k@ta, Vol. 27, No. 1, June 2025, 18-33
ISSN 1411-2639 (Print) / 2302-6294 (Online)

DOI: https://doi.org/10.9744/kata.27.1.18-33

OPEN ACCESS
k@ta
https://kata.petra.ac.id

Machine Translation Quality Evaluation and Post-Editing Efficiency: The Case of *Abadis Translator*

Hamidreza Abdi
Universitat Pompeu Fabra, Barcelona, SPAIN

Article Info:
Received: 4 December 2024
Revised: 15 January 2025
Accepted: 23 April 2025
Published Online: 30 June 2025

Keywords:
Machine translation,
Abadis translator,
ChatGPT,
Machine translation evaluation,
Wilss' matrix

Corresponding Author:
Hamidreza Abdi
ORCID iD: 0000-0002-0164-2887
Universitat Pompeu Fabra,
SPAIN
Email: hamidreza.abdi@upf.edu

ABSTRACT

Machine translation evaluation ensures that systems meet both linguistic and functional standards, providing end-users with reliable tools. This study examines the domestic machine translation system, *Abadis Translator*, focusing on the syntax, semantics, and pragmatics included in Wilss' (1982) matrix. An 80-statement translation test was developed, consisting of 60 instrumental, descriptive, and argumentative statements, along with 20 idiomatic statements to evaluate pragmatic accuracy. The statements were translated using *Abadis Translator* and assessed using Wilss' criteria. A Likert-scale questionnaire, with options such as *incorrect, inappropriate, undesirable, correct,* and *appropriate*, was employed. ChatGPT was used to post-edit the incorrect and inappropriate translations to enhance quality, helping to assess the reliability of both *Abadis Translator* and ChatGPT as an Artificial Intelligence (AI) editor. The results demonstrated that *Abadis Translator* exhibited strong grammatical accuracy, although it occasionally encountered subject-verb agreement issues in complex sentences. While it achieved moderate success in semantic translation, it struggled with pragmatic nuances, especially with idiomatic expressions and participle constructions. Nevertheless, post-editing with ChatGPT significantly improved overall translation quality by correcting grammatical errors and clarifying implied meanings.

INTRODUCTION

Machine translations (MTs) are popular computer-aided translation (CAT) tools that translators use to convert texts from one language to another. MTs are the oldest sub-disciplines and computer applications that examine the nature of language (Nirenburg, as cited in Abdi, 2016). Fauziah and Putri (2023) describe MT as a task that predicts the best target language (TL) sentence that matches the meaning of a source language (SL) sentence.

While recent advancements in deep neural networks have significantly enhanced MT performance (Lyu et al., 2024), critical issues persist. These include errors in terminology, syntax, and semantics that reduce translation quality, as highlighted by Makhachashvili et al. (2023). Moreover, global MT developers often neglect languages like Persian, prioritizing widely spoken languages such as English, French, and German. For Persian MTs that do exist, their quality often lags behind, posing challenges for users and translators in Iran who depend on reliable translations for professional and personal purposes.

In response to this gap, Iranian developers have created domestic MT systems such as *Targoman Translator*, *Abadis Translator*, and *Fastdic Translator*. These tools aim to provide localized solutions, yet little research has assessed their effectiveness compared to leading MTs like Google Translator or Microsoft Bing Translator. The extent to which these domestic MTs meet quality standards and address the linguistic needs of Persian-speaking users remains largely unexplored.

Previous studies have primarily relied on automated evaluation metrics such as BLEU, NIST, and METEOR (Wang, 2019), which offer efficiency but lack strong correlation with human judgments of translation quality (Gala et al., 2023). Human evaluation is generally considered a golden standard to understand how good a translation is, because it gives us a more thorough understanding of whether the translation is accurate, fluent, and acceptable (Han, 2018). A framework created by Wilss (1982), which assesses syntax, semantics, and pragmatics, is especially useful for getting a complete picture of translation quality. Nevertheless, not many studies have used this framework to evaluate Persian MT systems, leaving a gap in understanding their strengths and weaknesses.

This research tries to fill in these gaps by using Wilss's (1982) framework to evaluate *Abadis Translator*, paying close attention to its syntactic, semantic, and pragmatic dimensions. It also investigates whether ChatGPT can be helpful as a post-editor to boost translation quality and possibly serve as a cost-effective alternative to human editors. By carefully examining *Abadis Translator* and checking out ChatGPT's possible benefits, this study aims to help make this domestic MT system better and emphasizes how important post-editing is for fixing translation problems.

This research is part of a broader project aimed at evaluating the performance of Iranian software developers in creating local MT tools, such as *Targoman Translator*, *Abadis Translator,* and *Fastdic Translator*, with a focus on their quality and the need for post-editing. In this study, the results of *Abadis Translator* were presented. To achieve the objectives of the present study the following questions are raised:

- What was the output quality of *Abadis Translator* in terms of syntax, semantics, and pragmatics according to Wilss' (1982) matrix?
- Was ChatGPT deemed a reliable and valid online Artificial Intelligence (AI) editor for end-users and translators, potentially replacing human editors?

REVIEW OF LITERATURE

A Brief History of MT Evaluation

MT evaluation presents a significant challenge for evaluators, as they must assess the quality of MT output in terms of adequacy, acceptability, and comprehensibility to determine its overall effectiveness. In this context, Dorr et al. (2010) note that MT evaluation relies on the systematic assessment of a new system's quality to demonstrate its superiority over existing systems. Chan (2004) views MT evaluation as largely enjoyable, as it involves measuring the system's quality and effectiveness.

The history of MT evaluation goes backs to 1954, with some pretty big moves from Georgetown University and IBM. However, as Hutchins and Somers (1992) pointed out, people got a little too excited about MT's potential early on as its output failed to meet the expectations of clients, users, and developers. They argue that during this early phase, it was difficult to identify effective methods and technologies for evaluating MT. Orr and Small's (1967) research found that MT-produced translations lacked quality due to issues with accuracy, readability, and usefulness for target readers. They concluded that MT evaluation at that time was not worthwhile.

By contrast, van Slype (1979) introduced a more comprehensive framework for evaluating MT systems by considering both macro and micro levels. At the macro level, the evaluation aimed to address the needs of potential users by focusing on four key dimensions: *cognitive*, *economic*, *linguistic*, and *operational*. On the micro level, the evaluation sought to identify errors in MT output and propose corrective measures.

Over time, various methods for MT evaluation have been proposed by researchers and scholars. For instance, the Defense Advanced Research Project Agency (DARPA) (1990) conducted a comparative evaluation of two MT systems, CANDIDE and PANGLOSS, by comparing their output with human translations. White et al. (1993) evaluated this method in terms of its "sensitivity of measurement, efficiency, and the human time and effort required" (p. 208). Another evaluation method, the KANT method, was developed by Nyberg et al. (1994) based on a knowledge-based approach to MT, with a focus on error analysis. Their evaluation considered two main paradigms: correctness and style. The error analysis included assessment of analysis coverage, analysis correctness, generation coverage, and generation correctness.

Hirschman and Thompson (1997) came up with a unique method to evaluate MTs based on three main areas. First, there is adequacy evaluation, which is all about understanding what the end users really need, taking into account things like cost and its reliability. Then there is diagnostic evaluation, which is like giving the MT system a check-up by looking at its output. The third area, performance evaluation, , measures how well the MT system does its job in specific translation tasks. The main reason for evaluating MT is to figure out how good the translations are, pointing out both the strengths and weaknesses to inform end users.

Abadis Translator

Abadis started its activity in the Information Technology (IT) world with its first endeavor in 2005: building the *Abadis Dictionary* website (https://abadis.ir/translator/). This online dictionary has seen numerous enhancements over the years, evolving with new additions. The Abadis group also created a convenient browser extension for Firefox and Chrome users, a mobile app, and even a Telegram bot for dictionary and translation needs.

The *Abadis Dictionary* features a diverse collection of sections. It includes multilingual dictionaries covering English, Farsi, and Arabic, along with more than twenty specialized dictionaries. It includes features such as English and American pronunciation, translation of abbreviations and acronyms, and an online text translator. It also provides access to renowned resources such as the Dehkhoda and Moin dictionaries, and other dictionary types. Users can also explore a dictionary of synonyms and antonyms, Persian equivalents of foreign words, phrases approved by the Persian Language and Literature Academy, and a general encyclopedia.

For local MT development, applying Wilss' (1982) framework could enhance the quality of Persian translations, making them more accurate and contextually appropriate. On a global scale, refining *Abadis* through such an evaluation could contribute to the advancement of MT for Persian and other regional languages, setting a standard for addressing linguistic complexities in MT and improving the broader MT landscape.

Matrix for MT Evaluation

Trancism, a term coined by Abdi (2021b) and used as an alternative to Translation Criticism, is a contentious issue among scholars in Translation Studies, as Ntamwana and Munandar (2024) imply. Abdi (2024) argues that trancism involves a thorough examination of both the source and target texts, conducted using a structured model that equips the critic with comprehensive instructions to arrive at an objective judgment.

Wilss (1982) developed a framework for *trancism* that encompasses three key aspects: syntax, semantics, and pragmatics, providing critics with a systematic method to evaluate translated texts. Wilss (1982) believes this approach is beneficial for evaluating MTs since it mirrors the process of translating from the SL to the TL. This process entails decoding the SL and then encoding the TL through technological and algorithmic methods. However, he notes that this approach often falls short, as algorithmic simulations tend to be incomplete, representing more of a promise than a fulfilled reality.

According to Wilss (1982), the initial efforts in MT evaluation were characterized by a lexical approach supplemented by a few syntax rules oriented towards the TL, which was viewed as a success and raised inflated expectations. However, these expectations proved unrealistic due to the ambiguous issues related to interlingual structural divergences, which present computers with more complex challenges than those faced by human translators in Source Language text (SLT) analysis. This discrepancy arises because computers can only employ a text-internal approach based on logically explicit instructions, whereas human translators can integrate both text-internal and text-external considerations when working with the text to be translated.

The issue was largely addressed by integrating both syntactic and semantic approaches, leading to an analysis of deep-structural language levels. This method offers a significant advantage in that it is not restricted to a specific language pair and can be applied across various pairs. However, the pragmatic challenge in MT persists, particularly in translating statements that contain implied meanings or complex structures, including what Wilss (1982) refers to as participle constructions. He suggests that any participle construction requires cognitive interpretation to fill in the missing semantic determinants, which is essential for establishing a semantically unambiguous connection between the participle construction and its superordinate clause.

Wilss (1982) argues that recognizing the communicative intent of a sentence with an embedded participle construction is more challenging for computers than for human translators for two main reasons: the limited combinatorial power of the human translator's mind and the absence of extralinguistic or situational information. These factors are critical for ensuring a coherent link between recognition and application during the translation process. To address this issue, MT research has sought to employ techniques like the surface replacement of SL sign combinations with their corresponding TL counterparts.

According to Wilss (1982), for MT research to achieve translations that are considered good enough, it is essential to develop a program that combines syntactic analysis and synthesis with a complementary semantic program. This integrated approach would provide the computer with the necessary information to effectively compensate for its lack of linguistic competence, thereby creating the conditions for adequate MT results.

Wilss (1982) laid out very systematic criteria to pinpoint linguistic problems in MT outputs. He centered his framework around three fundamental linguistic aspects: syntax, semantics, and pragmatics. This framework provides a deep dive into MT output evaluation. It starts by checking the syntax, making sure the translation follows the grammar rules of the TL. Then, it moves on to the semantic side, confirming that the SL text's meaning is accurately brought over into the TL. Finally, Wilss (1982) tackles pragmatics, which is all about grasping the communication purpose and context of the translated text. By looking at these three layers, his framework is great at spotting typical linguistic snags like structural differences, ambiguities, and contextual misunderstandings. It provides a thorough method to assess how well an MT system copes with tricky translation tasks across different languages although this approach has its advantages when it comes to handling various language pairs.

Recent Studies in the Field

MT has recently become a subject of significant interest among researchers, leading to an increase in both empirical studies and comprehensive reviews. A notable contribution comes from Mercan et al. (2024), who investigated in depth the historical development of MT evaluation, considering the ethical implications. They found that MT has evolved through different stages, with each phase bringing new techniques and ideas that have greatly enhanced our understanding of MT and computer-assisted translation (CAT) tools. Deng and Yu (2022) also carried out a thorough review of MT in language learning, pinpointing undergraduate and graduate students as the main users. Their work uncovered that both teachers and students hold mixed views about MT, suggesting that incorporating MT into language education should be incorporated systematically, beginning with an introduction, proceeding to a demonstration, assigning tasks, and concluding with reflection on the experience.

Tan et al. (2020) contributed to the literature by examining various approaches to neural machine translation (NMT). They focused on design aspects of the architecture, different decoding methods, and ways to augment data. Their analysis highlighted key challenges within the NMT landscape, including a deeper understanding of NMT mechanisms, developing improved architectures, effectively utilizing monolingual data, and incorporating prior knowledge This research significantly underscores the complexity involved in improving NMT technology and the importance of continued investigation into these topics. Moreover, a study by Tosun (2024) looked into whether bilingual Turkish and English speakers who are not professionals can assess the accuracy of MT outputs. The study found that late bilinguals were more adept at detecting translation accuracy than their early bilingual peers, particularly regarding firsthand evidence translations.

Additionally, Yan et al. (2023) provided insights into the performance of various automatic metrics used in training MT systems, revealing robustness issues linked to adversarial translations and distribution biases in training datasets. Zerva and Martins (2024) addressed the biases inherent in estimated confidence intervals in translation quality assessment, demonstrating that current uncertainty quantification methods often underestimate uncertainty. Their findings suggested that employing conformal prediction can enhance coverage accuracy and fairness in evaluating translation quality across different language pairs, highlighting the necessity for improved methodologies in the field.

The contributions of Iranian researchers in the MT domain, though limited in number, are noteworthy. Aghai's (2024) study examined the extent to which automatic translation systems like ChatGPT and Google Translate render literary translations from Persian into English. He highlighted the challenges these systems have in capturing the true meaning and cultural subtleties found in literature. Following a comparable line of inquiry, Safari et al. (2023) explored the precision with which Google Translate and Bing Translator convert Persian to English. Their findings indicated that Google offered more dependable translations, especially in medical contexts. This research highlights the essential role of humans in ensuring translations are both accurate and culturally sensitive.

Several studies have delved into the effectiveness of MT in various contexts. Sartipi et al. (2023), for instance, examined how well different parallel corpora worked for translating between Persian and English, and they found that the BLEU scores varied significantly depending on the dataset used. Mirzaeian and Oskoui (2022) looked into what Iranian EFL student teachers thought about using MT in academic language learning, discovering that many of them felt at ease using these tools. Abdi (2021a) investigated whether Google Translate might replace human translators and determined that although it does a decent job in some areas, it still falls short in terms of producing natural-sounding translations.

Finally, Saghayan et al. (2021) studied the effect of MT on the ability to detect fake news, showing that it actually made classifying news correctly more difficult, especially when dealing with multiple categories. Collectively, these studies underscore the evolving role of MT in language education and its implications for translation quality and accuracy.

What distinguishes this study from prior research is its emphasis on the local machine MT system, *Abadis Translator*. It evaluates the output quality of *Abadis Translator* across three key areas: syntax, semantics, and pragmatics—an aspect that has not been explored in earlier studies. Furthermore, this research also considers another significant feature of this domestic MT: assessing its effectiveness in post-editing. This presents a critical challenge for ChatGPT, namely whether it can be regarded as a reliable online AI editor for end-users and translators.

METHOD

Participants

The participants in this study were evaluators invited to evaluate the output quality of the domestic MT system, *Abadis Translator*. The 20 evaluators were randomly selected from professional translators who agreed to participate in the study and specialized in English-Persian and Persian-English translations, each with over five years of translation experience. These translators were selected from the following websites: www.iacti.ir/members.html and www.proz.com. They were informed about the study's objectives, the research topic, and the vital role they would play in achieving the desired outcomes and enhancing the functionality of the domestic MT system.

Instrument

A translation test consisting of 80 statements was created for data collection. The 60 statements were selected according to Mistrík's (1997) classification of text types, which includes narrative, descriptive, and argumentative categories. These text types lend themselves to distinct linguistic and stylistic features that can be effective in evaluating the translation quality of any MT system. Each category comprised 20 statements drawn from various sources. Narrative statements were extracted from Rowling's (1997) *Harry Potter and the Sorcerer's Stone*, descriptive statements were sourced from Burke and Maxwell's (2012) *Lonely Planet: Iran (Travel Guide)*, which covers diverse aspects of Iranian culture and tourism, and argumentative statements were taken from scholarly articles, including works by Abdi (2021a), Abdi (2021b), and Dalaslan (2015). To assess *Abadis Translator*'s pragmatic accuracy, 20 idiomatic statements with implied meanings were added, sourced from *Book of Idioms* (Defense Language Institute English Language Center, 2011) that features various current American English idiomatic expressions.

The test underwent validation by a panel of experts who evaluated the statements to ensure their appropriate selection and organization, as well as their effectiveness in testing the output quality of *Abadis Translator*. In this way, they were asked to evaluate the test based on three paradigms, namely domain coverage, relevance, and clarity. This panel consisted of five specialists in the relevant fields who were conveniently selected and had more than five years of experience in teaching translation. Based on their feedback, revisions were implemented, including reordering and replacing certain statements.

For the pilot validation phase, 15 translators with similar characteristics to the evaluators, including specialization and experience, were selected. This number was considered sufficient to yield meaningful insights within the study's time constraint. The test was then given to 15 translators, and we calculated how their responses lined up with the overall score. The test was then given to 15 translators, and the alignment of their responses with the overall score was calculated. The results indicated the correlation coefficients ranged from a low of 0.54 to a high of 0.83, with an average of 0.76. All of these were within the good range of ±0.50 to ±1, suggesting that the test is sufficiently accurate at measuring what it is intended to. Furthermore, every item demonstrated p-values below 0.05 ($p < 0.05$), providing additional confirmation of the test's overall validity.

Data Collection and Analysis

For data collection, the translation test was initially translated from English into Persian using *Abadis Translator*. Both the Persian translations and the original 80 English statements were then distributed to the evaluators, who assessed the output quality of *Abadis Translator* based on three paradigms: syntax, semantics, and pragmatics, following Wilss's (1982) trancism/evaluation matrix.

The evaluators utilized a Likert-scale questionnaire with five response options: wrong, inappropriate, undecidable, correct, and appropriate. The Likert-scale responses were likely weighted on a scale from 1 to 5 to reflect the evaluators' subjective judgments of translation quality. Specifically, Wrong (1) indicated the lowest quality, Inappropriate (2) represented slightly better than Wrong, but still a poor translation, Undecidable (3) suggested that the evaluator could not decide on the translation's quality, Correct (4) showed a good translation, but not necessarily perfect, and Appropriate (5) signified the highest quality translation. This scale facilitated the evaluation of how effectively the domestic MT system produced accurate translations concerning syntax, semantics, and pragmatics. The agreement among the 20 raters was measured using Fleiss' kappa (k).

Additionally, ChatGPT was employed to post-edit the statements classified as wrong, or inappropriate by the evaluators, to assess whether the quality of these translations could be improved. In this way, such statements, alongside their Persian translations, were provided to ChatGPT for post-editing. Then, the post-edited outputs were reviewed for quality before evaluation. This involved a comparison was made between the original wrong and inappropriate Persian statements and their post-edited versions to determine whether there were improvements in terms of syntax, semantics, and pragmatics.

Data analysis included calculating the percentage and mean score for each category, which were then presented in tables. Moreover, the Wilcoxon signed-rank test was used because of its effectiveness in comparing pairs of observations across two related situations. This is especially useful when working with ordinal data or when the data does not meet the rules for parametric tests, such as the paired t-test. Using the Wilcoxon signed-rank test provides confidence that any observed improvements are not merely attributable to random chance, but are statistically meaningful, a key objective of this study. Thus, the test was performed separately for syntactic, semantic, and pragmatic paradigms to determine whether the evaluators' assessments of the Persian translations generated by *Abadis Translator* were significantly consistent within each paradigm.

RESULTS AND DISCUSSION

Inter-Rater Reliability Test

Fleiss' kappa (k) is designed for multiple raters, taking chance agreement into account. Given the involvement of 20 raters, agreement was established among them. After conducting the reliability analysis with SPSS statistics, Table 1 represents the k value and associated statistics. The test results showed that the k value was 0.681, thus demonstrating good strength of agreement. Additionally, the level of p value was less than 0.05, namely 0.000 ($p < 05$). This shows that the k coefficient was indeed significantly different from 0 (zero).

Table 1. Summary of interrater reliability tests among the 20 raters

Overall Kappa						
	k	Asymptotic Standard Error	*z*	*p*	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
Overall	.681	.78	6.729	.000	.253	.752

Syntactic Evaluation

As shown in Table 2, the evaluators expressed their agreement regarding the syntactic correctness of the translations produced by *Abadis Translator* for both narrative and descriptive statements. Specifically, *Abadis Translator* accurately translated 34% of narrative statements and 25% of descriptive statements. Additionally, 66% of narrative statements and 75% of descriptive statements were deemed appropriate by the evaluators. Table 2 further indicates that the evaluators agreed on the correct translation of more than half (54%) of the argumentative statements. Overall, the evaluators concurred with the syntactic correctness of a significant majority (85%) of the 60 Persian translations generated by *Abadis Translator.*

Table 2. Frequencies and percentages of the evaluators for syntactic evaluation of the 60 statements

Type of Statements	Wrong		Inappropriate		Undecidable		Correct		Appropriate		*N*		
	f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*M*
Narrative	-	-	-	-	-	-	136	34.0	264	66.0	400	100.0	4.65
Descriptive	-	-	-	-	-	-	98	25.0	302	75.0	400	100.0	4.75
Argumentative	69	17.0	107	27.0	7	2.0	144	36.0	73	18.0	400	100.0	3.13
Total	69	6.0	107	9.0	7	1.0	378	32.0	639	53.0	1200	100.0	4.17

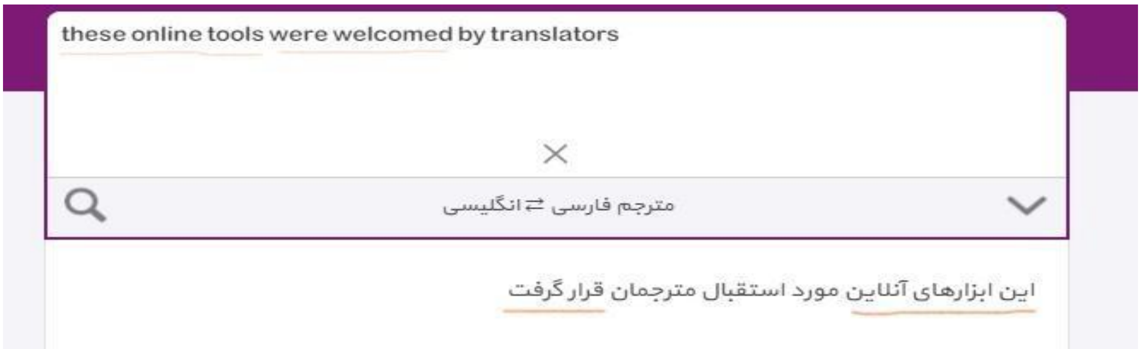
A one-sample Wilcoxon signed-rank test was conducted to assess whether there was a significant relationship between the evaluators' assessments and the syntactic correctness of the Persian translations produced by *Abadis Translator* for each type of statement. The results of this test, presented in Table 3, indicate that the *p*-values for all types of statements were below .05 ($p < .05$). Additionally, the average scores for each kind of statement were more than the midpoint of 2.5, showing that the evaluators mostly agreed with the grammatical rightness of the Persian translations. The exact average scores were: narrative statements (MN = 4.65), descriptive statements (MD = 4.75), and argumentative statements (MA = 3.13), all pointing to higher-than-average grammatical correctness.

Table 3 also reveals that the *p*-value for the general evaluation of the 60 Persian translations was 0.0, which is below .05 ($p < .05$). The overall average score for all the statements was 4.17 out of 5, going beyond the midpoint (4.17 > 2.5). This implies that, generally speaking, the evaluators largely concurred about the mainly grammatical accuracy of the Persian translations produced by *Abadis Translator.*

Table 3. One-sample Wilcoxon signed-rank test for Syntactic evaluation of the 60 statements

Statement Types	*N*	*MDN*	*p*
Narrative	20	5	0.003
Descriptive	20	5	0.000
Argumentative	20	4	0.002
Total	60	5	0.000

The greatest problem for *Abadis Translator* was arguably issues related to subject-verb agreement in complex sentences. As Figure 1 shows, the MT system incorrectly translated the original statement *these online tools were welcomed by translators* into *این ابزارهای آنلاین مورد استقبال مترجمان قرار گرفت.* This occurred because the MT used the singular verb *قرار گرفت* (was welcomed) for the plural subject *این ابزارهای آنلاین* (these online tools). *Abadis Translator* similarly erred when it translated the SL item *translations Google Translate produced were acceptable* into *ترجمه های تولید شده توسط گوگل ترنسلیت قابل قبول بوده است.* Here, it employed the singular verb *قابل قبول بوده است* (was accepted) with the plural subject *ترجمه های* (translations). Such grammatical errors, as those previously identified, are frequently observed in *Abadis Translator*.



these online tools were welcomed by translators

این ابزارهای آنلاین مورد استقبال مترجمان قرار گرفت

Figure 1. Syntactic performance of *Abadis Translator*

Semantic Evaluation

Table 4 demonstrated that the evaluators lacked consensus on the semantic adequacy of most narrative statements (65%) generated by *Abadis Translator*. The evaluators also had divergent views on the descriptive and argumentative statements. To be precise, they found nearly half of the argumentative statements (45%) to be incorrect and approximately one-third of the descriptive statements (35%) to be questionable. In general, they believed that *Abadis Translator* was able to produce more than half of the 60 statements correctly.

Table 4. Frequencies and percentages of the evaluators for semantic evaluation of the 60 statements

Type of Statements	Wrong		Inappropriate		Undecidable		Correct		Appropriate		*N*		
	f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*M*
Narrative	90	22.0	150	38.0	19	5.0	90	22.0	51	13.0	400	100.0	2.63
Descriptive	47	12.0	91	23.0	15	4.0	149	37.0	98	25.0	400	100.0	3.35
Argumentative	55	14.0	123	31.0	13	3.0	146	37.0	63	16.0	400	100.0	3.1
Total	192	16.0	364	30.0	47	4.0	385	32.0	212	18.0	1200	100.0	3.02

A one-sample Wilcoxon test was used to see how much the evaluators agreed on the meaning accuracy of each statement type. It is also run to find out if there was a meaningful link between their opinions and how well the Persian translations captured the original meaning. As Table 4 illustrates, the *p*-values for the descriptive and argumentative translations were below .05 ($p < .05$). This suggests a significant relationship between the evaluators' assessments and the semantic adequacy of these translations.

The average scores for these translations were also higher than the expected midpoint of 2.5, suggesting they were higher than average in terms of meaning accuracy: the descriptive translations had an average score ($MD = 3.35$), and the argumentative ones had an average score ($MA = 3.1$).

On the other hand, with the narrative sentences, the statistical significance level was above 0.05 ($p > 0.05$), indicating that the evaluators' opinions were not noticeably different from the expected midpoint. Additionally, the mean score for these statements was nearly equal to the theoretical mean/median ($2.63 \approx 2.5$), suggesting average semantic adequacy for the narrative statements.

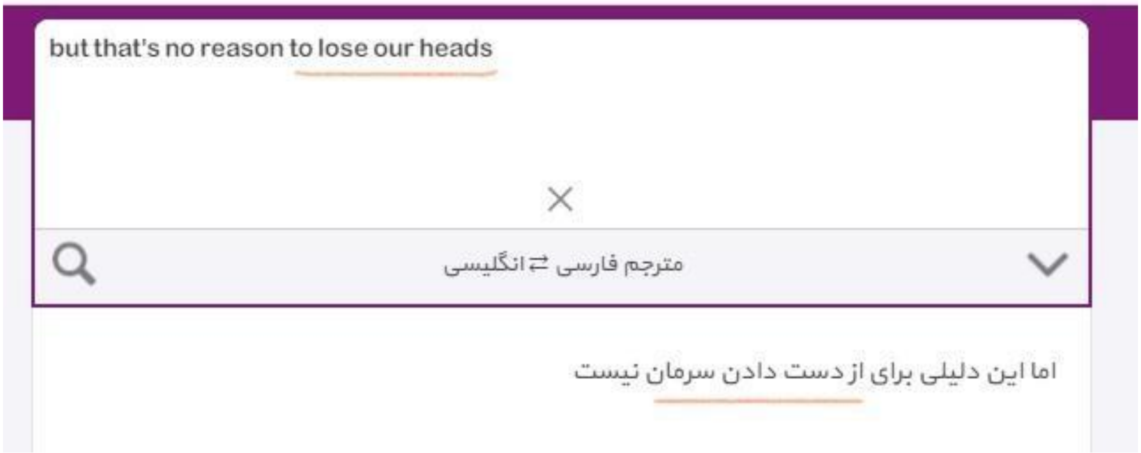
Moreover, Table 5 showed that the *p*-value for all 60 Persian translations was below 0.05 ($p < 0.05$). The overall mean score of 3.02 was higher than the theoretical mean/median of 2.5. This finding indicates that, overall, there was significant agreement among evaluators regarding the semantic adequacy of more than half of the Persian translations produced by *Abadis Translator*, demonstrating above-average semantic adequacy ($MT = 3.02$). This suggests that *Abadis Translator* was generally able to handle the semantic nuances of the 60 statements, particularly in the case of narrative statements.

Table 5. One-sample Wilcoxon signed-rank test for semantic evaluation of the 60 statements

Statement Types	*N*	*MDN*	*p*
Narrative	20	2	0.08
Descriptive	20	4	0.000
Argumentative	20	4	0.000
Total	60	4	0.003

Finding suitable equivalents for terms in the SL, especially in the case of idioms was another difficulty faced by *Abadis Translator*. This led to translations that were not just semantically inaccurate, but also awkward, to the extent that they made it more difficult for Persian readers to grasp the meaning. Figure 2 indicates that, in the idiom *but that's no reason to lose our heads,* the underlined phrase was translated into Persian literally as *از دست دادن سرمان* , which failed to reflect the figurative sense in which it is used. However, after it was post-edited by ChatGPT, it was changed to *خونسردیمان را از دست بدهیم*, which correctly conveyed the intended sense. This example illustrates the challenge that *Abadis Translator* faces when translating idiomatic expressions, as idiomatic translation demands a deeper understanding of both context and cultural subtleties.

Abadis Translator also made an error when translating the SL phrase *Mr. and Mrs. Dursley, of number four* into Persian as *آقا و خانم دورسلی، ازشماره چهار.* While it does translate as *آقا و خانم دورسلی، از شماره چهار*(see Figure 3). The latter explains, more contextually, where the Dursleys reside. *Abadis Translator* adhered to a rigid, literal translation instead of aiming for a more natural, meaning-based translation. ChatGPT's post-editing provided a more accurate version: *آقا و خانم دورسلی، ساکن شماره چهار*, which correctly conveyed that the Dursleys lived at that address. This example highlights the challenges *Abadis Translator* faces when translating phrases that need a more flexible, context-aware approach.

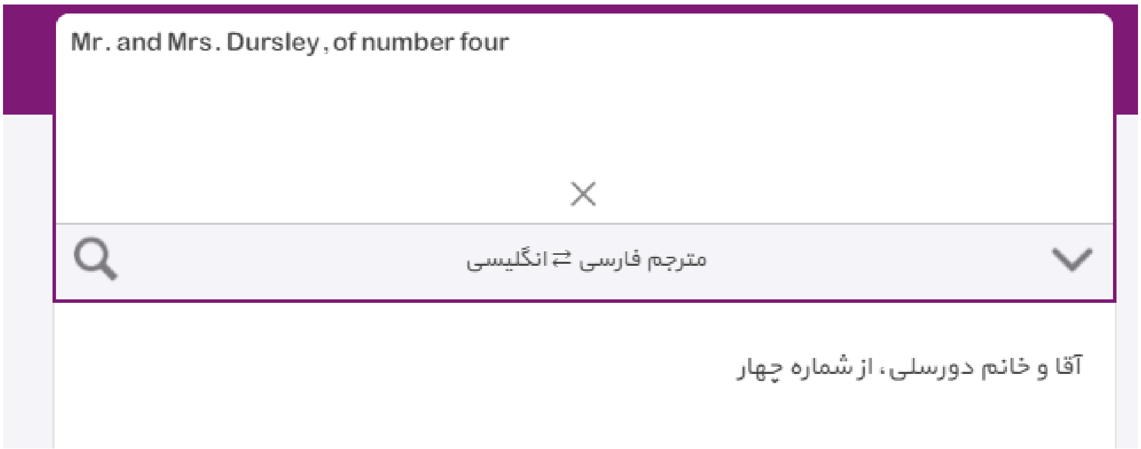


but that's no reason to lose our heads

مترجم فارسی ⇄ انگلیسی

اما این دلیلی برای از دست دادن سرمان نیست

Figure 2. Semantic performance of *Abadis Translator*



Mr. and Mrs. Dursley, of number four

مترجم فارسی ⇄ انگلیسی

آقا و خانم دورسلی، از شماره چهار

Figure 3. Semantic performance of *Abadis Translator*

Pragmatic Evaluation

According to Table 6, the evaluators expressed disagreement regarding the pragmatic correctness of all 20 idiomatic statements (100%) produced by *Abadis Translator.*

Table 6. Frequencies and percentages of the evaluators to pragmatic evaluation of the 20 statements

Type of	Wrong		Inappropriate		Undecidable		Correct		Appropriate		N		
Statements	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*M*
Idiomatic	320	80.0	80	20.0	-	-	-	-	--	-	400	100.0	1.59

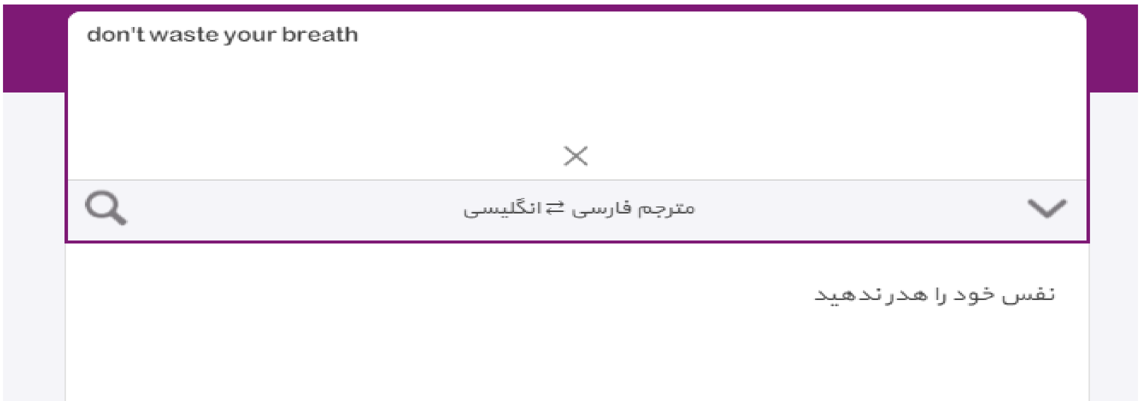
A one-sample Wilcoxon signed-rank test was conducted to assess whether there was a significant relationship between the evaluators' assessments and the pragmatic adequacy of the 20 Persian translations generated by *Abadis Translator.* Table 7 indicates that the *p*-value for the idiomatic statements was below 0.05 ($p < 0.05$). Additionally, the mean score for this type of statement was below the theoretical mean/median, suggesting that evaluators did not consider the idiomatic translations to be pragmatically adequate. That is to say, the MT system failed to capture pragmatic nuances, especially idiomatic expressions, because it focused on formal correspondence instead of functional equivalence.

In the pragmatic assessment, *Abadis Translator* demonstrated significant weaknesses. According to Figure 4, the idiom *don't waste your breath*, when translated literally, results in نفس خود را هدر ندهید, which is quite confusing for the target reader. The original idiomatic sense, meant to express the futility of arguing or discussing a certain point, was completely lost. The translation was far too literal, rendering it obscure within the flow of the conversation. These instances clearly highlighted a deficiency in psycholinguistic processing. As Wilss (1982) pointed out, the MT struggled to grasp the semantic nuances of underlying syntactic connections and could not produce sentences that would be natural sounding in a human translation.

Table 7. One-sample Wilcoxon signed-rank test for pragmatic evaluation of the 20 statements

Statement Types	*N*	*MDN*	*p*
Idiomatic	20	1	0.001

The translation *وقتان را هدر ندهید*, which ChatGPT post-edited, more accurately captured the meaning *don't waste your time*. This illustrates the difficulty *Abadis Translator* experienced in grasping and communicating the underlying message, a common challenge in pragmatic translation tasks.


don't waste your breath

نفس خود را هدر ندهید

Figure 4. Pragmatic performance of *Abadis Translator*

Discussion

The findings of this study shed light on the performance of the *Abadis Translator* in handling narrative, descriptive, and argumentative texts translated into Persian. Although the system showed reasonable syntactic accuracy, especially in terms of following basic grammatical structures like word order and subject-verb agreement, its capability to convey meaning accurately was less steadfast. This problem is not limited to the *Abadis Translator* but is a general drawback that MT systems face in their endeavor to render complex linguistic structures accurately, especially idiomatic expressions and context-dependent language.

The translation of idiomatic expressions has always posed considerable difficulty to MT research. Wilss (1982) explains that idioms rely on implicit meanings that require linguistic and cultural awareness. The study found that *Abadis Translator* quite frequently produced literal translations which left the meaning of idiomatic expressions obscure, for example, *a tough row to hoe*, meaning a difficult task, and *paddle his own canoe*, meaning to act independently, were literally translated into *دعوای سختی برای هو کردن* (a fierce argument for whistling) and *قایق خودش را پارو بزند* (row the boat itself), respectively. These mistranslations bear witness to certain fundamental limitations of statistical and pattern-based MT approaches in coping with non-literal language. Such results have also been documented by other MT studies (Abdi, 2021a; Aghai, 2024), thus making a case for the development of more contextually aware translation models.

Moreover, the intent of the study aligns with many of the wider concerns about why MT systems often fail to grasp pragmatic, cultural, and contextual nuances. The use of more literal translations rather than more context-aware forms of treatment accounts for a huge proportion of the errors made by the system. This limitation is not unique to *Abadis Translator*, as it is encountered by many MT systems, both domestic and international. More advanced neural models have difficulty capturing idiomatic language because these idioms and cultural references are so deeply embedded in the contextual meanings that statistical correlations alone can rarely catch them. Safari et al. (2023) pointed out the fact that Google Translate was superior to *Abadis Translator* in medical translation, suggesting that the quality of MTs varies across different domains, requiring specific tuning-based approaches for accurate results.

The study has also pointed out how human monitoring ensures the quality of a translation. MT may still be evolving, yet human input is vital mainly in post-editing tasks. The results seem to depict examples, such as ChatGPT, which, if well applied, would help in improving translation accuracy through correction of grammatical errors and in clarifying implied meanings. Although there were cases that ChatGPT relied on literal translation and produced inappropriate translations, these were corrected when it was asked to retranslate them. Nonetheless, this does not detract one bit from the credibility of ChatGPT as a powerful and robust AI editor for both end-users and translators.

From a developmental perspective, the study recommends that domestic MT developers should set their goals on improving grammatical precision of complex sentences and interpretation of implied meanings and idiomatic expression. For example, training models on linguistically diverse corpora; using contextual embeddings; and using feedback from users in the iterative cultivation of models would improve the veracity of translations. Wilss's (1982) observation of functional dependencies in translation comes into play here, whereby translated text depends not only on syntactical accuracy, but also on a correct rendering of intended meaning across linguistic and cultural barriers.

For translation students, research projects analyzing MT system performance will provide necessary insights into translation quality and post-editing problems. Being aware of recurrent MT errors, especially in the area of idiomatic expressions and contextual nuances, will enable students to increase their proficiency in post-editing while working with AI-assisted translation tools. These endeavors would improve MT accuracy while developing a more profound understanding of translation as a cognitive and cultural process.

CONCLUSION

The present study aimed to critically evaluate the output quality of *Abadis Translator*, a domestic MT system, focusing on syntax, semantics, and pragmatics. This evaluation utilized Wilss's (1982) matrix of translation evaluation to identify the system's strengths and weaknesses. This study is part of a larger project on the performance of Iranian software developers in developing nationwide MT systems: *Targoman Translator*, *Abadis Translator*, and *Fastdic Translator*, with special attention to translation quality and post-editing processes.

The results showed that the *Abadis Translator* performed well when it came to grammatical accuracy and was able to translate a large proportion of the 60 sentences into Persian; however, it was sometimes failing subject-verb agreement, which occurred particularly in more complicated sentences. Additionally, the results suggested that *Abadis Translator* was capable of producing translations that were somewhat semantically accurate. However, the pragmatic evaluation revealed that it frequently failed to provide accurate translations of idiomatic expressions and participle constructions, as all such statements were rendered inappropriately.

The primary weakness of *Abadis Translator* lies in its inability to manage implied meanings and select suitable equivalents for the SL items, which necessitate more advanced cognitive processing. Wilss (1982) underscores the importance of recognizing functional dependencies rather than merely formal structures for effective TL reproduction of such items. This limitation indicates that post-editing is essential for the translations produced by *Abadis Translator*. In conclusion, *Abadis Translator* performed well in syntax but struggled with semantics and pragmatics, especially idiomatic expressions. Moreover, ChatGPT improved translation accuracy, demonstrating its potential as a reliable AI post-editing tool.

The broader application of Wilss' (1982) Matrix in evaluating MT systems could provide a standardized, robust framework for assessing translation quality across different languages and technologies, promoting consistency in research and practice globally. In-depth research on AI post-editing, above all into ChatGPT, could revolutionize the translation industry considering it is one economical and expandable solution for improving the quality of translations, where skilled translators are restricted in number. Such developments would potentially set up an efficient and accurate global translation infrastructure that hopefully would be beneficial to businesses, governments, and individuals in multilingual settings.

This study is limited to *Abadis Translator* and ChatGPT, which may not represent all MT systems or AI tools, an instance that limits its findings' generalizability. The effectiveness of ChatGPT as post-editor can vary depending on the type of the text or the context, and the use of Wilss's (1982) matrix for the evaluation may exclude alternative frameworks or modern methods. The conclusion would provide a brighter perspective on strengths and weaknesses related to MT or, whichever the case may be, lacks the direct comparison with human translators. Furthermore, the current work does not capture the variety of texts that could render evidence of the nature of the challenges the process faces, and the revelations described could quickly become dated, with the ever-evolving nature of AI technologies.

Future research could investigate the various efficiencies of MT on other systems apart from *Abadis Translator,* with the influence of new AI technologies on translation quality and review of support-type editor systems perhaps to a comparative review of post-editing skills supported by ChatGPT against human translators across unique domains as a means of further depth for the strength and weaknesses that characterizes these systems. Further, improvements to evaluation frameworks, the addition of more methods, and the examination of the evolution of MT and post-editing over time would contribute to advancing the field.

REFERENCES

Abdi, H. (2016). *Translation and technology: A study of Iranian freelance translators.* Lambert Academic Publishing.

Abdi, H. (2021a). Considering machine translation (MT) an aid or a threat to the human translator: The case of Google Translate. *Journal of Translation and Language Studies*, *2*(1), 19-32. https://doi.org/10.48185/jtls.v2i1.122

Abdi, H. (2021b). Examining the appropriateness of Reiss's functionalist-oriented approach to trancism. *Theory and Practice in Language Studies, 11*(5), 561-567. http://dx.doi.org/10.17507/tpls.1105.15

Abdi, H. (2024). Negative analytic of the Persian translation of Atwood's *The Blind Assassin* using Berman's model of trancism. *International Journal of Philology and Translation Studies, 6*(1), 27-40. https://doi.org/10.55036/ufced.1434250

Aghai, M. (2024). ChatGPT vs. Google Translate: Comparative analysis of translation quality. *Iranian Journal of Translation Studies, 22*(85), 85-100. https://dorl.net/dor/20.1001.1.17350212.1403.22.1.9.2

Burke, A., & Maxwell, V. (2012). *Lonely planet: Iran (Travel Guide)*. Lonely Planet Publications.

Chan, S. W. (2004). *A dictionary of translation technology*. Chinese University Press.

Dalaslan, D. (2015). *An analysis of the English translation of Erzurum folk riddles in the light of Raymond Van den Broeck's translation criticism model* [Master's thesis, Hacettepe University Graduate School of Social Sciences]. Hacettepe University Digital Archive. http://www.openaccess.hacettepe.edu.tr/xmlui/handle/11655/1285

Defense Advanced Research Projects Agency. (1990). *Defense advanced research projects agency: Technology transition.* United States Department of Defense. https://apps.dtic.mil/sti/tr/pdf/ADA434135.pdf

Defense Language Institute English Language Center. (2011). *American language course - Book of idioms: Slang, special expressions, & idiomatic language*. Lackland Air Force Base.

Deng, X., & Yu, Z. (2022). A systematic review of machine-translation-assisted language learning for sustainable education. *Sustainability, 14*(13), Article 7598. https://doi.org/10.3390/su14137598

Dorr, B., Snover, M., & Madnani, N. (2010). Machine translation evaluation. In J. McCary, C. Christianos & J. P. Olive (Eds.), *Handbook of natural language processing and machine translation* (pp. 801-814). Springer.

Fauziah, I. M., & Putri, R. W. P. A. (2023). Book review of Zakaryia Almahasees's *Analyzing English-Arabic Machine Translation*. *International Journal of Arabic-English Studies, 23*(2), 443-444. https://doi.org/10.33806/ijaes.v23i2.471

Gala, J., Chitale, P., Raghavan, A. K., Gumma, V., Doddapaneni, S., Kumar, A., Nawal, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M., & Kunchukuttan, A. (2023). IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research, 12*, 1-95. https://doi.org/10.48550/arXiv.2305.16307

Han, L. (2018, September 19). *Machine translation evaluation resources and methods: A survey* [Conference presentation]. Ireland Postgraduate Research Conference (IPRC), Dublin, Ireland. https://arxiv.org/abs/1605.04515

Hirschman, L., & Thompson, H. S. (1997). Overview of evaluation in speech and natural language processing. In R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), *Survey of the state of the art in human language technology*. Cambridge University Press

Hutchins, J., & Somers, H. (1992). *An introduction to machine translation*. Academic Press.

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Aji, A. F., Wong, D., Liu, S., & Wang, L. (2024, May 20-25). A paradigm shift: The future of machine translation lies with large language models [Conference presentation]. *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation,* Torino, Italy. https://aclanthology.org/2024.lrec-main.120

Makhachashvili, R., Mosiyevych, L., & Kurbatova, T. (2023). Challenges of machine translation application to teaching ESP to construction students. *Social Sciences and Humanities, 3*, 1-11. https://doi.org/10.55056/cs-ssh/3/04005

Mercan, H., Akgun, Y., & Odacioglu, M. C. (2024). The evolution of machine translation: A review study. *International Journal of Language and Translation Studies, 4*(1), 104-116.

Mirzaeian, V. R., & Oskoui, K. (2022). Investigating Iranian EFL student teachers' attitude toward the implementation of machine translation as an ICALL tool. *Journal of English Language Teaching and Learning, 14*(30), 165-179. https://doi.org/10.22034/ELT.2022.52038.2496

Mistrík, J. (1997). *Štylistika*. Slovenské Pedagogické Nakladateľstvo.

Ntamwana, S., & Munandar, A. (2024). Translation criticism: Implementation of House's TQA model on Sweet Hour of Prayer into Indonesian. *k@ta, 26*(1), 62-74. https://doi.org/10.9744/kata.26.1.62-74

Nyberg, E., Mitamura, T., & Carbonell, J. (1994, August 5). *Evaluation metrics for knowledge-based translation*. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics* (pp. 95–99). Kyoto, Japan. https://aclanthology.org/C94-1013.pdf

Orr, D. B., & Small, V. H. (1967). Comprehensibility of machine-aided translations of Russian scientific documents. *Mechanical Translation and Computational Linguistics, 10*(1-2), 1-10.

Rowling, J. K. (1997). *Harry Potter and the sorcerer's stone.* Scholastic Press.

Safari, M., Pourhaji, M., Fathi-Alishah, S., Sajjadi, S., & Mohammadi, M. (2023). Translating medical texts from Persian to English: Accuracy of machine translation. *Archives of Advances in Bioscience, 14*(E43067), 1-6. https://doi.org/10.22037/aab.v14i1.43067

Saghayan, M. H., Ebrahimi, S. F., & Bahrani, M. (2021, May 18-20). Exploring the impact of machine translation on fake news detection: A case study on Persian Tweets about COVID-19. *Proceedings of the 29th Iranian Conference on Electrical Engineering (ICEE),* 540–544. https://doi.org/10.1109/ICEE52715.2021.9544409

Sartipi, A., Dehghan, M., & Fatemi, A. (2023, February 1). *An evaluation of Persian-English machine translation datasets with transformers*. arXiv preprint arXiv:2302.00321. https://doi.org/10.48550/arXiv.2302.00321

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open, 1*, 5-21. https://doi.org/10.1016/j.aiopen.2020.11.001

Tosun, S. (2024). Machine translation: Turkish-English bilingual speakers' accuracy detection of evidentiality and preference of MT. *Cognitive Research: Principles and Implications, 9*(10), 1-12. https://doi.org/10.1186/s41235-024-00535-z van Slype, G. (1979). *Critical study of methods for evaluating the quality of machine translation: Final report*. Commission of the European Communities Directorate General Scientific and Technical Information and Information Management. https://aei.pitt.edu/39751/

Wang, L. (2019). *Discourse-aware neural machine translation* [Doctoral thesis, Dublin City University School of Computing]. Dublin City University. https://doras.dcu.ie/22903

White, J. S., O'Connell, T. A., & Carlson, L. M. (1993, March 21-24). Evaluation of machine translation [Conference presentation]. *Human Language Technology Workshop*, Plainsboro, New Jersey. https://aclanthology.org/H93-1040

Wilss, W. (1982). *The science of translation: Problems and methods.* Gunter Narr.

Yan, Y., Wang, T., Zhao, C., Huang, S., Chen, J., & Wang, M. (2023). BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5428–5443. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.297

Zerva, C., & Martins, A. F. T. (2024). Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*, 12, 1460–1478. https://doi.org/10.1162/tacl_a_00711