

ASSESSING NEURAL MACHINE TRANSLATION ERRORS AND POST-EDITING PRACTICES ON LITERARY TRAITS OF EXHORTATION TEXTS

Bahia Zemni ¹ 

Translation Department, College of Languages, Princess Nourah bint Abdulrahman University,
Saudi Arabia

baalzemni@pnu.edu.sa

Abstract: Despite the challenges and limitations, it is undeniable that machine translation (MT) remains a crucial tool for translators, especially as technology developers endeavor to improve current systems and theorists strive to optimize concepts and processes, such as post-editing (PE). In this regard, the present study is dedicated to a critical and statistical analysis of the PE operations carried out by university students to improve the quality of an Arabic-English MT of a motivational text produced by the SYSTRAN Pro trial version. A total of 27 Master's students specializing in translation participated in the study, and a mixed-method approach combining descriptive statistics and *one-way ANOVA* was employed to evaluate the data. The study focuses on categorizing and searching for statistically significant differences among the types of corrections proposed by students, and then analyzing them. The motivational text was chosen for its literary, religious, and cultural elements, which present particularly challenging translation problems for NMT systems. The results indicate limitations in participants' proficiency in PE. The researchers identified four recurring correction patterns: error detection without correction (Case 1), inappropriate correction (Case 2), partially accurate correction (Case 3), and fully relevant correction (Case 4). Statistical analysis revealed significant differences among these four cases. Notably, Case 4 exhibited a significantly higher frequency. These findings highlight the need for dedicated PE training within translation curricula to enhance students' competence in revising NMT outputs.

Keywords: Arabic-English translation; motivational text; Neural Machine Translation (NMT); NMT Quality; Post-editing; SYSTRAN;

How to cite this article:

Zemni, B. (2026). Assessing neural machine translation errors and post-editing practices on literary traits of exhortation texts. *Journal of Studies in Language, Culture, and Society (JSLCS)*, 9(2), 125-145

¹ Corresponding author: Bahia Zemni Author's ORCID ID: <https://orcid.org/0000-0002-6238-7509>

1. Introduction

The contemporary world is characterized by the ubiquitous integration of technology, exerting a profound influence across all domains including translation. Advancements in artificial intelligence and neural networks are propelling translation daily workflow into a highly automated and autonomous epoch. However, MT-generated documents are susceptible to errors (Canfora & Ottmann, 2020), influenced by factors such as language pairs (Solyman et al., 2021; Li et al. 2023), translation objectives, domains, text types (Afzaal et al., 2022) and the specific MT system employed (Marzouk, 2021). As a result, translators' involvement in the translation process has significantly decreased, and their roles primarily involve pre-editing (Pre E) and post-editing (PE) operations (Marzouk & Hansen-Schirra, 2019). These operations, encompassing various linguistic and non-linguistic modifications, such as correction, reorganization, and the substitution of lexical and cultural elements (Li et al. 2023), are necessary to fix MT imperfections.

In this regard, this study aims to help teachers and learners quickly develop machine-translation literacy, to design classroom activities incorporating MT tools (Cotelli Kureth & Summers, 2025), and to clarify the types, quality, and efficiency of post-editing operations among students in particular. Expanding upon previous studies (Abdelaal & Alazzawie, 2020; Aldawsari, 2023 among others), that have investigated various technological facets and their repercussions within the field of translation, addresses the following research questions:

1. What is the overall proficiency level of students in mastering PE tasks?
2. Which approaches (e.g., detecting mistakes only, detecting and proposing corrections) do students most frequently adopt during PE operations? Are there statistically significant differences in the adoption rates of these approaches?
3. What types of errors do students commonly make when correcting mistakes produced by NMT, specifically SYSTRAN Pro? Are there statistically significant differences in the frequency or types of errors made by students?

The investigation has established the following null hypotheses:

1. Students demonstrate a moderate to weak level of proficiency in mastering PE tasks, as measured by accuracy, fluency, and efficiency in correcting machine-translated texts.
2. Students most frequently adopt the approach of detecting and proposing corrections during PE operations, as it allows for more comprehensive improvements to the text.

Null Hypothesis (H_0): There is no statistically significant difference in the adoption rates of different PE approaches.

3. Students most commonly make errors related to over-editing or under-editing when post-editing SYSTRAN Pro outputs.

Null Hypothesis (H_0): There is no statistically significant difference in the frequency or types of errors made by students when correcting mistakes produced by SYSTRAN Pro.

By addressing these research questions and testing the corresponding hypotheses, this study aims to achieve the following objectives:

- To evaluate the overall proficiency level of students in PE tasks.
- To determine the most frequently adopted PE approaches among students and to analyse whether there are statistically significant differences in the adoption rates of these approaches.

- To investigate the types of errors students commonly make when post-editing SYSTRAN Pro outputs and to examine whether there are statistically significant differences in the frequency or types of these errors.

2. Literature Review

The rapid advancement of translation technologies over the past decade has profoundly transformed the field of translation studies, creating new opportunities for scholarly inquiry. Among these advancements, NMT has garnered significant attention, particularly in relation to PE. While NMT has demonstrated remarkable progress in certain language pairs, its effectiveness varies considerably depending on linguistic structures and textual complexities. This section provides a comprehensive review of some studies on PE strategies in Arabic translation, with the aim of identifying existing research gaps and justifying the significance of the present study.

Several studies (Benková & Benko, 2020; Stahlberg, 2020; Balashov, 2022) identify NMT as one of the most notable advancements in deep learning and artificial intelligence, utilizing an encoder-decoder framework to enhance translation quality. To further refine NMT-generated texts, PE techniques are often employed. Pérez (2024) highlights that the effectiveness of PE largely depends on the quality of the initial MT output, as higher-quality translations require minimal intervention. Furthermore, Costa-Jussà et al. (2024) assert that contemporary NMT systems can now leverage multilingual capacities and even perform zero-shot translation, yielding promising results in terms of language coverage and translation accuracy. Comparative studies have further demonstrated NMT's superiority over earlier translation models. Koponen et al. (2019) conducted an analysis of post-edited English-to-Finnish translations generated by NMT, statistical machine translation (SMT), and rule-based machine translation (RBMT). Their findings revealed that NMT consistently produced superior translations, requiring significantly less PE than SMT and RBMT. Similarly, Valdez and Lomeña Galiano (2021) investigated the effectiveness of NMT in translating journalistic texts from Spanish to French. Their study found that NMT effectively preserved the semantic content of lexemes and made appropriate translation choices, particularly in the processing of anaphoric expressions. This success was attributed to linguistic and cultural similarities between Spanish and French, as well as to the efficiency of NMT's neural architecture. This architecture employs a single large neural network to transform source sentences into target sentences (Stahlberg, 2020).

Despite its progress, NMT performance remains inconsistent when applied to morphologically complex and structurally diverse languages such as Arabic. Aldawsari (2023) examined NMT's ability to process Arabic homonyms, heteronyms, and polysemes, highlighting morphosyntactic challenges that often lead to ambiguity and translation errors. Similarly, Aldelaa and Malkawi (2024) assessed NMT accuracy in translating Arabic corpora into English, identifying key deficiencies such as unwanted literal translations of idiomatic expressions and omissions or misinterpretations of culture-bound elements. They also reported inaccurate lexical predictions leading to a lack of cohesion, ambiguity resulting from homonymy and polysemy, and inadequate word choices in colloquial and figurative expressions. Furthermore, Almekhlafi and Nagi (2024) analyzed NMT performance in translating academic texts, particularly English-to-Arabic abstracts of educational research papers. Their findings underscore persistent lexical and syntactic inaccuracies, emphasizing the need for more extensive Arabic-language corpora and improved NMT training data to enhance translation quality.

Beyond assessing NMT accuracy, some studies have explored PE strategies among Arabic-speaking students. Sammarraie and Ghassemiazghandi (2024) investigated variations in translation abilities among Iraqi university students, while Mohammed and Al-Rubai'i (2024) examined the PE performance of fourth-year translation students from Al-Mustansiriyyah and Mosul Universities. Their research assessed students' ability to post-edit an English journalistic text translated into Arabic via Google Translate, focusing on the types of errors made in PE tasks and their impact on overall translation quality. Findings revealed that accuracy-related errors were the most frequent, followed by terminology inconsistencies, highlighting the difficulties students face in mastering PE techniques.

While existing research has made valuable contributions to NMT accuracy and PE methodologies, several gaps remain. First, the number of studies focusing on Arabic NMT and PE strategies is still limited, particularly given the linguistic complexity of Arabic. Second, despite the growing recognition of PE training, further research is needed to enhance students' proficiency in PE methodologies and to refine Arabic-specific PE strategies. Finally, and most importantly, only a limited number of studies have examined Arabic-English outputs generated by NMT. To address these gaps, the present study investigates NMT efficiency and PE practices in translating an Arabic motivational text, with a specific focus on linguistic, cultural, and stylistic challenges.

3. Key Concepts

3.1. MT: ups and downs

MT has become an indispensable tool in the contemporary multilingual landscape, offering significant potential to support human translators. While MT has limitations and continues to be compared to human translations, recent advancements have significantly enhanced its quality and transformed the field of translation research. NMT, a more sophisticated approach that utilizes recurrent neural networks, has gained prominence due to its ability to capture long-distance relationships and complex word alignment data. This has led businesses such as Google, SYSTRAN, and Microsoft to incorporate it into their translation technologies, reflecting its growing importance in the industry.

NMT is a computational process that involves numerical operations executed by artificial neural networks. These systems convert a given sentence into a series of numerical representations (Pérez-Ortiz et al., 2022). Indeed, NMT relies on neural networks utilizing a recurrent neural architecture based on the encoder-decoder model to generate translations (Rivera-Trigueros, 2022). During the encoding phase, the system processes each word of the source text independently, assigning it a neural representation or embedding. These word-level representations are then aggregated to form a sentence-level representation (Pérez-Ortiz et al., 2022). The entire sentence is subsequently encoded into a fixed-length vector (Rivera-Trigueros, 2022), which integrates individual contextualized representations that adapt dynamically based on their linguistic context. In the decoding phase, this sentence-level representation is incrementally processed to predict each word of the target sentence sequentially. The encoder and decoder responsible for these operations are artificial neural networks that function as interconnected components within a unified neural network architecture (Pérez-Ortiz et al., 2022).

As MT quality has improved, the evaluation process has become more sophisticated. Establishing a universally accepted standard for MT evaluation remains a challenging endeavor due to the diverse linguistic and non-linguistic elements involved. Translation, a complex interplay of language, culture, and technology, lacks standardized assessment criteria, hindering the quantification of translation quality. This difficulty stems from the cognitive,

linguistic, social, cultural, and technical processes inherent in translation (Castilho et al., 2018).

Calude (2003) identifies two primary factors that hinder the establishment of universally accepted MT quality standards: The variability of correct translations and the inherent complexities in assessing translation quality. Human translators often produce varying translations of the same text, making it difficult to construct a definitive benchmark. Additionally, the malleability, continuous nature, creativity, and diverse functions of language cannot be fully captured within a mathematical framework. Consequently, the primary goal of MT has shifted from producing flawless translations to generating useful and contextually appropriate translations that meet the specific needs of users (Zanón, 2005).

The evaluation of MT quality remains a critical area of research, particularly for languages with complex morphological structures like Arabic. MT systems often encounter challenges in handling Arabic, leading to errors in grammatical and linguistic correspondences (Jaber F., 2015). The quality of MT output is influenced by factors such as language pairs, translation direction, domain, and the specific MT system used (Marzouk, 2021). Pre E and PE rules should be adjusted accordingly to optimize MT performance. Arabic, a Semitic language with a complex grammar, extensive morphology, and limited corpora, presents unique challenges for MT systems. Its vast vocabulary, twice that of English (Solyman et al., 2021), contributes to the difficulty of accurate translation. Common errors in Arabic-to-other language MT include deletions, insertions, pronoun errors, subject person inflection errors, word order problems, and plurality errors (Abdelaal and Alazzawie, 2020). Translation strategies at the lexical and cultural levels often follow a sequence of literal, semantic, and communicative approaches (Jibreel, 2023).

3.2. PE: definition, functioning and features

Given the inherent complexity, diversity, instability, and nonlinearity of language, it is imperative to employ specific guidelines and anticipate potential translation errors to ensure accurate text transmission and reader comprehension. This necessitates a flexible and adaptable framework, known as PE. Therefore, PE involves the evaluation of MT output against the original source text, with the objective of rectifying errors to meet predefined quality standards (Guerberof Arenas, 2020). As defined by the International Organization for Standardization, PE is a process aimed at enhancing the quality of MT to render it acceptable (François, 2022). In essence, PE entails modifying a machine-translated or even human-translated text to improve its acceptability, accuracy, and adherence to specified criteria set by the translator, client, or final receiver.

Hence, PE can be categorized based on the extent of editing: complete PE, which yields high-quality human-level translations, and light PE, which produces translations that are ‘good enough’ or understandable but may contain stylistic, grammatical, or spelling issues (Guerberof Arenas, 2020). Blain (2013) distinguishes between these categories, noting that light PE focuses on ensuring accurate semantic translation, avoiding unintended additions or omissions, and maintaining basic spelling conventions while allowing for stylistic modifications. In contrast, complete PE addresses grammar and syntax errors, ensures accurate translation of crucial terminology, and adheres to fundamental principles of spelling, punctuation, and hyphenation.

The choice between full and light PE depends on the desired level of quality and the specific requirements of the text. For texts that require minimal updates and precise stylistic and semantic conveyance, full PE is recommended. Conversely, light PE is suitable for documents with limited access, frequent updates, and a primary focus on clarity and meaning (François, 2022).

The adoption of MT and PE varies across countries, language combinations, and the quality of MT systems. Low-quality MT output, such as free or raw versions, is not suitable for PE (Koponen et al., 2019). The language pairs used significantly influence MT quality and, consequently, the required level and characteristics of PE. Several studies (Bentivogli et al., 2018, among others) have demonstrated that NMT performs better for certain language pairs, such as English-to-French and English-to-German. NMT exhibits fewer errors in morphology, word order, verb forms, and various linguistic features. However, NMT struggles with prepositions, ambiguous words, lexical choices, and specific cases requiring semantic understanding. Morphologically complex language pairs, such as Arabic, present significant challenges for MT systems due to differences in person, gender, tense, orthography, spelling, vocabulary, syntax, and aspect. These disparities necessitate a full PE at multiple levels to address the complexities of Arabic translation (Koponen et al., 2019; Almutawa & Izwaini, 2015).

4. Methodology

This research investigates the effectiveness of PE by focusing on meaning, spelling, grammar, stylistic fluency and adherence to the appropriate terminology. The study was conducted during the spring semester of an MT course at a department of translation. Participants were 27 Master's students specializing in translation, comprising 2 males and 25 females. Their ages ranged from 21 to 22 years old, and their English proficiency levels were categorized as low, moderate, and high.

The gender distribution of the sample (25 females and 2 males) reflects the enrolment demographics of translation programs in Saudi Arabian universities, where female students constitute the majority. While this composition limits the generalizability of findings across gender, the study's primary focus on PE competencies rather than gender-based differences mitigates this concern. Future studies should aim for more balanced samples to explore potential gender-related variations in PE performance.

It is important to note that participants were selected based on specific criteria. The sample consisted of translation students who had completed three years of academic training, during which they studied key MT concepts, including their types, evaluation methods, and related processes such as PE. The sample size was determined by participant availability.

The motivational text genre was selected because it uniquely combines literary devices (metaphors, imagery), religious references (Quranic verses, Hadith), and cultural expressions (proverbs, idiomatic phrases), which collectively present a particularly challenging set of translation problems for NMT systems. Unlike technical or journalistic texts, where errors tend to be primarily terminological or syntactic, motivational texts generate errors at multiple intersecting levels—semantic, cultural, stylistic, and pragmatic—thus providing a rich context for examining the full spectrum of PE challenges.

The selection of a motivational text was based on its incorporation of literary, religious, and cultural elements, which presented specific translation challenges i.e. more potential MT errors and subsequently more PE corrections. Regarding the use of a single 170-word text, this decision aimed to minimize student fatigue and maintain concentration, particularly as the PE task was conducted within a single session. Furthermore, the chosen text contained a sufficient number of challenging elements, enabling a meaningful assessment of both NMT outputs and students' PE strategies.

While these design choices were motivated by practical considerations — minimising participant fatigue and ensuring task completion within a single session — they constrain the generalisability of the findings. This study should therefore be interpreted as a preliminary

investigation that establishes a methodological framework for future research incorporating multiple texts of varying difficulty levels and a larger, more diverse sample.

Participants were tasked with revising the output of a trial version of SYSTRAN Pro, an Arabic-English translated text from Dr. Aaidh ibn Abdullah al-Qarni's motivational work, "Don't Be Sad" (2002). "Don't Be Sad" is a motivational and spiritual self-help book published in 2003, addressing various personal and life challenges in a relatable and predominantly literary religious manner. The book's literary quality and its global success, with over 10 million copies sold worldwide (Omer & Oakes, 2019), highlight its significance.

The text and its machine-translated version were provided to participants during a designated session, with the requirement that they submit their post-edited versions by the end of the session. To facilitate the PE process, participants were encouraged to utilize a comprehensive range of resources, including expert consultations, dictionaries, online concordancers, and other relevant tools. Prior to the task, the researchers informed participants that their performance would be assessed based on fluency, grammatical accuracy, and appropriateness, following the evaluation criteria outlined in the grading rubric. Participants were informed that their corrections would be included in a scientific study intended for publication. The researchers ensured complete anonymity of participants' identities throughout the study and after publication, in strict accordance with prevailing ethical guidelines.

Two experienced evaluators participated in the assessment process. Both hold doctoral degrees in translation studies, with over 10 years of professional experience in Arabic-English translation and MT evaluation. To ensure consistency, inter-rater reliability was assessed using Cohen's *Kappa* coefficient, yielding a value of $\kappa = 0.82$, indicating strong agreement. Discrepancies were resolved through discussion until consensus was reached. The scoring system assigned one point for each correctly identified error and an additional point for each appropriate correction, for a maximum of 80 points. Deductions of 0.5 points were applied for incorrect corrections to penalize misguided interventions while still crediting partial effort.

A data collection table was created, requiring students to identify inaccuracies. The table included columns for error identification, error characterization (fluency, grammar, appropriateness to the context), and error correction. After data collection, a mixed-method approach was employed to address the study questions and gain a deeper understanding of student performance.

A quantitative analysis utilizing statistical tools, including the mean (*M*), standard deviation (*SD*), mode (*Mo*), median (*Md*), and quartiles (*Q*), was conducted to evaluate students' competencies and limitations. These analyses facilitated the identification of similarities and differences in students' PE practices, and revealed underlying attitudes to inform potential solutions.

To validate the descriptive findings, evaluate the equality of student practices, and identify specific distinctions (Morrell & Carroll, 2010), a *one-way ANOVA* was conducted to compare the number of corrections for different error types (Cases 1, 2, 3, and 4). *Tukey's HSD post hoc* test was used to determine pairwise differences in cases of significant findings. All statistical tests were conducted at a significance level of $p < .001$.

The distributed document contained 40 errors. An assessment grid was used to assign a score of one for each identified error and an additional point for each correct rectification, resulting in a maximum possible score of 80 points. Deductions of 0.5 points were applied for incorrect answers.

A content analysis of students' corrections was conducted to assess their practices toward PE. The participants' work was evaluated by experienced translators and educators, who were briefed on the study's scope, objectives, and evaluation criteria. Drawing on their expertise, assessments, and the translation of the book by Faisal Ibn Muhammad Shafeeq (2005), the researchers analysed students' revisions alongside evaluators' feedback. This analysis informed the discussion of results and facilitated the identification of the most suitable translations.

5. Results

The data collected from the edits of students were subjected to an initial analysis. The descriptive statistics of the students' PE scores are presented in Table 1:

Table 1

Descriptive Statistics of Students' PE Scores

Participants	$\bar{X}$	SD	Mo	Md	Q1	Q3	Min–Max
27	20.88	9.63	21	21	14	24.5	5–39.5

The Shapiro-Wilk test confirmed normality of the score distribution ($W = 0.96$, $p = .38$), and Levene's test indicated homogeneity of variance across groups ($F(3, 104) = 1.87$, $p = .14$). These results justify the use of parametric tests (*one-way ANOVA test*) for subsequent analyses.

The mean ($\bar{X} = 20.88$) indicates a low level of proficiency, while the standard deviation (SD = 9.63) suggests a moderate to relatively high spread of scores within the sample. The mode (Mo = 21) shows that a score of 21 occurred most frequently, and the median (Md = 21) reveals that 50% of students scored below 21 out of 80. Together, these values indicate that the scores are distributed symmetrically around the mean, consistent with a normal distribution.

The results indicate that participants possess a modest level of expertise in PE. The mean score was 20.88 out of 80, with 50% (13 individuals) scoring below 21, and 75% of participants scoring 24.5 or below. In addition, none of the participants attained the necessary average score (the best score obtained was 39.5 out of 80). This limitation can be attributed to the insufficient training in PE. While the institution provides modules in MT, it offers none dedicated specifically to PE.

Furthermore, the researchers identified four recurring correction patterns: error detection without correction (Case 1), inappropriate correction (Case 2), partially accurate correction (Case 3), and fully relevant correction (Case 4). The number, mean, and standard deviation of corrections for each case are presented in Table 2.

Table 2

Students' Descriptive Statistics of Each Correction Case

	Minimum	Maximum	Sum	$\bar{X}$	SD
Case 1	0	3	20	0.74	0.94
Case 2	0	6	69	2.56	1.48
Case 3	0	8	72	2.67	2.17
Case 4	1	16	184	6.81	4.21

To assess the presence of statistically significant differences among these cases, an analysis of variance (*one-way ANOVA test*) was conducted, as presented in Table 3:

Table 3

Students' Corrections Cases One-way ANOVA Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	534.991	3	178.330	28.019	< .001
Within Groups	661.926	104	6.365		
Total	1196.917	107			

The *one-way ANOVA test* yielded a large effect size ($\eta^2 = 0.45$), indicating that correction type accounted for a substantial proportion of the variance in correction frequencies.

The results in Table 3 indicate statistically significant differences among the four correction types, confirmed by *Fisher's F-test*, $F(3, 104) = 28.019$, $p < .001$. Since $p < .05$, we can conclude that the null hypothesis is rejected and the alternative hypothesis is accepted. This indicates statistically significant differences among the four correction types. To examine the specific variations in number, the researchers conducted a *Tukey's HSD post hoc test*. The findings are displayed in Table 4:

Table 4

Students' Corrections Cases Tukey's HSD Post Hoc Test

(I) Case	(J) Case	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-1.815*	.687	.046	-3.61	-0.02
	3	-1.926*	.687	.030	-3.72	-0.13
	4	-6.074*	.687	.000	-7.87	-4.28
2	1	1.815*	.687	.046	0.02	3.61
	3	-0.111	.687	.998	-1.90	1.68
	4	-4.259*	.687	.000	-6.05	-2.47
3	1	1.926*	.687	.030	0.13	3.72
	2	0.111	.687	.998	-1.68	1.90
	4	-4.148*	.687	.000	-5.94	-2.36
4	1	6.074*	.687	.000	4.28	7.87
	2	4.259*	.687	.000	2.47	6.05
	3	4.148*	.687	.000	2.36	5.94

* The mean difference is significant at the 0.05 level.

Tukey's HSD post hoc test comparisons were used for the pairwise tests. Case 1 ($M = 0.74$) produced significantly fewer corrections than both Case 2 ($M = 2.56$) and Case 3 ($M = 2.67$), $p = .46$ and $p = .030$, respectively, indicating that students rarely went beyond merely flagging the error. The sharpest contrast separated Case 1 from Case 4 ($M = 6.81$), $p < .001$: fully relevant corrections far outnumbered bare error detection. Case 2 and Case 3 did not differ, $p = .998$, whereas Case 4 significantly exceeded both, $p < .001$. To summarize, the *Tukey's HSD post hoc* analysis revealed that the significant difference found in the *one-way*

ANOVA test was mainly attributable to Case 1 and Case 4, which had a much higher number of adjustments of the latter kind than the other cases. These findings indicate a notably high frequency of corrections in the third case, which could account for the poor scores achieved. Hence, a subsequent examination of the third case was conducted to ascertain the specific nature of the mistakes committed by the participants. The researchers noted that these errors can be categorized into three distinct groups: fluency errors (Type A), grammatical errors (Type B), and translation or appropriateness errors (Type C). A second *one-way ANOVA test* was run to test for differences across the three Case-3 error types (Table 5):

Table 5
Case 3 Types One-way ANOVA Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	18.840	2	9.420	7.659	.001
Within Groups	95.926	78	1.230		
Total	114.765	80			

For the Case 3 error types *one-way ANOVA test*, the effect size was moderate ($\eta^2 = 0.16$), suggesting a meaningful but smaller effect compared to the main correction types analysis.

The results shown in Table 5 indicate that there are statistically significant differences among the three error types. This is supported by the *Fisher F-test* value ($F(2,78) = 7.659$) at a significance level of $p < .001$. The differences across the three error types were statistically significant. To determine the specific locations of these discrepancies, the researchers used a *Tukey's HSD post hoc test*. The findings are displayed in Table 6:

Table 6
Case 3 Types Tukey's HSD Post Hoc Test

(I) Type	(J) Type	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
A	B	.111	.302	.928	-0.61	0.83
	C	-.963*	.302	.006	-1.68	-0.24
B	A	-.111	.302	.928	-0.83	0.61
	C	-1.074*	.302	.002	-1.80	-0.35
C	A	.963*	.302	.006	0.24	1.68
	B	1.074*	.302	.002	0.35	1.80

* The mean difference is significant at the 0.05 level.

The *Tukey's HSD post hoc test* comparisons for the three error types pointed to a single source of variation. Type A ($M = 0.59$) and Type B ($M = 0.48$) were statistically indistinguishable, $p = .928$, so fluency and grammatical errors occurred at comparable rates. Type C ($M = 1.55$), by contrast, stood apart from both, surpassing Type A ($p = .006$) and Type B ($p = .002$) and confirming that appropriateness errors dominated. This confirms that Type C errors occurred markedly more often than Type B errors. In sum, the *Tukey's HSD post hoc* results show that the main contrast lies in Type C, which recorded the highest error count among the three types.

Overall, the participants achieved poor scores due to either a limited number of identified faults or insufficient corrections. These issues were primarily related to translation problems and, to a lesser extent, style and language problems. In the upcoming section, the researchers will conduct a descriptive analysis of the corrections proposed by the participants.

The use of descriptive statistics, one-way *ANOVA*, and *Tukey's HSD post hoc tests* effectively aligns with the study's objectives by providing a comprehensive analysis of students' proficiency in post-editing (PE) tasks. Descriptive statistics, such as the mean score of 20.88 out of 80 highlight students' modest level of expertise, underscoring the need for targeted PE training. The results revealed statistically significant differences, particularly between Case 1 and Case 4, indicating variations in students' ability to correct MT errors effectively. Additionally, the high frequency of Case 3 (partially accurate corrections) suggests that students struggle with refining MT outputs, which was statistically linked to lower overall PE scores. The subsequent section will offer an elaborate elucidation of specific identified mistakes of each category.

5.1. Fluency errors

ST: مزجت بالكدر، و خلطت بالنكد

SYSTRAN Pro output: mingled with cadre, and mingled with grudge

Correction: mingled, mixed

In this instance, SYSTRAN Pro utilized the term “mingle” to convey two distinct verbs: مزجت and خلطت, despite the fact that these verbs are not entirely equivalent. In Arabic, the first term refers to the act of creating a uniform mixture, whereas the second term pertains to the creation of a mixture that is not homogeneous. Regarding the participants, the majority of them observed this mistake and rectified it accurately.

ST: وأنت منها في كبد

SYSTRAN Pro output: you're part of it

Correction: wherein

The error in this particular instance resides in the structure. In this instance, the computer maintained the initial structure, producing a statement that does not adhere to the norms of fluency in the English language. Concerning this fluency error, characterized by the adoption of an Arabic structure due to a literal translation, it was observed that only a small number of participants were aware of it. As a modification, they suggested the following constructions: “you are in it suffering” and “suffering from it.” We consider both options to be inconsequential, as even if the translator chooses a literal translation, the phrase “it makes...” would be more suitable in terms of structure.

These fluency errors reflect what Guerberof Arenas (2020) describes as structural interference from the source language, a common challenge in PE where the post-editor must recognize and restructure calques that NMT systems transfer from the source text's syntactic patterns.

5.2 Grammar errors

ST: ولن تجد والداً أو زوجة، أو صديقاً، أو نبيلاً، و لا مسكناً و لا وظيفة إلا وفيه ما يكدر

SYSTRAN Pro output: You will not find a father or wife, or a friend or a noble man, or a house or a job unless it is provided with enough resources

Correction: You will not find a father or a wife, a friend or a rich man, a house or a job which is free from problems

This is a non-prioritized horizontal list of objects. One can utilize the method outlined above, employing commas and incorporating the conjunction “or” prior to the final part.

ST: وحزن، ثم يصفو الخير كله و الصالح و السرور في الجنة ...أراد الله

SYSTRAN Pro output: God ... sadness, then they describe all good and goodness and happiness in heaven

Correction: God ... sadness, and then

It is crucial to include the conjunction “and” before “then” in order to effectively link the two ideas, guarantee their logical progression, and create a seamless presentation of them. None of the participants have commented on this inaccuracy in this particular case.

ST: ثم يصفو الخير كله و الصالح و السرور في الجنة، و يجمع الشر كله و الفساد و الحزن في النار

SYSTRAN Pro output: they describe all good and goodness and happiness in heaven, and gather all evil and corruption and sadness in hell

Correction: all good, goodness and happiness in heaven, and gather all evil, corruption and sadness in hell

Only two participants noticed that the computer maintained the Arabic structure, wherein coordinating conjunctions are employed to list elements within a phrase. In contrast to English, which employs punctuation such as commas and includes the conjunction “and” before the final part.

The participants’ failure to detect grammatical errors aligns with Budiharto’s (2019) observation that L1 structural interference often goes unnoticed when post-editors rely on intuitive rather than systematic revision strategies.

5.3 Appropriateness errors

ST: منغصة اللذات

SYSTRAN Pro output: infested with breath

Correction: The pleasures of life are short-lived

The source text implies that there is no enduring pleasure or satisfaction in this existence. The machine produced a translation that was completely inaccurate. The machine failed to comprehend the meaning of the word منغصة, which translates to “bothersome”. As a result, it mistakenly suggested the word “infest,” which refers to the act of propagating in a troublesome manner. In this particular instance, the participants proposed corrections of various kinds. One example of a correction is the phrase “Infests joys with sorrow,” which falls under Case 4. Another correction falls under Case 2 and is the phrase “is full of pleasures,” as it has the opposite connotation of the original phrase. Another participant suggested the phrase “full and filled.” This correction falls under Case 3 Type C, as the participant performed an addition.

ST: جاهمة المحيا

SYSTRAN Pro output: infallible

Correction: stuffy

The term “infallible” refers to someone or something that is completely incapable of making mistakes, when the intended meaning is that life is consistently characterized by a negative disposition and abundant difficulties. The author employs a metaphor to liken life to a woman who is stuffy. Consequently, the machine failed to comprehend the significance and

suggested a wholly inaccurate translation. The participants, however, suggested revisions that primarily fall under Case 3 Type C. Their suggestions were mostly literal, such as: life is frowning upon us, life has an angry face, and life has a surly face.

ST: كثيرة التلون

SYSTRAN Pro output: colourful

Correction: wherein change is constant

In this particular instance, the computer failed to differentiate between the literal and metaphorical connotations of the term تلون. Indeed, the term denotes a phenomenon in which an object undergoes a change in color or exhibits several colors. However, in the specific context of this text, it pertains to an individual or entity (such as life) that is distinguished by its fluctuating and unpredictable nature. In a similar manner, the participants put forward corrections for Case 3. However, this time, some participants suggested translations of Type B that were semantically relevant but lacked stylistic appropriateness, such as “sometimes good sometimes bad,” which is syntactically incorrect. Other participants proposed partially accurate translations that failed to capture the precise meaning, such as “has many fluctuations,” “with many faces,” or “unpredictable.”

ST: مزجت بالكدر

SYSTRAN Pro output: mingled with cadre

Correction: mingled with chagrin

In the given case, the machine opted to borrow the word الكدر instead of providing a translation, even though it is not a technical term. The intended message is that life is invidious and characterized by sorrow, disappointment, and bad experiences. The researchers observed that the majority of the participants’ propositions fall into Case 3 Type C. They have either added, such as in the phrase “sadness and anxiety,” or omitted, as in the phrase “mingled and mixed of sorrow,” or qualitatively impoverished, as in “sadness” or “misfortune.”

ST: خلطت بالنكد

SYSTRAN Pro output: mingled with grudge

Correction: mingled with grouch

The intended meaning of this example is that of a situation where an individual is constantly disrupted by the challenges, difficulties, and injustices of life. The machine proposed the term “grudge,” which denotes an individual who harbors anger and negative sentiments against someone who has previously caused him/her damage. Participants suggested less suitable modifications in terms of meaning, such as pessimism, bile, or bitterness.

ST: أنت منها في كبد

SYSTRAN Pro output: you’re part of it with a liver

Correction: hardships are relentless

This example features a religious statement derived from the Quran. The word is mentioned in verse four of Surah Al-Balad, which means “We have certainly created man into hardship,” representing a true test. This nuance is accurately translated as “hardship.” Nevertheless, the machine was unable to understand the meaning and use the verse to provide an accurate translation, instead opting for a literal translation. This translation is explained by

the graphical similarities of “kabad” and “kabid” which share the same script but differ in their diacritical marks, altering their pronunciation and meanings. The phrase expressing the desired connotation is written with a “fatha” and means “with difficulty,” whereas the machine used the word with a “kasra,” which means “liver.” This distinction was overlooked since the diacritical markings were not included in the source text. Participants suggested various suggestions. Case 3 has appeared several times, either by proposing a correction of Type B with an adequate meaning but a grammatical error (tense): “you will suffer”, or by proposing a correction of Type C with a more or less inaccurate meaning: “pain” (qualitative impoverishment). By contrast, the participants provided grammatically accurate and semantically adequate Case 4 corrections, such as: it’s causing great discomfort.

ST: لتتجو رأساً برأس

SYSTRAN Pro output: so that he may escape upside down

Correction: to survive head-to-head

Once more, this is a translation of a cultural expression. Indeed, this adage was spoken by the famous Arab poet Al-Farazdaq. This proverb describes an individual’s profound gratitude and contentment with their existence. The intended meaning in this example is to promote benevolence as a means to prevent bad repercussions, as what goes around comes around. The machine erroneously misunderstood the meaning and proceeded to perform a literal translation without taking the context into account. Participants acknowledged that the term is a cultural phenomenon and aimed to maintain its significance by suggesting the English cultural expression “an eye for an eye.” Furthermore, certain participants opted to elucidate the significance of the term by suggesting “they will do the same to you.”

ST: الجروح قصاص

SYSTRAN Pro output: the wounds are cold

Correction: wounds are legal retribution

In this example, the author utilized a religious expression derived from verse 45 of Surah Al-Maaida. Hence, the intended connotation is a purposeful retribution or harm inflicted as a response to an offense one has caused or committed against another. The software’s algorithm accurately translated only the first word and suggested the adjective “cold” for the second word. Through an online search, it was discovered that the actor Liam Neeson had been in two movies titled “Cold Pursuit” and “Retribution.” Consequently, we hypothesized that the AI-powered system may have mistakenly linked the intended meaning with this particular situation and suggested the term “cold” as a translation. The majority of participants suggested the aforementioned modification, however a few failed to comprehend the intended meaning and instead suggested “wounds will be healed.”

ST: يصفو الخير كله ثم

SYSTRAN Pro output: then they describe all good

Correction: All good is filtered

The machine has confused يصفو “yasfou,” which means “to filter,” and يصف “yasif,” which means “to describe.” The two words are written similarly, with the first word containing one additional letter. Due to the absence of diacritics, the machine failed to comprehend the correct meaning. Every participant identified this error, although they did not reach a consensus on how to rectify it. Some individuals advocated for the verb “to ascend,” while others preferred the verbs “to be” or “to remain.”

ST: لا تسرح مع الخيال

SYSTRAN Pro output: and don't let go of the fantasy

Correction: live according to your reality

In this instance, the machine suggested a meaning that is contrary to the intended one. The phrase “don't let go” denotes the act of holding onto something and not abandoning it, while the original text advises against living or being absorbed in a fictional world and instead embracing reality. The majority of the participants rectified this mistake by suggesting “don't live,” although a few committed grammatical errors, such as using “with” instead of “in” in the phrase “don't wander with imagination.”

ST: وطَّوع نفسك لمعايشتها

SYSTRAN Pro output: and volunteer yourself to live it

Correction: Accept life as it is and adapt accordingly to all circumstances

The machine has wrongly paired the meanings of **وطَّع**, which signifies “adapting and accommodating,” with **وتطَّع**, which denotes “volunteering.” The two words are nearly identical, with the second term containing one additional letter. Due to this similarity, the system failed to comprehend the correct meaning. The majority of the participants identified this error and proposed multiple corrections. Within these suggestions, we have: get used, train yourself and teach yourself. It is evident that these modifications fall under Case 3, Type C, as they do not accurately convey the intended meaning of the original text.

ST: الكمال و التمام

SYSTRAN Pro output: perfection and perfection

Correction: perfection and completeness

The machine iterated the term “perfection” to translate two distinct words. The author has employed a religious phrase borrowed from the Holy Quran once more. In verse 3 of Surah Al-Maaida, the words are used. The machine experienced confusion due to the subtle nuances between the two words. While many people perceive them as interchangeable, there exists a subtle distinction. While both phrases share a similar meaning of removing a deficiency or flaw, **التمام** specifically suggests achieving completeness in addressing a weakness, whereas **الكمال** signifies the enhancement or perfection of something that is already lacking nothing. The majority of participants, with the exception of a handful who omitted the term, rectified it proficiently.

It is important to mention that the machine excluded a Hadith of the Prophet that the author employed. Neither the term “Hadith” nor any synonymous term, nor the actual content of this Hadith is present. The cause for this exclusion remains undisclosed.

Regarding the term **الله**, certain participants expressed a preference for maintaining the cultural and religious aspects of the source language by suggesting the use of the word “Allah” as a gesture of reverence towards the concept of the unity of God. This is due to the fact that any other title, such as God, Lord, Master, etc., can be pluralized, which would undermine the absolute uniqueness of God and pose a challenge to the presentation of Islam's most significant dogma.

These findings support Forcada's (2017) observation that NMT systems tend to produce fluent but semantically inaccurate output, which can mislead post-editors into accepting surface-level correctness at the expense of meaning fidelity. The prevalence of Type C errors further aligns with Mesa-Lao's (2014) finding that PE tasks involve lower ST attention

compared to human translation from scratch, which may explain why participants failed to fully grasp the intended cultural and religious connotations.

6. Discussion

Based on the first outcomes, it can be inferred that the findings can be categorized into two distinct segments. The initial section is dedicated to the translations generated by the machine, while the subsequent section is allocated for the behaviours of the participants. The researchers deemed it necessary to first explain the MT errors before analysing the results related to students' PE practices.

The researchers discovered that the machine has some success in accurately translating the given text and but struggles to comprehend meaning and identify the appropriate rephrasing in the target language. The reasons, as indicated by the examined examples, are numerous and varied. The system struggles to translate figurative language, such as the metaphor *جاهمة المحيا*. The use of metaphor, a stylistic device that relies on imagery, is a significant difficulty for machines and can lead to issues related to intercultural communications (Jaber I.N., 2008). It should be noted that, in order to comprehend the meaning of mental images in the source language, a human translator must contextualise them. This specifically means gaining an understanding of the origins of these ideas within their cultural framework and studying how individuals perceive them as a result of their interactions with the surrounding environment (Jaber I.N., 2008).

In addition, the MT has been unable to accurately translate cultural idioms, such as proverbs like *رأساً برأس*, or religious terms found in the source text. A cultural expression can be defined as the most basic element that conveys a cultural information (Cuciuc, 2019). Religious expressions can take several forms, including those that closely resemble verses or sections of verses, or those that follow a similar structure to the Quran but with little variations that make them somewhat different (Zemni et al., 2022). MT has faced difficulties in comprehending and preserving this particular cultural phenomenon and its theological manifestations. To successfully complete this task, a human translator needs to possess a deep understanding of the cultures associated with both the source and target languages. To effectively understand and interpret the implied meanings of a religious phrase, expression, or concept, as well as accurately relay the intended intentions of the author, one needs to possess a specific level of intercultural competence (Elewa, 2014). Only through this method can readers gain deeper insights into the original culture (religion). We hold the belief that, thus far, MT lacks the capability to achieve this.

The examination of the aforementioned examples has additionally revealed that MT frequently has difficulties in distinguishing words with identical spelling (sometimes homonyms), particularly in the lack of diacritical marks, which are crucial for discerning meaning in Arabic. If the diacritical sign is absent from a term, MT is unable to comprehend its intended meaning and instead selects the most often used word with a similar spelling.

The second section of findings of this study indicates that participants exhibit a modest level of expertise in PE, as reflected in a mean score of 20.88 out of 80. Notably, half of the participants scored below 21, and 75% achieved a score of 24.5 or lower. The analysis of correction patterns revealed four distinct cases: (1) identification of errors without making corrections, (2) provision of inappropriate corrections, (3) partially accurate corrections, and (4) relevant corrections. The significant differences observed in the *one-way ANOVA test* were primarily driven by Case 1 and Case 4. Furthermore, the high frequency of partially accurate corrections (Case 3) likely contributed to the overall poor scores, particularly in Type C errors, where participants identified errors but corrected them inadequately in terms of grammar, fluency, or meaning. Additionally, none of the participants detected grammatical

errors, and although many recognized errors in metaphoric, cultural, and religious expressions, their corrections were semantically relevant yet lacked stylistic and grammatical appropriateness.

These findings align with those reported by Mohammed and Al-Rubai'i (2024), who also found accuracy-related errors to be the most frequent among Iraqi translation students. However, studies on language pairs with greater typological similarity, such as Spanish-French (Valdez & Lomeña Galiano, 2021), report fewer PE challenges, suggesting that the morphological and structural complexity of Arabic significantly amplifies PE difficulty. Future comparative studies across language pairs would help isolate the role of linguistic distance in PE performance.

These results can be attributed to several key factors. First, inadequate training in PE techniques appears to be a major contributing factor. As De Faria Pires (2020) asserts, structured PE instruction can improve students' ability to apply optimal strategies tailored to different text types. Given that students demonstrated difficulty in refining stylistic and grammatical aspects, PE training should emphasize the distinction between light and full PE, as well as the importance of careful source-text (ST) consultation to prevent meaning shifts and omissions. Furthermore, achieving proficiency in PE requires sufficient training, particularly as PE has become an integral component of modern translation processes (Hernandez Morin, 2023).

Second, underdeveloped linguistic competencies significantly influenced participants' performance. As Budiharto (2019) notes, participants' reliance on linguistic structures from their native language (Arabic) likely affected the quality of their translations. This issue is further exacerbated by the absence of dedicated PE courses in their curriculum, despite the central role of PE in ensuring meaning accuracy, grammatical correctness, fluency, and terminology precision (François, 2022). Without systematic linguistic training, students may struggle to produce high-quality post-edited texts, particularly in cases where stylistic refinements are required.

Third, misleading machine translation (MT) proposals may have influenced participants' corrections. Neural machine translation (NMT) systems, while improving fluency, often prioritize readability at the expense of accuracy, resulting in translations that may appear correct but deviate from the intended meaning (Forcada, 2017). This could explain why participants detected errors but failed to correct them appropriately. Furthermore, as Mesa-Lao (2014) and Carl et al. (2011) suggest, PE generally involves lower attention to the ST compared to human translation, reinforcing the need to stress ST consultation during PE training to enhance accuracy.

Fourth, participants demonstrated insufficient familiarity with effective revision strategies. Guerberof Arenas (2019) emphasizes that high-quality PE should preserve as much of the raw MT output as possible while ensuring accuracy, grammatical correctness, and cultural appropriateness. Given that participants failed to detect grammatical errors and struggled with stylistic refinement, their limited exposure to revision strategies likely contributed to these deficiencies. Additionally, PE encompasses multiple verification tasks (François, 2022), requiring systematic training to develop competencies beyond mere error identification.

Finally, a lack of motivation or engagement in the PE process may have played a role in participants' performance. Some translators may correct only major errors, believing that they should retain as much of the MT output as possible, while others may aim to humanize the text by correcting even minor errors (Guerberof Arenas, 2019). The tendency to leave errors

uncorrected despite recognizing them may be attributed to either insufficient training or a lack of motivation to engage deeply with the PE process.

To address these challenges, a multifaceted approach should be implemented. First, integrating comprehensive PE training into curricula is essential. Boukhentache (2024), for instance, stresses the need to acquaint teachers with new translation techniques so that their pedagogical value can be properly demonstrated and adopted. In this respect, structured training modules should focus on both error identification and correction, incorporating hands-on translation exercises into and from Arabic across diverse text types to develop students' cultural, translational, linguistic, and editorial skills. Furthermore, explicit instruction on fluency and grammatical refinement should complement translation accuracy. Adopting ISO 18587 standards or TAUS guidelines can help establish robust training programmes that emphasise key competencies such as linguistic proficiency, research skills, cultural awareness, and technical expertise (Nitzke & Hansen-Schirra, 2021).

Second, linguistic and editing competencies should be enhanced through supplementary courses or workshops on advanced grammar, stylistics, and editing techniques. Comparative analysis exercises and engagement with bilingual corpora, Arabic ones in particular, can improve contextual understanding and refine students' ability to produce high-quality translations.

Third, systematic feedback and reflection should be incorporated into PE instruction. Peer review sessions can encourage collaborative learning, while detailed instructor feedback can highlight areas for improvement. Encouraging self-reflection on PE decisions can help students develop a more critical and analytical approach to PE.

Fourth, the integration of technology and automated evaluation tools can facilitate skill development. AI-based translation evaluation tools can provide real-time feedback on PE accuracy and fluency, while comparative studies between human and machine PE can strengthen students' analytical skills. Familiarizing students with industry-standard PE software can also bridge the gap between academic training and professional practice.

Finally, fostering motivation and engagement is crucial to improving PE performance. Designing practical, real-world PE projects related to professional contexts, such as journalistic or business translation, can increase students' investment in the process. Collaboration with professional translators or internships in translation agencies can further enhance students' motivation and practical experience.

We acknowledge that implementing these recommendations may face institutional constraints, including limited curriculum space and instructor expertise in PE. We therefore propose a phased approach: initially integrating PE modules within existing MT courses, followed by the gradual development of standalone PE courses as institutional capacity allows. Collaboration with translation industry partners could also provide cost-effective access to professional PE tools and real-world training materials.

In conclusion, the findings underscore the necessity of structured PE training, enhanced linguistic competencies, effective revision strategies, and increased motivation to improve students' PE performance. By implementing these targeted solutions, students can develop the necessary skills to engage in professional PE practices, ensuring higher translation quality and greater alignment with industry standards.

7. Conclusion and Future Research

The findings of this study highlight a modest level of expertise in post-editing (PE) among participants, as reflected in their predominant reliance on partial or inappropriate corrections. The high frequency of errors, particularly in grammar, fluency, and meaning, underscores deficiencies in PE training, linguistic competencies, and revision strategies. Additionally, participants demonstrated limited engagement with source-text consultation and a lack of familiarity with effective revision techniques, which further hindered their performance. These challenges necessitate a comprehensive approach to PE instruction, including structured training modules, enhanced linguistic and editorial skill development, and systematic feedback mechanisms. Incorporating technology-driven evaluation tools and fostering greater motivation through practical, real-world projects can also improve students' engagement with PE tasks. By implementing these targeted solutions, students can acquire the essential skills required for professional PE, ensuring higher translation accuracy and alignment with industry standards. From a practical standpoint, the study's main contribution lies in informing the design of PE training curricula for Arabic translation programmes, offering concrete guidance for the development of structured modules, revision protocols, and assessment criteria tailored to the specific challenges identified.

Future research should employ larger and more diverse participant samples, incorporate multiple text types (e.g., journalistic, legal, technical), and compare PE performance across different NMT systems to enhance the external validity of the findings. Additionally, cross-linguistic comparative studies examining PE performance across different language pairs would help isolate the specific role of Arabic's morphological complexity in PE difficulty. Methodologically, future studies could employ eye-tracking experiments and think-aloud protocols to provide deeper insights into the cognitive processes underlying PE decisions.

The current study contributes to the understanding of PE when dealing with Arabic and serves as a foundation for future research in this area.

References

- Abdelaal, M. N., & Alazzawie, A. (2020). Machine translation: The case of Arabic-English translation of news texts. *Theory and Practice in Language Studies*, 10(4), 408–418. <https://doi.org/10.17507/tpls.1004.09>
- Afzaal, M., Imran, M., Du, X., & Almusharraf, N. (2022). Automated and human interaction in written discourse: A contrastive parallel corpus-based investigation of metadiscourse features in machine-human translations. *Sage Open*, 12(2), 1–10. <https://doi.org/10.1177/21582440221142210>
- Aldawsari, H. A. H. (2023). Comparing the performance of Google Translate and SYSTRAN on Arabic lexical ambiguity. *Arab World English Journal for Translation & Literary Studies*, 7(3), 19–34. <https://doi.org/10.24093/awejtls/vol7no3.2>
- Aldelaa, A. S., & Malkawi, M. A. K. M. (2024). Investigating problems related to the translation of idiomatic expressions in the Arabic novels using neural machine translation. *Theory and Practice in Language Studies*, 14(1), 71–78. <https://doi.org/10.17507/tpls.1401.08>
- Almekhlafi, H. A., & Nagi, K. A. (2024). Fine-grained evaluation of English to Arabic Neural Machine Translation: A case study of educational research abstracts. *Al-Andalus Journal for Humanities & Social Sciences*, 95(11), 198–218.
- Almutawa, F., & Izwaini, S. (2015). Machine translation in the Arab World: Saudi Arabia as a case study. *trans-kom*, 8(2), 382–414.
- Al-Qarni, A. (2002). *La tahzan [Don't be sad]*. Maktabat Al-Obeikan / Maktabat Ibn Hazm.
- Al-Qarni, A. I. A. (2005). *Don't be sad* (F. Ibn Muhammad Shaffeq, Trans.). International Islamic Publishing House.

- Balashov, Y. (2022). The boundaries of meaning: A case study in neural machine translation. *Inquiry: An Interdisciplinary Journal of Philosophy*, 68(7), 1651–1684. <https://doi.org/10.1080/0020174X.2022.2113429>
- Benková, L., & Benko, L. (2020). Neural machine translation as a novel approach to machine translation. Paper presented at *DIVAI 2020*, Štúrovo, Slovakia.
- Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2018). Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Computer Speech & Language*, 49, 52–70. <https://doi.org/10.1016/j.csl.2017.11.004>
- Blain, F. (2013). *Modèles de traduction évolutifs* [Unpublished doctoral thesis]. Le Mans University. <https://theses.hal.science/tel-01142926>
- Boukhentache, S. (2024). Evaluating translation in the Algerian English language public school textbooks. *Journal of Studies in Language, Culture, and Society*, 7(3), 26–34.
- Budiharto, A. (2019). Native language interference on target language writings of Indonesian EFL students: An exploratory case study. *Indonesian EFL Journal*, 5(1), 107–116. <https://doi.org/10.25134/ieflj.v5i1.1630>
- Calude, A. S. (2003). Machine translation of various text genres. *Linguistic Society of New Zealand (Inc.)*, 46, 67–94.
- Canfora, C., & Ottmann, A. (2020). Risks in neural machine translation. *Translation Spaces*, 9(1), 58–77. <https://doi.org/10.1075/ts.00021.can>
- Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment* (pp. 9–38). Springer.
- Costa-Jussà, M. R., Cross, J., Çelebi, O., et al. (2024). Scaling neural machine translation to 200 languages. *Nature*, 630(8018), 841–846. <https://doi.org/10.1038/s41586-024-07335-x>
- Cotelli Kureth, S., & Summers, E. (2025). Policy on the use of Machine Translation (MT): A good model for wider policies on Generative AI (GenAI)? *Journal of Studies in Language, Culture, and Society (JSLCS)*, 8(1), 30–48.
- Cuciuc, N. (2011). Traduction culturelle : transfert de culturèmes. *La Linguistique*, 47(2), 137–150.
- Elewa, A. (2014). Features of translating religious texts. *Journal of Translation*, 10(1), 25–33.
- François, F. (2022). Post-édition, mode d'emploi. *Traduire*, 246, 5–9. <https://doi.org/10.4000/traduire.2754>
- Guerberof Arenas, A. (2020). Pre-editing and post-editing. In E. Angelone, et al. (Eds.), *The Bloomsbury companion to language industry studies* (pp. 333–360). Bloomsbury Academic.
- Jaber, F. (2015). The landscape of translation movement in the Arab World. *Arab World English Journal (AWEJ)*, 6(4), 128–140.
- Jaber, I. N. (2008). Translating metaphoric expressions: Translating culture. *Journal of College of Education for Women*, 19(2), 1–16.
- Jibreel, I. (2023). Online machine translation efficiency in translating fixed expressions between English and Arabic. *Theory and Practice in Language Studies*, 13(5), 1148–1158.
- Koponen, M., Salmi, L., & Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, 33(1–2), 61–90.
- Li, Z., Gao, S., Li, X., & Asghari, H. B. (2023). Strengthening research of SYSTRAN in the field of artificial intelligence automatic translation. In J. H. Abawajy, et al. (Eds.), *Proceedings of ICATCI 2022* (pp. 626–633). Springer.

- Marzouk, S. (2021). An in-depth analysis of the individual impact of controlled language rules on machine translation output. *Machine Translation*, 35, 167–203.
- Marzouk, S., & Hansen-Schirra, S. (2019). Evaluation of the impact of controlled language on neural machine translation. *Machine Translation*, 33(3), 179–203.
- Mohamed, S. A., Elsayed, A. A., Hassan, Y. F., & Abdou, M. A. (2021). Neural machine translation: Past, present, and future. *Neural Computing and Applications*, 33, 15919–15931.
- Mohammed, S. S., & Al-Rubai'i, A. M. A. (2023). Post-editing of machine translation of an English-into-Arabic text by translation students. *The Journal of the University of Duhok*, 26(1), 1227–1235.
- Morrell, P., & Carroll, J. (2010). *Conducting educational research: A primer for teachers and administrators*. Sense Publishers.
- Nitzke, J., & Hansen-Schirra, S. (2021). *A short guide to post-editing*. Zenodo. <https://doi.org/10.5281/zenodo.5646896>
- Omer, A., & Oakes, M. (2019). Writing styles of Salwa and Al-Qarni. In M. El-Haj, et al. (Eds.), *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics* (pp. 16–21). Association for Computational Linguistics.
- Pérez, C. R. (2024). Re-thinking machine translation post-editing guidelines. *JoSTrans: The Journal of Specialised Translation*, 41, 26–47.
- Pérez-Ortiz, J. A., Forcada, M. L., & Sánchez-Martínez, F. (2022). How neural machine translation works. In D. Kenny (Ed.), *Machine translation for everyone* (pp. 141–164). Language Science Press.
- Pires, L. F. (2020). Master's students' post-editing perception and strategies. *FORUM*, 18(1), 26–44.
- Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*, 56(2), 593–619.
- Sammarraie, H. S. M. A., & Ghassemiazghandi, M. (2024). Enhancing post-editing machine translation skills among Iraqi undergraduate students. *Arab World English Journal for Translation and Literary Studies*, 8(2), 165–182.
- Solyman, A. (2021). Synthetic data with neural machine translation for automatic correction in Arabic grammar. *Egyptian Informatics Journal*, 22(3), 303–315.
- Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69, 343–418.
- Valdez, C., & Lomeña Galiano, M. (2021). Exploration de la traduction automatique neuronale espagnol-français. *Traduction et Langues*, 20(1), 86–112.
- Zanón, N. T. (2005). Evaluating the output quality of machine translation systems: SYSTRAN 4.0. *Philología Hispalensis*, 19(1), 189–201.