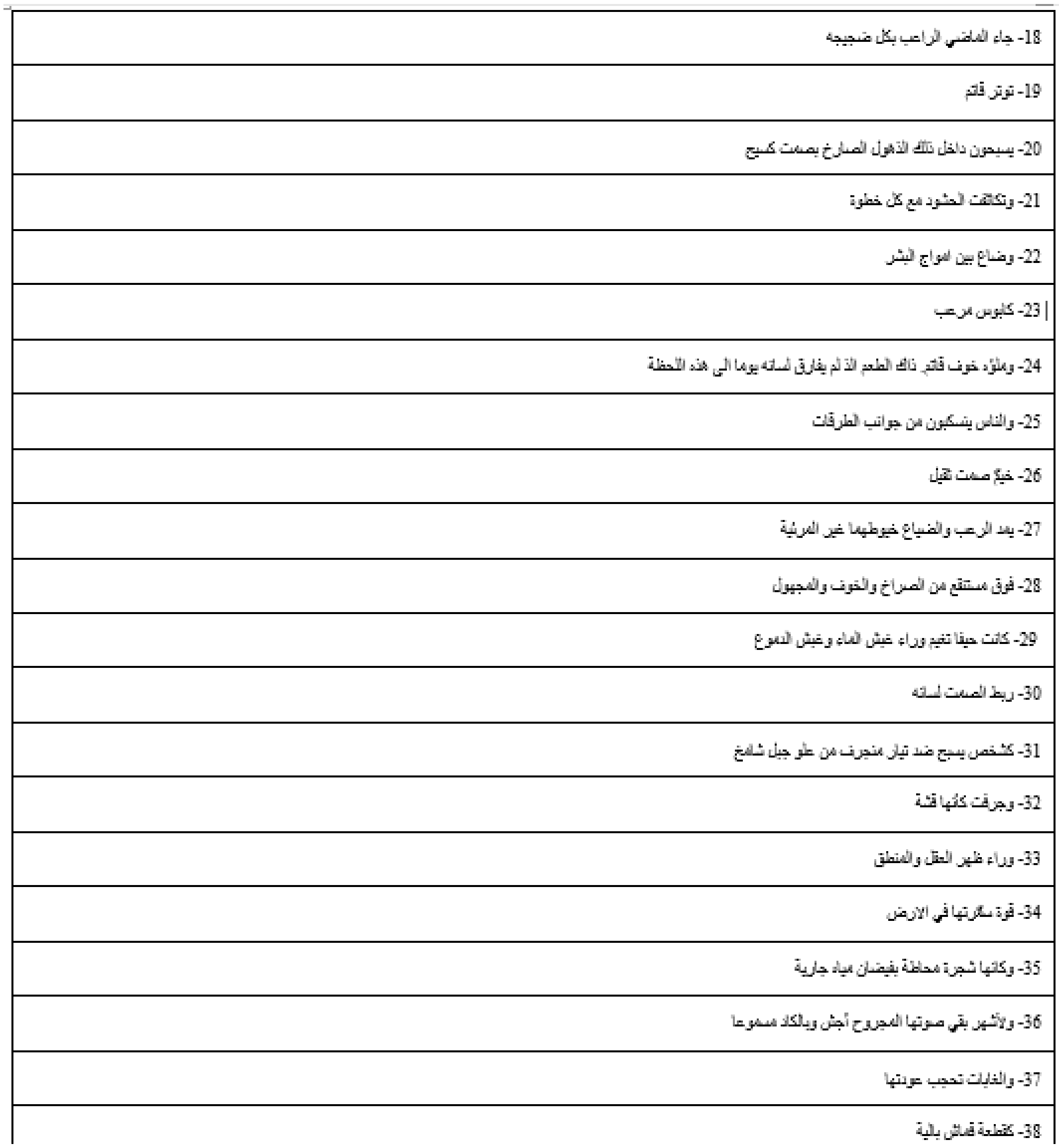
https://doi.org/10.1353/book6617

Youssef, A., & Albarakati, M. (2023). Translators' preferences for literal translation of figurative language. *International Journal of Translation*, *35*(1), 15-29.

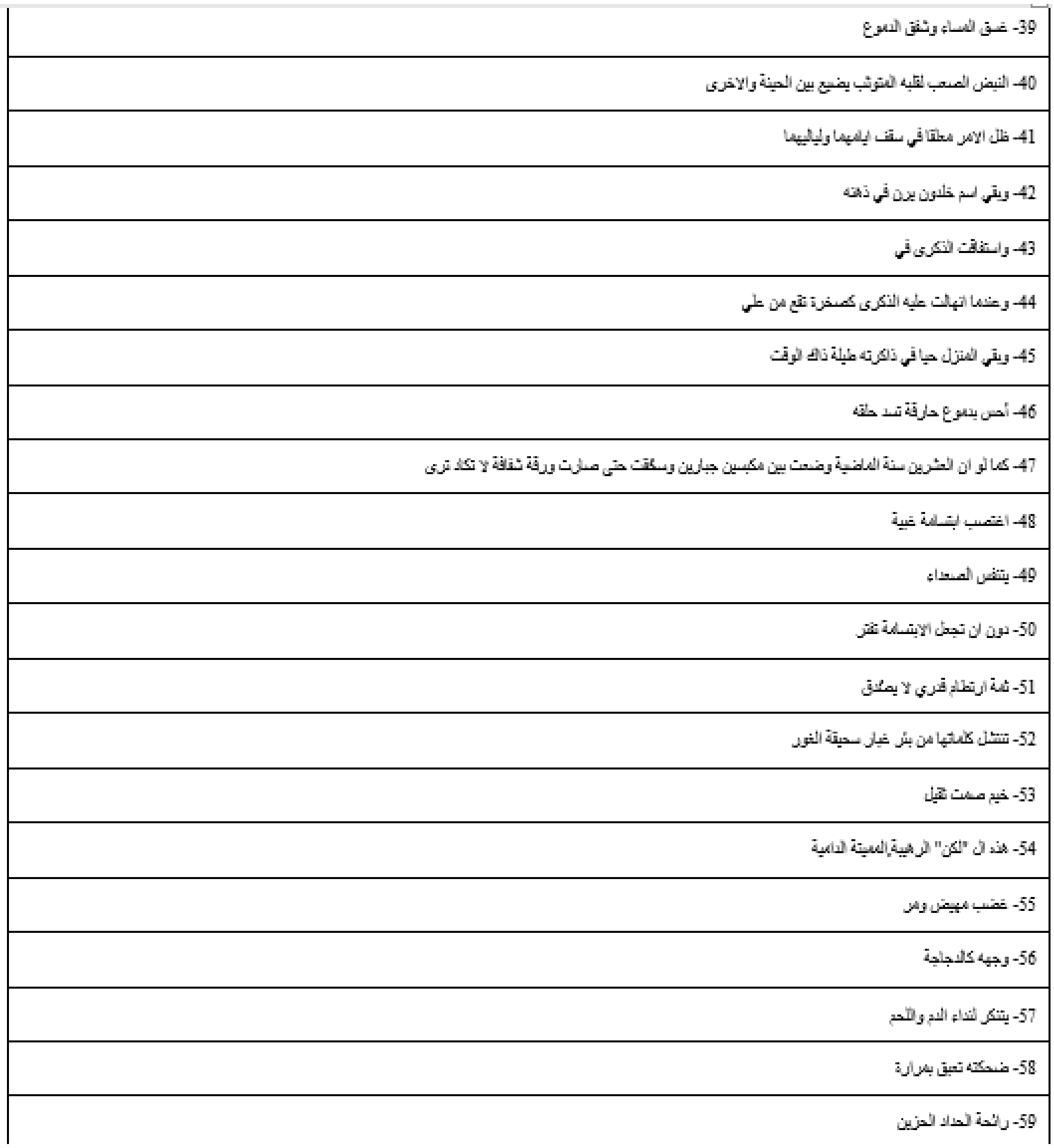
Appendix 1



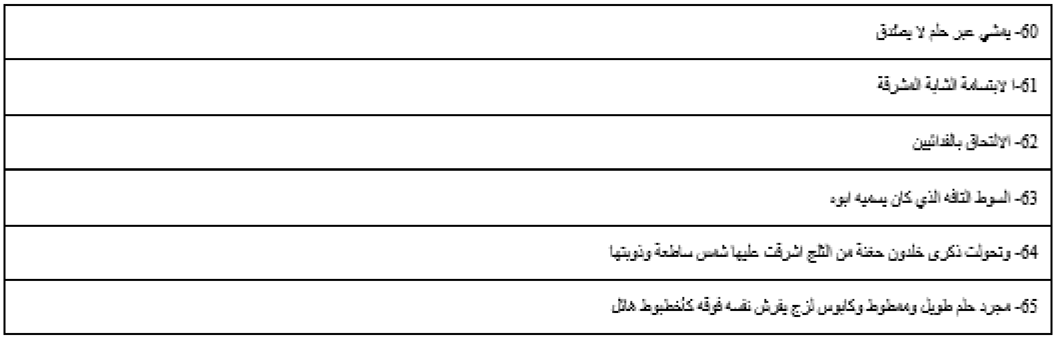
Source Figurative Expressions (Arabic)
1- احس أن شيئا ما ربط لسانه
2- شعر بالأسى يشلقه من الداخل
3- انهالت الذاكرة داخل رأسه كحبات المطر مثلما تنهال الصخور واحدة تلو الاخرى
4- اخذت الاحداث تتساقط فوق بعضها وتملأ جسده
5- كانت الحقول تتسرب تحت نظره
6- احس بجبهة تلتهب
7- كان الاسفلت يشتعل تحت عجلات سيارته
8- وشمس حزيران الحزين تصب نار غضبها على الأرض
9- ضاعت الطريق وراء ستار من الدموع
10- كانت رائحة الحرب ما تزال هناك
11- ينتصب عملاقا لا يُصدق في أحشائها
12- وبدأت الأسماء تنهال كالمطر على رأسه وكأن طبقة غبار كثيفة
13- أسواد عينيك وعيني
14- ينبثق الماضي كما يندفع البركان
15- غارق في تجاعيد وجهها
16- يأتي على ركام الهزيمة المريرة
17- وفجأة انقض عليه الماضي حاد كالسكين
18- جاء الماضي الراعب بكل ضجيجه
19- توتر قاتم



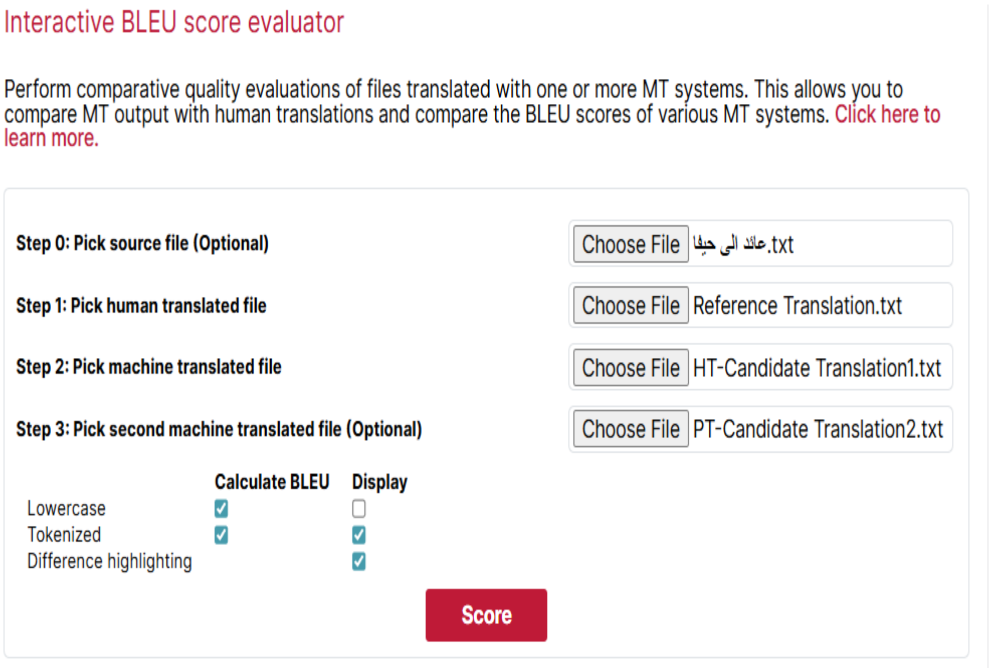
18- جاء الماضي الراعب بكل ضجيجه
19- توتر قاتم
20- يسبحون داخل ذلك الذهول الصارخ بصمت كسيح
21- وتكاثفت الحشود مع كل خطوة
22- وضاع بين امواج البشر
23- كابوس مرعب
24- وملؤه خوف قاتم. ذاك الطعم الذ لم يفارق لسانه يوما الى هذه اللحظة
25- والناس ينسكبون من جوانب الطرقات
26- خيمّ صمت ثقيل
27- يمد الرعب والضياع خيوطهما غير المرئية
28- فوق مستنقع من الصراخ والخوف والمجهول
29- كانت حيفا تغيم وراء غبش الماء وغبش الدموع
30- ربط الصمت لسانه
31- كشخص يسبح ضد تيار منجرف من علو جبل شاهق
32- وجرفت كأنها قشة
33- وراء ظهر العقل والمنطق
34- قوة سمّرتها في الارض
35- وكأنها شجرة محاطة بفيضان مياه جارية
36- ولأشهر بقي صوتها المجروح أجش وبالكاد مسموعا
37- والغابات تحجب عودتها
38- كقطعة قماش بالية



39- غسق المساء وشفق الدموع
40- النبض الصعب لقلبه المتوثب يضيع بين الحينة والاخرى
41- ظل الامر معلقا في سقف ايامهما ولياليهما
42- وبقي اسم خلدون يرن في ذهنه
43- واستفاقت الذكرى في
44- وعندما انهالت عليه الذكرى كصخرة تقع من علي
45- وبقي المنزل حيا في ذاكرته طيلة ذاك الوقت
46- أحس بدموع حارقة تسد حلقه
47- كما لو ان العشرين سنة الماضية وضعت بين مكبسين جبارين وسحقت حتى صارت ورقة شفافة لا تكاد ترى
48- اغتصب ابتسامة غبية
49- يتنفس الصعداء
50- دون ان تجعل الابتسامة تفتر
51- ثمة ارتطام قدري لا يصدّق
52- تنتشل كلماتها من بئر غبار سحيقة الغور
53- خيم صمت ثقيل
54- هذه ال "لكن" الرهيبة,المميتة الدامية
55- غضب مهيض ومر
56- وجهه كالدجاجة
57- يتنكر لنداء الدم واللحم
58- ضحكته تعبق بمرارة
59- رائحة الحداد الحزين



60- يمشي عبر حلم لا يصدّق
61-ا لابتسامة الشابة المشرقة
62- الالتحاق بالفدائيين
63- السوط التافه الذي كان يسميه ابوه
64- وتحولت ذكرى خلدون حفنة من الثلج اشرقت عليها شمس ساطعة وذوبتها
65- مجرد حلم طويل وممطوط وكابوس لزج يفرش نفسه فوقه كأخطبوط هائل



Interactive BLEU score evaluator

Perform comparative quality evaluations of files translated with one or more MT systems. This allows you to compare MT output with human translations and compare the BLEU scores of various MT systems. Click here to learn more.

Step 0: Pick source file (Optional) Choose File عائد الى حيفا.txt

Step 1: Pick human translated file Choose File Reference Translation.txt

Step 2: Pick machine translated file Choose File HT-Candidate Translation1.txt

Step 3: Pick second machine translated file (Optional) Choose File PT-Candidate Translation2.txt

	Calculate BLEU	Display
Lowercase	☑	☐
Tokenized	☑	☑
Difference highlighting		☑